Project: Creditworthiness

By Faouzi NAJEH

Step 1: Business and Data Understanding

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

This project looks into a binary classification problem solution using various Alteryx tools & technics in an attempt to building different classification models capable of predicting whether a customer is approved or not for a loan for a set of defined records, to finally choose the most accurate or performant model and score the customer_to_score new dataset.
The business is about a small bank that want judge for approval/disapproval in their customers loan applications, the paramount work is to determining if customers are creditworthy or not-creditworthy to give a loan to.
Note that the manager is extremely interested on how accurately can we identifying whether a customer is approved or not approved for a loan rather than the customer number in each of two classes.
Our goal in this project is to help make the business decision and give a final customers most accurate classification to the manager in order to make the most beneficial decision for the subject bank.

Key Decisions:

Answer these questions

- **What decisions needs to be made?**

  Answer 01: The decision needs to be made is which customers in our available dataset are approved and not approved for a loan, we need to classify all new customers and figure out the target result for each record?
- **What data is needed to inform those decisions?**

  Answer 02: To make this decision, we have 2 datasets: On the one hand, we have the data of all past applications. We have used this dataset to create and train the model. On the other hand, we have the list of customers who have applied to get a loan. This

dataset has been scored with the model to get the list and number of final customers that are creditworthy to get a loan.

A set of features related to customers are available to make our decision, such as 'Account-Balance', 'Duration-of-Credit-Month', 'Length-of-current-employment', 'Payment-Status-of-Previous-Credit', 'Credit-Amount, Credit-Amount', 'Concurrent-Credits' and the customer age, all those independent variables may be good predictors for our target variable which is the 'Credit-Application-Result'.

For numeric features we will investigate about linear relationships and high correlations.

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

  Answer 03: For this problem we will use a binary classification since the target variable represent whether or not a customer is approved for a loan from our bank.

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".

Answer 01: As mentioned in the Pearson correlation analysis presented below, we have investigated numeric features linear relations to see if we have high correlations, but in most cases, features have negative or positive weak linearity between each other, except the 'Credit-amount' and the 'Duration-of-Credit-month' have relatively good positive relation but not 'high' with a correlation of 0.57.

Pearson Correlation Analysis

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Most.valuable.available.asset	Age.years	Occupation	No.of.dependents
Duration.of.Credit.Month	1.0000000	0.5739797	0.2998549	-0.0641970	NaN	-0.0652690
Credit.Amount	0.5739797	1.0000000	0.3255454	0.0693159	NaN	0.0039858
Most.valuable.available.asset	0.2998549	0.3255454	1.0000000	0.0862334	NaN	0.0464542
Age.years	-0.0641970	0.0693159	0.0862334	1.0000000	NaN	0.1177356
Occupation	NaN	NaN	NaN	NaN	1.0000000	NaN
No.of.dependents	-0.0652690	0.0039858	0.0464542	0.1177356	NaN	1.0000000

Matrix of Corresponding p-values

	Duration.of.Credit.Month	Credit.Amount	Most.valuable.available.asset	Age.years	Occupation	No.of.dependents
Duration.of.Credit.Month		0.0000e+00	7.5764e-12	1.5175e-01	NaN	1.4502e-01
Credit.Amount	0.0000e+00		8.3045e-14	1.2164e-01	NaN	9.2916e-01
Most.valuable.available.asset	7.5764e-12	8.3045e-14		5.3979e-02	NaN	2.9987e-01
Age.years	1.5175e-01	1.2164e-01	5.3979e-02		NaN	8.4080e-03
Occupation	NaN	NaN	NaN	NaN		NaN
No.of.dependents	1.4502e-01	9.2916e-01	2.9987e-01	8.4080e-03	NaN	

Figure 01: Pearson Correlation Analysis Results

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed.

Answer 02: While assessing our dataset by the field summary tool with Alteryx, we have realized that two fields ('Age_years', 'Duration-in-Current-address') contains missing values, note that approximately 68% of data in the 'Duration-in-Current-address' field are missing, while around 2.5% of data in the 'Age-years' field are missing values.

- 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.

Answer 03: Based on the result of the summary field tool mentioned previously, we have around 7 fields that represent low variability in our dataset: 'Occupation', 'Concurent-Credits', 'Guarantors', 'No-of-dependents, 'Telephone' and 'Foreign-Worker'.

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

As expected, after finishing the cleaning process we have our dataset contains 13 columns and 500 records with an average age of 36.

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.

Answer: During the cleaning process, we have removed one field which is 'Duration-in-Current-address' and we decided to remove it because more than half of the data in this field (68%) is missed, presented in red in figure 02.

On the other hand, we have imputed the 'Age-years' field by the median, since we just have 2.5% of it are missing values, presented in red in figure 03, so we should impute these values in favor of our analysis and modelling steps.

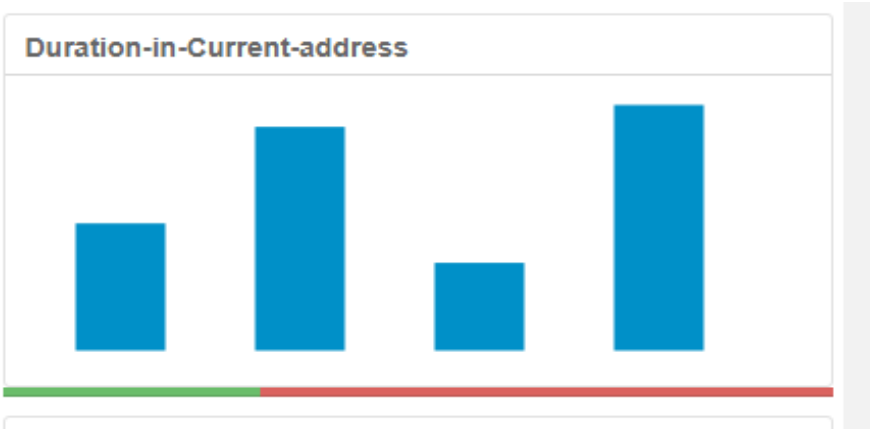

Figure 02: Duration in Current address distribution

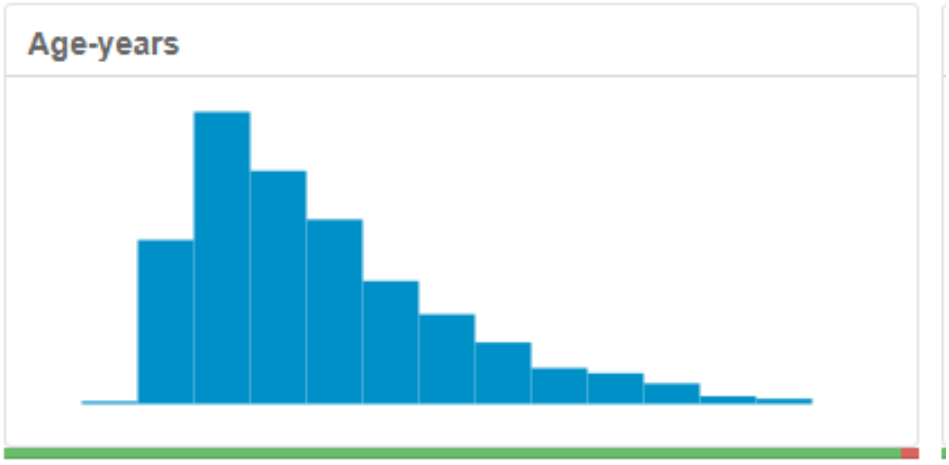

Figure 03: Age in years distribution

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.
- **Answer**: After building our four models, based on models reports we can conclude about feature importance and figure out best predictors for our target.

1. **Logistic Regression:** It seems that we only have some significant features that could be good predictors for our target variable, whose are: 'Account-Balance', 'Payment-status', 'Purpose', 'Credit-amount', 'Instalment-per-cent' and 'Most-valuable-available-asset', as mentioned in the logistic regression report next in figure 04, these variables have all p-values less than 0.05 which give us statistical significance, but most of variables have p-values greater than 0.05 and does not present statistical significance to be good predictors. Note that the 'Age-years' does not seems to be a good predictor in our logistic regression model.

Report for Logistic Regression Model Logistic_Regression

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 +
Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, 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	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
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	

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

Figure 04: Logistic Regression Report

2. **Decision Tree:** For the decision tree model, based on the feature importance chart presented in figure 05, the most important features are 'Account-Balance', 'Duration-of-Credit-Month', 'Credit-amount', 'Value-Savings-Stocks', 'Age-years', so we can conclude that these variables are good predictors for our target by the decision tree model.

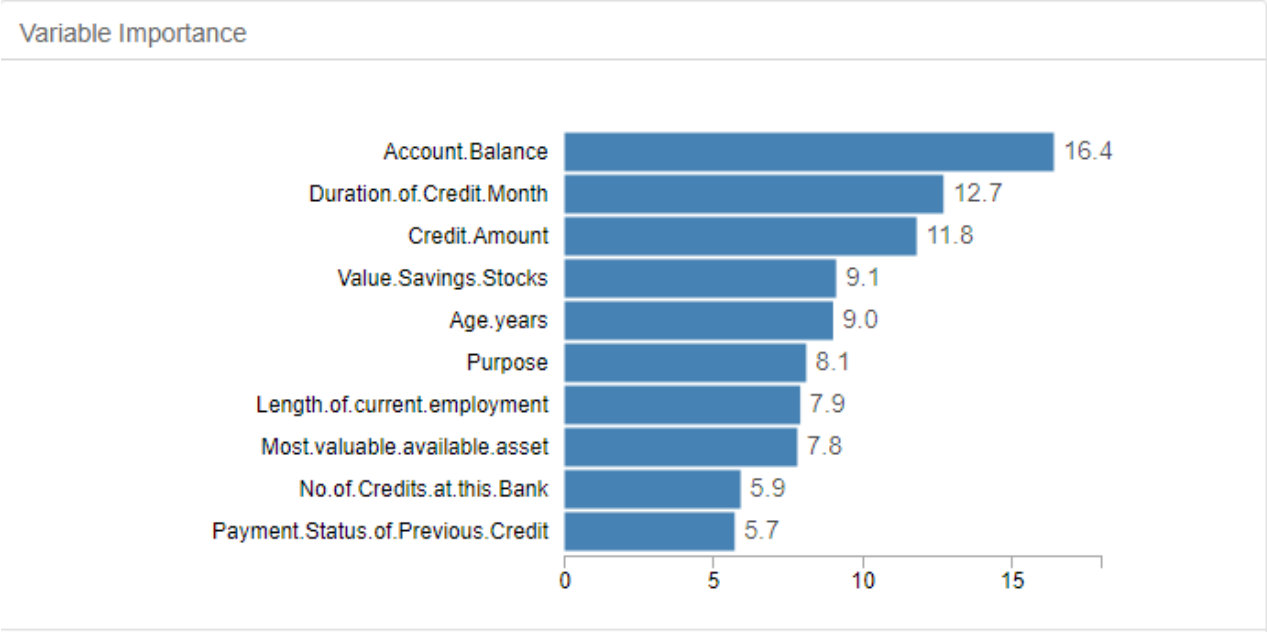

Figure 05: Decision Tree Variable Importance Chart

3. **Random Forest Classifier**: Concerning the random forest classification which was performed by 500 trees, we have obtained the feature importance chart presented in figure 06.

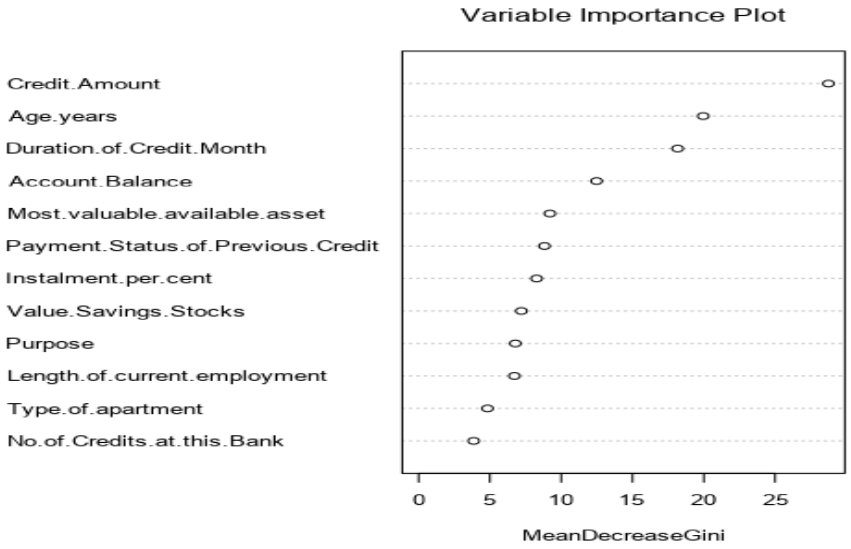

Figure 06: Random Forest Variable Importance Chart

From the previous chart, the most important variables with the random forest classification are: 'Credit-amount', 'Age-years', 'Duration-of-Credit-Month', 'Account-Balance', and 'Payment-status' with a relatively low degree of importance. Therefore, with the random forest model these variables are considered as good predictors in order to determine whether a customer is approved or not approved for a loan.

4. **Boosted Model:** Finally for the boosted model, results shown that the 'Credit-amount' and the 'Account-Balance' features are the most important predictors for the target variable.

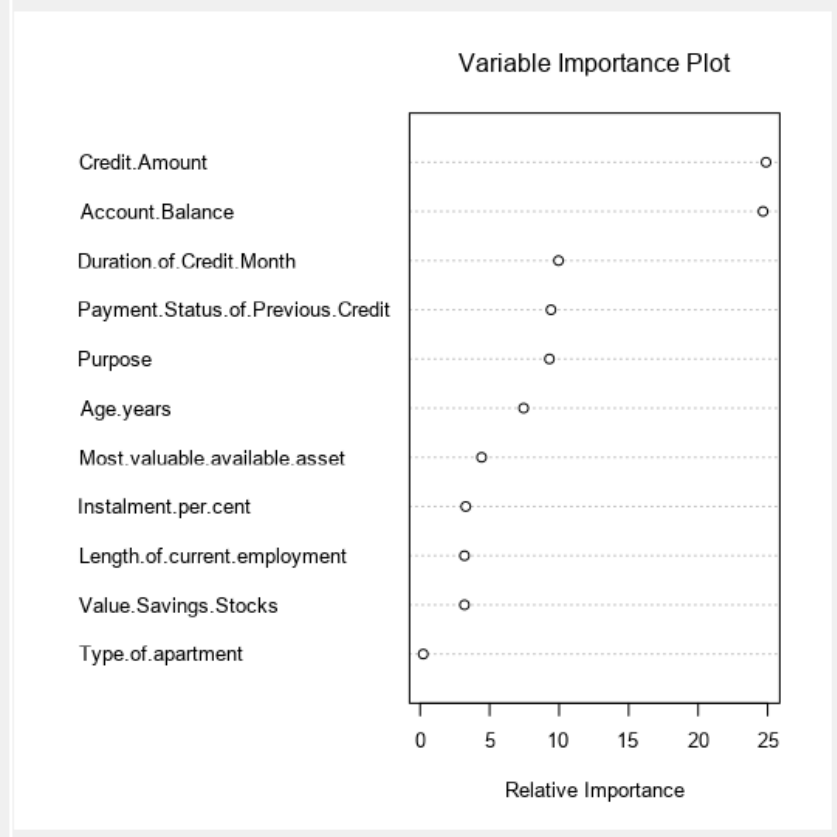

Figure 07: Boosted Model Variable Importance Chart

Conclusion: At the end of the cleaning and training steps, the 'Credit-Amount', 'Account-Balance' features seems to be the common best predictors for all our four models, then comes 'Age-years' and 'Duration-of-Credit-Month' differently ranged for each model and with lower effect comparing to 'Credit-Amount' and ''Account-Balance' predictors.

- 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)

- **Answer**: At this step we have validated each one of our models against the validation set, by using the union tool and a model comparison. Based on the model comparison report and the confusion matrix for each model we have figure out interpretations as follow:

1. **Logistic Regression:** The model is doing well at predicting our target variable with an overall accuracy around 78%. By looking at all accuracy percent we realized that the model predicts much more accurately the creditworthy records (approved applications) with an accuracy of 90%, while for Non-Creditworthy 49% of predicting accuracy as shown in the model comparison report(fig.08). That could be explained by the low number of not approved records.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree	0.6733	0.7721	0.6296	0.7905	0.4000
Random_Forest_Classifier	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7490	0.9619	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.

Figure 08: Model Comparison Report

2. **Decision Tree:** The overall accuracy for the decision tree is relatively lower than the logistic regression and our tree predict the target by a 67% accuracy which is slightly low value comparing to other models.
3. **Random Forest Classifier:** The random forest classifier model seems also to be a good model, it predicts our target variable with an overall accuracy of 79% (fig.08).
4. **Boosted Model:** Our boosted model accuracy is almost the same as the decision tree, we predict the target with a 78% accuracy. By looking at the confusion matrix the random forest and the boosted model performing equally.

Confusion matrix of Boosted_Model

	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	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of Logistic_Regression

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

Confusion matrix of Random_Forest_Classifier

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

Figure 09: Confusion Matrix

Looking for Bias for each Model:

Logistic Regression: The overall percent accuracy of the Logistic model is 76% which is strong.

PPV=True_positives\(True_positives+False_positives)=95/(95+23)=0.80

NPV=True_Negatives\(True_Negatives+False_Negatives)=22/(22+10)=0.68

So after checking the confusion matrix there is bias seen in the logistic model prediction.

Decision Tree: The overall percent accuracy of the DecisionTree model is 67%.
PPV=True_positives\(True_positives+False_positives)=83/(83+27)=0.75
NPV=True_Negatives\(True_Negatives+False_Negatives)=18/(18+22)=0.45
After checking the decision tree confusion matrix, we can conclude that there is bias seen in the model's prediction to Creditworthy.

Random Forest Classifier: The overall percent accuracy of the random forest model is 79%.
PPV=True_positives\(True_positives+False_positives)=102/(102+28)=0.78
NPV=True_Negatives\(True_Negatives+False_Negatives)=17/(17+3)=0.85
After checking the decision tree confusion matrix, we can conclude that there is no bias seen in the model's prediction.

Boosted Model: The overall percent accuracy of the Boosted model is 79%.
PPV=True_positives\(True_positives+False_positives)=101/(101+28)=0.78
NPV=True_Negatives\(True_Negatives+False_Negatives)=17/(17+4)=0.80
So after checking the confusion matrix there is no bias seen in the logistic model prediction.

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

Answer: In order to help make decision and classify our set of applications in two segments Approval/Not-Approval, I have chosen the **Random Forest** model as a classifier to predict whether a new customer is creditworthy or not to give a loan to.

The logistic regression and decision tree models presenting bias in predictions so we won't work with any of these two models.

In term of overall accuracy, the random forest model is 79% accurate which is considered as a good accuracy and it is so close to the 'Boosted' model value which is 78%. However, the 'random forest' predicts the CrediWorthy with a perfect accuracy and also for the Non-CreditWorthy segment we have more important accuracy compared to the boosted model, and this point may be important because our models are predicting the Non-CreditWorthy segement with more difficulty.

Also, by having a close glance to the ROC curve results, presented in figure 10, the random forest classifier reaches the top the quickest and it has the overall highest curve of all. Added to that, our random forest classifier has the highest AUC among our four models. We can say that the 'random forest' classifier and the 'Boosted model' are performing at the same level and are the better models comparing to the rest of models. Related on the point that our 'random forest model' predicts more accurately overall and for the Non-CreditWortyh segment, and the fact of AUC and ROC we have chosen the 'random forest model' to predict our target in this project and get a final classification solution that give us applications results.

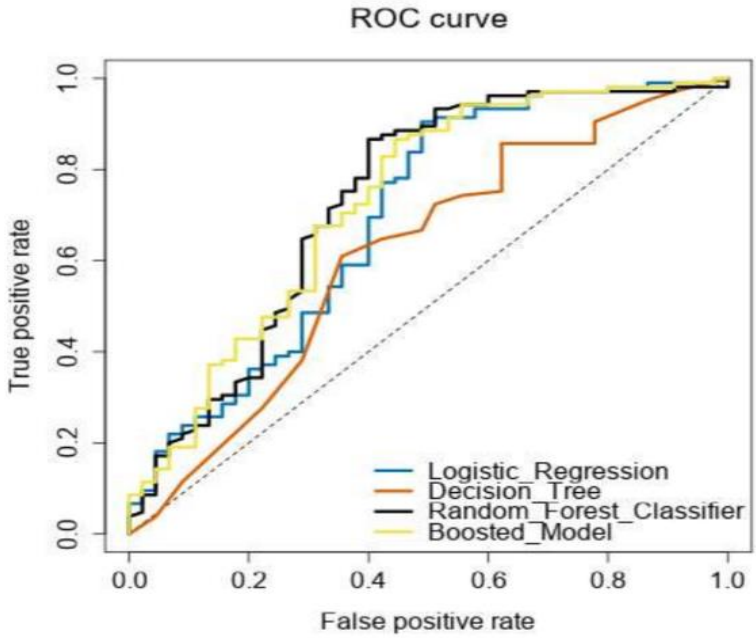

Figure 10: ROC Curve

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

- How many individuals are creditworthy?

Answer: 408

2 of 2 Fields | Cell Viewer | 1 record displayed, 1,100 bytes

Record	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	408	92

Figure 11: Boosted Model Final Score

Alteryx Workflow

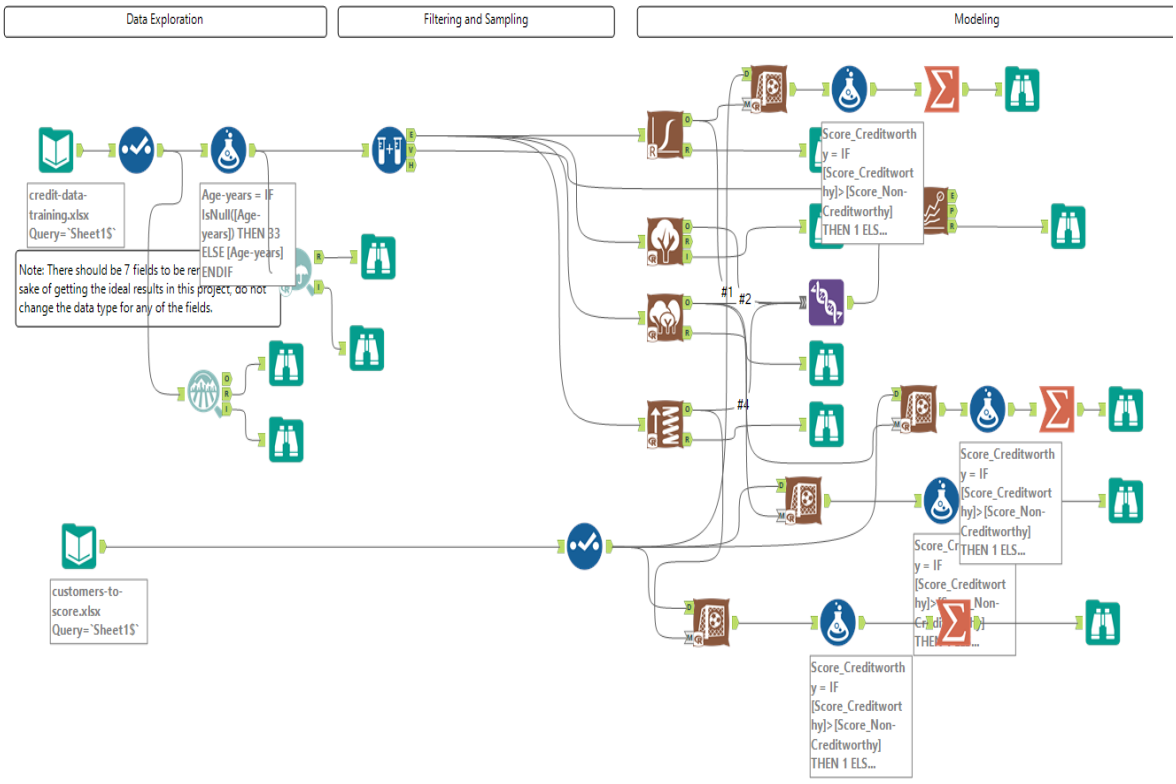

Figure 12: Predicting Default Risk Alteryx Workflow