Project: Creditworthiness

Step 1: Business and Data Understanding

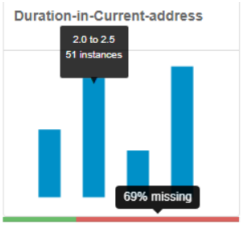
The recent financial scandal has led to a huge increase in the number of people applying for loans. A more efficient system is required to deal with almost double the typical weekly number of loan approval requests. This project will focus on the steps taken to create an efficient model that predicts the viability of all the new customers for loan approval.

The model is trained using a training set consisting of historic loan approval data. The variables identify key customer demographic fields to determine potential predictor variables. The target variable is binary in nature, as appraising a customer for a loan is an equitant of a yes or no answer. A binary model will have to be used to accurately make the prediction.

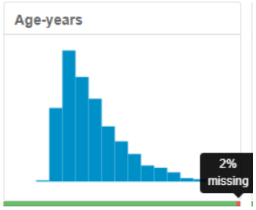
Step 2: Building the Training Set

As a first step, we used the Alteryx Field Summary tool to view the distribution of each and every variable. A lot of decisions to remove or keep fields were possible with this simple initial step:

- The categorical variable ‘Duration-in-Current-address’ does not contain data for 69% of its records and was removed for this reason.

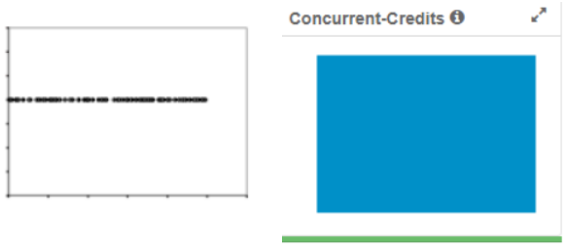


- 12 records (out of 500) were null for the Age-years field. A *impute* tool was used to substitute the null values with the median Age-years value. A median is a better aggregated value due to the skewed nature of this data

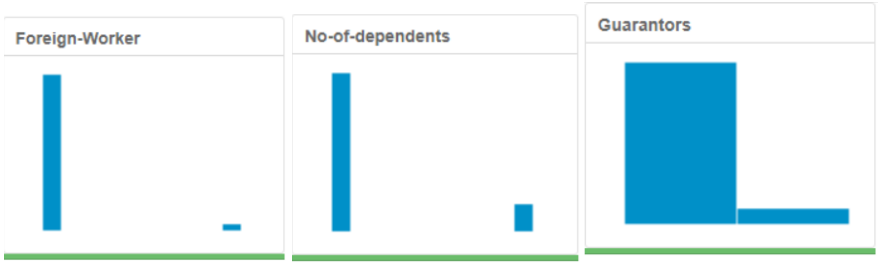


- A logical decision to exclude ‘Telephone’ seem eminent as we could not determine a possible way this being part of the customer profile would constitute towards credit worthiness.

- The fields for ‘Concurrent credits’ and ‘Occupation’ were also removed as this dataset did not provide any variability in the types for each. As shown in the charts below, all the data is the same, and show zero variability.

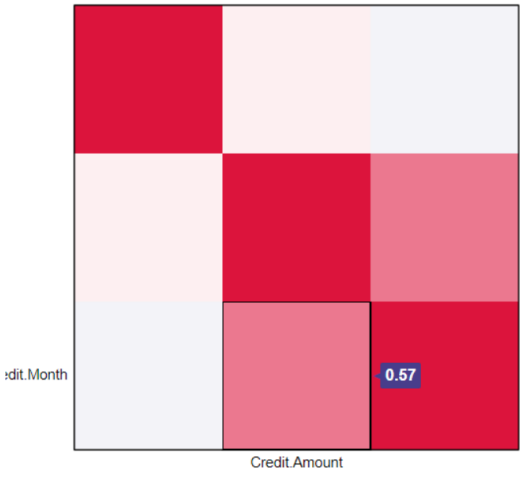


- And finally, the fields for ‘Guarantors’, ‘Foreign-Worker’ and ‘No-of-dependents’ were removed as they show very little variability by being skewed to a single categorical value



We also checked the validity of the training set by checking the validity of the three continuously numeric predictors in the model

- We verified that there were no significant outliers
- We also verified that these variables were not highly correlated variables or duplicating each others behavior. The correlation coefficient is under 0.7, the fields were not removed.



In the end we are left with a training dataset with 13 columns that will be used to train our classification model.

Step 3: Train your Classification Models

We tested out training set against 4 different Classification Models*: Logistic Regression, Decision Tree, Forest Model, Boosted Model*. Results for each model and their validations are compared using the 'Model Comparison' tool in Alteryx.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepwiseModel_Predict_Loans	0.7600	0.8364	0.7306	0.8762	0.4889
DTModel_Predict_Loans	0.7467	0.8273	0.7054	0.8667	0.4667
FMModel_Predict_Loans	0.8000	0.8707	0.7361	0.9619	0.4222
BoosterModel_loans	0.7067	0.8254	0.7050	0.9905	0.0444

You can see from the results above that the 'Forest Model' had the best result. Each model was also validated against an independent dataset which resulted in the accuracy calculations. Below we look at highlights from the reports of each model, where you can see the predictor variables. You will also see that each model has bias resulting from false positives and false negatives.

Logistic Regression (using Stepwise)

- Overall Accuracy: 76%
- Top 3 predictors:
  - 'Account Balance' with class 'Some Balance'
  - 'Purpose' with class 'New car'
  - 'Credit Amount'
- P-values which also shows other predictors with more than 95% confidence interval

	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	.

- Confusion Matrix

Confusion matrix of StepwiseModel_Predict_Loans

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

- CONCLUSION: Bias is seen in prediction
  - PPV = 92 / (92+23) = 0.8
  - NPV = 22 / (13+22) = .63

Decision Tree

- Accuracy: 75%
- Predictor variables (shown in order of importance)

- Confusion Matrix

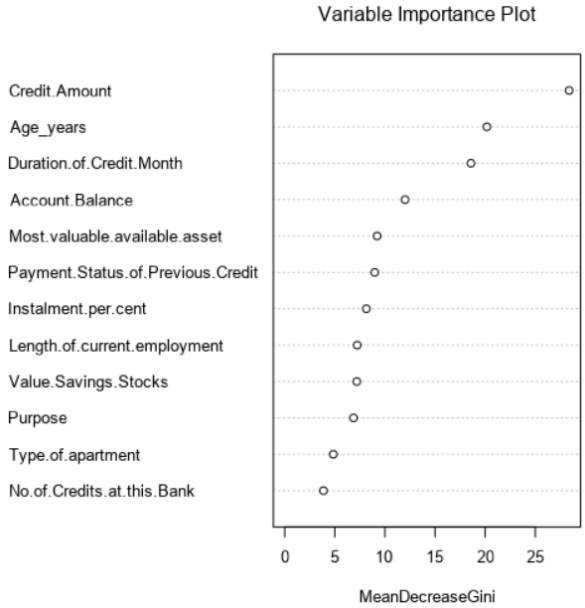
Confusion matrix of DTModel_Predict_Loans

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

- CONCLUSION: Bias is seen in prediction
  - PPV = 91 / (91+24) = 0.79
  - NPV = 21 / (14+21) = 0.6

Forest Model

- Accuracy: 80%
- Predictor Variables (also shown in order of importance like Decision tree above)



- Confusion Matrix

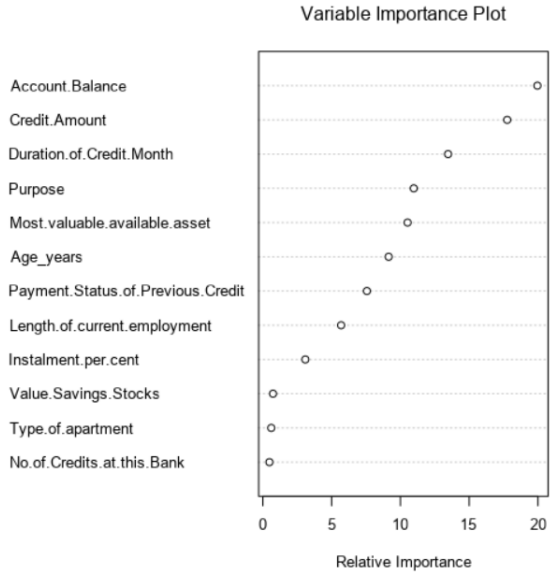
Confusion matrix of FMModel_Predict_Loans

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

- CONCLUSION: No Bias as PPV and NPV values are similar
    - PPV = 101 / (101+26) = 0.8
    - NPV = 19 / (4+19) = 0.83

Boosted Model

- Accuracy: 71%
- Predictor Variables (shown in order of importance)



- Confusion Matrix

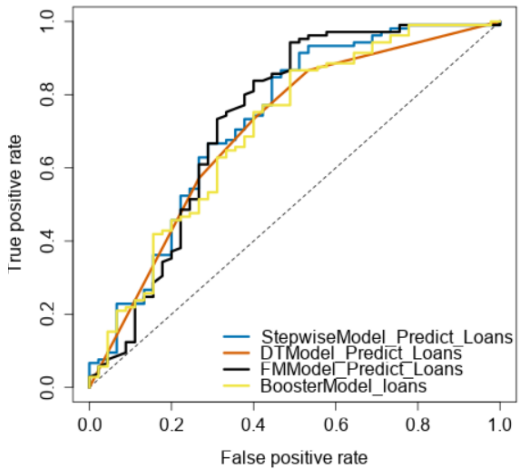
Confusion matrix of BoosterModel_loans

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	35
Predicted_Non-Creditworthy	2	10

- Bias is seen in prediction
    - PPV = 103 / (103+35) = 0.75
    - NPV = 10 / (2+10) = 0.83

Step 4: Writeup

The ‘Model Comparison’ tool consolidated all the accuracies for the models and revealed that the ‘Forest Model’ is the most accurate with 80%. The ROC curve (below) also confirms the validity of choosing the Forest model over the others as the plot for the Forest Model true positive rate reaches 1 faster than the other models



As an added measure, we also see that the AUC (or Area under the Curve) is also the highest amongst the models.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepwiseModel_Predict_Loans	0.7600	0.8364	0.7306	0.8762	0.4889
DTModel_Predict_Loans	0.7467	0.8273	0.7054	0.8667	0.4667
FMModel_Predict_Loans	0.8000	0.8707	0.7361	0.9619	0.4222
BoosterModel_loans	0.7533	0.8477	0.7073	0.9810	0.2222

The Forest model is the only model that does not show significant bias towards predicting a client as being ‘Creditworthy’. This can be seen with the numbers for ‘Positive Predictive Value’ (PPV) being close in value to its respective ‘Negative Predictive Value’ (NPV), while all other models showing a significant difference in their NPV and PPV values. [PPV= true positives \ (true positives + false positives) and NPV= true negatives \ (true negatives + false negatives)]

Finally, with the best model select, we scored the data of 500 new loan applicants against the Forest Model. **This resulted in 406 applicants being deemed as being credit worthy**.