Yifei Li 251134118 & Luchang Liu 251107314

1. Introduction

In 2007, the occurrence of the financial crisis plunged the U.S. economy into a prolonged downturn. This period of global stagnation starting in 2007 was called the "Great Recession". Small businesses have been an essential source in the U.S. credit market (Li et al., 2018). Nevertheless, they have been greatly impacted by the market collapse during the "Great Recession". Small businesses are provided with loans from banks by the U.S. government, and several small businesses benefit and succeed from this subsidy. However, there were still some cases where some small businesses defaulted the loans granted

In 1953, U.S. Small Business Administration (SBA) was established to promote and help small businesses in the U.S. credit markets. Advisory and financial assistance was provided by SBA to small businesses, including direct government loan services, with the functions of protecting Small or Medium Enterprises (SMEs), maintaining competitive free enterprise, strengthening the nation's overall economy, and assisting in the recovery of the economic

In this project, we investigate a variety of variables to find how they influence the state of loans of small businesses during the Pre-Financial Crisis Period. We aim to determine what small businesses should be granted loans in California at the beginning of the financial crisis. The decision making of whether to grant loans leads to a binary classification problem. Back to the present, human beings experience a global economic stagnation due to the impact of the COVID-19 pandemic. Under this environment, we could use the model we create in our paper to solve the same question in the future.

2. Data Management

In this chapter, we introduce the dataset and present details of data management. The dataset was firstly published by SBA in 2014 and can be downloaded from Kaggle at https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied. The original data contains 27 variables (see Appendix A) with 899,164 data points. In what follows, we describe each step of data manipulation.

Step (1). We subset the year 2005 as our start point since in Figure 2.1, the year 2015 is the start point of economic recession reflected by a sharp drop in the Gross Domestic Product (GDP) growth rate. In addition, the original sample size is too large to analyze, so in this project, we only focus on data from the largest state in the US, California (coded as *CA*). As a result, the current data set includes 9,821 cases from *CA* in 2005.

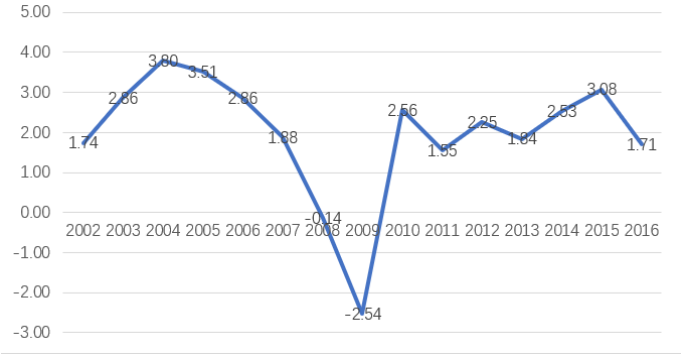


Figure 2.1: Line graph of U.S. GDP growth (annual %). Data from the World Bank (2020).

Step (2). We perform data cleaning and subset the numbers of loan terms which are not larger than 108 months (*Term* <= 108) because without this operation, it will result in a lower default rate for smaller businesses with longer terms. Some small businesses may default after 108 months but still show 'PIF' as the latest year the dataset records is 2014. We then exclude eight data with missing values in *MIS_Status*, one error entry data of *LowDoc*, and nine empty data points in *UrbanRural*. We eliminate the variable *RevLineCr* as it contains over 30% of error entries. We delete variables with the same values (*FranchiseCode* and *LowDoc*) and finally have 11 variables and 7,029 data points left in the dataset.

Step (3). We then reconstruct the variables. Discount Retailer, Health Care, Food & Restaurants, Freights & Logistics, and DIY & Repairs were these industries that best survived during the global financial crisis in the United States (The Investopedia Team, 2020). We believe that small businesses within these industries are less likely to default so we replace *NAICS* with a new variable *Industry* based on the *NAICS* number and the industry sector graph (see Appendix B). We labelled small businesses in "Agriculture, forestry, fishing and hunting" (sector 11), "Professional, scientific, and technical services" (sector 54), and "Health care and social assistance" (sector 62) as category "1" and all other industries are labelled as "0". We modify *MIS_Status* by replacing "CHGOFF" with category '1' since we focused more on the default status, and we replaced "PIF" with '0' to reveal that the business has paid its loan in full.

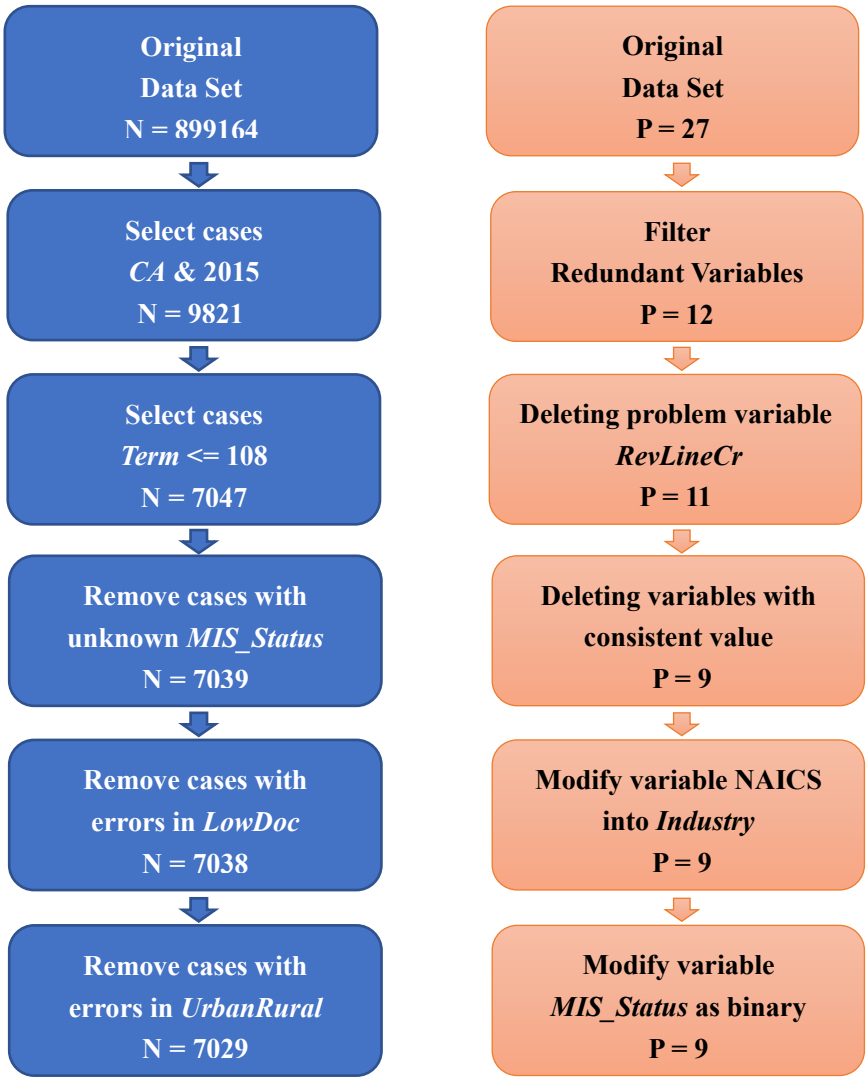


Figure 2.2: Flow chart of data management

Figure 2.2 summarizes the above operations and finally, there are 7029 data with nine variables (*Term, NoEmp, NewExist, CreateJob, RetainedJob, UrbanRural, DisbursementGross, MIS_Status,* and *Industry*) for us to explore in the dataset. We randomly separate the dataset (*N = 7029*) into a training set with a total number of 4921 and a testing set (*N = 2108*) with a ratio of 7:3.

3. Method

In this chapter, we present the methodology of our project including some exploratory data analysis (EDA) with data visualization technologies, and the model applied to predict whether a loan defaults or not.

3.1 EDA

We first compare the main characteristics (median and interquartile range (IQR) for numeric variables; size and proportion for categorical variables) of the training set and testing set to ensure that there is no significant difference between these two datasets. A Pearson's R correlation graph for all variables is then conducted to detect any potential collinearity among them. Histogram plots, scatter plots, and bar charts are produced to identify their characteristics. Due to the limitations of Pearson's R in detecting correlation between categorical variables, we then introduce Cramer's V (Cramer et al., 1946) as a complement which can identify levels of association among categorical variables. The Cramer's V falls within [0.1, 0.3], [0.4, 0.5], (5, ∞) will be considered weak, medium, and strong association respectively, while no association if the value is less than 0.1 (Wu et al., 2014).

3.2 Main Data Analyses

In this section, logistic regression (faraway, 2016), penalized logistic regression (Santosa & Symes, 1986), and Random Forests (Ho, 1995) will be introduced. Additionally, we will present the metrics that will be applied to evaluate the performance.

3.2.1 Logistic Regression

Since we aim to predict the status of default, which is a binary response according to our definition, it is a good idea to create a logistic regression to fit the data. Research conducted by Luo and Lei (2008) showed that the logistic model can predict corporation credit default probability with around 76% of accuracy, showing great predicted performances. The logistic regression belongs to the generalized linear model family where a logit link (2) is used. The model can be expressed as follows.

$$\eta_i = \beta_1 + x_{i1} + \cdots + \beta_q x_{iq} \qquad i = 1, 2, \ldots, n \tag{1}$$

$$\eta = \log\left(\frac{p}{1-p}\right) \tag{2}$$

$$Odds = \frac{p}{1-p} \tag{3}$$

where n is the total number of observations with covariates x_1, x_2, …, x_q, and p is the probability of 'success'. In our project, 'success' is defined as the default of SBA loans. For

interpretation purpose, one can use log odds of (success) loan defaulting, which make the logistic regression stand out from most the machine learning method. To estimate the coefficient in the model, the negative log-likelihood ($L_{log} = -\ln(L) = -\sum_{i=1}^{n}[y_i\eta_i - \log(1+e_i^{\eta})]$) will be minimized.

3.2.1.1 Modelling

We obtain a logistic regression model using all available predictors in the training set. We use the glm() function in R with binomial family to fit the logistic model, and the stepwise method is applied to do variable selection using the Akaike information criterion (AIC). Q-Q plot and half-normal plot are used to check normality and outliers.

3.2.1.2 Goodness of Fit

There are two general approaches to identifying how well the model fits the data. One is to test whether the model needs to be more complex, specifically, whether it needs additional nonlinearities and interactions to satisfactorily represent the data. The other is to get a measure of how well we can predict the dependent variable based on the independent variables (Allison, 2013). The first approach we mentioned was used to test the goodness of fit of the logistic model. The Hosmer–Lemeshow (HL) test is used and test statistic is shown as equation (4).

$$\chi^2_{HL} = \sum_{j=1}^{J} \frac{\left(y_j - m_j p_j\right)^2}{m_j p_j (1 - p_j)} \tag{4}$$

where *j* is the number of the bins of the observations divided based on the linear predictor, y_j is the mean response in the j^{th} bin, p_j is the mean predicted probability of m_j observations on the bin. The null hypothesis for HL test is that there is no lack of fit in the model.

Pseudo-R-Squared is another measure used in logistic regression of how well we can predict the dependent variable based on the independent variables. Faraway (2016) proposed that Nagelkerke's method was a good statistic for R^2, and other methods also widely used were McFadden's R^2 and the Cox-Snell's R^2. For Cox-Snell's R^2, the statistic $R^2_{C\&S}$ is

$$R^2_{C\&S} = 1 - (L_0/L_M)^{2/n} \tag{5}$$

where L_0 is the value of the likelihood function for a model with no predictors, L_M is the likelihood for the model being estimated. However, a big problem with Cox-Snell's R^2 is that it has an upper bound that is less than 1.0. Nagelkerke modified Cox-Snell's R^2 by dividing it by its upper bound and got a new statistic R^2_N.

$$R^2_N = \frac{1 - (L_0/L_M)^{2/n}}{1 - {L_0}^{2/n}} \tag{6}$$

Allison (2022b) commented this correction was purely ad hoc, and it greatly reduced its capability of extending to other kinds of regression estimated by maximum likelihood, e.g., negative binomial regression. The author suggested McFadden's R^2 as the best choice with statistic R^2_{McF}.

$$R^2_{McF} = 1 - \ln(L_0)/\ln(L_M) \tag{7}$$

Because of McFadden's R^2's intuitive appeal, e.g., its upper bound is 1.0, closely relating to R^2 definitions for linear models and satisfying almost all Kvalseth's (1985) eight criteria for a good R^2. Based on these advantages, we will use McFadden's R^2 as our measure. Domencich et al. (1975) stated that McFadden's R^2 from 0.2 to 0.4 indicates excellent fits.

3.2.2 Penalized Logistic Regression

Instead of what we did in section 3.2.1, maximizing the log-likelihood, here we explored Ridge (8) and LASSO (9) regularization work by adding a penalty term to the log-likelihood function.

$$L_{log} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{8}$$

$$L_{log} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{9}$$

where *j* is the number of variables in the model and λ is a free parameter selected to minimize the out-of-sample error. The difference between them is that the ridge regression uses the squared value of coefficients and LASSO uses the absolute value of coefficients as their own penalty term. However, LASSO allows variables to be selected while ridge does not. In this project, we use cross-validation to find the best λ and applied it to create models.

3.2.3 Random Forests

Random Forests was first proposed by Tin Kam Ho (1995) and further developed by Leo Breiman (2001). This method overcomes the fundamental limitation in the complexity of decision tree classifiers but with good simplicity. In research by Zhu et al. (2019), the accuracy of Random Forests algorithm is 98% in predicting default samples, indicating better performances than logistic regression (73%), decision tree (95%) and other machine learning algorithms. We will create a new model using Random Forests with the '*randomForest*' package in R (Liaw & Wiener, 2002).

When modelling using the '*randomForest*' package, '*mtry*' and '*ntree*' are the decisive inputs that significantly influence the performance and economization of the forest (Faraway, 2016). We first consider '*ntree*', the number of the B trees used. A larger number of B trees usually brings better performance.

3.2.4 Evaluation Metrics

To measure and compare those models we discussed previously, Sensitivity (Sens), Accuracy (Acc), and Area under the ROC Curve (AUC) will be the main metrics to evaluate the performance as we focus more on how well this model can detect small businesses that will not pay the loan in full. As for supplements, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and balanced accuracy (Bal-Acc) will be calculated in each model. We include the Matthews Correlation Coefficient (MCC) which is thought to be better than to Acc and F1 score in evaluating binary classification tasks (Chicco & Jurman, 2020).

$$sensitivity = \frac{TP}{TP + FN} \qquad specificity = \frac{TN}{TN + FP} \qquad accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$positive\ predictive\ value = \frac{TP}{TP + FP} \qquad negative\ predictive\ value = \frac{TN}{TN + FN}$$

$$F1\ score = \frac{2TP}{2TP + FP + FN} \qquad balanced\ accuracy = \frac{sensitivity + specificity}{2}$$

Definitions of evaluation metrics are provided along with a sample of confusion matrix (see Appendix D) including values of True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN).

4. Results

We use the stepwise method to exclude two variables with high collinearity relationships found during EDA. We then construct optimized models in LASSO, Ridge and Random Forests. There is no significant difference among LASSO, Ridge and Logistic models in metrics except for sensitivity and specificity. Random forests perform better than the other models.

4.1 EDA Results

All the characteristics shown in the training and testing sets are approximately the same (see Appendix C). We noticed the medians of the variable *Term* are different between the training set and the testing set, with values of 75 and 85 respectively.

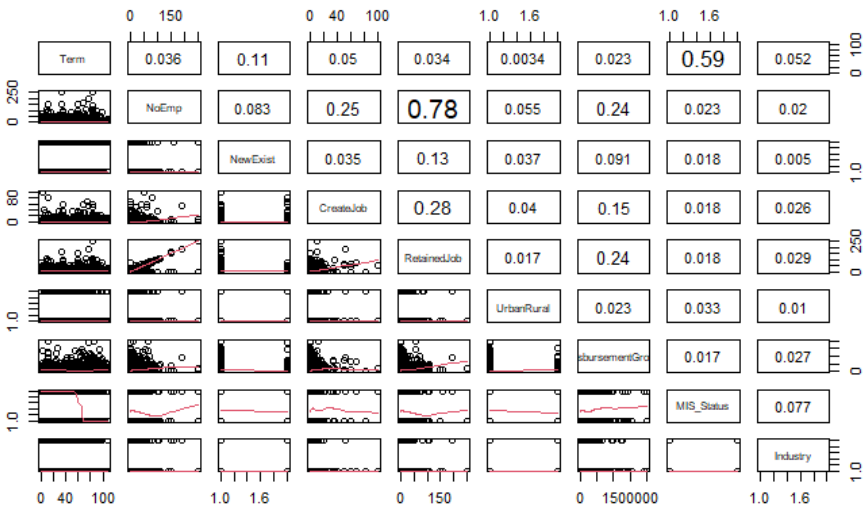


Figure 4.1 Pearson's R correlation matrix

We notice from Figure 4.1 that the retained job (*RetainedJob*) is the most correlated with the number of employees in a small business (*NoEmp*) with a 0.78 correlation. Additionally, the created job (*CreateJob*) and the disbursement gross are likewise related to the number of employees at 0.25 and 0.24 correlation ships, respectively. Thus, we generate figure 4.2 to explore the correlation between the three variables above and the employees' numbers.

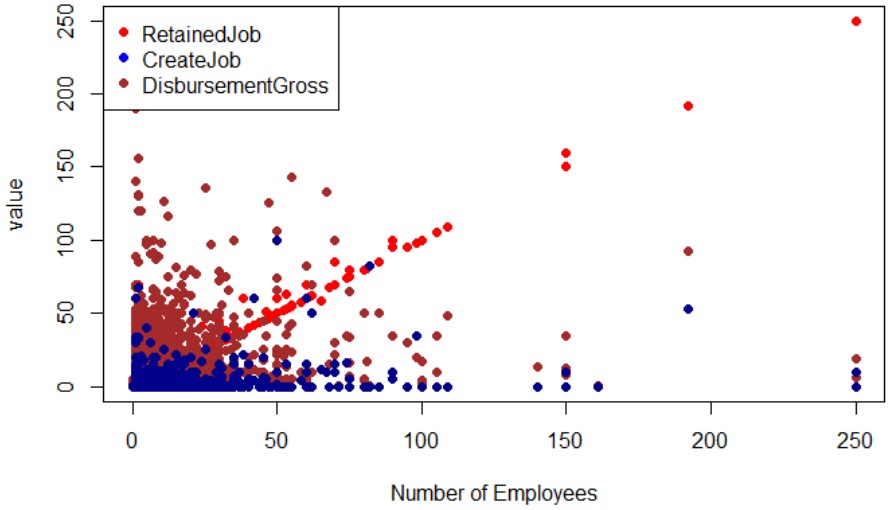


Figure 4.2 Scatter plot of *RetainedJob*, *CreateJob*, *DisbursementGross* (values downscaled by 10,000), and *NoEmp*

As Figure 4.2 shown, we find a positive linear relationship between the variables *NoEmp* and *RetainedJob* as the number of the retained jobs increases following the increase of the number of the employees in the business. Furthermore, the distribution of scatter points of created jobs is similar to the distribution of disbursement; the points in the plot are more concentrated in a range of employees' counts between 0 and 50 when the numbers of retained jobs are from 0 to 50.

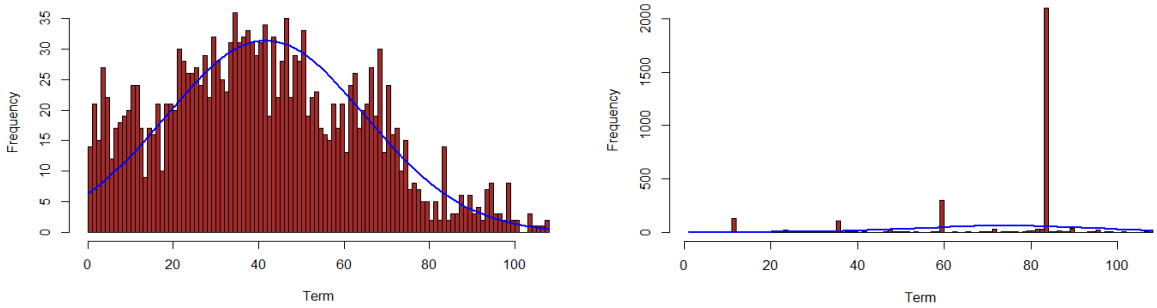


Figure 4.3 Frequency histograms of *Term* when *MIS_Status* = 1 (left) and *MIS_Status* = 0 (right)

In Figure 4.3, the graph on the left presents that the frequency of the companies over the number of terms approximately follows a normal distribution. Nevertheless, we can observe from the graph on the right that there are over 2000 small businesses in which the terms of loans are 83 months indicating the most frequent number.

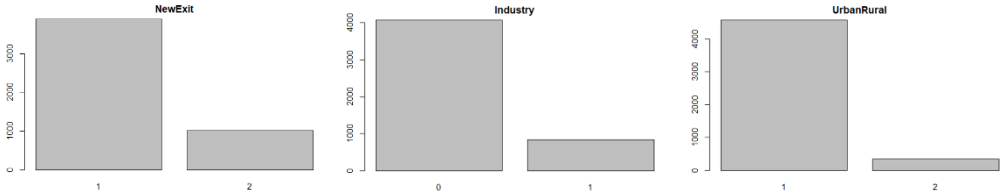


Figure 4.4 Bar charts of *NewExit* (left), *Industry* (middle), and *UrbanRural* (right).

From Figure 4.4, we observed that there is similar distribution in *Industry*, *NewExit*, and *UrbanRural*. To test the collinearity, Table 4.1 is then constructed to show the value of Cramer's V among these variables in pairs.

Table 4.1 Cramer's V correlation matrix

	Industry	**UrbanRural**	**NewExit**
Industry	-	-	-
UrbanRural	0.0103	-	-
NewExit	0.0050	0.0374	-
MIS_Status	0.0770	0.0374	0.0185

It is clear from Table 4.1 that all of these values are smaller than 0.1 so we can conclude that there is no significant paired correlation among categorical variables.

4.2 Modelling Results

In this section we present modelling results via Logistic, LASSO, Ridge and Random Forests. We decide to exclude two variables detected by AIC as it does not impair the model's performance. We find that regular logistic regression has larger deviance in value compared to LASSO and Ridge. A comparison of model performance with evaluation metrics introduced in section 3.2.4 is finally presented

4.2.1 Logistic Regression and Penalized Regression

From Figure 4.5 we notice from the Q-Q normal plot that the scatters do not go along the dash line in the graph. The pattern on the plot suggests a bimodal distribution which is the same as our expectations. In the half-normal plot, we can detect some outliers that are very far from the majority of points. However, these points are not errored entries after we checked them one by one.

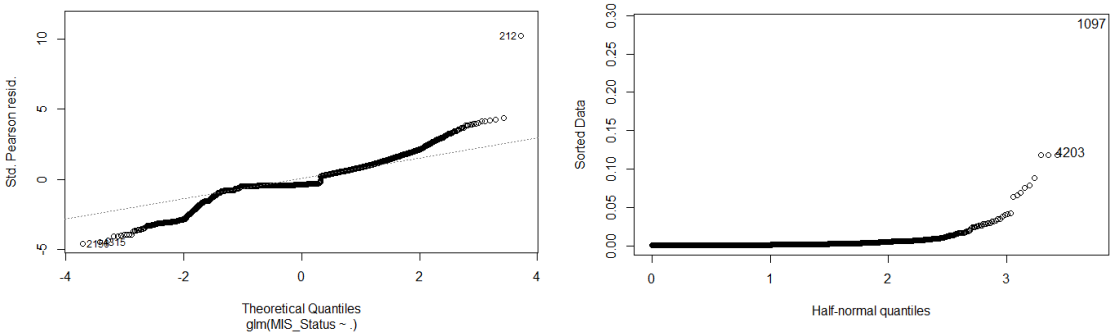


Figure 4.5 The Q-Q Normal Plot (left) and the Half-Normal Plot

We exclude variables *CreateJob* and *RetainedJob* after applying the stepwise method and we notice from the table (see Appendix E) that the coefficients of the predictors change slightly. Table 4.2 summarize the metrics on model performance. There is no significant change for the metrics after deleting these variables. Considered the similar performances in full model and the nested model, we decide to use the nested model as it is simpler.

Table 4.2 Evaluation metrics on logistic model

	AUC	Acc	Sens	Spec	PPV	NPV	MCC	F1score
Full	0.8560	0.8146	0.6771	0.9439	0.9190	0.7566	0.6477	0.7797
Reduced	0.8558	0.8151	0.6781	0.9431	0.9177	0.7581	0.6479	0.7799

Interpretation. The odds ratio of *Term* is exp(-5.585e-02) = 0.9457 (95% Confidence Interval = [-5.8968e-02, -5.2800e-02]), showing that with other factors fixed, one unit increase in the term of the loans will decrease the **odds** of defaulting the loans by approximately 5.43%. The odds ratio of the number of employees in the small firm is exp(-1.240e-02) = 0.9877 (95% Confidence Interval = [-1.9388e-02, -5.8327e-02]) indicating that with other factors fixed, one unit increase in employees in the company will decrease the odds of defaulting the loans by roughly 1.23%. The odds ratio of the new small firm is exp(3.356e-01) = 1.3988 (95% Confidence Interval = [1.572793e-02, 5.1346-02]) with other factors fixed, which means if other factors are close to 0 and fixed, compared to the business which is already existing, the new company is approximately 39.88% higher in the odds of defaulting the loans, with other factors are fixed. The odds ratio of the rural business is exp(-4.303e-01) = 0.6503 (95% Confidence Interval = [-7.3339e-02, -1.3348e-02]) with other factors fixed. If other factors are close to 0 and fixed, compared to urban companies, the rural small business has approximately 34.98% lower in the odds of defaulting the loans. With other factors fixed, the odds ratio of the loan disbursement is exp(4.169e-07 * 10,000) = 1.0041 (95% Confidence Interval = [-1.3633e-02, 9.6285e-02]) and an increase in 10,000 dollar in disbursement will increase the odds of defaulting on the loans by about 0.41%. In addition, the odds ratio of the small business in thriving industries is exp(-4.077e-01) = 0.6652 (95% Confidence Interval = [-6.1115e-02, -2.0719e-02]) with other factors fixed. We can say the small businesses within industries sector 11, 54, and 62 are approximately 33.48% lower in the odds of defaulting the loans compared with the other industries, with other factors fixed.

Goodness of fit. We then use diagnostic plots (Figure 4.6) of the binned residuals to investigate some potential problems in the goodness of fit. Ideally, the line should cross 95% or more of these vertical lines, however, the line crosses a small proportion of them, which may be a bad sign. However, diagnostic plots can not determine how well the model fits (Faraway, 2016).

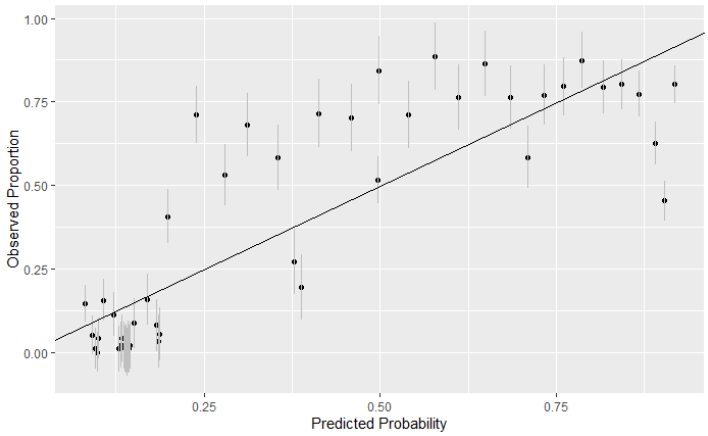


Figure 4.6 Diagnostic plots between observed and predicted proportions

HL test was applied and found that the p-value is less than 2.2e-16 which is small, so we reject the null hypothesis. The result suggests that this model is a poor fit. By using McFadden's method, we got the R2 of 0.2893 (for the full model) and 0.2891 (for the reduced model) indicating excellent model fit as falling within the range from 0.2 to 0.4. However, this contradictory situation does not surprise us because of the limitation of HL test in solving large data sets.

We compared the logistic regression with LASSO and ridge and find that LASSO exclude all the predictors excepted for *Term*. The coefficients of Ridge also different with logistic model even though it include all variables. The coefficients of *Term, CreateJob, RetainedJob,* and *UrbanRual* increase significantly compared to the full model. However, the values of deviance do not differ between LASSO (4721.683) and Ridge (4725.906). Regular logistic regression has a smaller deviance (4620.181), suggesting a better fit.

4.2.2 Random Forests

From the graph (see Appendix F), there is a rapid decrease in this mean squared error (MSE) when progressing from a small number of B trees. MSE becomes stabler from 1000 B trees but still at a certain level of fluctuation and finally be stabilized after 4000 B trees. Considering the size of the dataset and our computing power, we finally decide on 5000 as our setting for '*ntree*'.

We then decide the input for *mtry*, the size of the subsample of predictors selected at each node. Out-of-bag (OOB) error rates for each *mtry* are computed to determine the size we should choose. A small value in OOB error rate indicates a better performance. It is obvious that choosing 4 subsamples can have the least OOB error rate and we finally set *mtry* = 4 and *ntree* = 5000 to do the modelling.

Table 4.3 Corresponding values in *oob.value* (OOB error rate) and *mtry*

mtry	1	2	3	4	5	6	7	8	9	10
oob.value	0.279	0.100	0.087	0.084	0.085	0.085	0.086	0.088	0.087	0.087

Figure 4.7 expresses how much accuracy the model losses by excluding each variable. From the graph below we can find that predictor *Term* has the highest Mean Decrease GINI, followed by *DisbursementGross*. indicating the highest and second-highest importance in the model respectively.

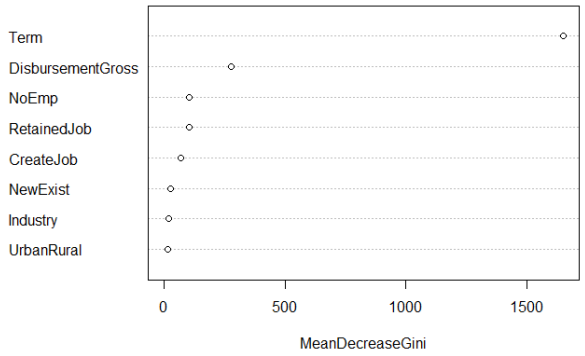


Figure 4.7 The Mean Decrease Accuracy plot

4.2.3 Models' Performance Comparison

From the table (see Appendix G), the Random Forests method had the highest scores among these four models in any metrics, showing the best performance. Values in metrics are similar among Logistic, LASSO, and Ridge. However, Logistic regression has the higher values in PPV (0.9190), MCC (0.6477), and Specificity (0.9439) compared with LASSO and Ridge, while LASSO and Ridge performed better in Sensitivity (7% higher than Logistic) and NPV (10% higher). There is no noticeable difference between LASSO and Ridge, suggesting an equivalent in performance.

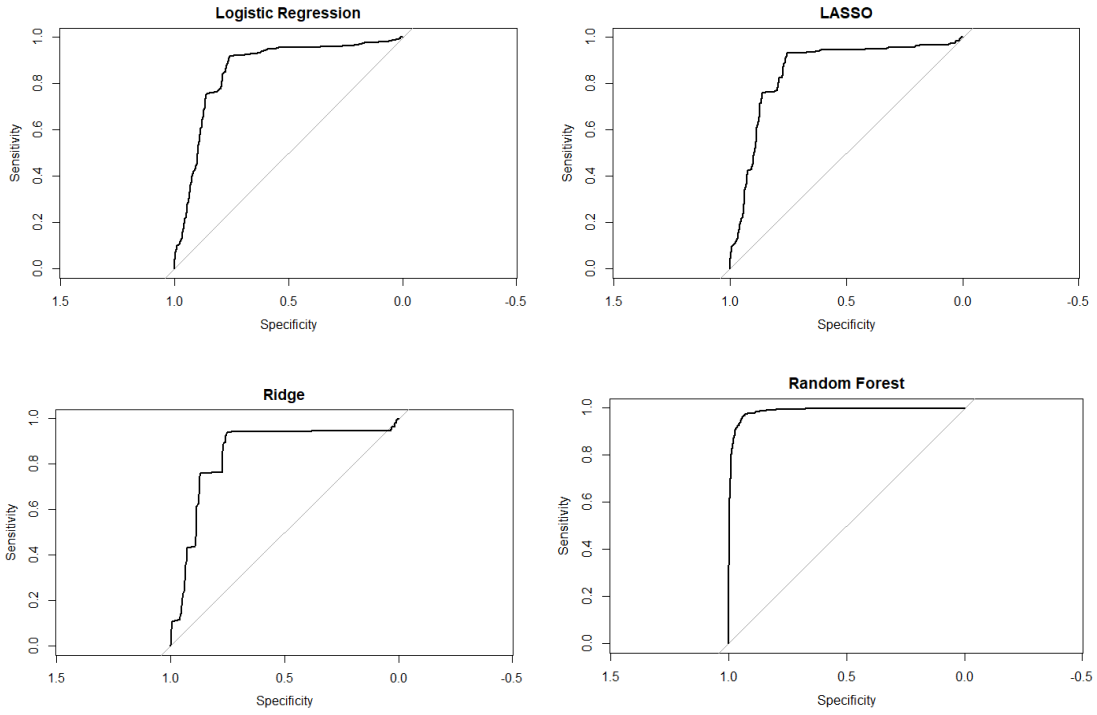


Figure 4.8 ROC graphs of Logistic, LASSO, Ridge, and Random Forests models

From Figure 4.8, we can see that the ROC curves of the Logistic Regression and LASSO show the same distribution; the ROC curve of Ridge is similar to the ones of the previous models, but there are more ladders showing on the left side of the curve for the Ridge model. However, the ROC curve of the Random Forests is the closest to the top left corner of the plot, which means both the true positive rate and the true negative rate are the closest to 1.

Overall, the Random Forests will be top priority to be considered as our predictive model because of its highest scores in metrics among the four models. For the remaining models, we may prefer LASSO or Ridge as it had higher accuracy and sensitivity which we are more interested in.

5. Conclusion, Discussion and Future work

In the result section, we use the HL test and conclude that our logistic model is a poor fit. However, we also notice that the model's accuracy is over 80%, indicating that this model can successfully predict 80% more of companies' status in the test set which seems to contradict the conclusion we previously get. This is due to the limited capability of the HL test in solving

large data sets. Allison (2022a) found that small differences from the proposed model are considered significant when using HL test in a very large data set. Even though they proposed a standardizing method of the HL test to improve, it still works poorly when the sample size is larger than 25,000 (Yu et al., 2017). Additionally, Allison (2022a) believed that the most troubling problem of the HL test is that the results are highly influenced by the number of groups set, and there is no theory to guide the choice of that number. The author showcased an example that a change in one unit of group set can result in a fluctuation of the p-value by 0.5. Due to the limitations of the HL test, we use Pseudo-R-Squared as our alternative to test whether this logistic model is a good fit. The result from Pseudo-R-Squared showed that our model fits well, which is consistent with its performance in prediction.

We also find that using LASSO and Ridge do not significantly improve the accuracy, which is not out of our surprise. We believe that one of the reasons is that LASSO performs better when solving sparse coefficients. Regulization is designed to solve 'overfitting' for complicated models. In our dataset, the number of variables is only nine which is not large enough to show the advantages of LASSO and Ridge. We find that Random Forests performs the best among four models and followed by LASSO and Ridge because they have better values in sensitivity. It is out of our expectation that Random Forest significantly outperforms the other three models, 10% on average higher in all of the evaluation metrics. This project shows us the power of Machine Learning methods. However, logistic regression has a good ability to explain and predict real-world business.

In the future, an elastic net can be used, which is a combination of LASSO and Ridge. We are also interested in applying our findings to similar backgrounds, for example, other financial crises and the economic recession under COVID-19 pandemic.

Reference

Allison, P. (2022a, February 17). *What's the Best R-Squared for Logistic Regression?* Statistical Horizons. https://statisticalhorizons.com/r2logistic/

Allison, P. (2022b, February 17). *Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression*. Statistical Horizons. https://statisticalhorizons.com/hosmer-lemeshow/

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7

Chi-Square and Cramer's V: What do You Expect? (2015). *Statistics for Political Analysis: Understanding the Numbers*, 245–272. https://doi.org/10.4135/9781483395418.n9

Cramer, H., Cramér, H., & Karreman Mathematics Research Collection. (1946). *Mathematical Methods of Statistics*. Amsterdam University Press.

Cramér's V Coefficient. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, And*. https://doi.org/10.4135/9781506326139.n162

Domencich, T. A., McFadden, D., & Charles River Associates. (1975). *Urban Travel Demand*. Van Haren Publishing.

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)* (2nd ed.). Chapman and Hall/CRC.

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*, *2*(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

A gentle introduction to logistic regression and lasso regularisation using R. (2017, October 6). Eight to Late. https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

Li, M., Mickel, A., & Taylor, S. (2018). "Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines. *Journal of Statistics Education*, *26*(1), 55–66. https://doi.org/10.1080/10691898.2018.1434342

Luo, J. H., & Lei, H. Y. (2008). Empirical Study of Corporation Credit Default Probability Based on Logit Model. *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*. https://doi.org/10.1109/wicom.2008.2276

Paul, P., Pennell, M. L., & Lemeshow, S. (2012). Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Statistics in Medicine*, *32*(1), 67–80. https://doi.org/10.1002/sim.5525

Santosa, F., & Symes, W. W. (1986). Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*, *7*(4), 1307–1330. https://doi.org/10.1137/0907087

Wu, B., Zhang, L., & Zhao, Y. (2014). Feature Selection via Cramer's V-Test Discretization for Remote-Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(5), 2593–2606. https://doi.org/10.1109/tgrs.2013.2263510

Yu, W., Xu, W., & Zhu, L. (2017). A modified Hosmer–Lemeshow test for large data sets. *Communications in Statistics - Theory and Methods*, *46*(23), 11813–11825. https://doi.org/10.1080/03610926.2017.1285922

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503–513. https://doi.org/10.1016/j.procs.2019.12.017

Appendix A

Name	**Role**	**Type**	**Unit**	**Description**
LoanNr_ChkDgt	Predictor	Categorical	N/A	The identifier: the Primary Key
Name	Predictor	Categorical	N/A	The name of the small business
City	Predictor	Categorical	N/A	The city of the small business
State	Predictor	Categorical	N/A	The state of the small business
Zip	Predictor	Categorical	N/A	The zip code of the small business
Bank	Predictor	Categorical	N/A	Bank name
BankState	Predictor	Categorical	N/A	Bank State

NAICS	Predictor	Categorical	N/A	North American Industry Classification System Code (See Appendix)
ApprovalDate	Predictor	Date	N/A	The date that SBA released the promise
ApprovalFY	Predictor	Categorical	N/A	The fiscal year of the promise
Term	Predictor	Numerical	Month	The period of the loan, unit in months.
NoEmp	Predictor	Numerical	Count	The number of employees in the business.
NewExist	Predictor	Categorical	N/A	The status of business: **1** if the business is already existing; **2** if it is a new one.
CreateJob	Predictor	Numerical	Count	The number of jobs that are created in the business.
RetainedJob	Predictor	Numerical	Count	The number of jobs that are retained in the business.
FranchiseCode	Predictor	Categorical	N/A	If the company does not have a franchise, then the code should be **00000** or **00001**.
UrbanRural	Predictor	Categorical	N/A	The status of location: **1** if the small business is urban; **2** if it is a rural business.
RevLineCr	Predictor	Categorical	N/A	The revolving line of credit of the small business: **Y (= Yes)** or **N (= No)**.

LowDoc	Predictor	Categorical	N/A	The program of Low Documentation Loan: **Y (= Yes)** or **N(= No)**.
ChgOffDate	Predictor	Date	N/A	The date when the loan of the small business is announced to be in default.
DisbursementDate	Predictor	Date	N/A	The date that the loan is paid
DisbursementGross	Predictor	Numerical	USD	The amount of loan that was paid.
BalanceGross	Predictor	Numerical	USD	The gross outstanding amount
ChgOffPrinGr	Predictor	Numerical	USD	The amount of the loan which is charged off
GrAppv	Predictor	Numerical	USD	The total amount of the loan that is approved by the bank
SBA_Appv	Predictor	Numerical	USD	The amount of the approved loan that is guaranteed by SBA
MIS_Status	Response	Categorical	N/A	The status of the loan: **CHGOFF** if the loan is charged off; **PIF** if the loan is paid in full.

Appendix B

Sector	Description
11	Agriculture, forestry, fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities

23	Construction
31-33	Manufacturing
42	Wholesale trade
44-45	Retail trade
48-49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services
55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other Services (except public administration) 92 Public administration

Appendix C

		TRAIN	TEST
MIS_Status	N (default)	842	365
	Proportion (default)	17.11%	17.31%
Term	median	75	85
	interquantile	[39,84]	[40,84]
NoEmp	median	3	3
	interquantile	[2,7]	[2,6]
NewExist	N (New)	1021	472
	Proportion (New)	20.77%	22.38%
CreateJob	median	0	0
	interquantile	[0,1]	[0,1]
RetainedJob	median	3	2
	interquantile	[1,6]	[1,6]
UrbanRural	N (rural)	338	155
	Proportion (rural)	6.87%	7.35%
DisbursementGross	median	50000	50000
	interquantile	[25000,104539]	[25000,102000]
Industry	N (= 1)	842	365
	Proportion (N = 1)	17.11%	17.31%

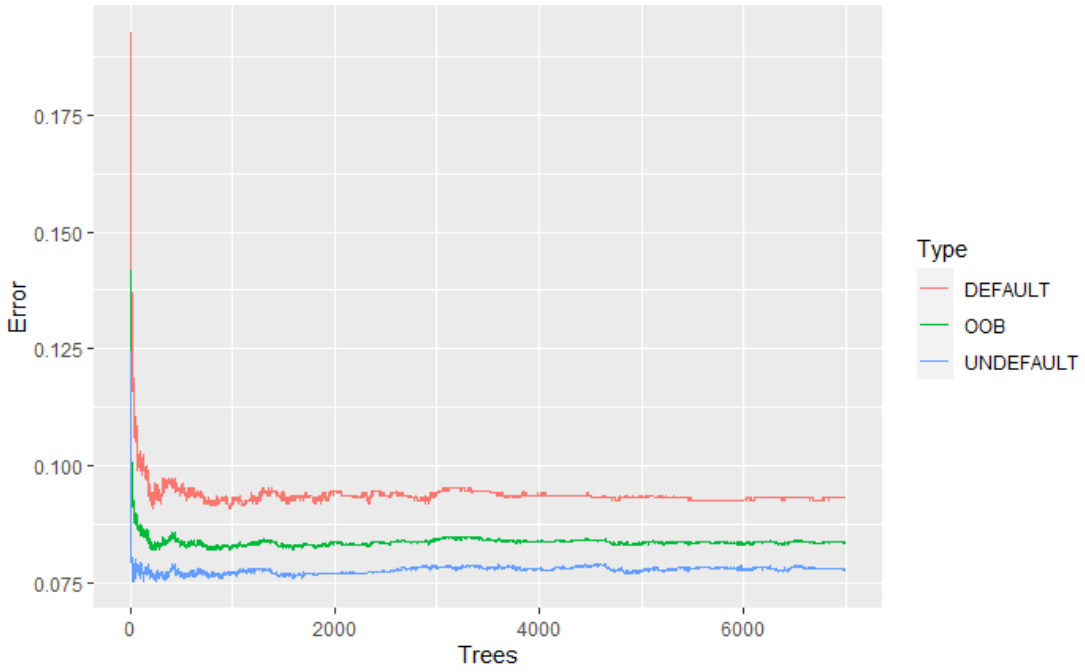
Appendix D Table 3.1 The sample of confusion matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive,** Type I error
	Predicted condition negative	**False negative,** Type II error	**True negative**

Appendix E Coefficients changes after excluding variables using stepwise (compared with Ridge).

Variable	Full Model	Reduced Model	Ridge
Term	-0.0559	-0.0558	1.9918
NoEmp	-0.0094	-0.0124	-0.0407
NewExist2	0.3327	0.3357	0.1579
CreateJob	-0.0035	NA	2.9736
RetainedJob	-0.0042	NA	-3.8359
UrbanRural2	-0.4342	-0.4303	-2.7831
DisbursementGross	4.44E-07	4.17E-07	2.39E-7
Industry1	-0.4105	-0.4077	-0.3034

Appendix F



Appendix G

	AUC	Acc	Sens	Spec	PPV	NPV	MCC	F1
Logistic	0.856	0.8146	0.6771	0.9439	0.919	0.7566	0.6477	0.7797
LASSO	0.852	0.8227	0.7592	0.8563	0.7371	0.8702	0.6114	0.748
Ridge	0.846	0.8212	0.7513	0.8597	0.7463	0.8628	0.6101	0.7488
Random Forests	0.974	0.9739	0.9748	0.9735	0.9532	0.9858	0.9436	0.9639

1. Introduction
2. Data Management
    CODE A
3. Method
4. Results

    4.1 EDA Results

        CODE B (Appendix C)
        CODE F (Figure 4.1)
        CODE G (Figure 4.2)
        CODE C (Figure 4.3)
        CODE D (Figure 4.4)
        CODE I (Figure 4.5)
        CODE L (Figure 4.6)
        CODE E (Table 4.1)

    4.2 Modeling Results

        4.2.1 Logistic Regression and Penalized Regression

            CODE H
            CODE J
            CODE K
            CODE M
            CODE P
            CODE R

        4.2.2 Random Forests

            CODE O
            CODE Q

        4.2.3 Models' Performance Comparison

            CODE N
            CODE S

Section 0

Package

```
library(MASS)
library(rcompanion)
library(randomForest)
library(caret)
library(pROC)
library(faraway)
library(dplyr)
library(ResourceSelection)
library(ggplot2)
library(pscl)
```

```
library(nortest)
library(insight)
library(cowplot)
library(glmnet)
```

Section 2.1

CODE A

```
#Modified factor
#select train and test sample
#train set : test set = 7:3
set.seed(123)
seq <- rnorm(4920)
train <- Cleaned_Data[sample(nrow(Cleaned_Data),4920), ]
test <- Cleaned_Data[-sample(nrow(Cleaned_Data),4920), ]

#Modified factor
as.factor(Cleaned_Data$Industry)

train$UrbanRural <- as.factor(train$UrbanRural)
train$NewExist <- as.factor(train$NewExist)
train$MIS_Status <- ifelse(train$MIS_Status == 'CHGOFF', 1,0)
train$Industry <- as.factor(train$Industry)
train$MIS_Status <- as.factor(train$MIS_Status)

test$UrbanRural <- as.factor(test$UrbanRural)
test$NewExist <- as.factor(test$NewExist)
test$MIS_Status <- ifelse(test$MIS_Status == 'CHGOFF', 1,0)
test$Industry <- as.factor(test$Industry)
test$MIS_Status <- as.factor(test$MIS_Status)

#check each variables is in the appropriate form
summary(train)
```

Section 4.1

##CODE B

```
#CODE B
#Appendix C
```

```
summary(train)
summary(test)
```

##CODE C

```
#Figure 4.3
#left graph
x <- TTold$Term
h<-hist(x, breaks=100, col="brown", xlab="Disbursement",
   main="Frequency of Disbursion")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)

#right graph
x <- train$Term
h<-hist(x, breaks=100, col="blue", xlab="Term",
   main="Fequcey of Term")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```

CODE D

```
#Figure 4.4
#Plot of NewExit
plot(train$NewExist,
     main = 'NewExit')

#Plot of UrbanRural
plot(train$UrbanRural,
     main = 'UrbanRural')

#Plot of Industry
plot(train$Industry,
     main = 'Industry')
```

CODE E

```
#Table 4.1
xtabs(~ Industry + NewExist + UrbanRural, train)
```

```r
Cor_Ind1 <- xtabs(~ NewExist + Industry, train)
cramerV(Cor_Ind1, ci = TRUE)

Cor_In2 <- xtabs(~ Industry + UrbanRural, train)
cramerV(Cor_In2, ci = TRUE)

Cor_In3 <- xtabs(~ NewExist + UrbanRural, train)
cramerV(Cor_In3, ci = TRUE)

Cor_Mis4 <- xtabs(~ MIS_Status + Industry, train)
cramerV(Cor_Mis4)

Cor_Mis5 <- xtabs(~ MIS_Status + NewExist, train)
cramerV(Cor_Mis5)

Cor_Mis6 <- xtabs(~ NewExist + UrbanRural, train)
cramerV(Cor_Mis6)
```

CODE F

```r
#Pearson Correlation
#Figure 4.1
pairs(train)

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    Cor <- abs(cor(x, y)) # Remove abs function if desired
    txt <- paste0(prefix, format(c(Cor, 0.123456789), digits = digits)[1])
    if(missing(cex.cor)) {
        cex.cor <- 0.4 / strwidth(txt)
    }
    text(0.5, 0.5, txt,
         cex = 1 + cex.cor * Cor) # Resize the text by level of correlation
}

pairs(train,
      upper.panel = panel.cor,    # Correlation panel
      lower.panel = panel.smooth) # Smoothed regression lines
```

CODE G

```
#Figure 4.2

plot(train$RetainedJob,train$NoEmp,
     main = 'relationship between RetainedJob & Noemp')
boxplot(train$Term,train$MIS_Status,
     main = 'relationship between Term & MIS_Status')


filter = which(train$MIS_Status == 0)
filter2 = which(train$MIS_Status == 1)
TTnew <- train[filter,]
TTold <- train[filter2,]
summary(TTnew$Term)



xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)

plot(train$NoEmp,train$DisbursementGross,
     main = 'relationship between Disbursion & Noemp')
plot(train$RetainedJob,train$DisbursementGross,
     main = 'relationship between Disbursion & RetainedJob')

plot(train$CreateJob,train$RetainedJob)




#create scatterplot of x1 vs. y1
plot(train$NoEmp, train$RetainedJob, col='red', pch=19)


#add scatterplot of x2 vs. y2
points(train$NoEmp, train$DisbursementGross/10000, col='brown', pch=19)

#add scatterplot of x2 vs. y2
points(train$NoEmp, train$CreateJob, col='darkblue', pch=19)

#add legend
```

```
legend('topleft', legend=c('RetainedJob', 'CreateJob','Disbursemen'), pch=c
(19, 19,19), col=c('red', 'blue','brown'))
```

#Section 4.2.1 #Code H

```
#Create Logistic Model
glm <- glm(MIS_Status ~ ., family = binomial, train)
sumary(glm)
beta <- coef(glm)
drop1(glm,test = 'Chi')
confint(glm_r)

linpred <- predict(glm)
predprob<- predict(glm, type = 'response')
head(predprob)

rawres <- train$MIS_Status - predprob
```

CODE I

```
#Figure 4.5

qqnorm(residuals(glm))
halfnorm(hatvalues(glm))

filter(train, hatvalues(glm) > 0.015) %>% select (Term,NoEmp,NewExist,Creat
eJob,RetainedJob,UrbanRural,DisbursementGross,MIS_Status,Industry)
```

Stepwise

##CODE J

```
glm_r <- step(glm, trace = 0)
sumary(glm_r)

#Note that we have excluded variables with high correlation
```

##Using HL Test ##CODE K

```
#For glm
predict_y <- predict(glm,x_test,type = 'response')
predict_y <- unname(predict_y)
preY <- ifelse(predict_y >= 0.1968882   , 1,0)
hoslem.test(exaY, preY)
```

```
#For glm_r
predict_y_r <- predict(glm_r,x_test,type = 'response')
predict_y_r <- unname(predict_y_r)
preY_r <- ifelse(predict_y_r >= 0.1982326   , 1,0)
hoslem.test(exaY, preY_r)
```

Graph

CODE L

```
#Figure 4.6
linpred <- predict(glm,train)
wcgsm <- na.omit(train)
wcgsm <- mutate(train,residuals=residuals(glm),predprob=predict(glm,train,t
ype = 'response'))
gdf <- group_by(wcgsm, cut(linpred, breaks=unique(quantile(linpred,(1:50)/5
1))))
hldf <- summarise(gdf, y=sum(residuals), ppred=mean(predprob), count=n())

hldf <- mutate(hldf, se.fit=sqrt(ppred*(1-ppred)/count))
ggplot(hldf,aes(x=ppred,y=y/count,ymin=y/count-2*se.fit,ymax=y/count+2*se.
fit))+geom_point()+geom_linerange(color=grey(0.75))+geom_abline(intercept=
0,slope=1)+xlab("Predicted Probability")+ylab("Observed Proportion")
```

Using McFadden's Pseudo - R^2

CODE M

```
#McFadden's

pR2(glm_r)
pR2(glm)
#values of 0.2 to 0.4 for ρ2 represent EXCELLENT fit."
```

Calculate Accuracy and Specificity

CODE N

```
#For glm
confusionMatrix( as.factor(exaY) , as.factor(preY), positive = "1")
```

```r
#For glm_r
confusionMatrix( as.factor(exaY) , as.factor(preY_r), positive = "1")
```

#Random Forest # CODE O

```r
set.seed(123)
#Random Forest (default mtry = 4, trees = 500)

rf <- randomForest(MIS_Status~.,data = train)
print(rf)

preY_RF <- predict(rf,test)

confusionMatrix(preY_RF,exaY, positive = "1")


#Find the suitable No. of variables tried at each split

##Tree = 500
oob.values <- vector(length=10)
for(i in 1:10) {
  temp.model <- randomForest(MIS_Status ~ ., data = train, mtry=i, ntree=500)
  oob.values[i] <- temp.model$err.rate[nrow(temp.model$err.rate),1]
}
oob.values

##Tree = 1000
oob.values <- vector(length=10)
for(i in 1:10) {
  temp.model <- randomForest(MIS_Status ~ ., data = train, mtry=i, ntree=1000)
  oob.values[i] <- temp.model$err.rate[nrow(temp.model$err.rate),1]
}
oob.values

##Tree = 5000
oob.values <- vector(length=10)
for(i in 1:10) {
  temp.model <- randomForest(MIS_Status ~ ., data = train, mtry=i, ntree=5000)
  oob.values[i] <- temp.model$err.rate[nrow(temp.model$err.rate),1]
}
oob.values
```

```
##Tree = 7000
oob.values <- vector(length=10)
for(i in 1:10) {
  temp.model <- randomForest(MIS_Status ~ ., data = train, mtry=i, ntree=70
00)
  oob.values[i] <- temp.model$err.rate[nrow(temp.model$err.rate),1]
}
oob.values

###So we should choose no. = 4 with 5000 trees



#Modify the No. of variables tried at each split

##Explore more trees (Try 5000 Trees)
rf_new <- randomForest(MIS_Status ~.,data = train,mtry = 4, ntree = 5000)
print(rf_new)

preY_RF_new <- predict(rf_new,test)

confusionMatrix(preY_RF_new,exaY, positive = "1")

##graph
obb.error.data <- data.frame(
  Trees = rep(1:nrow(rf_new$err.rate), times = 3),
  Type = rep(c('OOB','UNDEFAULT','DEFAULT'), each = nrow(rf_new$err.rate)),
  Error = c(rf_new$err.rate[,'OOB'],
    rf_new$err.rate[,'0'],
    rf_new$err.rate[,'1'])
)

ggplot(data = obb.error.data, aes(x = Trees, y = Error)) + geom_line(aes(co
lor = Type))
```

CODE P

```
deviance(lasso.model)
deviance(ridge.model)
deviance(glm_r)
```

IMPORTANCE

CODE Q

```
varImpPlot(rf_new)
```

LASSO, RIDGE & ELASTIC NET

CODE R

LASSO

```
set.seed(123)
# Predictor variables
x <- model.matrix(MIS_Status~., train)[,-1]
# Outcome variable
y <- train$MIS_Status

cv.lasso <- cv.glmnet(x, y, alpha = 1, type.measure = 'deviance', family = "
binomial")
plot(cv.lasso)

cv.lasso$lambda.min
coef(cv.lasso, cv.lasso$lambda.min)
coef(cv.lasso, cv.lasso$lambda.1se)

# Final model with lambda.min
lasso.model <- glmnet(x, y, alpha = 1,type.measure = 'deviance', family = "
binomial",
                      lambda = cv.lasso$lambda.1se)
# Make prediction on test data
x.test <- model.matrix(MIS_Status ~., test)[,-1]
pre_lasso <- lasso.model %>% predict(newx = x.test, type = 'response')
preL <- ifelse(pre_lasso >= 0.4, 1,0)
# Model accuracy
confusionMatrix(as.factor(exaY) , as.factor(preL), positive = "1")
```

RIDGE

```
cv.ridge <- cv.glmnet(x, y, alpha = 0, type.measure = 'deviance', family = "
binomial")
plot(cv.ridge)
```

```
cv.ridge$lambda.min
coef(cv.ridge, cv.ridge$lambda.min)
coef(cv.ridge, cv.ridge$lambda.1se)

# Final model with lambda.min
ridge.model <- glmnet(x, y, alpha = 0,type.measure = 'deviance', family = "
binomial",
                      lambda = cv.ridge$lambda.1se)
# Make prediction on test data
x.test <- model.matrix(MIS_Status ~., test)[,-1]
pre_ridge <- ridge.model %>% predict(newx = x.test, type = 'response')
preL <- ifelse(pre_ridge >= 0.4, 1,0)
# Model accuracy
confusionMatrix(as.factor(exaY) , as.factor(preL), positive = "1")
```

ROC

CODE S

```
set.seed(123)

pre_rf_pro <- predict(rf,test,type = 'prob')
pre_rf_pro_new <- predict(rf_new,test,type = 'prob')

#For Random Forest
RF.testing.ROC <- roc(test$MIS_Status ~ pre_rf_pro[, 1], quiet = TRUE)
plot(RF.testing.ROC)

#For Random Forest (NEW)
RF.testing.ROC_new <- roc(test$MIS_Status ~ p1_new[, 1], quiet = TRUE)
plot(RF.testing.ROC_new)

#For glm (NOT glm_r)
GLM.testing.ROC <- roc(test$MIS_Status ~ predict_y,quiet = TRUE)
plot(GLM.testing.ROC)
coords(GLM.testing.ROC, "best", "threshold")

#For glm_r
GLM_R.testing.ROC <- roc(test$MIS_Status ~ predict_y_r,quiet = TRUE)
plot(GLM_R.testing.ROC)
coords(GLM_R.testing.ROC, "best", "threshold")
```

```r
#For LASSO
GLM_LASSO.testing.ROC <- roc(test$MIS_Status ~ pre_lasso[,1],quiet = TRUE)
plot(GLM_LASSO.testing.ROC)
coords(GLM_LASSO.testing.ROC, "best", "threshold")

#For Ridge
GLM_ridge.testing.ROC <- roc(test$MIS_Status ~ pre_ridge[,1],quiet = TRUE)
plot(GLM_ridge.testing.ROC)
coords(GLM_ridge.testing.ROC, "best", "threshold")


##Put glm_r, glm, And LASSO together
plot(GLM.testing.ROC,type="l",col="red")
lines(GLM_R.testing.ROC,col="blue")
lines(GLM_LASSO.testing.ROC,col="blue")


#Notebook refer a graph with threshold and Proportion
#It also mention about R2 attributed to Nagelkerke, we can refer it and do a
 comparation

#Should be 1 - specificity (remember to correct it)



#Calculate AUC
auc(GLM_R.testing.ROC)
auc(GLM.testing.ROC)
auc(RF.testing.ROC)
auc(RF.testing.ROC)
auc(GLM_LASSO.testing.ROC)
auc(GLM_ridge.testing.ROC)
```