Predictive Analytics for Business

Project #4 Predicting Default Risk classification models

Name: Marwan Saeed Alsharabbi

Date: 12–Jan–2020

2020

INTRODUCTION

Classification models

Classification model: A **classification model** tries to draw some conclusion from the input values given for training. It will predict the class abeles/categories for the new data. Feature: A feature is an individual measurable property of a phenomenon being observed.

Broadly speaking, there are four **types of classification**. They are:

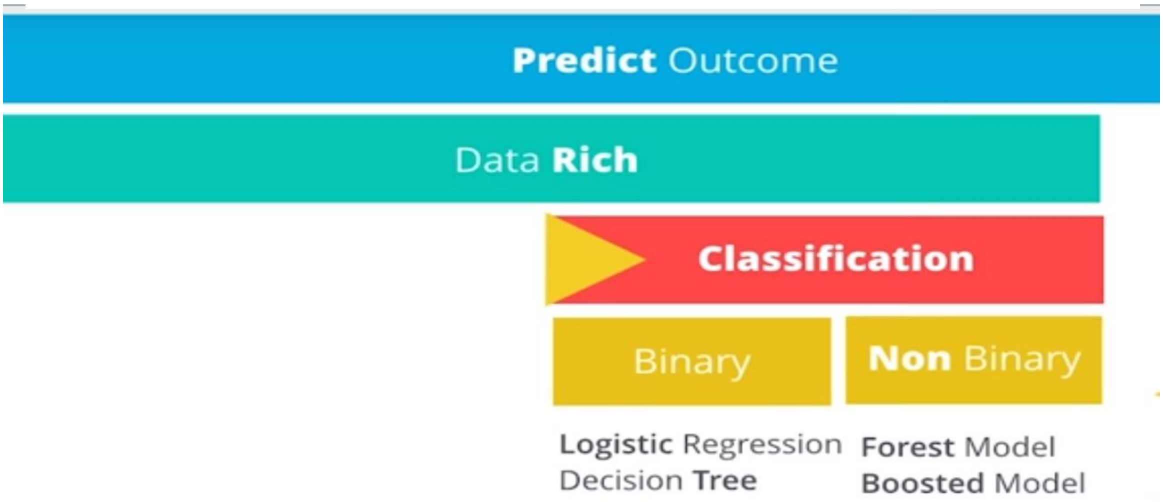
(i) Geographical **classification**,

(ii) Chronological **classification**,

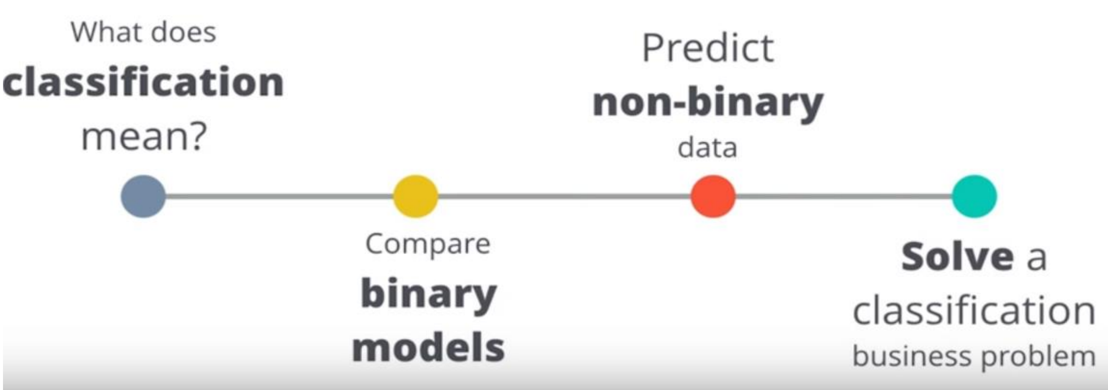
(iii) Qualitative **classification**,

(iv) Quantitative **classification**.

There are a number of classification models. Classification models include **Binary (logistic regression**, **decision tree)**, **Non-Binary** (**random forest**, **boosted tree** Model in the course

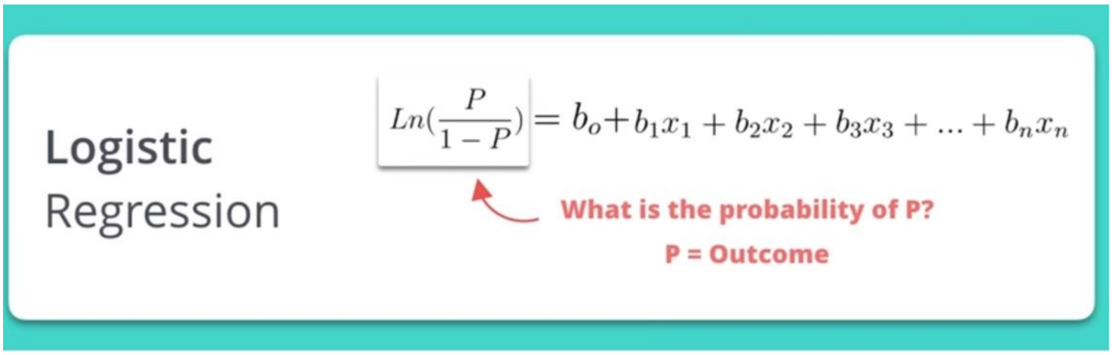


What is going on within this Classification section Model



1-Logistic Regression

Logistic Regression is a statistical method used to predict binary outcomes by analyzing the outcome’s relationship with one or more predictor variables.



STEP 1: SELECT TARGET AND PREDICTOR VARIABLES

Target Variable: The target variable is the variable we are trying to predict with the model. This should be a binary variable: yes/no, true/false, 0/1, etc.

Predictor variables: The predictor variables are used to help predict the target variable. Predictor variables should be: (1) Relevant to the target variable, (2) not highly correlated to other predictor variables, and (3), do not have a high number of missing values

Useful Alteryx tool: Association Analysis

STEP 2: PREPARE DATA

Preparing the data includes dealing with issues such as missing, dirty, or duplicate data; removing outliers; blending and formatting data, etc. Your final dataset should include one row for each outcome and set of predictor variables.

Estimation and validation samples: Next, split the data set into two parts: one part for Estimation (for training the model) and one part for Validation (to help us verify that we are creating a useful model).

Useful Alteryx tool: Create Samples

STEP 3: BUILD AND RUN THE MODEL

Run the model with the target and predictor variables. Observe the statistical significance of each of the predictor variables by looking at the p-value in the output. If it’s below 0.05, then the relationship between the target and predictor variable is statistically significant. If not, it is not significant and can be excluded from the model. R-squared is an estimate between 0 and 1 of the explanatory power of them model and can be used to compare models and select the best one.

Using a technique called “stepwise regression” can automatically identify the best combination of predictor variables.

Useful Alteryx tools: Linear Regression, Stepwise

STEP 4: MODEL VALIDATION

Apply the model to the validation sample and observe how accurately the model predicts the outcomes. This step helps avoid overfitting and helps you understand how accurate your predictions will be on new data.

Useful Alteryx tool: Model Comparison

STEP 5: APPLY THE MODEL TO MAKE PREDICTIONS

Apply the model to a new dataset to make predictions. This dataset should have all the predictor variable values, which are passed through the model to predict the unknown target variable value. The prediction will be a number between 0 and 1, representing the likelihood of positive outcome.

Useful Alteryx tool: Score

2-Decision Tree and Forest Models

These are two classification models. These models help identify what group a data point belongs to. Decision Tree and Forest models can help predict classification of categorical or continuous variables.

STEP 1: Create sample
In any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data.

Useful Alteryx tool: Create Sample

STEP 2: Model Settings
Select a target variable and predictor variables, you can include as many predictor variables as you would like because the model will only use variables that work best. Specify the number of records needed to allow for a split, the smaller the number the more splits you will get. In the Forest Model you can choose the number of trees to use.

Useful Alteryx tool: Forest, Decision Tree

STEP 3: Interpreting the Report
Root Node Error in the Decision Tree model is the percentage of how many of the data points went to the incorrect terminal node (predicted incorrectly) when all the data points are validated against themselves within the entire training set (the

Estimation dataset). The Pruning Plot lists out the levels in the decision tree with their related error terms with cross-validation samples.

The Variable Importance Plot is a bar graph that's length indicates the importance of the predictor variables. The Confusion Matrix is a matrix (or table) that lists out all of the possible prediction results when we validate our model against itself.

The Out of the Bag Error Rate for the Forest Model explains how well the model performed with the cross-validation set in the estimation data. Similar to R-squared. The Percentage Error for Different Number of Trees graph helps us see what the correct number of trees is to use, so we can avoid over computing in the future. What we are looking for where does the graph flatline?

Useful Alteryx tools: Forest, Decision Tree

STEP 4: Model Comparison

Use the fit and error measures, Accuracy which represents the overall accuracy, the number of correct predictions of all classes divided by total sample number. The F1 score is calculated the following way, precision * recall / (precision   recall) You can read more about precision and recall. There will also be a confusion matrix in this report to show how the models compared to the validation set. This confusion matrix is one of the best methods to review the accuracy and precision of your model as well as to understand any model bias in classifying your data points.

Useful Alteryx tool: Model Comparison

STEP 5: Score Data

Apply the model by attaching a score tool to the data set you are trying to classify and the model object.

Useful Alteryx tool: Score

3-Boosted Models

STEP 1: Create sample
In any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data.

Useful Alteryx tool: Create Sample

STEP 2: Model Settings
Select a target variable and predictor variables, you can include as many predictor variables as you would like because the model will only use variables that work best. For a Boosted model it is best to set the target type in the model customization tab. Your options are Continuous, Count, Binary Categorical or Multinomial Categorical.

Useful Alteryx tool: Boosted Model

STEP 3: Interpreting the Report
The Variable Importance Plot is a bar graph that's length indicates the importance of the predictor variables. The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specified assessment criteria

Useful Alteryx tools: Boosted Model

STEP 4: Model Comparison
Use the fit and error measures, Accuracy which represents the overall accuracy, the number of correct predictions of all classes divided by total sample number. The F1 score is calculated the following way, precision * recall / (precision   recall) You can read more about precision and recall.

The Confusion Matrix is a matrix (or table) that lists out all of the possible prediction results when we validate our model against our validation set. This confusion matrix is one of the best methods to review the accuracy and precision of your model as well as to understand any model bias in classifying your data points.

Useful Alteryx tool: Model Comparison

STEP 5: Score Data
Apply the model by attaching a score tool to the data set you are trying to classify and the model object.

Useful Alteryx tool: Score

Project Details: -

The Business Problem

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

Steps to Success

Step 1: Business and Data Understanding

Your project should include a description of the key business decisions that need to be made.

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

Note: If you decide to impute any data field, for the sake of consistency in the data cleanup process, impute the data using the median of the entire data field.

Step 3. Train your Classification Models

You should 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

Step 4. Writeup

Compare all of the models' performance against each other. Decide on the best model and score your new customers.

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

Step 1: Business and Data Understanding

What decisions needs to be made?

determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand.
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, identify people who qualify and do not qualify for loans for this problem.

: Awesome: Good job identifying the key decision to be made.

What data is needed to inform those decisions?

Data on all past applications and list of customers that need to be processed in the next few days.

credit-data-training.xlsx - This file contains all credit approvals from your past loan applicants the bank has ever completed.
customers-to-score.xlsx - This is the new set of customers that you need to score on the classification model you will create.

: Correct! These datasets should provide us with useful data to carry out our analysis. These should include data related to the customer's current length of employment, income, credit score, if the customer carries a credit balance from month to month, and their current savings.

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

Binary classification models (logistic regression, decision tree, forest model, boosted model) are needed to help make these decisions and select the best model.

: Awesome: The correct model type has been identified.

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:

1-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".
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)

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.

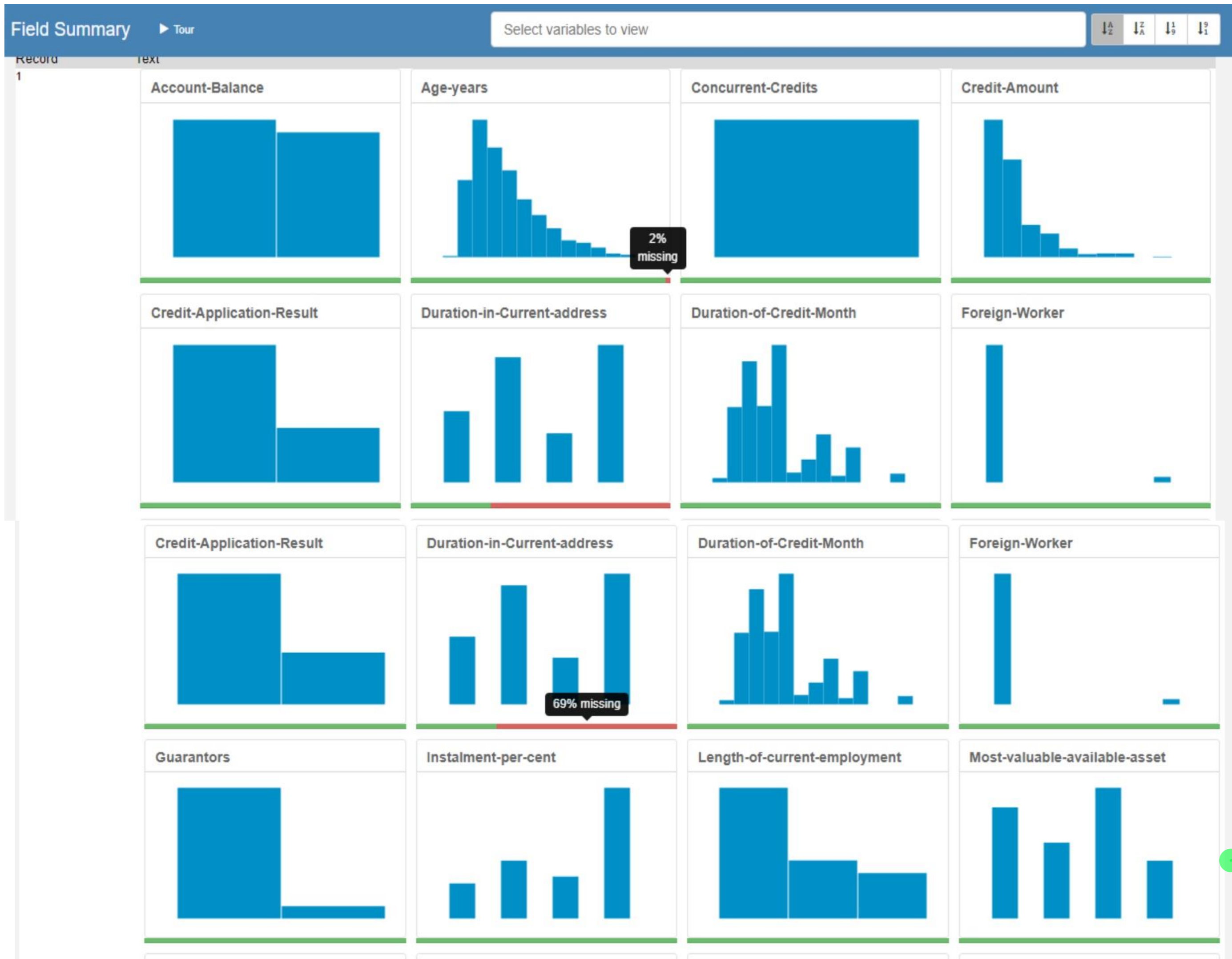
In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields.
Duration in Current Address has 69% of the data missing. Since fields with a lot of missing data should be removed this variable has been removed.
The histography of the variable **Guarantors, Foreign-worker and No-of dependents** shows that majority of the data is heavily skewed towards one type of data. Also, **Concurrent Credits** and **Occupation** have that are entirely uniform and there are no other variations of the data. All these variables have been removed due to low variability.
Telephone does not have any predictive ability to the credit application result, so this field should also be removed.
Age-Years has 2% of the data missing. The missing data of this variable has been imputed using the median, 33 of the entire data field. Please see the Visualizations in next page

: Correct! The "Duration in current address" has too many missing data to be useful in our analysis, and hence should be dropped.

: Awesome: The low variability variables have been correctly identified.

: Yes, including the "Telephone" field does not seem logical for our analysis.

: Awesome: This decision is correct. We impute because not much data is missing here. And we use the median because of the presence of a slight skew in the age field's distribution.



: Nice work also including the visualizations for each of the fields.

: Awesome: You did an excellent work removing all the appropriate fields alongwith their correct justifications.

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

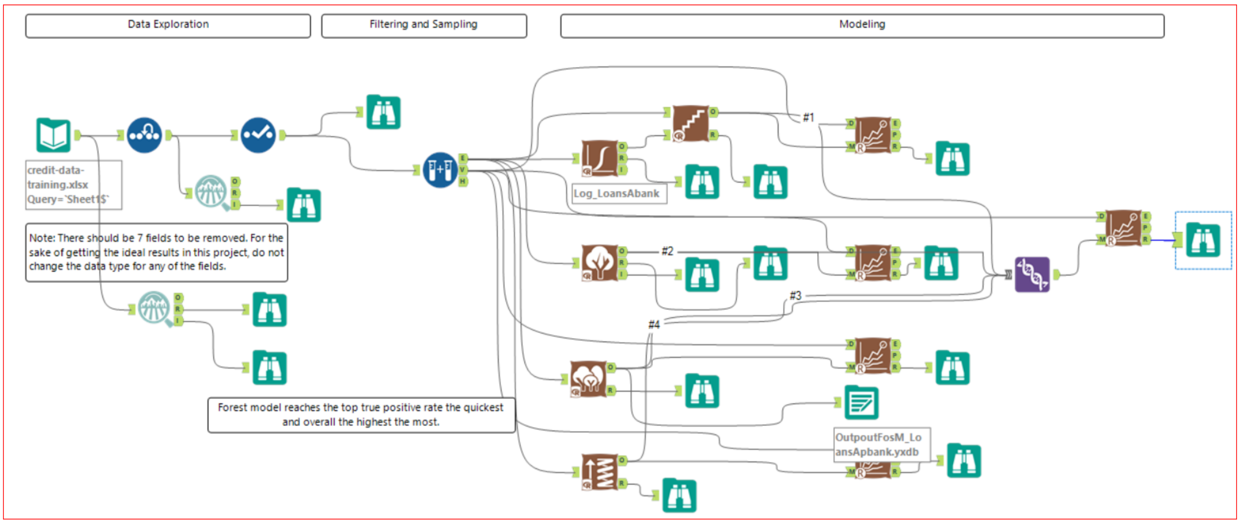
*Answer these questions for **each model** you created:*

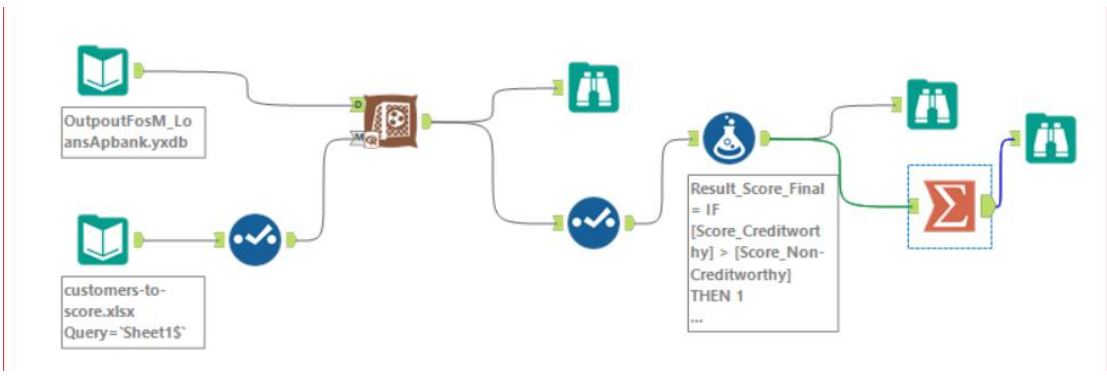
Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

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?

Train your Classification Models:

Alteryx Workflow





1-Logistic Regression

Logistic regression is one of the most basic forms of regression modeling. It's part of a family of "generalized linear models" or GLM for short. This basically means that the formula is very similar to that of a linear regression. when executing the logistic regression model, we see the emergence of multiple variables and classes where I value the R-Squared = 0.2199 that we need to use (Stepwise) in order to reduce the number of variables and the result is accurate. See the report Logistic regression

: Awesome: Good job using the stepwise tool here. Recall from the lesson that stepwise automates the process of coming up with the best predictor variables, thereby improving our overall efficiency of coming up with the final solution.

Report

Report for Logistic Regression Model Log_LoansAbank

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, 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	-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

2- Logistic Regression -Stepwise

The Stepwise Regression tool needs to figure out all of the possible variables it can calculate first and it takes this list of possible variables from the Logistic Regression Tool output. When we see the implementation report, the number of variables decreased due to return and deletion. Some variables changed value R-square =0.2048 For this logistic regression (stepwise) model, Account Balance, Payment status of Previous Credit, and Purpose are three of the most significant variables. The overall accuracy is 76%.

The result Comparison Report Logistic Regression -Stepwise

Report for Logistic Regression Model SW_LoansApbank

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial("logit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

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

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: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SW_LoansApbank	0.7600	0.8364	0.7306	0.8762	0.4889

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.

Confusion matrix of SW_LoansApbank

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

Performance Diagnostic Plots

Using the confusion matrix,
accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 92/ (92+23) = 0.8, 80% while
accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 22/ (13+22) = 0.6286, 62.86%
The model seems to be slightly biased towards predicting customers as non-creditworthy.

: Correct! And because of the bias, we should choose this model, or else we would deny loans to many individuals who are creditworthy.

3-Decision Tree

Decision Tree can help predict classification of categorical or continuous variables in any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data in the decision tree model, in this project Account Balance, Duration of Credit Month, and Value Saving Stocks are
three of the most significant variables. The overall accuracy is 74.67%.

Report

Summary Report for Decision Tree Model DT_LoansApbank

Call:
rpart(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, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)

Model Summary

```
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

Leaf Summary

```
node), split, n, loss, yval, (yprob)
    * denotes terminal node
1) root 350 97 Creditworthy (0.7228571 0.2771429)
  2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
  3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
    6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
    7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
      14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
      15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *
```

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_LoansApbank	0.7467	0.8273	0.7054	0.8667	0.4667

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.

Confusion matrix of DT_LoansApbank

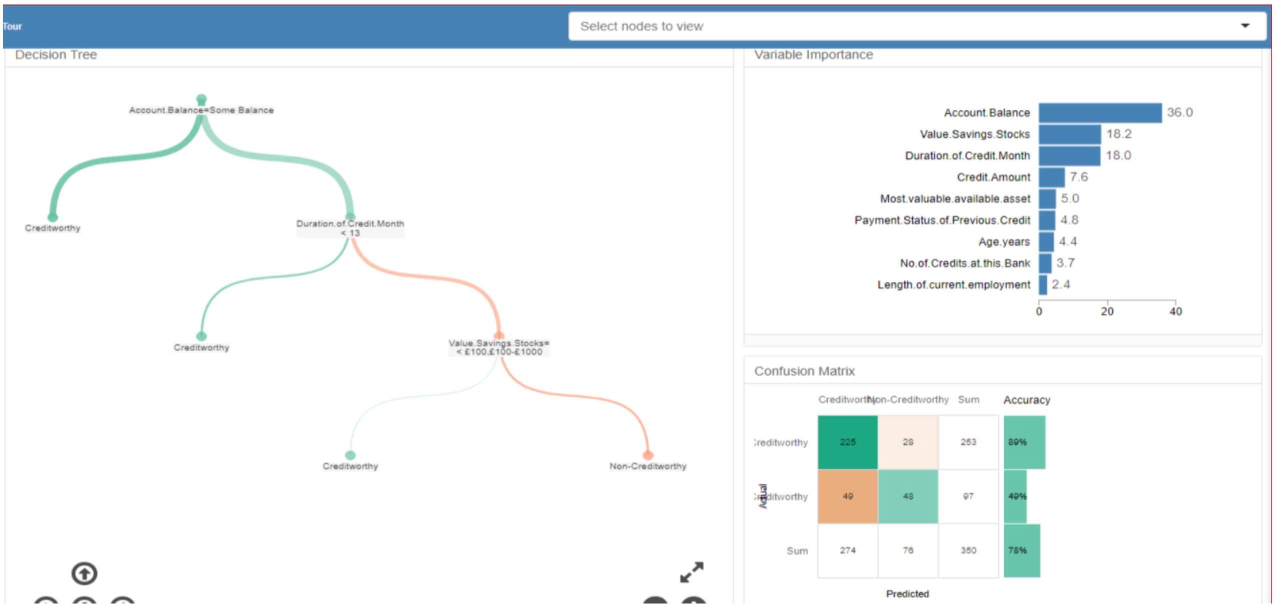
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Performance Diagnostic Plots

Using the confusion matrix,
accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 91/ (91+24) = 0.7913, 79.13%
accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 21/ (14+21) = 0.6, 60%

The model seems to be biased towards predicting customers as non-creditworthy

: Good job again! Similar to the logistic regression model, we shouldn't choose the decision tree or else we would deny loans to many individuals who are creditworthy.



4-Forest Models

Forest models can help predict classification of categorical or continuous variables in any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data in the decision tree model, in this project Credit.Amount , Duration of Credit Month, and Amount Balanse ,Age-years Four of the most significant variables. The overall accuracy is 80.00%.

Report

Basic Summary

Call:

randomForest(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, data = the.data, ntree = 500, replace = TRUE)

Type of forest: classification

Number of trees: 500

Number of variables tried at each split: 3

OOB estimate of the error rate: 24%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.639	62	35

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FosM_LoansApbank	0.8000	0.8707	0.7361	0.9619	0.4222

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*.

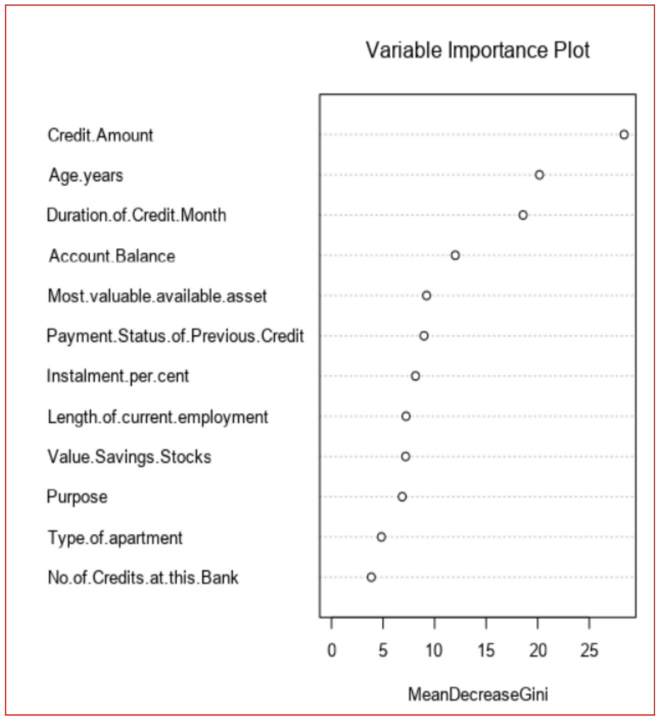
AUC: area under the ROC curve, only available for two-class classification.

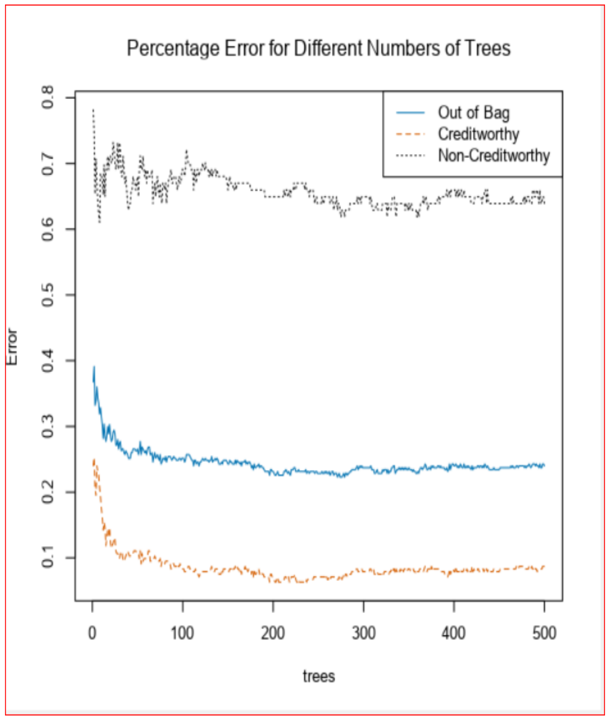
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of FosM_LoansApbank

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

Using the confusion matrix,
accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 101/ (101+26) = 0.7952, 79.52%
accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 19/ (19+4) = 0.8260, 82.60%
Since accuracies for creditworthy and non-creditworthy are comparable 79.52% and 86.37% respectively, this model isn't biased





: Awesome: Well done coming up with the correct set of the most significant variables and also identifying the model to be not biased.

5-Boosted Models

In any classification problem you will need to set an estimation sample and a validation sample of your data. This helps us compare different classification models to see which better fit the data. In this boosted model, Account Balance, Credit Amount and Credit Month three of the most significant variables. The overall accuracy is 78.67%.

Report

Summary Report for Decision Tree Model DT_LoansApbank

Call:
rpart(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, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)

Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

Leaf Summary

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 350 97 Creditworthy (0.7228571 0.2771429)
  2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
  3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
    6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
    7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
     14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
     15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *
```

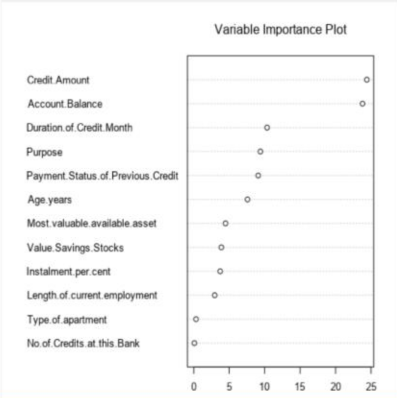
Records 1 to 5

Report for Boosted Model BoostM_LoansApbank

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BoostM_LoansApbank	0.7867	0.8632	0.7524	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.

Confusion matrix of BoostM_LoansApbank

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

Performance Diagnostic Plots

Using the confusion matrix,
accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 101/ (101+28) = 0.7952, 79.52%
accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 17/ (17+4) = 0.8095, 80.95%
Since accuracies for creditworthy and non-creditworthy are comparable 79.52% and 80.95% respectively, this model isn't biased

: Correct! The boosted model is also not biased.

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.

Forest Model has been chosen since it has the highest accuracy of 80% among all four classification models. Also, accuracies for creditworthy and non-creditworthy are among the highest of all

: Awesome: The best model has been correctly chosen and the decision appropriately justified.

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_LoansApbank	0.7467	0.8273	0.7054	0.8667	0.4667
FosM_LoansApbank	0.8000	0.8707	0.7361	0.9619	0.4222
BoostM_LoansApbank	0.7867	0.8632	0.7524	0.9619	0.3778
SW_LoansApbank	0.7600	0.8364	0.7306	0.8762	0.4889

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.

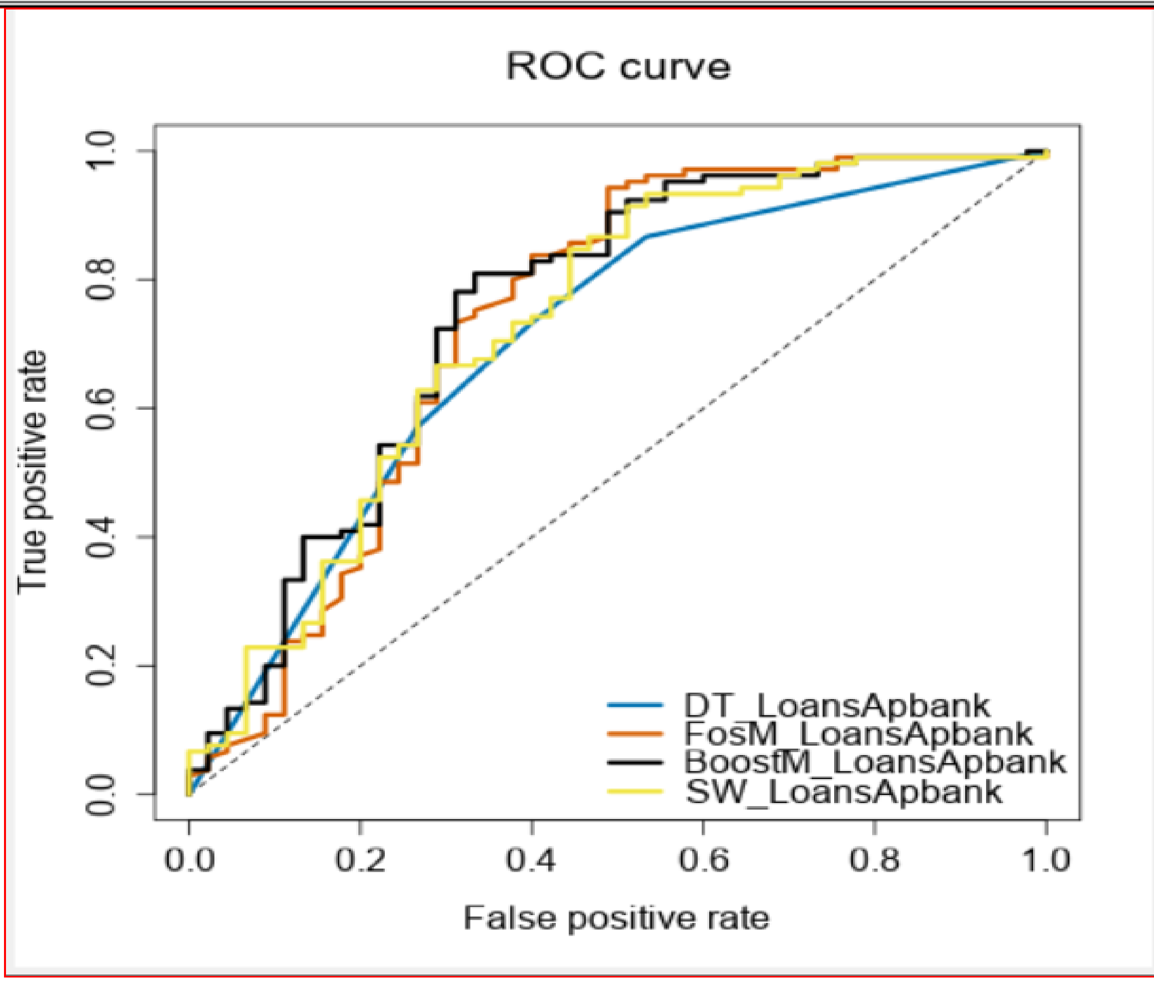
Confusion matrix of BoostM_LoansApbank	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of DT_LoansApbank	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of FosM_LoansApbank	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of SW_LoansApbank	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

: Awesome: Excellent work coming up with the correct confusion matrices.



: Awesome: Nice work coming up with the correct ROC curve.

The forest model reaches the highest real positive rate and is the fastest and most comprehensive ever. Despite the results between Boosted Model and Forest Model

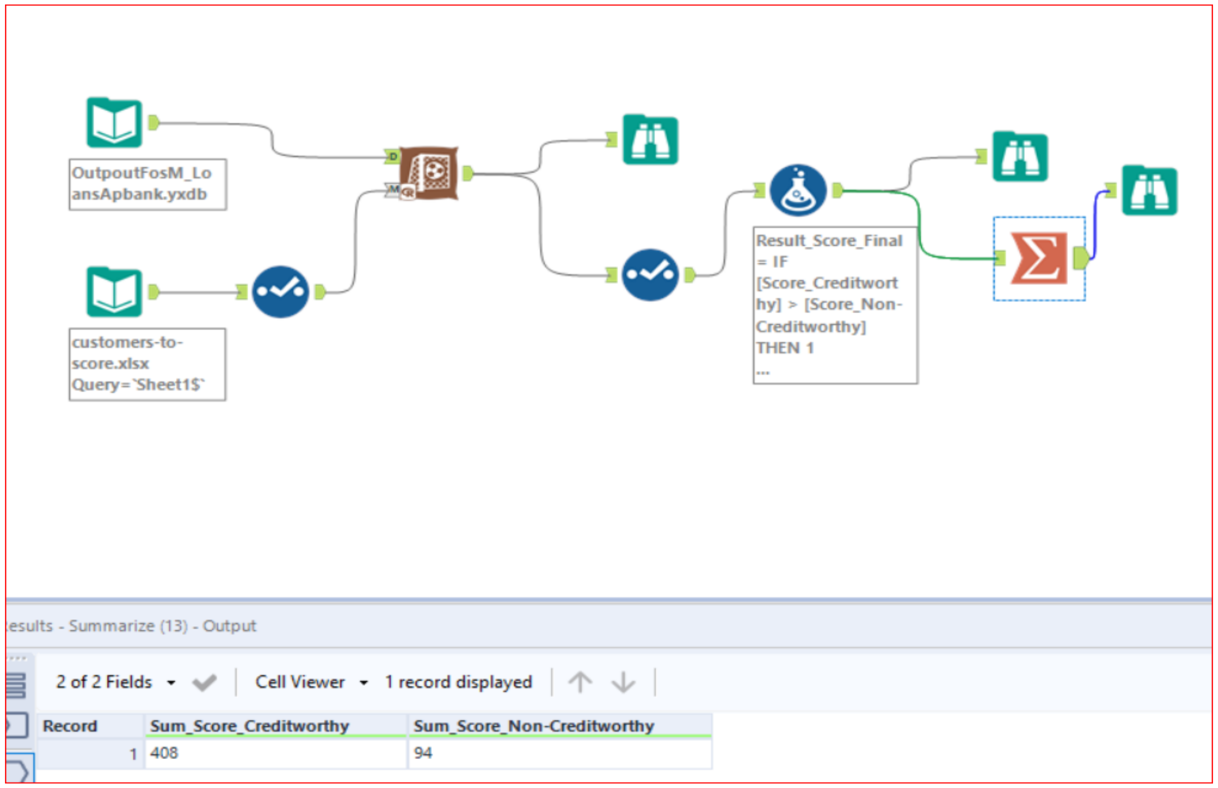
: Awesome: Well done correctly interpreting the ROC curve. This means that we are getting a higher rate of true positive rates vs. false positives. This is important because we do not want to extend loans to people who are not creditworthy.

: Comment: In addition, as you correctly noted, the forest model is low in bias, which is another advantage. We shouldn't be choosing models that are biased towards predicting individuals who are creditworthy, as they do not predict individuals who are not creditworthy nearly at the same level as those who are. This is bad for 2 reasons: 1. Loans will be extended to people who are not creditworthy leading towards bad loans 2. Opportunity will be missed by not extending loans to people who are creditworthy.

How many individuals are creditworthy?

There are 408 creditworthy new customers that we could approve for a loan and 94 noncreditworthy customers that should not be approved for a loan.

: Awesome: The final number of creditworthy individuals is absolutely correct.



I hope to be home to the project requirements despite the valuable information that we learned from the lessons and also the project, but we do not know the exact correct result

Help resources Forums :https://knowledge.udacity.com
https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

I wish success to all.

Marwan Saeed Alsharabbi