Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

We need to decide from the 500 loan applicants who is creditworthy, in other words whom to give loan and whom to reject.

We need data about our previous customers (including previous loan amount, payment status, loan duration, customer's age, etc.) that we actually know whether were creditworthy or not in order to build a predictive model, and then we need the same information about the new 500 customers to make the key decision based on our built model.

As our target variable is the state of a customer to be creditworthy or not, so we have two possible outcomes of a target variable, we need to create a binary model.

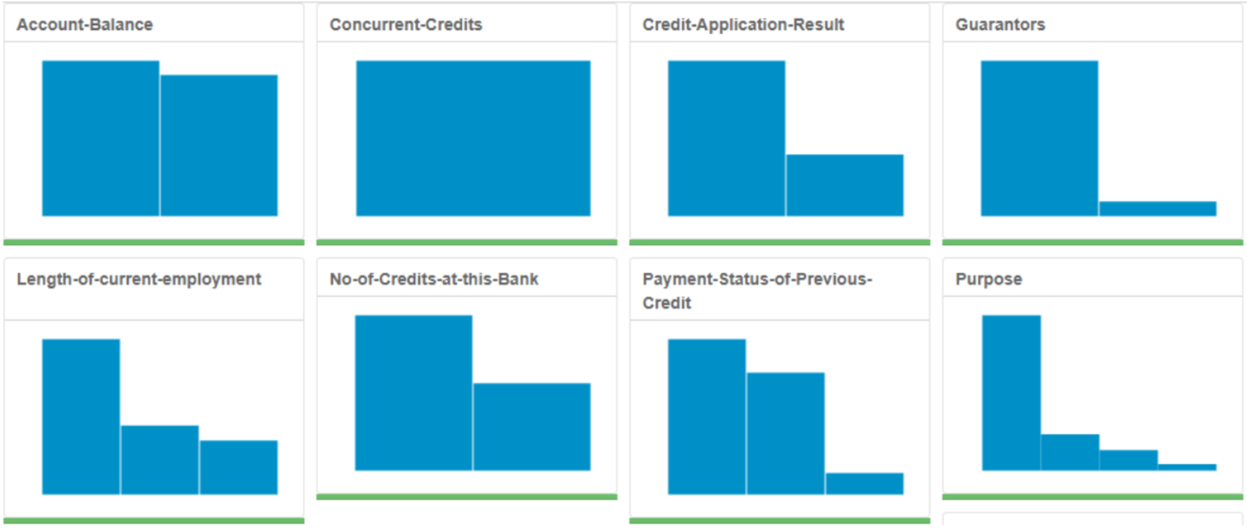
Step 2: Building the Training Set

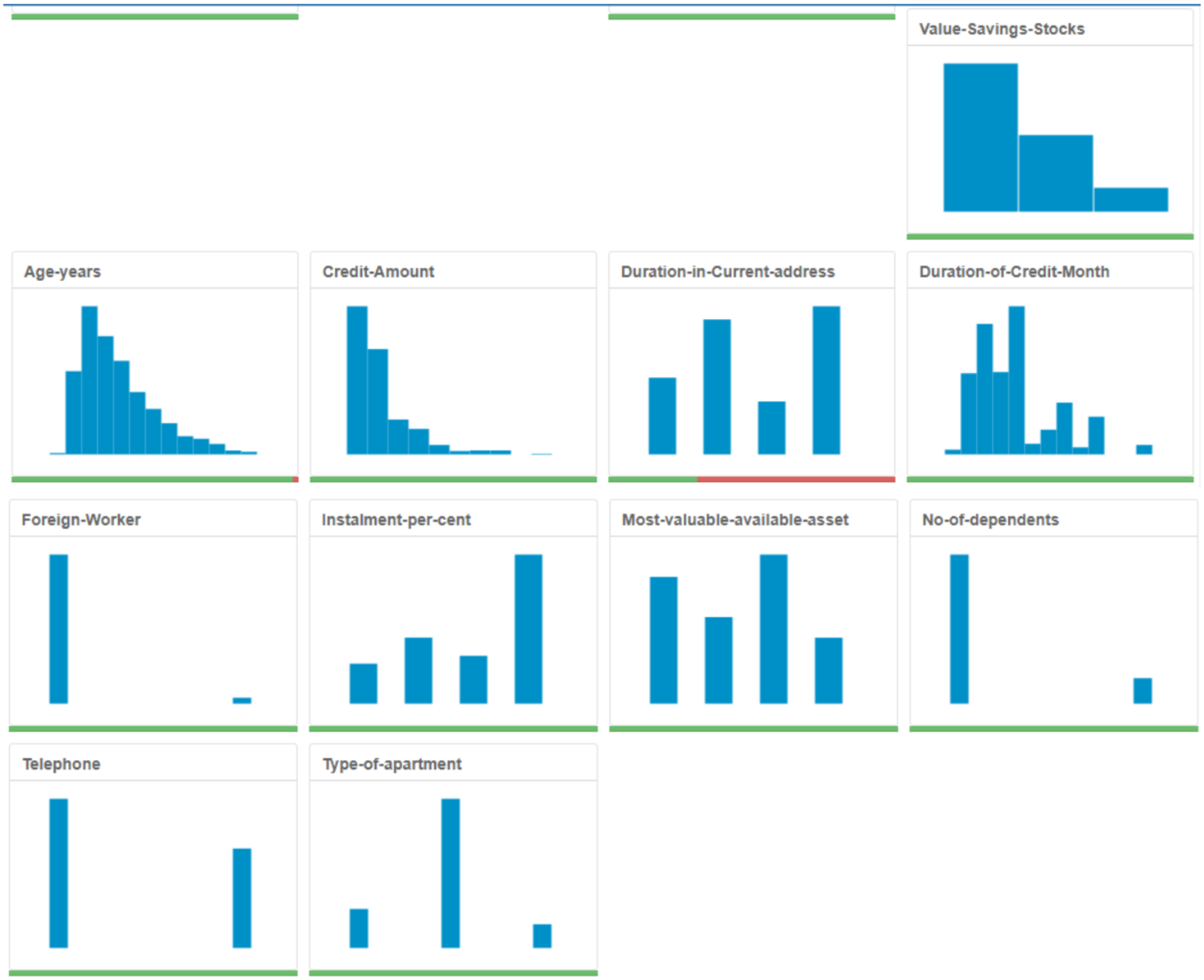
Using MS Excel Correlation tool, we find out that there are no highly correlated numeric fields in our data set.

	Duration-of-Credit-M	*Credit-Amount*	*lment-per-*	*-in-Current*	*able-availc*	*Age-years*	*-of-apartn*	*Occupation*	*of-depende*	*Telephone*	*reign-Worker*
Duration-c	1										
Credit-Am	0.57397971	1									
Instalment	0.068105529	-0.2888515	1								
Duration-i	-0.050649341	-0.1580691	0.173393	1							
Most-valu	0.29985487	0.32554538	0.081493	0.109297	1						
Age-years	-0.066318945	0.06864301	0.04054	0.301966	0.085437	1					
Type-of-ap	0.15251629	0.17007119	0.074533	-0.15755	0.373101	0.333075	1				
Occupatio	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1			
No-of-dep	-0.065269009	0.00398578	-0.12589	-0.05665	0.046454	0.117735	0.170738	#DIV/0!	1		
Telephone	0.1431762	0.28633845	0.029354	0.084925	0.203509	0.176479	0.101443	#DIV/0!	-0.04856	1	
Foreign-W	-0.115915664	0.02549285	-0.13341	-0.03659	-0.14601	-0.00328	-0.08985	#DIV/0!	0.065943	-0.05552	1

Using Field Summary tool in Alteryx we can see that a few fields need to be removed to have more accurate results in our analysis. Among these fields are 'Concurrent Credits', 'Guarantors', 'Foreign Workers', 'Occupation' and 'No of dependents' as they have low variability (see below histograms).
We will also remove the fields 'Duration in Current Address' fields as it has 68.8% missing values which is too high to keep or input. Finally, we will remove the field 'Telephone' as it does not have a logical connection with the target variable. We will input the field 'Age years' as only

2.4% of values is missing. For this input we will choose the median value and not mean, as the histogram is right-skewed and the median value will better represent the entire dataset.





Step 3: Train your Classification Models

Logistic Regression

After running the Logistic Regression model we use the Stepwise tool in Alteryx to find out the best predictor variable in our model.

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 *
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296 **
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 .
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 .

As we can see from the Coefficients table, the most significant fields for the target variables are 'Account Balance – Some Balance' with the highest level, then 'Credit Amount' and 'Purpose' – New car, as well as 'Payment Status of Previous Credit – Some Problems', 'Length of Current Employment - <1yr' and 'Installment Percent' with lower level of significance. The drawback of this model is that it has low R-squared: 0.2048.

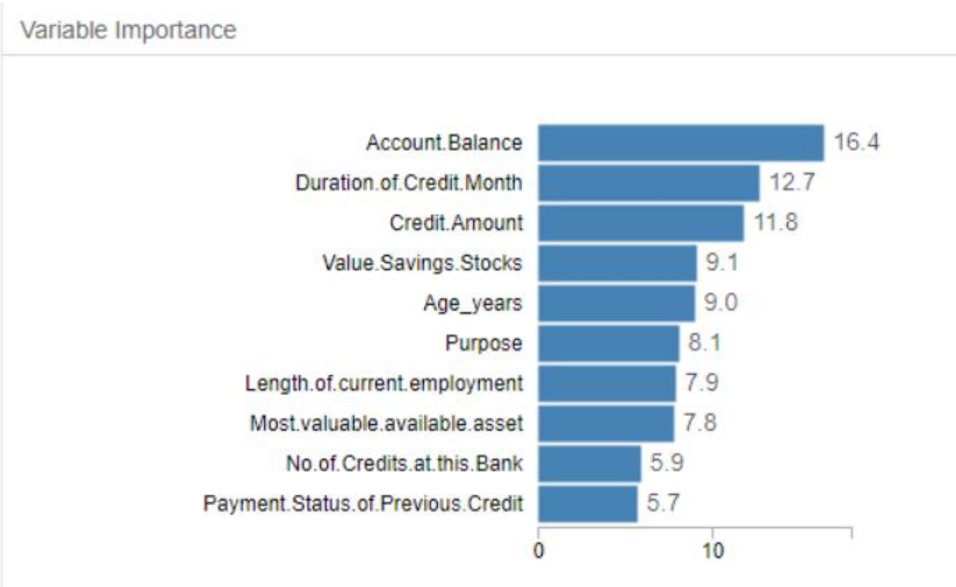
The overall accuracy of the model is 78%, the accuracy of creditworthiness is 90% quite, however, the accuracy of non-creditworthiness is 48%. This makes the model biased as there is a gap between the accuracy of 'creditworthy' and 'non-creditworthy' customers. In other words, the model has problem in correctly identifying creditworthy customers and predicts many creditworthy customers as non-creditworthy which significantly lowers the accuracy of predicting non-creditworthy customers. Below is the confusion matrix:

Confusion matrix of creditworthy_LR

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

As we can see, the number of False Negatives is quite high - 10, which means the model does not predict non-creditworthy customers accurately.

Decision Tree



From the variable importance we can see the most important predictor variables are ‘Account Balance’, ‘Duration of Credit Month’, ‘Credit Amount’, ‘Value Savings Stock’, ‘Age years’ and ‘Purpose’. As we can notice, the variables ‘Type of apartment’ and ‘Installment Percent’ are not significant at all.

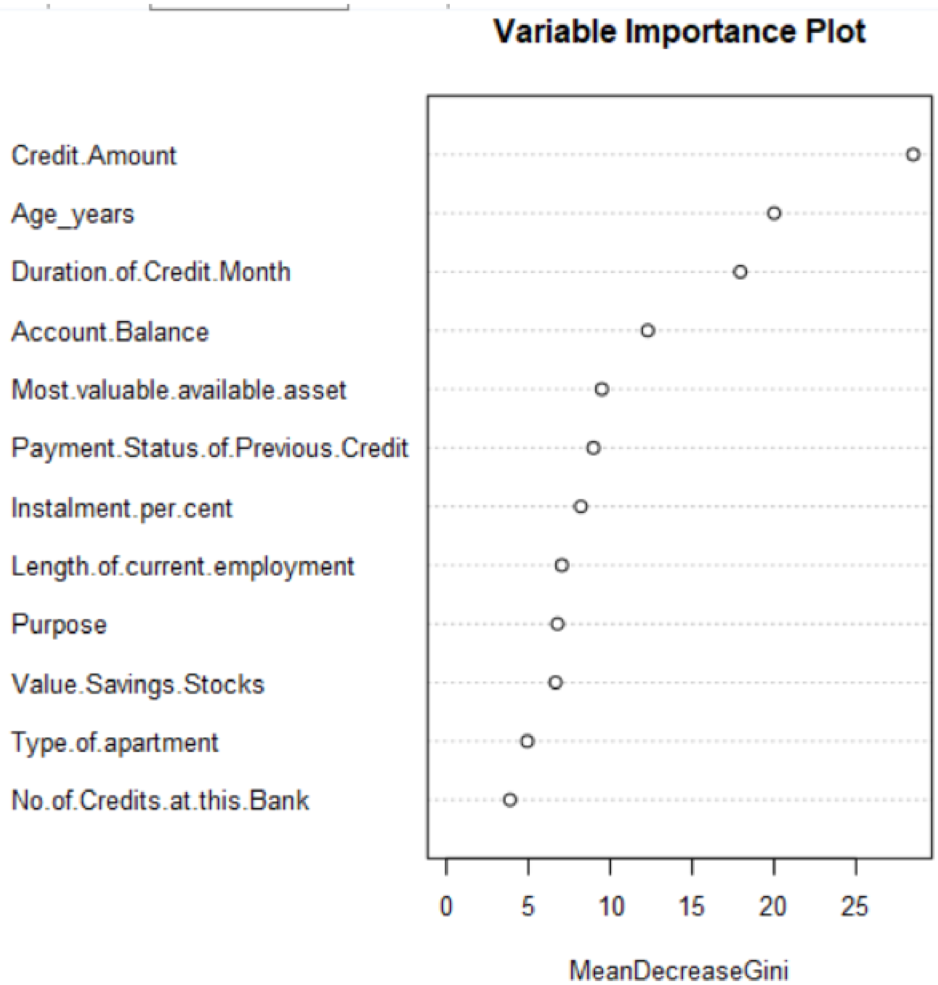
The overall accuracy of the model is 66% which is quite low, compared to Logistic Regression, accuracy of creditworthiness is 79% and the accuracy of non-creditworthiness is 37%.

This model has the same problem of being biased as there is a big gap between the accuracies of creditworthiness and non-creditworthiness. Here again the number of False Negatives is very high 14 in case when True Negatives is 20 (see confusion matrix).

Confusion matrix of creditworthy_DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	25
Predicted_Non-Creditworthy	14	20

Forest Model



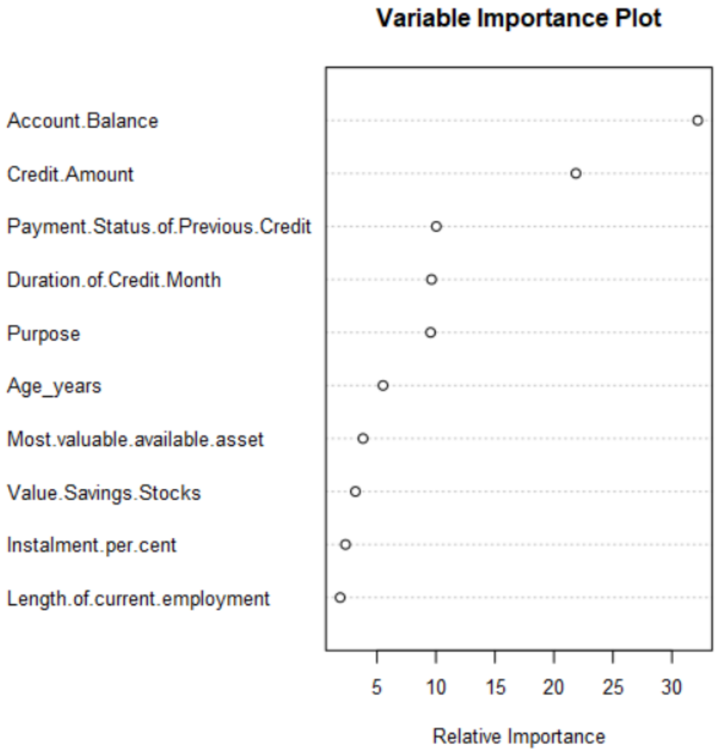
Here we can see, that the most important predictor variables are ‘Credit Amount’, then ‘Age years’, ‘Duration of Credit Month’ and ‘Account Balance’. The rest of the variables have relatively lower level of importance.

The overall accuracy of the model is 80%, including 97% percent of creditworthiness accuracy and 42% of non-creditworthiness accuracy. Although, here we also have a gap between the percentages of accuracies in case of predicting creditworthy and non-creditworthy customers, however, compared to previous two models the overall level of accuracy and the accuracy of creditworthiness is higher. And if we look at the confusion matrix, we can see that the number of False Negatives is too low, only 3.

Confusion matrix of creditworthy_FM

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	31
Predicted_Non-Creditworthy	3	14

Boosted Model



The most important variables in this model are 'Account Balance' and 'Credit Amount'. The variables "No of credits at this bank" and "Type of apartment' are not important at all for the Boosted Model.

The overall accuracy of the model is 78%, the accuracy of creditworthiness is 96% and the accuracy of non-worthiness is 37%. In case of this model the gap between creditworthy and non-creditworthy accuracies is the highest, which makes this model biased as well.

Confusion matrix of creditworthy_BM

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

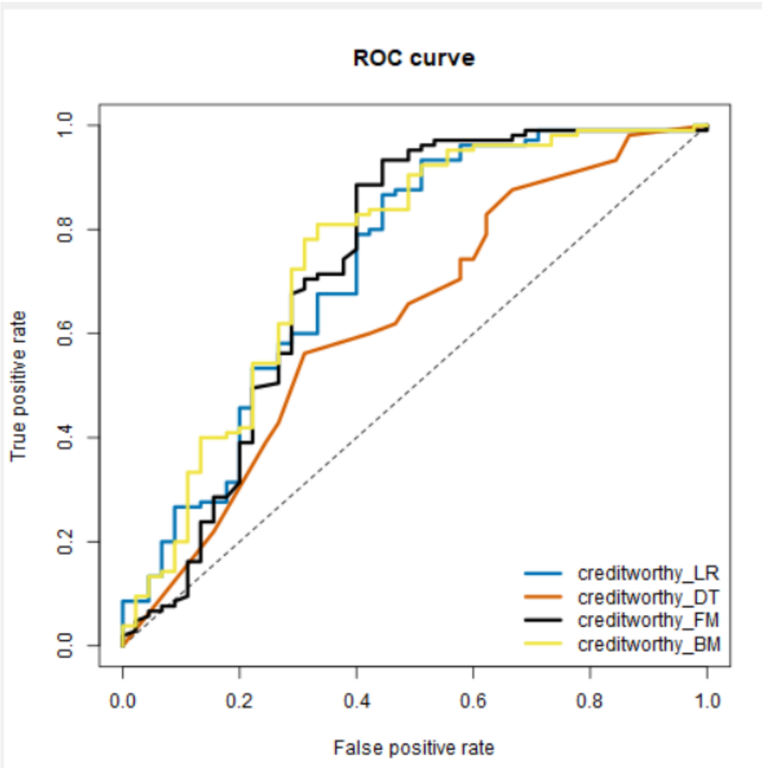
Step 4: Writeup

For the final analysis we use Union tool in Alteryx to compare all four models side by side.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
creditworthy_LR	0.7800	0.8520	0.7314	0.9048	0.4889
creditworthy_DT	0.6667	0.7685	0.6272	0.7905	0.3778
creditworthy_FM	0.8067	0.8755	0.7343	0.9714	0.4222
creditworthy_BM	0.7867	0.8632	0.7524	0.9619	0.3778

As we can see, all four models have relatively the same level of overall accuracy against our validation set, however, the Forest Model has the highest 80.67%, Logistic Regression and Boosted Model has slightly less 78% and Decision Tree has the lowest 66%. The highest level of creditworthy accuracy has again Forest Model, and with the non-creditworthy accuracy percentage it is on the second place with 42% after Logistic Regression model 48%. Based on the Fit and error measures report we can definitely exclude the Decision Tree model, as it has the lowest accuracy in all the categories, and if we compare Logistic Regression and Boosted Model, we still cannot decide which one is better as one is better at predicting creditworthy and the other non-creditworthy customers. So, we need to have a look at other reports, as ROC curve which is used to identify the quality of the model: the higher the positioning of the True Positive rate/False Positive rate curve (which means that the Area under the curve (AUC) will have a higher value) the better will be the model for analysis. As we can see the Forest Model and Boosted Model has the highest left positioning for most of the graph, maybe the Boosted Model has a slightly better one with higher AUC (0.75) compared to Forest Model (0.73)



Finally, comparing the confusion matrices of all four models we see that the best two models are Forest Model and Boosted Model. However, if we compare all four categories of confusion matrix totally, Forest Model has better accuracy than Boosted model.

Confusion matrix of creditworthy_BM

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

Confusion matrix of creditworthy_DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of creditworthy_FM

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of creditworthy_LR

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

Comparing all four models and corresponding reports, for our analysis we will use Forest Model to have more accurate results in predicting creditworthy customers.
After final analysis the total number of creditworthy customers is 409.

Below are Alteryx workflow screenshots made for this analysis: