A Study On Statistical Models Of Credit Risk

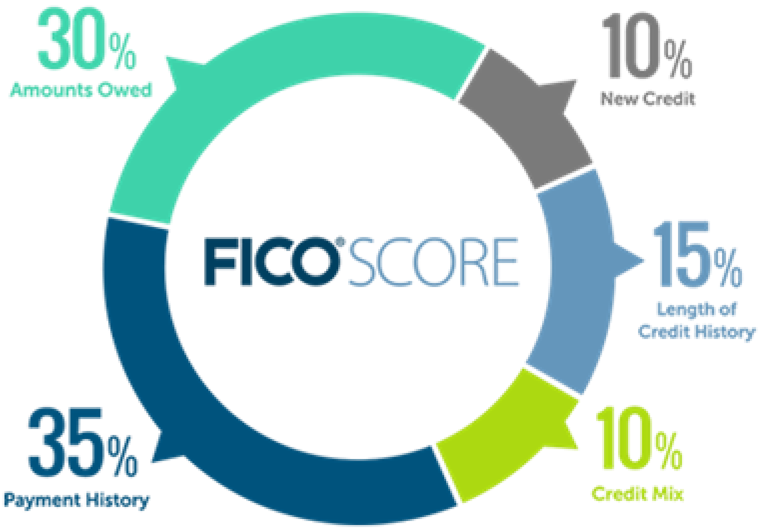
Ajay Pirabaharan, Alexander Watson, Jake Denton,
Jinghan XI, Liping Wen,
Mingyang LI

UNITED KINGDOM · CHINA · MALAYSIA

Introduction

Context

- A **credit score** is a metric which acts as a measure of an **individual's risk of defaulting (failing to repay) on a loan**.
- **Everyone over the age of 18** has a credit score, which not only determines whether or not a loan application is approved but also how much it will cost to borrow the money (interest).
- So, what **factors** are considered when a credit score is evaluated?



Motivation

- Organisations need to define the specific set of rules that **classifies a loan as "bad".**
- Basel Committee on Banking Supervision defines default essentially as a **delinquency stage of 90 days or more.**

Structure

- The variables are studied in order to get a preliminary understanding and then the **data is cleaned.**
- Models themselves are fitted and analysed, including **Logistic Regression, Basic Decision Trees** and **Random Forests.**
- **Compare** to existing models used in industry, how the models may be **appropriate** and how to **improve** the models.

Understanding the data

Variable	1	2	3	4	5	6	7	8	9	10	11
Min	0	0	0	0	0	0	0	0	0	0	0
1st Q	0	0.03	41	0	0.2	3400	5	0	0	0	0
Median	0	0.15	52	0	0.4	5400	8	0	1	0	0
Mean	0.06	6.05	52.3	0.42	353	6670	8.45	0.27	1.02	0.24	0.76
3rd Q	0	0.56	63	0	0.9	8249	11	0	2	0	1
Max	1	50708	109	98	329664	3008750	58	98	54	98	20
N/A's	0	0	0	0	0	29731	0	0	0	0	3924

Table 1: The Summary Data for all variables.

- Define the **response variable** as the number of times a person experienced 90 days past due delinquency or worse (1).
- Observe that the **mean** for this variable is 0.06, implying the dataset is **imbalanced**.
- **29731** observations which have at least one variable **missing** - belongs to monthly income (6) and the number of dependents (11).
- The missing data will be replaced by the **median values.**
- The median is a sensible prediction as it **isn't skewed** by unusually small/large data.

Cleaning the Data

The **Pearson correlation** between the variables is shown in the table.

All correlations **higher than 0.4** are explored in further detail.

Variables	1	2	3	4	5	6	7	8	9	10	11
1	1	0.24	-0.1	0.12	0.06	-0.02	-0.03	0.11	-0.02	0.1	0.04
2	0.24	1	-0.26	0.11	0.06	-0.03	-0.17	0.1	-0.08	0.09	0.08
3	-0.1	-0.26	1	-0.05	-0.08	0.03	0.18	-0.05	0.06	-0.04	-0.21
4	0.12	0.11	-0.05	1	-0.03	-0.01	-0.05	0.98	-0.03	-0.98	0
5	0.06	0.16	-0.08	-0.03	1	-0.05	0.35	-0.05	0.52	-0.05	0.1
6	-0.02	-0.03	0.03	-0.01	-0.05	1	0.09	-0.01	0.14	-0.01	0.06
7	-0.03	-0.17	0.18	-0.05	0.35	0.09	1	-0.07	0.42	-0.06	0.04
8	0.11	0.1	-0.05	0.98	-0.05	-0.01	-0.07	1	-0.04	0.99	-0.01
9	-0.02	-0.08	0.06	-0.03	0.52	0.14	0.42	-0.04	1	-0.04	0.13
10	0.1	0.09	-0.04	0.98	-0.05	-0.01	-0.06	0.99	-0.04	1	-0.01
11	0.04	0.08	-0.21	0	0.1	0.06	0.04	-0.01	0.13	-0.01	1

Table 2.3: The Pearson correlation between each of our variables.

- Observation of **Age being 0** is initially removed.
- **High correlation** between the variables representing the number of times someone pays a specified number of days past due (4,8,10).
- This is caused by **cluster of observations** with high values of 98 and **removed** to lower correlation.
- Values of the variable representing the Revolving Utilization above **1 are changed to 1.**
- This is due to restriction that the max can be 1 from the description of the variable.

Logistic Regression

1. Model Introduction

Logistic Function: $p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_{10} X_{10}}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_{10} X_{10}}}$

Logistic Regression Model: $\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{10} X_{10}$

The Likelihood Function: $l(\boldsymbol{\beta}) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=1} (1 - p(x_{i'}))$

Predictors: $X_1, X_2, \ldots, X_{10},$

Coefficients: $\beta_0, \beta_1, \ldots, \beta_{10},$

Response Variable: Y=0 or 1

The probability of default: $p(X) = p(Y = 1|X)$

Log-Odds: $\log\left(\frac{p(X)}{1-p(X)}\right)$

2. Checking the assumptions

- The observations should be independent of each other.
- There should be little or no multicollinearity between the independent variables.
- The independent variables should be linearly related to the log odds.
- A large sample size.

Our data satisfies these 4 assumptions.

3. The fitted Logistic Regression model

Intercept& predictors	Estimate	Odds ratio	z value
(Intercept)	-3.407	0.033	-50.698
RevolvingUtilizationOfUnsecuredLines	2.049	7.759	49.017
age	-0.018	0.982	-16.177
NumberOfTime30.59DaysPastDueNotWorse	0.426	1.531	31.809
DebtRatio	-3.184e-05	0.99997	-2.471
MonthlyIncome	-2.259e-05	0.99998	-6.185
NumberOfOpenCreditLinesAndLoans	0.031	1.032	10.108
NumberOfTimes90DaysLate	0.702	2.017	34.766
NumberRealEstateLoansOrLines	0.097	1.102	7.486
NumberOfTime60.89DaysPastDueNotWorse	0.598	1.818	21.782
NumberOfDependents	0.036	1.037	3.007

Preserving all the predictors, fit a model using **training data** (70% of the cleaned data), the summary findings are shown. The estimated coefficient > 0, or the odds ratio > 1: an **increase in the predictor** is associated with an **increase in the probability of default.**

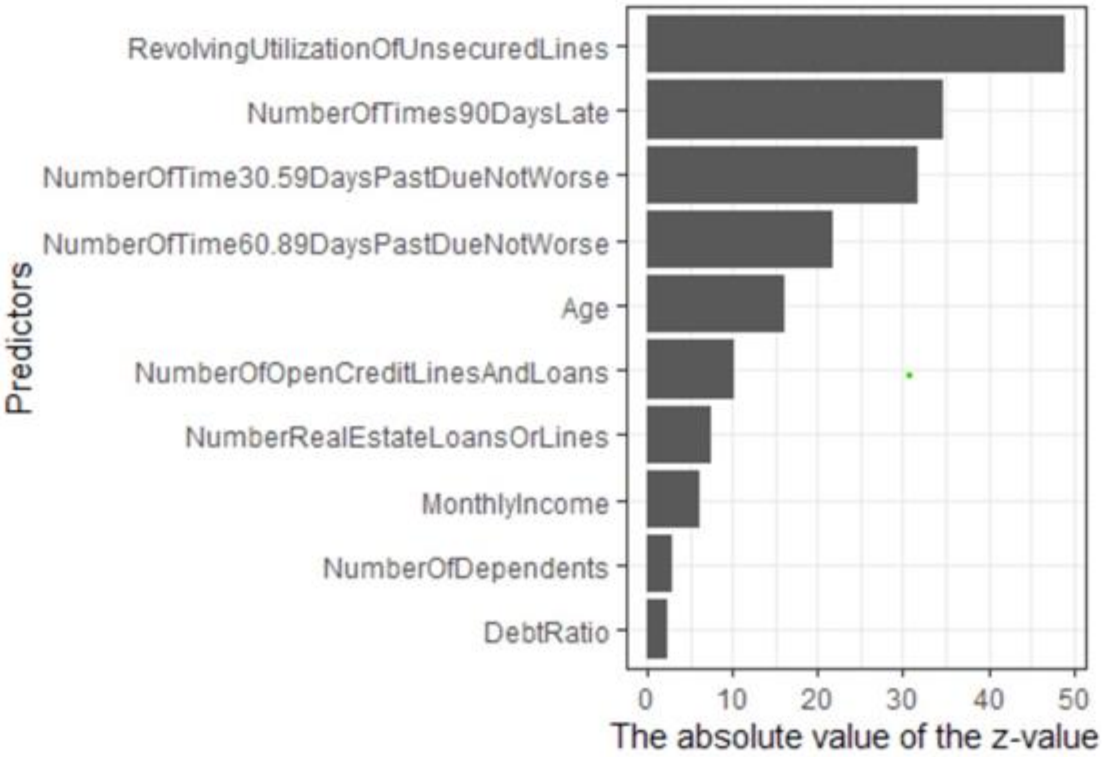
3. The fitted Logistic Regression model

According to the absolute value of the z statistic, the importance of the predictors are shown in the figure.

The top 4 important predictors are:

- RevolvingUtilizationOfUnsecuredLines,
- NumberOfTimes90DaysLate,
- NumberOfTime30.59DaysPastDueNotWorse,
- NumberOfTime60.89DaysPastDueNotWorse,

The last three of which all belong to the late payment history.



4. Predictive Ability

Use the test data (30% of the cleaned data) to **predict** the probability of default.

Probability $\xrightarrow{\text{Threshold}}$ Binary Variable → Confusion Matrix → ROC Curve → Threshold, Sensitivity, Specificity

Probability threshold : an observation is predicted as default, i.e. Y=1, if the probability exceeds the threshold.
Confusion matrix:

Confusion Matrix	True 0	True 1
Predicted 0	TN	FN
Predicted 1	FP	TP

Y=1 is called Positive, denoted by P; Y=0 is called Negative, denoted by N, Correct classifications on the diagonal.

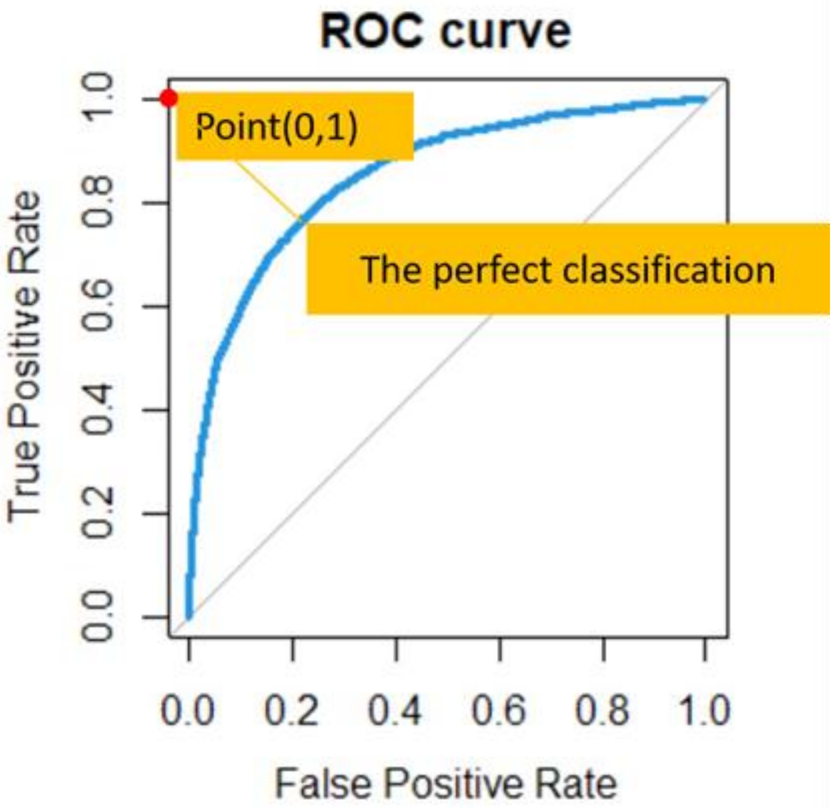
Two performance metrics :

- **Sensitivity (TPR)** = True Positive Rate: TP/(TP+FN)
- **1-specificity (FPR)**=False Positive Rate: FP/(FP+TN)

(Accuracy does not make much sense in this highly unbalanced data set.)

4. Predictive Ability

ROC curve: created by plotting TPR against FPR for various thresholds.



Select a threshold of 0.5:

	True 0	True 1
Predict 0	41624	2481
Predict 1	331	482

- **Sensitivity = 16.3% & Specificity = 99.2%**

Choose the point on the top left corner suggesting the threshold of 0.06336881:

	True 0	True 1
Predict 0	32578	683
Predict 1	9377	2280

- **Sensitivity = 76.9% & Specificity = 77.6%**

Since ROC curve does not deal with the different cost between the false negatives and false positives, **more analysis is needed**, if the specific weight of the cost was given.

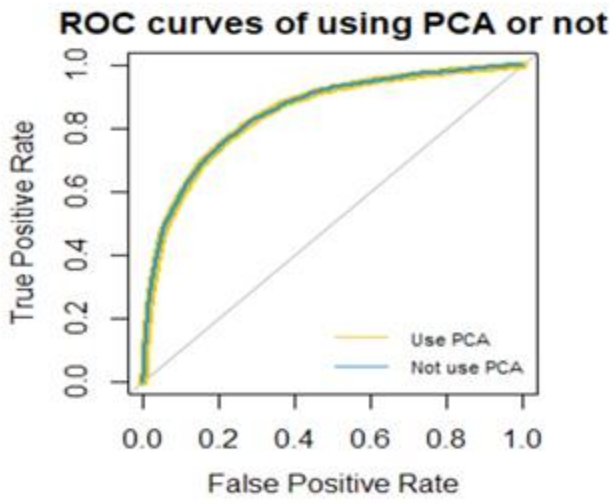
5. Methods tried to improve the model and compared with the fitted model

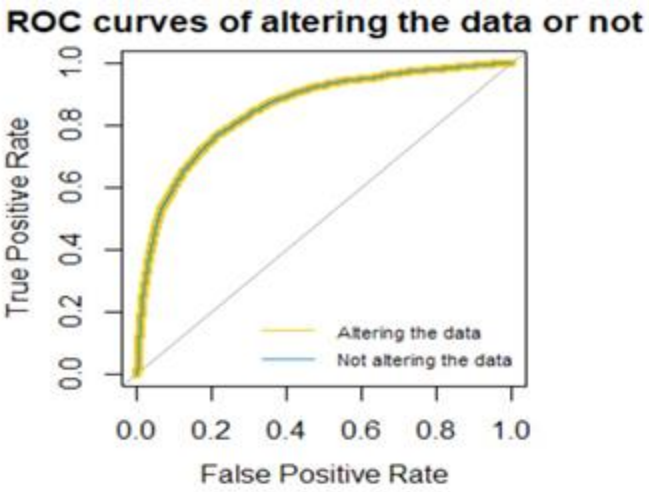
- **Principal Component Analysis (PCA)**
  After using the PCA, there are still 8 predictors, with the model becoming much harder to interpret.
- **Altering the training data set**
  Weight the 1s and 0s in the training data, such that n times as many 0s as 1s, n=1,2,…14, n=5 is chosen to be compared with the fitted model.

Comparison:

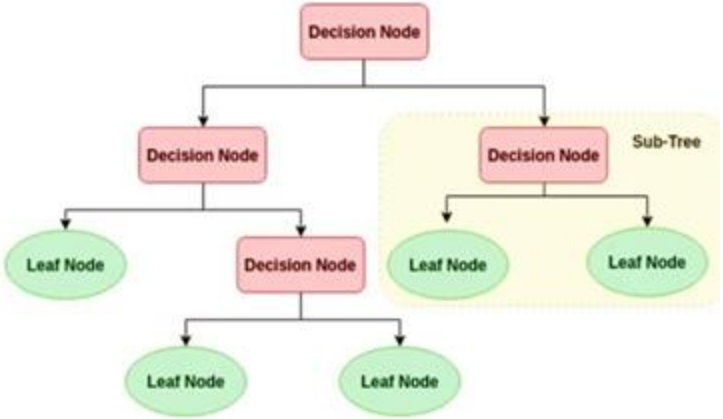
The ROC curves are almost the same, and the metrics are very close.

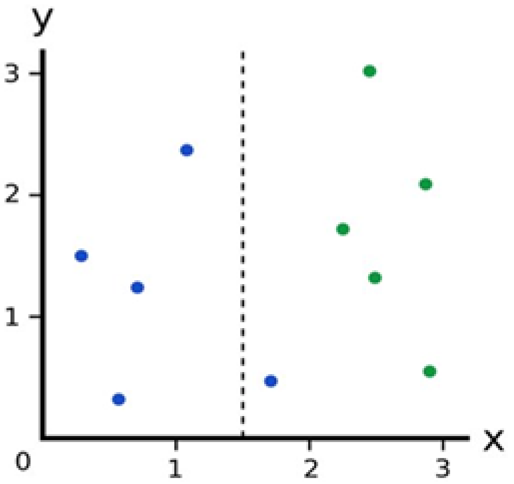
Models	Sensitivity	Specificity
The fitted model	76.9%	77.6%
PCA	76.2%	78.5%
Altered Data	78.5%	76.5%





Decision Trees and Gini Impurity



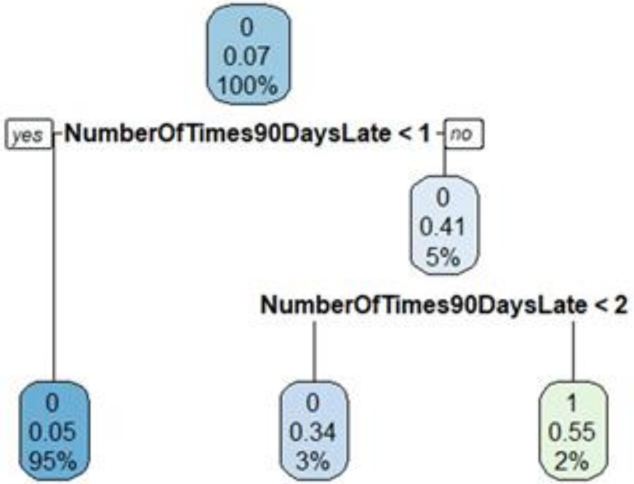


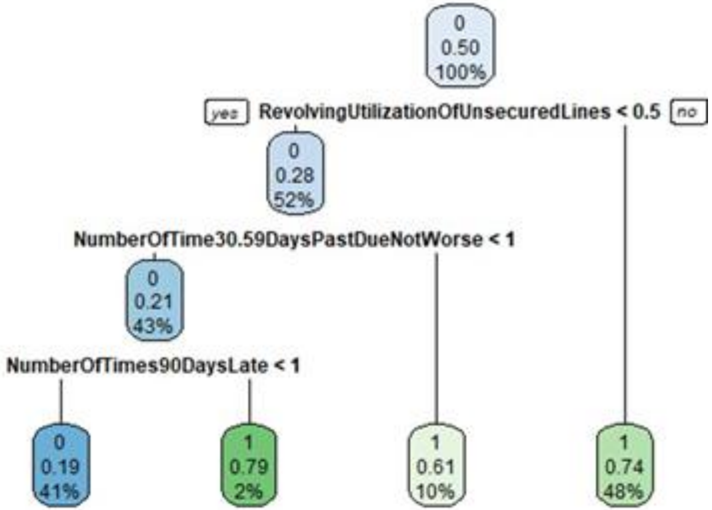
- **Root→Decision→Leaf**
- Data passed through tree based upon **conditions** at each decision node.
- At the leaf node, the **prediction** is given by the dominant class of the response variable in the group.
- **Aim**: Find a condition that splits the two classes of data into the purest subsets.

Need to maximise the greedy algorithm!

$$G = \sum_{i=1}^{C} p(i)\,(1 - p(i))$$

$$Gini\ Gain = G_{before} - \sum_{i=1}^{C} Proportion * G_{after}$$





Tree 1:

- Fitted using 60% representative sample
- Only one condition for the prediction
- Not accurate with 1s (low sensitivity 17%)
- Accurate with 0s (high specificity 99%)
- 5.6% offered loan have delinquency

	True 0	True 1
Predict 0	**27696**	1640
Predict 1	274	336

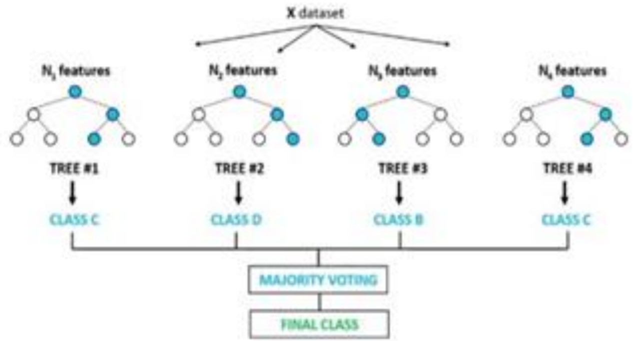
Tree 2:

- Fitted using balanced data (7500 0/1s)
- Up to three variables conditioned upon
- Improved sensitivity (85.2%)
- Reasonable specificity (67.6%)
- 1.5% offered loan have delinquency

	True 0	True 1
Predict 0	**18912**	291
Predict 1	9058	1685

Random Forest

Random Forest consists of a **multitude of decision trees** that operate as an ensemble, with the data passing through all the trees individually. The output is based upon the number of trees that predict value 1 or value 0.



Variables	Mean decrease in Gini coefficient
RevolvingUtilizationOfUnsecuredLines	968.03
Age	103.95
NumberOfTime30.59DaysPastDueNotWorse	508.37
DebtRatio	17.58
MonthlyIncome	21.03
NumberOfOpenCreditLinesAndLoans	44.09
NumberOfTimes90DaysLate	516.99
NumberRealEstateLoansOrLines	17.13
NumberOfTime60.89DaysPastDueNotWorse	231.29
NumberOfDependents	2.26

	True 0	True 1
Predict 0	19554	326
Predict 1	8416	1650

100 trees are used, each allowed up to 8 leaf nodes at first. (Model 1)

	True 0	True 1
Predict 0	19596	289
Predict 1	8347	1687

100 trees are used, each allowed up to 20 leaf nodes at first. (Model 2)

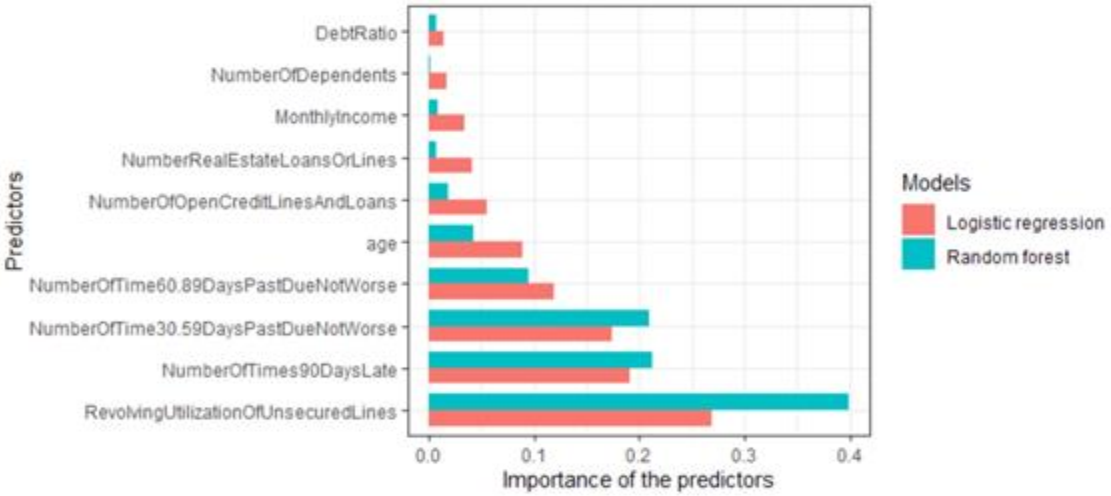
Advantages:

- Solving both classification and regression problems
- The model is also good at estimating missing data values

Disadvantages:

- A black box model

Comparing the Models



Model	Sensitivity %	Specificity %
Logistic	76.9	77.6
Decision Tree	85.3	67.6
Random Forest	85.4	70.1

Importance Of Variables

- Most important predictors are proportion of available credit used and payment history in all models
- Logistic regression predictions are influenced more by the other predictors

Other Considerations

- White-box/black-box?
- Probability or prediction?
- Number of variables utilised

Discussions

Model Issues:

- Imbalanced data… Combine variables?
- Variables only take **2 years** of payment history into account
- Alternatives for variables/other variables to consider?

How Do The Models Compare To FICO?

- Payment History (35%)
- Amounts Owed (30%)
- Other aspects…

Further Research

How to improve our models

- Combining the variables for the number of times someone is 90 days late with serious delinquency
- Increasing the number of 1s in the dataset
- Weighting 0 observations less than the 1s (using for example grid search)
- Cost function

Other popular models

- Neural Networks
- Support Vector Machine

Conclusion

Aim

- To build and analyse algorithms which are able to predict the probability that an individual will default on a loan.

Method and outcome

- Decision Tree (85.3% sensitivity and 67.6% specificity).
- Random Forest (85.4% sensitivity and 70.1% specificity).
- Logistic Regression (76.9% sensitivity and 77.6% specificity with a threshold of 0.06336881) .
- Logistic Regression gives more information regarding a single individual's risk and has flexibility in picking a balanced sensitivity and specificity.

The Most Important Variables

- The revolving utilisation variable.
- The variables which represent history of late payments.

How could models be improved

- Combining the variables for delinquency with 90 days late payments
- Adding variables
- Considering alternatives to variables like age