1 Business and Data Understanding

" " " " "

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand. Due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week! Your manager sees this new influx as a great opportunity and wants you to figure out how to process all of these loan applications within one week. Fortunately for you, you just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants. For this project, you will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to your manager in the next two days. You have the following information to work with:

1. *Data on all past applications*
2. *The list of customers that need to be processed in the next few days*

" " " " "

From the above business problem, the main purpose of this analysis project is to find the creditworthiness of the customer from a list of 500 applicants who had applied for a loan from the bank where I'm working at.

1.1 AIM

The main aim is to reduce the amount of time taken for selecting the creditworthy customers from a pile of applications.

1.2 Key Decision Need to Be Made

To sort out whether an applicant is worthy enough to provide him with a loan or not.

Since, our goal is to find whether a customer is creditworthy or not (to classify customers under two different categories), it clearly comes under the category of binary classification models.

2 Explore and Cleanup the Data

""""

To properly build the model, and select predictor variables, you need to explore and cleanup your data.

Here are some guidelines to help you clean up the data:

1. *Are any of your numerical data fields highly-correlated with each other? The correlation should be at least .70 to be considered "high".*
2. *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed*
3. *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.*
4. *Your clean data set should have 13 columns where the Average of Age Years should be 36 (rounded up)*

""""

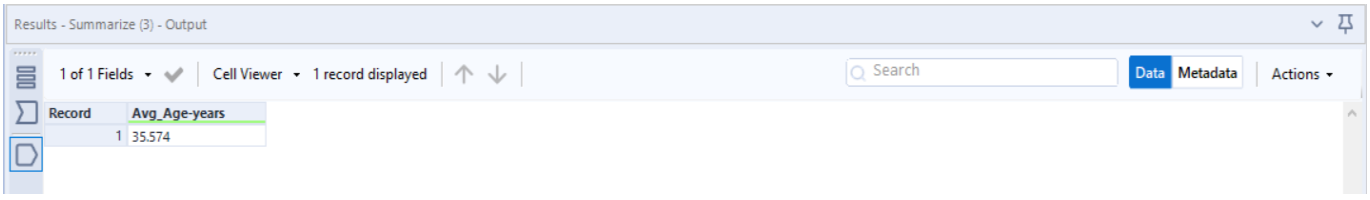
After having done a few analyses on the provided dataset, I chose to move on with removing few columns which are not going to useful for me in building a classification model to find whether a customer is worthy or not from the provided historical dataset. Furthermore, I am going to fill all null values in the provided dataset with the median value of the entire column.

- Columns dropped – "Telephone" – (unrelated), "No-of-dependents" (low variability), "Duration-in-Current-address" (Large null values), "Guarantors" (low variability), "Foreign-Worker"(low variability), "Concurrent-Credits" (Unique Values), "Occupation" –(Unique values).
- All null values in the column "Age-years" has been replaced with the median value 33.

As mentioned in the data cleaning guidelines

"Your clean data set should have 13 columns where the Average of Age Years should be 36 (rounded up)"

The average of all values in the "Age-years" after preprocessing is 36 app.



3. *Train your Classification Models*

"""""

Choose 70% to create the Estimation set and 30% to create the Validation set. Set the Random Seed to 1 if you're using Alteryx.

Train your dataset using these models:

- *Logistic Regression*
- *Decision Tree*
- *Forest Model*
- *Boosted Tree*

"""""

Once the data cleansing part is done, the next step is to build a classification model which can able to identify whether a customer is creditworthy or not from the new dataset (new instances).

Here, I first built four different classification models in order to choose the one with higher accuracy since my manager cares only about how accurate my model can identify the qualified customers.

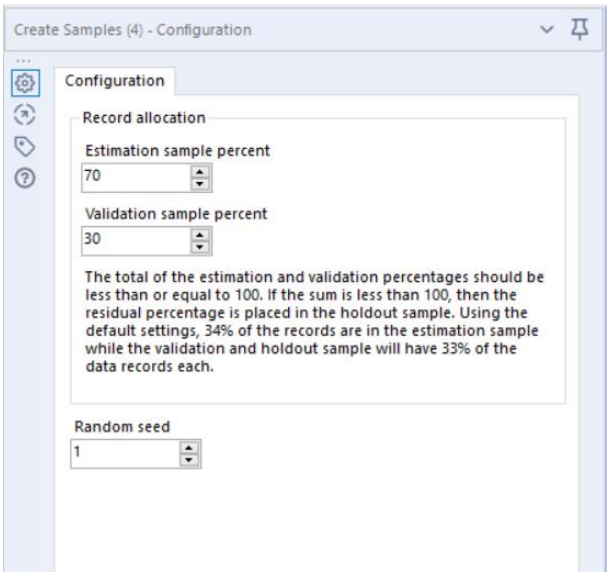
"Your manager only cares about how accurate you can identify people who qualify and do not qualify for loans for this problem.

Models built:

1. Logistic Regression.
2. Forest Model.
3. Decision Tree.
4. Boosted Tree.

STEP 1: Splitting Cleansed dataset into train and test datasets.

In this step, I split the dataset into train set and test set with 70% of the dataset for training purpose while 30% of the dataset for testing purpose.



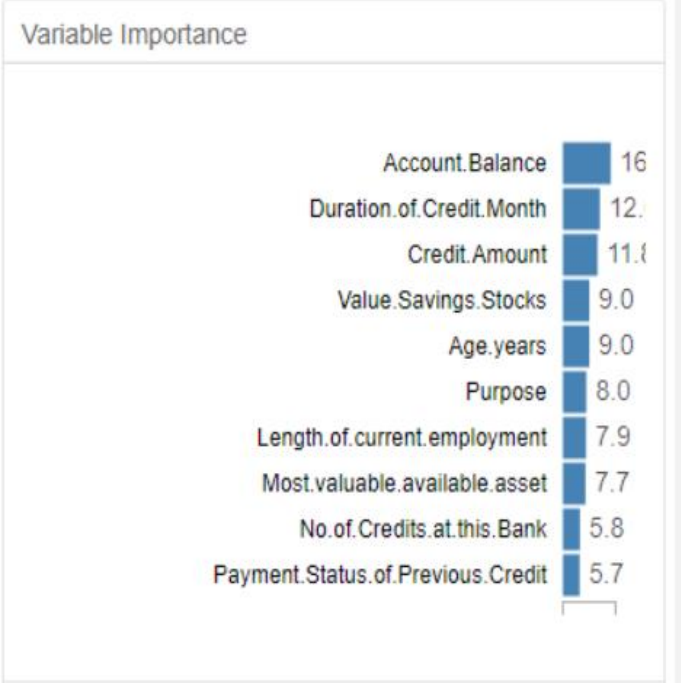
STEP 2: Training our model with our Train dataset:

Once we done with cleansing and splitting our dataset into train and test sets, we then built a classification model and trained it with the help of our training dataset. However, since there are lots of clustering algorithms available out there in the industry, we have to choose the one which suits for our purpose with high accuracy without overfitting. “So, how do we do that?” – one better way to do that is by building and training a model with all the available clustering algorithms and then choose the one with best accuracy by comparing all of them.

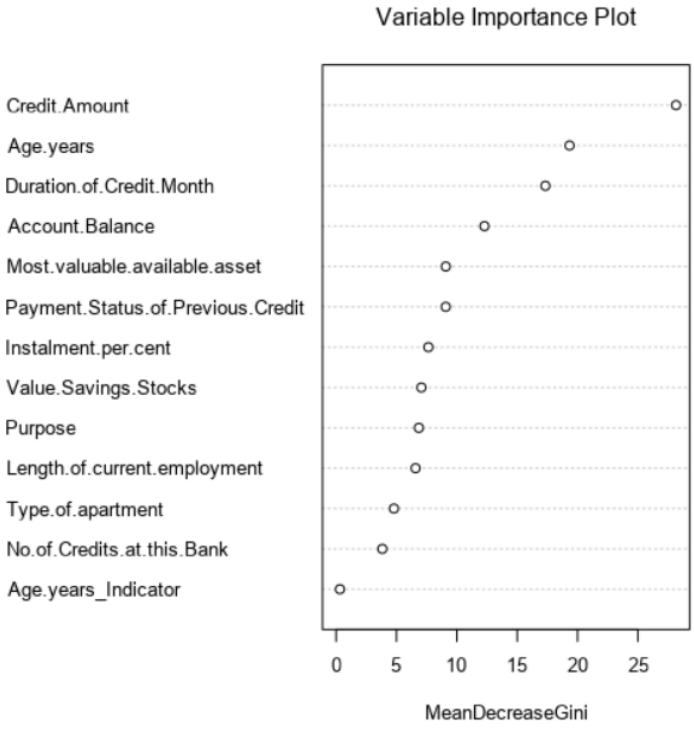
So, here I built four different clustering algorithms such as Logistic regression, forest model, decision tree and boosted model, and then I compared all four of them to see which one performs better while comparing it with all the other three models with the help of a metric known as “Accuracy”.

Here are the variable importance plots for Decision Tree, Boosting, Logistic Regression and Forest models, which shows how important each variable are for each models to predict the outcomes.

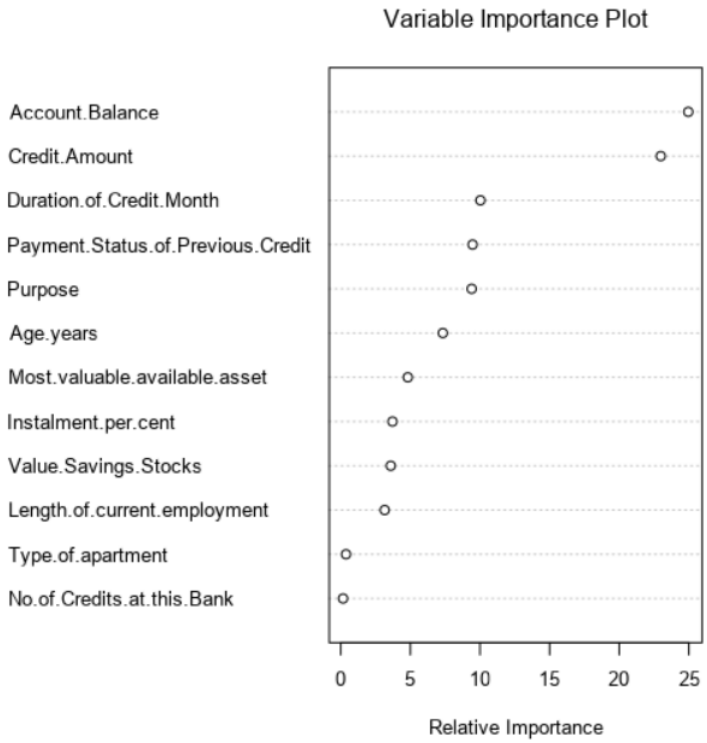
DECISION TREE:



FOREST:



BOOSTING



LOGISTIC REGRESSION

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.924e+00	6.960e-01	-4.20099	3e-05	***
Payment.Status.of.Previous.CreditPaid Up	2.956e-01	3.046e-01	0.97036	0.33187	
Payment.Status.of.Previous.CreditSome Problems	1.238e+00	5.173e-01	2.39324	0.0167	*
Account.BalanceSome Balance	-1.627e+00	3.130e-01	-5.19792	2.01e-07	***
Credit.Amount	1.753e-04	5.927e-05	2.95778	0.0031	**
Instalment.per.cent	2.831e-01	1.367e-01	2.07158	0.0383	*
PurposeNew car	-1.717e+00	6.182e-01	-2.77705	0.00549	**
PurposeOther	1.858e+01	1.779e+03	0.01045	0.99167	
PurposeUsed car	-7.890e-01	4.015e-01	-1.96530	0.04938	*
Most.valuable.available.asset	2.602e-01	1.453e-01	1.79085	0.07332	.
Length.of.current.employment4-7 yrs	2.331e-01	4.651e-01	0.50106	0.61633	
Length.of.current.employment< 1yr	8.235e-01	3.900e-01	2.11134	0.03474	*
Age.years_Indicator	-3.494e+01	2.134e+03	-0.01637	0.98694	

ACCURACY OF MODELS:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest	0.8000	0.8718	0.7424	0.9714	0.4000
Logistic_Regression	0.7600	0.8364	0.7418	0.8762	0.4889
Boosted	0.7867	0.8632	0.7515	0.9619	0.3778
Decision_Tree	0.6667	0.7685	0.6272	0.7905	0.3778

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.

From the above report, it is very much clear that the **accuracy of Forest Model** is superior while comparing it with the other three models, followed by boosted, logistic regression and decision tree.

CONFUSION MATRIX:

A method which is being used to visualize the performance of the classification models. It is normally used to find how many values are identified correctly and how many of them are not.

The below table is the confusion matrix of all the four classifiers in the order of boosted, decision tree, forest model and logistic regression.

Confusion matrix of Boosted

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest

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

Confusion matrix of Logistic_Regression

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

BOOSTED MODEL:

PPV = TP/(TP+FP) = 101/(101+28) = .78 || NPV = TN/(TN+FN) = 17/(17+4) = .80

This model doesn't show bias towards any particular sides.

DECISION TREE: PPV = 83/(83+28) = .78 || NPV = 17/(17+22) = 17/(39) = .43

This model is based towards creditworthy.

FOREST: PPV = 102/(102+27) = .78 || NPV = 18/(18+3) = 18/(21) = .85

This model appears to be slightly biased, but still the difference is not too high. Hence we can conclude that it's not biased.

LOGISTIC REGRESSION: PPV = 92/(92+23) = .80 || NPV = 22/(22+13) = .62

This model is clearly biased towards Non-creditworthy class.

4. Write up

""""

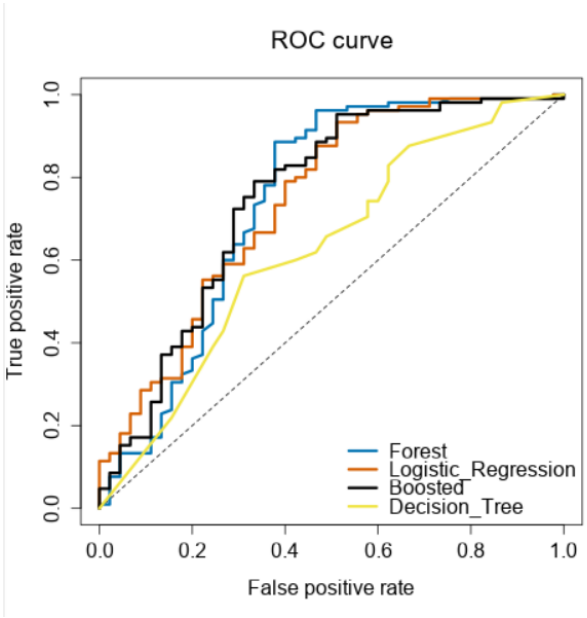
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

""""

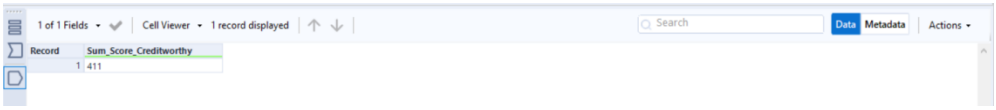
Based on what have mentioned in the above statement, we call an applicant as creditworthy if his/her "Score_Creditworthy" > "Score_NonCreditworthy". In other words, we can call an applicant as creditworthy applicant if his/her "Score_Creditworthy" value is above 0.5.

We have chosen to go ahead with Forest model because of its higher accuracy (0.80) comparing to other models. Also forest model has the highest "creditworthy_accuracy" of 0.97 while comparing to other models.

Here is the ROC curve graph which compares all the models which I have used, with each other. From this graph, it is very much clear that Forest model performs better than all the others.



Finally, Based on the output which we have a got by analyzing the new applicants dataset using our forest model, there are total of **411 creditworthy applicants** are there, for whom we can provide a loan with.



Alteryx Model: