Samadipa Saha
Uma Sreeram
Brett Watanabe

Determining Potential Credit Risks from Loan Data

Abstract

Experimented with various models to predict if an individual is good or bad credit risk based on a wide number of predicting variables. Logistic regression provided with a good interpretation of how the predicting variables are related to the response variable and which variables might be more significant in explaining the variability of the response. A bimodal normal histogram plot of residuals was observed, indicating that a significant predicting variable or interaction term is missing. Since it is difficult to identify good interaction terms when there are a large number of categorical variables using logistic regression, we used decision trees. Also tried other techniques like random forest, SVM and KNN. Although the accuracy was similar for quite a few of our models. The specificity, ie., the true negative/ bad credit risk rate was highest for lasso regression. This is particularly important, because wrongly classifying bad credit risk individuals might be more costly for the bank.

Introduction

Reasons for our analysis:

When a customer wants to purchase something, but does not have the funds, they may go to a bank to get a loan. Banks charge interest on loans so it is in their best interest to loan money to customers as long as those customers pay them back. However, there are customers who do not pay back their loans and cause the bank to lose money. If banks can identify customers that are good or bad credit risks, they can be more selective about who they loan money to and save themselves money.

Project goals:

In this study, we use bank data from a German bank to identify customers who are good and bad credit risks. We use several classification techniques including logistic regression, support vector machines (SVM), k-nearest neighbors (KNN), decision trees, and random forests.

A Priori Expectations:

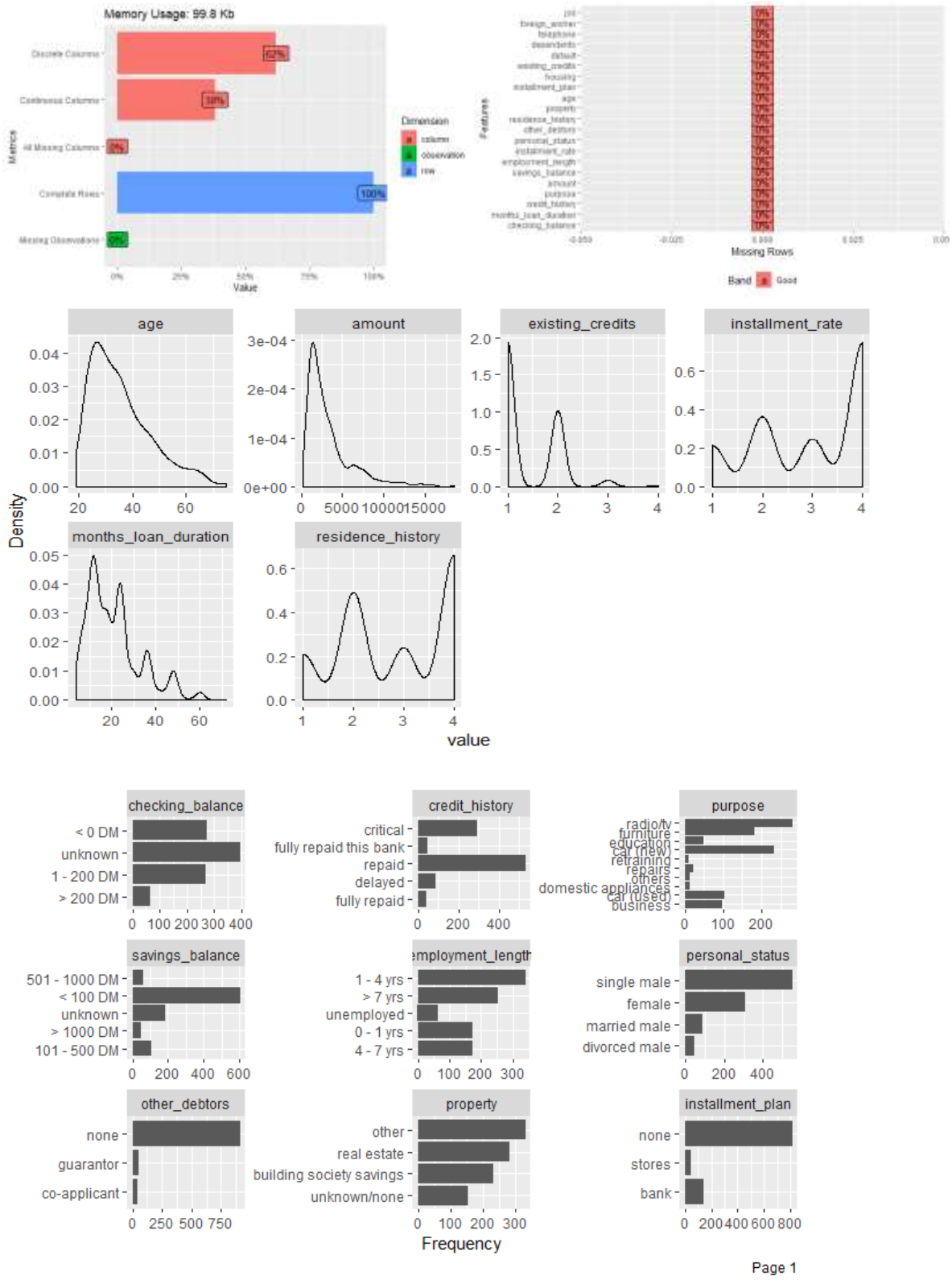
We expect that logistic regression will provide valuable insights into the relationship of the response variable with the predicting variables and random forest will provide good predictability.

Description of the Data:

This data came on the UCI Machine Learning Data Repository at the following url:
https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
The dataset has 1000 rows with 21 columns. The original dataset has an unusual labeling system so we reformatted it to be human-readable.
We are using the following variables:
Response variable:
not_default - indicating if the person is good credit risk (1) or if the person is bad credit risk (0)

Predicting Variables	Type	Description
checking_balance	Categorical	Indicates if the checking balance is 200 DM or not
months_loan_duration	Quantitative	Duration of the loan in months
credit_history	Categorical	Credit status for if the customer has paid their debts
purpose	Categorical	Purpose for the loan
amount	Quantitative	Amount of the loan (DM)
savings_balance	Categorical	Ranges of money in savings account
employment_length	Categorical	Ranges of years employed
installment_rate	Categorical (Quantitative)	Installment rate in percentage of disposable income
personal_status	Categorical	Gender and marital status
other_debtors	Categorical	None, guarantor, or co-applicant
residency_history	Categorical (Quantitative)	Years at current residence
property	Categorical	Type of residence
age	Quantitative	Age of customer
installment_plan	Categorical	None, bank, or stores
housing	Categorical	Rent, own, or for free
existing_credits	Categorical (Quantitative)	Number of existing credits at bank
dependents	Categorical	Number of people being liable to provide maintenance for
telephone	Categorical	If customer has a telephone
foreign_worker	Categorical	If customer is a foreign worker
job	Categorical	Job type

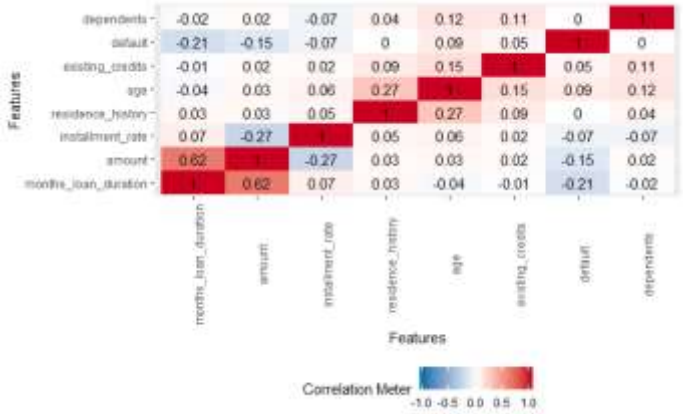
Exploratory Data Analysis:



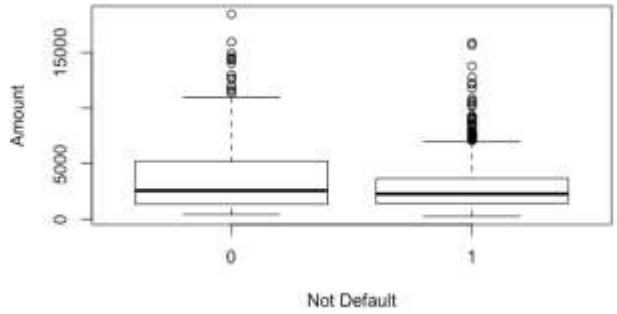
Inferences:

- 70% of the rows in our dataset has information about people classified as good credit risks.
- The dominant demographic includes:
    - Single Males

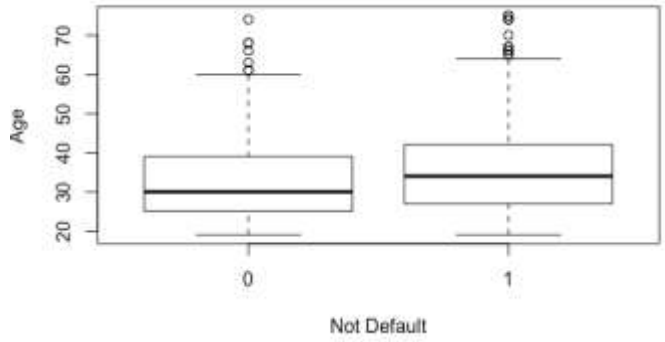
    - Employment length: mostly less than 1-4 years. A significant portion with experience greater than 7 years
    - Highly skilled
    - Foreign workers
    - Fully repaid their previous loans
- Majority of the population sampled are not enrolled in any installment_plan,have dependents and have their own housing
- Most of the loans taken in the past by the people sampled were taken individually (no co-debtors/co-applicants cited)
- Purpose of the loans availed so far were majorly for radio/tv and new cars.
- Majority of the loans availed were for amounts in the range of 0-5000(units)
- Number of existing_credits (lines of credits) are mostly 1 or 2
- Studying the correlation matrix of continuous variables, we can conclude that:
    - Amount has an obvious higher correlation with month_loan_duration as higher the loan, it would take longer to pay off
    - Month_loan_duration also has a slight negative correlation with our response variable 'default', showing that most loans defaulted were higher in month_loan_duration.
    - There is a slightly positive correlation between age and residence history (Older the person, more likely they are to have lived in a place for a long time)
    - Interestingly, installment_rate has a slight negative correlation with amount
      This is probably due to the bank offering more attractive rates to people availing loans of a higher amount.



- We also plotted boxplot for the quantitative variables against the response variable.
    - The boxplot of amount vs response shows that the mean amount for default (0) is greater than the mean amount for no default which indicates individuals taking higher loan amounts are more likely to default than individuals taking lower loan amounts.



- The boxplot of age vs response shows that the mean age for default (0) is lower than the mean amount for no default which indicates that younger individuals are more likely to default.



Methods used and Model Assumptions

We classified customers as good or bad credit risks using logistic regression, decision trees, random forest, support vector machines(SVM), and the k-nearest neighbors algorithm (KNN). We also applied variable selection to several of the models to see if that improved classification rates. Our data was split into a training set and a test set. The training set consisted of 80% (800 rows) of the original data and the test set consisted of 20% (200 rows). The training set was used to train our models and the test set was used to calculate the classification accuracy.

Logistic Regression

Logistic regression is a supervised learning algorithm used for classification. It models the probability of success given predictors and it links the probability of success given predictors to the predicting variables using a non-linear logit link function.

$$p(x_1, \ldots, x_p) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

We started with logistic regression because it is simple and easier to interpret. We can see the relative importance of each factor in the model visually.

We tried several models for logistic regression as summarized below.

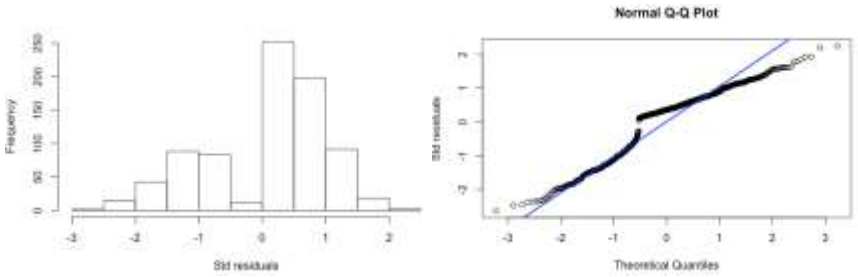
1. We fitted logistic regression model to the full dataset, after factoring the categorical variables into dummy variables.
    - Only 15 out of the 50 total variables (including dummy variables) were significant at 0.05 level significance. The p-value for the test of overall regression was 0, which indicates that at least one variable is significant in explaining the variability of the

response. The significant variables are checking_balance> 200 DM, checking_balance unknown, months_loan_duration, credit_historyfully repaid, credit_historyfully repaid this bank, credit_history repaid, purpose car (used), amount, savings_balance> 1000 DM, savings_balanceunknown, employment_length4 - 7 yrs, installment_rate4, installment_plannone, existing_credits and foreign_worker yes.

We make the key observations:

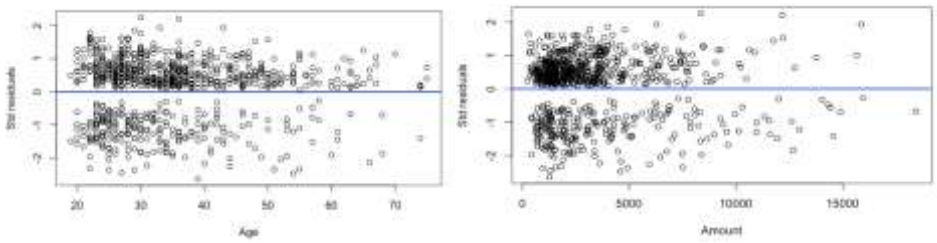
- ❏ The coefficient for amount is -1.384e-04. One unit increase in amount decreases the log odds ratio of good credit by -1.384e-04, given other predictors in the model. This satisfies our earlier exploratory data analysis claim using boxplot of amount vs not default.
- ❏ The coefficient for month loan duration is -3.155e-02. One unit increase in month loan duration decreases the log odds ratio of good credit by -3.155e-02, given other predictors in the model. This satisfies our earlier exploratory data analysis claim using correlation matrix.
- ❏ The coefficient for employment length 4-7 years is 7.448e-01. For individuals with employment length 4-7 years vs employment length >7 years, the log odds ratio of good credit risk increases by 7.448e-01, given other predictors in the model. This doesn't supports our intuition that individuals with a longer employment history would be more likely to be good credit risk.
- ❏ The coefficient for foreign worker yes is -1.308e+00. For individuals who are foreign workers vs who are domestic workers, the log odds ratio of good credit risk decreases by -1.308e+00, given other predictors in the model. This supports our intuition that foreign workers might be bad credit risk because they can flee away from the country without much legal obligation.

- We performed goodness of fit tests to verify if the logit link function was appropriate fit for the data.Using both pearson and deviance residuals, the p-value was greater than 0.05, hence we do not reject the null hypothesis of good fit.
- Performing the test for overdispersion, the over-dispersion metric was less than 2. Thus there was no over-dispersion.
- The histogram plot and qq plot of the residuals showed a bimodal distribution, indicating an important predicting variable or interaction term was missing from our data.

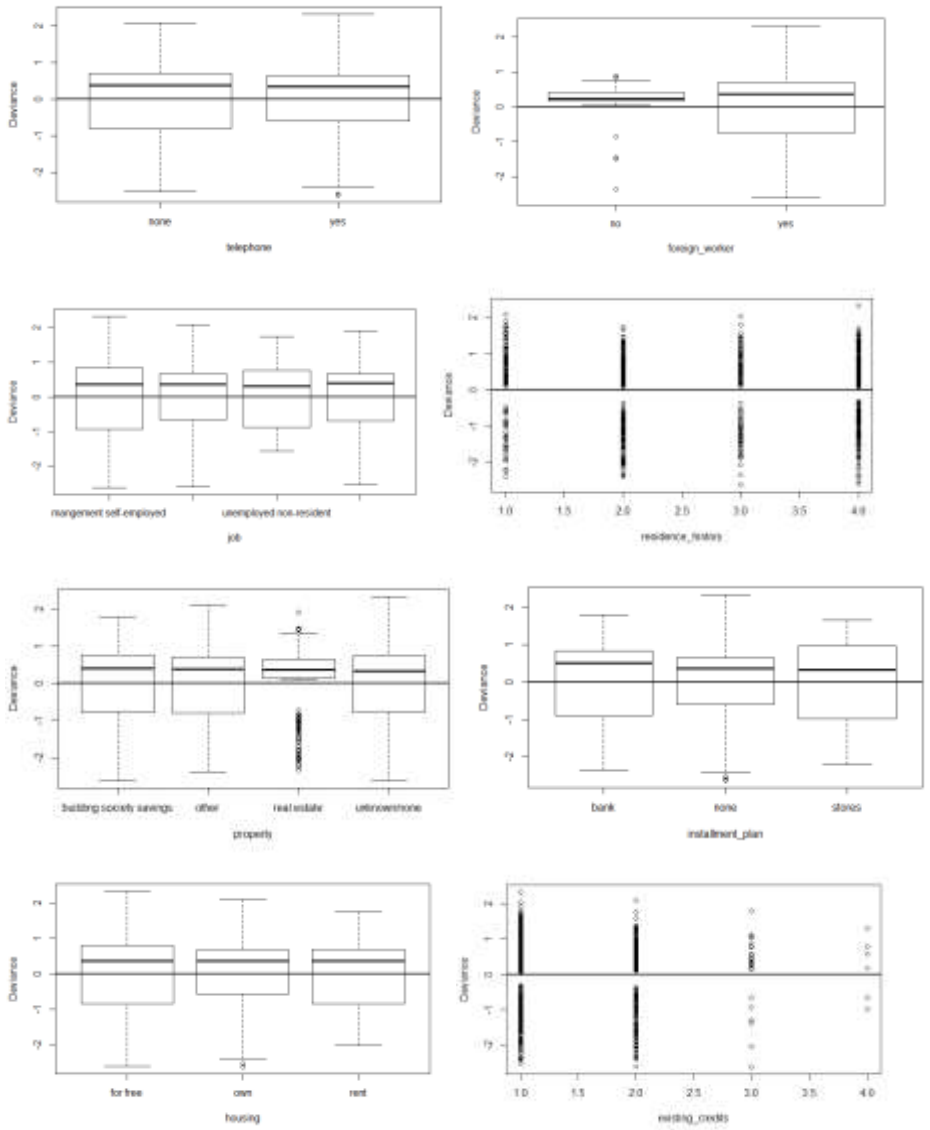


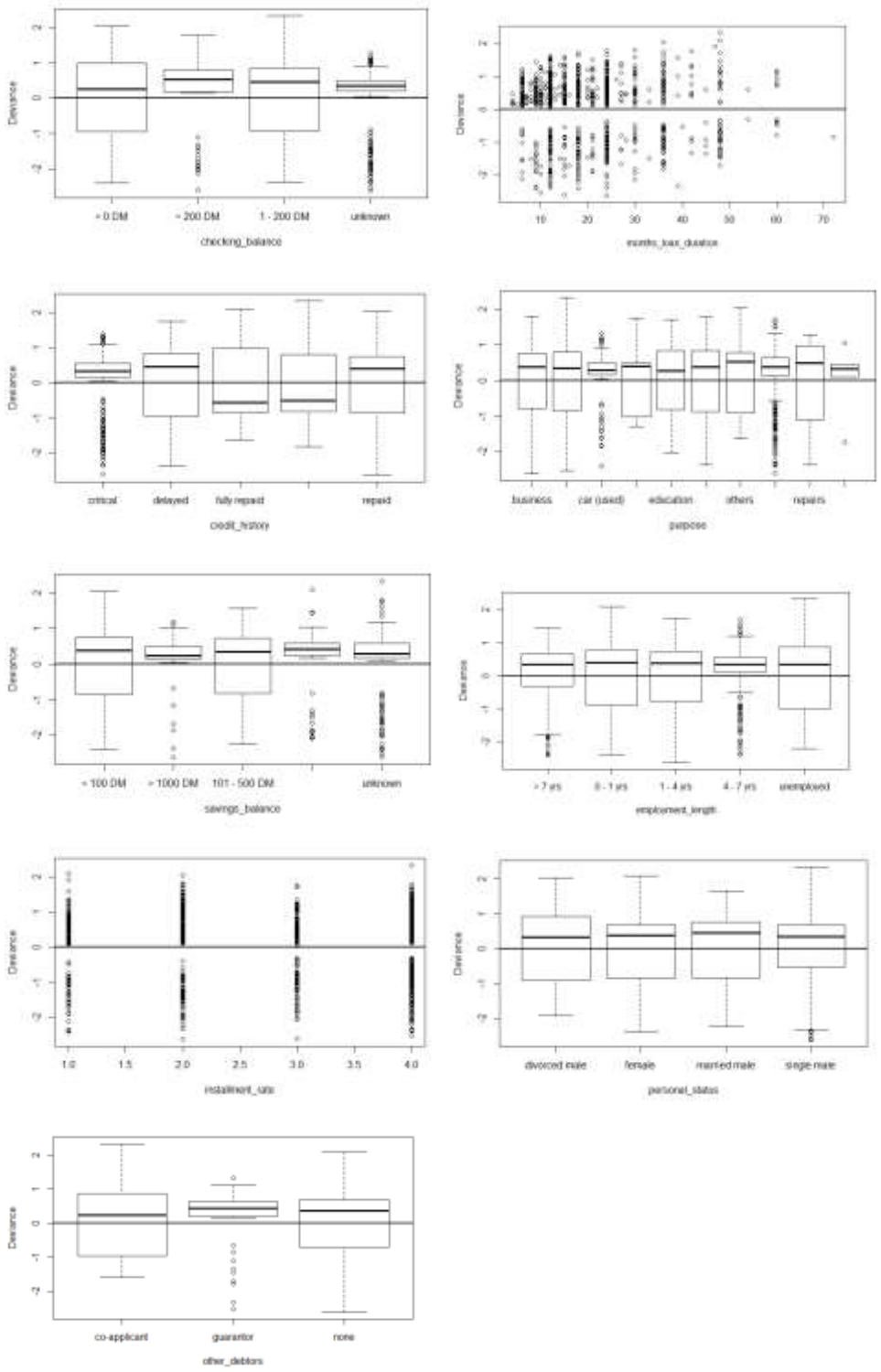
- The residual plots against the predictors for some variables namely age,amount,residence history, working credits,months loan duration and installment

rate are distributed equally equally about the zero line. However, we can see some clustering among the residuals, suggesting that the independence assumption might not hold.

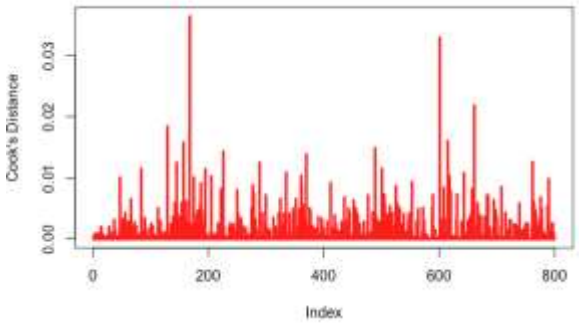


- Categorical variables like telephone, foreign worker, job, property, installment plan, housing, checking balance, purpose ,employment length, personal status, other debtors have the median residual value of different categories on one side of the zero line indicating some bias





2. Plotting the cook's distance for each observation provided us with the outlier observations. We fitted our second model by removing the outlier observations from our dataset. 23 out of 50 variables are significant at the 0.05 significance level.



3. We fitted the next model with the probit link function on all the training data and compared the model performance against the full model fitted earlier with the logit link function.
4. The fourth model was fitted using forward stepwise regression. This was done to reduce the large feature set and to avoid possible multicollinearity among the predicting variables. For forward stepwise regression, we start with a model with only the intercept and add variables incrementally to the model which minimizes the criterion value. The criterion used is the sum of the prediction risk and the complexity penalty. $R_{tr}(S) + Complexity\ Penalty$ Three kinds of complexity penalty can be used.
    - Mallow's cp

      $$\text{Complexity Penalty} = \frac{2|S|\hat{\sigma}^2}{n}$$

      where |S| is the number of predictors and $\hat{\sigma}^2$ is the estimated variance based on the full model.

    - Akaike Information Criteria (AIC)

      $$\text{Complexity Penalty} = \frac{2|S|\sigma^2}{n}$$

      where σ^2 is the true variance of the model which can be replaced with an estimate from the full model or submodel.

    - Bayesian Information Criteria (BIC)

      $$\text{Complexity Penalty} = \frac{|S|\sigma^2 \log(n)}{n}$$

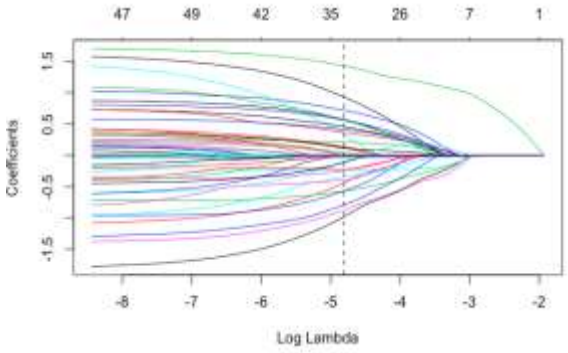
      where σ^2 is the true variance of the model which can be replaced with an estimate from the full model or submodel. BIC penalizes complexity most and is most preferred in model selection.

    41 out of 50 variables were selected by the forward stepwise regression model. Out of the 41 selected variables, only 15 variables were significant at the 0.05 significance level.

5. The fifth model was fitted using the backward stepwise regression and a comparison was made between the variables selected using forward stepwise regression and backward stepwise regression.For backward stepwise regression, we start with a model with only the intercept and remove variables incrementally from the model which minimizes the criterion

value. The criterion used is the sum of the prediction risk and the complexity penalty as discussed before in forward stepwise regression. 41 out of 50 variables were selected by the backward stepwise regression model. Out of the 41 selected variables, only 15 variables were significant at the 0.05 significance level.

6. We tried to perform feature selection using lasso regression to optimize for the bias variance tradeoff. Lasso regression estimates the predictor coefficients by minimizing the penalized sum of squares errors. Numerical algorithms are used to estimate the coefficients as there is no closed form solution. 32 out of 50 variables were selected using lasso regression.
   The regression coefficient path of the lasso regression model was also plotted to visualize which coefficient enters the model early.



7. Next, we performed elastic net regression because it removes the limitation on the variables selected as imposed by lasso regression and also stabilizes the L1 regularization path. We performed elastic net regression with alpha = 0.5.

Support Vector Machine

A support vector machine (SVM) is a supervised learning model used for classification. The model represents each sample as a point in an R^n space in such a way that samples of a different category can be divided by a gap in that space. New samples that are on one side of that gap are classified as one category and samples on the other side of the gap are classified as another category. We used SVMs to classify customers as good or bad credit risks.

We created models that used several types of algorithms used for pattern analysis called kernels. We used linear, polynomial, Gaussian RBF, hyperbolic tangent, Laplacian, and Bessel kernels.

K-Nearest Neighbors

K-nearest neighbors algorithm is a classification method. When predicting the class of a new data point, it looks at that point's k nearest neighbors in the data set's feature space and assigns the data point the class that the majority of its neighbors have. We used k-nearest neighbors to classify customers as good or bad credit risks with k being the odd integers between 3 and 25.

Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences.

We got a flowchart-like structure in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents the probability of default. As the result we obtained a vector which was the probability of the record being a good credit risk(1).

Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.This assigns a probability of a record being a good credit risk(1) after polling all its candidate trees.

Results

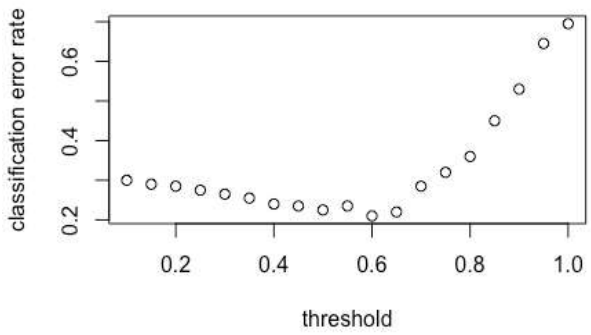
Logistic Regression

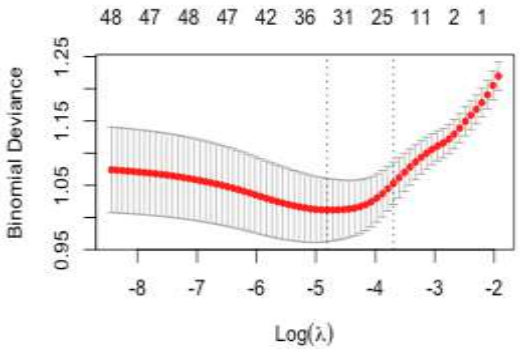
We fitted each of the above discussed logistic regression model on the training set and evaluated the model performance on the testing set. The prediction returned with probabilities of each observation being in class 1, which were then converted to class labels by using the threshold value r, where the classification error rate was minimized. All observations with probabilities > r are classified as 1 (good credit risk) and <= r as bad credit risk. The following table shows the performance of various logistic regression models fitted on the training data.

Model	Accuracy	Sensitivity	Specificity	AUC
Full Model	0.79	0.8540146	0.6507937	0.7359
Full Model w/o outliers	0.79	0.8776978	0.5901639	0.7339
Full Model with probit	0.79	0.8561151	0.6393443	0.7477
Forward step model	0.79	0.8776978	0.5737705	0.7257
Backward step model	0.79	0.8776978	0.5737705	0.7257
Lasso	0.79	0.8201439	0.7213115	0.7707
Elastic Net	0.77	0.8417266	0.5901639	0.7159

From the table above we can see that the overall accuracy is almost similar across all the models. But an interesting fact to note is that 1. the specificity for the **lasso** regression model is much higher compared to the other models. Specificity corresponds to the true negative rate. In our case the true bad risk rate. This is particularly important if wrongly identifying a bad credit risk individual is much more costlier to the the loaning bank than wrongly classifying a good credit risk individual as bad credit risk. Also, 2. the area under ROC curve for the model fitted with lasso regression is higher than the other models.

The classification error rate for various threshold values r for lasso regression model is plotted in figure (i) below :



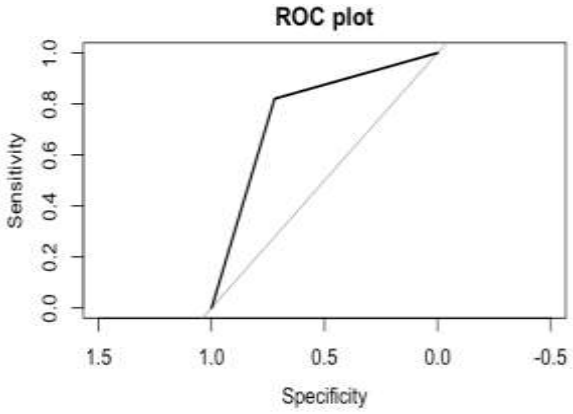


(i) (ii)

The plot (ii) above displays the cross-validation error according to the log of lambda. The left dashed vertical line indicates that the log of the optimal value of lambda is approximately -5, which is the one that minimizes the prediction error. This lambda value will give the most accurate model. The exact value of lambda was found to be **0.00813013**.

The confusion matrix for predictions using lasso regression model on the test set and the ROC plot is given below. Please not that 0 corresponds to bad credit risk and 1 corresponds to good credit risk.

```
          Reference
Prediction  0   1
         0 44  25
         1 17 114
```



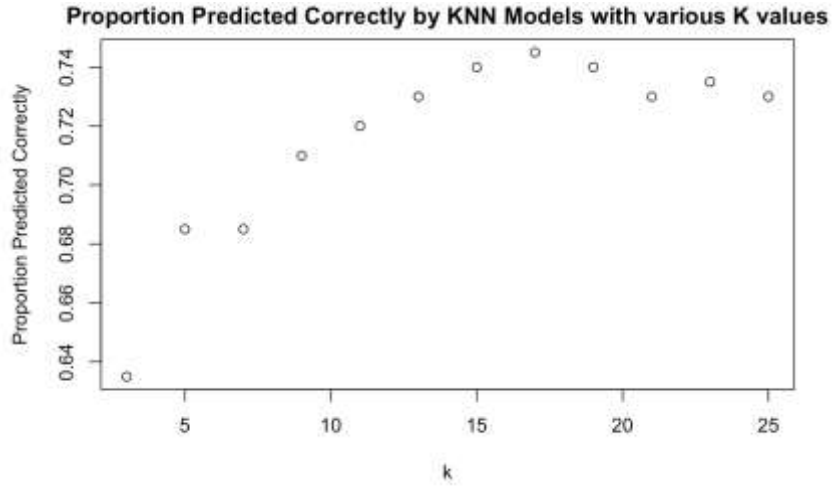
SVM

Of all the kernels we used, the linear kernel performed the best.

Kernel	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Linear	0.730	0.8705036	0.4098361	0.6401698
Polynomial	0.730	0.8705036	0.4098361	0.6401698
Laplacian	0.710	0.8705036	0.3442623	0.6073829
Gaussian RBF	0.675	0.8057554	0.3770492	0.5914023
Hyperbolic tangent	0.660	0.7625899	0.4262295	0.5944097
Bessel	0.535	0.6762590	0.2131148	0.4446869

KNN

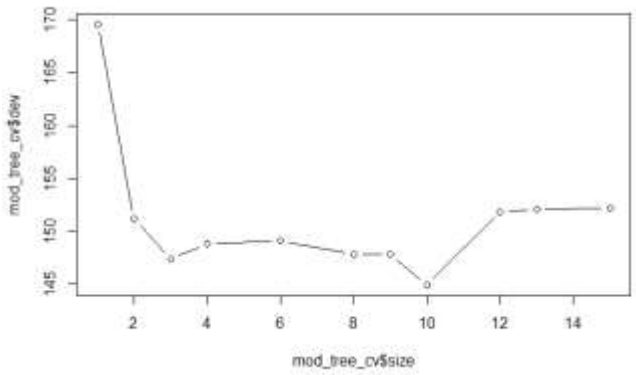
Of all the k's that we used, the highest accuracy was returned when k = 17. When k = 17, the accuracy, sensitivity, sensitivity, and AUC are 0.745, 0.9209, 0.3443, and 0.6326 respectively.



Decision Tree

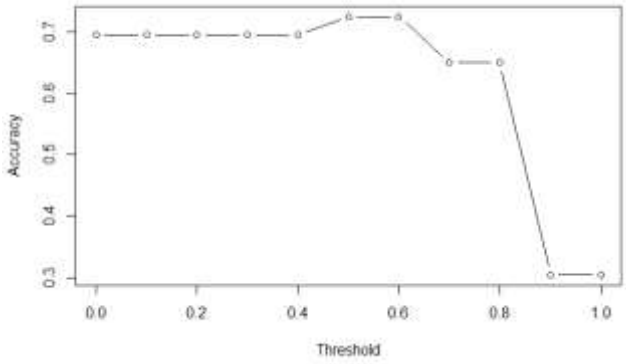
With an initial fit, we obtained a decision tree with a depth of 6. We then performed cross-validation on the training data to try different options of depth. We checked the deviance residuals at different options and found that we obtain the least residuals at depth 3.

Deviance at different depths:



Pruned tree:

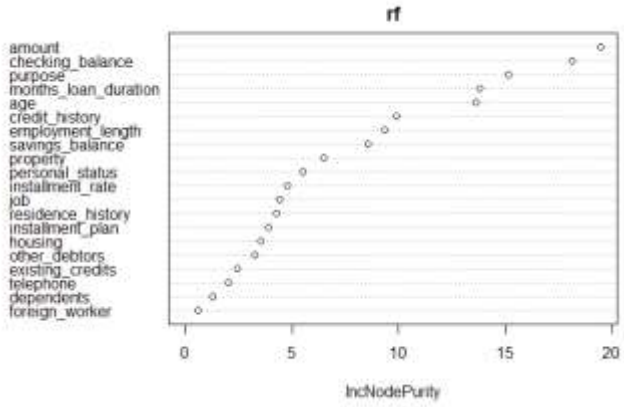
We then pruned the tree and calculated accuracy on the testing data at different thresholds.
We got the highest accuracy of 72.5% at threshold 0.5

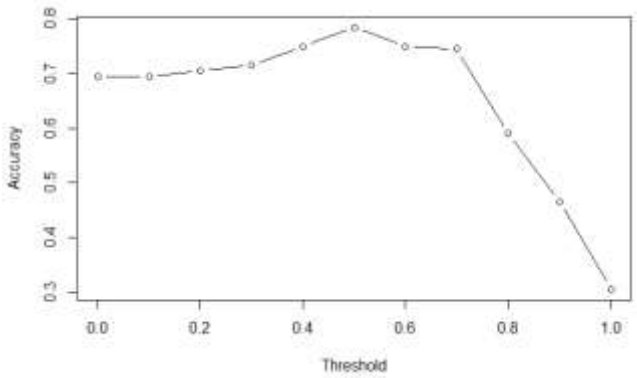


Sensitivity: 0.8345324
Specificity: 0.4754098
AUC: 0.655

Random Forest

While fitting the random forest, we plotted the node purity to check which are the attributes chosen by most decision trees within the forest to form the splits. We see amount, checking balance, purpose and months loan duration to be good candidates. Two of these (checking balance and months loan duration) were also observed in our previous decision tree. The candidate trees form the splits in different combinations of the attributes and the final classification is determined by polling.





We got the best accuracy of 78.5% when the threshold was 0.5

Sensitivity: 0.9496403
Specificity: 0.4098361
AUC: 0.6797

Discussion

Subject Matter Implications

Of all the models we tried, lasso regression performed the best, in terms of AUC and specificity. We could not see a notable difference in the overall accuracy among the various logistic regression models that we tried and logistic regression models in general outperformed the other models. Although we had expected decision trees to work well with interaction variables, however the overall accuracy might be low because of overfitting. Random forest corrects for this overfitting, and we get higher accuracy. SVM performance is possibly poor because we did not experiment with the misclassification parameter, implying there might be possible overfitting. Thus, it can be implied that the better performance of the lasso regression can accredited to the model optimizing the bias-variance tradeoff, and thus avoiding overfitting unlike the other models.

Limitations and Next Steps

There are several things that we could try in future studies to improve our classification. In this study, we only had 1000 data points which is not that many. We may get better results if we get a larger data set. When doing logistic regression, our data did not meet the assumptions needed to make inferences. Notably, we found a bimodal distribution in the model's residuals. If we find a dataset with more predictors, we may be able to find one that can explain the bimodality and allow us to make inferences.