Project: Creditworthiness

Step 1: Business and Data Understanding

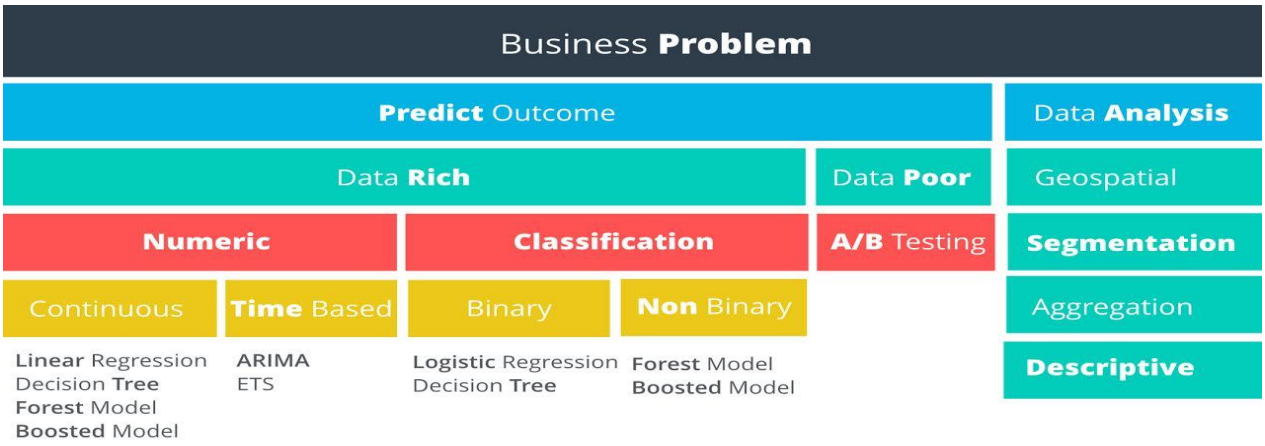
Key Decisions:

1. What decisions need to be made?
   The bank has a sudden influx of 500 loan applications which is higher than the normal. However, all the applications cannot be approved, we need to **decide which customers are creditworthy** and hence their loan applications can be approved.

2. What data is needed to inform those decisions?
   - Data on all past applications, and
   - The list of customers that need to be processed in the next few days.

   These should have the information such as Purpose of the loan, Duration of Credit Month, Account Balance, Credit Amount, Value Savings Stocks, Length of current employment, Instalment percent, Age, Number of credits at this bank, etc.,

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
   Going by the methodology map, the steps would be as follows
   a. We have a business problem of whose loan applications should be approved.
   b. For this, we need to predict whether the customer is creditworthy.
   c. We do have sufficient data to predict this.
   d. We must classify a customer as either Creditworthy or Non-Creditworthy, so this would be a ***Binary Classification***.



Step 2: Building the Training Set

Used Association Analysis tool in Alteryx to quickly check if there are any fields that highly-correlate with each other and the correlation is at least 0.70. But, *none of the numerical data fields are highly correlated*.

Full Correlation Matrix

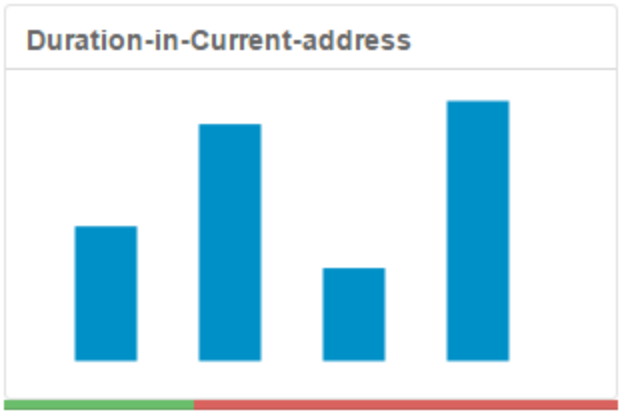
	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Duration.in.Current.address	Most.valuable.available.asset	Age.years
Duration.of.Credit.Month	1.000000	0.565054	0.145637	-0.032494	0.128814	-0.018171
Credit.Amount	0.565054	1.000000	-0.253286	-0.136621	0.457147	0.040486
Instalment.per.cent	0.145637	-0.253286	1.000000	0.131231	0.115114	0.111456
Duration.in.Current.address	-0.032494	-0.136621	0.131231	1.000000	-0.047386	0.301966
Most.valuable.available.asset	0.128814	0.457147	0.115114	-0.047386	1.000000	0.123579
Age.years	-0.018171	0.040486	0.111456	0.301966	0.123579	1.000000
Type.of.apartment	0.126967	0.100413	0.178926	-0.163386	0.182744	0.208552
No.of.dependents	-0.185180	0.082721	-0.293380	-0.036814	0.019435	0.046996
Telephone	0.238437	0.192532	0.038515	0.055112	0.083395	0.141103
Foreign.Worker	-0.207298	-0.045994	-0.155458	-0.015787	0.071932	-0.020939

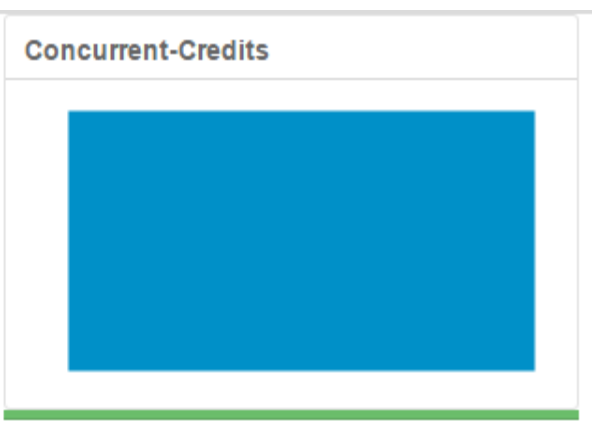
	Type.of.apartment	No.of.dependents	Telephone	Foreign.Worker
Duration.of.Credit.Month	0.126967	-0.185180	0.238437	-0.207298
Credit.Amount	0.100413	0.082721	0.192532	-0.045994
Instalment.per.cent	0.178926	-0.293380	0.038515	-0.155458
Duration.in.Current.address	-0.163386	-0.036814	0.055112	-0.015787
Most.valuable.available.asset	0.182744	0.019435	0.083395	0.071932
Age.years	0.208552	0.046996	0.141103	-0.020939
Type.of.apartment	1.000000	-0.010189	0.179688	-0.026742
No.of.dependents	-0.010189	1.000000	-0.097632	0.218454
Telephone	0.179688	-0.097632	1.000000	-0.168472
Foreign.Worker	-0.026742	0.218454	-0.168472	1.000000

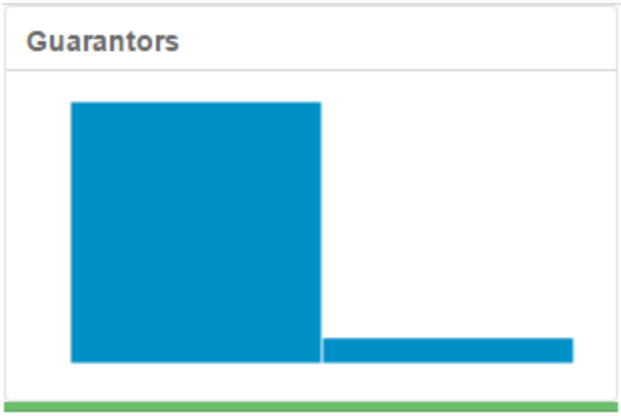
1. In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

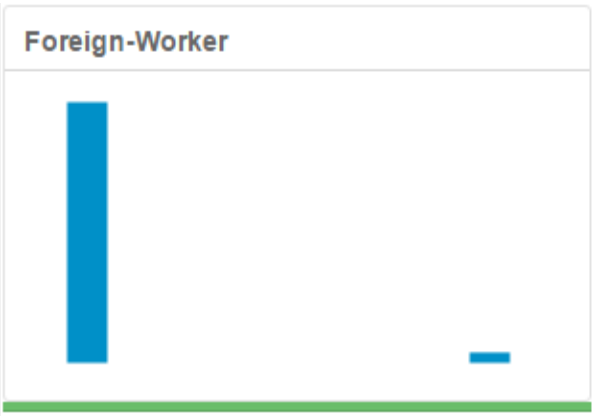
   To decide which fields to be removed/imputed, Field Summary tool in Alteryx was used.

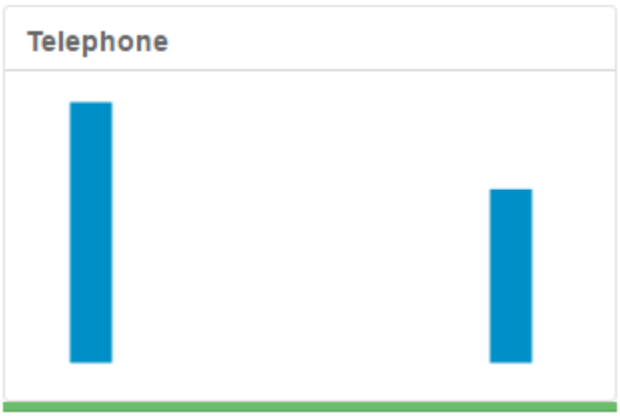
Removed Variables	Reason/Description
Guarantors	Low Variability, has 91.4% of 'None' and 8.6% of 'Yes', so this is removed.
Duration-in-Current-address	Missing Data, 69% of the values are missing. This cannot be imputed as it will skew the model, hence this is removed.
Concurrent-Credits	This is also a form of Low Variability, has only one value (Other Banks/Depts) through out, so this variable is removed.
Occupation	The data is uniform, i.e., only one value is present and there are no other variations of the data. This is also a kind of low variability; hence this cannot be added to the model.
Foreign-Worker	Low Variability, has 96.2% of '1' and 3.8% of '2', so this is removed.
No-of-dependents	Low Variability, 85.4% of '1' and 14.6% of '2' which is not proper value distribution, so this is removed.
Telephone	Used Association Analysis with Target Variable (Credit Application Result) to decide which one to remove. After examining the results, it is observed that Telephone has the lowest correlation to the target variable (high p-values) and there is no logical reason to include this variable.

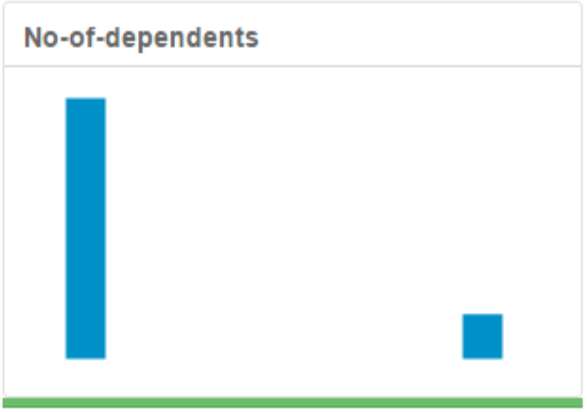








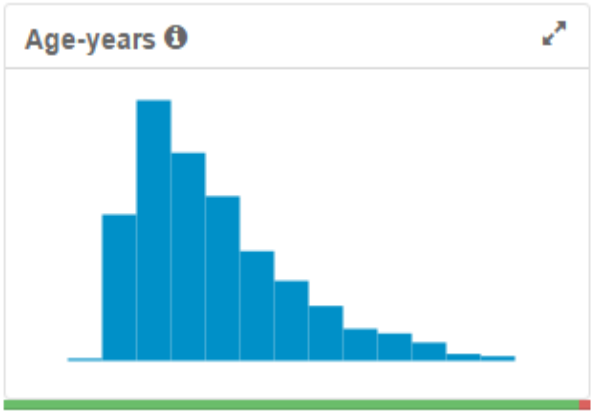


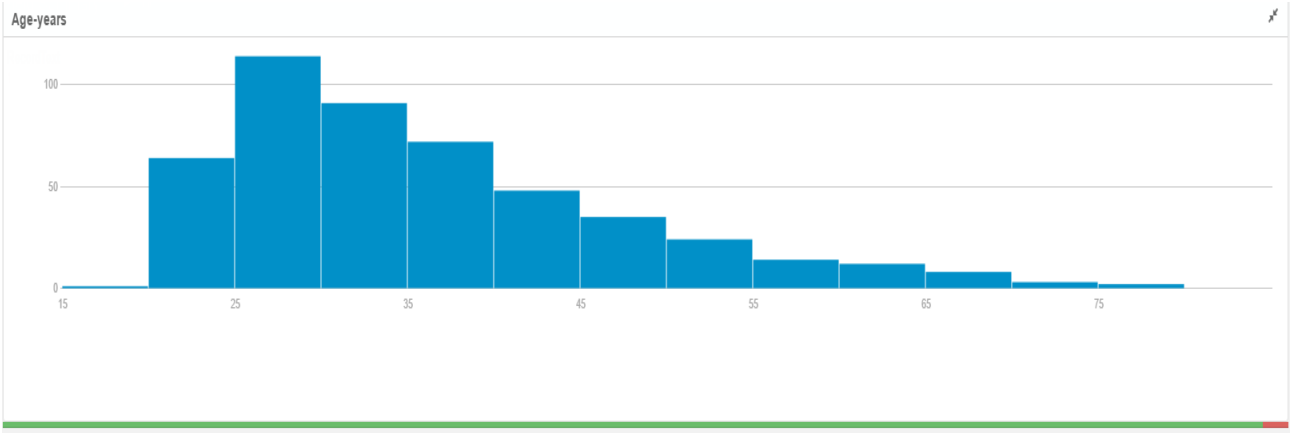


Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value
Most.valuable.available.asset	-0.232248	0.0050930 **
Duration.of.Credit.Month	-0.215149	0.0096065 **
Instalment.per.cent	-0.130496	0.1190020
Age.years	0.123088	0.1416213
Credit.Amount	-0.092205	0.2717004
Foreign.Worker	0.072525	0.3876717
Duration.in.Current.address	0.067284	0.4229716
Type.of.apartment	-0.039360	0.6395134
No.of.dependents	0.038037	0.6508161
Telephone	0.030838	0.7136766

The field Age-years has 2.4% of data missing. Since this is a numerical field we can impute the missing values using either the average or the median value. Since the data distribution is skewed, Median would be a better representation of the data instead of average/mean. So, missing values in the field Age-years is imputed with the median value.





Statistics	
min	19
Q1	27
median	33
Q3	42
max	75

Summary	
Count	500
Valid	488
Unique	53
Mean	35.64
Stdev	11.5

Step 3: Train your Classification Models

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

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 *
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
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747

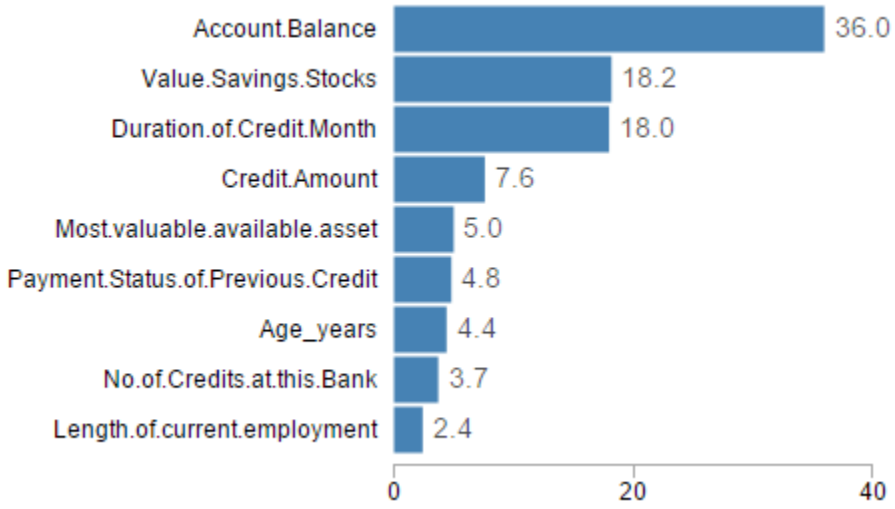
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

The significant predictor variables in the Logistic Regression model are Balance, Purpose, Credit Amount, Payment Status, Length of Current Employment, Instalment per cent, Most valuable available asset.

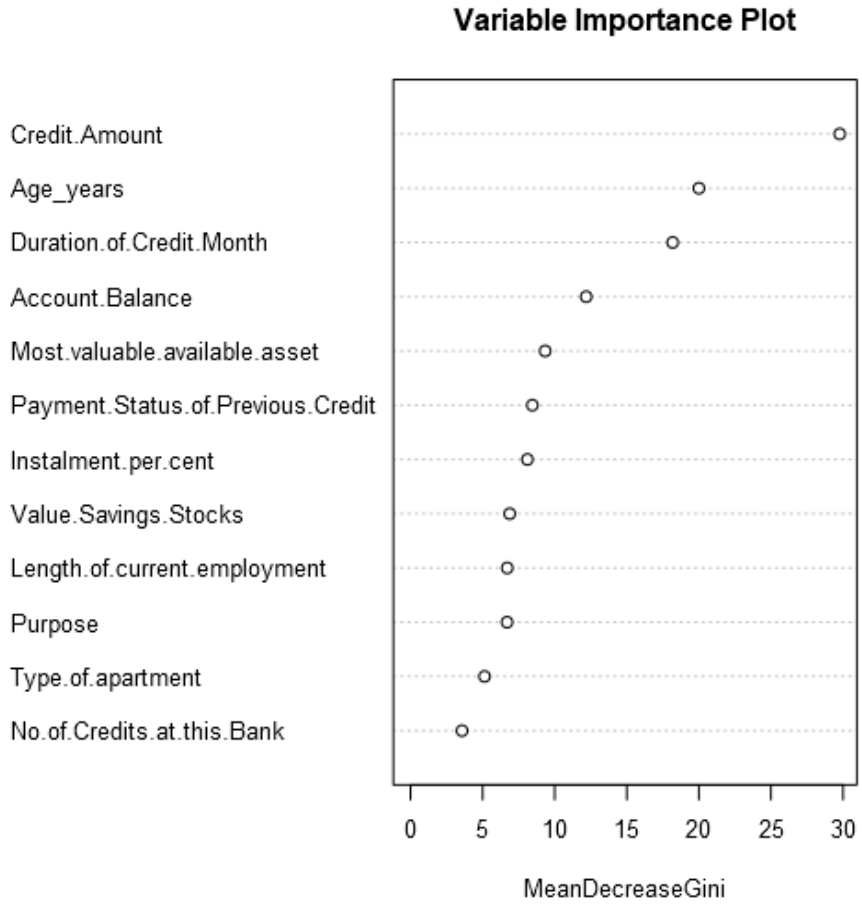
Decision Tree

The predictor variables as per Decision Tree are Account Balance, Value Savings Stocks, Duration of Credit Month, Credit Amount, Most Valuable available asset, Payment status of previous credit, Age_years, Number of credits at this Bank, and Length of current employment.



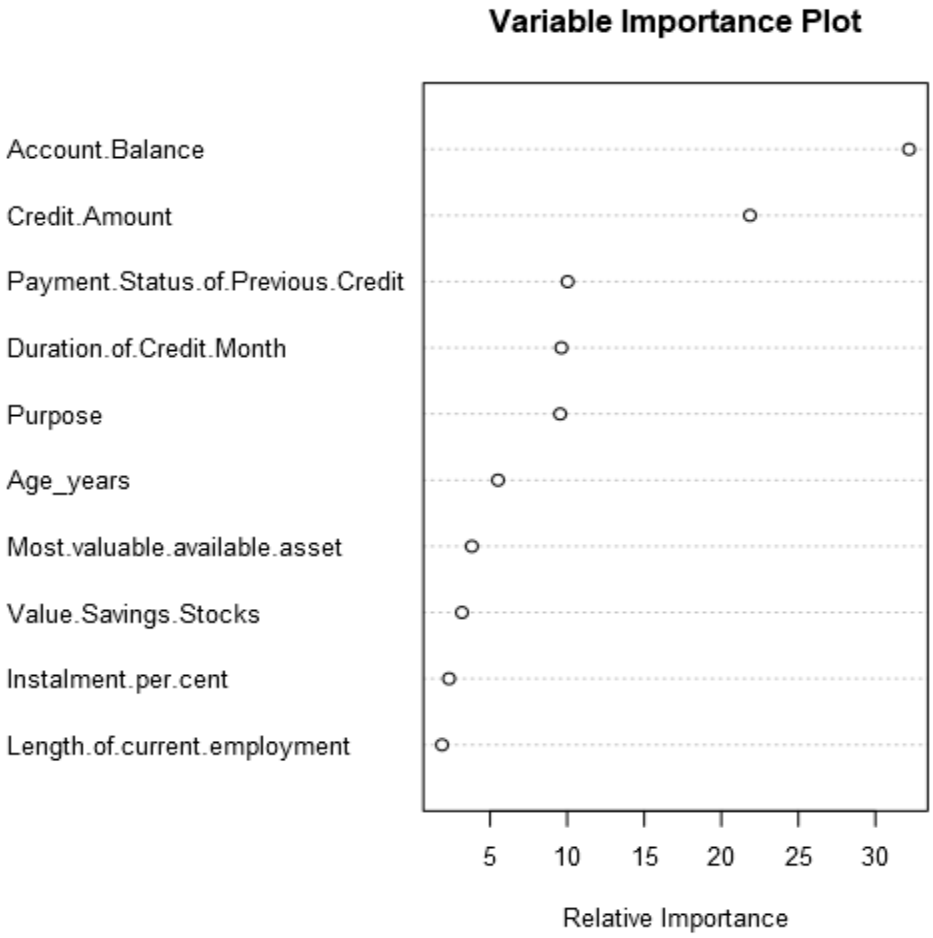
Forest Model

The significant predictor variables as per Forest Model are Credit Amount, Age_Years, Duration of credit month, Account Balance, Most valuable available asset, Payment status of previous credit, Instalment per cent, Value savings stocks, Length of current employment, purpose, Type of apartment, and Number of credits at this Bank.



Boosted Model

The significant predictor variables in Boosted Model are Account Balance, Credit Amount, Payment Status of previous credit, Duration of Credit Month, purpose, Age_years, Most valuable available asset, Value savings stocks, Instalment per cent, and Length of current employment.



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

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.8051	0.6875
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Accuracy in Logistic Regression against the Validation set is 0.7800 or 78%
Accuracy in Decision tree model against the Validation set is 0.7467 or 75%
Accuracy in Forest Model against the Validation set is 0.8000 or 80%
Accuracy in Boosted Model against the Validation set is 0.7867 or 79%

Confusion Matrices:

Logistic Regression:

Confusion matrix of Logistic_Regression

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

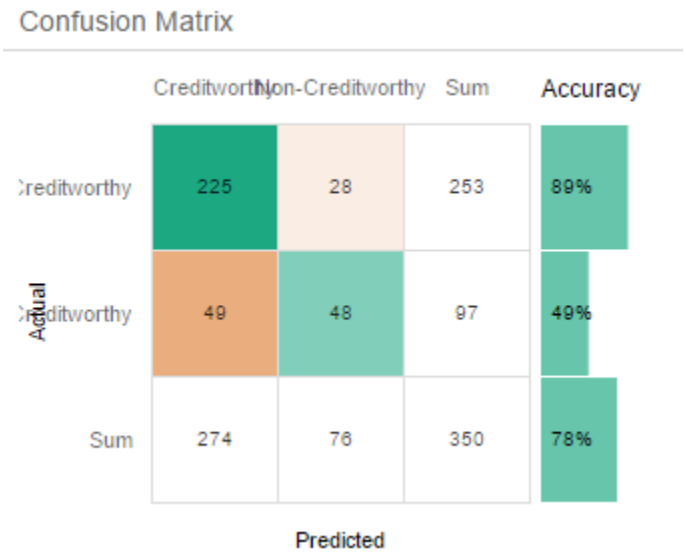
With the above confusion matrix and the comparison against the validation sample, it can be observed that there is a slight bias in the model towards predicting non-creditworthy status for clients who are actually creditworthy.

Decision Tree:

This model similar to Logistic Regression has a bias towards predicting non-creditworthy status for clients who are creditworthy. This is evident in the visualization below (Orange box), where 49 out of 97 applicants were falsely predicted as non-creditworthy.

Confusion matrix of Decision_Tree

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



Forest Model:

This model predicts better Non-creditworthy applicants (83%), and it does good in predicting the Creditworthy applicants(80%) as well. So, there doesn't seem to be any bias in this model.

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Boosted Model:

This model predicts 78% creditworthy and 81% non-creditworthy applicants. So, there doesn't seem to be any bias in this model.

Confusion matrix of Boosted_Model

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

Step 4: Writeup

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    c. ROC graph
    d. Bias in the Confusion Matrices

To decide the model that successfully predicts the creditworthiness, we use union tool to get all the models together and run against the validation sample. We get the consolidated report as shown below.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.8051	0.6875
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Logistic_Regression

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

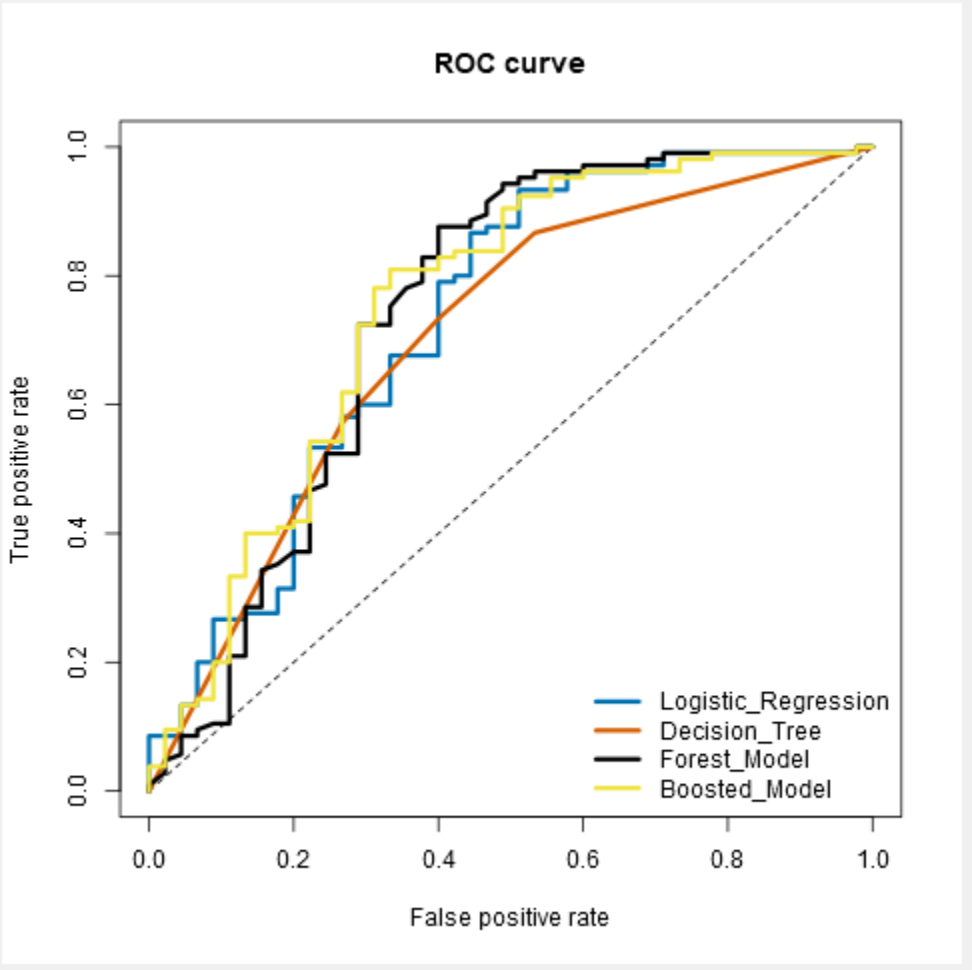
Looking at the Model Comparison Report, it appears that the Forest Model has the highest accuracy (0.8000) trailed by the Boosted Model (0.7867), the Logistic Regression Model (0.7800) and the Decision Tree Model (0.7467).

While considering the outcomes in detail, we can see that all models perform practically the same (0.78 to 0.80 being the maximum) for the actual Creditworthy category, the top two being the Logistic Regression and Forest models. The prediction for Non-creditworthy category shows that Logistic Regression and Decision Models are not sufficiently reliable as they perform

inadequately in this category (0.6000 – 0.6875). The highest accuracy is achieved by the Forest Model (0.8261) and then by the Boosted Model (0.8095).

Based on the confusion matrices shown above, there doesn't seem to be any bias in the Forest and Boosted Model.

Receiver Operating Characteristic(ROC) curve is a graphical plot that helps to visualize the performance of a binary classifier. Area Under the Curve(AUC) is the sign of a better performing classifier.



The above graph shows both the and Forest model (black curve) and the Boosted model (yellow curve) are really close and perform well. This can be understood by the AUC values in the above Model Comparison Report, where Forest model's AUC is 0.7419 and the Boosted model's AUC is 0.7524.

Considering all dimensions of the analysis, we can conclude that ***Forest Model*** performs best with the highest overall prediction accuracy for Creditworthy and Non-Creditworthy segments which the bank manager is interested in.

2. How many individuals are creditworthy?

Using Score tool in Alteryx, it is predicted that *415* applicants out of the 500 are Creditworthy.

Record #	Sum_Creditworthy
1	415

Alteryx Workflow: