

Project: Creditworthiness

Step 1: Business and Data Understanding

Due to a financial scandal that hit a competitive bank, there is an influx of new people applying for loans for our bank instead of the other bank in the city. All of a sudden, we have nearly 500 loan applications to process this week! And we have to process all of these loan applications within one week and provide a list of creditworthy customers to the manager in the next two days.

Key Decisions:

Answer these questions

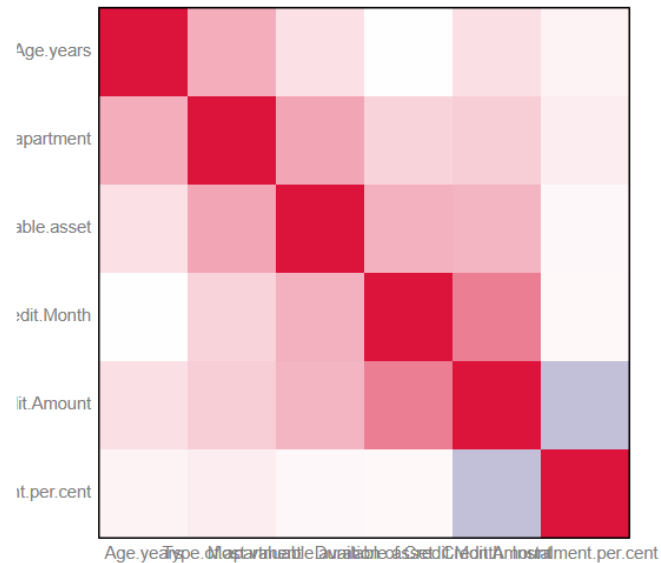
- The decision that needs to be made is whether a person is credit-worthy or not credit-worthy for our bank loan.
- The data that is needed for us to make such prediction is:
 - Data on all past applications (credit-data-training.xlsx - This file contains all credit approvals from the past loan applicants the bank has ever completed).
 - The list of customers that need to be processed in the next few days (customers-to-score.xlsx - This is the new set of customers that you need to score on the classification model you will create).

This data contains information about the clients such as (**Credit-Application-Result, Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Guarantors, Duration-in-Current-address, Most-valuable-available-asset, Age-years, Concurrent-Credits, Type-of-apartment, No-of-Credits-at-this-Bank, Occupation, No-of-dependents, Telephone, Foreign-Worker**)

- As the decision that we will be making is one of two choices, a person is either credit-worthy or not-credit-worthy. The model that we will be using to help us decide is a **binary** model.

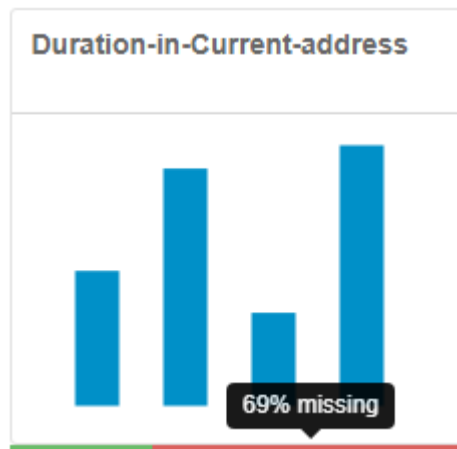
Step 2: Building the Training Set

- To check for correlation between numeric variables, I have used the spearman correlation matrix:



No high correlations to be found in the matrix.

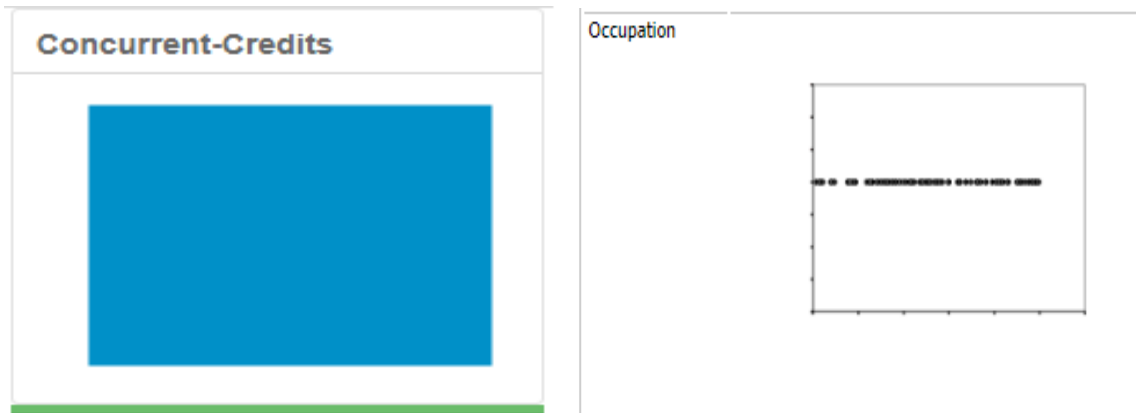
- There is only one data field in the data set that contains a lot of **missing data** and should be removed:



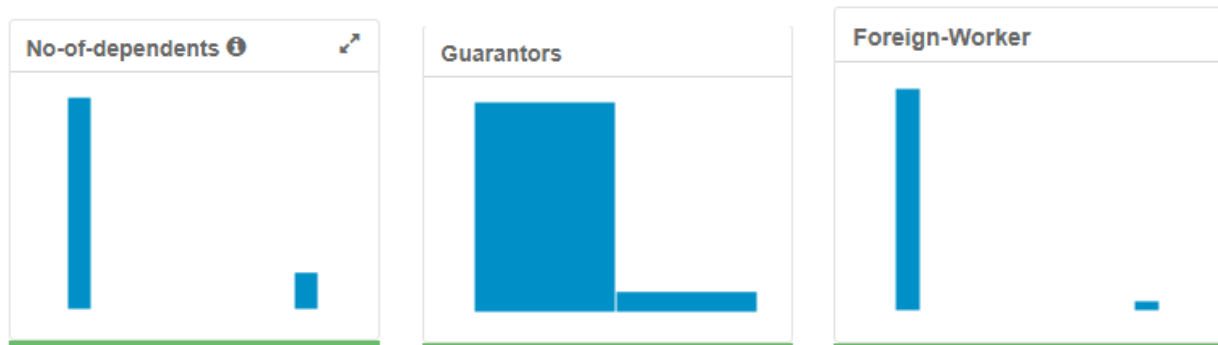
The "Duration_in_current_address" has **69%** of **missing data**, that is a very high number and this field should be removed.

- Other fields that should be removed from the data set because of other reasons such as low variability or the data is uniform.

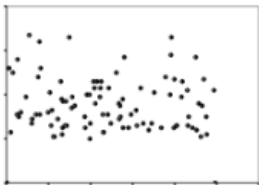
Both fields “concurrent_credits” and “occupation” should be removed because they are **uniform** and have only one value.



- The other fields such as “No_of_dependents”, “Guarantors” and “Foreign_worker” have **low variability** and should be totally removed.



- We will be also removing the field of “Telephone” because there is no logical reason for including the variable
- We have a few missing data in the “age_years” field, the missing data will be imputed with the **median** (33); I have chosen the median over the mean because the mean might sometimes skew to some very high or very low data points.

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502

We are left with **13** Columns in our data set; this cleaned data will be used to train models.
And the average value of “age_years” is **36** (rounded up).

Step 3: Train your Classification Models

We used four different models to be used to train on our data: Logistic Regression, Decision Tree, Forest Model and Boosted Model.

1. Logistic Regression

- For the logistic regression model, the significant predictor variables are (Account.Balance, Payment.Status.of.Previous.Credit, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent)
As we can see the P_values in the following table:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183 *
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566 **
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618 .
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296 **
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596 *
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549 *
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289 .

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom

Residual deviance: 328.55 on 338 degrees of freedom

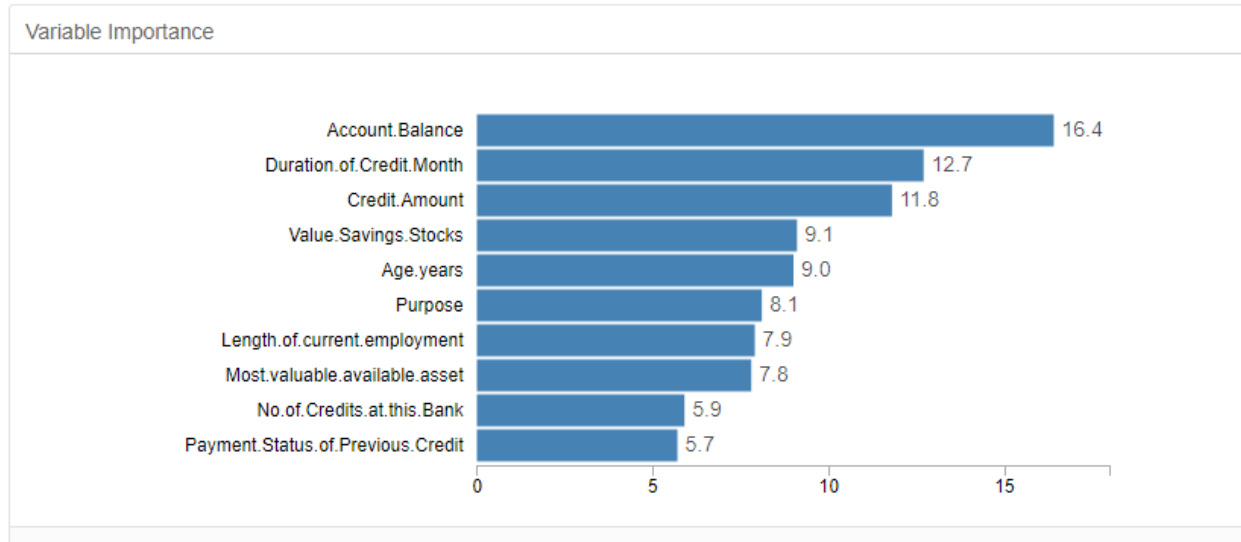
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

2. Decision Tree:

- For the Decision Tree model, the significant predictor variables are (Account.Balance, Credit.Amount, Duration_of_credit_month)

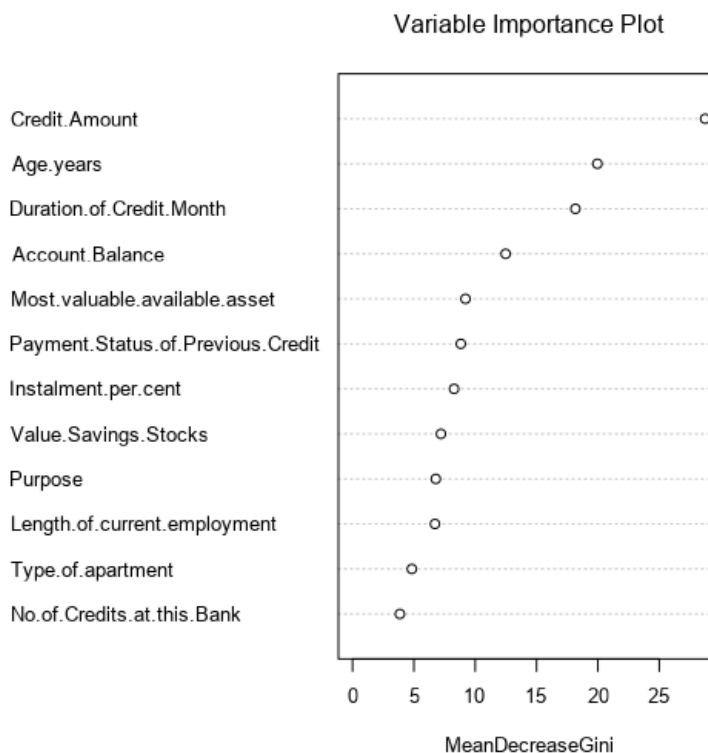
As we can see the following variable importance histogram:



3. Forest Model:

- For the Forest model, the significant predictor variables are (Account.Balance, Credit.Amount, Duration_of_credit_month, Age_years)

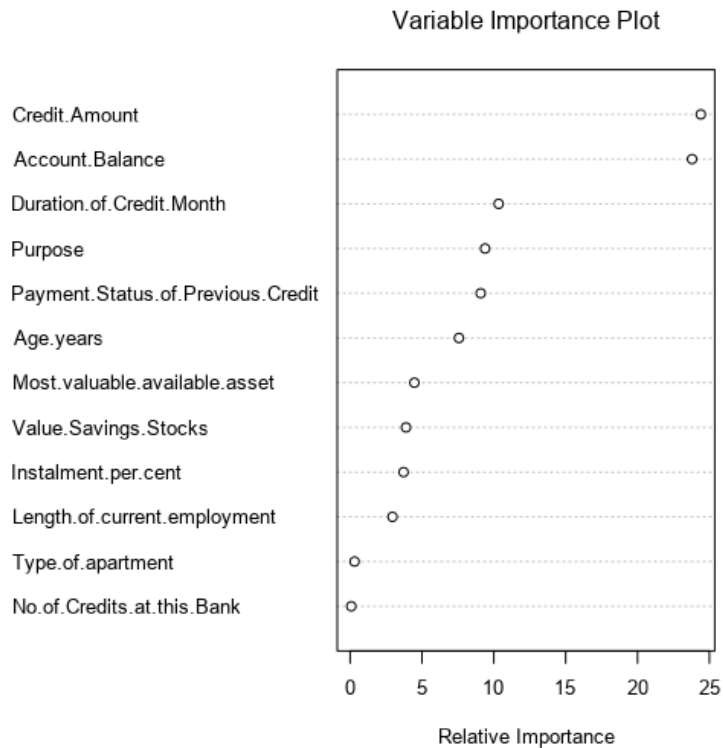
As we can see the following variable importance chart:



4. Boosted Model:

- For the Boosted model, the significant predictor variables are (Account.Balance, Credit.Amount, Duration_of_credit_month, purpose, Payment.status.of.previuos.credit, age_years)

As we can see the following variable importance chart:



Now we will see the overall accuracy of all four models with their confusion matrices:

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
decision_tree_model	0.6733	0.7721	0.6296	0.7905	0.4000
forest_model	0.7933	0.8681	0.7368	0.9714	0.3778
boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778
log_reg	0.7600	0.8364	0.7306	0.8762	0.4889

We can notice that the forest model has the highest accuracy.
And the decision tree model has the poorest accuracy.

Confusion matrix of boosted_model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of decision_tree_model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of forest_model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of log_reg		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

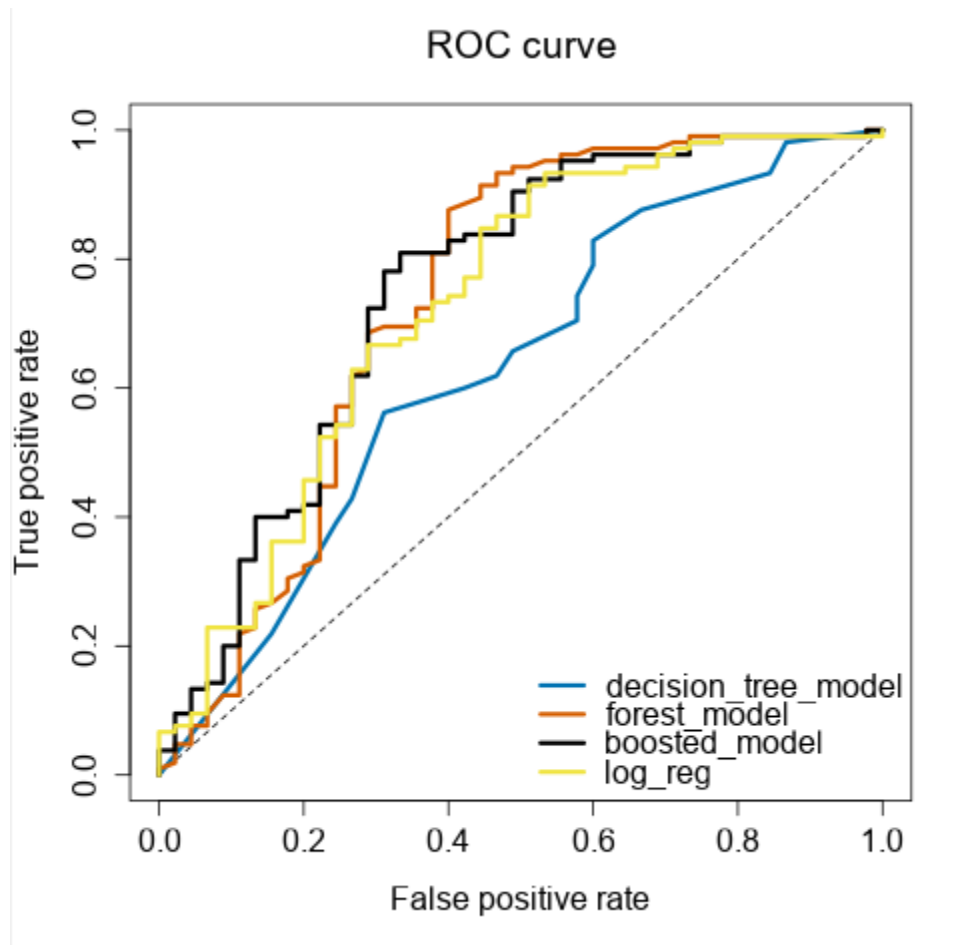
Looking at the confusion matrices, it seems that the models are biased towards predicting creditworthy. However, this is because our data set contains a lot more creditworthy people than Non-creditworthy people.

Step 4: Writeup

After removing some bad data, imputed other missing data, and then trained four different classification models on our data set. We came up with the decision of choosing the forest model because of its high accuracy and F1 score, the forest model has the highest accuracies when it comes to Accuracies within “Creditworthy” and “Non-Creditworthy” segments

Looking at the confusion Matrices above, we can also notice the boosted model and the forest model are less biased than the two other models.

Looking at the ROC curve below, we can see that both the forest model and the boosted model do a fine job.



- After using the scoring tool on our 500 new customers data, our model predicts **408** people to be creditworthy and 92 to be Non_creditworthy.

Record	Credit Worthiness	Count
1	Creditworthy	408
2	Non_Creditworthy	92