Econometrics Final Project

Kulindu Cooray
kulindu.coorays@gmail.com

Abstract - Credit companies process numerous applications annually and must assess whether to approve or deny credit to each applicant. This paper investigates various methods and algorithms for classifying applicants as creditworthy or not. Using a credit approval dataset from the University of California, Irvine's Machine Learning Repository, several statistical models employing diverse techniques are applied. The study aims to identify accurate models and pinpoint the key factors influencing the likelihood of credit approval.

I. Introduction

Credit approval relies on evaluating an individual's financial history to determine their risk of default. Lenders gather applicant data to make informed decisions about whether to extend credit. This process requires statistical analysis and the development of predictive models to improve decision-making and ensure efficiency. By analyzing historical data, these models can provide accurate predictions about the likelihood of repayment. In this study, all statistical analyses will be conducted using R and RStudio.

II. Data

This paper uses the Credit Approval dataset from the University of California, Irvine Machine Learning Repository. The dataset originally includes 690 entries and 16 variables.

```
> head(data)
  V1    V2    V3 V4 V5 V6 V7   V8 V9 V10 V11 V12 V13   V14 V15 V16
1  b 30.83 0.000  u  g  w  v 1.25  t   t   1   f   g 00202   0   +
2  a 58.67 4.460  u  g  q  h 3.04  t   t   6   f   g 00043 560   +
3  a 24.50 0.500  u  g  q  h 1.50  t   f   0   f   g 00280 824   +
4  b 27.83 1.540  u  g  w  v 3.75  t   t   5   t   g 00100   3   +
5  b 20.17 5.625  u  g  w  v 1.71  t   f   0   f   s 00120   0   +
6  b 32.08 4.000  u  g  m  v 2.50  t   f   0   t   g 00360   0   +
```

However, it can't be analyzed as-is because of issues like multi class variables, differences in the scales of continuous variables, and missing data. Here's how these problems were addressed.

A. *Formatting*

To make the dataset easier to interpret, descriptive names were assigned to the original variable names. For example, V1 was converted into a binary gender variable, and V2 was converted into a age variable. The changes made are as follows:

Old Name	New Name	Description
V1	Gender	Male or Female
V2	Age	Age of applicant
V3	Debt	Debt held by applicant
V4	MartialStatus	Whether applicant is is married or not
V5	BankCustomer	Bank the applicant is a customer of
V6	EducationLevel	Education level of applicant
V7	Ethnicity	Ethnicity of applicant
V8	YearsEmployed	Number of years the applicant has been employed
V9	PriorDefault	Whether the applicant has previously defaulted before
V10	Employed	Whether the applicant is currently employed
V11	CreditScore	The applicant's credit score
V12	DriversLicense	Whether the applicant has a drivers license
V13	Citizen	Whether the applicant is a citizen
V14	ZipCode	The zipcode of the applicant
V15	Income	The year income of the applicant
V16	Approved	Whether or not the application was approved

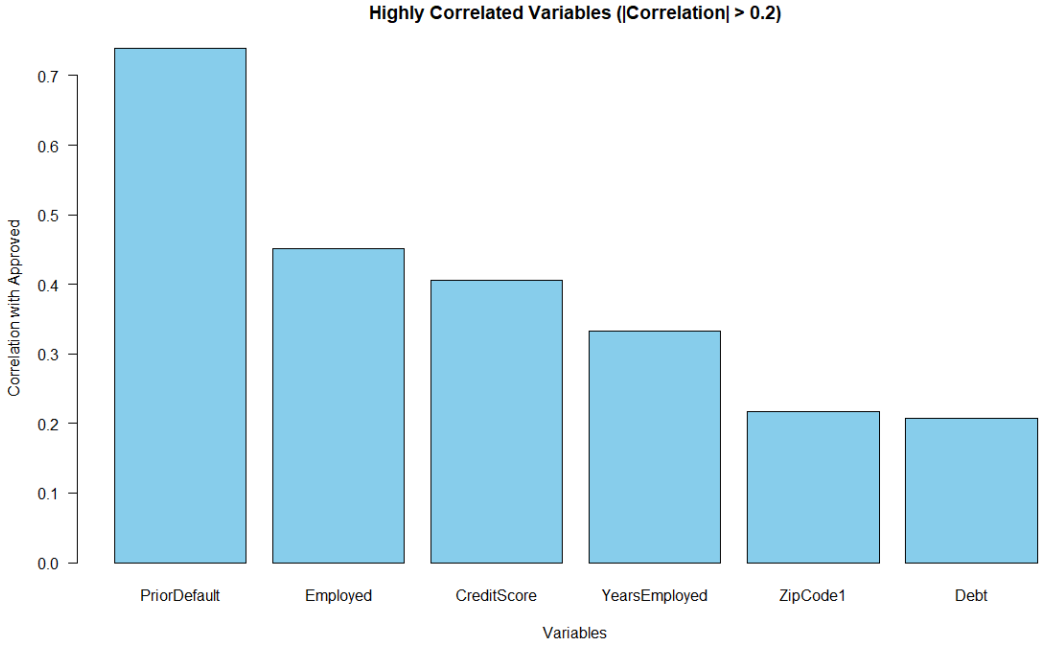
Next, discrete variables with multiple categories, such as education level, were split into separate binary variables to allow analysis in R. This process increased the total number of variables from 16 to 41.Missing data is a common challenge in large datasets, and the Credit Approval dataset contained 37 missing values. To address this, all observations with missing data were removed, reducing the dataset size to 653 complete entries.

III. Significant Variables

With the data preprocessed and ready for analysis, we can begin exploring the dataset. Among the 41 features available for prediction, not all will significantly contribute to determining whether an application is approved. Here, we outline methods to identify which predictors are most relevant. Additional techniques will be applied within specific models.

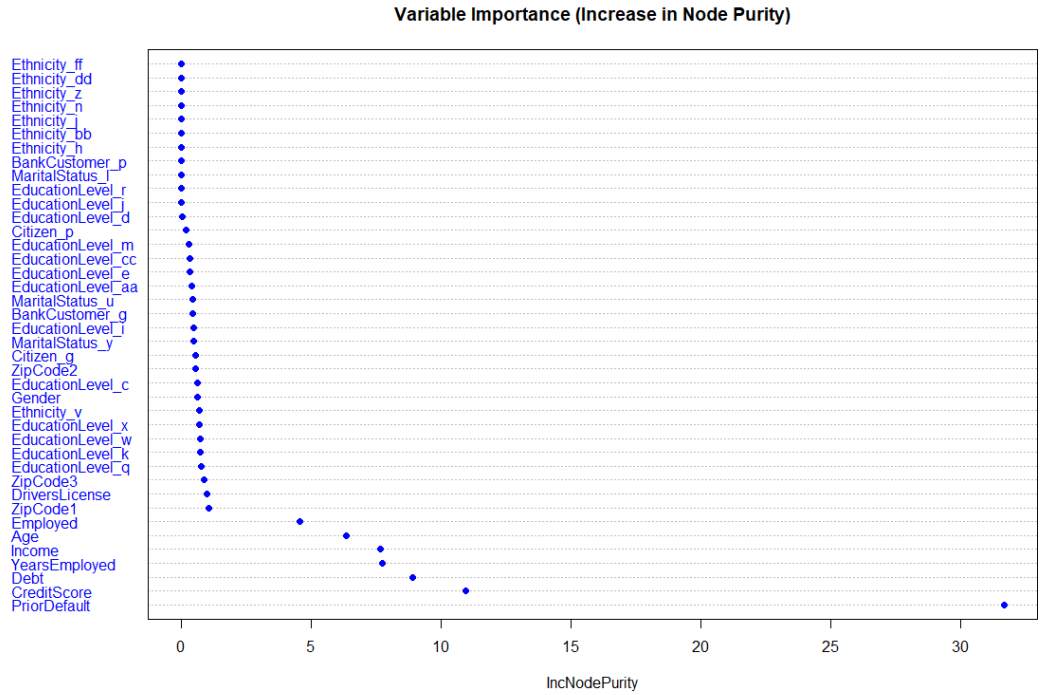
A. Correlated Variables

To start, we compute the correlation matrix for the dataset, focusing on the relationship between each variable and the Approved variable. Variables with a correlation greater than 0.2 are identified as the most relevant predictors. A bar chart visualizing these highly correlated variables is presented:



B. Random Forest

A random forest model offers another effective way to identify important predictors. By examining the increase in node purity, we can see which variables are most influential in determining whether an application is approved.



Higher Increasing Node Purity values indicate greater reductions in the Gini score, signifying higher relevance to the dependent variable. From the graph, it is evident that the most significant predictors are: PriorDefault, CreditScore, YearsEmployed, Income, Debt, Age, and Employed.

This finding aligns closely with the results from the correlation analysis, with the exception of ZipCode1. When comparing ZipCode1 to other Zip Code-related variables based on Increasing Node Purity, it becomes clear that location is not a significant factor in determining application approval. Therefore, it will be excluded from further analysis.

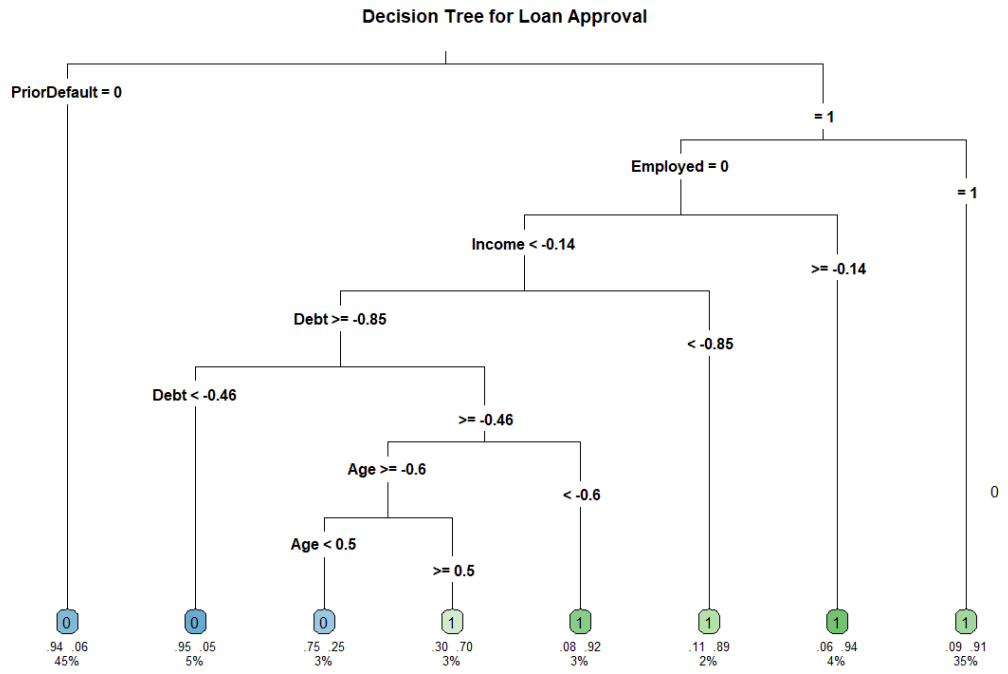
For the classification models developed in this paper, the following variables will be used as the Significant Predictors:

PriorDefault, CreditScore, YearsEmployed, Income, Debt, Age, Employed

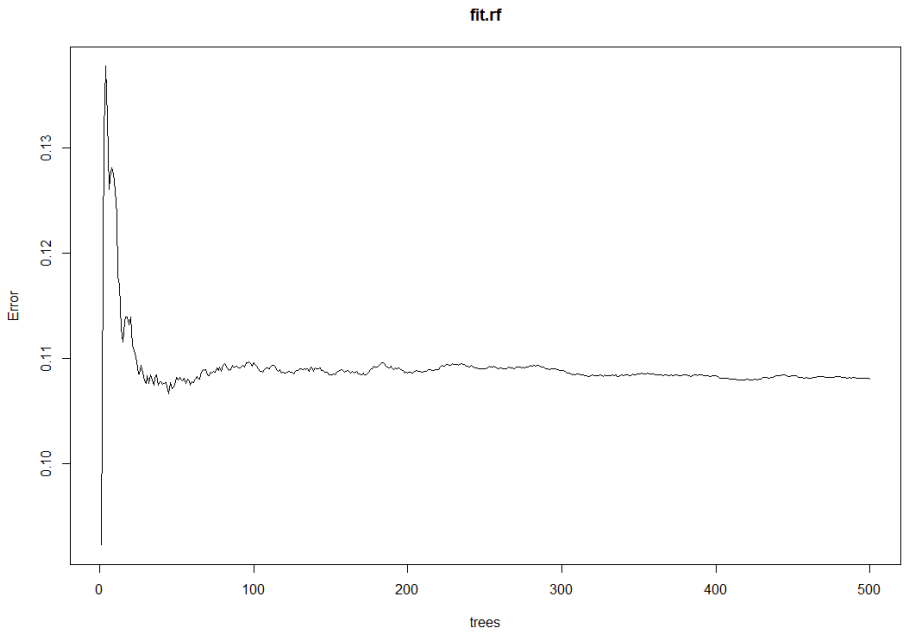
IV. Classification Models

A. Random Forest

Decision trees are a non-parametric method for classification. The tree splits the data by asking a series of yes/no questions. Each split evaluates a specific condition and the process goes further down the tree structure. Below is an example of a decision tree generated using the `rpart` library.



Single decision trees often show high variance, meaning their predictions can differ based on small changes in the data. One method to reduce this variance is to average the predictions of multiple trees. Instead of averaging an arbitrary number different trees, random forests improve this approach by using de-correlation. The predictors are chosen randomly for each split. This randomness helps ensure that individual trees are less similar to one another. Averaging the predictions in large numbers of random trees reduces test error and provides a more accurate model.

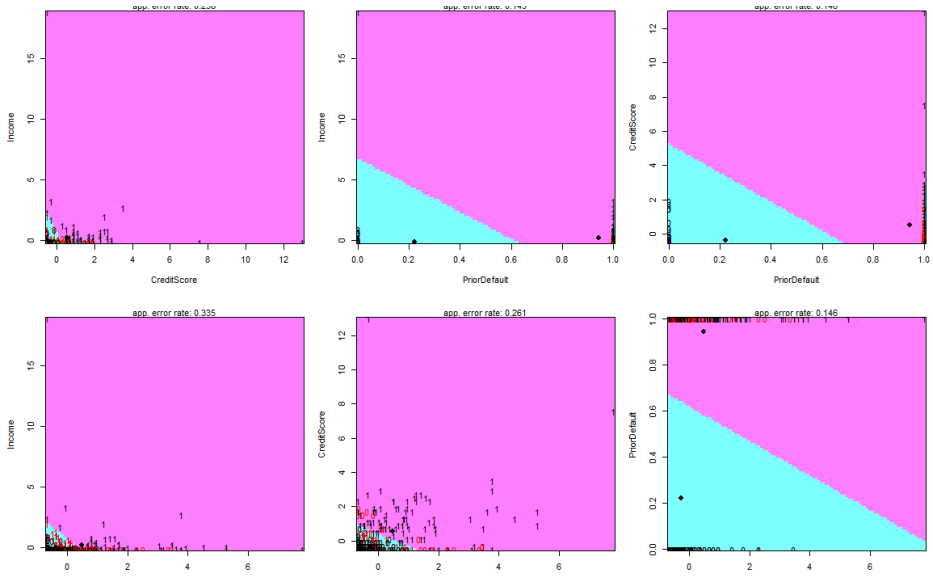


As shown above, increasing the number of trees significantly reduces the test error, particularly up to approximately 100 trees. Beyond this point, additional trees provide little to no improvement in performance. The random forest classification method achieves an accuracy of 0.8588.

B. Discriminant Analysis

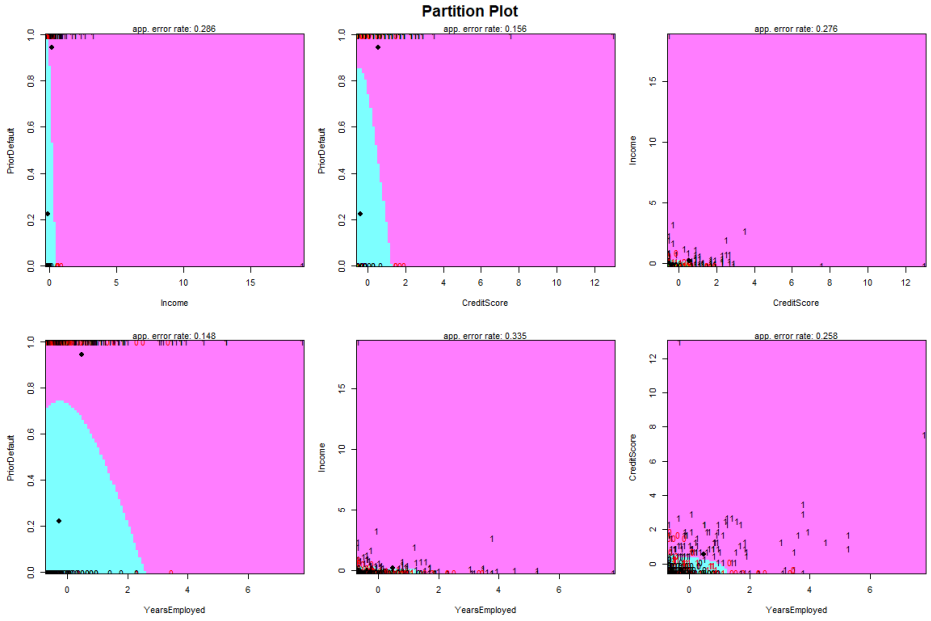
Next, linear and discriminant analysis methods are applied to classify the data by creating boundaries that separate the application statuses. For simplicity and consistency with prior model performance, the analysis will focus on the reduced subset of variables.

1. Linear:



Accuracy: 0.878

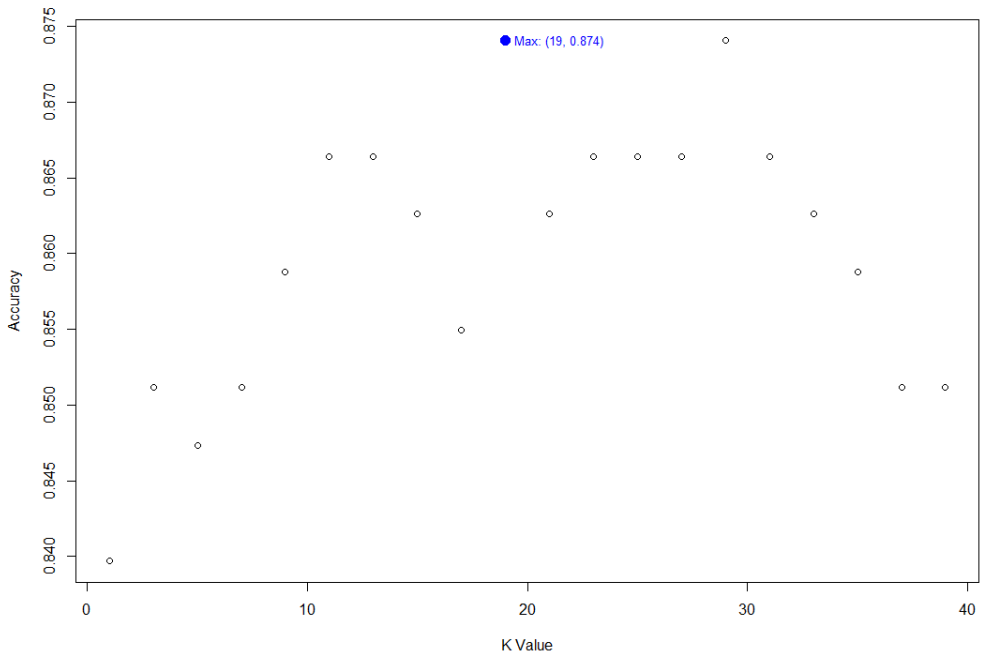
2. Quadratic:



Accuracy: 0.744

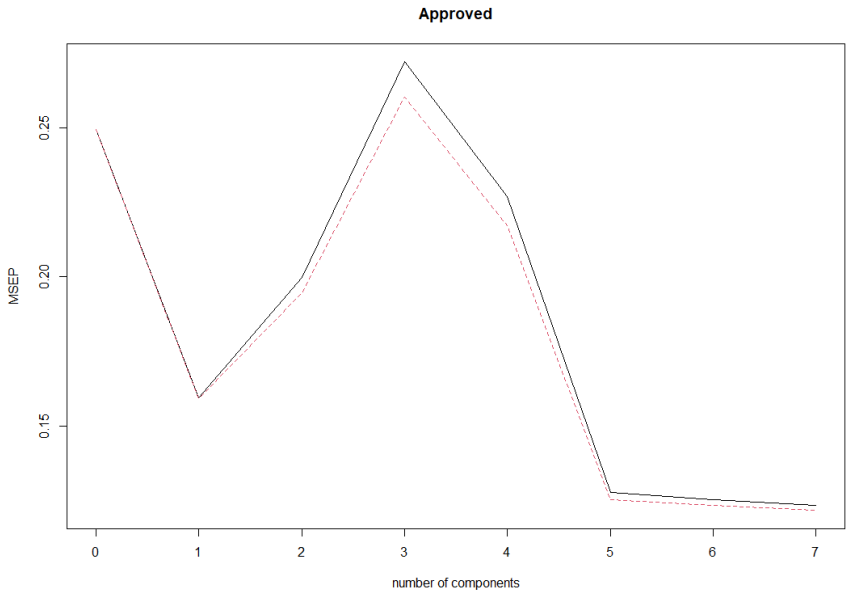
C. K-Near Neighbors (KNN)

Next we are using K nearest neighbors (KNN) method for classification. This algorithm works by calculating the distance to the k nearest data points and classifying an applicant based on the greatest magnitude among those points. The choice of K heavily influences the performance of the model, as it determines the number of neighbors considered in the decision.



The optimal value of K is 19 with an accuracy of 0.874

D. Principal Component Regression (PCR) Principal Components Regression simplifies a linear model using new linear components which are combined from the original predictors. The method decreases the number of dimensions in a model by selecting only the major components. Then, from these components, it goes back to mapping this with the original predictors. We can thus fit a PCR model with the optimal number of components that best represents the original data:



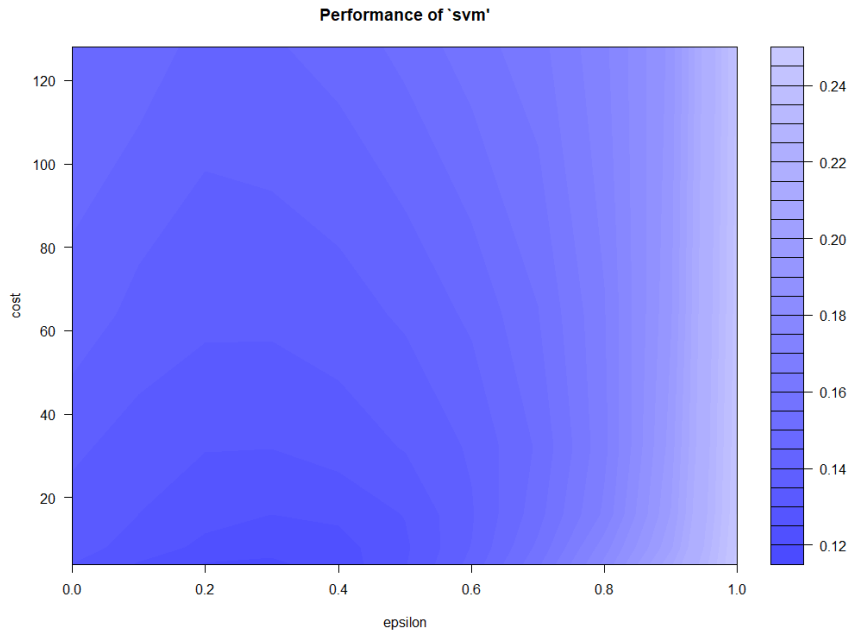
It is apparent that the MSE decreases most significantly when using five components. Further beyond this point, additional components show almost negligible performance improvement. The underlying coefficients for our feature set are the following:

Kernel	Accuracy
Linear	0.8778
Polynomial	0.8321
Radial	0.8740
Sigmoid	0.8664

The model has an accuracy of 0.8778

E. Support Vector Machine (SVM)

The SVM models work by finding a hyperplane in the feature space such that it separates it into distinct regions where data from different classes are placed. A model will find an optimum boundary to divide data samples. Here, I used four SVM models, using different kernels to define the shape of the decision boundary. Below is the tuning process for the linear kernel:

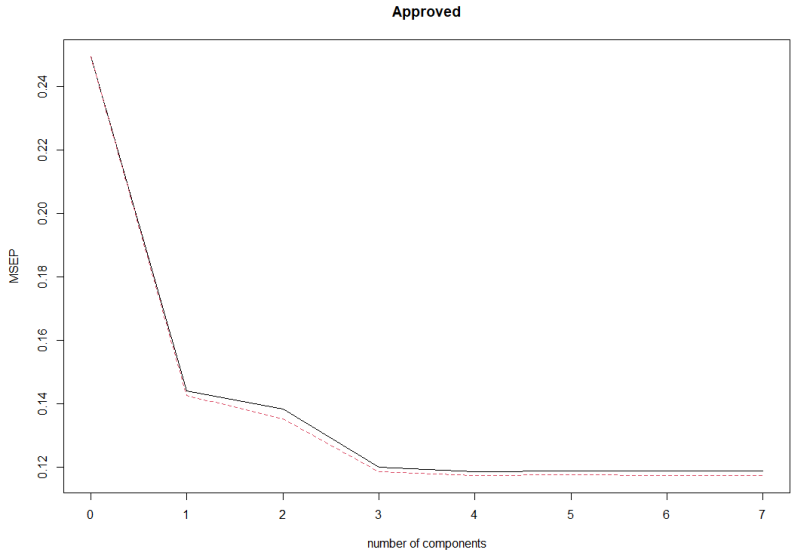


Dark blue values indicate higher accuracy. From the visualization, the optimal values for epsilon and cost to determine the boundary are identified as (0.3, 4).

The linear model has the best performance and barely outperforms the radial and sigmoid kernel functions.

F. Partial Least Squares (PLS)

Unlike Principal Component Regression (PCR), which selects principal components in an unsupervised manner, PLS incorporates the response variable when identifying components. The model assigns greater weight to predictors that are more strongly correlated with the response variable. The results of running the model are as follows:



We see that three components are sufficient to reach the optimal value of mean squared error. Also, PLS regression gives the same coefficients and has the same accuracy as PCR (0.8778), which means neither is better than the other in this analysis.

G. Linear

$$Y = X\beta + \varepsilon$$

For the linear probability model, I fit it using OLS. I will assess its performance in comparison with results from the subset of significant variables and the best subset as determined through the `regsubset()` function.

Linear Model Statistics

	Dependent variable:	
	Approved	
	SIG. VARS	BEST SUBSET
PriorDefault	0.612***	0.620***
	(0.038)	(0.037)
Income	0.063***	0.126***
	(0.016)	(0.045)
YearsEmployed	0.028	
	(0.018)	
ZipCode1		0.079**
		(0.038)
Citizen_p		-1.372
		(0.909)
MaritalStatus_l		
Ethnicity_ff		
Employed	0.141***	0.134***
	(0.041)	(0.040)
CreditScore	0.036*	0.040**
	(0.019)	(0.018)
Age	0.011	
	(0.019)	
Debt	0.004	
	(0.018)	
DriversLicense		-0.043
		(0.033)
Constant	0.052*	0.056*
	(0.029)	(0.033)
Accuracy	0.8778625954198470.877862595419847	
Observations	391	391
R^2	0.585	0.591
Adjusted R^2	0.578	0.583
Residual Std. Error (df = 383)	0.324	0.322
F Statistic (df = 7; 383)	77.182***	78.904***
Note:		*p<0.1;* ***p<0.05;*** p<0.01

The results show that a number of the variables highlighted in the previous model as being significant are insignificant in this model. However, the accuracies of the Significant Variables model and the Best Subset model are almost the same. Removing the insignificant variables from the first model yields an accuracy of 0.877, which is not a significant improvement. As a result, the Best Subset model is the best linear model for the task of predicting approval.

H. Logistic

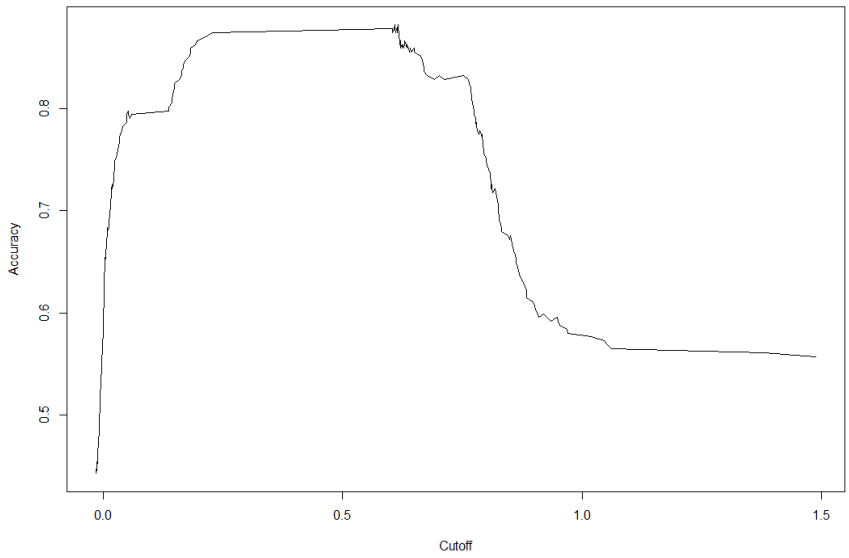
$$Y = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

In the logistic regression analysis, I assess the performance of three distinct models created using the `glm()` function. These models differ based on their selection of predictor variables, as outlined below:

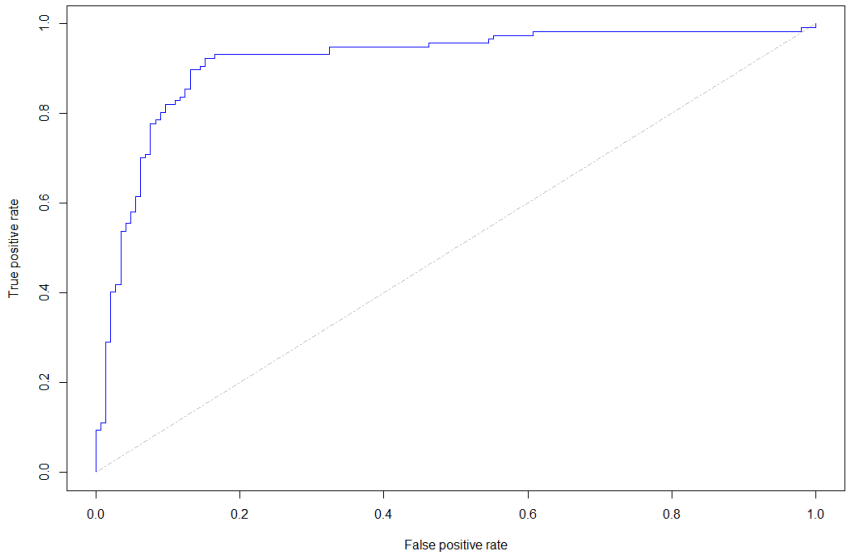
Logistic Model Results

	Dependent variable:		
	Approved		
	SIG. VARS	REDUCED VARS	PAPER VARS
PriorDefault	0.612***	0.632***	0.662***
	(0.038)	(0.036)	(0.036)
Income	0.063***	0.063***	0.062***
	(0.016)	(0.016)	(0.016)
YearsEmployed	0.028		
	(0.018)		
Employed	0.141***	0.140***	
	(0.041)	(0.041)	
CreditScore	0.036*	0.047***	0.076***
	(0.019)	(0.018)	(0.016)
Age	0.011		
	(0.019)		
Debt	0.004		
	(0.018)		
Constant	0.052*	0.043	0.088***
	(0.029)	(0.028)	(0.026)
Accuracy	0.8778625954198470.8778625954198470.877862595419847		
Observations	391	391	391
Log Likelihood	-111.368	-113.529	-119.440
Akaike Inf. Crit.	238.736	237.058	246.879
Note:			*p<0.1;* ***p<0.05;*** p<0.01

The first model includes the important variables from above, the second model eliminates all the variables that were determined to be completely unimportant, and the third model includes variables suggested by Deepesh and Khaneja [1]. To identify a potential threshold at which to determine an applicant to be approved or rejected, we compare the accuracy of the first model containing only the initially significant variables for different thresholds:



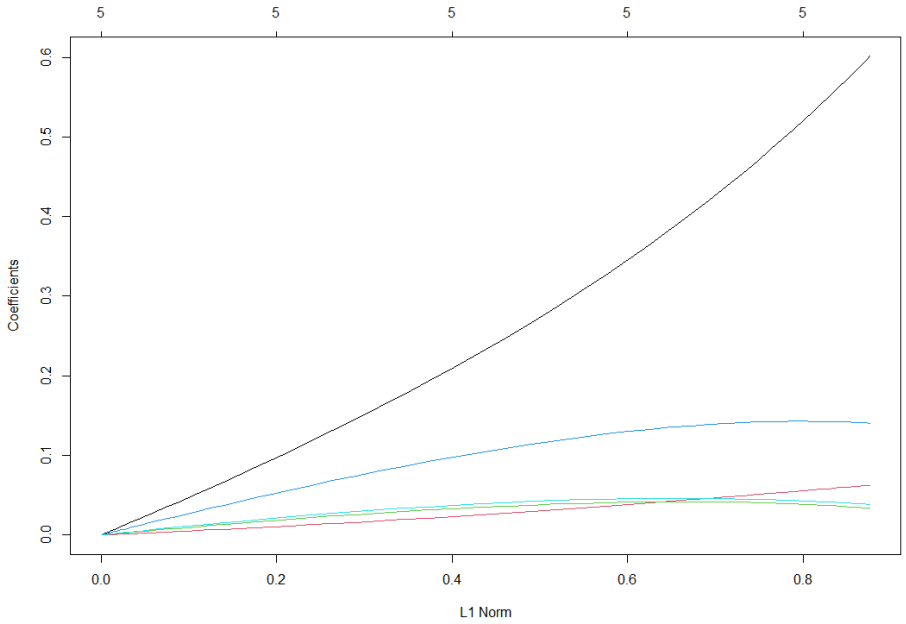
As can be seen, a cutoff value of around 0.5 yields the best accuracy. The three models all produce the same predictive accuracy of 87.79%. This is because, although the variables involved in the different models are different, the elimination of the insignificant variables does not have a strong enough impact to change the classification of the model for a cutoff value of 0.5.



The ROC curve of the regression model with significant variables is very good, having a 0.915 Area Under the Curve, suggesting that the model has great accuracy in differentiating between approved and rejected applications.

I. Ridge

Ridge regression shrinks variable coefficients in an effort to reduce mean squared error while addressing the strong relationship between predictors. The following are observations made by fitting a ridge regression model on the significant variable set:



The graph shows how the model determines the optimal coefficient values for each predictor. Among them, the **PriorDefault** variable stands out as the most significant (black line). The model has an accuracy of 0.8778.

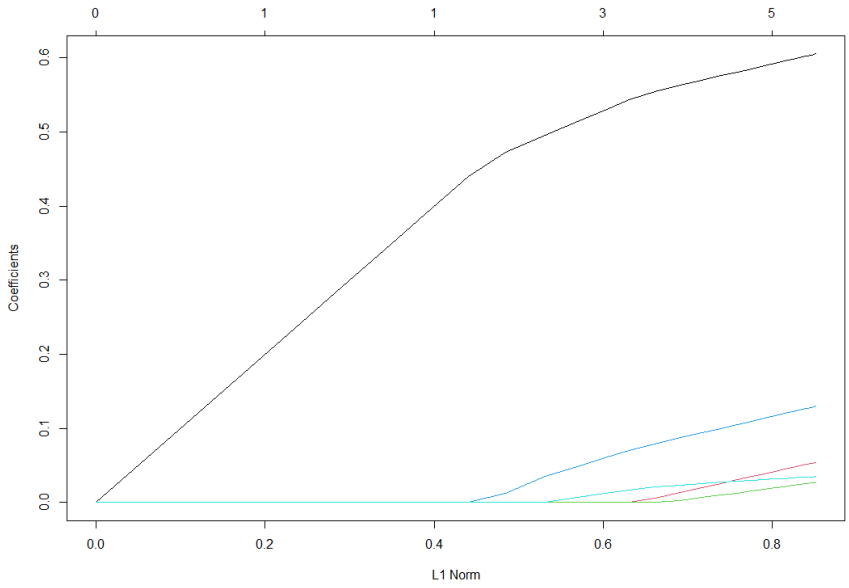
J. LASSO

The important limitation of the ridge is that it keeps all the variables in the model and does not eliminate any variables even if those variables have minimum or no importance. Opposite to this, LASSO is used to shrink some coefficients towards zero and eliminates the unimportant variables by making their coefficients exactly zero. By implementing LASSO regression to the dataset, we got the following coefficients which highlight the most important predictors while eliminating insignificant ones:

LASSO Regression
Coefficients

Variable	Coefficient
(Intercept)	0.08600
YearsEmployed	0.00727
PriorDefault	0.57906
Employed	0.10073
CreditScore	0.02744
Income	0.01628
Zip.Code1	0.01219
MaritalStatus_1	0.20701

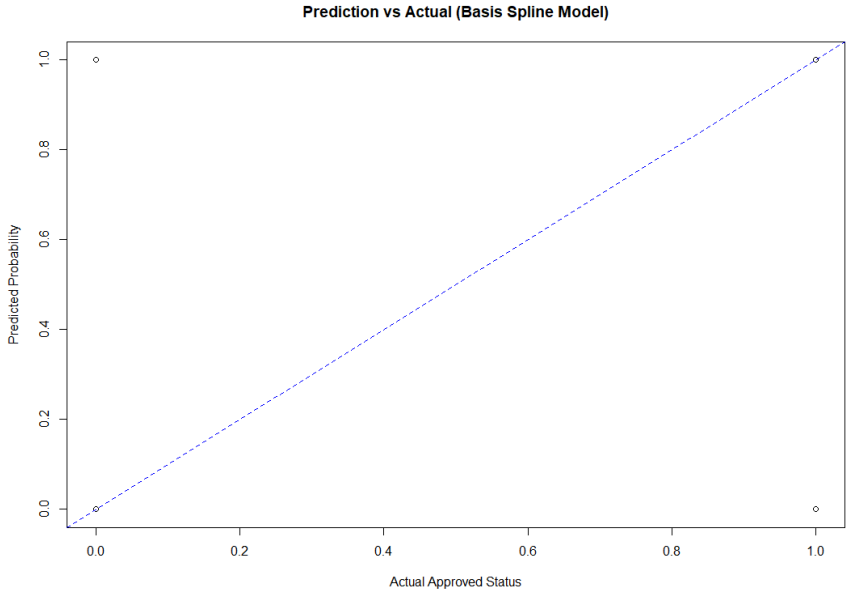
As can be seen, a cutoff value of around 0.5 yields the best accuracy. The three models all produce the same predictive accuracy of 87.79%. This is because, although the variables involved in the different models are different, the elimination of the insignificant variables does not have a strong enough impact to change the classification of the model for a cutoff value of 0.5.



The LASSO model finds the optimal values of coefficients, which minimize the cost function, and at the same time performs variable shrinkage and selection. The results include several variables that seem highly unlikely to bear a tight relationship with application approval, as was also the case when using best subset selection in the linear model. The accuracy yielded by the LASSO model is 0.8778, the same as from ridge regression.

K. Splines

Next, the basis spline is fitted to the data. With this approach, by using more than one function, different patterns emerge at different points within the data, which provides a flexible fit. Using the `bs()` function, the model is fitted to the significant variable set. The basis spline model results in an accuracy of 0.771.



The dashed blue line is the ideal that the predictions are equal to the actual values. Points falling on this line are perfect predictions; points off this line are prediction errors. The closer the points are to the line, the better the model performs because the predictions will be closer to the real observations. The larger scatter or deviations from the line indicate poor predictions of the model from the actual data.

V. Conclusion

Model	Accuracy
Linear	0.8854
Logistic	0.8778
LDA	0.8778
QDA	0.7442
KNN	0.8740
Ridge	0.8778
LASSO	0.8778
PCR	0.8778
PLS	0.8778
Splines	0.7709
RF	0.8587
SVM	0.8778

The table highlights that among all the models tested, the best subset linear model achieves the highest test accuracy.

The analysis indicates that PriorDefault is the most relevant variable in terms of acceptance or rejection of credit applications. Further, income and credit score are the other major variables affecting application outcomes. However, education, race, and marital status are not significant from a statistical perspective.

Many of the regression techniques returned very similar test performances, indicating that an 87% accuracy might be close to the best possible performance for this size dataset. More flexible models would likely continue to improve in their accuracy with more data points, allowing them to capture more nuanced patterns in the data.

VI. References

1. Deepesh Khaneja, "Credit Approval Analysis using R", Technical Report, November 2017
2. Ms.D.Jayanthi,"Credit Approval Data Analysis Using Classification and Regression Models", International Journal of Research and Analytical Reviews, vol5 (3), 2018
3. Fu, Z and Z, Liu, "Classifier Comparisons On Credit Approval Prediction", Final project of Course CS229 in 2014 at Stanford University

VII. Code Appendix

R file is available in the attached zip folder.