Project 4

Creditworthiness

Predicting Default Risk

https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

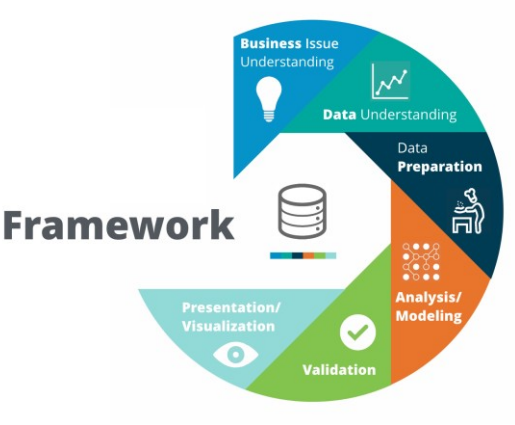
The Business Problem

- I am a loan officer at a small bank operating for two years that needs to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not.
- Generally my team typically receives 200 loan applications per week and approves them by hand.

- Due to a financial scandal that hit a competitive bank last week, I suddenly had an influx of new people applying for loans at our bank instead of the other bank in your city. All of a sudden I had nearly 500 loan applications to process this week.
- My manager sees this new influx as a great opportunity and wants me to figure out how to process all of these loan applications within one week.
- Fortunately I just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants.

Details

For this project, I will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to my manager in the next two days.



Problem Solving Framework

Skills required

In order to complete this project, I must be able to

- Cleanup, format, and blend a wide range of data sources.
- Build predictive classification models using *Logistic Regression*, *Decision Tree*, *Random Forest*, and *Boosted Model*.

Step 1, Business and Data Understanding

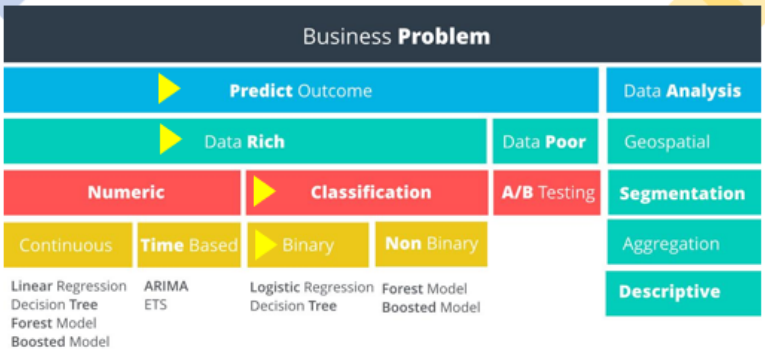
A description of key business decisions that need to be made. (250 word limit)

Key Decisions

What decisions needs to be made?

Key decision is providing to my manager the number of creditworthy applicants.

I will develop a list of creditworthy customers, summarizing it to my manager in the next two days. I will do that analyzing data related to past applications and to the list of customers that applied for a loan. The main purpose of business case is predicting whether loan applicants are creditworthy or non creditworhthy. This is a Binary Classification Business Model.



1. What data is needed to inform those decisions?

- *credit-data-training.xlsx*

  Credit approvals from past loan applicants the bank has ever completed. I present variables name and description.

	Variable Name	Description
1	Credit-Application-Result	If applicant is Creditworthy or Non-Creditworthy
2	Account-Balance	Account balance of the applicant: Some Balance, No Account
3	Duration-of-Credit-Month	Months of credit applied for
4	Payment-Status-of-Previous-Credit	Applicant status related to previous credit process: No Problems (in this bank), Paid Up or Some Problems
5	Purpose	Purpose for which loan is being taken: Home Related, New car or Used car
6	Credit-Amount	Credit amount applied for
7	Value-Savings-Stocks	Range of savings
8	Length-of-current-employment	Range of employment length: < 1yr, 1-4 yrs or4-7 yrs
9	Instalment-per-cent	Instalment percent: 1, 2, 3 or 4
10	Guarantors	Whether applicant has guarantors or not: None or Yes

	Variable Name	Description
11	Duration-in-Current-address	Time in current address
12	Most-valuable-available-asset	Most valuable asset available: 1, 2, 3 or 4
13	Age-years	Age in years
14	Concurrent-Credits	Concurrent Credits
15	Type-of-apartment	Type of aparment: 1, 2 or 3
16	No-of-Credits-at-this-Bank	Whether applicant has 1 or more than 1 credit at our Bank
17	Occupation	Occupation
18	No-of-dependents	Number of dependents
19	Telephone	Telephone: 1 or 2
20	Foreign-Worker	Whether applicant is foreign or not

- *customers-to-score.xlsx*

  Set of customers used to score classification model.

  Variables are the same, with the difference that there 19 instead of 20. The reason is that *Credit-Application-Result* is the expected result. That is the variable that will be predicted.

Step 2, Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

Here are some guidelines to help guide your data cleanup:

- 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".
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
- 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.
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

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.

There is no correlation greater then 0.7 for numerical data fields. Neither before nor after imputation of Age Years.

BEFORE Age Years imputation

Pearson Correlation Analysis

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Idade	Type.of.apartment
Duration.of.Credit.Month	1.000000	0.570441	0.079515	0.304734	-0.066319	0.153141
Credit.Amount	0.570441	1.000000	-0.285631	0.327762	0.068643	0.168683
Instalment.per.cent	0.079515	-0.285631	1.000000	0.078110	0.040540	0.082936
Most.valuable.available.asset	0.304734	0.327762	0.078110	1.000000	0.085437	0.379650
Idade	-0.066319	0.068643	0.040540	0.085437	1.000000	0.333075
Type.of.apartment	0.153141	0.168683	0.082936	0.379650	0.333075	1.000000

Matrix of Corresponding p-values

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Idade	Type.of.apartment
Duration.of.Credit.Month		0.0000e+00	7.9292e-02	6.0352e-12	1.4350e-01	6.8791e-04
Credit.Amount	0.0000e+00		1.2929e-10	1.1013e-13	1.2996e-01	1.8138e-04
Instalment.per.cent	7.9292e-02	1.2929e-10		8.4757e-02	3.7152e-01	6.7164e-02
Most.valuable.available.asset	6.0352e-12	1.1013e-13	8.4757e-02		5.9299e-02	0.0000e+00
Idade	1.4350e-01	1.2996e-01	3.7152e-01	5.9299e-02		4.1744e-14
Type.of.apartment	6.8791e-04	1.8138e-04	6.7164e-02	0.0000e+00	4.1744e-14	

AFTER Age Years imputation

Full Correlation Matrix

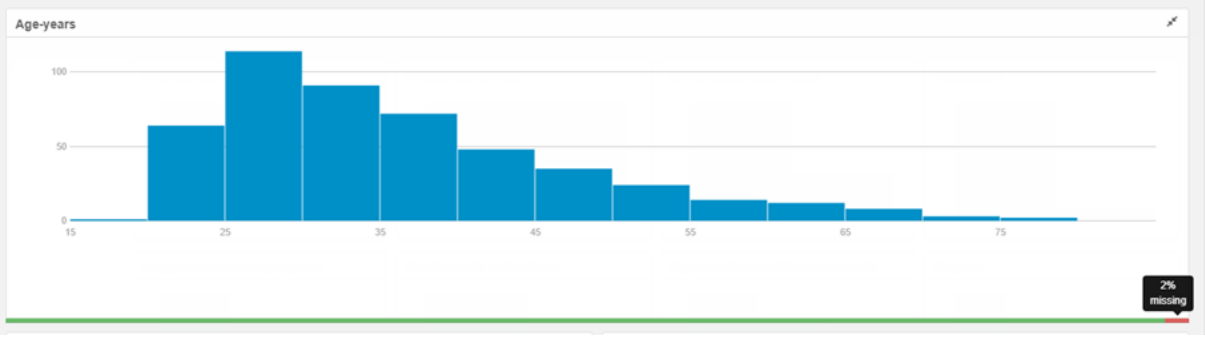
	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Idade	Type.of.apartment
Duration.of.Credit.Month	1.000000	0.573980	0.068106	0.299855	-0.065320	0.152516
Credit.Amount	0.573980	1.000000	-0.288852	0.325545	0.068110	0.170071
Instalment.per.cent	0.068106	-0.288852	1.000000	0.081493	0.040187	0.074533
Most.valuable.available.asset	0.299855	0.325545	0.081493	1.000000	0.083642	0.373101
Idade	-0.065320	0.068110	0.040187	0.083642	1.000000	0.327461
Type.of.apartment	0.152516	0.170071	0.074533	0.373101	0.327461	1.000000

Matrix of Corresponding p-values

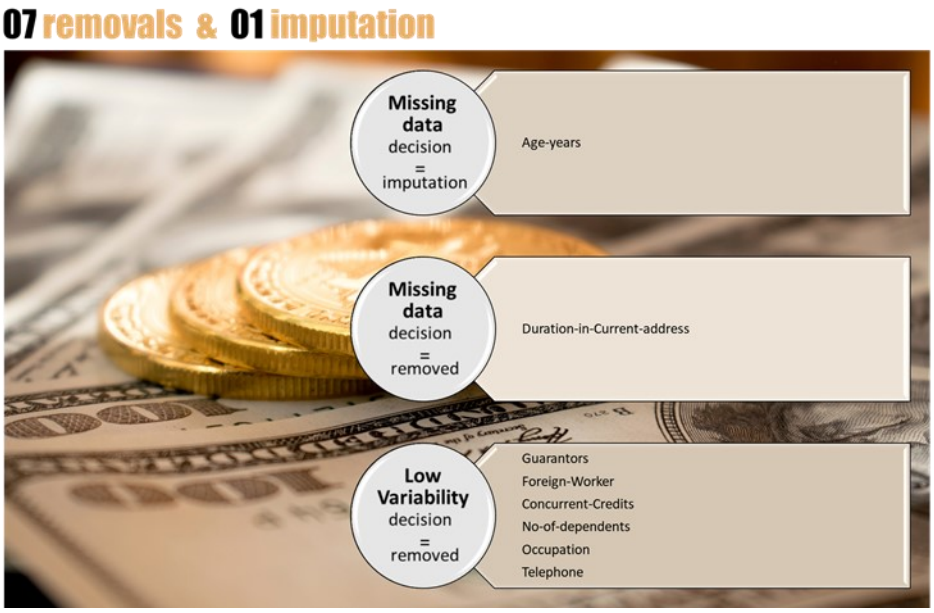
	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Idade	Type.of.apartment
Duration.of.Credit.Month		0.0000e+00	1.2830e-01	7.5764e-12	1.4470e-01	6.2192e-04
Credit.Amount	0.0000e+00		4.5919e-11	8.3045e-14	1.2828e-01	1.3277e-04
Instalment.per.cent	1.2830e-01	4.5919e-11		6.8653e-02	3.6987e-01	9.5961e-02
Most.valuable.available.asset	7.5764e-12	8.3045e-14	6.8653e-02		6.1639e-02	0.0000e+00
Idade	1.4470e-01	1.2828e-01	3.6987e-01	6.1639e-02		5.8176e-14
Type.of.apartment	6.2192e-04	1.3277e-04	9.5961e-02	0.0000e+00	5.8176e-14	

For this project the instruction was removing 7 variables from the dataset, based on those that have low variability, that have unique value or that are irrelevant to the case. I analyzed variable distribution using *Field Summary Tool* at O and I outputs.

Age-years There were 2% missing values. For those I did imputation using 36 rounded as median as oriented. If had used the average here, the values would be a bit less accurate. Imputation was developed because there were missing data and one of the more basic reasons why we should care about missing data is that some statistical algorithms just won´t if there are values missing. And missing values can add bias to results. Observation: I changed field name Age-years to *Idade*.



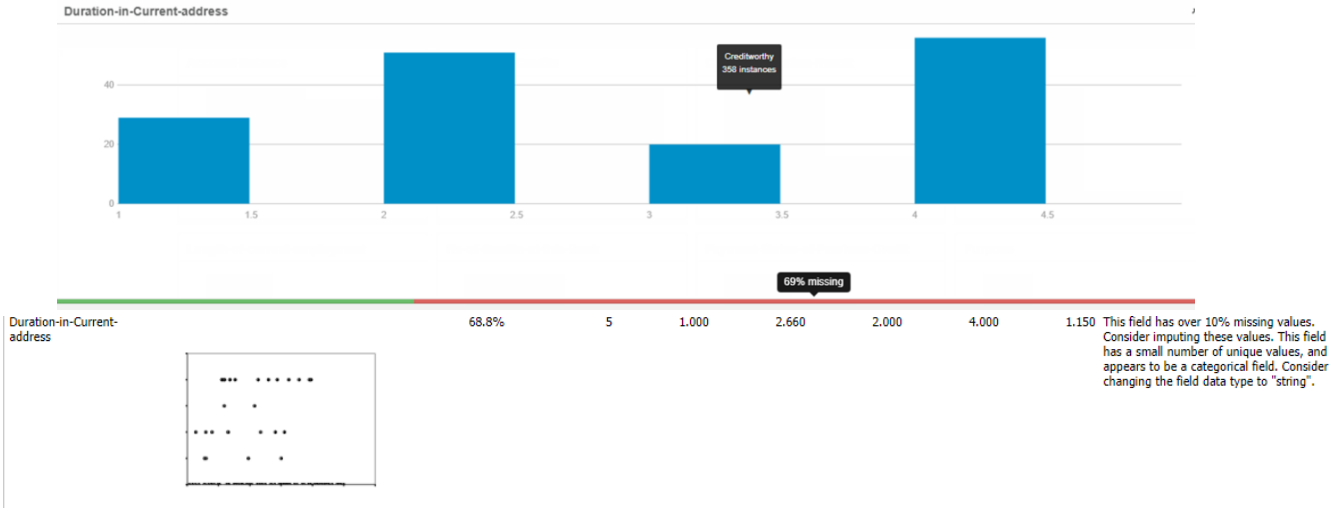
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	



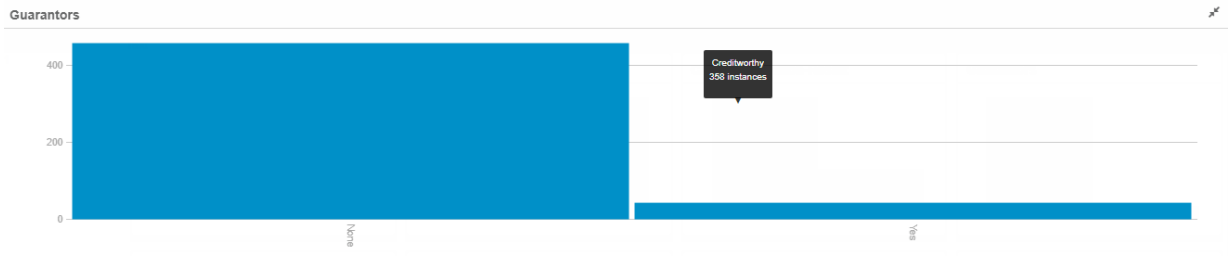
I decided to remove seven variables using Select Tool.

① Duration-in-Current-address

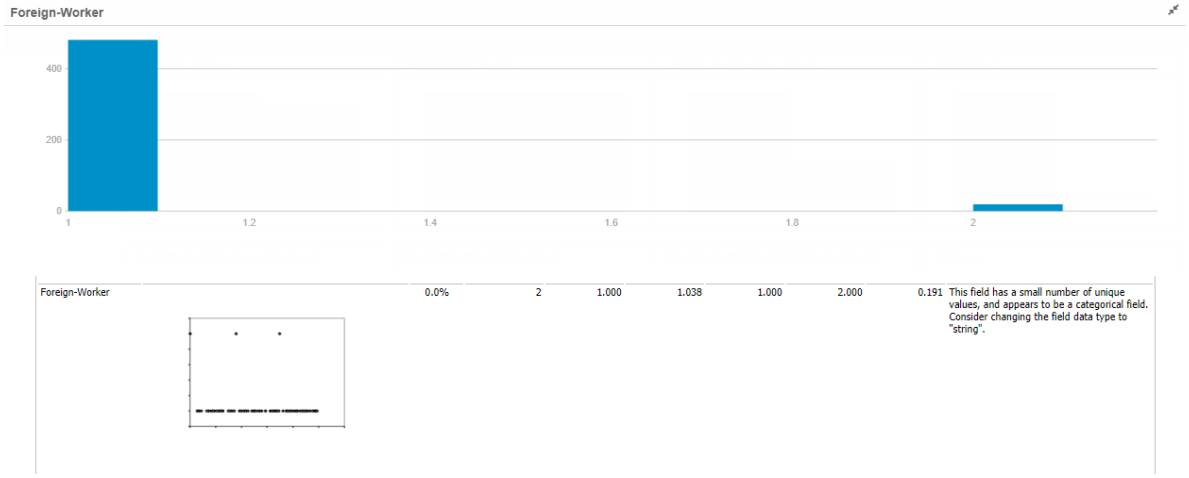
68.6% of data is missing. Decided to remove the data for the entire variable.



② Guarantors, 457 x 43, low variability. Data is skewed towards *None*.



③ Foreign-Worker, 481 x 19 = low variability. This field has a small number of unique values. Data is skewed towards *1*.



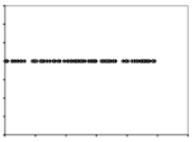
④ Concurrent-Credits, 500 = low variability. Field has only one value.

⑤ No-of-Dependents, 427 x 73 = low variability. The majority of fata is skewed to 1.

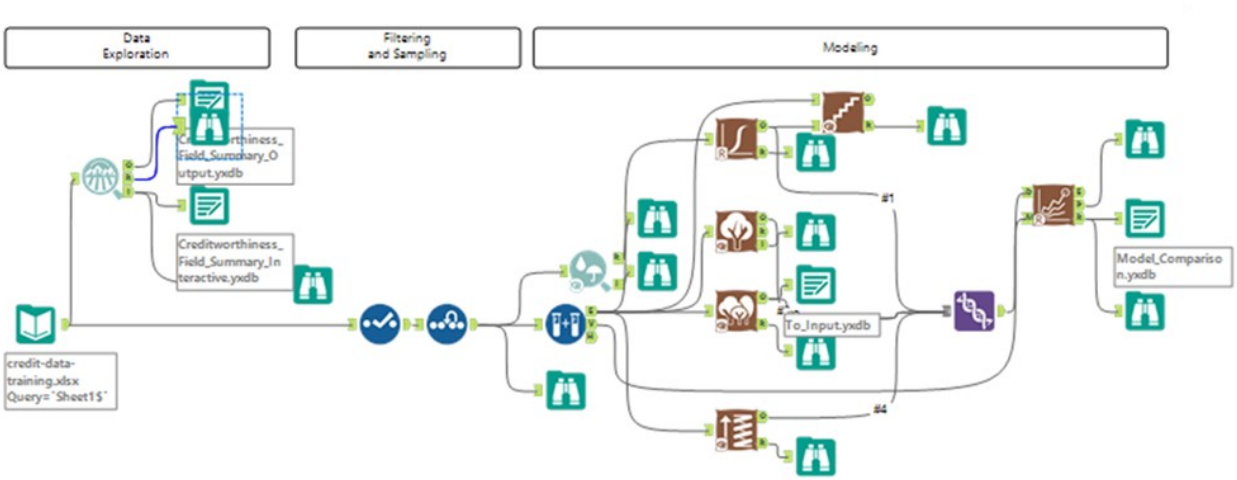
⑥ Telephone, 300 x 20 = low variability. Removed because presents low variability and doesn´t explain anything about the creditworthiness of the applicant.

⑦ Occupation, low variability. Field only has value 1.

Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".



I built the following Alteryx workflow.



Step 3, Train your Classification Models

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.

- 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?

You should have four sets of questions answered. (500 word limit)

- I built four different models aiming to perform a comparative analysis using Alteryx: Logistic, Decision Tree, Forest and Boosted Model running them over prepared sampled data.

- I connected all model outputs to a Union Tool.

- Afterwards I used Model Comparison Tool in order to evaluate which one performed best
  - validated models
  - compared accuracy and
  - studied significant variables or the most important variables.

Following the *Model Comparison Report* a summary of *Accuracy*.

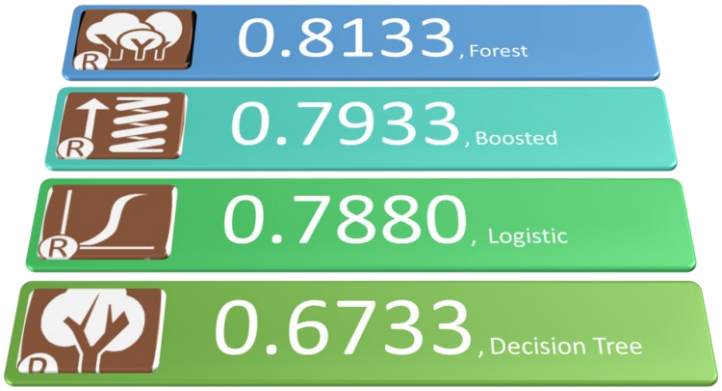
Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic.Model	0.7800	0.8520	0.7302	0.9048	0.4889
Decision.Tree.Model	0.6733	0.7721	0.6296	0.7905	0.4000
Forest.Model	0.8133	0.8793	0.7399	0.9714	0.4444
Boosted.Model	0.7933	0.8670	0.7539	0.9619	0.4000

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.



Model´s Overall Accuracy

Following the *Confusion Matrixes* for Boosted, Decision Tree, Forest and Logistic Models.

Confusion matrix of Boosted.Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of Decision.Tree.Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of Forest.Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

Confusion matrix of Logistic.Model

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

Logistic Regression Model

Logistic regression model ranked third with overall accuracy of 79%, with the following significant variables.

		Significant Variables	
Account.BalanceSome Balance	***	1	Account-Balance
Payment.Status.of.Previous.CreditSome Problems	*	2	Payment.Status.of.Previous-Cred
PurposeNew car	**	3	Purpose
PurposeUsed car	*		
Credit.Amount	**	4	Credit-Amount
Instalment.per.cent	*	5	Instalment.per.cent
Most.valuable.available.asset	*	6	Most.valuable.available.asset

Confusion matrix of Logistic.Model

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

I concluded that model is biased to predict that applicants are Creditworthy.

	Actual Creditworthy	Actual Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

PPV, Positive Predictive Value	NPV, Negative Predictive Value
81%	69%

Model is biased towards Crediworthy.
NPV and PPV are not close to each other.

12%

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ 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 + Idade + Type.of.apartment + No.of.Credits.at.this.Bank, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.083	-0.719	-0.429	0.691	2.543

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9876970	1.013e+00	-2.9495	0.00318	**
Account.BalanceSome Balance	-1.5437093	3.233e-01	-4.7744	1.80e-06	***
Duration.of.Credit.Month	0.0063762	1.371e-02	0.4650	0.64195	
Payment.Status.of.Previous.CreditPaid Up	0.4026406	3.843e-01	1.0478	0.29472	
Payment.Status.of.Previous.CreditSome Problems	1.2595449	5.334e-01	2.3614	0.01821	*
PurposeNew car	-1.7552103	6.279e-01	-2.7955	0.00518	**
PurposeOther	-0.2858883	8.362e-01	-0.3419	0.73243	
PurposeUsed car	-0.7858570	4.124e-01	-1.9055	0.05672	.
Credit.Amount	0.0001771	6.842e-05	2.5890	0.00963	**
Value.Savings.StocksNone	0.6095558	5.099e-01	1.1954	0.23193	
Value.Savings.Stocks£100-£1000	0.1726278	5.650e-01	0.3056	0.75994	
Length.of.current.employment4-7 yrs	0.5321240	4.932e-01	1.0790	0.28059	
Length.of.current.employment< 1yr	0.7772931	3.957e-01	1.9644	0.04948	*
Instalment.per.cent	0.3104587	1.399e-01	2.2192	0.02648	*
Most.valuable.available.asset	0.3255698	1.557e-01	2.0915	0.03649	*
Idade	-0.0152253	1.539e-02	-0.9892	0.32258	
Type.of.apartment	-0.2537778	2.958e-01	-0.8578	0.391	
No.of.Credits.at.this.BankMore than 1	0.3627862	3.816e-01	0.9508	0.34172	

Confusion matrix of Logistic.Model

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

Decision Tree

Decision tree model came in last place in accuracy resulting in 67%.

Top three most important variables are Account-Balance, Duration-of-Credit-Month and Credit-amount.

I concluded that model is biased to predict that applicants are Creditworthy.

Confusion matrix of Decision.Tree.Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

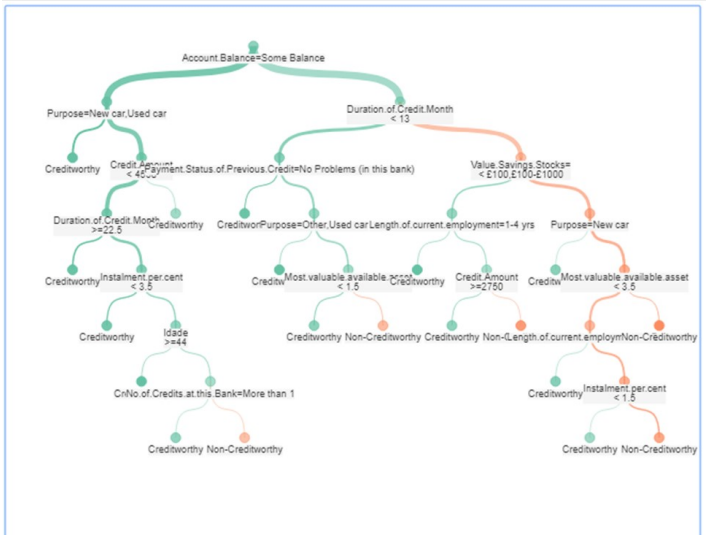
	Actual Creditworthy	Actual Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

PPV, Positive Predictive Value	NPV, Negative Predictive Value
75%	45%

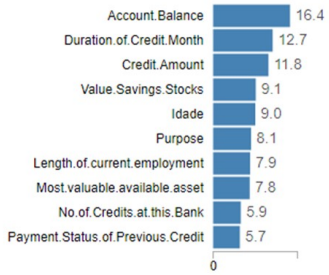
Model is biased towards Crediworthy.
NPV and PPV are not close to each other.

30%

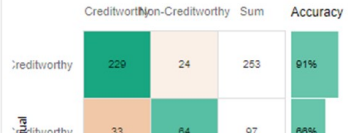
Decision Tree



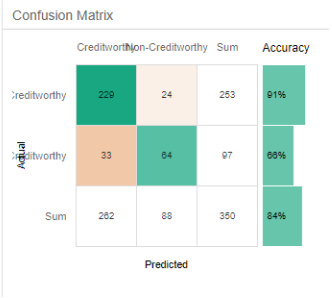
Variable Importance



Confusion Matrix



Confusion Matrix



Forest Model

The overall accuracy of Forest Model is 81% which is strong and the best performance of the four models analyzed.

Based on the variable importance plot the three most important variables are Credit_Amount, Idade (Age-Years) and Duration-of-Credit-Month.

Model is unbiased.

Confusion matrix of Forest.Model

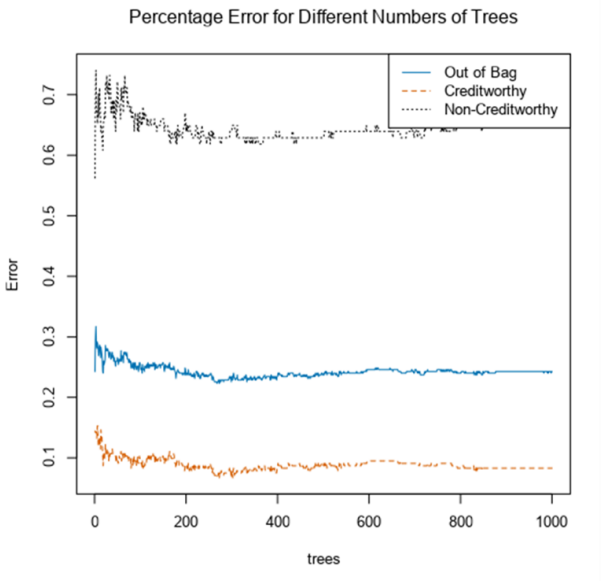
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

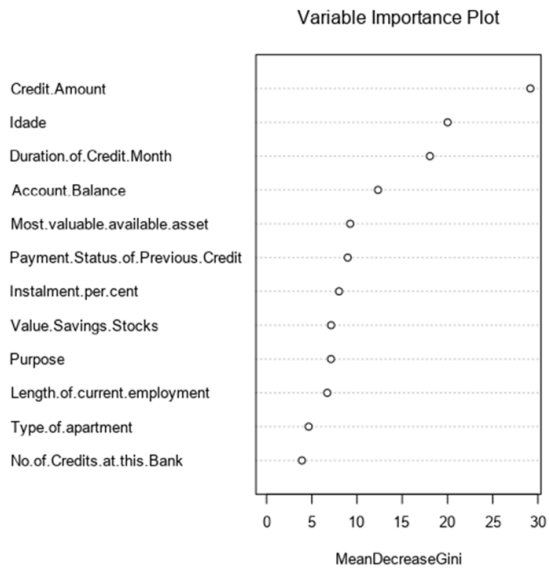
	Actual Creditworthy	Actual Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

PPV, Positive Predictive Value	NPV, Negative Predictive Value
80%	87%

Model is unbiased.
NPV and PPV are close to each other.

7%





Boosted Model

The accuracy of Boosted Model is 79%, the 2nd best performance in relation to the four models tested.

Based on the variable importance plot, the top 3 important predictive variables are Amount-Balance, Credit-Amount and Duration-of-Credit-Month.

Model is unbiased.

Confusion matrix of Boosted.Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

	Actual Creditworthy	Actual Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

PPV, Positive Predictive Value	NPV, Negative Predictive Value
79%	82%

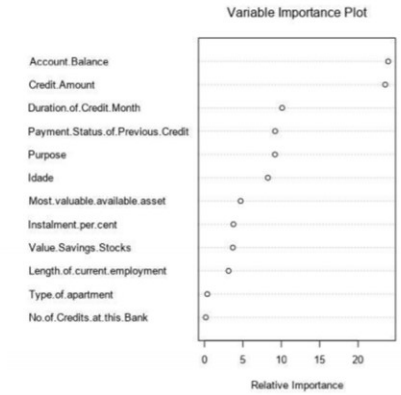
Model is unbiased.
NPV and PPV are close to each other.

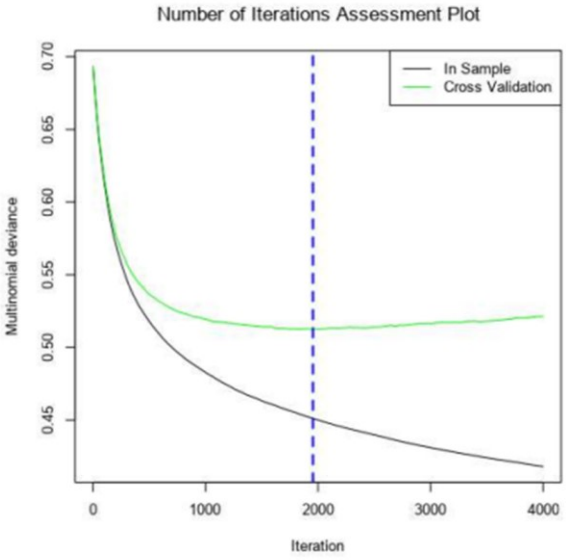
3%

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1955

Plots:





Step 4, Writeup

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)
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:
    - Overall Accuracy against your Validation set.
    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    - ROC graph
    - Bias in the Confusion Matrices

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.

I decided using Forest Model based on model comparison report.
Since my "manager only cares about overall accuracy" I decided by Forest Model, which resulted in the highest overall accuracy = 0.8133.

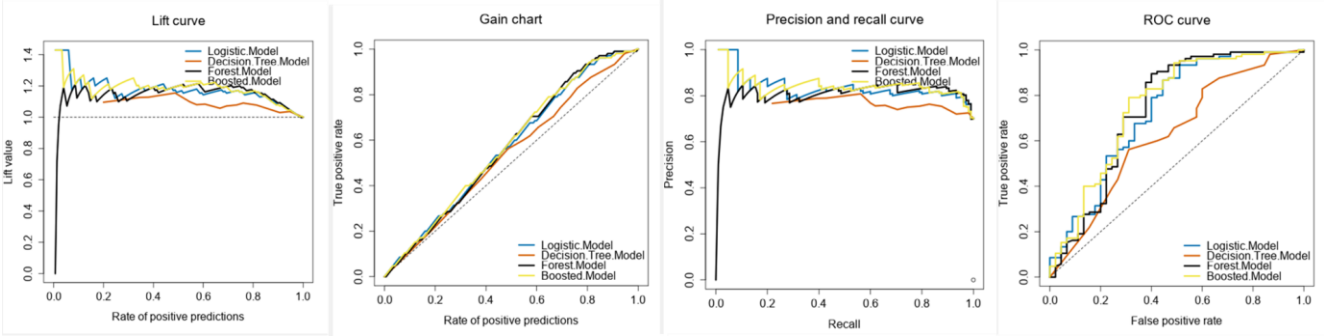
Forecast model stands out! I analyzed PPV, NPV, ROC curve and Gain chart. At ROC curve and Gain chart, Forest model is represented by the black line and is with high performance facing the other models. Forest model is unbiased. Logistic Regression, Boosted and Decision Tree presented biased. Forest model also has the highest F1 (0.8793).

Layout

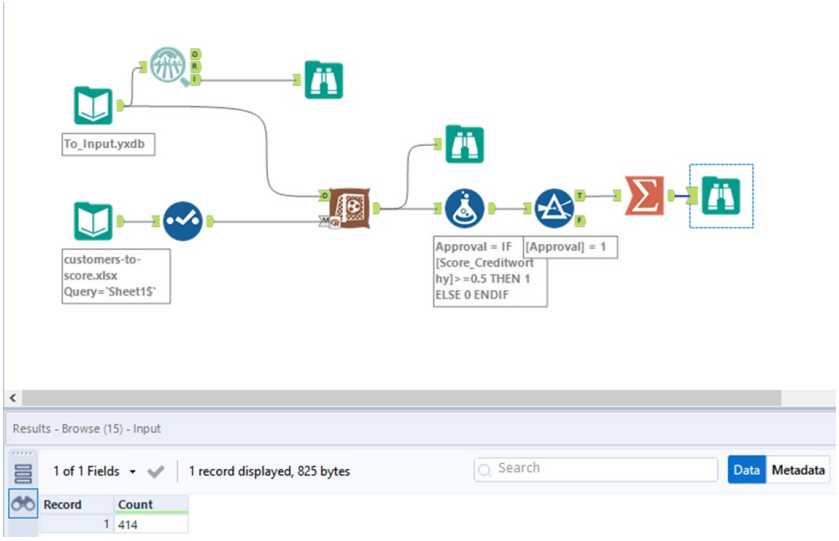
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic.Model	0.7800	0.8520	0.7302	0.9048	0.4889
Decision.Tree.Model	0.6733	0.7721	0.6296	0.7905	0.4000
Forest.Model	0.8133	0.8793	0.7399	0.9714	0.4444
Boosted.Model	0.7933	0.8670	0.7539	0.9619	0.4000



I developed the following Alteryx workflow to predict applicants creditworthiness, based on data generated on the previous one.



Considering calculations performed by workflow,

414

out of 500

are creditworthy

= 83%.