Project: Creditwrthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?

I work in a small bank and am responsible for determining if customers are creditworthy to give a loan to. There is an influx of nearly 500 loan applications to process this week and my manager asked me how to proceed.

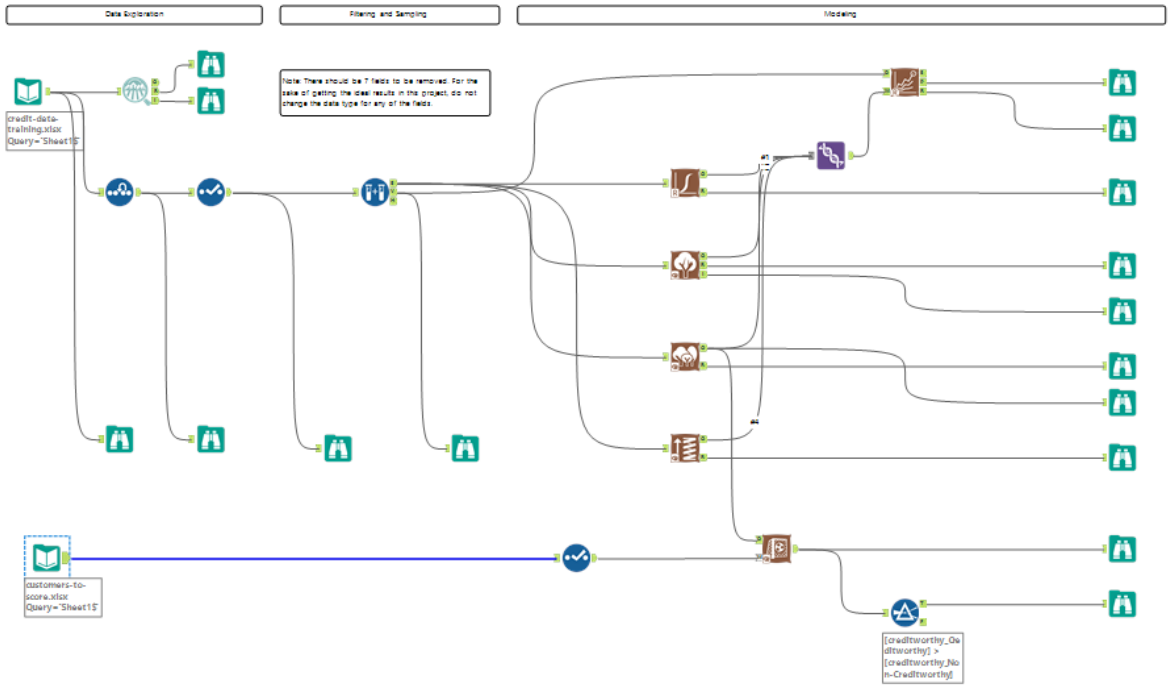
- What data is needed to inform those decisions?

Data on all past applications and data of customers that need to be processed.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

Considering that the options qualify and no-qualify, in other words only two outcomes then the best model is Binary.

Alterys model



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)*

Note: *For students using software other than Alteryx, please format each variable as:*

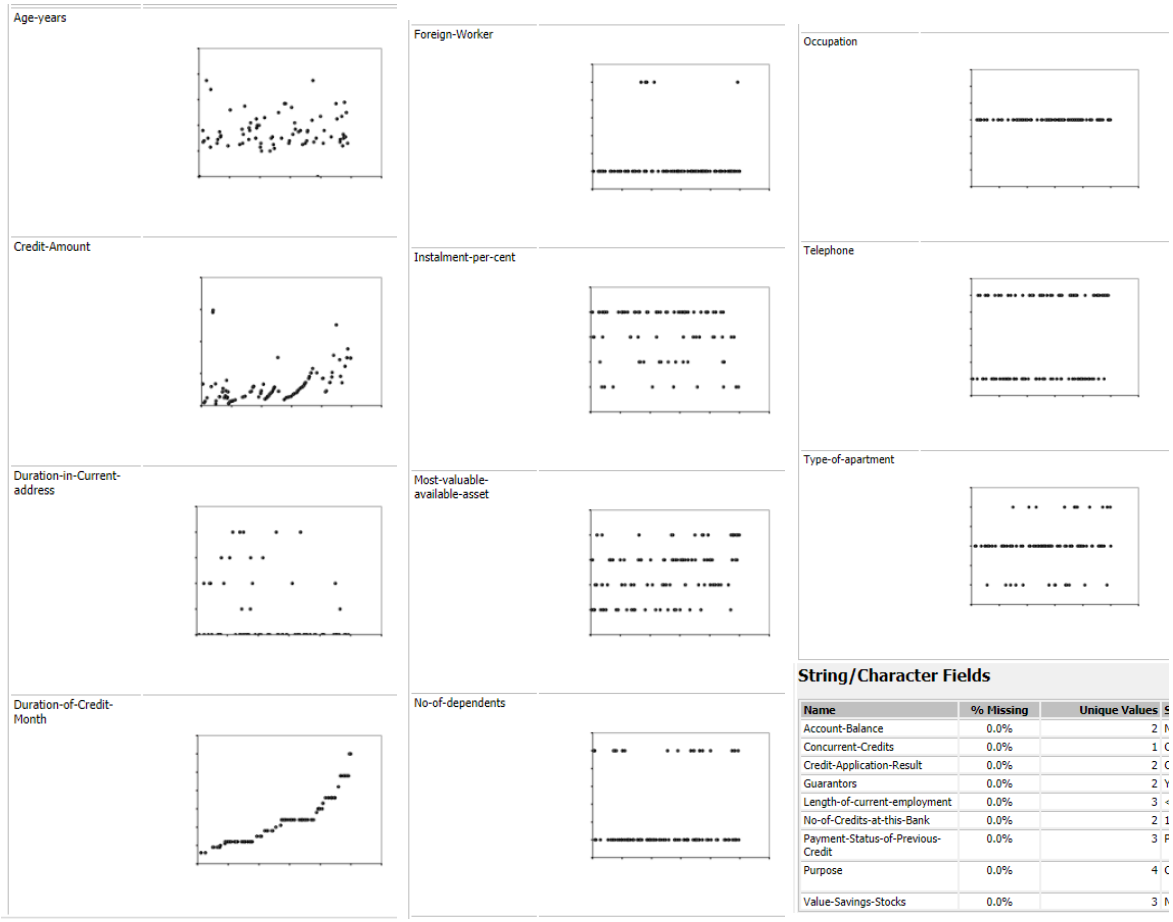
Variable	Data Type
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

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.

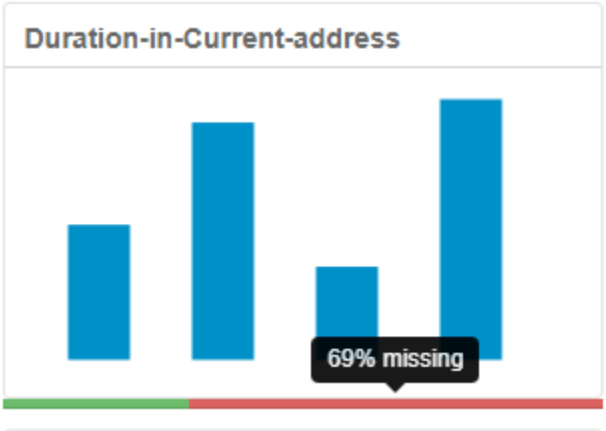
Using the field summary tool is possible to get a better view of the data and analyze each field independently to evaluate if it is worth it to include in the analysis.



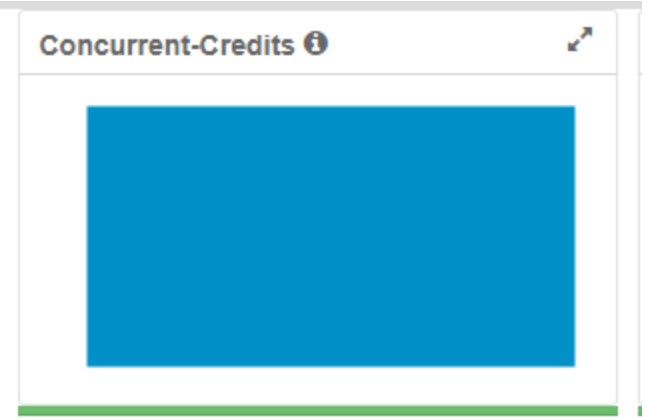
String/Character Fields

Name	% Missing	Unique Values
Account-Balance	0.0%	2
Concurrent-Credits	0.0%	1
Credit-Application-Result	0.0%	2
Guarantors	0.0%	2
Length-of-current-employment	0.0%	3
No-of-Credits-at-this-Bank	0.0%	2
Payment-Status-of-Previous-Credit	0.0%	3
Purpose	0.0%	4
Value-Savings-Stocks	0.0%	3

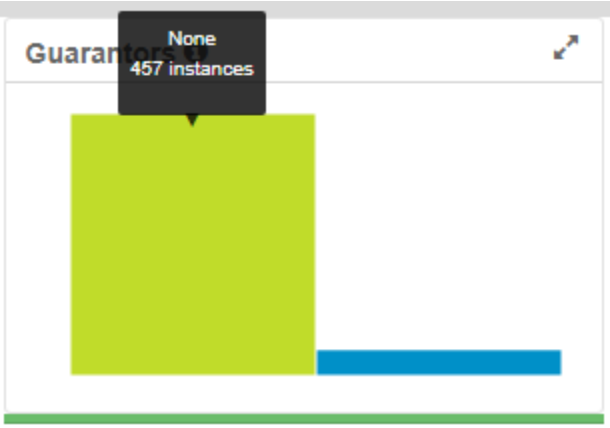
Duration of Current Address
as we can see in the image below from the records it is is missing 69% this attribute, considering this immuting would not help due to the high amount of missing data.



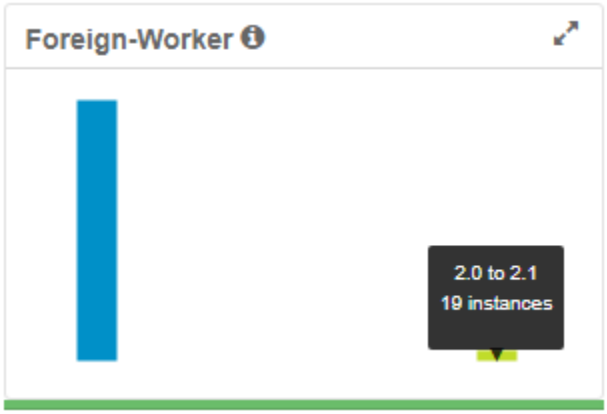
Concurrent Credits
As we can see in the image below all the 500 instances of the data have the same value, therefore this cannot help to predict if the client qualifies or not-qualifies. Therefore it would be convenient to remove it.



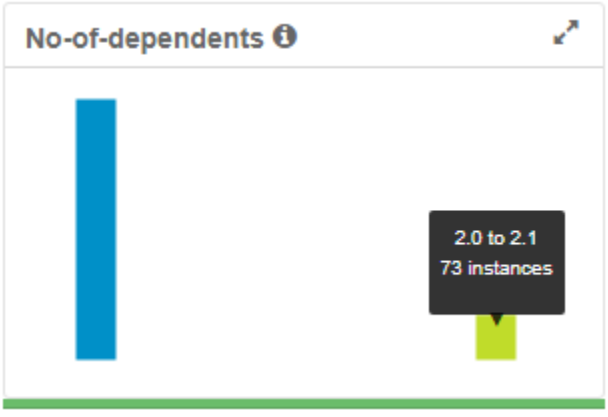
Guarantors
Even though this field has no missing data, it has the value of none for 457 out 500 instances, that would not help to predict if the client qualifies or no-qualifies, therefore it would be convenient to remove it.



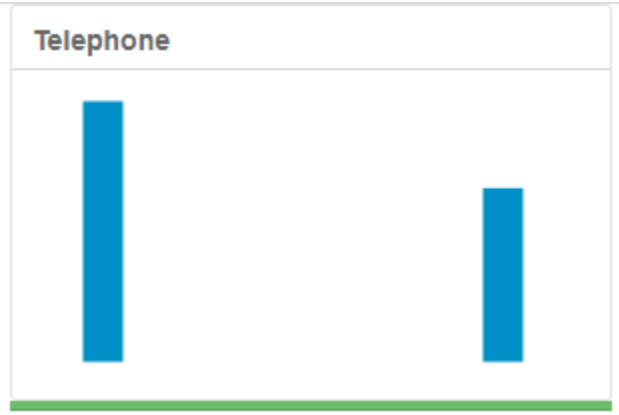
Foreign-Worker
Even though this field has no missing data, it has the value of '1' for 481 out 500 instances, that would not help to predict if the client qualifies or no-qualifies, therefore it would be convenient to remove it.



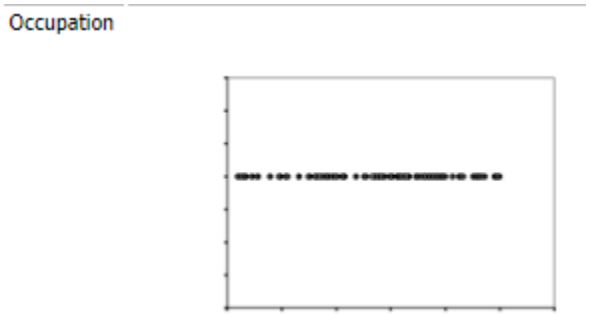
No-of-dependents
Even though this field has no missing data, it has the value of '1' for 427 out 500 instances, that would not help to predict if the client qualifies or not-qualifies. therefore it would be convenient to remove it.



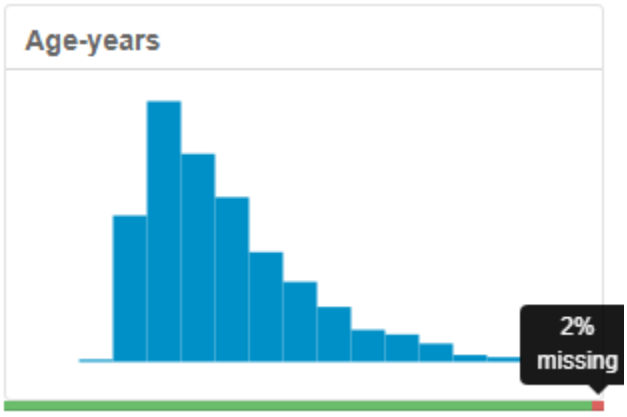
Telephone
Even though this field has no missing data due to the nature of the data, it does not help to predict if the client qualifies or not-qualifies. Therefore it would be convenient to remove it.



Occupation
As we can see in the image below all the 500 instance of the data have the same value ‘1’, therefore this cannot help for to predict if the client qualify or no-qualify., therefore it would be convenient to remove it



Age-years
It is important to mention that even though Age years is missing 2% of the inputs this can be balanced with the use of mutation to replace nulls with the median.



After the reviewing each field, it is possible to conclude that the from the 20 fields we could remove 7 that would not help to create a efficient model, these fields to remove are:

1. Duration of Current Address
2. Concurrent Credits
3. Guarantors
4. Foreign-Worker
5. No-of-dependents
6. Telephone
7. Occupation

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)

Logistic Regression

It is possible to identify in the report below the significant predictive variables are Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Length of Current Employment, Installment per Cent and Most valuable available asset.

Report for Logistic Regression Model AO_logistic

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 + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years_ImputedValue, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	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_ImputedValue	-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)

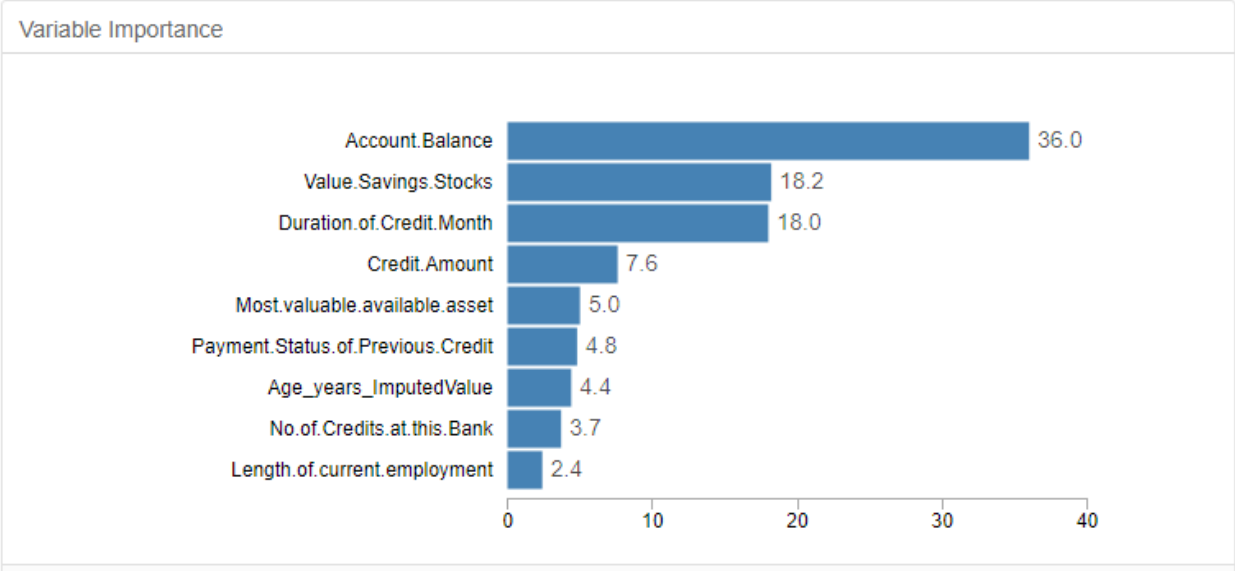
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Decision Tree

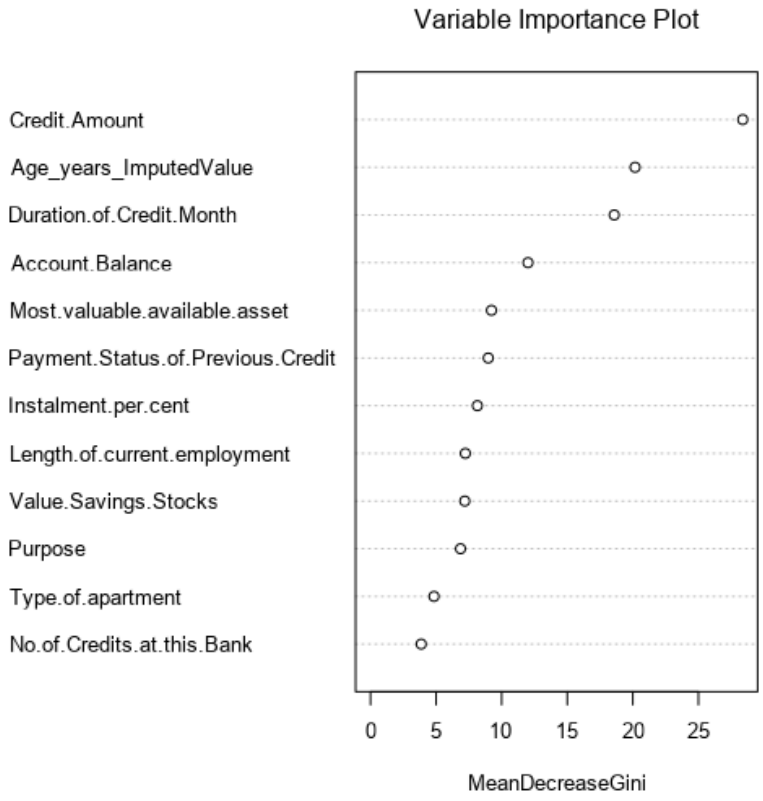
It is possible to identify in the graph of variable importance below the top 3 predictive variables are Account Balance, Value Savings Stocks, and Duration of Credit Month.



Forest Model

It is possible to identify in the graph below, the 3 most relevant predictive variables are Credit Amount, Age Years, and Duration of Credit Month.

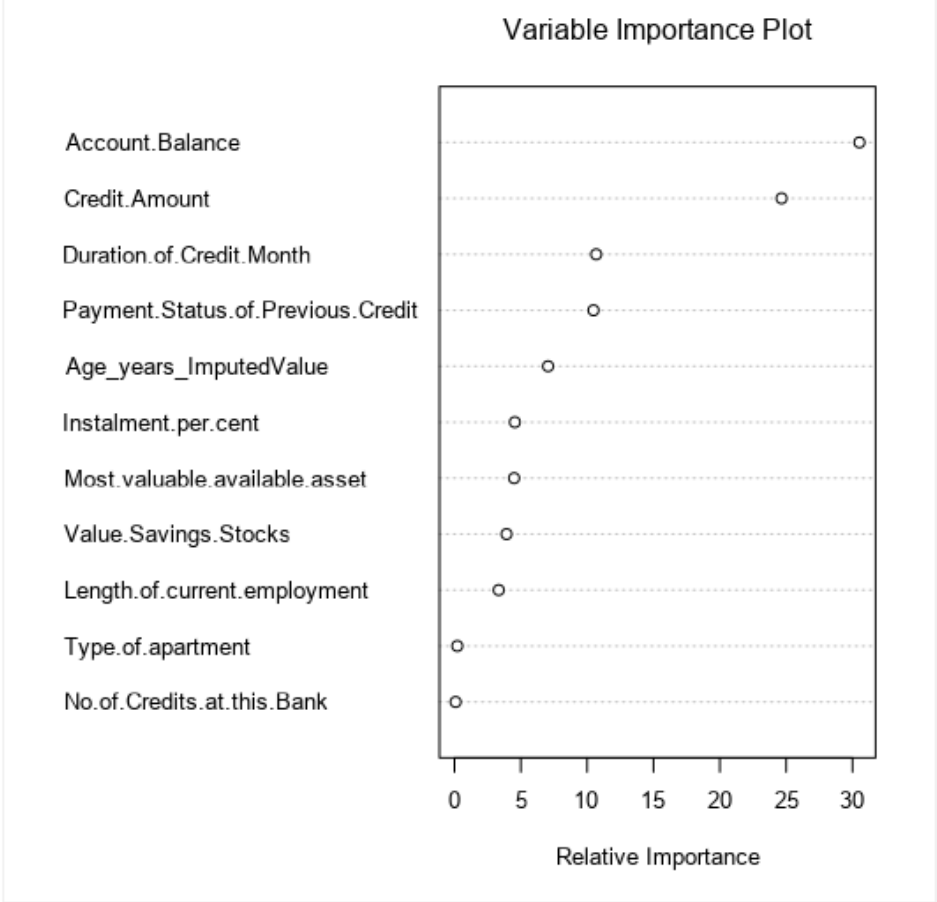
9



Boosted Model

It is possible to identify in the graph below, the 3 most relevant predictive variables are Account Balance, Credit Amount, and Duration of Credit Month.

2 Plots:



By using the Model Comparison Tool it is possible to get the report below that allows to compare the accuracy of the multiple models. The accuracy for the models is as follow:

- Logistic Regression model is 0.7800.
- Decision Tree model is 0.7467.
- Forest Model is 0.8000.
- Boosted Model is 0.7933.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
AO_logistic	0.7800	0.8520	0.7314	0.8051	0.6875
AO_DecisionTree	0.7467	0.8273	0.7054	0.7913	0.6000
AO_Forest	0.8000	0.8707	0.7361	0.7953	0.8261
AO_Boosted	0.7933	0.8681	0.7522	0.7846	0.8500

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 AO_Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of AO_DecisionTree

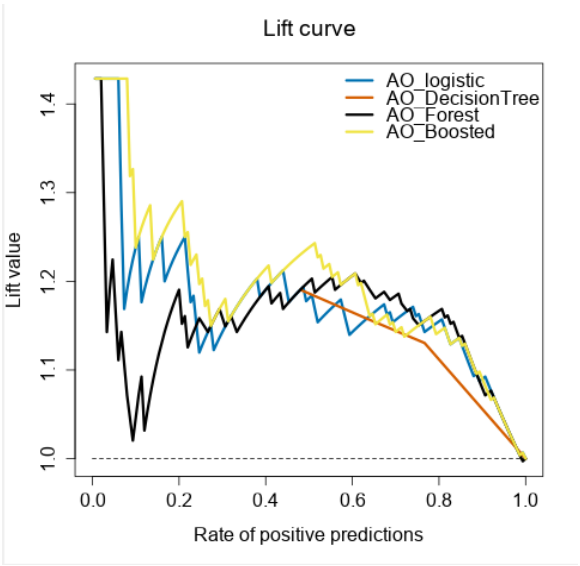
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

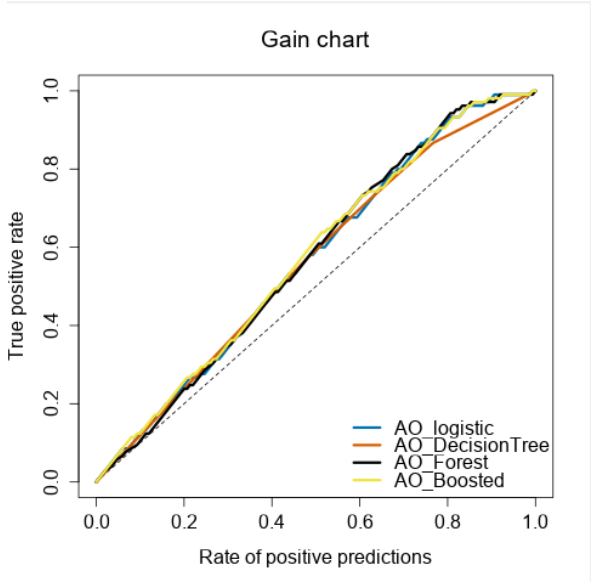
Confusion matrix of AO_Forest

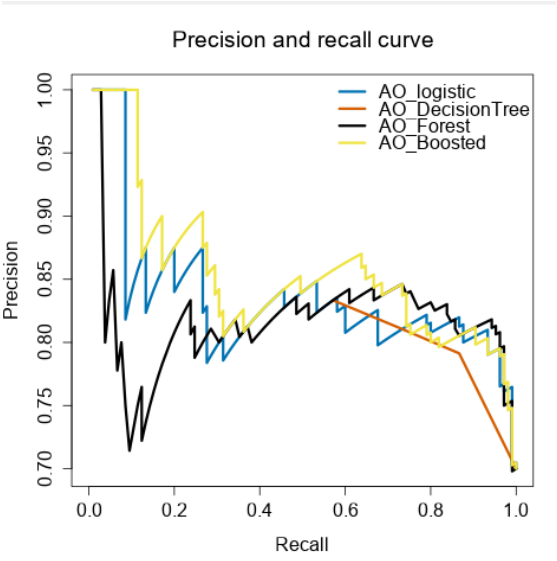
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

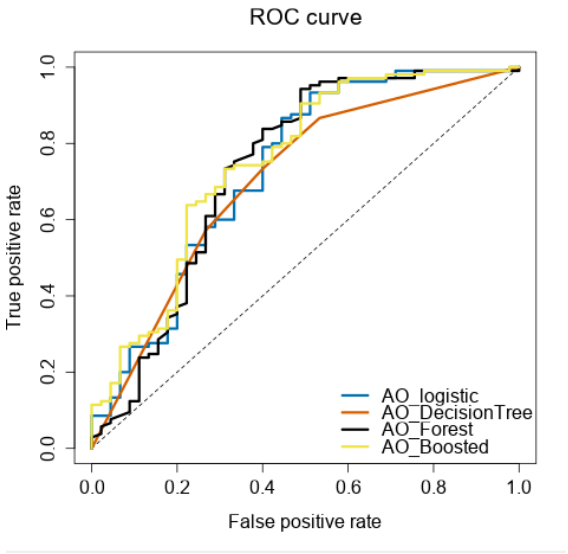
Confusion matrix of AO_logistic

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









It is possible to identify in the report and graphs before that there is bias considering thgat there are many creditworthy applicants than non-creditworthy applicants therefore we need to look for the model that offers more accuracy. the F1 or precision measure that is the percentage of actual members that were predicted to be divided by the total number of cases predicted will allows to select the model that is more accurate for this analysis, in this case it is the Forest model that get 0.8707 of F1 score precision and it is above all other models.

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:

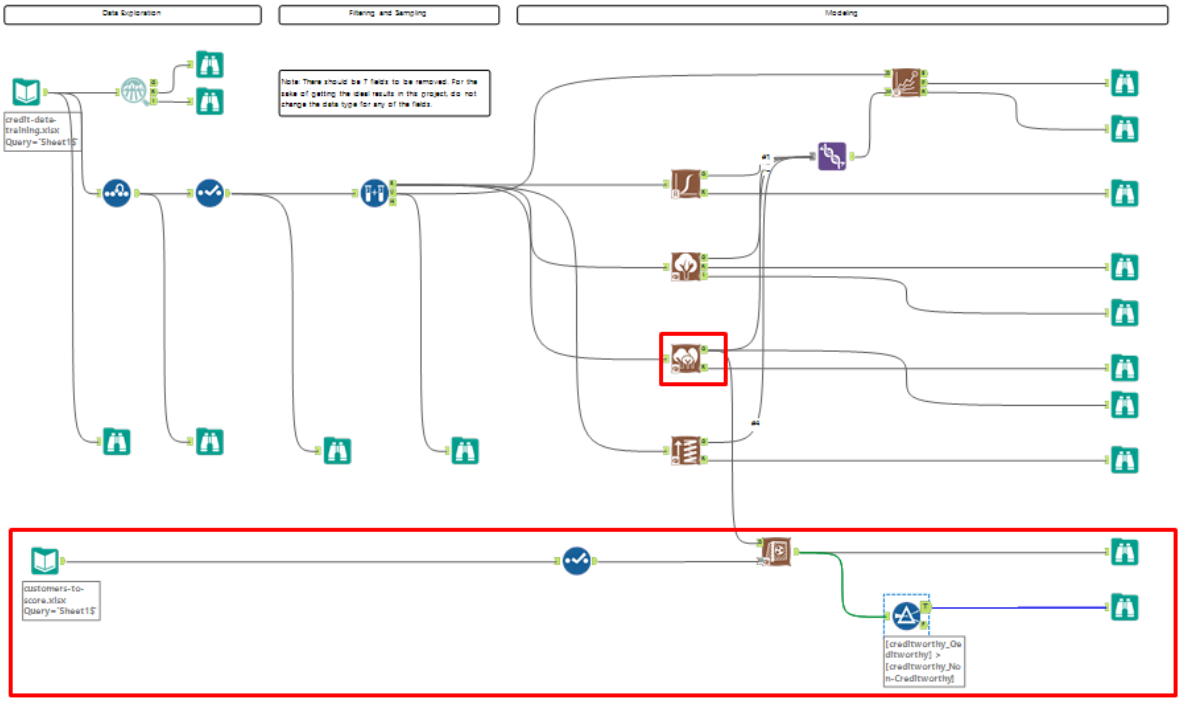
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    - ROC graph
    - Bias in the Confusion Matrices

It is possible to identify in the report and graphs in the previous point, that there is bias considering that there are many creditworthy applicants than non-creditworthy applicants therefore we need to look for the model that offers more accuracy. the F1 or precision measure that is the percentage of actual members that were predicted to be divided by the total number of cases predicted will allows to select the model that is more accurate for this analysis, in this case it is the Forest model that get 0.8707 of F1 score precision and it is above all other models.

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

- How many individuals are creditworthy?

By using the score tool with the customers-to-score data and the forest model, then filtering the output, we get a result of 406 creditworthy individuals.



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.