Project: Creditworthiness

Answer these questions

- What decisions needs to be made?

  The decision that needs to be made by the bank is whether the 500 new applicants are creditworthy to be able to approve a bank loan.

- What data is needed to inform those decisions?

  What is needed to inform the decisions are the score from the data credit training and the score of customer data to determine if the clients are creditworthy based on the probabilities and considering the different variables as Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Most-valuable-available-asset, Age-years, Type-of-apartment, No-of-Credits-at-this-Bank

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

  The models that needs to be used to make this decision is Binary since we must classify every single one of the clients whether the loan applicant is creditworthy or non-creditworthy. Whether they can get approved for the bank loan or not, this decision depends on different variables, considering the variables and correlation with the Credit-Application-Result.

- 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”.

  Yes, according to the Logistic Regression Model, Boosted Model, the Decision tree, and with a bit less correlation on the Forest Model, there was a strong correlation between Account-Balance and the Credit-Application-Results.

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

  Yes, there is data missing in Age-years and 69% missing data in Most-valuable-available-asset.

- 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.

  Yes, there is low variability in the subset of Guarantors.

- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

  Yes, the Age Years rounded in my data set rounded up is 36.

:

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

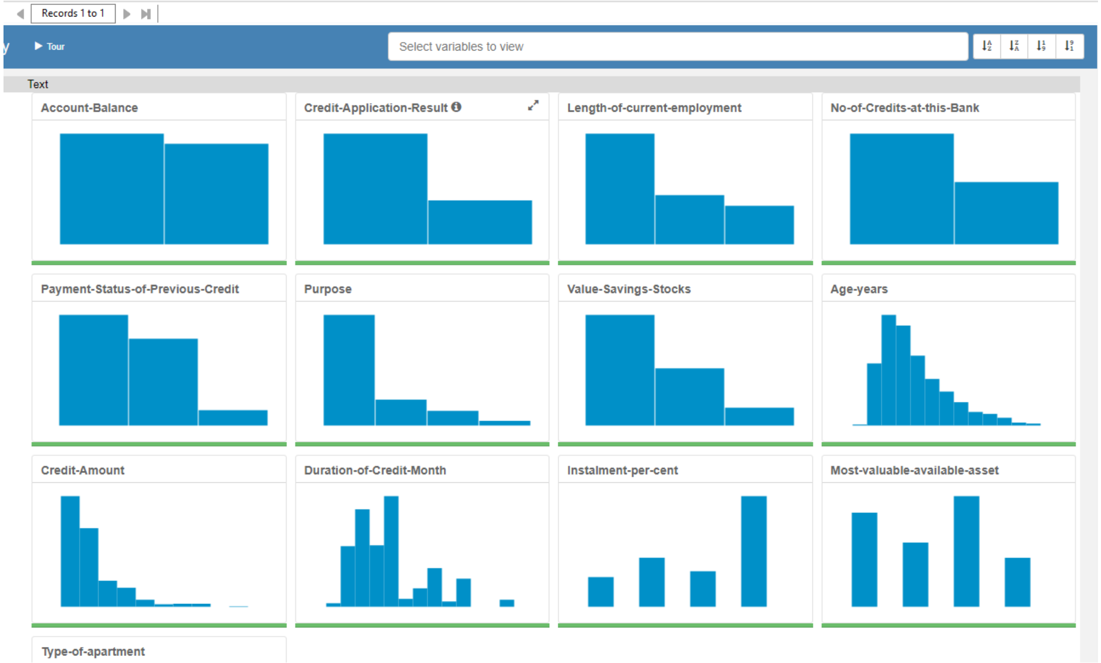
Telephone	Double
Foreign-Worker	Double

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 fields that where remove where: Guarantors, Duration-in-Current-address, Concurrent-credits, Telephone, Occupation, No-of-Dependents, Foreign-Worker. Guarantors, Foreign workers, No-of-Dependents are examples of uniformity distributed data skew to one of the values therefore are low variable. Concurrent-credits had only one value therefore it was removed because of its low variability, Telephone and Occupation is removed due to uniform value across the data. Duration-in-Current-address had many values missing therefore it was best to remove, the field that was imputed was Age-years to 33 as the median, because I wanted the model to be as closed as to the real data as possible, and the median represents the middle score less effected by outliers.



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.

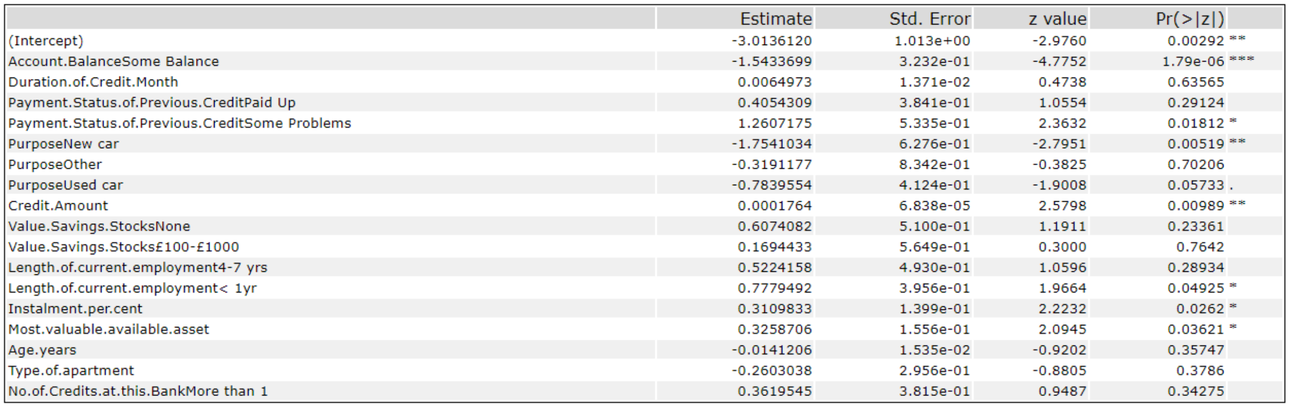
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*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.

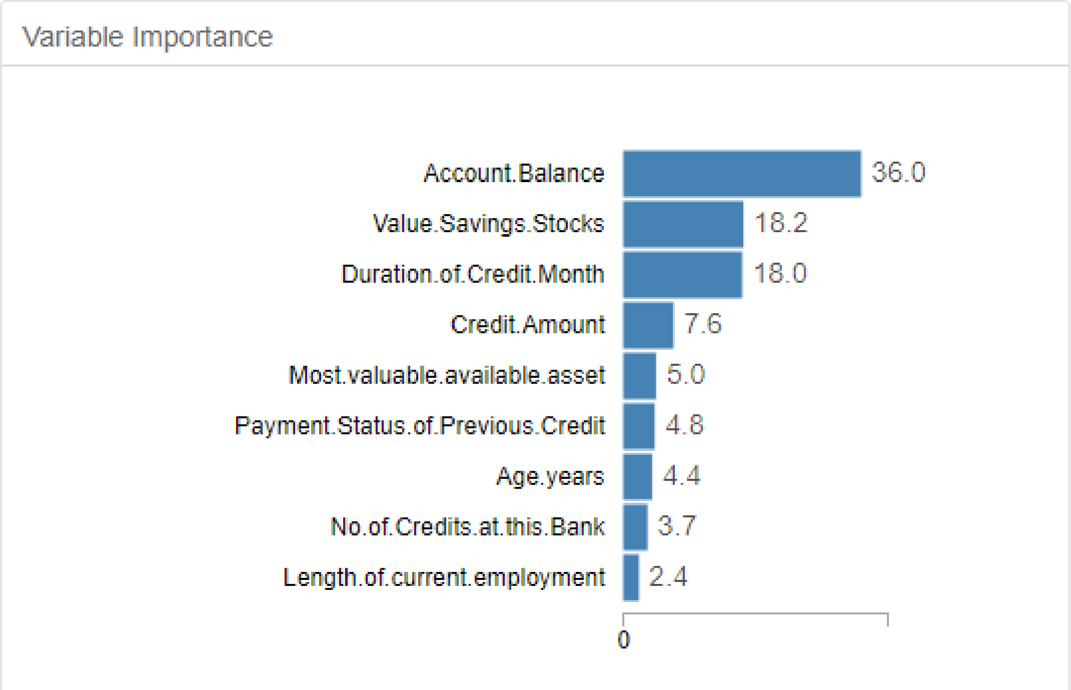
  The most important predictable variables are Account-Balance, Credit-Amount, and Payment-Status-of-Previous-Credit

Logistic Regression

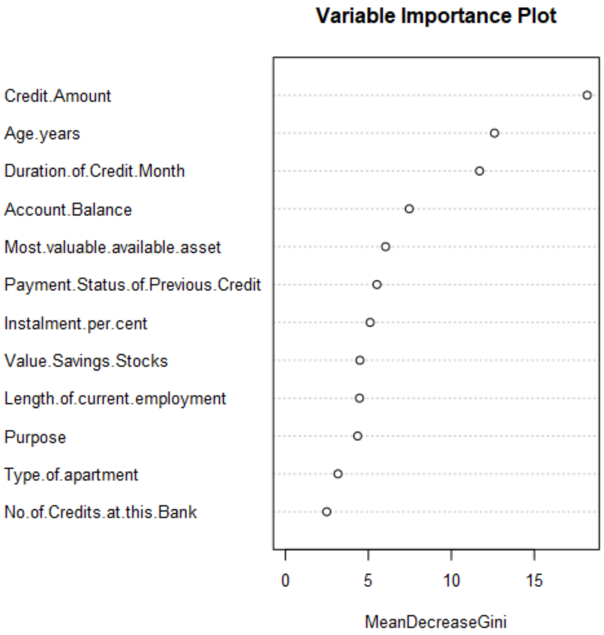


	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812	*
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519	**
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206	
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733	.
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989	**
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361	
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642	
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934	
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925	*
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262	*
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786	
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275	

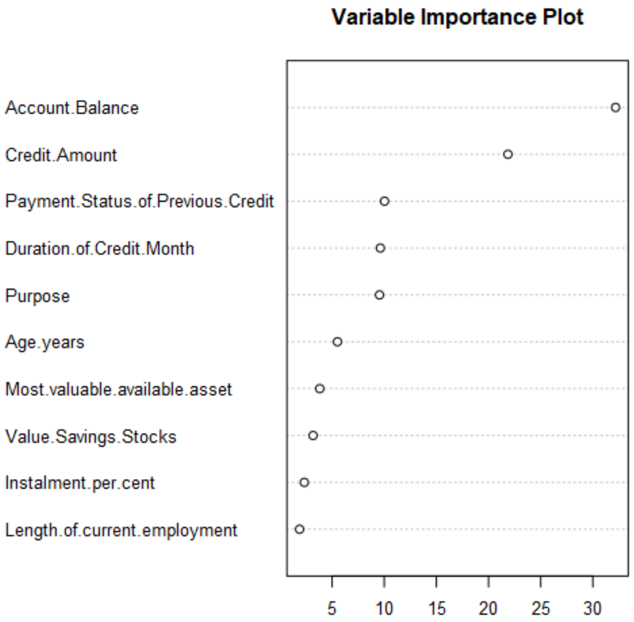
Decision Tree



Forest Model

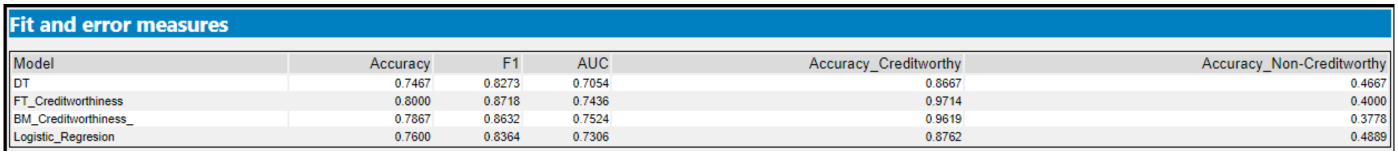


Boosted Model



- 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?

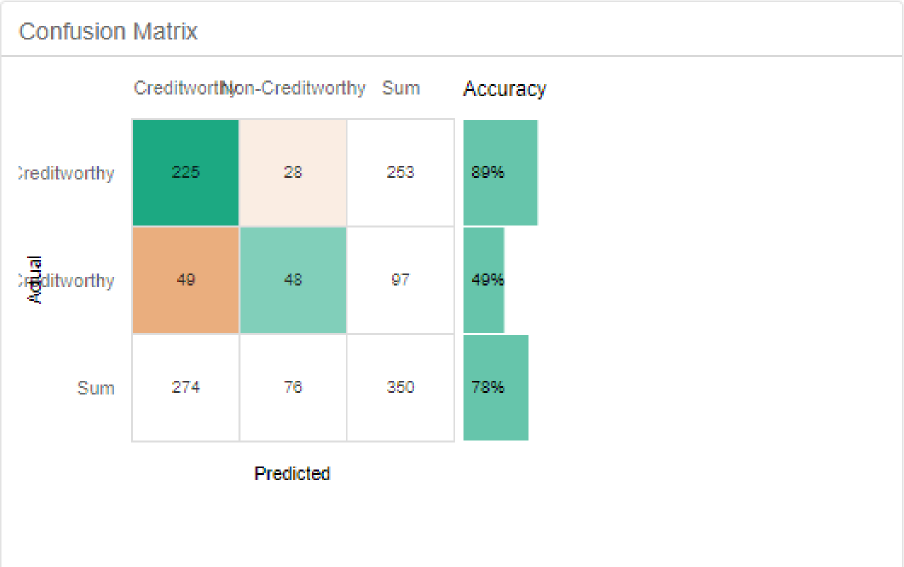
Overall percentage accuracy chart Per model



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8273	0.7054	0.8667	0.4667
FT_Creditworthiness	0.8000	0.8718	0.7436	0.9714	0.4000
BM_Creditworthiness_	0.7867	0.8632	0.7524	0.9619	0.3778
Logistic_Regresion	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion Matrix



Confusion Matrix

	Creditworthy	Non-Creditworthy	Sum	Accuracy
Creditworthy	225	28	253	89%
Non-Creditworthy	49	48	97	49%
Sum	274	76	350	78%

Bias in Model predictions

The Logistic Regression model has an overall accuracy of 76%, in correctly predicting creditworthy individuals 88% and the accuracy in correctly predicting non-creditworthy individuals is 49%. This means that this model has bias towards correctly predicting creditworthy individuals because its accuracy in the creditworthy segment is way higher than in the non-creditworthy.

The Decision Tree has an overall accuracy 75%, accuracy in predicting creditworthy individuals 87% and the accuracy in correctly predicting non-creditworthy individuals is 47%. This means that this model has bias towards correctly predicting creditworthy individuals because its accuracy in the creditworthy segment is way higher than in the non-creditworthy.

The Forest Model has an overall accuracy 80%, accuracy in predicting creditworthy individuals 97% and the accuracy in correctly predicting non-creditworthy individuals is 40%. This means that this model has bias towards correctly predicting creditworthy individuals because its accuracy in the creditworthy segment is way higher than in the non-creditworthy.

The Boosted Model has an overall accuracy 79%, in correctly predicting creditworthy individuals 96% and the accuracy in correctly predicting non-creditworthy individuals 38%. This means that this model has bias towards correctly predicting creditworthy individuals because its accuracy in the creditworthy segment is way higher than in the non-creditworthy.

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)

The classification model that came up was done through examination of data and validation of it, cleansing of data, and data forecasting based on different types of models uses that included Logistic Regression, Decision Tree, Forest Model and the Boosted Model. From the results of the model comparison report, the one chosen was the Logistic regression as it's among the highest in accuracy for the binary model, the total of new customer that will qualify for a loan is 406 individuals.

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:

  The model that I choose is the Logistic Regression

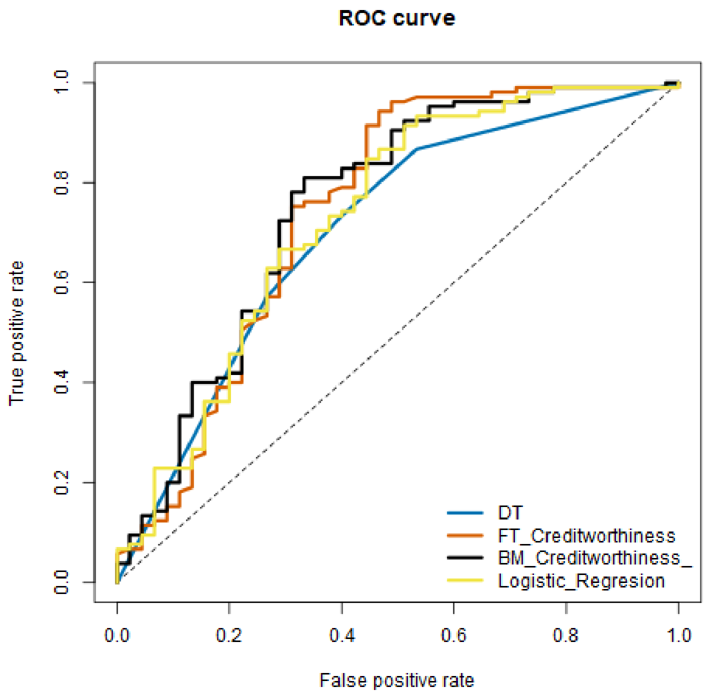
  - Overall Accuracy against your Validation set

    76% accuracy

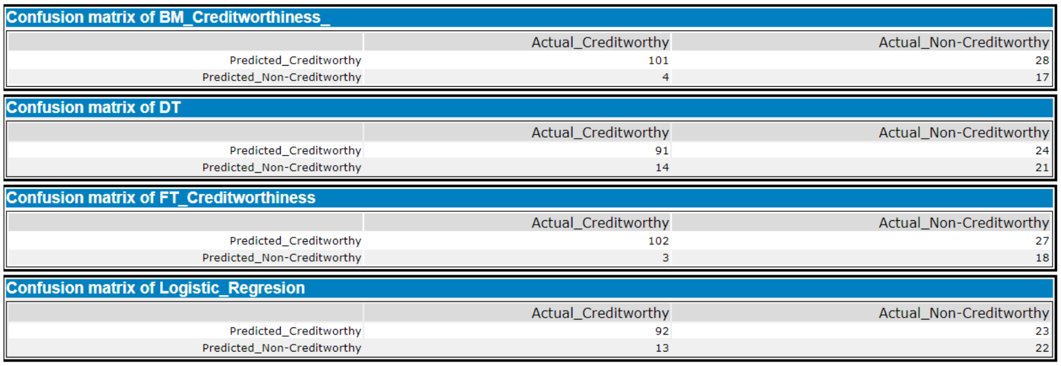
  - Accuracies within "Creditworthy" and "Non-Creditworthy" segments

    88% Creditworthy and 49% Non-Creditworthy

  - ROC graph



- Bias in the Confusion Matrices



Confusion matrix of BM_Creditworthiness_

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of FT_Creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Confusion matrix of Logistic_Regresion

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

- How many individuals are creditworthy?

406 individuals