Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?

  Predict whether customers applying for loan are creditworthy to approve load to in a data-driven and automated matter using a classification model. This model should speed up the approval process as the number of applications becomes higher and provide a more consistency than manual subjective approach.

- What data is needed to inform those decisions?

  Historical customer loan approval data relevant to credit worthiness which includes loan specifics, customer demographics and whether load is approved or not.

  Loan specifics includes

  - current account balance (ability to repay)
  - length of loan
  - interest rate
  - previous payment status (ability to repay)
  - purpose of loan (hot tub vs car loan to drive to work)
  - size of loan
  - other accounts at bank (other assets to repay)

  Customer demographics includes

  - age (older customers have more risk of non-repayment)
  - guarantors available (loan may be recouped from guarantor if default)
  - length of employment (reliability of customer)

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

  Binary classification to determine whether or not to approve the load.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

Here are some guidelines to help guide your data cleanup:

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called "low variability" and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Note: *For students using software other than Alteryx, please format each variable as:*

Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double

Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

  The statistic of each feature is described in the tablet below and the distribution of each feature is also plotted to visual any missing, constant and unbalance.

  For *Age-years*, there's 2% is missing. Given the small percentage and importance of this feature, best to impute the missing values with median since it's the most common value. Averages might skew the age up or down with just one extreme customer age.

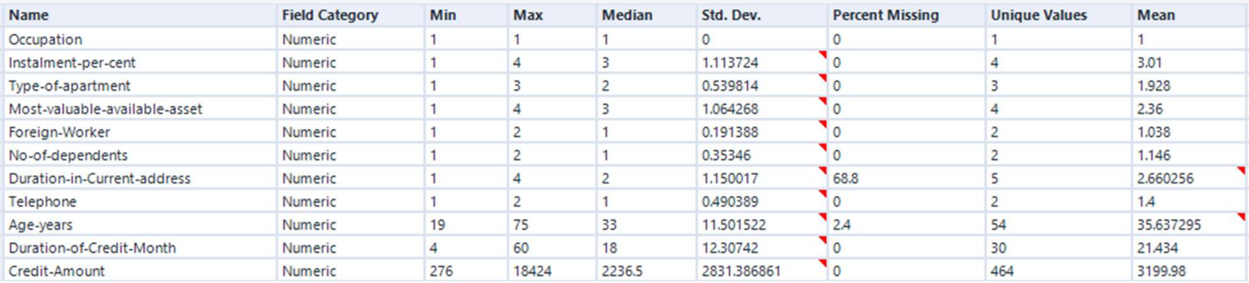
  Occupation and *Concurrent-Credits* feature is constant, same value for all customers in dataset, so do not contribute to predictiveness of model. As such, will be removed.

  Duration-in-Current-address has 69% missing values. Given such a high percent, this feature should be removed.

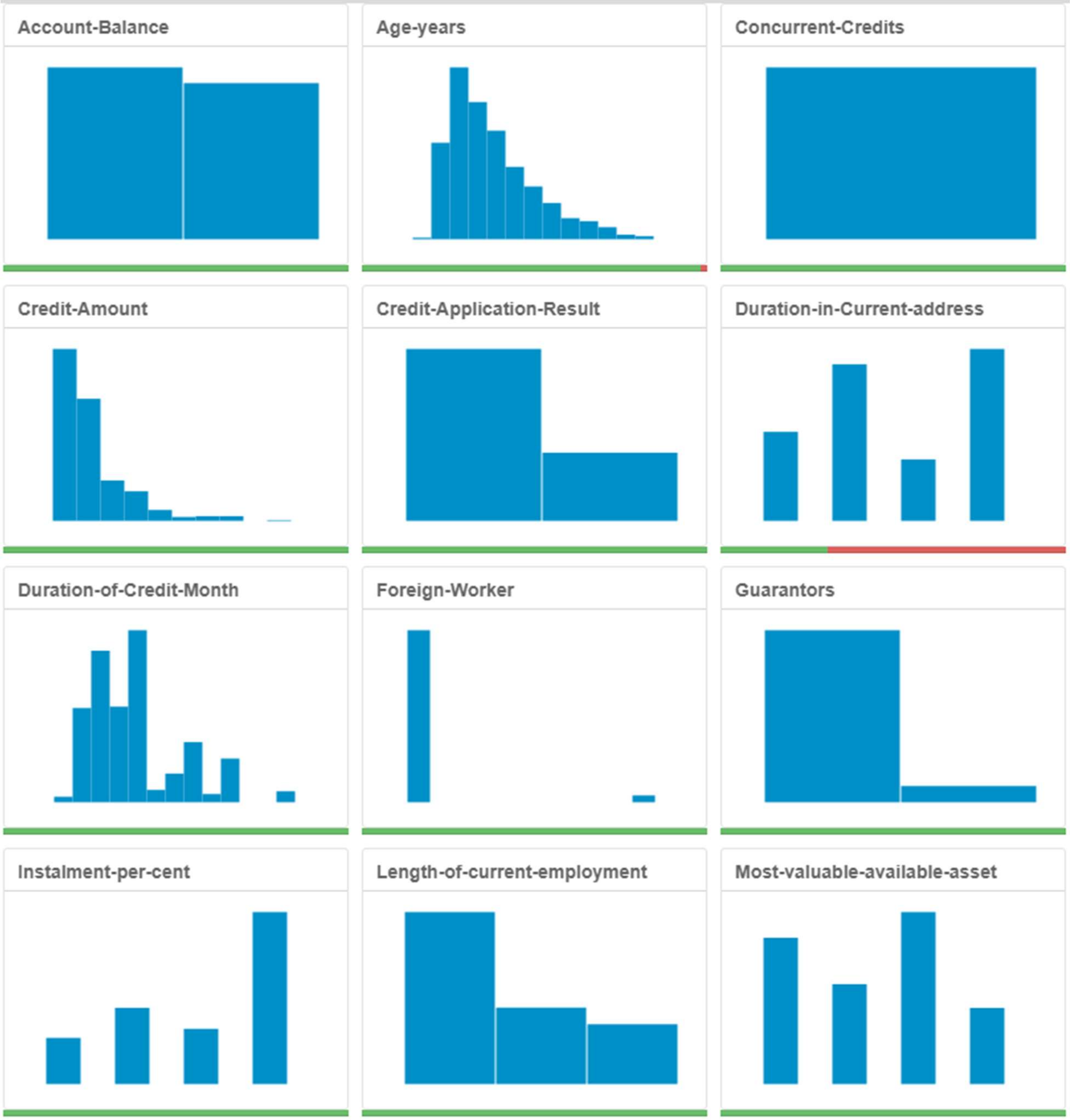
  Guarantors, No-of-dependents and *Foreign-Worker* are removed due to the highly skewed preponderance of one class over the other.

  Features that are not relevant to load application, *Telephone,* is removed.

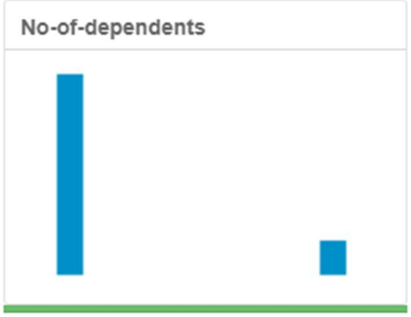
  In total, 7 features are removed, leaving 12 features to predict for the target, *Credit-Application-Result*.

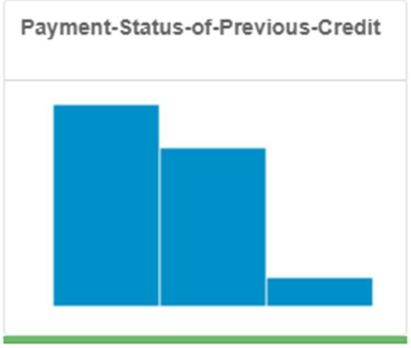


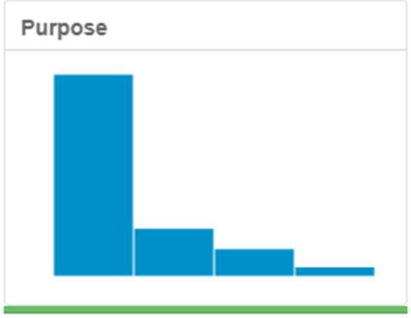
Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
Occupation	Numeric	1	1	1	0	0	1	1
Instalment-per-cent	Numeric	1	4	3	1.113724	0	4	3.01
Type-of-apartment	Numeric	1	3	2	0.539814	0	3	1.928
Most-valuable-available-asset	Numeric	1	4	3	1.064268	0	4	2.36
Foreign-Worker	Numeric	1	2	1	0.191388	0	2	1.038
No-of-dependents	Numeric	1	2	1	0.35346	0	2	1.146
Duration-in-Current-address	Numeric	1	4	2	1.150017	68.8	5	2.660256
Telephone	Numeric	1	2	1	0.490389	0	2	1.4
Age-years	Numeric	19	75	33	11.501522	2.4	54	35.637295
Duration-of-Credit-Month	Numeric	4	60	18	12.30742	0	30	21.434
Credit-Amount	Numeric	276	18424	2236.5	2831.386861	0	464	3199.98

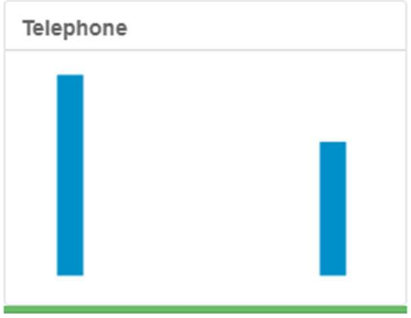


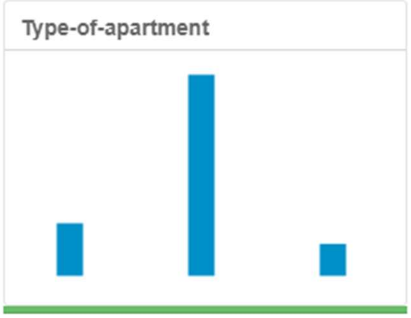
No-of-Credits-at-this-Bank











Value-Savings-Stocks

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*Answer these questions for **each model** you created:*

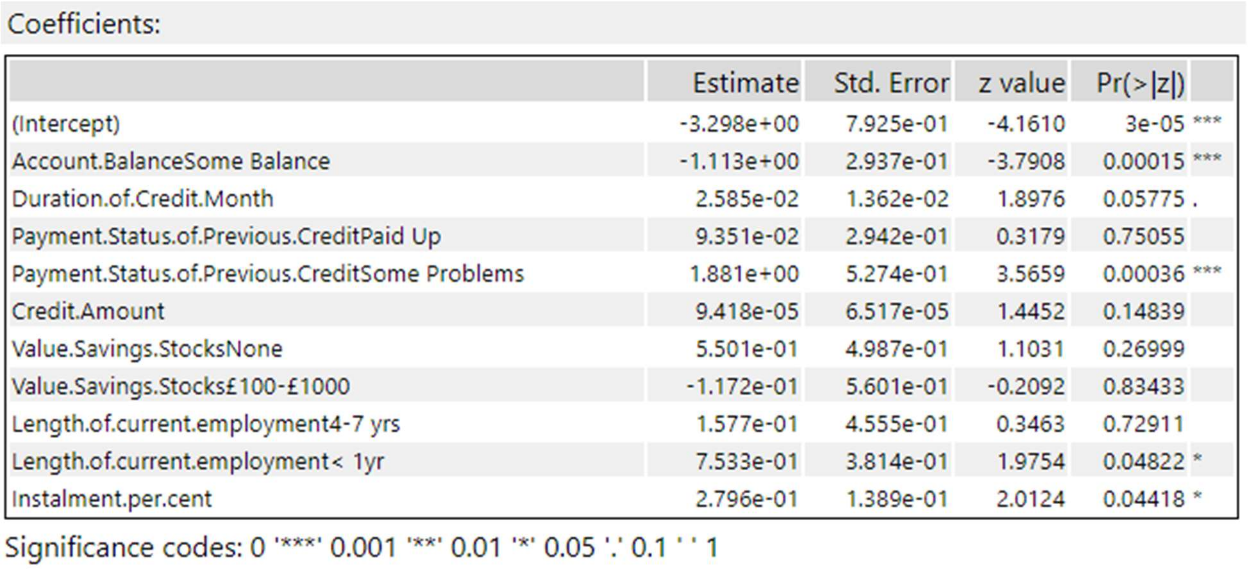
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

You should have four sets of questions answered. (500 word limit)

Logistic Regression:

Features of significance are *Account.Balance, Payment.Status, Length.of.current.employment,* and *Installment.per.cent* since their p values less than 0.05 as indicated by the asterisk code.



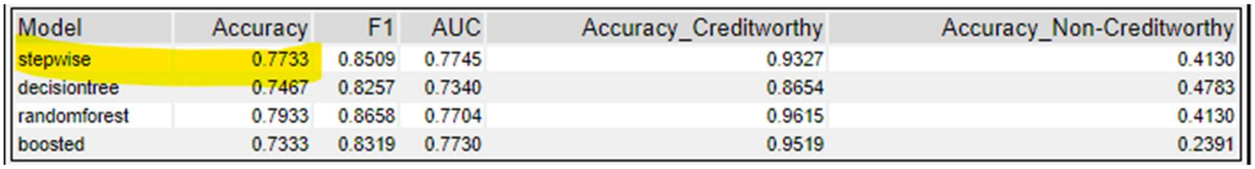
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.298e+00	7.925e-01	-4.1610	3e-05	***
Account.BalanceSome Balance	-1.113e+00	2.937e-01	-3.7908	0.00015	***
Duration.of.Credit.Month	2.585e-02	1.362e-02	1.8976	0.05775	.
Payment.Status.of.Previous.CreditPaid Up	9.351e-02	2.942e-01	0.3179	0.75055	
Payment.Status.of.Previous.CreditSome Problems	1.881e+00	5.274e-01	3.5659	0.00036	***
Credit.Amount	9.418e-05	6.517e-05	1.4452	0.14839	
Value.Savings.StocksNone	5.501e-01	4.987e-01	1.1031	0.26999	
Value.Savings.Stocks£100-£1000	-1.172e-01	5.601e-01	-0.2092	0.83433	
Length.of.current.employment4-7 yrs	1.577e-01	4.555e-01	0.3463	0.72911	
Length.of.current.employment< 1yr	7.533e-01	3.814e-01	1.9754	0.04822	*
Instalment.per.cent	2.796e-01	1.389e-01	2.0124	0.04418	*

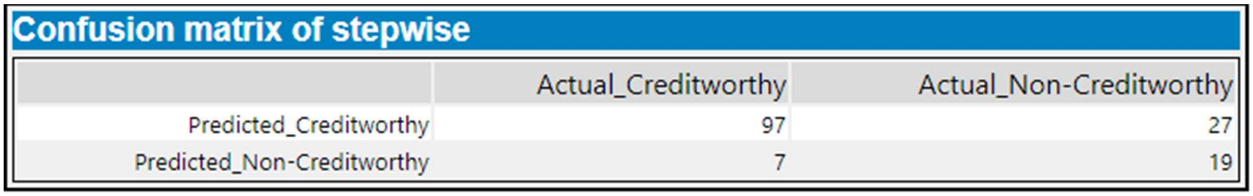
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

With the validation dataset, the logistic regression model (stepwise) has overall accuracy of 77.33%. The confusion matrix shows good performance for predicting creditworthiness (97 correct vs 7 wrong) but for non-creditworthiness there are more errors (19 correct vs 27 wrong).

Relative similarity between Precision and NPV indicates model has low bias for either classes.



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stepwise	0.7733	0.8509	0.7745	0.9327	0.4130
decisiontree	0.7467	0.8257	0.7340	0.8654	0.4783
randomforest	0.7933	0.8658	0.7704	0.9615	0.4130
boosted	0.7333	0.8319	0.7730	0.9519	0.2391



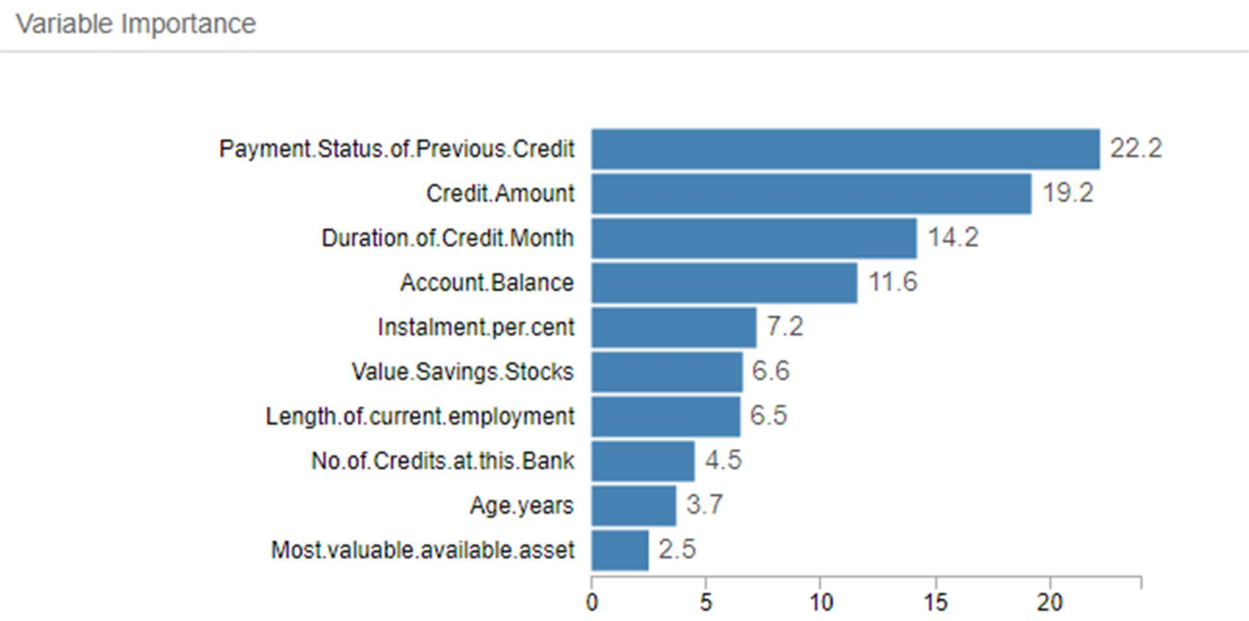
Confusion matrix of stepwise

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	97	27
Predicted_Non-Creditworthy	7	19

Precision	78.2%
Negative Predictive Value (NPV)	73.1%

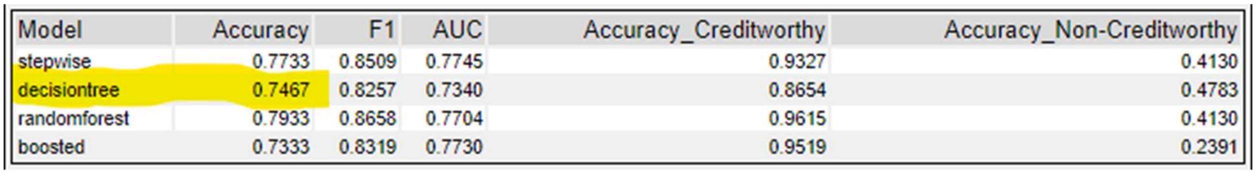
Decision Tree:

Top 4 features of importance are *Payment.Status, Credit.Amount, Duration.of.Credit.Month,* and *Account.Balance*.

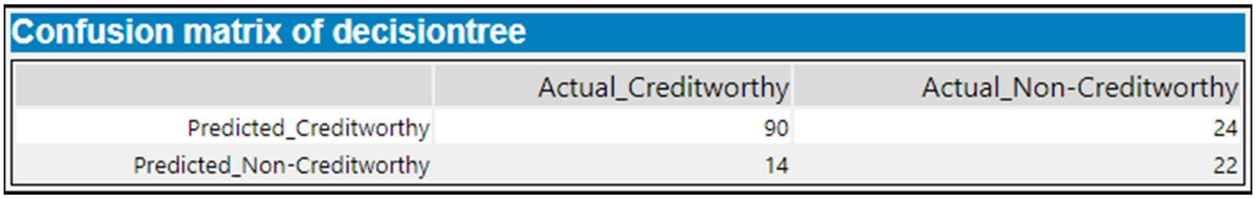


With the validation dataset, overall accuracy of decision tree model is 74.67%. The confusion matrix shows decent performance for predicting creditworthiness (90 correct vs 14 wrong) but poorly for non-creditworthiness there are almost equal errors (22 correct vs 24 wrong).

Large difference between Precision and NPV indicates model has higher bias for creditworthiness.



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stepwise	0.7733	0.8509	0.7745	0.9327	0.4130
decisiontree	0.7467	0.8257	0.7340	0.8654	0.4783
randomforest	0.7933	0.8658	0.7704	0.9615	0.4130
boosted	0.7333	0.8319	0.7730	0.9519	0.2391

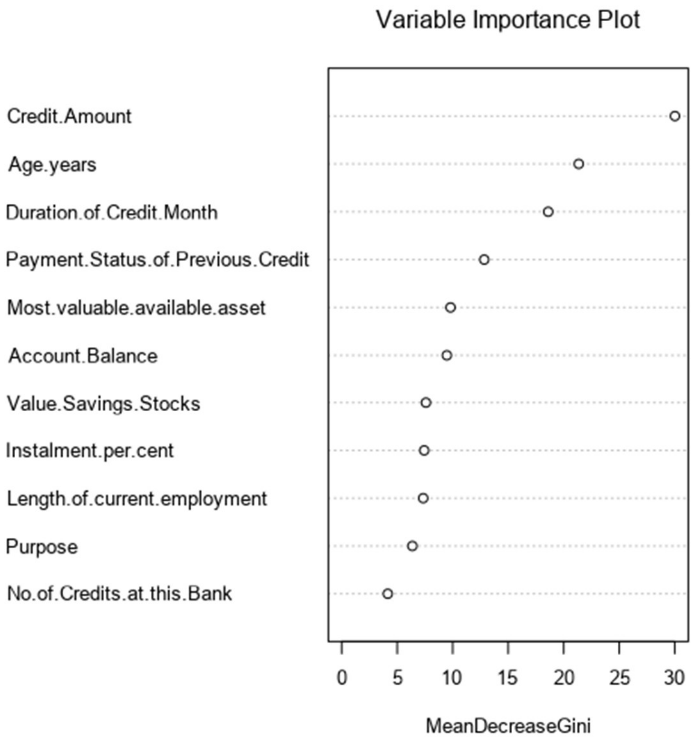


Confusion matrix of decisiontree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	90	24
Predicted_Non-Creditworthy	14	22

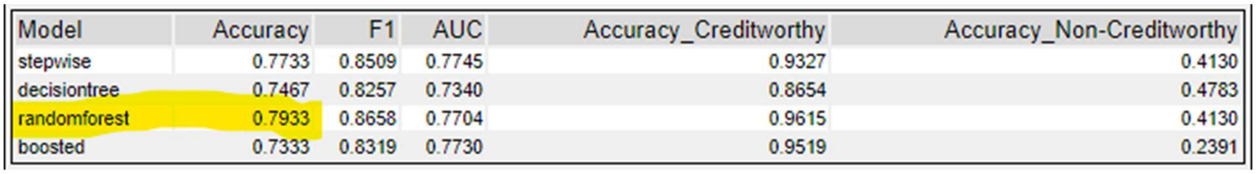
Precision	78.9%
Negative Predictive Value (NPV)	61.1%

Forest Model:
Top 3 features of importance are *Credit.Amount, Age.years, and Duration.of.Credit.Month.*

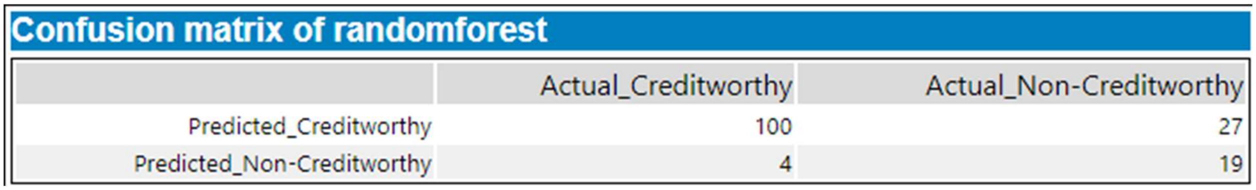


With the validation dataset, overall accuracy of decision tree model is 79.33%. The confusion matrix shows good performance for predicting creditworthiness (98 correct vs 6 wrong) but poorly for non-creditworthiness there are almost equal errors (20 correct vs 26 wrong).

Relative similarity between Precision and NPV indicates model has low bias for either classes.



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stepwise	0.7733	0.8509	0.7745	0.9327	0.4130
decisiontree	0.7467	0.8257	0.7340	0.8654	0.4783
randomforest	0.7933	0.8658	0.7704	0.9615	0.4130
boosted	0.7333	0.8319	0.7730	0.9519	0.2391

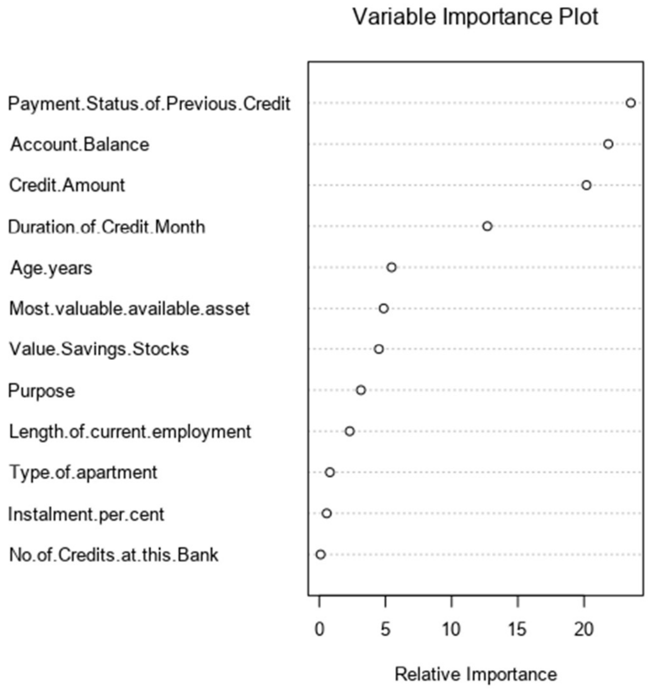


Confusion matrix of randomforest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	4	19

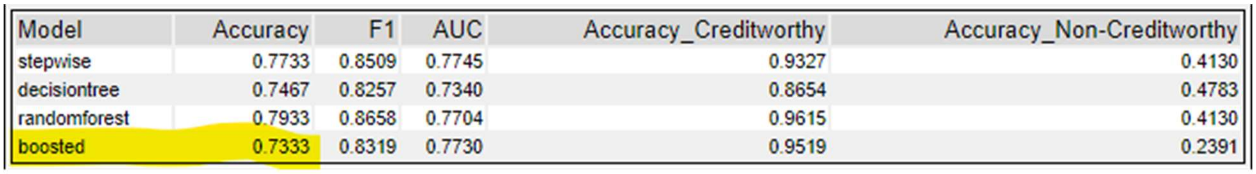
Precision	78.7%
Negative Predictive Value (NPV)	82.6%

Boosted Model:
Top 3 features of importance are *Payment.Status, Account*.Balance, *Credit.Amount, and Duration.of.Credit.Month.*

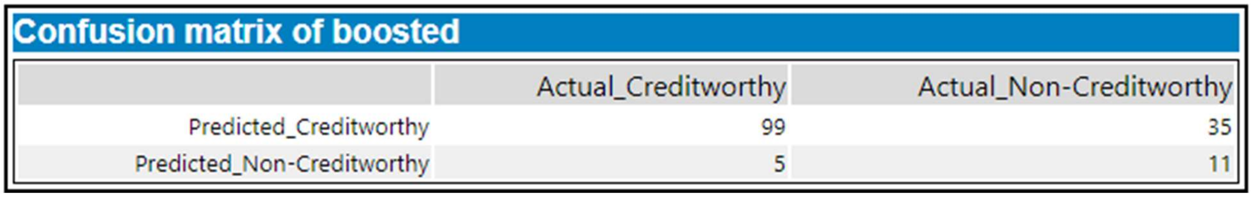


With the validation dataset, overall accuracy of decision tree model is 73.33%. The confusion matrix shows good performance for predicting creditworthiness (99 correct vs 5 wrong) but very poorly for non-creditworthiness where there are more errors (12 correct vs 34 wrong).

Relative similarity between Precision and NPV indicates model has low bias for either classes.



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stepwise	0.7733	0.8509	0.7745	0.9327	0.4130
decisiontree	0.7467	0.8257	0.7340	0.8654	0.4783
randomforest	0.7933	0.8658	0.7704	0.9615	0.4130
boosted	0.7333	0.8319	0.7730	0.9519	0.2391



Confusion matrix of boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	99	35
Predicted_Non-Creditworthy	5	11

Precision	73.9%
Negative Predictive Value (NPV)	68.8%

Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    - ROC graph
    - Bias in the Confusion Matrices

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

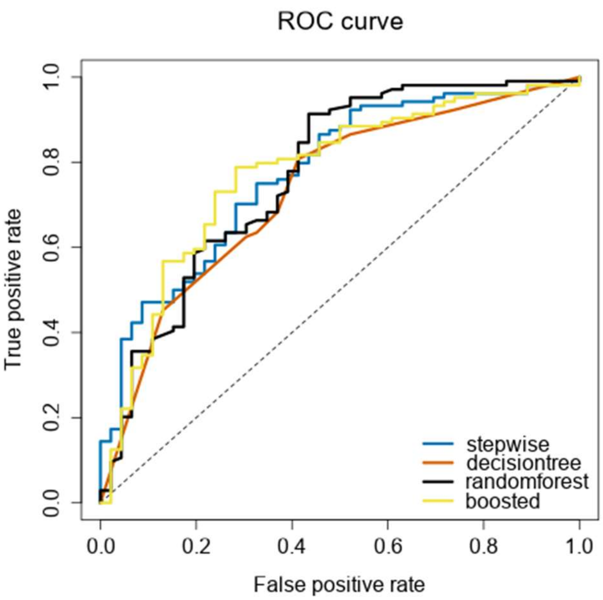
Based solely on overall prediction accuracy on the validation dataset, the model with the highest value is Random Forest.

The highest accuracy for creditworthiness, 96.2%, is Random Forest model, while its accuracy for non-creditworthiness is second highest at 41.3%. The highest accuracy non-creditworthiness is 47.8% with the Decision Tree model but its accuracy for creditworthiness is the lowest at 86.5%. Random Forest model perform well overall for accuracy in creditworthiness (top 1) and non-creditworthiness (top 2).

From the ROC curve, Random Forest model reaches the maximum of True positive rate the fastest.

From the comparison between Precision and NPV for each model, high bias is found in Decision Tree model while low biases are in Boosted, Random Forest and Logistic Regression models.

So Random Forest model has low bias, reaches peak of ROC curve fastest and has highest overall prediction accuracy over the other models.



- How many individuals are creditworthy?

  Using the Random Forest model to predict creditworthiness, 411 of the 500 new applicants for loan are creditworthy.