11-10-2020

Statistical Modeling of Private Sector Participation in Disaster Risk Reduction Data

Wupeng Yin
wyin002@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

STATISTICAL MODELING OF PRIVATE SECTOR PARTICIPATION IN DISASTER RISK REDUCTION DATA

A thesis submitted in partial fulfillment of

the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

Wupeng Yin

2020

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This thesis, written by Wupeng Yin, and entitled Statistical Modeling of Private Sector Participation in Disaster Risk Reduction Data, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

Florence George

B. M. Golam Kibria

Wensong Wu, Major Professor

Date of Defense: November 10, 2020

The thesis of Wupeng Yin is approved.

Dean Michael R. Heithaus
College of Arts, Sciences and Education

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2020

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to Dr. Wu, who has supported me through every step of the thesis with her infinite patience and erudition. Dr. Wu has taught me not only the academic attitude and scientific method, but also a positive attitude towards life, especially in this pandemic situation. I also wish to thank the members of my committee for their support and guidance, Dr. George and Dr. Kibria. I wish to acknowledge Dr. Sarmiento and his team in the Extreme Event Institute at Florida International University for providing data and other useful sources of information.

I wish to thank Dean Donnelly for her wise suggestions and comments. I would like to convey my deepest gratitude to Dr. Zahedi, not only for his instructive wisdom and constructive comments, but also for his tremendous encouragement and support through my study.

I would like to thank Mery Castro for her efficient coordination help with the Dean's office. I also would like to thank Dr. Roudenko for her continuous support. I profoundly appreciate the support of the Department of Mathematics and Statistics. I am very grateful for having a great learning experience at Florida International University. I would especially like to acknowledge my friend Lauren Rodriguez and Francisco Diaz who have kept me motivated during this journey through shared laughter and camaraderie.

Finally, I thank Jinghua and my dog Cairo for their unconditional love and company throughout my studies.

ABSTRACT OF THE THESIS

STATISTICAL MODELING OF PRIVATE SECTOR PARTICIPATION IN DISASTER RISK REDUCTION DATA

by

Wupeng Yin

Florida International University, 2020

Miami, Florida

Professor Wensong Wu, Major Professor

The impacts of disaster on the private sector are inevitable, but their risks can be managed and reduced by preventively evaluative measures. Disaster risk reduction index (DRRI) and Disaster Experience (DE) variables were investigated in a survey study in six Western Hemisphere cities within the private sector of various business sizes. Our thesis built and evaluated 16 predictive models of DRRI with 36 categorical predictors and N = 1162 observations. Four statistical methods for linear regression and five for classification as well as seven machine learning methods were utilized. We also used stepwise selection and regulation methods for variable selection. They improved the performance of some models. To evaluate and compare the prediction performance among all models, we used resampling 5-fold cross-validation (CV) to estimate the true mean squared error (MSE) and classification accuracy. The results indicated that the neural network was outperformed among all the predictive models with the highest classification accuracy.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER I. INTRODUCTION

1.1 Background

All types of businesses can be impacted significantly by disasters (Asgary et al., 2012). Such disaster is "a serious disruption of the functioning of a community or a society at any scale due to hazardous events interacting with conditions of exposure, vulnerability and capacity, leading to one or more of the following: human, material, economic and environmental losses and impacts" (United Nations, 2016: 13). As a result of the COVID-19 pandemic, 41.3% of businesses were temporarily closed and 1.8% of businesses were permanently closed considering the sample draws from US-based business (Bartik et al., 2020). As many unprecedented disasters are becoming frequent, businesses have no choice but to confront disaster- induced direct and indirect losses and consequently try to find the appropriate business continuity plans (Asgary, 2016). To get a better understanding how well the private sector is going to get ready for the future impact of disasters, a survey study on private sector participation in Disaster Risk Reduction was conducted in six Western Hemisphere cities in 2012 (Sarmiento et al., 2012). A secondary data analysis (Sarmiento et al., 2019) investigated the relationship between disaster experiences and business readiness capabilities.

Many statistical methods are being constantly developed to control for the variables aiming at improving the better understanding of the model relationships and widely used for prediction purposes under certain criteria of optimal fit (Dawes, 2001). Statistical methods use mathematical models and techniques to help improve the estimates of

uncertainty analysis. The application of statistical methods is using different models to extract information from the predictors in the high dimensional data set and provides access to the model robustness and prediction accuracy of the targeted responses (Datta-Gupta & Mishra, 2017).

Descriptive statistics is still a valuable and substantial method to summarize and overall describe the data set. Sarmiento et al. (2012) provided descriptive statistics analysis in detail on the disaster risk reduction data set and later, they adopt the classical linear regression to the same data set with the purpose of developing the relationship between business' disaster experience and the disaster risk reduction (Sarmiento et al., 2019). Through the reported result of using descriptive statistics along with typical linear methods for regression on such a high dimensional data set, we assumed that in addition to linear regression, the logistic regression approaches might be applicable to this data set and bring enhancements on the model performance and accuracy under different model selection and regulation methods. We explored the complex relationships between the predictors and the response in the data by evaluating different statistical models. On the other hand, we put our efforts on the performance of the different statistical modeling approaches and showed that whether they serve the data set well or outperform the classical linear regression in that way by comparing their prediction accuracy.

The data used in this study were generously provided and originally developed by Dr. Sarmiento and his team in the Extreme Event Institute of Florida International University.

1.2 Literature Review

Risk management has gradually been understood and the way disaster risk has been approached has significantly evolved over the past half-century. According to the researches of Sarmiento et al. (2012, 2019), they focused on which and how much impact disaster experiences may impose on the readiness capabilities, considering different business sizes and various city locations of the private sectors.

The original study of private sector participation in Disaster Risk Reduction was conducted from June to November of 2012 in six Western Hemisphere cities: Bogotá, Colombia; Kingston, Jamaica; Miami, Florida, USA; San José, Costa Rica; Santiago, Chile; and Vancouver, British Columbia, Canada (Sarmiento et al., 2012). The survey interviewed senior managers, personnel, or directors of private sector companies with the questionnaire involved in three main sectors, and it resulted in 1197 responses on 210 question items.

In a secondary study (Sarmiento et al., 2019), seventeen disasters experienced (DE) by businesses and three disaster readiness indexes had been extracted and measured from the raw data of the survey. The DEs include supply chain disruption, power outage, damaged facilities/ equipment/ inventories, etc. The three disaster readiness indexes are defined as DRRI (Disaster Risk Reduction Index), BCI (Business Continuity Index), and CSRI (Corporate Social Responsibility Index), respectively. The DRRI corresponds to the value of measures taken to control risks and reduce potential damage and losses as a result. The BCI values the measures taken to ensure business safety and continuity of time-sensitive

operations. The CSRI corresponds to the value business commitment to contribute to economic development, quality of life of the workforce, their families, and the local community. The descriptive statistics, multivariable linear regression, and stepwise model selection were applied to the data set with six predictors and disaster readiness index (DRRI) as the dependent variable. It concluded that business size played an important role in the models and disaster experience has a positive effect on the response.

The evaluation of the original linear regression approach in the study of Sarmiento et al. (2019) is valuable for providing an overall inspection of the data set. Accordingly, from the aspect of statistics, we believed that more information could be explored and extracted from the data set by using various statistical methods. As a matter of fact, we consider all DEs as a whole along with the interaction terms as the predictors instead of choosing six of DEs before fitting them to the linear model. In addition, some statistical modeling approaches and machine learning methods were utilized to achieve the integral comprehension of the data set.

1.3 Data Preparation and Research Aims

Data Preparation

We used the seventeen induced DE variables as well as business size and city location as predictors while DRRI is used as the response. In the previous study (Sarmiento et al., 2019) has revealed that business size has an impact on the relationship between disaster experiences and the responses. Cox (1984) noted that “Large component main effects are more likely to lead to appreciable interactions than small components. Also, the

interactions corresponding to larger main effects may be in some sense of more practical importance." In consequence, we have a compelling reason to include the interactions between seventeen DE variables and business size as predictors as well.

Therefore, the overall organization of our data set composed of 36 predictors, 1 independent response, and 1197 observations in the raw data set. All predictors are categorical: DE variables are of 0/1 scale, business size is of three levels (small, medium and large), and city locations are of six levels. The response variable was calculated by four items extracting from the survey with equal increment in 0.25 each, taking values of 0, 0.25, 0.5, 0.75, and 1. Although the response DRRI is quantitative, it can also be categorized as a factor with five levels because the previous study did not justify the equal increments. It is noticed that there were a few missing values in DRRI, which are 0.7% in city location and 2.9% in business size. Since the percentage of the missing values is relatively small, we deleted the corresponding observations of any missing values listwise and prepared the data set with 1162 complete observations.

The abbreviation of the 17 DEs', business size, and city were defined in the thesis:

* LI: Loss of IT
* SCD: Supply chain disruption
* Dea: Deaths
* LAS: Loss of access to site
* EC: Extreme conditions (high/low temperatures, flood/high winds)
* DC: Damage to corporate image/reputation/brand

* LTC: Loss of telecommunications
* PG: Pressure groups
* PO: Power outage
* IA: Industrial action
* WO: Water outage
* EI: Environmental incident
* CH: Customer health/product safety issue/incident
* LKSP: Loss of key skills and personnel
* NP: Negative publicity/coverage
* DF: Damaged facilities/equipment/inventories
* OT: Other
* City: City locations
* BS: Business size

Research Aims and Objectives

In this age of information, people have an intensified desire to use readily available information to make decisions for future events such as “Do I need the umbrella today?” “When is a good time to invest in real estate?” or “Who should get the COVID-19 vaccine first?” Predictive modeling is a process that uses statistical methods to generate, process, and validate a model that helps us to make the best decisions by forecasting future outcomes. Prediction accuracy is usually used to guide the decisions. Comparisons

of predictive accuracy provide an understanding of the robustness among the competing predictive models.

As we look into the dependent variables of the data set, they can be treated as quantitative or categorical. Therefore, both regression and classification models can be applied. Effective and widely used traditional statistical methods applied for prediction include Multiple Linear Regression (Efroymson, 1960; Garside, 1965; Andrews, 1974); Multinomial Logistic Regression (Engel, 1988; Böhning, 1992), Ordinal Logistic regression (McCullagh, 1980; Winship et al., 1984), and Linear Discriminant Analysis (Fisher, 1936; Friedman, 1989) for classification. In addition to the aforementioned linear methods for regression and classification, machine learning methods such as Random Forest (Ho, 1995), Support Vector Machine (SVM) (Cortes & Vapnik, 1995), Neural Networks (NN) (Hopfield, 1982, 1984), See5/C5.0 (Quinlan, 1993), Stochastic Gradient Boosting (SGB) (Friedman, 2002) and k-Nearest Neighbors algorithm (k-NN) (Altman, 1992) can be applied to this dataset for the predictive purposes. Random forest is an algorithm that combines bagging with random feature selection and focuses on ensembles of decision trees. The support vector machine uses multidimensional surfaces to define the relationship between predictors and responses. The neural network's concepts are borrowed from an understanding of human brains to model arbitrary functions (Lantz, 2013). The See5/C5.0 is a more advanced version of C4.5 algorithm which is used to generate decision trees by using the concept of information entropy (Quinlan, 1993). The stochastic gradient boosting is a variation of the boosting approach to regularization of boosting models based on decision trees with random sampling at each iteration

(Friedman, 2002). The k-NN is a non-parametric approach consists of the k-closest samples from the training set (Kuhn & Johnson, 2013). All these methods result in nonlinear predictive functions.

The 36 predictors may or may not contribute to the predictive models. Excluding or minimizing the effects of the variables which are less contributed to the model may improve the prediction accuracy and model interpretability (James et al., 2013). Within the framework of linear regression modeling, some approaches can perform variable selection among high dimensional predictors, such as stepwise selection (Efroymson, 1960), the Lasso (Santosa & Symes, 1986; Tibshirani, 1996) as a shrinkage method, cross-validation (Allen, 1974; Stone, 1974; Stone, 1977), and principal component regression (Kendall, 1957; Spurrell, 1963; Massy, 1965) as a dimension reduction method. After important variables were selected, all statistical models were finalized, fitted, and interpreted using the entire dataset.

The performance of the predictive models could be assessed in cross-validation, where is a resampling approach involves randomly dividing the observations into k folds of approximately equal size with k-times repetitions. Each time the procedure uses a diverse fold and treat it as a validation set, then computes the mean squared error on the k-1 folds. The prediction error could be calculated via mean squared error and percent of incorrect predictions. The cross-validated prediction error and its standard error for all models were estimated and compared numerically. We also used stacking (Wolpert, 1992), an ensemble machine learning algorithm that is used to combine the predictions from the bottom layer models with the purpose of enhancing the predictive performance.

In summary, the research objectives of the thesis are:

* Objective 1: Build predictive models by statistical methods and machine learning methods.
* Objective 2: Apply model selection and regulation for better prediction accuracy and easier interpretation.
* Objective 3: Assess the performance of the predictive models with the resampling approach of k-fold cross-validation.

All data analyses in the thesis were implemented with RStudio (version 1.3.1073) for Mac (RStudio Team, 2020).

The organization of the rest of the thesis was as follows: Chapter II provides the mathematical descriptions of the statistical models and machine learning methods. Chapter III focuses on analyzing the real data by using the models and methods we illustrate in Chapter II. Chapter IV explains and interprets the models and the results we obtain from the data analysis in Chapter III. Finally, we summarize our work in the thesis and point out the future works in Chapter V.

CHAPTER II. STATISTICAL METHODOLOGIES

In Chapter II, we will offer an overview of all the methods or approaches we are going to use in the thesis. There are four main sections in Chapter II. The first section illustrates the statistical approaches for the classical linear regression, followed by the statistical methods for classification in the second section. The statistical methods list in these two sections are linear approaches. On the other hand, machine learning methods have flourished impressively over the past decades, so we decide to implement and compare them to the statistical methods. As the preference of the thesis is to put the emphasis on the statistical methods instead of the machine learning methods, we briefly enumerate seven machine learning methods in section three. Model comparison methods take part in the last section corresponding to the purpose of understanding the model performance.

2.1 Linear Methods for Regression

Linear regression is a simple approach for supervised learning and a useful tool for predicting a numeric response. A linear regression model assumes that the relationship between the predictors and the mean of the response is linear. For prediction purposes, though it might be considered less complex than some of the contemporarily nonlinear models, linear regression is still a widely used statistical method providing an efficient and interpretable description of how the predictors affect the response/responses (Hastie et al., 2001).

There are four principal assumptions associated with a linear regression model (Tamhane & Dunlop, 2000):

a. Linearity: The mean of the response is a linear combination of the predictors.
b. Normality: The errors are normally distributed.
c. Statistical independence: The errors are uncorrelated with each other.
d. Homoscedasticity: constant variance of the errors.

In this section, we will discuss the multiple linear regression, the principal components regression as a dimension reduction method, AIC-based forward-stepwise variable selection, and the Lasso as a regularization method. These methods serve the purpose of the data analysis in Section 3.1, where the dependent variable under the consideration of quantitative output. Among the four methods, AIC-based forward-stepwise variable selection and the Lasso can select important variables as in Objective 2.

2.1.1 Multiple Linear Regression

Multiple linear regression is a statistical method that attempts to determine the linear relationship between more than one predictor and a quantitative response. Given a $n \times p$ matrix **X** and quantitative response Y, in general, the multiple linear regression model takes the form (Hastie et al., 2001):

$$f(X) = E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \qquad (2.1)$$

Where the input vector $X^T = (X_1, X_2, \ldots, X_p)$. The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in (2.1) are unknown parameters and must be estimated. The jth predictor X_j can be various forms obtained from different sources (Hastie et al., 2001). Although X_j must be

the quantitative inputs, transformations of the quantitative input, basis expansion of the quantitative inputs and interaction terms between the quantitative inputs are all welcomed.

As reviewed in Chapter I, the study of Sarmiento et al. (2019) revealed that business size seemed to be associated with disaster experiences. To include the interaction effect of business size on disaster experiences in our predictive model, we added the interaction between business size and 17 DEs (introduced in Chapter I) to extend the model. In general, consider a multiple linear regression model with only two predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{2.2}$$

Where $\varepsilon \sim N(0, \sigma^2)$ is a random error term. The model with the inclusion of the interaction term, which is the product of X_1 and X_2, is still linear (James et al., 2013):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \tag{2.3}$$

Typically, the estimation method of least square approach is used to estimate the parameters. Hence, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are chosen to minimize the residual sum of squares (Hastie et al., 2001):

$$RSS(\beta) = \sum_{i=1}^{N} \big(y_i - f(x_i)\big)^2 = \sum_{i=1}^{N} \left(y_i - \beta_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{2.4}$$

Where y_i denotes the ith observation of Y, x_{ij} denotes the (i, j)th element in $\mathbf{X}$. We can rewrite the $RSS(\beta)$ in the vector form:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{2.5}$$

The $\mathbf{X}$ denotes the $n \times (p+1)$ matrix with additional columns added to the input matrix as the first column with all 1s. Here $\mathbf{y}$ denotes the n-vector of outputs. The least square estimator of β that minimizes $RSS(\beta)$ is:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.6}$$

Predicted values in vector form, where $\mathbf{H}$ is the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y} \tag{2.7}$$

Now we can make the interpretation using the formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{2.8}$$

Each $\hat{\beta}_i$ is an unbiased estimator of β_i: $E\left[\hat{\beta}_i\right] = \beta_i$. The fitted output vector $\hat{y}$ is a linear combination of the column vector x_i's.

In interpreting this model to describe the ith row coefficient for the predictor x_i, we would say "A one-unit increase in the predictor x_i would yield a $\left|\hat{\beta}_i\right|$-unit increase/decrease in the predicted $\hat{y}$. The increase or decrease is based on the positive/negative sign of the coefficient $\hat{\beta}_i$.

2.1.2 Stepwise Variable Selection

In statistics, the process of determining which predictors contribute most to the model and then selecting a subset of relevant variables is becoming more critical to serve the high-dimensional data analysis. A classic example is stepwise selection (Efroymson, 1960) for linear regression. The direction of stepwise selection can be forward, backward, or both. Since it requires finding the least square estimates of all candidate models, the

forward-stepwise selection, which consider the relatively simpler models, is the least computational expansive. Therefore, we will focus on forward-stepwise selection. Each step of predictor evaluating in the forward-stepwise selection procedure is using the preassigned base model. The detailed iterations of the forward selection are:

1. Create a null model contains only the interception term;
2. Add the predictors one-at-a-time until all candidate predictors are included in the model and evaluate these sub-optimal models with the smallest residual sum of squares (RSS) or the highest R-squared;
3. Choose the final optimal model using AIC score, BIC score or other criteria.

The Akaike information criterion (AIC) is a commonly used criterion. It was first formulated by Hirotugu Akaike. The Akaike information criterion (AIC) is a penalized version of the maximized log likelihood function of the residual sum of squares (RSS) (Kuhn & Johnson, 2013):

$$\text{AIC} = n \log\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + 2d \tag{2.9}$$

Where d is a measure of model complexity. The Akaike information criterion (AIC) becomes large when training error is large, or the model is overfitting and too complex. Hence, we want to choose the final optimal model with the smallest AIC score. We will adopt AIC based forward-stepwise selection in the data analysis in Chapter III. Forward-stepwise selection is fast computationally and no limitation on the dimension of the data set, and have lower variance, but may increase the risk of getting more bias.

2.1.3 The Lasso

In statistics, the lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996) is a regression analysis method that could shrink the regression coefficients towards exactly zero by using the L_1 penalty. In this case, the lasso performs the variable selection just like the forward-stepwise selection we introduced above.

The lasso minimizes RSS with the L_1 penalty:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P}|\beta_j| = RSS + \lambda \sum_{j=1}^{P}|\beta_j| \tag{2.10}$$

Where $\sum_{j=1}^{P}|\beta_j|$ is the L_1 lasso penalty and $\lambda \geq 0$. The standardized tuning parameters of the lasso coefficients are $s = t/\sum_1^p|\hat{\beta}_j|$. There is one-to-one mapping between parameter λ and t. When $s = 1$, the lasso coefficients are the same with the least squares estimates (Hastie et al., 2001).

In order to choose the optimal tuning parameter s, we adopt k-fold cross-validation (CV) and one-standard-error rule. We would introduce details of CV as a resampling method for model comparison in Section 2.4. We focus on the selection process for now:

1. Split the whole data set into training and test set.
2. Calculate the CV's mean squared error of the test set for each candidate of s.
3. Plot the CV's mean square error (MSE) of the test set versus all the candidates of s and find the smallest MSE (blue line, Fig.1) on the curve.

4. Select the optimal tuning parameter s (green line, Fig.1) using the CV's MSE on the curve within one standard error (red line, Fig.1) on the left of the smallest one.

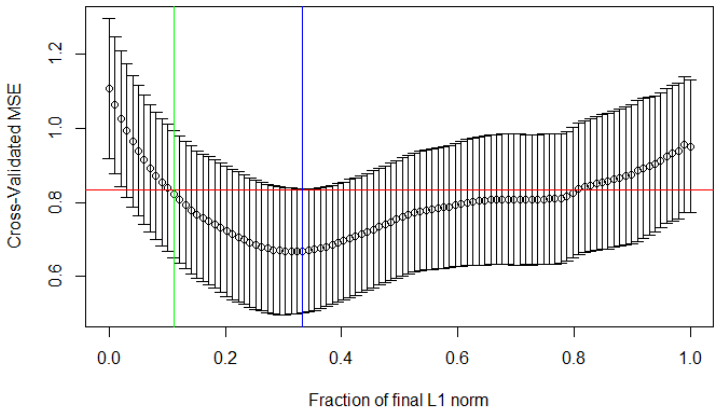

Figure 1: CV MSE vs. Tuning Parameter in Lasso.

The lasso approach usually provides relatively less prediction error with less computationally cost.

2.1.4 Principal Components Regression

When doing predictions on the big, real-world data, by looking at the dictionary defining what the predictors are, we notice that the predictors sometimes could be correlated with each other over the redundant information collecting. The impact of the multicollinearity issue could increase the variability and vitiate the stability of the performance on the ordinary least square solution for the multiple linear regression (Kuhn & Johnson, 2013). One widely applied approach to this problem is principal component regression (PCR) (Massy, 1965), which is using the principal component analysis (PCA) first for pre-processing in order to obtain the uncorrelated combinations of the original predictors and then use them as the predictors to perform regression.

Principal components regression (PCR) is a two-stage dimension reduction method (Kuhn & Johnson, 2013). The first stage is using PCA, an unsupervised approach to extract the low-dimensional representation of the data set, that is, fewer principal components, that capture as much as possible the of the variation in the predictors. In this stage, the response does not have any influence on PCA, and the principal components are orthogonal to each other. Stage two is to select enough principal components as independent predictors and perform the regression analysis.

The variables in the data set should be centered and scaled beforehand. Assume we have a data matrix X with n observations and p features, then the first principal component of X is the linear combination of the predictors:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip} \tag{2.11}$$

That maximized the sample variance:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{maximize}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2\right\} \tag{2.12}$$

subject to $\sum_{j=1}^{p}{\phi_{j1}}^2 = 1$, where the elements $\phi_{11},\ldots, \phi_{p1}$ are the loadings and $z_{11},\ldots, z_{n1}$ are the scores of the first principal component (James et al., 2013). Then the kth principal component can be defined and uncorrelated to all previous ones.

$$z_{ik} = \phi_{1k}x_{i1} + \phi_{2k}x_{i2} + \cdots + \phi_{pk}x_{ip} \tag{2.13}$$

That to maximize the sample variance of the n values of z_{ik}:

$$\underset{\phi_{1k},\ldots,\phi_{pk}}{maximize}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jk}x_{ij}\right)^2\right\} \tag{2.14}$$

The problem in Equation (2.12) can be solved by the eigen decomposition in linear algebra, but we do not discuss the details in the thesis.

We can interpret PCA with the above linear relationship and plot the principal component scores for better visualization. Moreover, the proportion of variance explained by the kth principal component is defined (James et al., 2013):

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jk}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}{x_{ij}}^2} \tag{2.15}$$

We usually look into the cumulative proportion of variance explained by the principal components to decide how many components we would like to use for the regression. The basic idea is to choose the less amount of the principal components which can explain a desirable amount of variation. So, PCR provides a dimension reduction solution to high dimensional linear regression, but it does not select variables.

As illustrated above, the response of the data set is absent in PCA approach, so the PCR may have difficulties to detect a predictive relationship related to the response variability (Kuhn & Johnson, 2013).

2.2 Linear Methods for Classification

When we deal with the real-world data, we notice that the responses (or dependent variables in a classical way) are usually in different types, and the common types of responses are typically presented in quantitative or qualitative measurements.

In the above section, we have illustrated the linear regression for the quantitative response. An alternative way to regress the data set is to represent the response numerically by codes and assign them to K classes or categories (Hastie et al., 2001). Although it would be fine if we convert a binary response into quantitative and mapping the output back to the 0-1 categories after the regression, this regression approach may not be able to find the correct boundaries for more than two classes response.

In this context, we decided to convert the quantitative response of the data set into qualitative variable because it was a discrete variable with the values of 0, 0.25, 0.5, 0.75 and 1, and we believe there is an ordering between these values. For these reasons, it is preferable to use the classification approach on our data set such as multinomial logistic regression, ordered logistic regression and Linear Discriminant Analysis (LDA). Also, we use the subset selection and shrinkage fitting methods for better prediction accuracy and model performance as in Objective 2.

2.2.1 Multinomial Logistic Regression

Multinomial logistic regression is considered as the multiple-class extension of binomial logistic regression that the model contains K-1 terms that provide the posterior

probabilities of each K class by using the linear relationship with the predictors. We have the model (Hastie et al., 2001):

$$\log\frac{Pr(G=1|X=x)}{Pr(G=K|X=x)}=\beta_{10}+{\beta_1}^T x$$
$$\log\frac{Pr(G=2|X=x)}{Pr(G=K|X=x)}=\beta_{20}+{\beta_2}^T x$$
$$\vdots$$
$$\log\frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)}=\beta_{(K-1)0}+{\beta_{(K-1)}}^T x \quad (2.16)$$

Where *G* denotes the qualitative outputs. We show the model in K-1 log-odds terms with the probabilities sum to exactly one and fit the model by maximizing the likelihood function under the 0-1 loss (Hastie et al., 2001):

$$\ell(\theta)=\sum_{i=1}^{N}\log p_{g_i}(x_i;\theta) \quad (2.17)$$

Where $p_k(x_i;\theta)=Pr(G=k|X=x_i;\theta)$ and $\theta=\{\beta_{10},{\beta_1}^T,\dots,\beta_{(K-1)0},{\beta_{(K-1)}}^T\}$, and g_i denotes the ith observation of *G*. The x_i is the ith observed value of the input variables that is defined as a scalar or vector with $i=1,2,\dots,N$, given the $N\times p$ input matrix. The coefficients are unbiased and estimated by maximizing the likelihood function which is optimized by an algorithm called iteratively reweighted least squares (IRLS).

By using the multinomial logistic regression, we assumed that our dependent variable is categorical without any order. In the next section, we will introduce the ordinal logistic regression because the dependent variable in our data set shows that there is an ordering among the values.

2.2.2 Stepwise Variable Selection

As we mentioned in Section 2.1.3, we want to use the alternative fitting procedure to yield better prediction accuracy and model interpretability than the multinomial logistic regression. We have stated the steps of the stepwise selection in Section 2.1.3 for linear regression. For classification, we choose to use the same AIC statistic for model selection (Hastie et al., 2001):

$$\text{AIC} = -\frac{2}{N} \cdot \text{loglik} + 2 \cdot \frac{d}{N} \tag{2.18}$$

Where the "loglik" refers to the maximized log-likelihood:

$$\text{loglik} = \sum_{i=1}^{N} \log Pr_{\hat{\theta}}(y_i) \tag{2.19}$$

We choose the optimal model with the minimum AIC over the candidate models.

2.2.3 Elastic Net

Zou and Hastie (2005) introduced the elastic-net penalty, a regularized regression method that compromises between the L_1 of the lasso and L_2 of ridge. The elastic-net penalty can be used for regression or classification model.

For the multinomial problem, the elastic-net penalty has the form (Hastie et al., 2001):

$$\max_{\{\beta_{0k},\beta_k \in \mathbb{R}^p\}_1^K} \left\{ \sum_{i=1}^{N} log\ Pr(g_i|x_i) - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\alpha|\beta_{kj}| + (1-\alpha){\beta_{kj}}^2\right) \right\} \tag{2.20}$$

Where β_{kj} denotes the (k, j)th element in the coefficients set. The advantage of elastic net model is that the L_1 penalty (Lasso) enable the effect of carriable selection while the L_2 (ridge) penalty shrinks the coefficients towards 0. The value of the parameter α enables effective regularization bridges between the pure lasso penalty (when $\alpha = 1$) and a pure ridge-type penalty (when $\alpha = 0$) (Kuhn & Johnson, 2013). The total amount of penalization has been controlled by another tuning parameter λ. When $\lambda = 0$, then there is no regularization in the model.

2.2.4 Ordinal Logistic Regression

In statistics, the ordinal logistic regression (also known as ordered logistic regression or ordered logit model) is a sub-type of the logistic regression for an ordinal dependent variable. Ordinal logistic regression model has been first considered since the 1980s (McCullagh & Nelder, 1989) but becomes popular recently in many fields.

In the model, the observed ordinal response Y is a function of another continuous laten variable Y^* with various threshold values μ_j, that is, Y^* is observed in a discrete form of Y through a censoring mechanism (Echaniz et al., 2019):

$$Y_i = 0 \;\; if \;\; Y_i^* \leq \mu_0$$

$$Y_i = 1 \;\; if \;\; \mu_0 < Y_i^* \leq \mu_1$$

$$\vdots$$

$$Y_i = J \;\; if \;\; \mu_{J-1} < Y_i^* \leq \mu_J \tag{2.21}$$

The model estimates $Z_i = \sum_{k=1}^{K} {\beta_k}^T X_{ki} = E(Y_i^*)$.

Under the assumption of parallel slopes (Borooah, 2001), the ordinal logistic regression model can be written as:

$$P(Y_i > j \mid X_i) = \frac{\exp\left(\beta^T X_i - \mu_j\right)}{1 + \left\{\exp\left(\beta^T X_i - \mu_j\right)\right\}} \ , \ j = 1,2 \dots, J-1 \qquad (2.22)$$

Estimate the above model by maximizing the log-likelihood function under the 0-1 loss (Echaniz et al., 2019):

$$\log L = \sum_{i=0}^{N} \sum_{j=0}^{J} m_{ij} \log\left\{F\left(\mu_j - {\beta_k}^T x_{ki}\right) - F\left(\mu_{j-1} - {\beta_k}^T x_{ki}\right)\right\} \qquad (2.23)$$

Where m_{ij} is a binary variable, $m_{ij} = 1 \ if \ Y_i = j$ and $m_{ij} = 0$ otherwise.

The ordinal logistic regression has been widely used and has been developed in many statistical software packages. However, the fitting progress may encounter problems with the large numbers of the predictors because too many parameters is needed to evaluate and set as the initial parameters beforehand.

2.2.5 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a statistical approach to model each class density as multivariate Gaussian and assumes that the classes have a common covariance matrix Sigma. When we compare LDA to logistic regression, the LDA is more stable than the latter if the sample size is small and the distributions of the independent variables are

approximately normal. Moreover, the LDA has been popularly used for the multi-class classification.

As mentioned above, the LDA assumes $\Sigma_k = \Sigma$ for all $k = 1, \dots, K$. Variances and correlations are the same across all classes with only the center locations are different, that is, the class Gaussian distributions are shifted versions of each other. We have the linear discriminant function (Hastie et al., 2001):

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k{}^T \Sigma^{-1} \mu_k + \log(\pi_k) \tag{2.24}$$

Estimated the model:

$$\hat{G}(x) = \arg\max_k \delta_k(x) \tag{2.25}$$

The LDA methods approximates the Bayes classifier by estimating the following parameters with the training data set (Hastie et al., 2001):

* $\hat{\pi}_k = N_k/N$, where N_k is the number of class-k samples.
* $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$, the sample mean vector of class-k samples, where x_i denotes the ith training sample.
* $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N - K)$, the pooled covariance matrix.

For two-class problem, LDA is the same as linear regression of indicator response matrix but the intercept may be different. For more than two-class the solutions are different.

Linear Discriminant Analysis (LDA) could be powerful when the data set has met the assumption. However, it seems like the LDA makes unrealistic assumptions about the real-world data (multivariate Gaussian of the set of independent variables, homoscedasticity) and is also very sensitive to outliers.

2.3 Machine Learning Methods and Ensemble Learning

As information and computational capacity have become more easily available in decades, we are more desirable to use tools to find the patterns and make decisions using big data. The process of study interested in the development of these tools is known as machine learning (Kuhn & Johnson, 2013). Here, we introduce a few machine learning methods that we choose to use for our data. These methods are C5.0, gradient boosting machines (GBM), K-Nearest Neighbors (KNN), neural network (NN), random forest, and support vector machines (SVM). It is a practical way to process the data set with these non-linear methods and compare their output with the linear methods we considered in the previous sections. However, we only intend to focus on the result of using these methods instead of the algorithms themselves, so the theoretical mathematics of the machine learning methods is beyond the scope of the thesis. Moreover, we explicate ensemble learning which is an approach to improve the prediction performance of the machine learning methods.

All the machine learning approaches we illustrate in the thesis can be used for regression as well as for classification problems.

2.3.1 C5.0

The C5.0 algorithm is an improved version of C4.5 algorithm (Quinlan, 1993) that provides a tree-based and rule-based model. It performs well for most problems and is easy to understand and use. To build decision trees, C5.0 uses the concept of entropy or information gain as the splitting criterion. Then C5.0 post-prunes the decision tree. The process is to grow a tree large enough to overfit the training data and then remove the nodes and branches that have little effect on the classification errors to reduce the size of the decision tree to a more appropriate level (Lantz, 2013).

2.3.2 Gradient Boosting Machines (GBM)

Gradient boosting machines (GBM) was first called by Friedman (Friedman 2001), a typically tree-based boosting method. The principal idea of GBM is to seek an additive model to minimize the loss function, given a loss function and a weak learner.

The GBM algorithm usually initialized the response with the mean of the response in regression or the sample log-odds for classification. After the opposite of the gradient of the fitting error has been calculated, a model is used to fit the pseudo-residuals as the outcome in order to minimize the loss function. Finally, the current model is added to the previous model weighted by the optimal step size with repetitive iterations. The simple GBM has two tuning parameters (tree depth and a number of iterations) when trees are used as the base learner (Kuhn & Johnson, 2013).

2.3.3 K-Nearest Neighbors (KNN)

The k-nearest neighbors algorithm (KNN) is one of the simplest supervised machine learning algorithms proposed by Thomas Cover (Altman, 1992). As a model-free method, the KNN uses the k-closest training samples to predict a new sample by measuring the distance between two samples. Euclidean distance and Minkowski distance are the most commonly used metrics. In KNN regression, the predicted response assigned to the new sample is the mean or median of the values of k neighbors' response. In KNN classification, the predicted response is classified by the majority vote of its k-nearest neighbors (Kuhn & Johnson, 2013).

Generally, the standardized predictors are required because the performance of KNN approach relies on the distance between samples. The resampling methods can be used to determine the tuning parameter *k*. Like the procedure to find the optimal tuning parameter for the lasso, the optimal *k* can be achieved across the candidate values with the minimum MSE.

2.3.4 Neural Networks (NN)

Neural networks (NN) (Bishop 1995) are one of the most popular machine learning methods that use the multi-layer networks of neurons to mimic how the human brain works. The hidden layer of the NN is the intermediate layer between the predictors and output. The hidden layer consists of one or multiple hidden units, and their variables are unobserved. Usually, the relationship between the hidden unit and partial or all set of the predictors is non-linear because a non-linear function such as logistic (i.e., sigmoidal) is

used to transform the original linear relationship. Once we set the numbers of the hidden unit, a linear combination is used to connect each unit to the outcome when the NN is for regression. For classification, another non-linear function, sigmoidal, is used for the combination.

Over-fitting usually occurs during NN training the data set due to the overly complex model with a large number of the parameters. Weight decay, an additional term in the weight, is a penalized approach for NN to avoid this problem. Also, model averaging is generally used on NN for a more stable prediction (Kuhn & Johnson, 2013).

2.3.5 Random Forest

Random forest (Breiman, 2001), an ensemble learning method that provides improved performance of bagged trees by reducing the correlation among trees. In the ensemble, each model uses the random sample extracted from the original data to generate a prediction for a new sample during training, and then use the average of these m predictions for regression or the majority votes for classification to give the forest's final prediction.

At each split, the algorithm randomly selects m predictors as split candidates instead of including the full set of the original predictors in the bagged trees method. The value of m is typically a prerequisite for the model (Kuhn & Johnson, 2013):

* For regression, the recommended setting of m value is one-third of the predictors and suggested to start with at least five values to tune m.

* For classification, the recommended setting of m value is the square root of the number of predictors and suggested to start with at least five values to tune m.

As a starting point, at least 1000 trees are suggested.

2.3.6 Support Vector Machines (SVM)

In machine learning, SVM is a supervised learning model first developed by Vladimir Vapnik (Kuhn & Johnson, 2013). The principle of SVM is to amplify the feature space by using kernel function (James et al., 2013). The SVM can perform either linear or non-linear regression/classification with the implement of different types of kernel function, such as polynomial, radial basis function and hyperbolic tangent.

In SVM, the cost value and the parameters of the kernel function should be tuned accordingly to avoid over-fitting and under-fitting. Usually, the model may under-fit the data when the cost value is low, and vice versa. The resampling approach for SVM to find the optimal parameters is commonly used for a balance between over-fitting and under-fitting of the data (Kuhn & Johnson, 2013).

2.3.7 Ensemble Learning

In statistics and machine learning, ensemble learning is used to train the multiple models (often called "weak learners" or "base models") for the same problem and combine the strength of them to get better results. The first step of ensemble learning is to develop the base models as the building blocks. Mostly, these base models may not perform well by themself and usually have large prediction error as a result of high variance or bias.

Therefore, the second step of ensemble learning is to make an effort to reduce the variance and/or the bias by combining weak learners to form a composite learner (ensemble model) for better performance. There are three commonly used approaches to combine weak learners: bagging (bootstrap aggregating), boosting, and stacking (Hastie et al., 2001).

The bagging method often generates same base learners and train them parallelly with the bootstrap resampling data that extract from the original dataset, and then combines base learners' predictions either using averaging for regression or casting a vote for classification. In this way, bagging methods can obtain an ensemble model with lower variance. For example, the random forest method we introduced in the above sub-section, is a bagging method that generates single trees as the base learners to independently fit on bootstrap samples, then combines their predictions to produce an output with lower variance.

The boosting approach generates the same base learners but trains them sequentially, that is, for each iteration, train the current ensemble model focused on the weakness of the previous one and then aggregate the current one weighted by its performance to the previous one to form a strong learner. The iterative strategy of the boosting method allows the ensemble model to get a lower bias. For example, the GBM method we introduced in the above sub-section is a boosting method. At each iteration, we fit a weak learner is to the pseudo-residuals and calculate the value of its weight by following the one-dimensional optimization process to, then update the ensemble model by adding the new weak learner multiplied by its weight.

Stacking is a relatively new approach compares with bagging and boosting. First, stacking considers various types of learning algorithms for base learners. Second, another algorithm or model is utilized to combine the predictions. In Chapter 3, we are going to use stacking as the ensemble learning approach to aggregate several machine learning algorithms to find out if it could help to enhance the model accuracy.

2.4 Model Comparison Methods

Now we have illustrated all the statistical learning methods and machine learning methods that we are going to use for our data set in Chapter III. The sequential task of deciding which method may outperform the others for this particular data set is important and necessary. By comparing the results of these models, we would have a better understanding of each learning method applied to our data set.

In this section, we discuss the way that we choose to separate the whole data set and the approach that we use to compare the prediction performance among the models.

2.4.1 Data Splitting and Cross-Validation

In the previous sections, we have seen that in the stepwise selection and the lasso, we need to select the variables or to choose the optimal tuning parameter by comparing the performance among the candidates. Moreover, after the appropriate level of flexibility of the model has been selected, what we really interested in is the accuracy of the predicted response that we obtain from our final model, that is, the evaluation of the model performance. Hence, there are two tasks we consider in here:

1. Model selection and regulation: the process of choosing the optimal model by estimating the performance of the candidates.
2. Model assessment: evaluating a model's performance by estimation its prediction error.

As our data set large with more than a thousand observations, the best approach to fulfill the above two tasks is to randomly separate the dataset into three parts: a training set, a validation set, and a test set. The training set is used to train the models; the validation set is used for model selection and regulation; the test set is set aside for the final model assessment (Hastie et al., 2001). The k-fold cross-validation (CV) is one of the resampling techniques that can ensure the samples are randomly separate into k sets with the same size, and researches showed that the repeating process of the k-fold cross-validation can effectively increase the accuracy of the prediction (Kuhn & Johnson, 2013). Therefore, we consider the reiterative k-fold CV as the resampling approach to randomly separate the whole data set into three subsamples in order to estimate the true prediction error.

Take the lasso as an example, suppose we first use 5-fold CV to split the whole data set into a modeling set and a test set and set the test set aside. Second, we use 3-fold CV again to re-split the modeling set into training and validation set, train the model with a grid of tuning parameter on the training set and predict the validation set to get the CV MSE. The optimal tuning parameter has been selected with the minimum CV MSE, as is the final model. Third, the final model is used to predict the test set and test error is

calculated. The external 5-fold CV provides 5 test errors. Finally, we can estimate the true prediction error by averaging the five test errors.

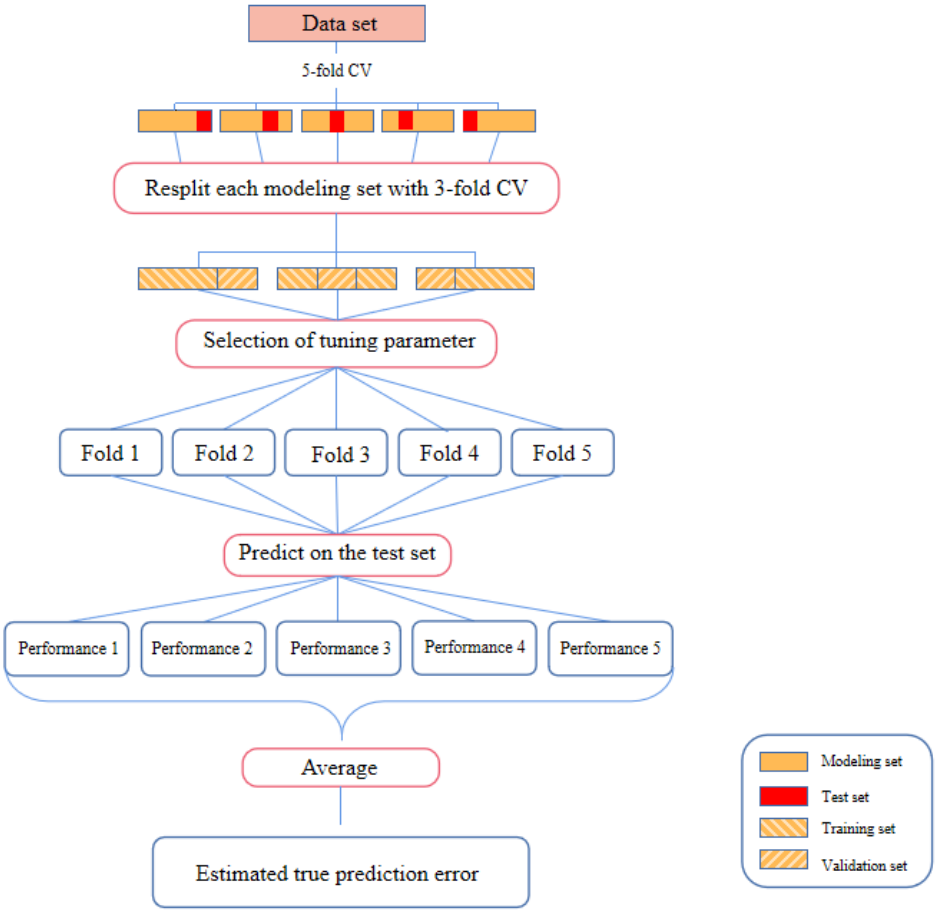

Figure 2: The Process of External 5-Fold CV and Internal 3-Fold CV.

In Chapter III, we use a 5-fold CV to evaluate the true prediction error for the multiple linear regression and the multinomial logistic regression approaches. The procedure of repeated CV, that is, the approach of an external 5-fold CV comes with an internal 5-fold CV as we illustrate in Figure. 2 is carried out for the rest of the models.

2.4.2 Prediction Performance Measures

1. Mean Square Error (MSE)

We have mentioned MSE a few times in the previous sections, now we would like to introduce it formally. In the regression, MSE is one of the most commonly used ways to measure how well the model fits the data, that is to say, a qualified estimator to measure the average squared difference between the prediction values and the actual value. For classification, we simply convert both prediction classes and observation classes into quantitative values to allow us to compute MSE for the classification approach as well. Moreover, for regression, we convert the quantitative response to the qualitative response by mapping a numeric range onto each class and then convert them back to numeric for calculating the MSE. Hence, the MSE for regression and classification is comparable in that way. We have the MSE formula (James et al., 2013):

$$\text{MSE} = \frac{2}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \tag{2.26}$$

Where $\hat{f}(x_i)$ is the prediction of the ith observation.

Generally speaking, we are more interested in the test MSE instead of the training MSE because we would like to know the accuracy of the prediction. Suppose that we have trained and fitted our model on the training data set and obtained the estimate $\hat{f}$, and we define (x_0, y_0) is an observation in the test data which we have set aside from the very

beginning. The expected test MSE can tell us whether $\hat{f}(x_0)$ is approximately equal to y_0. That is (James et al., 2013):

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left\{Bias\left(\hat{f}(x_0)\right)\right\}^2 + Var(\varepsilon) \qquad (2.27)$$

We use k-fold CV approach to estimate the overall test MSE (true prediction error) by computing the average of $E\left(y_0 - \hat{f}(x_0)\right)^2$.

In Equation (2.27), we see the relationship between bias, variances, and test MSE. Our goal is to select the final model with the test MSE as small as possible. However, this bias-variance trade-off can explain that when the model is with the high bias then there is probably an over-fitting, on the contrary, when the model is with the high variance then an under-fitting may occur. Hence, the test MSE is a necessary estimator when we compare flexible methods with classical but simpler methods.

2. Classification Accuracy Rate

Calculating the proportion of how many prediction outcomes match the true response of the observation is a common and direct approach to evaluate the prediction performance. We define the accuracy rate with a simple indicator function:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \qquad (2.28)$$

The accuracy rate shows the proportion of how many observations are classified correctly by our model.

Although the prediction outcomes are numeric of the regression approaches illustrated in Section 2.1, we convert the quantitative response to qualitative response by mapping a numeric range onto class. This mapping approach can ensure that every model we used in the thesis are compared under the same criteria.

We have explained the approaches along with the concepts that we are going to utilize to explore the private sector participation in disaster risk reduction data. The computational application on the data set will be shown in Chapter III.

CHAPTER III. APPLICATION

In Chapter III, our goal is to apply the statistical methods and machine learning approaches described in Chapter II to analyze our data set and present the outputs. The related results are shown in tables and figures. The comparison and discussion on the models' performance and variable selection will be provide in Chapter IV.

We use the RStudio environment, version 1.3.1073, for statistical computing. The R function and the package we used will be illustrated under each analysis approach.

3.1 Descriptive Statistics

The raw data contain 36 predictors, 3 response, and 1197 observations. All predictors are categorical: DE variables are of 0/1 scale, business size is of three levels (small, median and large), and city locations are of six levels. The response variables are calculated by four items extracting from the survey with equal increment in 0.25 each, taking values of 0, 0.25, 0.5, 0.75, and 1. We are only interested in the response of DRRI in the thesis. Although the response DRRI is quantitative, it can also be categorized as a factor with five levels because the previous study did not justify the equal increments. It is noticed that the predictor of the data set has a small number of missing values, which are 0.7% in city location and 2.9% in business size. Since the percentage of the missing cases is small, we consider these missing values can be ignored. After deleting the corresponding observations of any missing values listwise, we prepare the data set with 1162 complete observations, 36 predictors and 1 response of DRRI.

We give the abbreviation of each of the variables for more succinct outputs. The abbreviation of the 17 disaster experiences (DE), business size, city, and the response are:

* LI: Loss of IT
* SCD: Supply chain disruption
* Dea: Deaths
* LAS: Loss of access to site
* EC: Extreme conditions (high/low temperatures, flood/high winds)
* DC: Damage to corporate image/reputation/brand
* LTC: Loss of telecommunications
* PG: Pressure groups
* PO: Power outage
* IA: Industrial action
* WO: Water outage
* EI: Environmental incident
* CH: Customer health/product safety issue/incident
* LKSP: Loss of key skills and personnel
* NP: Negative publicity/coverage
* DF: Damaged facilities/equipment/inventories
* OT: Other
* City: City locations
* BS: Business size
* DRRI: Disaster Risk Reduction Index

Let us take a look at the between-predictor correlations:

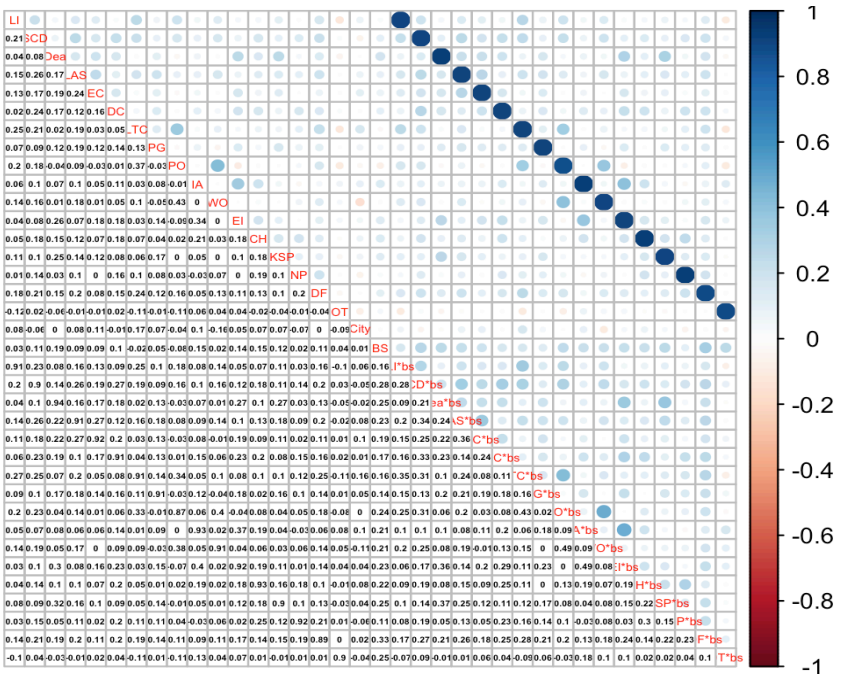

Figure 3: A Mixed Visualization of the Correlation Matrix. The Correlation Coefficients are Colored According to the Value. The Size and Shade of Each Circle Represent the Strength of Each Relationship, While the Color Represents the Direction, Either Negative or Positive.

The interaction term between each DE and business size is indicated as DE*BS. For example, LI*BS indicates the interaction between LI and BS. Clearly, there is a highly positive correlation between the interaction term and the main effect terms, the DE. The correlation between each of the interaction terms and its corresponding main effect is within a range of (0.8749, 0.9358). Other than these, the rest of the paired correlations which are bigger than 0.3 are as follows:

Table 1: The Pair of Variables with Correlation ≥ 0.3, Sorted from the Least to the Greatest.

Var1	Var2	Correlation
SCD*BS	PO*BS	0.3098
SCD*BS	LTC*BS	0.3140
Dea	LKSP*BS	0.3175

BS	DF*BS	0.3255
SCD*BS	DC*BS	0.3256
LTC	PO*BS	0.3350
PO	LTC*BS	0.3354
SCD*BS	LAS*BS	0.3375
IA	EI	0.3440
LI*BS	LTC*BS	0.3486
Dea*BS	EI*BS	0.3603
LAS*BS	EC*BS	0.3615
LTC	PO	0.3653
Dea*BS	LKSP*BS	0.3741
EI	IA*BS	0.3746
PO	WO*BS	0.3775
IA	EI*BS	0.3986
WO	PO*BS	0.4040
PO	WO	0.4265
LTC*BS	PO*BS	0.4269
IA*BS	EI*BS	0.4859
PO*BS	WO*BS	0.4936

There is no considerably high correlation showed in the table for the rest of the paired variables.

The distribution of the response is given as the frequency and relative frequency of each level.

Table 2: Distribution of DRRI.

DRRI					
Value	**0**	**0.25**	**0.5**	**0.75**	**1**
Frequency	209	326	351	237	39
Relative Frequency	0.1799	0.2806	0.3021	0.2040	0.0336

We notice that the frequency of value “1” is much less than the frequency of the other values. To avoid absence of value “1” in the test set, we chose the 5-fold external CV and 5-fold internal CV for the resampling, not higher fold.

All the 17 DEs are binary, City has 6 levels, BS has 5 levels, interaction term has 4 levels. We take SCD and its interaction as the example:

Table 3: Frequency Distribution of Selected Predictors.

SCD

Value	**0**	**1**
Frequency	1025	137
Relative Frequency	0.8821	0.1179

SCD*BS

Value	**0**	**1**	**2**	**3**
Frequency	1025	93	33	11
Relative Frequency	0.8821	0.0800	0.0284	0.0095

BS

Value	**1**	**2**	**3**
Frequency	915	201	46
Relative Frequency	0.7874	0.1730	0.0396

City

Value	**1**	**2**	**3**	**4**	**5**	**6**
Frequency	146	261	263	191	183	118
Relative Frequency	0.1256	0.2246	0.2263	0.1644	0.1575	0.1015

To illustrate the sample relationship between response and DE, the frequency table of DRRI with one of the predictors SCD is given as an example.

Table 4: Contingency Table of Variable DRRI and SCD.

		SCD		
		0	**1**	**Total**
DRRI	**0**	196	13	209
	0.25	299	27	326
	0.5	306	45	351
	0.75	196	41	237

1	28	11	39
Total	1025	137	1162

3.2 Linear Methods for Regression

In this section, we consider the independent variables and the dependent variable all as numeric. The independent variables are scaled, that is, having zero mean and unit variance.

Although the outcome is assumed to be continuously distributed for linear regression, we map a range of the quantitative output into the corresponding discrete value in order to compare with the original observation value. The mapping criteria:

$$\hat{y} \leq 0.125 \rightarrow \text{"0"}$$
$$0.125 < \hat{y} \leq 0.375 \rightarrow \text{"0.25"}$$
$$0.375 < \hat{y} \leq 0.625 \rightarrow \text{"0.5"}$$
$$0.625 < \hat{y} \leq 0.875 \rightarrow \text{"0.75"}$$
$$\hat{y} > 0.875 \rightarrow \text{"1"} \qquad (3.1)$$

We use this criterion to achieve the “mapping MSE”. In this way, we could fairly compare the prediction performance for both regression and classification approach later by the test set MSE.

3.2.1 Multiple Linear Regression

In order to fit a multiple linear regression model, we use function lm() that is available in the stats package. There is no tuning process for the multiple linear regression itself.

Therefore, we only use external 5-fold CV alone, and we obtain five folds from five external CV iterations. For example, the estimated coefficients for the fifth fold are as follows:

Table 5: Estimated Coefficients of the Multiple Linear Regression.

Estimated Coefficients			
(Intercept)	0.4122		
LI	-0.0178	BS	0.0348
SCD	0.0261	LI*BS	0.0162
Dea	-0.0294	SCD*BS	-0.0136
LAS	0.0179	Dea*BS	0.0507
EC	0.0006	LAS*BS	-0.0094
DC	-0.0206	EC*BS	0.0178
LTC	0.0381	DC*BS	0.0287
PG	0.0028	LTC*BS	-0.0012
PO	0.0081	PG*BS	0.0097
IA	-0.0131	PO*BS	0.0171
WO	-0.0238	IA*BS	0.0162
EI	0.0020	WO*BS	0.0186
CH	0.0070	EI*BS	0.0149
LKSP	-0.0062	CH*BS	-0.0156
NP	-0.0502	LKSP*BS	-0.0153
DF	0.0126	NP*BS	0.0509
OT	-0.0364	DF*BS	-0.0016
City	-0.0285	OT*BS	0.0376

The R-squared of the multiple linear regression model is 0.1422. The estimated model can be written as:

$$f(x) = 0.4122 - 0.0178\times LI + 0.0261\times SCD - 0.0294\times Dea + 0.0179\times LAS + 0.0006\times EC - 0.0206\times DC + 0.0381\times LTC + 0.0028\times PG + 0.0081\times PO - 0.0131\times IA - 0.0238\times WO + 0.0020\times EI + 0.0070\times CH - 0.0062\times LKSP - 0.0502\times NP + 0.0126\times DF - 0.0364\times OT - 0.0285\times City + 0.0384\times BS + 0.0162\times(LI*BS) - 0.0136\times(SCD*BS) + 0.0507\times(Dea*BS) -$$

$$0.0094\times(LAS*BS) + 0.0178\times(EC*BS) + 0.0287\times DC*BS - 0.0012\times(LTC*BS) + 0.0097\times(PG*BS) + 0.0171\times(PO*BS) + 0.0162\times(IA*BS) + 0.0186\times(WO*BS) + 0.0149\times(EI*BS) - 0.0156\times(CH*BS) - 0.0153\times(LKSP*BS) + 0.0509\times(NP*BS) - 0.0016\times(DF*BS) + 0.0376\times(OT*BS) \quad (3.2)$$

The variables with the p-value coefficients ≤ 0.1 for each fold:

Table 6: Significant Coefficients of the Independent Variables for Multiple Linear Regression.

P-Values of The Significant Variables

Fold 1

(Intercept)	LI	LTC	NP	City	BS	LI*BS	NP*BS
0.0000	0.0454	0.0347	0.0048	0.0007	0.0006	0.0985	0.0120

Fold 2

(Intercept)	LTC	NP	City	BS	NP*BS
0.0000	0.0234	0.0630	0.0048	0.0101	0.0572

Fold 3

(Intercept)	LAS	LTC	NP	City	BS	LTC*BS	WO*BS
0.0000	0.0774	0.0009	0.0394	0.0003	0.0008	0.0471	0.0844

NP*BS
0.0254

Fold 4

(Intercept)	CH	NP	City	BS	NP*BS
0.0000	0.0611	0.0161	0.0003	0.0004	0.0169

Fold 5

(Intercept)	NP	City	BS	NP*BS
0.0000	0.0616	0.0018	0.0135	0.0853

It shows that different splitting of the entire data influences on the significance of the coefficients. The coefficients of the variables NP, City, BS, and NP*BS are significant for all folds. The coefficients of the main effect variables are not necessarily significant for both when the coefficients of their interaction term are significant.

The MSE of the training set, test set and after mapping:

Table 7: The MSE of the Multiple Linear Regression.

	MSE of Training	**MSE of Test Set**	**MSE of Mapping**
Fold 1	0.0663	0.0587	0.0644
Fold 2	0.0645	0.0662	0.0695
Fold 3	0.0624	0.0768	0.0855
Fold 4	0.0628	0.0761	0.0792
Fold 5	0.0633	0.0750	0.0811
Mean	0.0638	0.0705	0.0760
SD	0.0016	0.0079	0.0087

The rate of accuracy of each fold:

Table 8: The Rate of Accuracy of the Multiple Linear Regression.

	Rate of Accuracy
Fold 1	0.4204
Fold 2	0.2870
Fold 3	0.3117
Fold 4	0.2895
Fold 5	0.3193
Mean	0.3256
SD	0.0548

Across five folds, the rate of accuracy varies from the smallest 0.2870 to the largest 0.4204. Fold 1 has the smallest MSE of mapping and the highest accuracy at the same

time. Although fold 3 has the largest MSE of mapping, its rate of accuracy is not the lowest among the five folds.

3.2.2 Stepwise Variable Selection

We use the stepAIC() function in MASS package to perform stepwise selection with 5-fold external CV and 5-fold internal CV. The training set is used for the modeling, validation set is used to find the optimal model with the minimal CV MSE and the test set is used to evaluate the prediction performance.

For example, the estimated coefficients for the selected predictors in the fifth fold:

Table 9: Estimated Coefficients of the Stepwise Selection for Linear Regression.

	Estimated Coefficients
(Intercept)	0.4124
LTC	0.0363
DF	0.0160
OT	-0.0394
City	-0.0265
BS	0.0445
EC*BS	0.0231
PO*BS	0.0349
OT*BS	0.0388

The variables that have been selected in each fold:

Table 10: Selected Variables of the Stepwise Selection for Linear Regression.

Fold 1							
EC	LTC	NP	OT	City	BS	PG*BS	NP*BS
DF*BS	OT*BS						

Fold 2

LI	EC	LTC	PO	WO	NP	DF	City
BS	LI*BS	SCD*BS	Dea*BS	LTC*BS	WO*BS	NP*BS	

Fold 3

SCD	Dea	LAS	LTC	PO	CH	NP	City
BS	Dea*BS	EI*BS	NP*BS				

Fold 4

SCD	Dea	LTC	WO	NP	City	BS	SCD*BS
EC*BS	WO*BS	LKSP*BS	NP*BS	DF*BS			

Fold 5

LTC	DF	OT	City	BS	EC*BS	PO*BS	OT*BS

We calculate the frequency of the variables that have shown up on the above table:

Table 11: The Frequency of Selected Variables of the Stepwise Selection for Linear Regression.

	Variable	Frequency	Relative Frequency
1	BS	5	1
2	City	5	1
3	LTC	5	1
4	NP*BS	4	0.8
5	NP	4	0.8
6	Dea*BS	2	0.4
7	DF*BS	2	0.4
8	EC*BS	2	0.4
9	OT*BS	2	0.4
10	SCD*BS	2	0.4
11	WO*BS	2	0.4
12	Dea	2	0.4
13	DF	2	0.4
14	EC	2	0.4
15	OT	2	0.4
16	PO	2	0.4
17	SCD	2	0.4
18	WO	2	0.4
19	EI*BS	1	0.2

20	LI*BS	1	0.2
21	LKSP*BS	1	0.2
22	LTC*BS	1	0.2
23	PG*BS	1	0.2
24	PO*BS	1	0.2
25	CH	1	0.2
26	LAS	1	0.2
27	LI	1	0.2

The number of selected variables varies with the different folds. The 27 out of 36 variables have been selected through the stepwise selection. Fold 2 selects the largest number of fifteen variables and fold 5 selects the minimal numbers of eight variables. All five folds select the variables of BS, City and LTC.

The MSE of the training set, test set and the mapping:

Table 12: The MSE of the Stepwise Selection for Linear Regression.

	MSE of Training	**MSE of Test Set**	**MSE of Mapping**
Fold 1	0.0629	0.0748	0.0783
Fold 2	0.0657	0.0809	0.0913
Fold 3	0.0649	0.0755	0.0784
Fold 4	0.0646	0.0710	0.0768
Fold 5	0.0649	0.0588	0.0676
Mean	0.0646	0.0722	0.0785
SD	0.0011	0.0083	0.0084

The rate of accuracy of each fold:

Table 13: The Rate of Accuracy of the Stepwise Selection for Linear Regression.

	Rate of Accuracy
Fold 1	0.3188
Fold 2	0.2881
Fold 3	0.3333
Fold 4	0.3568

Fold 5	0.3273
Mean	0.3249
SD	0.0250

Among five folds, the rate of accuracy varies from the smallest 0.2881 to the largest 0.3568. Although Fold 5 has the smallest MSE of mapping, Fold 4 has the highest accuracy. Fold 2 has the largest MSE of mapping and lowest rate of accuracy at the same time.

3.2.3 The Lasso

In order to fit a lasso model, we use the cv.lars() and lars() function in lars package. The training set is used for the modeling, validation set is used to find the optimal tuning parameter by applying the one-standard-error rule on the minimal CV MSE. The test set is then used to evaluate the prediction performance for the selected model.

For example, the lasso coefficients for the predictors with the tuning parameter s = 0.0808 in the fifth fold:

Table 14: Estimated Coefficients of the Lasso for Linear Regression.

Lasso Coefficients			
(Intercept)	0.4103		
BS	0.0241	CH	0.0000
LTC*BS	0.0166	EI	0.0000
SCD*BS	0.0091	LKS	0.0000
PO*BS	0.0052	OT	0.0000
DF*BS	0.0023	NP	0.0000
EI*BS	0.0002	City	0.0000
DF	0.0000	LI*BS	0.0000
LI	0.0000	Dea*BS	0.0000
SCD	0.0000	LAS*BS	0.0000

Dea	0.0000	EC*BS	0.0000
LAS	0.0000	DC*BS	0.0000
EC	0.0000	PG*BS	0.0000
DC	0.0000	IA*BS	0.0000
LTC	0.0000	WO*BS	0.0000
PG	0.0000	CH*BS	0.0000
PO	0.0000	LKSP*BS	0.0000
IA	0.0000	NP*BS	0.0000
WO	0.0000	OT*BS	0.0000

Here we see that 30 of the 36 coefficient estimates are exactly zero, that is, this lasso model with tuning parameter 0.0808 chosen by CV only contains six variables.

The variables with non-zero estimated coefficient of each fold are:

Table 15: Selected Variables of the Lasso for Linear Regression.

Fold 1 with *s* = 0.0505

BS SCD*BS LTC*BS PO*BS

Fold 2 with *s* = 0.1010

BS SCD*BS LTC*BS PO*BS DF*BS

Fold 3 with *s* = 0.0707

BS LTC LTC*BS PO*BS DF*BS

Fold 4 with *s* = 0.0707

BS SCD*BS LTC*BS PO*BS DF*BS

Fold 5 with *s* = 0.0808

BS SCD*BS LTC*BS PO*BS EI*BS DF*BS

We calculate the frequency of the variables that have shown up on the above table:

Table 16: The Frequency of the Lasso for Linear Regression.

Variable	Frequency	Relative Frequency
BS	5	1
LTC*BS	5	1
PO*BS	5	1
DF*BS	4	0.8
SCD*BS	4	0.8
EI*BS	1	0.2
LTC	1	0.2

Although Fold 3 and Fold 4 have the same optimal tuning parameter 0.0707, the variables that have non-zero estimated coefficient of the two folds are not the same. In contrast, Fold 2 and Fold 4 have the exact same variables with the non-zero estimated coefficient but with different optimal tunning parameters. All five folds select the variables of BS, LTC*BS, and PO*BS and. The lasso shrinkage approach has the tendency of the preference of the interaction term compared to the stepwise selection.

The MSE of the training set, test set and the mapping:

Table 17: The MSE of the Lasso for Linear Regression.

	MSE of Training	MSE of Test Set	MSE of Mapping
Fold 1	0.0729	0.0632	0.0747
Fold 2	0.0686	0.0706	0.0687
Fold 3	0.0689	0.0707	0.0810
Fold 4	0.0682	0.0741	0.0789
Fold 5	0.0692	0.0722	0.0872
Mean	0.0695	0.0702	0.0781
SD	0.0019	0.0041	0.0069

The rate of accuracy of each fold:

Table 18: The Rate of Accuracy of the Lasso for Linear Regression.

	Rate of Accuracy
Fold 1	0.4027
Fold 2	0.3274
Fold 3	0.3158
Fold 4	0.2807
Fold 5	0.2605
Mean	0.3174
SD	0.0547

Among five folds, Fold 1 has the smallest MSE and the highest accuracy at the same time. Although Fold 4 has the largest MSE of the test set, Fold 5 has lowest rate of accuracy.

3.2.4 Principal Components Regression

We use the pcr() function in package pls and printcomp() function, part of the stats package, to perform PCR with 5-fold external CV and 5-fold internal CV. The training set is used for the modeling, validation set is used to find the optimal principal components with the minimal CV MSE, and the test set is used to evaluate the prediction performance for each fold.

For example, for Fold 5, we find the optimal principal components with the minimal CV MSE shown in Figure 4. The selected numbers of principal components for each fold are:

Table 19: Selected PCs of the PCA.

Number of PCs				
Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
20 comps	18 comps	19 comps	19 comps	18 comps

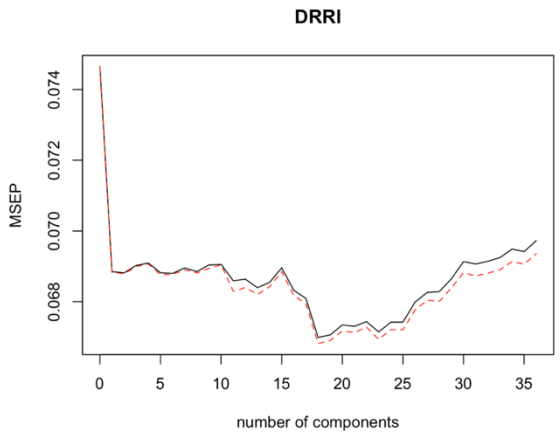

Figure 4: The CV Plot for Optimal Principal Components.

We have the estimated coefficients of Fold 5 with 18 principal components:

Table 20: Estimated Coefficients of the PCR.

PCR Coefficients			
(Intercept)	0.3145		
LI	-0.0048	BS	0.0390
SCD	0.0027	LI*BS	0.0034
Dea	0.0065	SCD*BS	0.0117
LAS	-0.0011	Dea*BS	0.0102
EC	0.0054	LAS*BS	0.0061
DC	0.0033	EC*BS	0.0112
LTC	0.0157	DC*BS	0.0047
PG	0.0051	LTC*BS	0.0239
PO	0.0049	PG*BS	0.0092
IA	0.0010	PO*BS	0.0176
WO	-0.0051	IA*BS	0.0043
EI	0.0059	WO*BS	0.0020
CH	-0.0059	EI*BS	0.0106
LKSP	-0.0140	CH*BS	-0.0024

NP	-0.0034	LKSP*BS	-0.0054
DF	0.0021	NP*BS	0.0009
OT	-0.0033	DF*BS	0.0107
City	-0.0277	OT*BS	0.0045

The percentage of the cumulated variance explained by the principal components:

Table 21: Cumulative Proportion of Variance Explained by Principal Components.

Cumulative Proportion of Variance Explained

1 comp	**2 comps**	**3 comps**	**4 comps**	**5 comps**	**6 comps**
0.1714	0.2670	0.3356	0.3997	0.4580	0.5130
7 comps	**8 comps**	**9 comps**	**10 comps**	**11comps**	**12comps**
0.5648	0.6128	0.6553	0.6969	0.7353	0.7711
13 comps	**14 comps**	**15 comps**	**16 comps**	**17 comps**	**18comps**
0.8056	0.8375	0.8689	0.8978	0.9233	0.9457
19 comps	**20 comps**	**21 comps**	**22 comps**	**23 comps**	**24 comps**
0.9669	0.9724	0.9760	0.9787	0.9811	0.9834
25 comps	**26 comps**	**27 comps**	**28 comps**	**29 comps**	**30 comps**
0.9856	0.9876	0.9894	0.9911	0.9927	0.9941
31 comps	**32 comps**	**33 comps**	**34 comps**	**35 comps**	**36 comps**
0.9953	0.9965	0.9975	0.9985	0.9994	1.0000

The selected numbers of 18, 19 and 20 components approximately explain 95% to 97% variance of the input space. And the first two components explain 27% of variance of the input data. The plot of the scores for the first and second components provides a visual

understanding. In Figure 5, each color of the circles represents each group. There is no distinct boundary to separate the five groups.

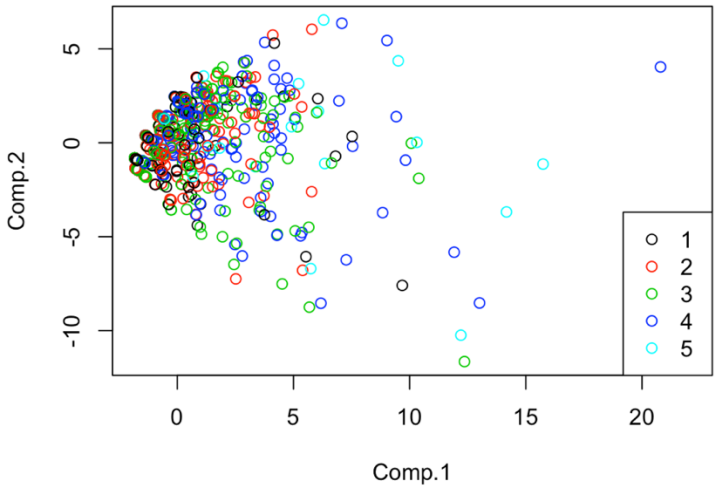

Figure 5: Visualization on the First Two Principal Components.

The MSE of the training set, test set and the mapping:

Table 22: The MSE of the PCR.

	MSE of Training	**MSE of Test Set**	**MSE of Mapping**
Fold 1	0.0678	0.0563	0.0595
Fold 2	0.0655	0.0662	0.0706
Fold 3	0.0640	0.0726	0.0822
Fold 4	0.0643	0.0702	0.0740
Fold 5	0.0642	0.0733	0.0785
Mean	0.0652	0.0677	0.0730
SD	0.0016	0.0070	0.0087

The rate of accuracy of each fold:

Table 23: The Rate of Accuracy of the PCR.

	Rate of Accuracy
Fold 1	0.4336
Fold 2	0.3229
Fold 3	0.2996
Fold 4	0.2982
Fold 5	0.3361
Mean	0.3381
SD	0.0558

Among five folds, Fold 1 has the smallest MSE of the test set and the highest accuracy at the same time. Although Fold 3 has the largest MSE of mapping, Fold 4 has the lowest rate of accuracy.

3.3 Linear Methods for Classification

In this section, we consider the dependent variable DRRI as a quantitative variable. The approaches of multinomial logistic regression, stepwise selection, elastic net, ordinal logistic regression and linear discriminant analysis are utilized for classification. The relative coefficients, optimal tuning parameter, selected variables, MSE of the training and test set and the accuracy rate of the model will be shown accordingly with each approach.

We prepare the data set as follows:

* The response DRRI is a five-level categorical variable: 0, 0.25, 0.5, 0.75 and 1.
* The 17 DEs are the two-level categorical variables: 0 and 1.

* The predictor City is a six-level categorical variable: 1,2,3,4,5 and 6.
* The predictor BS is a three-level categorical variable: 1,2 and 3.
* The 17 interaction terms are four-level categorical variables: 0,1,2,3 and 4.

3.3.1 Multinomial Logistic Regression

In order to fit a multinomial logistic model, we use function multinom() that is available in the nnet package. There is no tuning process for the multinomial logistic itself. Therefore, we only use external 5-fold CV alone, and we obtain five folds from five external CV iterations.

For example, the estimated coefficients of the fifth fold, where class 0 is considered as reference level:

Table 24: Estimated Coefficients of the Multinomial Logistic Regression.

	Estimated Coefficients			
	Class 0.25	**Class 0.5**	**Class 0.75**	**Class 1**
(Intercept)	0.6812	0.0743	-0.1834	-2.1937
LI1	-2.4338	-3.3057	6.4398	6.1830
SCD1	1.3971	2.4685	8.5076	4.5480
Dea1	0.2872	-4.0990	9.2415	4.5228
LAS1	0.1315	9.5837	-6.9741	1.7212
EC1	6.7975	4.3674	10.0002	6.4230
DC1	-3.9296	-7.5764	9.3893	-1.9918
LTC1	0.0353	-6.8168	1.3392	2.3355
PG1	-4.9977	8.3083	3.0806	2.9030
PO1	-4.3958	-0.9838	3.1050	3.0197
IA1	2.8530	5.8893	2.0339	6.5903
WO1	0.9735	-2.4385	-10.5801	-7.2146
EI1	2.0687	7.4865	2.6091	-0.6959
CH1	1.1676	0.5221	-8.5010	-10.5880
LKSP1	9.4290	8.9958	7.1961	4.3391
NP1	-4.6865	2.8491	7.6923	2.2269

DF1	-3.6247	1.9110	1.8806	1.4851
OT1	7.3147	6.9134	7.2630	8.0895
City2	0.2961	0.9696	0.2260	1.0395
City3	-0.6886	0.0941	-0.3297	-2.3065
City4	0.3081	0.1607	0.0115	-0.3026
City5	-0.2521	0.5970	0.4670	-2.2169
City6	-1.4814	-1.3280	-1.9984	-1.1054
BS2	-0.1731	0.5446	0.9286	-1.7076
BS3	-0.9134	-0.3381	-2.2824	0.8692
LI*BS1	2.7324	3.4979	-5.9690	-6.7099
LI*BS2	0.1879	-0.1582	-9.2921	-7.5206
LI*BS3	-5.3542	-6.6454	21.7009	20.4135
SCD*BS1	-1.3794	-1.8390	-8.3377	-3.2089
SCD*BS2	13.8563	13.2825	7.9545	8.4918
SCD*BS3	-11.0797	-8.9750	8.8908	-0.7349
Dea*BS1	-1.7768	3.6374	-9.5618	-11.3104
Dea*BS2	6.5757	13.5115	-0.0476	5.0023
Dea*BS3	-4.5117	-21.2480	18.8509	10.8308
LAS*BS1	-0.1438	-9.9556	7.7800	-1.4800
LAS*BS2	0.2771	-10.3995	7.6313	-3.9005
LAS*BS3	-0.0017	29.9388	-22.3854	7.1017
EC*BS1	-6.7089	-4.2511	-9.4024	-4.9463
EC*BS2	13.1274	16.6552	9.9521	13.7281
EC*BS3	0.3791	-8.0367	9.4504	-2.3588
DC*BS1	4.6603	7.3616	-9.2097	-16.5015
DC*BS2	-8.7910	-13.7086	4.7333	19.0187
DC*BS3	0.2011	-1.2293	13.8657	-4.5090
LTC*BS1	0.4414	7.5683	-0.3666	-1.4304
LTC*BS2	-0.3370	6.0731	-1.1701	-0.5962
LTC*BS3	-0.0691	-20.4582	2.8759	4.3621
PG*BS1	4.8394	-7.8976	-2.8737	-15.0047
PG*BS2	-9.3600	8.5331	12.1832	18.0831
PG*BS3	-0.4771	7.6727	-6.2289	-0.1754
PO*BS1	4.3385	1.1376	-2.8494	-3.0905
PO*BS2	5.4053	2.4713	-1.5032	-1.0423
PO*BS3	-14.1396	-4.5926	7.4577	7.1525
IA*BS1	-4.1557	-5.9692	-2.4222	-14.7098
IA*BS2	11.5387	9.4919	12.1613	7.1588
IA*BS3	-4.5300	2.3666	-7.7053	14.1412
WO*BS1	-1.3409	2.1377	9.9289	6.6314
WO*BS2	0.1573	2.1467	10.8104	10.1820
WO*BS3	2.1571	-6.7229	-31.3194	-24.0280
EI*BS1	-2.2044	-6.7582	-2.4047	-16.8448

EI*BS2	9.0565	1.5129	8.0291	15.0851
EI*BS3	-4.7834	12.7318	-3.0153	1.0637
CH*BS1	-3.0290	-1.4446	8.2573	-9.5827
CH*BS2	4.4127	6.5597	14.7440	16.6630
CH*BS3	-0.2161	-4.5929	-31.5022	-17.6684
LKSP*BS1	-10.4022	-9.9924	-8.0060	-4.7795
LKSP*BS2	5.3016	4.7561	5.7852	8.4672
LKSP*BS3	14.5295	14.2320	9.4169	0.6514
NP*BS1	5.4211	-2.7662	-8.7125	-2.3943
NP*BS2	-10.3086	6.8447	2.5392	9.1302
NP*BS3	0.2011	-1.2293	13.8657	-4.5090
DF*BS1	4.1217	-1.4462	-1.6922	-0.9204
DF*BS2	2.8297	-2.6224	-2.0854	-1.9478
DF*BS3	-10.5761	5.9796	5.6582	4.3533
OT*BS1	-6.9700	-6.8704	-8.2619	-8.0419
OT*BS2	-7.1015	-6.5944	-7.8516	-7.2677
OT*BS3	21.3863	20.3782	23.3765	23.3990

We consider DRRI = “0” as the baseline, therefore K-1 estimated models are presented:

ln (Pr (DRRI=0.25) / Pr(DRRI=0)) = 0.6812 - 2.4338(LI=1) - 1.3971(SCD=1) - 0.2872(Dea=1) + 0.1315(LAS=1) + 6.7975(EC=1) - 3.9296(DC=1) + 0.0353(LTC=1) - 4.9977(PG=1) -4.3958(PO=1) + 2.8530(IA=1) + 0.9735(WO=1) + 2.0687(EI=1) + 1.1676(CH=1) + 9.4290(LKSP=1) - 4.6865(NP=1) - 3.6247(DF=1) + 7.3147(OT=1) + 0.2961(City=2) - 0.6886(City=3) + 0.3081(City=4) - 0.2521(City=5) - 1.4814(City=6) - 0.1731(BS=2) - 0.9134(BS=3) + 2.7324(LI*BS=1) + 0.1879(LI*BS=2) - 5.3542(LI*BS=3) - 1.3794(SCD*BS=1) + 13.8563(SCD*BS=2) - 11.0797(SCD*BS=3) - 1.7768(Dea*BS=1) + 6.5757(Dea*BS=2) - 4.5117(Dea*BS=3) - 0.1438(LAS*BS=1) + 0.2771(LAS*BS=2) - 0.0017(LAS*BS=3) - 6.7089(EC*BS=1) + 13.1274(EC*BS=2) + 0.3791(EC*BS=3) + 4.6603(DC*BS=1) - 8.7910(DC*BS=2) + 0.2011(DC*BS=3) + 0.4414(LTC*BS=1) - 0.3370(LTC*BS=2) - 0.0691(LTC*BS=3) + 4.8394(PG*BS=1) -

9.3600(PG*BS=2) - 0.4771(PG*BS=3) + 4.3385(PO*BS=1) + 5.4053(PO*BS=2) - 14.1396(PO*BS=3) - 4.1557(IA*BS=1) + 11.5387(IA*BS=2) - 4.5300(IA*BS=3) - 1.3409(WO*BS=1) + 0.1573(WO*BS=2) + 2.1571(WO*BS=3) - 2.2044(EI*BS=1) + 9.0565(EI*BS=2) - 4.7834(EI*BS=3) - 3.0290(CH*BS=1) + 4.4127(CH*BS=2) - 0.2161(CH*BS=3) - 10.4022(LKSP*BS=1) + 5.3016(LKSP*BS=2) + 14.5295(LKSP*BS=3) + 5.4211(NP*BS=1) - 10.3086(NP*BS=2) + 0.2011(NP*BS=3) + 4.1217(DF*BS=1) + 2.8297(DF*BS=2) - 10.5761(DF*BS=3) - 6.9700(OT*BS=1) - 7.1015(OT*BS=2) + 21.3863(OT*BS=3) (3.3)

ln (Pr(DRRI=0.5) / Pr(DRRI=0)) = 0.0743 - 3.3057(LI=1) + 2.4685(SCD=1) - 4.0990(Dea=1) + 9.5837(LAS=1) + 4.3674(EC=1) - 7.5764(DC=1) - 6.8168(LTC=1) + 8.3083(PG=1) - 0.9838(PO=1) + 5.8893(IA=1) - 2.4385(WO=1) + 7.4865(EI=1) + 0.5221(CH=1) + 8.9958(LKSP=1) + 2.8491(NP=1) + 1.9110(DF=1) + 6.9134(OT=1) + 0.9696(City=2) + 0.0941(City=3) + 0.1607(City=4) + 0.5970(City=5) - 1.3280(City=6) + 0.5446(BS=2) - 0.3381(BS=3) + 3.4979(LI*BS=1) - 0.1582(LI*BS=2) - 6.6454(LI*BS=3) - 1.8390(SCD*BS=1) + 13.2825(SCD*BS=2) - 8.9750(SCD*BS=3) + 3.6374(Dea*BS=1) + 13.5115(Dea*BS=2) - 21.2480(Dea*BS=3) - 9.9556(LAS*BS=1) - 10.3995(LAS*BS=2) + 29.9388(LAS*BS=3) - 4.2511(EC*BS=1) + 16.6552(EC*BS=2) - 8.0367(EC*BS=3) + 7.3616(DC*BS=1) - 13.7086(DC*BS=2) - 1.2293(DC*BS=3) + 7.5683(LTC*BS=1) + 6.0731(LTC*BS=2) - 20.4582(LTC*BS=3) - 7.8976(PG*BS=1) + 8.5331(PG*BS=2) + 7.6727(PG*BS=3) + 1.1376(PO*BS=1) + 2.4713(PO*BS=2) - 4.5926(PO*BS=3) - 5.9692(IA*BS=1) + 9.4919(IA*BS=2)+ 2.3666(IA*BS=3) + 2.1377(WO*BS=1) + 2.1467(WO*BS=2) - 6.7229(WO*BS=3) - 6.7582(EI*BS=1) +

1.5129(EI*BS=2) + 12.7318(EI*BS=3) - 1.4446(CH*BS=1) + 6.5597(CH*BS=2) - 4.5929(CH*BS=3) - 9.9924(LKSP*BS=1) + 4.7561(LKSP*BS=2) + 14.2320(LKSP*BS=3) - 2.7662(NP*BS=1) + 6.8447(NP*BS=2) - 1.2293(NP*BS=3) - 1.4462(DF*BS=1) - 2.6224(DF*BS=2) + 5.9796(DF*BS=3) - 6.8704(OT*BS=1) - 6.5944(OT*BS=2) + 20.3782(OT*BS=3) (3.4)

ln (Pr (DRRI=0.75) / Pr(DRRI=0)) = - 0.1834 + 6.4398(LI=1) + 8.5076(SCD=1) + 9.2415(Dea=1) - 6.9741(LAS=1) + 10.0002(EC=1) + 9.3893(DC=1) + 1.3392(LTC=1) + 3.0806(PG=1) + 3.1050(PO=1) + 2.0339(IA=1) - 10.5801(WO=1) + 2.6091(EI=1) - 8.5010(CH=1) + 7.1961(LKSP=1) + 7.6923(NP=1) + 1.8806(DF=1) + 7.2630(OT=1) + 0.2260(City=2) - 0.3297(City=3) + 0.0115(City=4) + 0.4670(City=5) - 1.9984(City=6) + 0.9286(BS=2) - 2.2824(BS=3) - 5.9690(LI*BS=1) - 9.2921(LI*BS=2) + 21.7009(LI*BS=3) - 8.3377(SCD*BS=1) + 7.9545(SCD*BS=2) + 8.8908(SCD*BS=3) - 9.5618(Dea*BS=1) - 0.0476(Dea*BS=2) + 18.8509(Dea*BS=3) + 7.7800(LAS*BS=1) + 7.6313(LAS*BS=2) - 22.3854(LAS*BS=3) - 9.4024(EC*BS=1) + 9.9521(EC*BS=2) + 9.4504(EC*BS=3) - 9.2097(DC*BS=1) + 4.7333(DC*BS=2) + 13.8657(DC*BS=3) - 0.3666(LTC*BS=1) - 1.1701(LTC*BS=2) + 2.8759(LTC*BS=3) - 2.8737(PG*BS=1) + 12.1832(PG*BS=2) - 6.2289(PG*BS=3) - 2.8494(PO*BS=1) - 1.5032(PO*BS=2) + 7.4577(PO*BS=3) - 2.4222(IA*BS=1) + 12.1613(IA*BS=2) - 7.7053(IA*BS=3) + 9.9289(WO*BS=1) + 10.8104(WO*BS=2) - 31.3194(WO*BS=3) - 2.4047(EI*BS=1) + 8.0291(EI*BS=2) - 3.0153(EI*BS=3) + 8.2573(CH*BS=1) + 14.7440(CH*BS=2) - 31.5022(CH*BS=3) - 8.0060(LKSP*BS=1) + 5.7852(LKSP*BS=2) + 9.4169(LKSP*BS=3) - 8.7125(NP*BS=1) + 2.5392(NP*BS=2) + 13.8657(NP*BS=3) -

1.6922(DF*BS=1) - 2.0854(DF*BS=2) + 5.6582(DF*BS=3) - 8.2619(OT*BS=1) - 7.8516(OT*BS=2) + 23.3765(OT*BS=3) (3.5)

ln (Pr (DRRI=1) / Pr(DRRI=0)) = - 2.1937 + 6.1830(LI=1) + 4.5480(SCD=1) + 4.5228(Dea=1) + 1.7212(LAS=1) + 6.4230(EC=1) - 1.9918(DC=1) + 2.3355(LTC=1) + 2.9030(PG=1) + 3.0197(PO=1) + 6.5903(IA=1) - 7.2146(WO=1) - 0.6959(EI=1) - 10.5880(CH=1) + 4.3391(LKSP=1) + 2.2269(NP=1) + 1.4851(DF=1) + 8.0895(OT=1) + 1.0395(City=2) - 2.3065(City=3) - 0.3026(City=4) - 2.2169(City=5) - 1.1054(City=6) - 1.7076(BS=2) + 0.8692(BS=3) - 6.7099(LI*BS=1) - 7.5206(LI*BS=2) + 20.4135(LI*BS=3) - 3.2089(SCD*BS=1) + 8.4918(SCD*BS=2) - 0.7349(SCD*BS=3) - 11.3104(Dea*BS=1) + 5.0023(Dea*BS=2) + 10.8308(Dea*BS=3) - 1.4800(LAS*BS=1) - 3.9005(LAS*BS=2) + 7.1017(LAS*BS=3) - 4.9463(EC*BS=1) + 13.7281(EC*BS=2) - 2.3588(EC*BS=3) - 16.5015(DC*BS=1) + 19.0187(DC*BS=2) - 4.5090(DC*BS=3) - 1.4304(LTC*BS=1) - 0.5962(LTC*BS=2) + 4.3621(LTC*BS=3) - 15.0047(PG*BS=1) + 18.0831(PG*BS=2) - 0.1754(PG*BS=3) - 3.0905(PO*BS=1) - 1.0423(PO*BS=2) + 7.1525(PO*BS=3) - 14.7098(IA*BS=1) + 7.1588(IA*BS=2) + 14.1412(IA*BS=3) + 6.6314(WO*BS=1) + 10.1820(WO*BS=2) - 24.0280(WO*BS=3) - 16.8448(EI*BS=1) + 15.0851(EI*BS=2) + 1.0637(EI*BS=3) - 9.5827(CH*BS=1) + 16.6630(CH*BS=2) - 17.6684(CH*BS=3) - 4.7795(LKSP*BS=1) + 8.4672(LKSP*BS=2) + 0.6514(LKSP*BS=3) - 2.3943(NP*BS=1) + 9.1302(NP*BS=2) - 4.5090(NP*BS=3) - 0.9204(DF*BS=1) - 1.9478(DF*BS=2) + 4.3533(DF*BS=3) - 8.0419(OT*BS=1) - 7.2677(OT*BS=2) + 23.3990(OT*BS=3) (3.6)

The MSE of the training set and the test set:

Table 25: The MSE of the Multinomial Logistic Regression.

	MSE of Training	MSE of Test Set
Fold 1	0.1009	0.1289
Fold 2	0.0766	0.0956
Fold 3	0.0929	0.1217
Fold 4	0.0764	0.0962
Fold 5	0.0764	0.1132
Mean	0.0846	0.1111
SD	0.0115	0.0150

The rate of accuracy of each fold on the test set:

Table 26: The Rate of Accuracy of the Multinomial Logistic Regression.

	Rate of Accuracy
Fold 1	0.2920
Fold 2	0.3094
Fold 3	0.3239
Fold 4	0.3465
Fold 5	0.2983
Mean	0.3140
SD	0.0218

Among five folds, the rate of accuracy varies from the smallest 0.2920 to the largest 0.3465. Although Fold 2 has the smallest test MSE, Fold 4 has the highest accuracy. Fold 1 has the largest MSE and smallest rate of accuracy at the same time. The multinomial logistic regression mildly over-fit the data as the mean MSE of the training data is smaller than the mean MSE of the test data of 0.0265.

3.3.2 Stepwise Variable Selection

We use stepAIC() function, part of the MASS package, to perform stepwise selection with 5-fold external CV and 5-fold internal CV. Training set is used for the modeling, validation set is used to find the optimal model and test set is used to evaluate the prediction performance.

For example, the estimated coefficients of the selected predictors in the fifth fold:

Table 27: Estimated Coefficients of the Stepwise Selection for Classification.

	Estimated Coefficients			
	Class 0.25	**Class 0.5**	**Class 0.75**	**Class 1**
(Intercept)	0.6770	0.0741	-0.1866	-2.1904
LI1	-2.3559	-3.3963	6.4177	6.1530
SCD1	1.4573	2.8178	8.3719	4.6101
Dea1	0.3145	-3.8326	9.0788	4.1281
LAS1	0.0990	9.3856	-7.0093	1.9041
EC1	6.8263	4.5264	10.1261	6.6512
DC1	-3.6256	-7.0916	9.1537	-2.0353
LTC1	-0.0409	-6.7134	1.4078	2.4075
PG1	-4.7394	8.2863	3.1591	3.1089
PO1	-4.3604	-1.0130	2.9554	2.8824
IA1	2.7704	6.0032	2.1528	6.7084
WO1	1.1144	-2.2240	-10.0244	-6.8646
EI1	2.1362	7.4517	2.6038	-0.4991
CH1	1.2925	0.5368	-8.4955	-9.8655
LKSP1	9.0764	8.6386	6.9248	4.4174
NP1	-4.3397	2.9352	7.7294	2.5360
DF1	-3.6604	1.7883	1.7584	1.3497
City2	0.3009	0.9691	0.2284	1.0324
City3	-0.6855	0.0924	-0.3287	-2.3134
City4	0.3102	0.1576	0.0107	-0.3085
City5	-0.2502	0.5942	0.4661	-2.2187
City6	-1.4788	-1.3315	-1.9998	-1.1092
BS2	-0.1701	0.5501	0.9318	-1.6881
BS3	-0.9145	-0.3277	-2.2718	0.8737

LI*BS1	2.6582	3.5927	-5.9425	-6.6764
LI*BS2	0.0964	-0.0875	-9.2891	-7.5233
LI*BS3	-5.1105	-6.9016	21.6494	20.3527
SCD*BS1	-1.4401	-2.1879	-8.1975	-3.2714
SCD*BS2	13.0766	12.2097	7.3587	7.7348
SCD*BS3	-10.1792	-7.2040	9.2108	0.1468
Dea*BS1	-1.8043	3.3685	-9.3997	-11.6883
Dea*BS2	6.5204	13.2104	0.0873	5.3296
Dea*BS3	-4.4016	-20.4115	18.3913	10.4868
LAS*BS1	-0.1176	-9.7626	7.8098	-1.6660
LAS*BS2	0.2955	-10.2231	7.6494	-4.0610
LAS*BS3	-0.0789	29.3713	-22.4684	7.6312
EC*BS1	-6.7342	-4.4053	-9.5244	-5.1790
EC*BS2	13.3402	16.7373	10.0718	13.7513
EC*BS3	0.2203	-7.8057	9.5786	-1.9211
DC*BS1	4.3638	6.8838	-8.9675	-16.6235
DC*BS2	-8.1596	-12.6897	4.4525	18.4891
DC*BS3	0.1702	-1.2858	13.6687	-3.9009
LTC*BS1	0.5168	7.4642	-0.4365	-1.5019
LTC*BS2	-0.2634	5.9643	-1.2396	-0.6918
LTC*BS3	-0.2943	-20.1420	3.0839	4.6013
PG*BS1	4.5728	-7.8792	-2.9535	-14.6656
PG*BS2	-8.8460	8.6444	12.1966	17.9187
PG*BS3	-0.4662	7.5211	-6.0840	-0.1442
PO*BS1	4.3058	1.1708	-2.6946	-2.9451
PO*BS2	5.3817	2.5205	-1.3410	-0.8695
PO*BS3	-14.0478	-4.7044	6.9909	6.6970
IA*BS1	-4.0707	-6.0836	-2.5443	-14.9571
IA*BS2	11.6548	9.4060	12.0629	7.0675
IA*BS3	-4.8136	2.6808	-7.3658	14.5980
WO*BS1	-1.4864	1.9197	9.3658	6.2840
WO*BS2	0.0297	1.9355	10.2727	9.8228
WO*BS3	2.5711	-6.0792	-29.6629	-22.9714
EI*BS1	-2.2729	-6.7193	-2.3956	-16.1688
EI*BS2	9.0229	1.5751	8.0648	14.9218
EI*BS3	-4.6138	12.5958	-3.0655	0.7478
CH*BS1	-3.1542	-1.4625	8.2442	-9.3174
CH*BS2	4.6908	6.9379	15.1297	16.3334
CH*BS3	-0.2441	-4.9386	-31.8693	-16.8816
LKSP*BS1	-10.0443	-9.6325	-7.7345	-4.8625
LKSP*BS2	5.4769	4.9451	5.8818	8.2618
LKSP*BS3	13.6438	13.3259	8.7775	1.0181
NP*BS1	5.0723	-2.8521	-8.7564	-2.6961

NP*BS2	-9.5822	7.0731	2.8171	9.1329
NP*BS3	0.1702	-1.2858	13.6687	-3.9009
DF*BS1	4.1613	-1.3205	-1.5669	-0.7848
DF*BS2	2.8704	-2.4980	-1.9575	-1.8246
DF*BS3	-10.6921	5.6068	5.2828	3.9591
OT*BS1	0.3448	0.0423	-0.9968	0.0370
OT*BS2	0.2121	0.3184	-0.5816	0.8058
OT*BS3	25.9374	24.4940	27.8453	28.7170

The variables that are selected in each fold:

Table 28: Selected Variables of the Stepwise Selection for Classification.

	Selected Variables					
Fold 1						
LI	SCD	Dea	LAS	EC	DC	LTC
PG	PO	IA	WO	EI	LKSP	NP
DF	OT	City	BS	LI*BS	SCD*BS	Dea*BS
LAS*BS	EC*BS	DC*BS	LTC*BS	PG*BS	PO*BS	IA*BS
WO*BS	EI*BS	CH*BS				
Fold 2						
LI	SCD	Dea	LAS	EC	DC	LTC
PG	PO	IA	WO	EI	CH	LKSP
NP	DF	OT	City	BS	LI*BS	SCD*BS
Dea*BS	LAS*BS	EC*BS	DC*BS	LTC*BS	PG*BS	PO*BS
IA*BS	WO*BS	EI*BS	CH*BS	LKSP*BS	DF*BS	
Fold 3						
LI	SCD	Dea	LAS	EC	DC	LTC
PG	PO	IA	WO	EI	CH	LKSP
NP	OT	City	BS	LI*BS	SCD*BS	Dea*BS
LAS*BS	EC*BS	DC*BS	LTC*BS	PG*BS	PO*BS	IA*BS
WO*BS	EI*BS	CH*BS	LKSP*BS	NP*BS	DF*BS	
Fold 4						
LI	SCD	Dea	LAS	EC	DC	LTC
PO	IA	WO	LKSP	NP	OT	City
BS	LI*BS	SCD*BS	Dea*BS	LAS*BS	EC*BS	DC*BS
LTC*BS	PG*BS	EI*BS	CH*BS			

Fold 5

LI	SCD	Dea	LAS	EC	DC	LTC
PG	PO	IA	WO	EI	CH	LKSP
NP	DF	City	BS	LI*BS	SCD*BS	Dea*BS
LAS*BS	EC*BS	DC*BS	LTC*BS	PG*BS	PO*BS	IA*BS
WO*BS	EI*BS	CH*BS	LKSP*BS	NP*BS	DF*BS	OT*BS

The frequency and the relative frequency of the variables that have shown up on the above table:

Table 29: The Frequency of Selected Variables of the Stepwise Selection for Classification.

	Selected Variable	Frequency	Relative Frequency
1	CH*BS	5	1
2	DC*BS	5	1
3	Dea*BS	5	1
4	EC*BS	5	1
5	EI*BS	5	1
6	LAS*BS	5	1
7	LI*BS	5	1
8	LTC*BS	5	1
9	PG*BS	5	1
10	SCD*BS	5	1
11	BS	5	1
12	City	5	1
13	DC	5	1
14	Dea	5	1
15	EC	5	1
16	IA	5	1
17	LAS	5	1
18	LI	5	1
19	LKSP	5	1
20	LTC	5	1
21	NP	5	1
22	PO	5	1
23	SCD	5	1
24	WO	5	1
25	IA*BS	4	0.8
26	PO*BS	4	0.8
27	WO*BS	4	0.8
28	EI	4	0.8

29	OT	4	0.8
30	PG	4	0.8
31	DF*BS	3	0.6
32	LKSP*BS	3	0.6
33	CH	3	0.6
34	DF	3	0.6
35	NP*BS	2	0.4
36	OT*BS	1	0.2

Fold 4 selects the least numbers of predictors and Fold 5 selects the largest numbers of predictors. That is to say, 69% of variables are selected by Fold 4, 86% of variables are selected by Fold 1, 94 % of variables are selected by Fold 2 and Fold 3, and 97% of variables are selected by Fold 5. Although Fold 2 and 3 both select 34 out of 36 variables, variable NP*BS is not selected by Fold 2 and variable DF is not selected by Fold 3.

The MSE of the training set and test set for each fold:

Table 30: The MSE of the Stepwise Selection for Classification.

	MSE of Training	**MSE of Test Set**
Fold 1	0.0995	0.1137
Fold 2	0.0776	0.0928
Fold 3	0.0926	0.1194
Fold 4	0.0778	0.0916
Fold 5	0.0763	0.1132
Mean	0.0847	0.1061
SD	0.0106	0.0130

The rate of accuracy of each fold on the test set:

Table 31: The Rate of Accuracy of the Stepwise Selection for Classification.

	Rate of Accuracy
Fold 1	0.3363
Fold 2	0.2960
Fold 3	0.3239

Fold 4	0.3377
Fold 5	0.2983
Mean	0.3184
SD	0.0202

Among five folds, the rate of accuracy varies from the smallest 0.2983 to the largest 0.3377. Fold 4 has the smallest test MSE and the highest accuracy at the same time. Although Fold 3 has the largest MSE, Fold 2 has the smallest rate of accuracy.

Each MSE of training set is smaller than the MSE of the test. The mean MSE of the training data is smaller than the one of the test set of 0.0214. Hence, the stepwise selection model has a mild problem of over-fitting.

3.3.3 Elastic Net

We use the glmnet() and cv.glmnet() function, part of the glmnet package, to perform elastic net logistic regression with 5-fold external CV and 5-fold internal CV. The elastic net mixing parameter α is within the range $[0,1]$, so we conduct α as the set $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ beforehand. Training set is used for the modeling for each α, validation set is used to find the optimal tuning parameter λ by using the minimal MSE CV in one-standard-error rule, and the test set is used to evaluate the prediction performance for each α and λ in each fold.

For example, the estimated coefficients of the fifth fold with $\alpha = 0$ and it is optimal parameter $\lambda = 2.1899$:

Table 32: Estimated Coefficients of the Elastic Net Regularization for Classification.

	Estimated Coefficients				
	Class 0	**Class 0.25**	**Class 0.5**	**Class 0.75**	**Class 1**
(Intercept)	0.8265	1.2165	0.7663	-0.4623	-2.3470
LI	-0.0037	0.0012	-0.0255	0.0158	0.0122
SCD	-0.0317	-0.0291	0.0140	0.0256	0.0212
Dea	-0.0083	-0.0654	-0.0271	0.0550	0.0459
LAS	-0.0291	-0.0052	-0.0235	0.0455	0.0123
EC	-0.0114	-0.0304	-0.0087	0.0209	0.0296
DC	-0.0216	-0.0134	-0.0580	0.0460	0.0469
LTC	-0.0374	-0.0256	0.0102	0.0466	0.0062
PG	-0.0163	-0.0332	0.0184	0.0090	0.0221
PO	-0.0315	-0.0064	0.0108	0.0224	0.0047
IA	-0.0103	-0.0185	0.0146	-0.0141	0.0283
WO	-0.0283	0.0116	0.0049	-0.0012	0.0131
EI	-0.0329	-0.0283	0.0060	0.0114	0.0439
CH	-0.0032	-0.0507	0.0253	0.0196	0.0090
LKSP	0.0153	-0.0165	-0.0172	0.0090	0.0093
NP	-0.0213	0.0272	-0.0060	-0.0272	0.0274
DF	-0.0374	-0.0211	0.0204	0.0280	0.0101
OT	-0.0166	0.0306	0.0195	-0.0425	0.0090
City	0.0136	-0.0056	-0.0065	0.0013	0.0029
BS	-0.0263	-0.0392	-0.0025	0.0374	0.0306
LI*BS	-0.0056	-0.0072	-0.0260	0.0162	0.0226
SCD*BS	-0.0215	-0.0215	-0.0081	0.0308	0.0203
Dea*BS	-0.0136	-0.0279	-0.0176	0.0299	0.0291
LAS*BS	-0.0223	-0.0094	-0.0146	0.0315	0.0149
EC*BS	-0.0146	-0.0199	-0.0084	0.0176	0.0253
DC*BS	-0.0110	-0.0214	-0.0415	0.0437	0.0303
LTC*BS	-0.0256	-0.0220	-0.0050	0.0394	0.0131
PG*BS	-0.0133	-0.0308	0.0163	0.0038	0.0241
PO*BS	-0.0249	-0.0131	-0.0013	0.0262	0.0132
IA*BS	-0.0172	-0.0091	0.0105	-0.0062	0.0220
WO*BS	-0.0220	-0.0002	-0.0085	0.0115	0.0193
EI*BS	-0.0238	-0.0186	-0.0022	0.0108	0.0338
CH*BS	-0.0073	-0.0279	0.0119	0.0164	0.0069
LKSP*BS	-0.0079	-0.0129	0.0008	0.0105	0.0095
NP*BS	-0.0204	-0.0030	-0.0109	0.0059	0.0283
DF*BS	-0.0247	-0.0240	0.0098	0.0257	0.0132
OT*BS	-0.0157	0.0080	0.0121	-0.0165	0.0121

The optimal tuning parameter λ is selected corresponding to each α by choosing the minimal CV MSE with one-standard-error rule for each fold:

Table 33: The Optimal Tuning Parameter λ of the Elastic Net Regularization for Classification.

	Optimal Tuning Parameter λ					
	$\alpha = 0$	**$\alpha = 0.1$**	**$\alpha = 0.2$**	**$\alpha = 0.3$**	**$\alpha = 0.4$**	**$\alpha = 0.5$**
Fold 1	1.6183	0.2895	0.1095	0.1059	0.0872	0.0840
Fold 2	2.7733	0.3115	0.1558	0.1373	0.0855	0.0568
Fold 3	1.7046	0.2306	0.1266	0.1115	0.0762	0.0556
Fold 4	2.8630	0.2433	0.1335	0.0977	0.0882	0.0706
Fold 5	1.9902	0.2955	0.1622	0.0985	0.0811	0.0858
Mean	2.1899	0.2741	0.1375	0.1102	0.0836	0.0705
SD	0.5283	0.0314	0.0193	0.0145	0.0044	0.0129
	$\alpha = 0.6$	**$\alpha = 0.7$**	**$\alpha = 0.8$**	**$\alpha = 0.9$**	**$\alpha = 1$**	
Fold 1	0.0581	0.0498	0.0478	0.0353	0.0349	
Fold 2	0.0431	0.0488	0.0565	0.0605	0.0496	
Fold 3	0.0558	0.0478	0.0381	0.0281	0.0305	
Fold 4	0.0709	0.0459	0.0402	0.0357	0.0322	
Fold 5	0.0593	0.0463	0.0488	0.0476	0.0391	
Mean	0.0574	0.0477	0.0463	0.0415	0.0372	
SD	0.0089	0.0015	0.0066	0.0114	0.0068	

The MSE of the training set and test set of the paired α and λ in each fold:

Table 34: The MSE of the Elastic Net Regularization for Classification.

	MSE			
	$\alpha = 0$	**$\alpha = 0$**	**$\alpha = 0.1$**	**$\alpha = 0.1$**
Fold 1	0.0843	0.0741	0.0857	0.0741
Fold 2	0.0817	0.0849	0.0828	0.0858
Fold 3	0.0749	0.0805	0.0721	0.0843
Fold 4	0.0816	0.0880	0.0819	0.0883
Fold 5	0.0825	0.0888	0.0820	0.0890
Mean	0.0810	0.0833	0.0809	0.0843
SD	0.0036	0.0061	0.0052	0.0060

	$\alpha = 0.2$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.3$
Fold 1	0.0850	0.0752	0.0849	0.0755
Fold 2	0.0831	0.0852	0.0821	0.0858
Fold 3	0.0717	0.0843	0.0843	0.0815
Fold 4	0.0823	0.0880	0.0828	0.0880
Fold 5	0.0820	0.0893	0.0816	0.0898
Mean	0.0808	0.0844	0.0832	0.0841
SD	0.0053	0.0055	0.0014	0.0057

	$\alpha = 0.4$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.5$
Fold 1	0.0853	0.0722	0.0853	0.0719
Fold 2	0.0831	0.0852	0.0834	0.0886
Fold 3	0.0715	0.0840	0.0715	0.0840
Fold 4	0.0830	0.0883	0.0830	0.0883
Fold 5	0.0820	0.0893	0.0820	0.0893
Mean	0.0810	0.0838	0.0811	0.0844
SD	0.0054	0.0068	0.0055	0.0073

	$\alpha = 0.6$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.7$
Fold 1	0.0852	0.0725	0.0852	0.0725
Fold 2	0.0796	0.0852	0.0830	0.0866
Fold 3	0.0785	0.0820	0.0785	0.0820
Fold 4	0.0822	0.0883	0.0829	0.0880
Fold 5	0.0820	0.0893	0.0821	0.0911
Mean	0.0815	0.0834	0.0823	0.0840
SD	0.0026	0.0068	0.0024	0.0073

	$\alpha = 0.8$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 0.9$
Fold 1	0.0852	0.0725	0.0849	0.0758
Fold 2	0.0831	0.0866	0.0827	0.0858
Fold 3	0.0785	0.0820	0.0721	0.0883
Fold 4	0.0829	0.0880	0.0829	0.0880
Fold 5	0.0820	0.0893	0.0820	0.0893
Mean	0.0823	0.0837	0.0809	0.0854
SD	0.0024	0.0068	0.0051	0.0055

	$\alpha = 1$	$\alpha = 1$
Fold 1	0.0852	0.0725
Fold 2	0.0831	0.0866
Fold 3	0.0785	0.0820
Fold 4	0.0829	0.0880

Fold 5	0.0820	0.0893
Mean	0.0823	0.0837
SD	0.0024	0.0068

The rate of accuracy for each fold regularized by elastic net:

Table 35: The Rate of Accuracy of the Elastic Net Regularization for Classification.

	Rate of Accuracy			
	$\boldsymbol{\alpha = 0}$	$\boldsymbol{\alpha = 0.1}$	$\boldsymbol{\alpha = 0.2}$	$\boldsymbol{\alpha = 0.3}$
Fold 1	0.3142	0.3319	0.3009	0.3230
Fold 2	0.2870	0.2870	0.2960	0.2870
Fold 3	0.3320	0.3158	0.3158	0.3279
Fold 4	0.2632	0.2675	0.2719	0.2719
Fold 5	0.2689	0.2563	0.2521	0.2563
Mean	0.2930	0.2917	0.2873	0.2932
SD	0.0295	0.0318	0.0252	0.0314
	$\boldsymbol{\alpha = 0.4}$	$\boldsymbol{\alpha = 0.5}$	$\boldsymbol{\alpha = 0.6}$	$\boldsymbol{\alpha = 0.7}$
Fold 1	0.3274	0.3319	0.3230	0.3230
Fold 2	0.2960	0.2780	0.2780	0.2960
Fold 3	0.3198	0.3198	0.3279	0.3279
Fold 4	0.2675	0.2675	0.2675	0.2719
Fold 5	0.2521	0.2521	0.2521	0.2563
Mean	0.2926	0.2899	0.2897	0.2950
SD	0.0325	0.0344	0.0340	0.0312
	$\boldsymbol{\alpha = 0.8}$	$\boldsymbol{\alpha = 0.9}$	$\boldsymbol{\alpha = 1}$	
Fold 1	0.3230	0.3186	0.3230	
Fold 2	0.2960	0.2870	0.2960	
Fold 3	0.3279	0.3320	0.3279	
Fold 4	0.2719	0.2719	0.2719	
Fold 5	0.2521	0.2521	0.2521	
Mean	0.2942	0.2923	0.2942	
SD	0.0326	0.0329	0.0326	

The smallest mean among the test MSE is 0.0833 with $\alpha = 0$ (ridge) and the largest is 0.0854 with $\alpha = 0.9$. For the accuracy, the highest mean of the rate is 0.2950 with $\alpha =$

0.7 and the lowest mean rate is 0.2873 with $\alpha = 0.2$. The Elastic net approach does not have serious problems of over-fitting. The mean of the MSE and the mean of the accuracy rate do not change much among the elastic net models with the different tuning parameters.

3.3.4 Ordinal Logistic Regression

In this section, we use polr() function, part of the MASS package, to fit the ordinal logistic regression. We use the variables with non-zero coefficients selected by lasso from the above elastic net section to fit the model, that is, the average of the optimal tuning parameter $\lambda = 0.0372$ when $\alpha = 1$. As was the case with the multinomial logistic regression, there is no tuning process for the ordinal logistic itself. Therefore, we only use external 5-fold CV alone, and we obtain five folds from five external CV iterations. All the predictors and the response are not only categorized but also ordered.

The estimated coefficients of the elastic net regression with $\alpha = 1$ (lasso) and the average of the optimal parameter $\lambda = 0.0372$:

Table 36: Estimated Coefficients of the Elastic Net Regularization with $\alpha = 1$ and $\lambda = 0.0372$.

	Estimated Coefficients				
	Class 0	**Class 0.25**	**Class 0.5**	**Class 0.75**	**Class 1**
(Intercept)	0.1496	0.8063	0.7139	-0.1729	-1.4970
LI	0	0	0	0	0
SCD	0	0	0	0	0
Dea	0	0	0	0	0
LAS	0	0	0	0	0
EC	0	0	0	0	0
DC	0	0	0	0	0
LTC	-0.0020	0	0	0	0

PG	0	0	0	0	0
PO	0	0	0	0	0
IA	0	0	0	0	0
WO	0	0	0	0	0
EI	0	0	0	0	0
CH	0	0	0	0	0
LKSP	0	0	0	0	0
NP	0	0	0	0	0
DF	0	0	0	0	0
OT	0	0	0	-0.0614	0
City	0.0582	0	0	0	0
BS	0	-0.1338	0	0.0965	0.0108
LI*BS	0	0	0	0	0
SCD*BS	0	0	0	0.0197	0
Dea*BS	0	0	0	0	0
LAS*BS	0	0	0	0.0254	0
EC*BS	0	0	0	0	0
DC*BS	0	0	0	0	0
LTC*BS	0	0	0	0.2835	0
PG*BS	0	0	0	0	0
PO*BS	-0.0838	0	0	0.0085	0
IA*BS	0	0	0	0	0
WO*BS	0	0	0	0	0
EI*BS	0	0	0	0	0
CH*BS	0	0	0	0	0
LKSP*BS	0	0	0	0	0
NP*BS	0	0	0	0	0
DF*BS	-0.0115	-0.0008	0	0	0
OT*BS	0	0	0	0	0

There are nine variables with non-zero coefficient: LTC, PT, City, BS, SCD*BS, LAS*BS, LTC*BS, PO*BS and DF*BS. We use these variables as the predictors to fit the ordinal logistic regression model.

For example, the estimated coefficients of the fifth fold:

Table 37: Estimated Coefficients of the Ordinal Logistic Regression with Nine Predictors.

	Estimated Coefficients			
(Intercepts)	**0|0.25**	**0.25|0.5**	**0.5|0.75**	**0.75|1**
	-2.8619263	-1.3758656	0.1751465	2.6842740
BS.L	1.40769	1.40769	1.40769	1.40769
BS.Q	-0.33938	-0.33938	-0.33938	-0.33938
OT.L	-0.15998	-0.15998	-0.15998	-0.15998
LTC.L	0.39519	0.39519	0.39519	0.39519
City.L	-0.61533	-0.61533	-0.61533	-0.61533
City.Q	-0.58349	-0.58349	-0.58349	-0.58349
City.C	-0.44241	-0.44241	-0.44241	-0.44241
City^4	-0.49469	-0.49469	-0.49469	-0.49469
City^5	-0.14903	-0.14903	-0.14903	-0.14903
BS1:SCD.L	0.11565	0.11565	0.11565	0.11565
BS2:SCD.L	0.73199	0.73199	0.73199	0.73199
BS3:SCD.L	0.62274	0.62274	0.62274	0.62274
BS1:LAS.L	0.18897	0.18897	0.18897	0.18897
BS2:LAS.L	0.04373	0.04373	0.04373	0.04373
BS3:LAS.L	0.29222	0.29222	0.29222	0.29222
BS.L:LTC.L	0.11064	0.11064	0.11064	0.11064
BS.Q:LTC.L	0.47386	0.47386	0.47386	0.47386
BS1:DF.L	0.06806	0.06806	0.06806	0.06806
BS2:DF.L	0.63018	0.63018	0.63018	0.63018
BS3:DF.L	0.68357	0.68357	0.68357	0.68357
BS1:PO.L	0.16998	0.16998	0.16998	0.16998
BS2:PO.L	0.29146	0.29146	0.29146	0.29146
BS3:PO.L	1.11456	1.11456	1.11456	1.11456

The MSE of the training set and test set for each fold:

Table 38: The MSE of the Ordinal Logistic Regression.

	MSE of Training	**MSE of Test Set**
Fold 1	0.0731	0.0829
Fold 2	0.0754	0.0699
Fold 3	0.0725	0.0870
Fold 4	0.0784	0.0804
Fold 5	0.0733	0.0792
Mean	0.0745	0.0799
SD	0.0024	0.0063

The rate of accuracy of each fold on the test set:

Table 39: The Rate of Accuracy of the Ordinal Logistic Regression.

	Rate of Accuracy
Fold 1	0.3392
Fold 2	0.3347
Fold 3	0.2952
Fold 4	0.3722
Fold 5	0.3833
Mean	0.3449
SD	0.0348

Although Fold 2 has a significant small test MSE, Fold 5 has the highest accuracy rate among five folds. Fold 3 has the largest test MSE and lowest accuracy rate at the same time.

3.3.5 Linear Discriminant Analysis (LDA)

In this section, we use lda() function, part of the MASS package, to fit the LDA model. As in multinomial logistic regression and ordinal logistic regression, there is no tuning process for the LDA itself. Therefore, we only use external 5-fold CV alone, and we obtain five folds from five external CV iterations

The estimated coefficients of linear discriminants for the fifth fold:

Table 40: Estimated Coefficients of the LDA.

Estimated Coefficients				
	LD1	**LD2**	**LD3**	**LD4**
LI	-0.8645	-0.8570	0.0522	-1.0168
SCD	0.2544	0.5773	0.6185	0.7205
Dea	-0.2489	-0.7644	-0.2229	0.1503
LAS	0.0923	-0.0390	-0.2422	-0.4587
EC	0.2482	-0.1095	0.1354	0.7525
DC	-0.0759	0.1389	-0.0803	-0.3009
LTC	0.4304	-0.3732	1.2065	1.0667
PG	-0.2090	-0.4372	-0.1349	-0.3491
PO	-0.0325	-0.5482	-0.1967	-0.6300
IA	-0.0684	-0.0328	-0.3766	0.1708
WO	-0.3591	0.3294	0.4905	0.6619
EI	-0.2211	0.0048	0.5995	0.3123
CH	0.0418	-0.0662	-0.2530	-0.2092
LKSP	0.2451	0.2124	-0.7976	0.1620
NP	-0.4223	0.4638	0.3422	-0.0482
DF	0.0162	0.1751	0.1229	-0.2584
OT	-0.1721	0.4487	0.3583	-0.5929
City	-0.2206	0.1420	-0.5108	0.4042
BS	0.1851	-0.1600	0.1363	0.1095
LI*BS	1.0440	1.1597	-0.1478	0.8571
SCD*BS	-0.1772	-0.7659	-0.6140	-0.6022
Dea*BS	0.4930	0.8345	0.0468	-0.0179
LAS*BS	-0.0995	-0.1144	0.2499	0.1415
EC*BS	-0.0648	0.1358	-0.0848	-0.7563
DC*BS	0.1363	-0.0923	-0.1351	0.1658
LTC*BS	-0.2319	-0.0934	-1.0805	-1.2341
PG*BS	0.3548	0.4746	0.1662	0.5901
PO*BS	0.1885	0.4053	0.4839	0.5758
IA*BS	-0.0209	0.1828	0.6085	-0.2734
WO*BS	0.4641	-0.3098	-0.6177	-0.7801
EI*BS	0.4599	0.1142	-0.5859	-0.1821
CH*BS	-0.1184	0.0573	0.3915	0.4907
LKSP*BS	-0.4034	-0.2911	0.8353	-0.2863
NP*BS	0.5777	-0.1860	-0.3366	-0.0701
DF*BS	0.0500	-0.3895	0.0370	0.5507
OT*BS	0.2892	0.0247	-0.1126	0.4055

The proportion of the between-class variation explained by each discriminant function in the fifth fold:

Table 41: Proportion of Trace of the LDA.

Proportion of Trace			
LD1	**LD2**	**LD3**	**LD4**
0.5197	0.2535	0.1477	0.0791

Thus, the first and the second linear discriminant achieve about 52% and 25% of the separation respectively.

We can obtain a scatterplot of the best two discriminant functions:

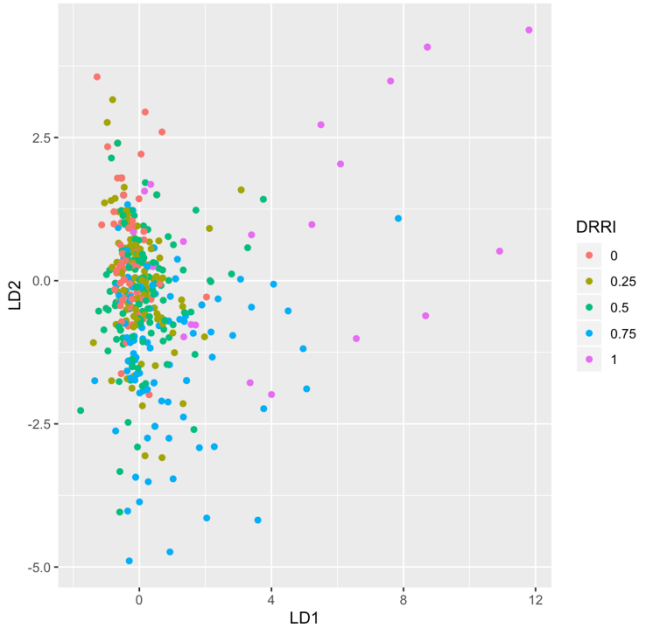

Figure 6: Scatterplot of the First Two Discriminant Functions.

Although the class overlap is quite considerable from the plot, the separation of class 0.75 (blue dot) and class 0 (red dot) is less ambiguous than the separation of class 0 and class

0.25 (yellow dot). Class 1 (pink dot) and Class 0.75 are more scattered compared to the rest of three classes.

The MSE of the training set and test set for each fold:

Table 42: The MSE of the LDA.

	MSE of Training	**MSE of Test Set**
Fold 1	0.0833	0.0976
Fold 2	0.0825	0.0939
Fold 3	0.0811	0.0967
Fold 4	0.0797	0.0938
Fold 5	0.0809	0.1095
Mean	0.0815	0.0983
SD	0.0014	0.0065

The rate of accuracy of each fold on the test set:

Table 43: The Rate of Accuracy of the LDA.

	Rate of Accuracy
Fold 1	0.3363
Fold 2	0.3094
Fold 3	0.3603
Fold 4	0.3465
Fold 5	0.3109
Mean	0.3327
SD	0.0223

Although Fold 4 has the smallest MSE of the test set, Fold 3 has the highest rate of accuracy. Although Fold 5 has the largest MSE of the test set, Fold 2 has the lowest rate of accuracy. As we compare the MSE of the training set to the MSE of the test set, LDA has a mild over-fitting problem.

3.4 Machine Learning Methods and Ensemble Learning

Contrary to popular belief, the machine learning approach has been around for many decades and has become a very rapidly moving field. The machine learning methods are favored and famous because they are designed to make the maximized accuracy possible instead of making the prediction model more interpretable.

In this section, we would like to know if the machine learning method can outperform the statistical methods and how much the machine learning method could boost the predictive performance on our data set. Hence, our purpose in this section is to focus on predictive performance. We will provide the output of the selected parameters, the MSE, and the rate of accuracy for each machine-learning algorithm.

The machine-learning algorithm of C5.0, GBM, KNN, NN, random forest and SVM are used to fit the data. We use stacking, an ensemble learning approach to find out if it can yield better performance than using a single model. The output DRRI is considered as the categorical output. Therefore, in this section, the machine learning approaches are utilized for classification.

The R package we use is a highly functional package called “caret”. We can fit the models by specifying the name of the methods using the train() function that is contained in caret package.

3.4.1 C5.0

We assign the argument of the method to “C5.0” in function train() to fit the C5.0 model. The internal 5-fold CV is used to choose the optimal parameters by evaluating the highest rate of accuracy in each validation set. The external 5-fold CV plays a role in the prediction of each chosen model in order to achieve the model performance.

The optimal parameters selected by interval CV for each fold:

Table 44: The Optimal Parameters of the C5.0.

	Parameters		
	trials	**model**	**winnow**
Fold 1	1	tree	TRUE
Fold 2	1	rules	FALSE
Fold 3	1	rules	FALSE
Fold 4	1	tree	FALSE
Fold 5	1	rules	TRUE

The “trails” is the boosting iterations, the “model” stands for the model type and “winnow” refers to the mechanism by analogy with the process for separating the wheat from the chaff.

The MSE of the training set and the test set:

Table 45: The MSE of the C5.0.

	MSE of Training	**MSE of Test Set**
Fold 1	0.0795	0.0843
Fold 2	0.0771	0.1009
Fold 3	0.0814	0.1053
Fold 4	0.0775	0.0984

Fold 5	0.0762	0.0932
Mean	0.0783	0.0964
SD	0.0021	0.0080

The rate of accuracy of each fold on the test set:

Table 46: The Rate of Accuracy of the C5.0.

	Rate of Accuracy
Fold 1	0.3186
Fold 2	0.3677
Fold 3	0.3563
Fold 4	0.3509
Fold 5	0.3613
Mean	0.3510
SD	0.0191

Fold 2 has the highest rate of accuracy with trails = 1, rule-based model and winnow of false. By comparison, the MSE of the training set is obviously lower than the MSE of the test set, the models are apparently over-fitting the data.

3.4.2 Gradient Boosting Machines (GBM)

We assign the argument of the method to “gbm” in function train() to fit the GBM model. The internal 5-fold CV is used to choose the optimal parameters by evaluating the highest rate of accuracy in each validation set. The external 5-fold CV plays a role in the prediction of each chosen model in order to achieve the model performance.

The optimal parameters selected by interval CV for each fold:

Table 47: The Optimal Parameters of the GBM.

	Parameters			
	n.trees	**interaction.depth**	**shrinkage**	**n.minobsinnode**
Fold 1	100	3	0.1	10
Fold 2	150	3	0.1	10
Fold 3	150	3	0.1	10
Fold 4	150	3	0.1	10
Fold 5	150	3	0.1	10

The "n.trees" indicates the number of gradient boosting iteration, "interaction.depth" is the number of splits of a tree, "shrinkage" is considered as a learning rate and "n.minobsinnode" stands for the minimum size of trees' terminal nodes.

The MSE of the training set and the test set:

Table 48: The MSE of the GBM.

	MSE of Training	**MSE of Test Set**
Fold 1	0.0802	0.0882
Fold 2	0.0749	0.0902
Fold 3	0.0714	0.1002
Fold 4	0.0701	0.0855
Fold 5	0.0671	0.0922
Mean	0.0727	0.0913
SD	0.0050	0.0056

The rate of accuracy of each fold on the test set:

Table 49: The Rate of Accuracy of the GBM.

	Rate of Accuracy
Fold 1	0.3761
Fold 2	0.3857
Fold 3	0.3522
Fold 4	0.3553

Fold 5	0.3403
Mean	0.3619
SD	0.0185

Again, Fold 2 has the highest rate of accuracy with parameters of n.trees = 100, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10. By comparison, the MSE of the training set is obviously lower than the MSE of the test set, the models have a seriously over-fitting problem.

3.4.3 K-Nearest Neighbors (KNN)

We assign the argument of the method to "knn" in function train() to fit the KNN model. The internal 5-fold CV is used to choose the optimal parameter by evaluating the highest rate of accuracy in each validation set. The external 5-fold CV plays a role in the prediction of each chosen model in order to achieve the model performance. The computation time of KNN is less than the above two approaches.

The optimal parameter selected by interval CV for each fold:

Table 50: The Optimal Parameter of the KNN.

	Parameter k
Fold 1	5
Fold 2	9
Fold 3	5
Fold 4	5
Fold 5	9

The "k" is the number of nearest neighbors.

The MSE of the training set and the test set:

Table 51: The MSE of the KNN.

	MSE of Training	MSE of Test Set
Fold 1	0.0921	0.0794
Fold 2	0.0886	0.0883
Fold 3	0.0842	0.1182
Fold 4	0.0849	0.1105
Fold 5	0.0824	0.0961
Mean	0.0864	0.0985
SD	0.0039	0.0159

The rate of accuracy of each fold on the test set:

Table 52: The Rate of Accuracy of the KNN.

	Rate of Accuracy
Fold 1	0.3938
Fold 2	0.3318
Fold 3	0.3198
Fold 4	0.3289
Fold 5	0.3361
Mean	0.3421
SD	0.0295

Fold 1 has the highest rate of accuracy with k = 5. By comparison, the MSE of the training set is lower than the MSE of the test set and there is a concern of the model over-fitting.

3.4.4 Neural Networks (NN)

We assign the argument of the method to “nnet” in function train() to fit the NN model. The internal 5-fold CV is used to choose the optimal parameters by evaluating the highest

rate of accuracy in each validation set. The external 5-fold CV plays a role in the prediction of each chosen model in order to achieve the model performance. The highest computational cost incurs for implementing the NN model among all used machine learning methods.

The optimal parameters selected by interval CV for each fold:

Table 53: The Optimal Parameters of the NN.

	Parameters	
	size	**decay**
Fold 1	9	0.6
Fold 2	9	0.6
Fold 3	9	0.6
Fold 4	9	0.6
Fold 5	9	0.6

The “size” is the number of hidden units and “decay” means the weights of the regularization. All five folds select the same parameters by the internal CV.

The MSE of the training set and the test set:

Table 54: The MSE of the NN.

	MSE of Training	**MSE of Test Set**
Fold 1	0.0805	0.0669
Fold 2	0.0789	0.0732
Fold 3	0.0770	0.0810
Fold 4	0.0779	0.0776
Fold 5	0.0748	0.0895
Mean	0.0778	0.0776
SD	0.0021	0.0085

The rate of accuracy of each fold on the test set:

Table 55: The Rate of Accuracy of the NN.

	Rate of Accuracy
Fold 1	0.5531
Fold 2	0.5112
Fold 3	0.5061
Fold 4	0.4912
Fold 5	0.4202
Mean	0.4964
SD	0.0483

Fold 1 has the smallest MSE of the test set and the highest rate of accuracy at the same time with parameters of size = 9 and decay = 0.6. Although NN model provides the surprisingly high rate of accuracy among all the approaches so far, there is no indications of over-fitting or under-fitting by comparing the mean of the training MSE with the mean of the test MSE.

3.4.5 Random Forest

We assign the argument of the method to “rf” in function train() to fit the random forest model. The internal 5-fold CV is used to choose the optimal parameter by evaluating the highest rate of accuracy in each validation set. The external 5-fold CV plays a role in the prediction of each chosen model in order to achieve the model performance.

The optimal parameter selected by interval CV for each fold:

Table 56: The Optimal Parameter of the Random Forest.

	Parameter mtry
Fold 1	2
Fold 2	19
Fold 3	19
Fold 4	2
Fold 5	19

The “mtry” defines the number of randomly selected predictors that are available at each split.

The MSE of the training set and the test set:

Table 57: The MSE of the Random Forest.

	MSE of Training	MSE of Test Set
Fold 1	0.0712	0.0702
Fold 2	0.0542	0.1006
Fold 3	0.0538	0.1154
Fold 4	0.0640	0.0844
Fold 5	0.0490	0.0924
Mean	0.0585	0.0926
SD	0.0090	0.0170

The rate of accuracy of each fold on the test set:

Table 58: The Rate of Accuracy of the Random Forest.

	Rate of Accuracy
Fold 1	0.3407
Fold 2	0.3274
Fold 3	0.3522
Fold 4	0.3377
Fold 5	0.3067
Mean	0.3329
SD	0.0171

Fold 3 has the largest test MSE, however, it also has the highest rate of accuracy. From the MSE table, we can see that random forest models seriously over-fit the training set. The test MSE in Fold 3 is more than the double of the training MSE.

3.4.6 Support Vector Machines (SVM)

We assign the argument of the method to “svmRadial” in function train() to fit the SVM model, that is, the SVM with radial basis function kernel. The internal 5-fold CV is used to choose the optimal parameter by evaluating the highest rate of accuracy in each validation set. The external 5-fold CV plays a role in the prediction of each chosen model in order to achieve the model performance.

The optimal parameters selected by interval CV for each fold:

Table 59: The Optimal Parameters of the SVM.

	Parameters	
	sigma	**C**
Fold 1	0.0523	1.0
Fold 2	0.0539	1.0
Fold 3	0.0530	1.0
Fold 4	0.0548	0.5
Fold 5	0.0530	1.0

The "sigma" is the parameter in the radial basis function (RBF) kernel determines the SVM decision boundary. The "C" refers to the cost function, which controls the training errors and margins.

The MSE of the training set and the test set:

Table 60: The MSE of the SVM.

	MSE of Training	MSE of Test Set
Fold 1	0.0693	0.0821
Fold 2	0.0668	0.0869
Fold 3	0.0572	0.0987
Fold 4	0.0697	0.0803
Fold 5	0.0592	0.0927
Mean	0.0644	0.0881
SD	0.0059	0.0076

The rate of accuracy of each fold on the test set:

Table 61: The Rate of Accuracy of the SVM.

	Rate of Accuracy
Fold 1	0.3363
Fold 2	0.3587
Fold 3	0.3198
Fold 4	0.3465
Fold 5	0.3193
Mean	0.3361
SD	0.0171

Although Fold 4 has the smallest test MSE, Fold 2 has the highest rate of accuracy. There is an over-fitting of the SVM model.

3.4.7 Stacking

In Chapter II, we introduced the ensemble learning of bagging, boosting, and stacking. For example, the machine learning method of random forest is a bagging algorithm and C5.0 and GBM are the boosting algorithms.

Our purpose of using ensemble learning is for a better prediction performance and indeed, several machine learning models have shown better performance in the above section. Moreover, we will utilize the stacking to combine multiple models in order to have better performance than any single model in the ensemble.

To carry out the stacking algorithm, we choose KNN and GBM approaches as the bottom layer models and the multinomial logistic regression approach as the top layer model. The external 5-fold CV splits the whole data set into a modeling set and test set and then the internal 5-fold CV splits the modeling set into a training set and validation set. Each bottom layer model is trained with the training set, and the optimal bottom layer model is chosen by the highest prediction accuracy in the validation set. The predicted outputs of these two optimal models (KNN and GBM) in the modeling set become the two predictors for the top layer model (multinomial logistic regression). Finally, the standby test set is used to evaluate the performance of the stacking ensemble.

The MSE of the training set and the test set for the bottom layer models and the top layer model:

Table 62: The MSE of the Bottom and Top Layer Models of the Stacking Ensemble.

	MSE of KNN		MSE of GBM	
	Training Set	**Test Set**	**Training Set**	**Test Set**
Fold 1	0.0921	0.0794	0.0802	0.0882
Fold 2	0.0886	0.0883	0.0749	0.0902
Fold 3	0.0842	0.1182	0.0714	0.1002
Fold 4	0.0849	0.1105	0.0701	0.0855
Fold 5	0.0824	0.0961	0.0671	0.0922
Mean	0.0864	0.0985	0.0727	0.0913
SD	0.0039	0.0159	0.005	0.0056

	MSE of Stacking	
	Training Set	**Test Set**
Fold 1	0.0863	0.0902
Fold 2	0.0778	0.0959
Fold 3	0.0737	0.1063
Fold 4	0.0712	0.0951
Fold 5	0.0708	0.0956
Mean	0.0759	0.0966
SD	0.0064	0.0059

The rate of accuracy of each fold on the test set:

Table 63: The Rate of Accuracy of the Bottom and Top Layer Models of the Stacking Ensemble.

	Rate of Accuracy		
	KNN	**GBM**	**Stacking**
Fold 1	0.3938	0.3761	0.4027
Fold 2	0.3318	0.3857	0.3946
Fold 3	0.3198	0.3522	0.3441
Fold 4	0.3289	0.3553	0.3509
Fold 5	0.3361	0.3403	0.3487
Mean	0.3421	0.3619	0.3682
SD	0.0295	0.0185	0.0280

As we compare the MSE of the training set to the MSE of the test set, the bottom layer model of GBM has over-fitted the data and the multinomial logistic regression, the top

layer model's over-fitting is ever worse. As we compare the accuracy rate for each layer's fold, even though the accuracy rate of stacking cannot be guaranteed as the highest among the three folds, it is not the lowest, either. Overall, the stacking ensemble improves the prediction performance of our data as it has the highest mean of the accuracy rate compared to the single bottom layer models.

CHAPTER IV. DISCUSSION

So far, all the proposed approaches were regressed onto the private sector participation data set for the analysis, and the outputs were illustrated in the above section for each model. Although there is no "best" model for the data, we are able to understand how much difference between the true response and the predicted output of our models by the evaluation of prediction performance. Other than that, with variable selection, we can explain the data in the simplest way by reducing the redundant predictors since they may add noise to the estimation that we are interested in.

We have two aims in this section. First, we compare the MSE and the rate of accuracy among all the models for an overall understanding of the prediction performance. Second, we analyze the significant variables of the multiple linear regression model and multinomial logistic model, the selected variables of the stepwise selection model, and the variables with non-zero estimated coefficients of the lasso model for a better understanding of the variables that are relatively important for the model.

4.1 Comparison on Prediction Performance

1. Mean Square Error (MSE)

We calculated the test MSE for each prediction model in Chapter III. To be specific, we used the MSE of mapping in linear regression which was used to simulate the predicted classes in order to compare the MSE of the test in classification approaches and machine learning methods. There were five MSE for each approach since there were five folds

yielded from 5-fold CV resampling. The mean of all five MSE was considered as the true prediction MSE of the approach.

Table 64: The MSE of 16 Prediction Models.

	MSE			
	Multiple Linear Regression	**Stepwise Linear Regression**	**Lasso Linear Regression**	**PCR Linear Regression**
Fold 1	0.0644	0.0783	0.0747	0.0595
Fold 2	0.0695	0.0913	0.0687	0.0706
Fold 3	0.0855	0.0784	0.0810	0.0822
Fold 4	0.0792	0.0768	0.0789	0.0740
Fold 5	0.0811	0.0676	0.0872	0.0785
Mean	0.0760	0.0785	0.0781	0.0730
SD	0.0087	0.0084	0.0069	0.0087
	Multinomial Classification	**Stepwise Classification**	**Lasso Classification**	**Ordinal Classification**
Fold 1	0.1289	0.1137	0.0725	0.0829
Fold 2	0.0956	0.0928	0.0866	0.0699
Fold 3	0.1217	0.1194	0.0820	0.0870
Fold 4	0.0962	0.0916	0.0880	0.0804
Fold 5	0.1132	0.1132	0.0893	0.0792
Mean	0.1111	0.1061	0.0837	0.0799
SD	0.0150	0.0130	0.0068	0.0063
	LDA Classification	**C5.0 Machine Learning**	**GBM Machine Learning**	**KNN Machine Learning**
Fold 1	0.0976	0.0843	0.0882	0.0794
Fold 2	0.0939	0.1009	0.0902	0.0883
Fold 3	0.0967	0.1053	0.1002	0.1182
Fold 4	0.0938	0.0984	0.0855	0.1105
Fold 5	0.1095	0.0932	0.0922	0.0961
Mean	0.0983	0.0964	0.0913	0.0985
SD	0.0065	0.0080	0.0056	0.0159

	NN Machine Learning	RF Machine Learning	SVM Machine Learning	Stacking Machine Learning
Fold 1	0.0669	0.0702	0.0821	0.0902
Fold 2	0.0732	0.1006	0.0869	0.0959
Fold 3	0.0810	0.1154	0.0987	0.1063
Fold 4	0.0776	0.0844	0.0803	0.0951
Fold 5	0.0895	0.0924	0.0927	0.0956
Mean	0.0776	0.0926	0.0881	0.0966
SD	0.0085	0.0170	0.0076	0.0059

From the above table, we have MSE of four linear regression models, five classification models, and seven machine learning models. For the linear regression models, the lasso has the smallest standard deviation of 0.0069, in contrast, the multiple linear regression and the PCR have the largest standard deviation of 0.0087 at the same time. On the other hand, the PCR also has the smallest mean MSE of 0.0730 and stepwise regression has the largest mean MSE of 0.0785. For the statistical approaches of classification, the ordinal logistic regression with variables selected by lasso has the smallest standard deviation of 0.0063 and the multinomial logistic regression has the largest standard deviation of 0.0150. Other than that, the ordinal logistic regression with variables selected by lasso has the smallest mean MSE of 0.0799 and the multinomial logistic regression has the largest mean MSE of 0.1111. For the machine learning approaches, the GBM model gives the smallest standard deviation of 0.0056 and the random forest model provides the largest standard deviation of 0.0170. Moreover, the NN model has the smallest mean MSE of 0.0776 and the largest mean MSE is from the KNN model.

Among all sixteen models, the GBM is with the smallest standard deviation of 0.0056 and the PCR provides the smallest mean MSE of 0.0730. In contrast, the random forest

model obtains the largest standard deviation of 0.0170 and the largest mean MSE of 0.1111 is from the multinomial logistic regression.

A boxplot graph can summarize the MSE data and show its interval scale and variability:

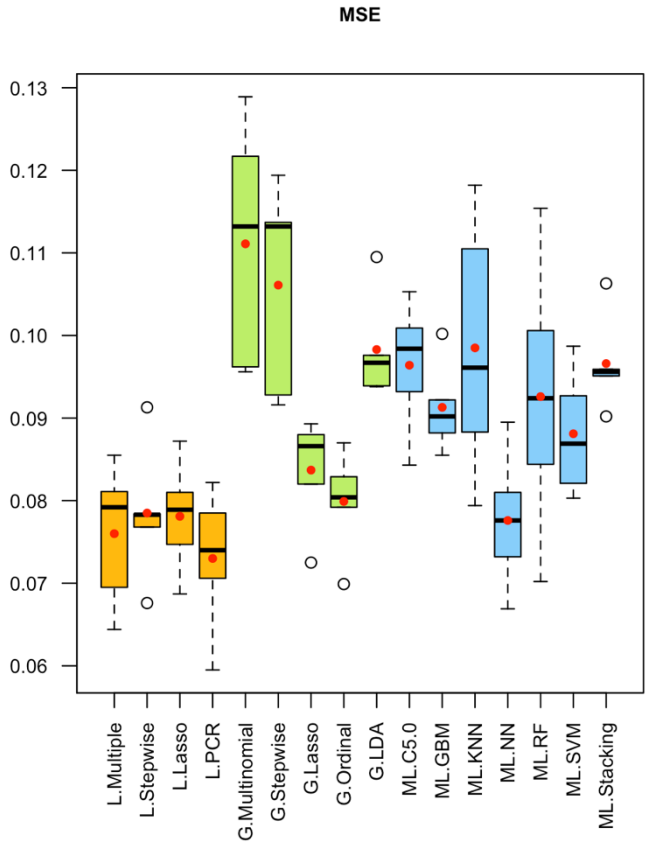

Figure 7: The Boxplots of the MSE. The Linear Regression Models, the Classification Regression Models and the Machine Learning Models are Showed in Orange, Green, and Blue.

The boxplots of the MSE of the linear regression models, the classification regression models, and the machine learning models are showed in orange, green, and blue respectively. The red dots indicate the mean of the MSE for each approach. The mean of the MSE in the linear regression models is lower than the classification models and machine learning models. We noticed that the multinomial logistic regression, the

stepwise regression for classification, the random forest, and the KNN indicate a wider range of MSE than the rest of the models do. The range of the stacking model is quite small.

The stepwise selection enhances the mean MSE for the linear regression and decreases the mean MSE for the classification, however, the range of the MSE does not change much. The lasso shrinkage approach decreases the mean MSE and the range of the MSE compares to the multinomial logistic regression, but it increases the mean MSE for the linear regression.

2. Classification Accuracy Rate

We calculated the rate of accuracy of the test set in Chapter III to provide another way of understanding the model performance for each prediction model. As in MSE, there were five accuracy rates for each approach since there were five folds yielded from 5-fold CV resampling. The mean of all five rates was considered as the true prediction accuracy rate of each approach.

Table 65: The Rates of Accuracy of 16 Prediction Models.

	Accuracy Rate			
	Multiple Linear Regression	**Stepwise Linear Regression**	**Lasso Linear Regression**	**PCR Linear Regression**
Fold 1	0.4204	0.3188	0.4027	0.4336
Fold 2	0.2870	0.2881	0.3274	0.3229
Fold 3	0.3117	0.3333	0.3158	0.2996
Fold 4	0.2895	0.3568	0.2807	0.2982
Fold 5	0.3193	0.3273	0.2605	0.3361
Mean	0.3256	0.3249	0.3174	0.3381
SD	0.0548	0.0250	0.0547	0.0558

	Multinomial Classification	Stepwise Classification	Lasso Classification	Ordinal Classification
Fold 1	0.2920	0.3363	0.3230	0.3392
Fold 2	0.3094	0.2960	0.2960	0.3347
Fold 3	0.3239	0.3239	0.3279	0.2952
Fold 4	0.3465	0.3377	0.2719	0.3722
Fold 5	0.2983	0.2983	0.2521	0.3833
Mean	0.3140	0.3184	0.2942	0.3449
SD	0.0218	0.0202	0.0326	0.0348

	LDA Classification	C5.0 Machine Learning	GBM Machine Learning	KNN Machine Learning
Fold 1	0.3363	0.3186	0.3761	0.3938
Fold 2	0.3094	0.3677	0.3857	0.3318
Fold 3	0.3603	0.3563	0.3522	0.3198
Fold 4	0.3465	0.3509	0.3553	0.3289
Fold 5	0.3109	0.3613	0.3403	0.3361
Mean	0.3327	0.3510	0.3619	0.3421
SD	0.0223	0.0191	0.0185	0.0295

	NN Machine Learning	RF Machine Learning	SVM Machine Learning	Stacking Machine Learning
Fold 1	0.5531	0.3407	0.3363	0.4027
Fold 2	0.5112	0.3274	0.3587	0.3946
Fold 3	0.5061	0.3522	0.3198	0.3441
Fold 4	0.4912	0.3377	0.3465	0.3509
Fold 5	0.4202	0.3067	0.3193	0.3487
Mean	0.4964	0.3329	0.3361	0.3682
SD	0.0483	0.0171	0.0171	0.0280

From the above table, we have accuracy rates of four linear regression models, five classification models, and seven machine learning models. For the linear regression models, stepwise regression yields the smallest standard deviation of 0.0250 and the PCR has the largest standard deviation of 0.0558. The lasso performs the lowest mean rate of

accuracy of 0.3174 and the PCR shows the highest mean rate of 0.3381. For the classification, again, stepwise regression has the smallest standard deviation of 0.0202, and the ordinal logistic regression with variables selected by lasso has the largest standard deviation of 0.0348. On the other hand, the lasso donates the lowest mean rate of 0.2942, just like it does for the linear regression. However, the ordinal logistic regression has the highest mean rate of 0.3449 among all classification models. For the machine learning approaches, the random forest and the SVM have the equivalently smallest standard deviation of 0.0171 while the NN models are in the largest standard deviation of 0.0483. Also, the random forest model contributes to the lowest mean rate of 0.3329, but the NN model provides a significantly high rate with a mean of 0.4963.

Among all sixteen models, the random forest and the SVM are both in the smallest standard deviation of 0.0171 while the PCR has the largest of 0.0558. The NN model shows a significantly high mean rate of 0.4964 while the lasso approach for classification is in the lowest mean rate of 0.2942.

Figure 8 shows the scale and variability of the accuracy rates. The boxplots of the accuracy rates of the linear regression models, the classification regression models, and the machine learning models are showed in orange, green, and blue respectively. The red dots indicate the mean of the rates for each approach. Although the linear regression methods show a distinctly low MSE, the rates of accuracy are pretty much at the same level as the classification approaches but with a wider range. The machine learning methods of the C5.0, the GBM, the NN, and the stacking ensemble contribute a higher mean rate compared to the linear regression and classification methods. Although the NN

model has the highest prediction rate of 0.4964, it means our "best" model is only capable of accurately predicting half of the outcomes. The stacking ensemble method has the second-highest mean rate of accuracy by combining the bottom layer models of the KNN and the GBM with the top layer model of multinomial logistic regression.

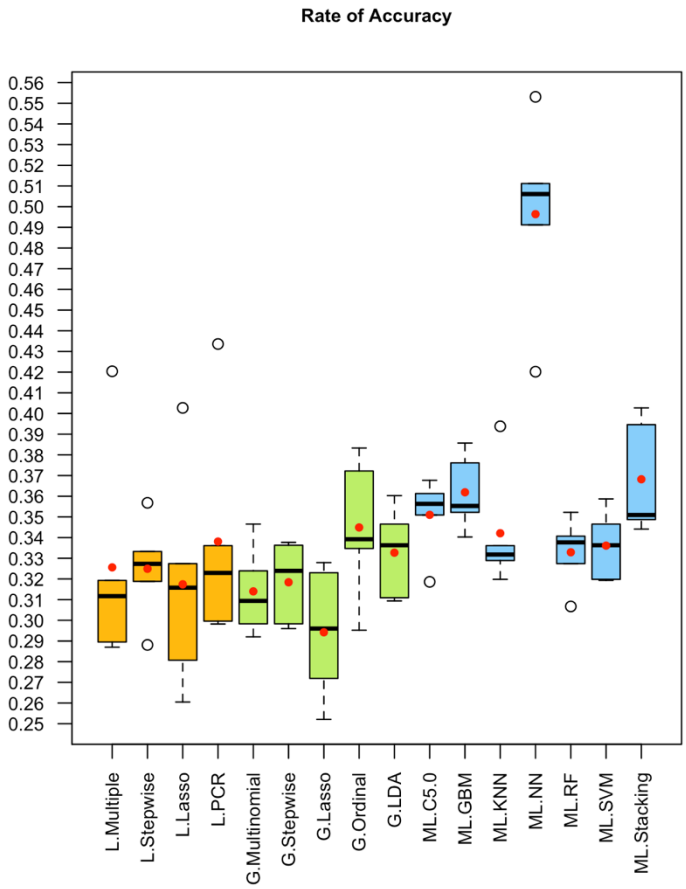

Figure 8: The Boxplot of the Rate of Accuracy. The Linear Regression Models, the Classification Regression Models and the Machine Learning Models are Showed in Orange, Green, and Blue

3. Discussion

We now have a comprehensive understanding of how our models perform on the private sector participation data. The predicted accuracy varies from 30% to 50%, and the estimated prediction error varies from 0.0730 to 0.1111. Given that the R-squared of the

multiple linear regression model is 0.1422 on a 0-1 scale, we do not expect high prediction accuracy and indeed, none of the models deliver good predictions. The relatively low R-squared value indicates that there is a lot of noise in the data.

Also, the high accuracy rate does not guarantee the lowest prediction error because the accuracy rate indicates how many predictions are matched the original response, and the prediction error tells how far away between the predicted value and the response. All linear regression models have low MSE but low accuracy. For example, although the PCR model has the smallest prediction error, its rate of accuracy is relatively low among the sixteen models. That is to say, the PCR has a large amount of the misclassifications with the estimated values that are very close to the true response, say, misclassified "1" by "0.75", or misclassified "0.25" by "0.5" or to "0", etc.

We are not surprised that the PCR has outperformed among the linear regression models. In the data set the 17 interactions are highly correlated to the main effects of the DEs, whereas the PCA is capable of reducing the dimensions for correlated variables and carrying out a set of components that are uncorrelated.

For the ordinal logistic regression, we used the predictors that are selected by the lasso approach. The smallest mean MSE and the highest rate of accuracy have been carried out from the ordinal logistic regression among all five classification regression models suggested that it is more justifiable to regard the response DRRI as an ordinal output.

Recall that the original data set contains three outcome variables of DRRI, BCI, and SCRI. The DRRI (Disaster Risk Reduction Index) corresponds to the value of measures

taken to control risks and reduce potential damage and losses as a result, the BCI (Business Continuity Index) values the measures taken to ensure business safety and continuity of time-sensitive operations, and the CSRI (Corporate Social Responsibility Index) is considered as the overlap of DRRI and BCI. We used the same models to predict the BCI as well but there is not enough time or space for us to discuss the detail in the thesis. Here we briefly illustrate the results of the model performance of the multiple linear regression, the multinomial logistic regression, and the NN approach where the BCI is the response for the models.

Table 66: Prediction Performance of the Response BCI.

	Multiple Linear Regression	**Multinomial Classification**	**NN Machine Learning**
MSE	0.0413	0.0504	0.0400
Rate of Accuracy	0.6857	0.7240	0.7648

These three models yield significantly lower mean MSE and higher mean rates of accuracy by using the same 36 variables to predict BCI, and there are more than a thousand observations in the data set that is sufficient enough to train the models. By comparison with the prediction performance of response BCI, we have the reason to believe that the variables of the input data may not qualified or efficient to explain and predict the response DRRI very well, but the performance is way much better when they are used to regress on the response BCI.

4.2 Comparison on Variable Selection

The second aim in this section is to select important variables from the 36 predictors of the data set. For the stepwise selection method, we choose the variables selected by the smallest AIC. For the lasso shrinkage method, the variables are selected by the corresponding non-zero coefficients with the tuning parameter 0.0707 for linear regression and 0.0372 for the classification. The entire data set is used for each model to perform and the checkmark symbol indicates whether the variable is selected or not. Although there is no model selection mechanism for the multiple linear regression approach and the multinomial logistic regression approach, we use the variables in which the p-value is smaller than 0.1 as the results for comparison.

The multiple linear regression, the multinomial logistic regression, stepwise selection for linear regression and classification, and the lasso shrinkage approach for linear regression and classification are abbreviated as MP, MN, SSL, SSC, LaL, and LaC respectively in the above table.

Table 67: Selected Variable of Six Models.

Model	**MP**	**MN**	**SSL**	**SSC**	**LaL**	**LaC**
LI		√		√		
SCD				√		
Dea				√		
LAS			√	√		
EC				√		
DC				√		
LTC	√		√	√		√
PG		√		√		
PO			√	√		
IA				√		

WO			√			
EI				√		
CH				√		
LKSP				√		
NP	√		√	√		
DF				√		
OT				√		√
City	√	√	√	√		√
BS	√	√	√	√	√	√
LI*BS		√		√		
SCD*BS				√	√	√
Dea*BS				√		
LAS*BS				√		√
EC*BS		√	√	√		
DC*BS		√		√		
LTC*BS					√	√
PG*BS		√		√		
PO*BS				√	√	√
IA*BS				√		
WO*BS			√	√		
EI*BS			√			
CH*BS		√				
LKSP*BS						
NP*BS	√	√	√			
DF*BS			√		√	√
OT*BS		√				

Among the six models, the stepwise selection for classification selects the most amount of the variables of 28 while the lasso for linear regression selects the less amount of the variables of 5 and the multiple linear regression has the less amount of the significant variables of 5 as well. Even though the number of variables selected by the stepwise selection for classification is many times more than the number of variables selected by the lasso for linear regression, there is no compelling difference in the model performance between these two models.

The multinomial logistic regression and the lasso shrinkage methods consider the interaction terms more important than the main effects in the model, however, the multiple linear regression and stepwise selection approaches value interaction term less important than the main effect. On one hand, the interaction term can be included or selected without its main effect of DE. On the other hand, another main effect of the interaction term, the variable of BS, is important for all six models. In Chapter III, we showed the high correlation between the interaction term and its main effect of DE. The appropriate variable selection methods were supposed to choose either the DE or its interaction term to avoid the high correlations between the predictors. However, the stepwise selection for classification has selected too many interaction terms and their main effect of DEs at the same time. This may cause the consequence for poor prediction performance.

The above table shows that the variable BS has the maximum six checkmarks, following by the variable City with five checkmarks. The LTC has four checkmarks, and the NP, the SCD*BS, the EC*BS, the PO*BS, the NP*BS and the DF*BS have three checkmarks. The variable of business size (BS) is chosen to be the most important predictors of DRRI since it has been selected 100% of the time using six different models. Our methods also identify other important predictors such as:

* City location (City)
* Loss of telecommunications (LTC)
* Negative publicity (NP)
* Interaction of negative publicity (NP) and business size

* Interaction of supply chain disruption (SCD) and business size
* Interaction of extreme conditions (EC) and business size
* Interaction of power outage (PO) and business size
* Interaction of damaged facilities (DF) and business size

The previous study (Sarmiento et al., 2019) has revealed that business size has an impact on the relationship between disaster experiences and the responses. From our result, it shows that five out of nine important variables are the interaction terms. As the considerations of the previous study, the interaction terms have necessarily contributed to the predictive ability of the regression models.

CHAPTER V. CONCLUSION

The thesis is to study a high dimensional data of the private sector participation in disaster risk reduction by using statistical models. According to the previous study (Sarmiento et al., 2012; 2019), the researchers utilized a linear regression approach to regress five out of seventeen disaster experiences (DE) and the business size as the input data set on the output of the disaster risk reduction index (DRRI). Other than that, the researches revealed that business size may have an impact on the relationship between DE and the response. Moreover, the descriptive statistical analysis showed that the DRRI can be considered not only as the numeric output but also as the categorical one. Therefore, we performed four linear regression models on the numeric response, five classification regression models, and seven machine learning approaches on the categorical response. The input data set included all seventeen DEs, business size (BS), city locations (City), and seventeen interactions between DE and business size.

The first aim of the thesis is to use different models to predict outcomes. For the linear regression, the multiple linear regression predicted the response with all 36 predictors. The stepwise selection and the lasso approach were using different ways to find the important variables among 36 predictors for the model. The PCR was another approach to combine the variables into principal components and then predict the response with the optimal numbers of the components. For the classification regression, the multinomial logistic regression used all 36 variables to predict the response. The stepwise selection and elastic net regularization approach were used to help choosing the important variables and used them to fit the model. The ordinal logistic regression considered the

response as the ordered outcome and used the variables that were selected by lasso. The LDA was another approach to separate the classes and predict on the response. All the machine learning methods that were utilized in the thesis are non-linear regression. We used seven machine learning methods to compare the prediction performance with the statistical approaches.

Our second aim is to access the prediction performance of the prediction models. In order to avoid over-fitting, ideally, we would like to train our model and test the effectiveness of the model with two separate data sets to estimate the true prediction error. The cross-validation (CV) technique is a useful tool to allow us to utilize our data better. The external 5-fold CV separated the whole data set into the modeling set and the test set. The modeling set was for training purpose and the test set was for prediction purpose. The internal CV was embedded whenever we needed to investigate the optimal tuning parameter for the model. The modeling was separated further into a training set and a validation set by the internal CV, and the final model would be decided if it came with the minimum MSE or the highest rate of accuracy of the validation set. Finally, the prediction performance was evaluated by the mean MSE and the mean rate of accuracy for each fold.

The last aim of the thesis is to understand which variables are more important for the prediction models. There are 36 variables in the input data set. Excluding or minimizing the effects of the variables which are less contributed to the model can yield better prediction accuracy and model interpretability. The variable selection approach, the shrinkage regularization methods were used to serve this aim.

Our results showed that the rate of accuracy varied from 29% to 50% among all the models. Regarding the prediction performance, the neural network approach contributed to the highest rate of accuracy. In contrast, the stepwise selection, the lasso shrinkage approach, and the elastic net regularization were not able to enhance the prediction performance as much as the machine learning approaches did. The ordinal logistic regression with the variables selected by the lasso shrinkage method outperformed other competitions of the statistical models. This result indicated that it would be more appropriate to consider the DRRI as the ordered and qualitative output. Also, the linear techniques for the dimension reduction approach, the PCA and the LDA, were provided a visualization of the data with the scatter plots. The plots are presented that there were neither clear boundaries for the classes nor separated clusters for the groups. Although we did not have enough space and time to show our work for the prediction of another response BCI, we did notice that when we used the same input data and the same statistical models to regress on the BCI instead of the DRRI, the prediction rate of accuracy was doubled and the MSE was halved. Therefore, we conclude that none of the proposed methods have yielded a satisfactory prediction performance for the outcome of the DRRI due to the high noise in the data, and the input data might not appropriate to predict the future output of the DRRI if we aim to achieve a desirable predicted accuracy.

Some of the results of the variable selection were expected to enhance the prediction performance and reduced model complexity. For instance, the ordinal logistic regression with nine variables selected by lasso provided higher prediction accuracy. However, the low consistency of the variable selection was shown under different criteria among the

models. Also, there was no significant difference in the prediction performance between the two models which were with distinctively different variables and different numbers of variables. Although there were variations of the selected variables, some interaction effects had been repeatedly chosen while their main effect was absent. The interaction term indeed had a different effect on the outcome of DRRI. Hence, it was reasonable to include the interaction term among some disaster experiences and business size in the predicted model. Another discovery of the variable selection was that in fact, both the business size and the city location had substantial effects on predicting the disaster risk reduction index, and there was no reason to exclude the variable of city location from the model beforehand.

In general, machine learning methods are usually more powerful on the prediction. From the perspective of prediction performance, some results of the machine learning algorithm were more desirable than the result of the traditional linear statistical approaches. In contrast, machine learning models were less interpretable than the statistical models. A more remarkable outcome was that we used the stacking ensemble technique to successfully enhance the prediction performance of three weak learners. This technique helped us to improve the model performance with no limitation on the type of learning algorithms to combine with.

Due to the restrictions of time and space, there were some limitations to this study. As a matter of fact, there were three indexes in the original data set, and the previous study explained that the relationship among these indexes was correlated. To study on only one index might not sufficiently reveal the relationship between the input data set and the

output data. Also, the thesis did not go further to explore how much the interaction term could affect the predicted output as we concluded that the interaction was needed for the prediction.

As future research, it should be first considered to use a multivariate regression model for the analysis. Additionally, the exploration of the variable selection is still necessary for better understanding and interpreting the prediction model. It is also important to analyze the interaction effects which could have more contributions to the model performance than their main effect could.

REFERENCES

[1] Allen, D. M. (1974). The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction. Technometrics. 16 (1): 125–127.

[2] Altman, N. S. (August 1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3):175–185.

[3] Andrews, D. F. (1974). A Robust Method for Multiple Linear Regression. Technometrics, 16:523–531.

[4] Asgary, A. (2016). Business Continuity and Disaster Risk Management in Business Education: Case of York University. AD-minister, 28:49-72.

[5] Asgary, A., Anjum, M. I., Azimi, N. (2012). Disaster Recovery and Business Continuity After the 2010 Flood in Pakistan: Case of Small Businesses. International Journal of Disaster Risk Reduction, 2:46–56.

[6] Bartik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., and Stanton, C. (2020). The Impact of Covid-19 on Small Business Outcomes and Expectations. Proceedings of the National Academy of Sciences, 117(30): 17656–17666.

[7] Bishop, C. (1995). Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

[8] Böhning, D. (1992). Multinomial Logistic Regression Algorithm. Ann. Inst. Statist. Math., 44:197–200.

[9] Borooah, V. K. (2001). Logit and Probit: Ordered and Multinomial Models. Thousand Oaks, CA: Sage.

[10] Breiman, L. (2001). Random Forests. Machine Learning, 45: 5–32.

[11] Cortes, C., Vapnik, V. (1995). Support Vector Networks. Machine Learning, 20:273-297.

[12] Cox, D. R. (1984). Interaction. International Statistical Review, 52: 1–31.

[13] Datta-Gupta, A., Mishra, S. (2017). Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences. Netherlands: Elsevier Science.

[14] Dawes, R. M. (2001). Clinical versus Actuarial Prediction. International Encyclopedia of the Social & Behavioral Sciences 2001: 2048-2051

[15] Echaniz, E., Ho, C.Q., Rodriguez, A., dell'Olio, L. (2019). Comparing Best-Worst and Ordered Logit Approaches for User Satisfaction in Transit Services. Transportation Research Part A: Policy and Practice, 130:752-769

[16] Efroymson, M. A. (1960). Multiple Regression Analysis. Mathematical Methods for Digital Computers. New York: Wiley:191–203.

[17] Engel, J. (1988). Polytomous Logistic Regression. Statistica Neerlandica, 42 (4):233–252.

[18] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7(2):179–188.

[19] Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5): 1189–1232.

[20] Friedman, J. H. (1989). Regularized Discriminant Analysis. Journal of the American Statistical Association, 84(405): 165-175.

[21] Friedman, J.H. (2002). Stochastic Gradient Boosting. Computational Statistics and Data Analysis, 38: 367–378.

[22] Garside, M. J. (1965). The Best Subset in Multiple Regression Analysis. Appl. Stat, 14:196–200.

[23] Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning; Data Mining, Inference and Prediction. New York: Springer.

[24] Ho, T. K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1:278–282.

[25] Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of National Academy of Science, USA, 79:2554-2558.

[26] Hopfield, J. J. (1984). Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons. Proceedings of National Academy of Science, USA, 81:3088-3092.

[27] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. New York: Springer.

[28] Kendall, M.G. (1957). A Course in Multivariate Analysis. London: Griffin.

[29] Kuhn, M., and Johnson, K. (2013). Applied Predictive Modeling. Springer, New York.

[30] Lantz, B. (2013). Machine Learning with R. Packt Publishing Ltd.: ISBN-978-1782162148.

[31] Massy, W.F. (1965). Principal Components Regression in Exploratory Statistical Research. J. Amer. Statist. Assoc., 60: 234–256.

[32] McCullagh, P. (1980). Regression Models for Ordinal Data. Journal of the Royal Statistical Society, B-42(2):109–142.

[33] McCullagh, P., and Nelder, J. A. (1989). Generalized Linear Models (2nd ed.). London, UK: Chapman and Hall.

[34] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers. San Mateo.

[35] RStudio Team (2016). RStudio: Integrated development for R. RStudio, Inc., Boston, MA.

[36] RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA.

[37] Santosa, F., Symes, W. W. (1986). Linear Inversion of Band-Limited Reflection Seismograms. SIAM Journal on Scientific and Statistical Computing, 7 (4): 1307–1330.

[38] Sarmiento JP, Hoberman G, Ilcheva M, Asgary A, Majano AM, Poggione S, Duran LR (2015): Private sector and disaster risk reduction: The Cases of Bogota, Miami, Kingston, San Jose, Santiago, and Vancouver. INT J DISAST RISK RE 14:225–237.

[39] Sarmiento, J. P., Hoberman, G., Ilcheva, M., Asgary, A., Majano, A. M., Poggione, S., Duran, L. R. (2012). Private Sector and Disaster Risk Reduction: The Cases of Bogota, Miami, Kingston, San Jose, Santiago, and Vancouver. Background Paper Prepared for the 2013 Global Assessment Report on Disaster Risk Reduction, UNISDR, Geneva, Switzerland.

[40] Sarmiento, J. P., Sarmiento, C., Hoberman, G., Chabba, M., Sandoval, V. (2019). Small and Medium Enterprises in the Americas, Effect of Disaster Experience on Readiness Capabilities. AD-minister, 35:117-136.

[41] Spurrell, D.J. (1963). Some Metallurgical Applications of Principal Components. Appl. Statist, 12: 180–188.

[42] Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36 (2): 111–147.

[43] Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. Journal of the Royal Statistical Society: Series B (Methodological), 39 (1): 44–47.

[44] Tamhane, A. C., and Dunlop, D. D. (2000). Statistics and Data Analysis. Prentice-Hall.

[45] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (methodological), Wiley, 58 (1): 267–88.

[46] United Nations (2016). Sustainable Development: Disaster Risk Reduction. Geneva: UN.

[47] Winship, C., Mare, R. D. (1984). Regression Models with Ordinal Variables. American Sociological Review, 49(4):512–525.

[48] Wolpert, D. H. (1992). Stacked Generalization. Neural Networks, 5(2):241–259.

[49] Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B, 67(2): 301–320.