Project: Creditworthiness

Step 1: Business and Data Understanding

We are a small bank and our teams typically get around 200m loan applications per week and we approve them by hand. Due to a recent financial scandal that hit a competitive bank last week, we suddenly have an influx of new people applying for loans to our bank instead of our competitor. While this is great for business, we are also responsible for determining the creditworthiness of the customer.

Management views this as an opportunity to use classification modeling to weed out the customers who don't pass the creditworthiness test.

Key Decisions:

Answer these questions

- What decisions needs to be made?
  - There is a group of 500 applicants and we need to decide if the applicants are creditworthy or not.
- What data is needed to inform those decisions?
  - We have access to past loan applicant's data which includes details about customer such as age and how long they are employed in the current job. We also have access to the individual's financial history.
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  - We would need a Binary model for this as it has only two outcomes – yes or no.

Step 2: Building the Training Set

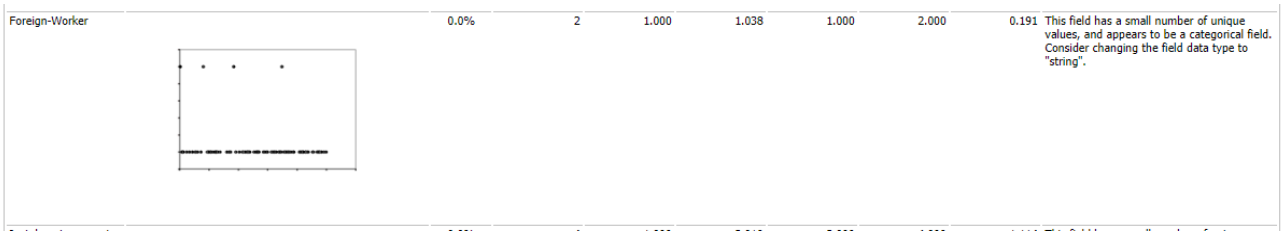
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  - I removed the field **Concurrent-Credits** and **Occupation**. This is because it has only one value and it would skew the data.
    Note: There seems to be a bug for the Field Summary tool Interactive output, so I am unable to visualize the field summary output as histograms.

https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Problem-with-Field-Summary/td-p/587812

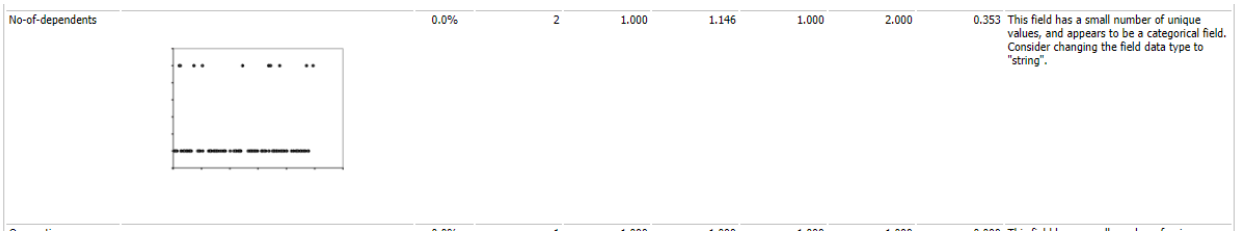
- I removed the field **Duration in Current Address** as there are too many null values.

Duration-in-Current-address	M
2	1
2	1
1	1

- I removed the fields **Guarantors** as it has just two options as results and one of the options is in majority. This would skew the results, so it has been removed.
- For the field **Age Years**, I used the median value to impute the data. We are using the median so that we can minimize skewing the data.
- The field **Foreign Worker** has small number of unique values



- The field **No. of Dependents** has low variability and more than 80% of values skews towards one value, so it must be removed.



- The field **Telephone** has also been removed as it is irrelevant to customer creditworthiness.

Step 3: Train your Classification Models

Logistic Regression

In the Logistic Regression model, we are using Credit Application Result as the target variable and some of the important predictor variables are **Account Balance, Credit Amount and Purpose**.

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

	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)

Below is the model comparison report for stepwise logistic regression model:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Logsitic	0.7600	0.8364	0.7306	0.8000	0.6286

Confusion matrix of Stepwise_Logsitic

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

The overall accuracy for this model is around 76.0%, the accuracy for creditworthy is higher (80%) than the accuracy for non-creditworthy (62.9%). This implies that this model is more likely to predict customers as non-creditworthy.

Bias Calculation:

Bias calculation needs us to look at the true negative rate and compare it with the ppv or the positive predictive value. If the true negative value (NPV) and PPV are close to each other, then we say that the model is unbiased, and if these values are not close then the model is biased. For the Logistic Regression Model:

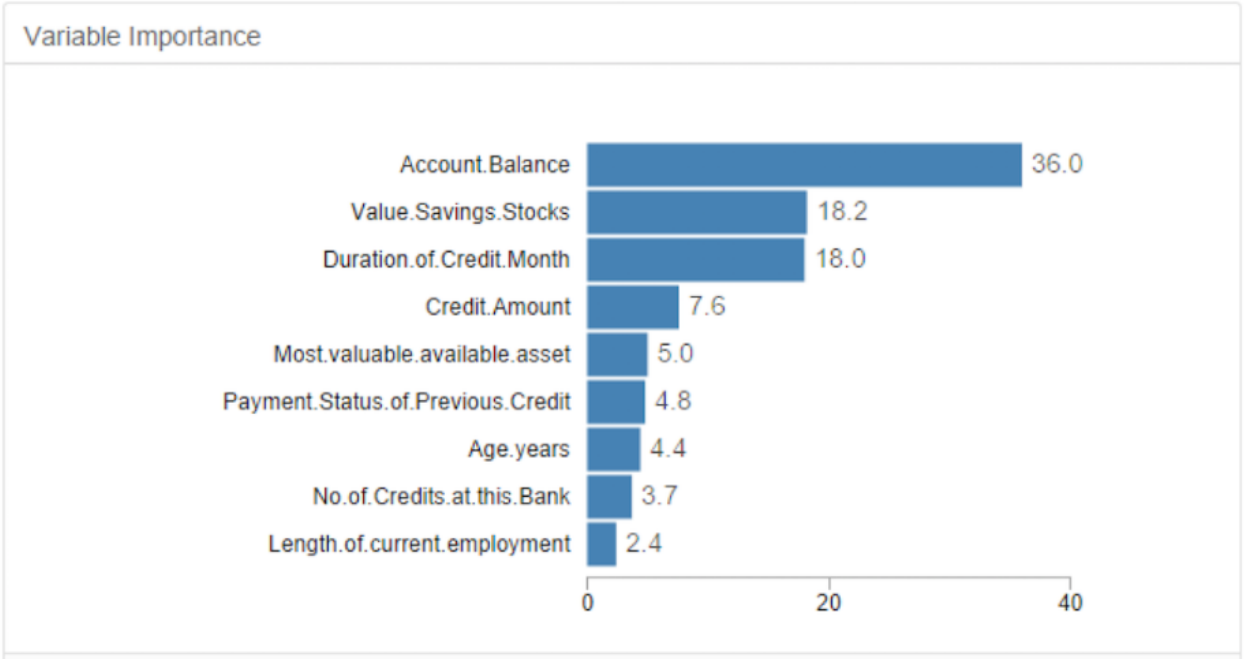
NPV = (No. of true negatives)/(true negatives + false negatives)
= 22/(22+13) = 22/35 = 62.85%

PPV = (No. of true positives)/(true positives + false positives)
= 92/(92+23) = 92/115 = 80%

Looking at the almost 20% difference between the two values we can conclude that this model is somewhat biased.

Decision Tree

Variable importance chart for the decision tree model is below:



The most important variables in the Decision Tree model are

- Account Balance
- Value Saving Stocks
- Duration of Credit Month

The model comparison for the Decision Tree model is below:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.7913	0.6000

Confusion matrix of DT_Credit

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

Bias Calculation:

Bias calculation needs us to look at the true negative rate and compare it with the ppv or the positive predictive value. If the true negative value (NPV) and PPV are close to each other, then we say that the model is unbiased, and if these values are not close then the model is biased. For the Logistic Regression Model:

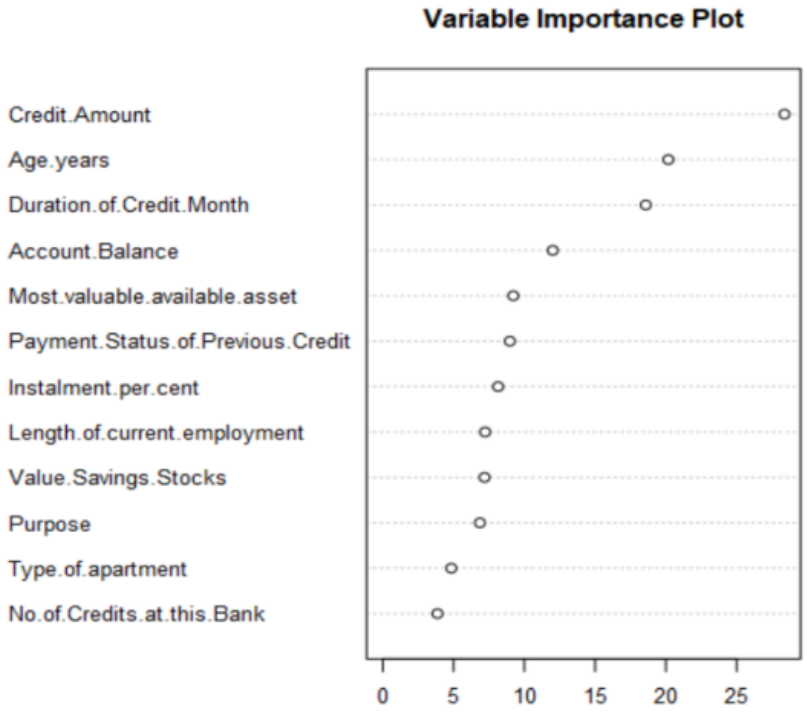
NPV = (No. of true negatives)/(true negatives + false negatives)
= 21/(21+14) = 21/35 = 60.0%

PPV = (No. of true positives)/(true positives + false positives)
= 91/(91+24) = 91/115 = 79.1%

Looking at the almost 20% difference between the two values we can conclude that this model is somewhat biased.

Forest Model

Variable importance chart for the **Random Forest** model is below:



The most important predictor variables are:

- Credit Amount
- Age in years
- Duration of Credit Month

Below is the model comparison report for the Forest Model:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Credit	0.8000	0.8718	0.7426	0.7907	0.8571

Confusion matrix of FM_Credit

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

The overall accuracy for this model is around 80%, the accuracy for creditworthy is lower (79%) than the accuracy for non-creditworthy (85.7%) which are comparable.

Bias Calculation:

Bias calculation needs us to look at the true negative rate and compare it with the ppv or the positive predictive value. If the true negative value (NPV) and PPV are close to each other, then we say that the model is unbiased, and if these values are not close then the model is biased. For the Logistic Regression Model:

NPV = (No. of true negatives)/(true negatives + false negatives)
= 18/(18+3) = 18/21 = 85.7%

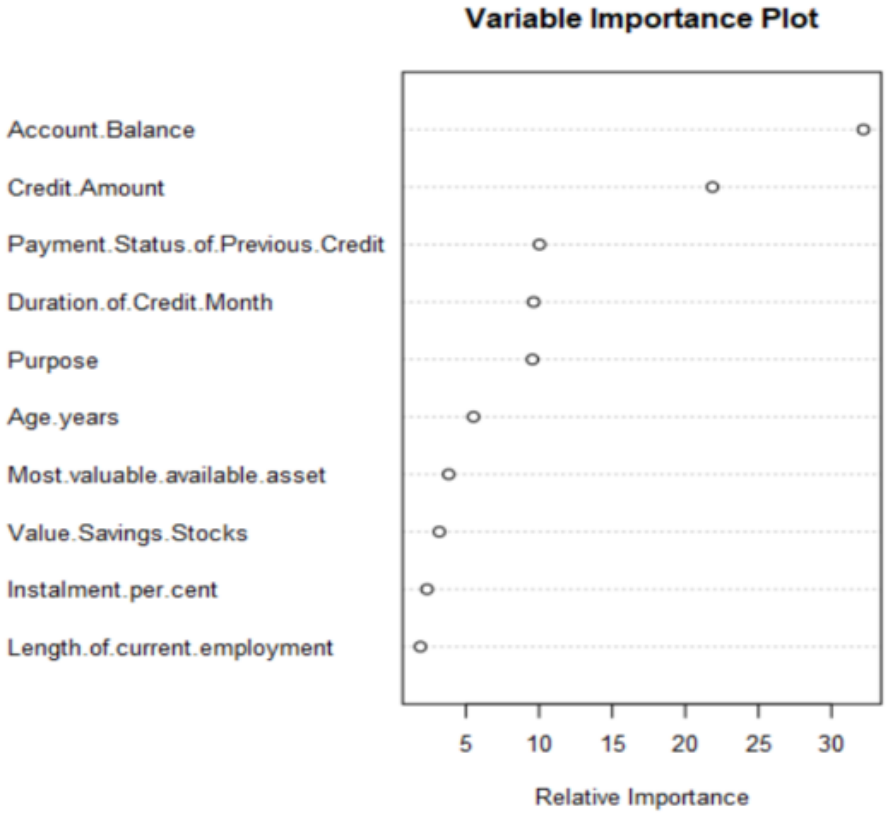
PPV = (No. of true positives)/(true positives + false positives)
= 102/(102+27) = 102/129 = 79%

Looking at the small difference between the two values we can safely conclude that this model has low bias, in other words, this model is unbiased.

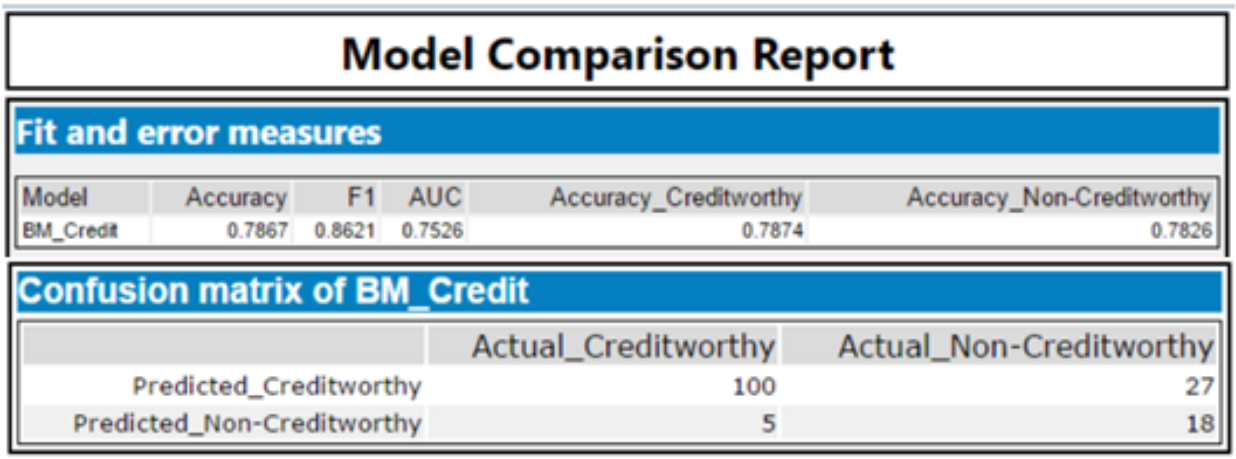
Boosted Model

The most important variables for the Boosted model are:

- Account Balance
- Credit Amount
- Payment Status of Previous Credit



Overall accuracy for this model is 78.67%.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_Credit	0.7867	0.8621	0.7526	0.7874	0.7826

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Bias Calculation:

Bias calculation needs us to look at the true negative rate and compare it with the ppv or the positive predictive value. If the true negative value (NPV) and PPV are close to each other, then we say that the model is unbiased, and if these values are not close then the model is biased. For the Logistic Regression Model:

NPV = (No. of true negatives)/(true negatives + false negatives)
= 18/(18+5) = 18/23 = 78.26%

PPV = (No. of true positives)/(true positives + false positives)
= 100/(100+27) = 100/127 = 78.74%

Looking at the almost non-existent difference between the two values we can safely conclude that this model has no bias, in other words, this model is unbiased.

Step 4: Writeup

Forest model offers the highest accuracy at 80%, hence that is the chosen model. Its predictions for credit worthiness are the highest among all the others. The accuracy difference between creditworthy and noncreditworthy is the most comparable and this makes sure that the right people get their loans approved and most importantly the bank does not lose money by approving the loans for the wrong people.

Model Comparison Report

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Linear_regresion	0.7800	0.8520	0.7314	0.8051	0.6875
Decision_Tree	0.6733	0.7721	0.6296	0.7545	0.4500
Forest_Model	0.7933	0.8681	0.7368	0.7846	0.8500
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_Credit	0.7867	0.8621	0.7526	0.7874	0.7826

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.7913	0.6000

Confusion matrix of DT_Credit

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

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Credit	0.8000	0.8718	0.7426	0.7907	0.8571

Confusion matrix of FM_Credit

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

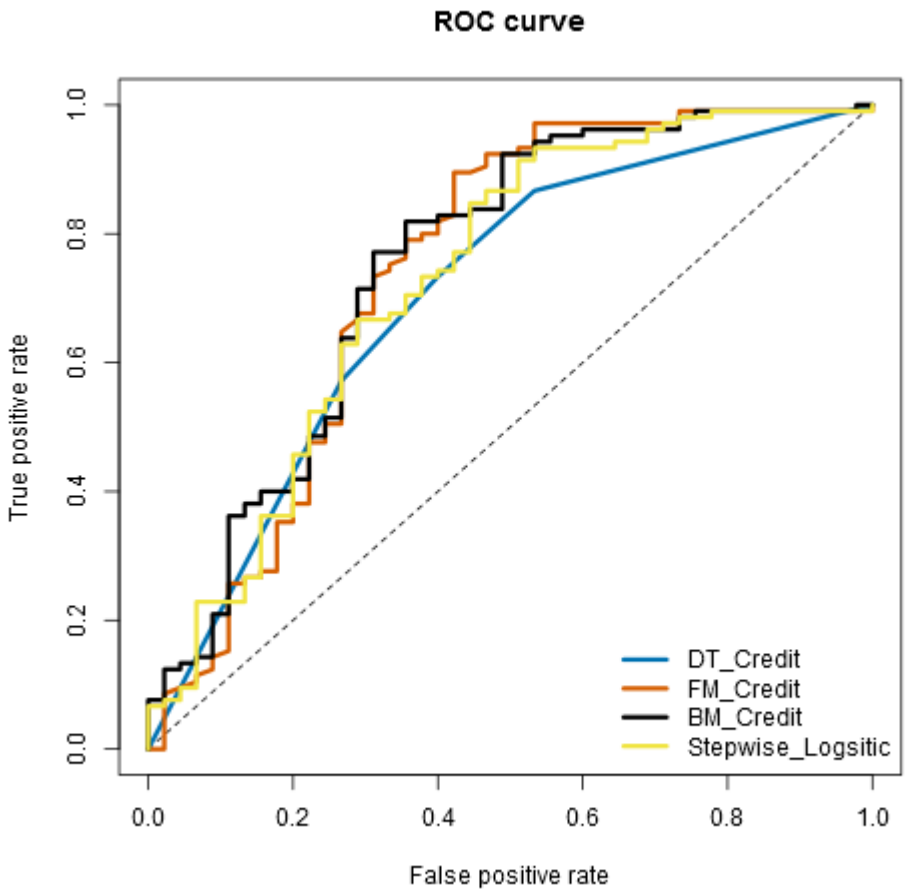
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Logsitic	0.7600	0.8364	0.7306	0.8000	0.6286

Confusion matrix of Stepwise_Logsitic

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

According to the forest model, **408 customers are creditworthy**.
Below is the ROC curve for all four models:



The **forest Model reaches true positive rate at the fastest rate**. It is the least biased towards any decisions as we can conclude from the accuracy difference between creditworthy and non-creditworthy. The M model also has the highest amount of area under it (AUC), thus proving that it is the best model for our case.