Group 1:

Name	Contribution
Jonathan Franko	Wrote paper and included insights from and MBA perspective, experimented with models such as random forest and SVM that proved to be more complicated and computationally intensive than required.
Poonam Mathur	Handled data cleaning. Tried experimenting with models like PCA, Adaaboost. Constructed models used in final classification and regression.
SiJing Li	Created presentation for explaining the construction of our models.
Group (all)	Much of the work was done in group meetings outside of class in addition to sharing scripts through google drive. Scripts were continuously revised and improved by group members helping each other resolve errors and improve how data was being handled between two models.

Introduction

The data provided consisted of 80,000 observations and 763 variables. With the exception of the target variable "loss" all variables were unnamed to prevent revealing personal financial information. The goal of this model was to determine the probability of default for a loan, and the loss it incurred. Two solve this problem it was determined that our team would need to take a two step approach, first classification to determine which observations would default, then regression to determine the total loss incurred from those predicted to default.

Data Preparation

Variables in the data set presented a unique challenge in preparing the dataset due to both the large number of variables, and the lack of information regarding the variables that could be

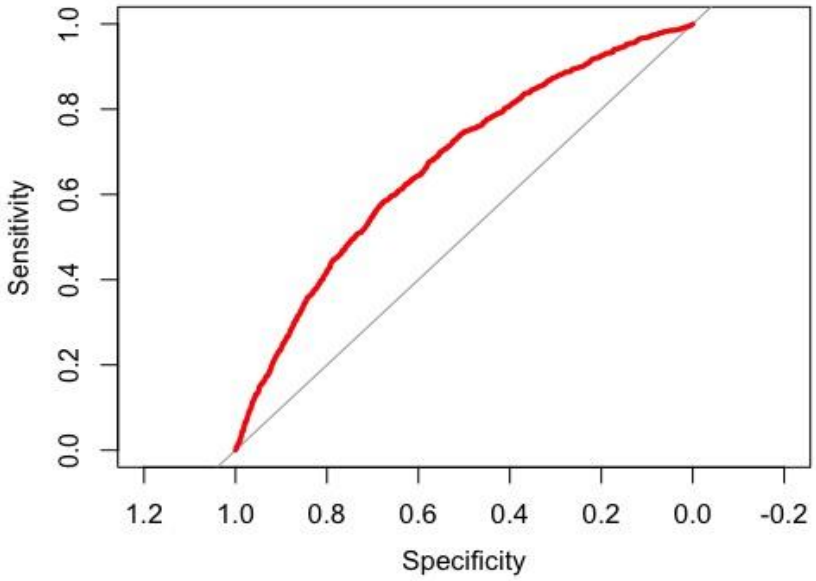
inferred. Our first goal was to reduce the number of variables that would need to be used to build a model. We noticed many observations had multiple NA values and decided to omit the observations which had NA values. Another option considered was imputing the median values to replace missing values, but due to the number of observations it was decided to sacrifice roughly half of the 80,000 observations instead of imputing a median value. We also noticed that many columns were identical, so variables that were repeated were removed from the dataset. The final step in data exploration was to check for correlation between variables, and omit any that were greater than 90% correlated to account for variables that were nearly identical. After removing observations with missing data and near zero variance variables we were left with 235 variables and 39,430 observations to build our model.

Building the models:

The approach taken was to build a classification model and a linear model. All data would be fed to the the classification model for training, and the linear model would only be trained on the observations that defaulted as we did not want the linear model to be biased towards the approximately 90% of observations that did not default. When running test data through the model it will first be determined whether or not the observation will default, those that did not default will be withheld from the regression model, and will have their predicted loss, “PD” in the script, set to zero. Those that were determined to default will be passed to the regression model to determine loss. After the data has been run through both models, the defaulted observations will be bound to the data that was held out of the regression model to create one output.

At this point we ran the test data through the linear model to determine the accuracy of combining the two models discussed. To do this we first split the data into a train and test data set. The first model built was a generalized linear model to predict probability of default for each observation using all available variables. The pROC package was used to calculate AUC, which was 0.67 for the classification model. Displayed below is the output and on the following page is the AUC curve

```
Data: V in 10579 controls (TestData$default 0) < 1250 cases (TestData$default 1).
Area under the curve: 0.67
```

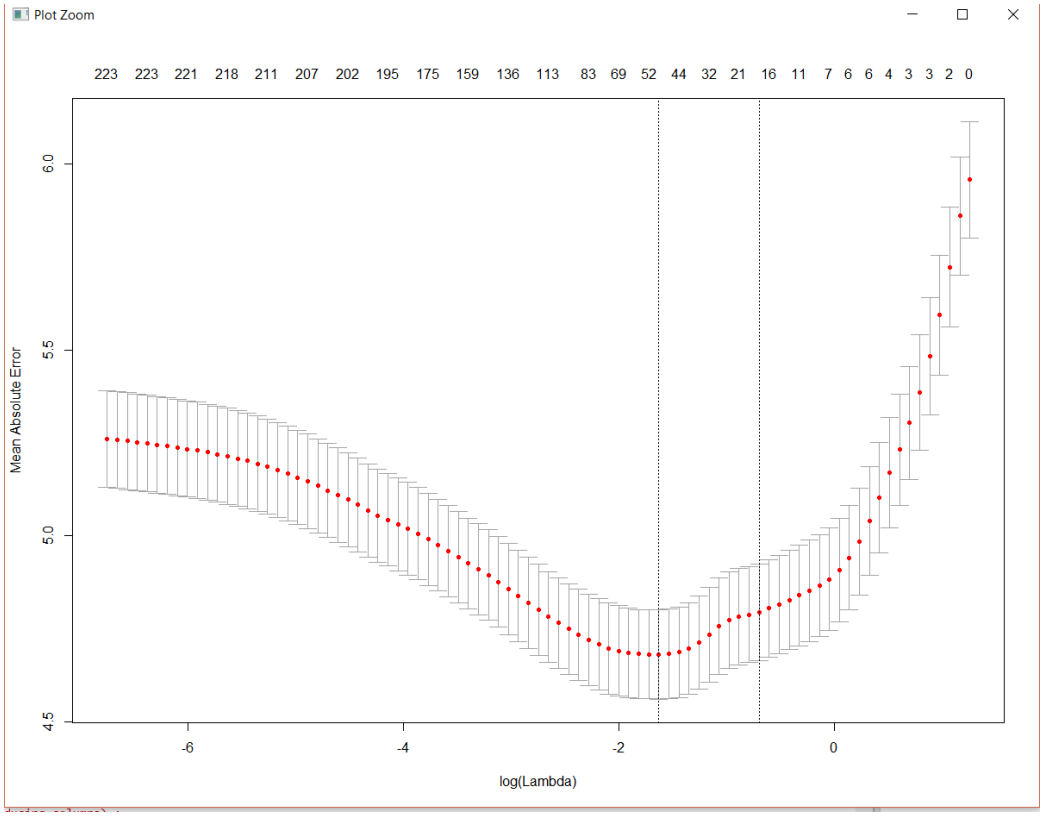


The next step we needed to take was to determine what the appropriate cutoff point would be between observations that were predicted as default and not default. To achieve this we took a partially calculated, partially trial and error approach. Looking at the training data it was noticed that approximately 10% of the observations defaulted. Based on this we hypothesized that it would be a similar case in the test data, and calculated the percentiles for the output of the classification model.

```
>   quantile(V, prob = seq(0, 1, length = 11), type = 5)
          0%          10%          20%          30%          40%          50%
6.476056e-06 3.087356e-02 4.479746e-02 5.713152e-02 6.981768e-02 8.422610e-02
         60%          70%          80%          90%         100%
1.003349e-01 1.198847e-01 1.466881e-01 1.905347e-01 1.000000e+00
```

It can be seen that at a 19% probability of default cutoff we would classify 10% of the data as default. Through trial and error we were able to lower the mean average error by decreasing the cutoff to 14.7%, which is where 80% of observations were classified as default.

The method chosen to build the model for predicting loss amount was a cross validated lasso regression model using cv.glmnet. The model was trained only from the observations that did default from the training set so that the model wouldn't be biased towards the observations that did not default.



At last, Test data set consisting of only values predicted to default based on the cutoff point, and into a holdout data set that would not be run through the model as they were determined to not default. We first changed the values into the binary number 0 and 1 and created one column "Default" and tried to predict who will default with how much percentage we extracted values only customer those who contain 1 in Default column. Once predicted values for loss were determined they were added to the test dataset in a column representing predicted loss (PL). The observations that were withheld had a PL variable added as zero, then the two datasets were combined into one output. Plotting the MAE of the model we can see that the model is most accurate when approximately 23-50 variables are used, and that on its own the model has an MAE of approximately 5.2 at a lambda value of 0.195. After changing the threshold to 0.20 we received MAE prediction which was reduced to 1.722

Conclusions

Because the variables are unmarked there is little we can conclude in regards to how and why certain variables affect the probability of default, however as we previously mentioned the lasso regression model was most accurate with only 23-50 variables. Future projects should be focused on determining what these variables are so that the cost of collecting information of the financial institution is reduced. The classification model had some predictive power, but should not be considered the only factor when determining whether or not to approve a loan. Displayed below is a confusion matrix of the predicted values vs the cutoff values of the determined cutoff point.

	True	
Predicted	0	1
FALSE	8696	777
TRUE	1883	473

It can be seen that although we were able to correctly identify 473 defaulted loans, we incorrectly identified 1883 as defaulting when they did not default, and failed to identify 777 that did default. If the financial were to use this model as gospel for determining whether or not to approve it would be turning down many potential opportunities. We chose to include a column in the prediction file that includes the default prediction in addition to the loss prediction. The reasoning for this is because that number can be extremely valuable in determining interest rates for loans that are at a higher risk for default.

What we were especially pleased with was how much the MAE was reduced between the training and test datasets when combining the classification and regression models. With this information the financial institution can not only get a better understanding over what loans will default, but a fairly accurate assessment of how much they are going to default by. This can be used in determining the types of loans approved, how much needs to be taken as collateral on loans, and what the interest rates will be on the loans so that the risk the bank is taking is compensated.