

# Predictive Analytics for Business

# Project #5 [A/B Test a New Menu Launch](https://classroom.udacity.com/nanodegrees/nd008-mena-connect/parts/458f8f11-d864-4894-b72a-e446001a7368#lesson-card-scroll-target-48312ab7-ffe9-415c-9cf1-7598f96bb361)

Name: Marwan Saeed Alsharabbi

Date: 12-Jan-2020

2020

**Introduction**

***A/B Testing Summary***

Summary: In an A/B test, a change is applied to a treatment group, and its performance is compared against a control group to estimate the impact of the change**.**

**STEP 1: SELECT A PERFORMANCE METRIC**

It’s important to understand the metric used to evaluate the results of the test. Whether the goal is to increase sales, profit, conversion rate, etc., this should be specified at the upfront**.**

**STEP 2: SELECT THE EXPERIMENT DESIGN**

Matched pair - when the sample size is small and/or the data is difficult to collect, a matched pair experiment should be used.Randomized design - when the sample size is large and the data is easy to collect, then a randomized experiment should be used. Randomized experiments are very common for web-based AB tests.

**STEP 3: SELECT TREATMENT AND CONTROL UNITS**

Each individual in the test is considered a unit. The unit can be a person, store, etc. In a test, units are split into two groups, the treatment group and control group. Treatment and control units are compared against each other

Useful Alteryx Tool: Formula

**STEP 4: SELECT EXPERIMENTAL AND CONTROL VARIABLES**

Experimental variable - The experimental, or treatment, variable(s) is the variable that is different between treatment and control units. For example, if you are testing a new price point, the experimental variable would be price.

Control Variables - The control variables are the variables that should remain constant between test and control groups. These variables ensure that the treatment and control groups are representative of each other and that the results will apply to the population. Control variables are used to match each treatment unit to one or more control units.

Useful Alteryx Tool: AB Controls

**STEP 5: DETERMINE TEST DURATION AND SAMPLE SIZE**

These two go hand in hand and contribute most directly to statistical significance. You can improve statistical significance by either increasing the sample size or test duration. Generally, the duration of a test should be at least as long enough to capture a representative sample.

STEP 6: RUN THE TEST AND PREPARE THE DATA

Now it’s time to run the test and collect the data. Preparing the data includes filtering for the dates of the test, ensuring there are no duplicate records, removing records with incomplete data, and removing outliers.

**STEP 7: ANALYZE RESULTS**

Lift - Compare the average performance between the two groups. It can also be useful to understand the distribution of the performance of the units.

Statistical Significance - Performing a t-test provides a p-value. P-values below 0.05, indicate statistically significant results. Paired t-test are used for matched pair experiments and unpaired t-test for randomized experiments.

Impact Estimation - In order to provide an expected impact of broad implementation of the treatment, apply the lift calculation to the entire population.

**Useful Alteryx Tool: AB Analysis**

### Project Overview

You're a business analyst for Round Roasters, a coffee restaurant in the United States of America. The executive team conducted a market test with a new menu and needs to figure whether the new menu can drive enough sales to offset the cost of marketing the new menu. Your job is to analyze the A/B test and write up a recommendation to whether the Round Roasters chain should launch this new menu.

#### How Do I Complete this Project?

This project uses skills learned throughout the "A/B Testing” course. To complete this project:

Go through the course

Apply the skills learned in the course to solve the business problem given in the project details section.

Use our guidelines and rubric to help build your project.

When you're ready, submit it to us for review using the submission template found in the supporting materials section.

#### Skills Required

In order to complete this project, you must be able to:

Cleanup, format, and blend a wide range of data sources

Plan and analyze A/B tests

**The Business Problem**

Round Roasters is an upscale coffee chain with locations in the western United States of America. The past few years have resulted in stagnant growth at the coffee chain, and a new management team was put in place to reignite growth at their stores.

The first major growth initiative is to introduce gourmet sandwiches to the menu, along with limited wine offerings. The new management team believes that a television advertising campaign is crucial to drive people into the stores with these new offerings.

However, the television campaign will require a significant boost in the company’s marketing budget, with an unknown return on investment ([ROI](http://www.investopedia.com/terms/r/returnoninvestment.asp)). Additionally, there is concern that current customers will not buy into the new menu offerings.

To minimize risk, the management team decides to test the changes in two cities with new television advertising. Denver and Chicago cities were chosen to participate in this test because the stores in these two cities (or markets) perform similarly to all stores across the entire chain of stores; performance in these two markets would be a good proxy to predict how well the updated menu performs.

The test ran for a period of 12 weeks (2016-April-29 to 2016-July-21) where five stores in each of the test markets offered the updated menu along with television advertising.

The comparative period is the test period, but for last year (2015-April-29 to 2015-July-21).

You’ve been asked to analyze the results of the experiment to determine whether the menu changes should be applied to all stores. The predicted impact to profitability should be enough to justify the increased marketing budget: at least 18% increase in profit growth compared to the comparative period while compared to the control stores; otherwise known as *incremental lift*. In the data, profit is represented in the *gross\_margin* variable.

You have been able to gather three data files to use for your analysis:

* Transaction data for all stores from 2015-January-21 to 2016-August-18
* A listing of all Round Roasters stores
* A listing of the 10 stores (5 in each market) that were used as test markets.

**Steps to Success**

**Step 1: Plan Your Analysis**

To perform the correct analysis, you will need to prepare a data set. Prior to rolling up your sleeves and preparing the data, it’s a good idea to have a plan of what you need to do in order to prepare the correct data set. A good plan will help you with your analysis. Here are a few questions to get you started:

* What is the performance metric you’ll use to evaluate the results of your test?
* What is the test period?
* At what level (day, week, month, etc.) should the data be aggregated?

**Step 2: Clean Up Your Data**

In this step, you should prepare the data for steps 3 and 4. You should aggregate the transaction data to the appropriate level and filter on the appropriate data ranges. You can assume that there is no missing, incomplete, duplicate, or dirty data. You’re ready to move on to the next step when you have weekly transaction data for all stores.

**Step 3: Match Treatment and Control Units**

In this step, you should create the trend and seasonality variables, and use them along with you other control variable(s) to match two control units to each treatment unit. Treatment stores should be matched to control stores in the same region. *Note: Calculate the number of transactions per store per week and use 12 periods to calculate trend and seasonality.*

Apart from trend and seasonality...

* What control variables should be considered? Note: Only consider variables in the RoundRoastersStore file.
* What is the correlation between your each potential control variable and your performance metric? (Example of correlation matrix below)
* What control variables will you use to match treatment and control stores?

**Step 4: Analysis and Writeup**

Conduct your A/B analysis and create a short report outlining your results and recommendations.

In an AB Analysis we use the correlation matrix to find the most correlated variable to the performance metric to include in the AB controls tool to help find the best matches.

### A/B Test Guide

The A/B test guide can help you map out the process you'll need to go through to complete an A/B test. This will be useful for the project, as well as any A/B tests you do in a professional setting.

**Data**

*round-roaster-stores.csv* - This file contains store information for each Round Roaster store in the USA.

*treatment-stores.csv* - This file contains store information for each store that offered the new menu items.

*round-roaster-transactions.zip* - This file contains transaction level information for all of Round Roaster's stores

**Aggregate and Export**

We recommend you save your aggregated transaction database as a separate file to further reduce your development time. You shouldn't need to re-aggregate the transaction database every time you want to test out a new workflow.

**Creating a week variable**

When calculating the week column, assign week 1 to the first week of the test period, so the week\_begin variable for week 1 would be your test start date. This will cause weeks prior to the test period to be negative, which is perfectly fine. See [here](https://classroom.udacity.com/nanodegrees/nd008-ent/parts/840e51ff-4c1b-41bd-bc02-0a659921b756/modules/bbde5353-d851-4624-bca9-75e8a07e6f10/lessons/ee1d5344-e0f5-41bd-90af-30695d0458a8/concepts/e93ee789-fb71-486c-8b0f-cd0de9b080e0) . If the first link doesn't work, please click [here](https://help.alteryx.com/current/Formula.htm) for an example of how to calculate the week variable using a formula tool in Alteryx.

**Number of Weeks**

The trend tool is used to create trend and seasonality variables to use as control variables. To do this, you need at least 52 weeks of data, plus the number of weeks you select in the tool to calculate trend, before the beginning of the test start date. In lesson 4, you used 6 weeks to calculate the trend, so you needed 58 weeks prior to the test start date. For the project, you are asked to use 12 weeks to calculate trend, so you'll need 64 weeks of data prior to the test start date. Since the test lasts for 12 weeks, this means you'll need a total 76 weeks of data.

Your filter at the beginning of your workflow should go back 76 weeks from the end of the test period 2016-July-21. In Alteryx the expression shuold look like [Invoice Date]>="2015-02-06" AND [Invoice Date]<"2016-07-22"

All stores should have 76 weeks of Data.

**AB Trend Tool**

The Test start date is 2016-April-29

The performance metric for this tool is the invoice count per week which represent weekly foot traffic. You had to create this variable with a summarize tool.

**AB Controls Tool**

You should use 3 numeric measures to match treatment and control stores.

1. Trend
2. Seasonality
3. AvgMonthSales (This should be determined by looking at the correlation between the appropriate numeric variables in the round roasters stores file AvgMonthSales and Sq\_ft with the performance metric gross margin.) - This variable is in the round-roaster-stores.csv file

**AB Analysis Tool**

Make sure to use weekly gross margin per store in all lift calculations and not total sales. Your data has gross margin in it but you will have to use a summarize tool to get weekly gross margin per store.

**Steps**

1. Filter the data to the proper date range
2. Aggregate the data to get the weekly gross margin and weekly traffic count (count of unique invoices)
3. Calculate Trend and Seasonality with the AB Trend Tool
4. Label the data as treatment and control stores
5. Calculate correlation between other numeric measures and the performance metric (gross margin)
6. Match Treatment to control stores per region using the AB Controls Tool
7. Calculate lift from control to treatment store with AB Analysis tool

If you run into errors in Alteryx or unexpected results from a tool we have a guide to help you figure out what is going on.

**Alteryx Debugging Guide:** See the Resources tab in the left most panel of your classroom for a downloadable PDF of the Alteryx Debugging Guide

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.

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.

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.

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.

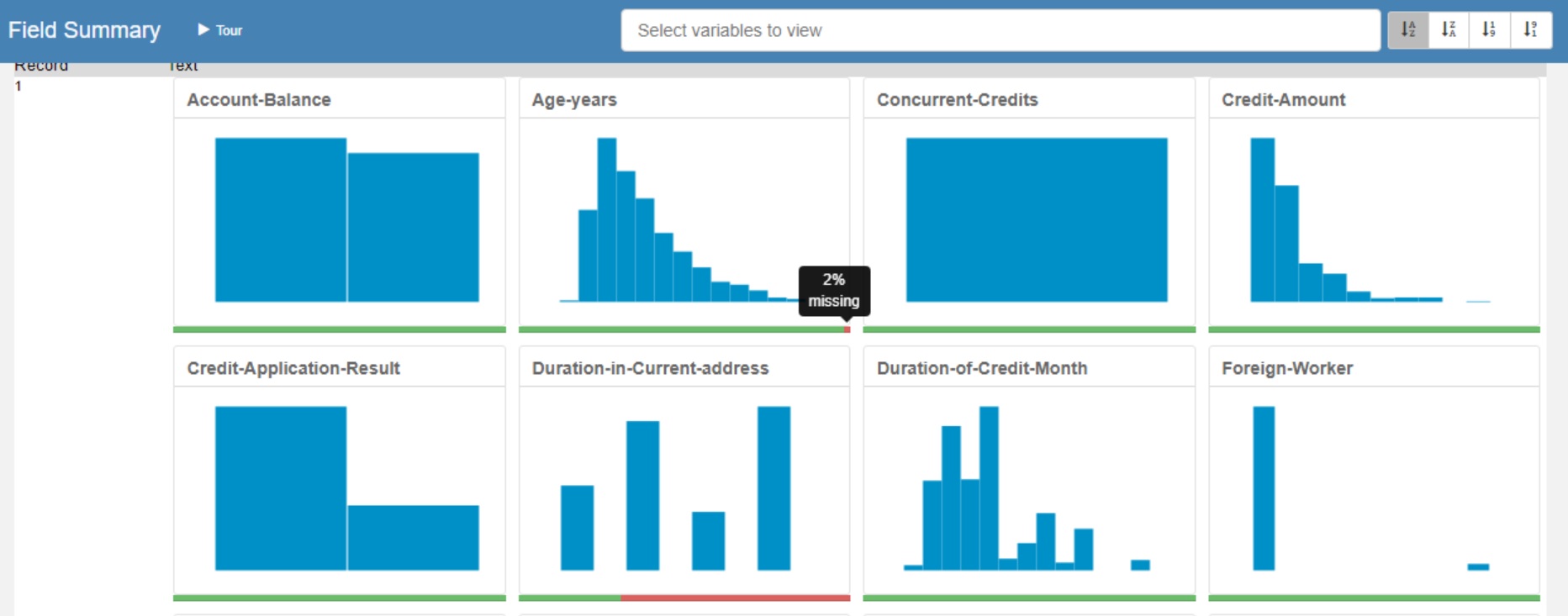
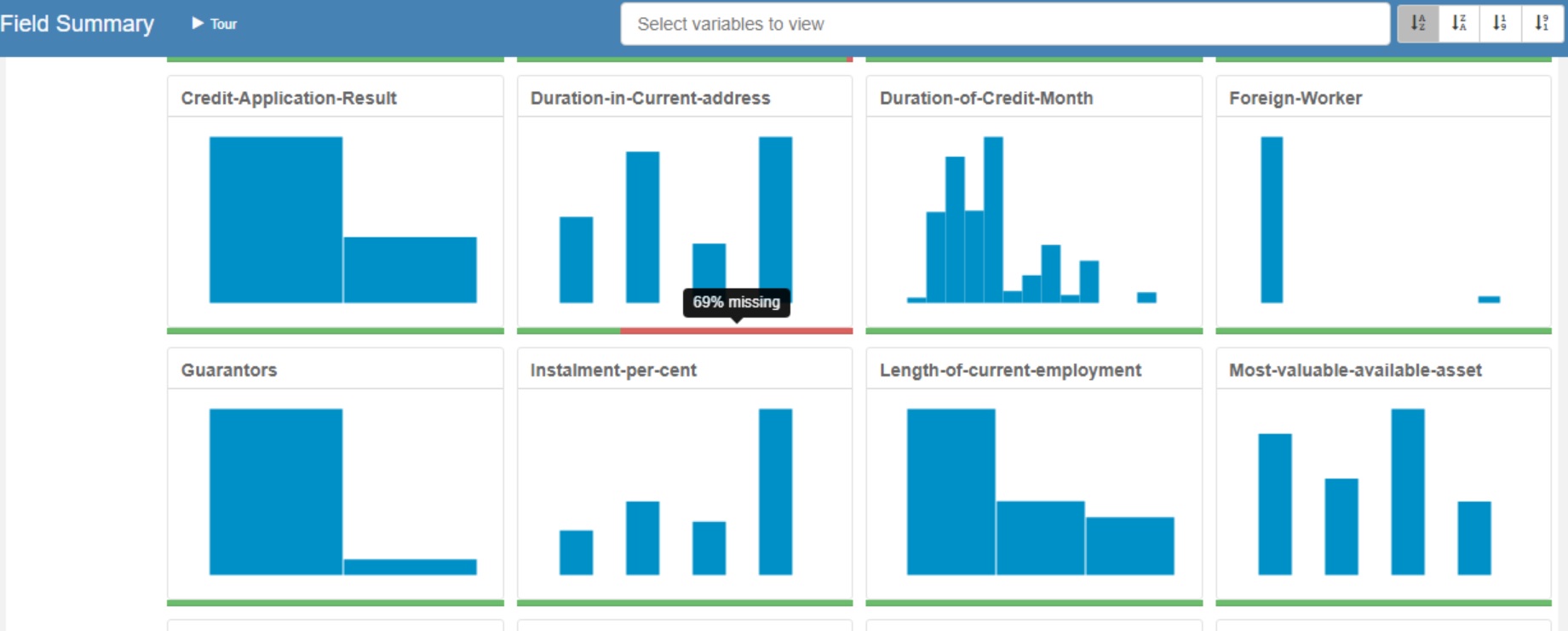
In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields.

**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



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

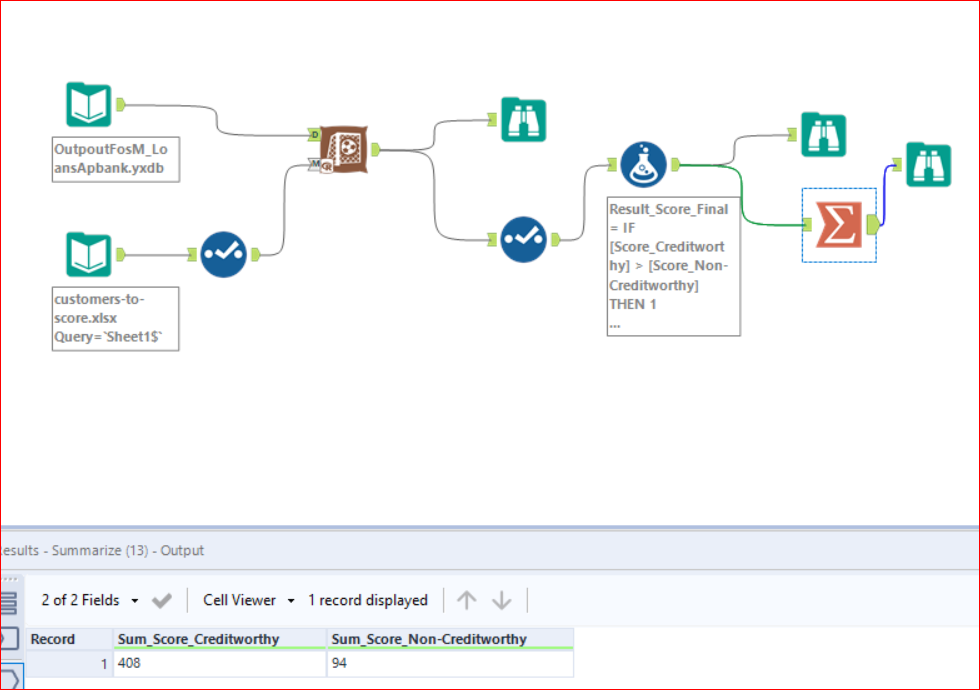
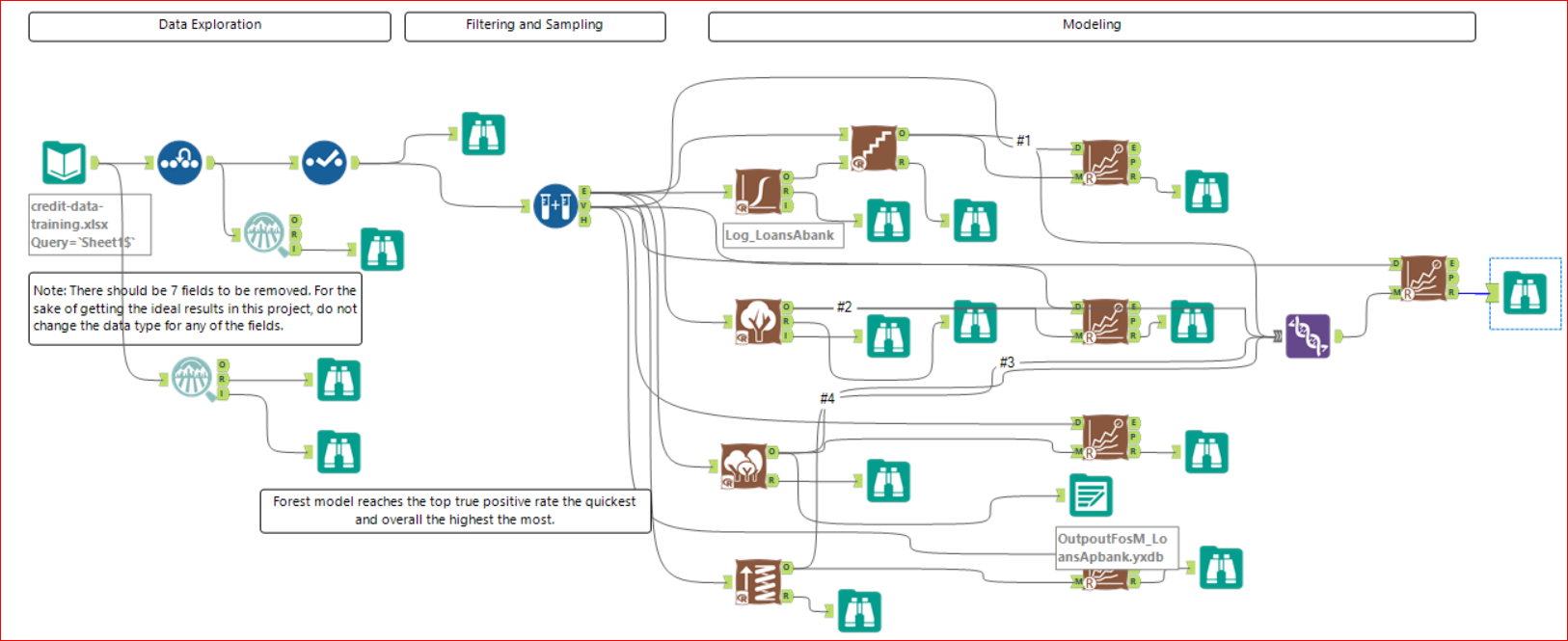
*Answer these questions for* ***each model*** *you created:*

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?

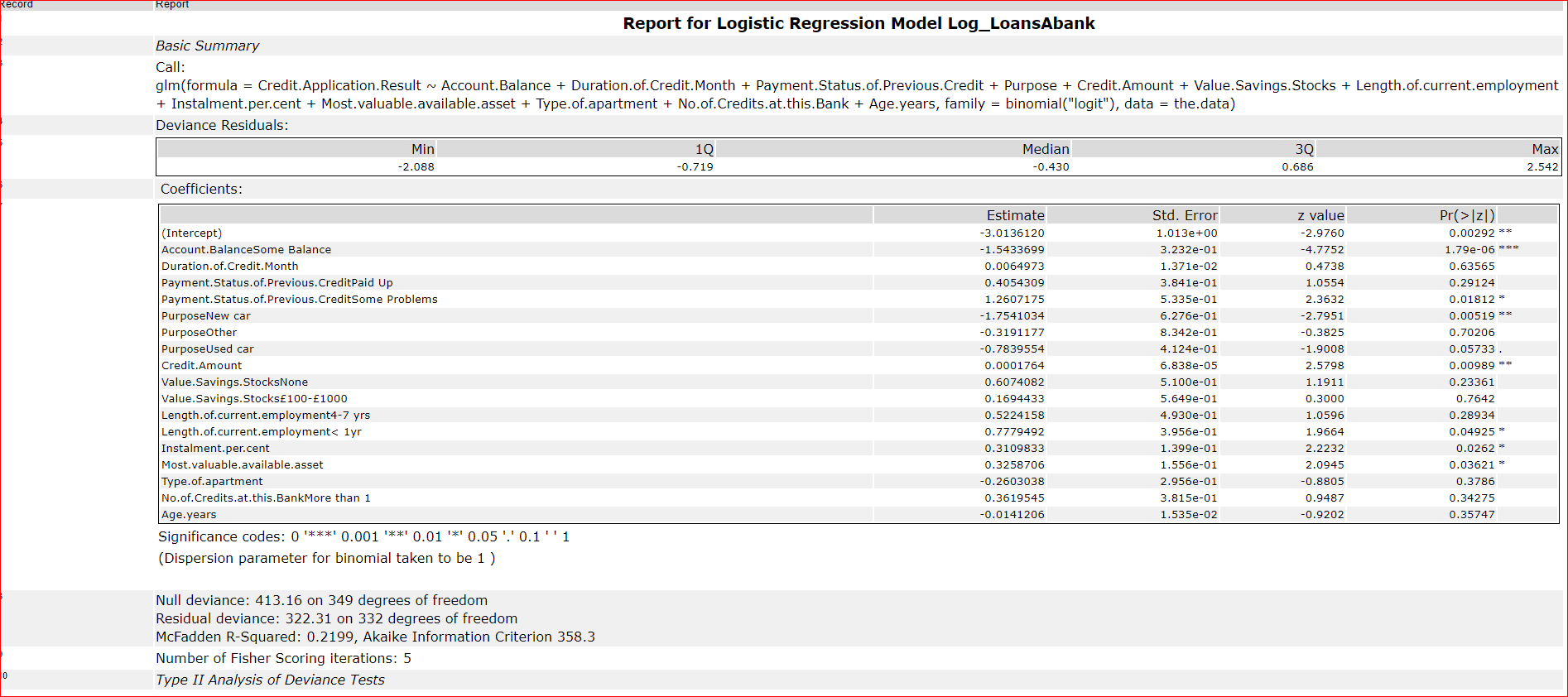
Train your Classification Models:

**Alteryx Workflow**



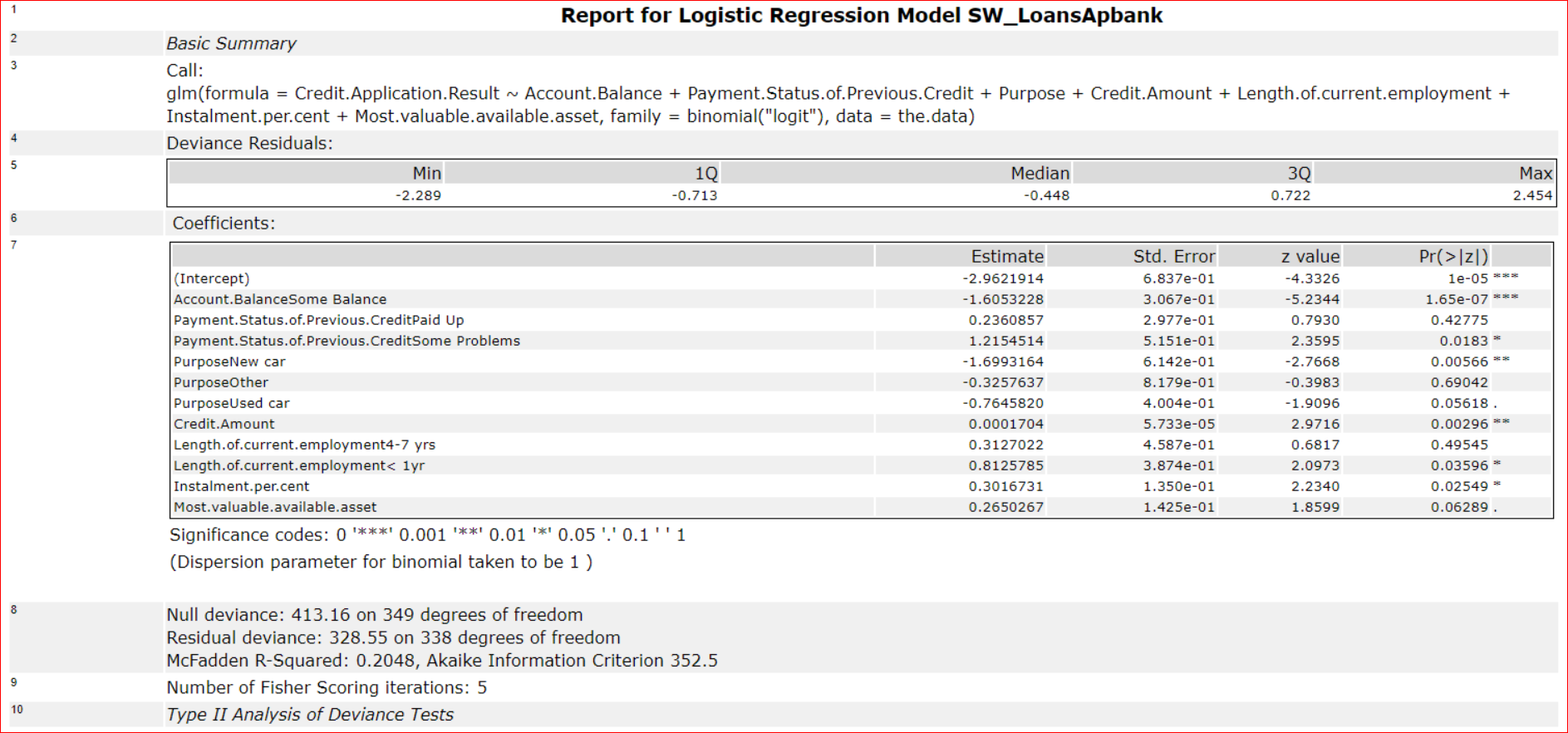
1-Logistic Regression

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

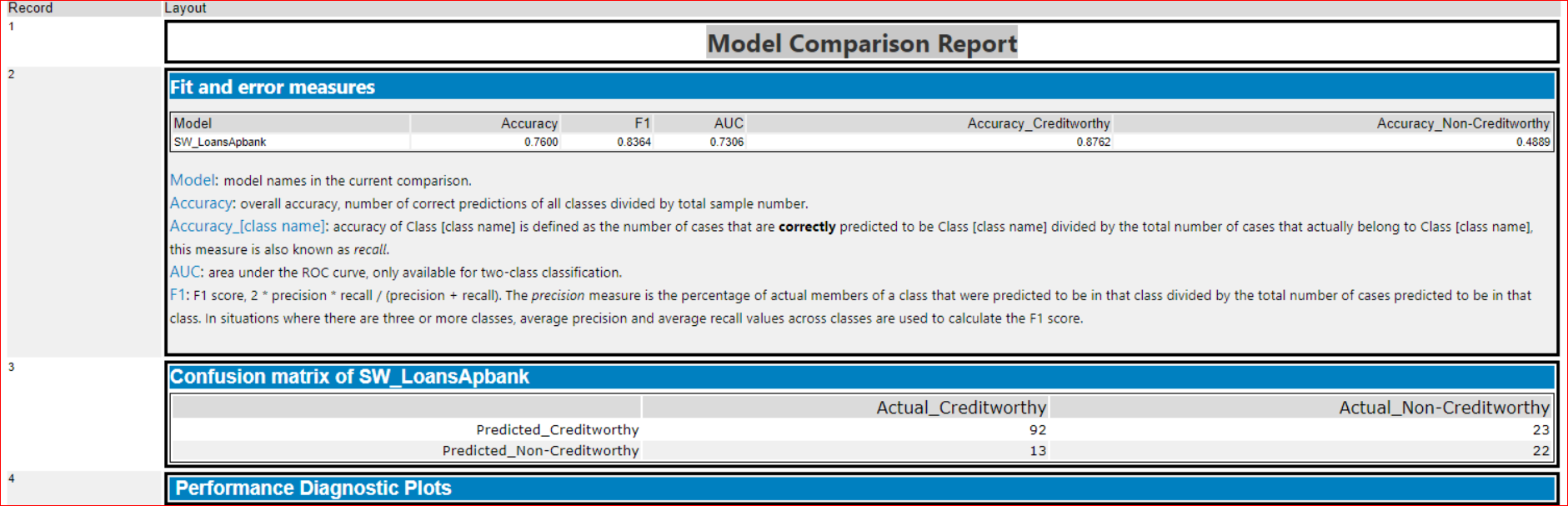


2- Logistic Regression -Stepwise

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



Using the confusion matrix,

accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 92/ (92+23) = 0.8, 80% while

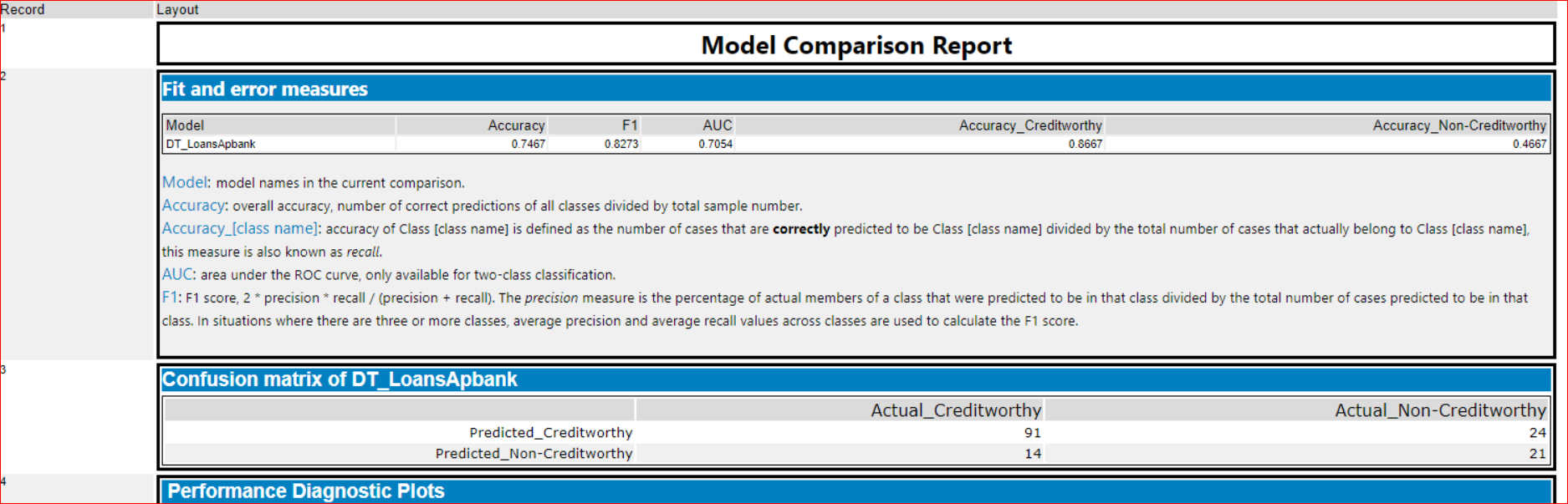
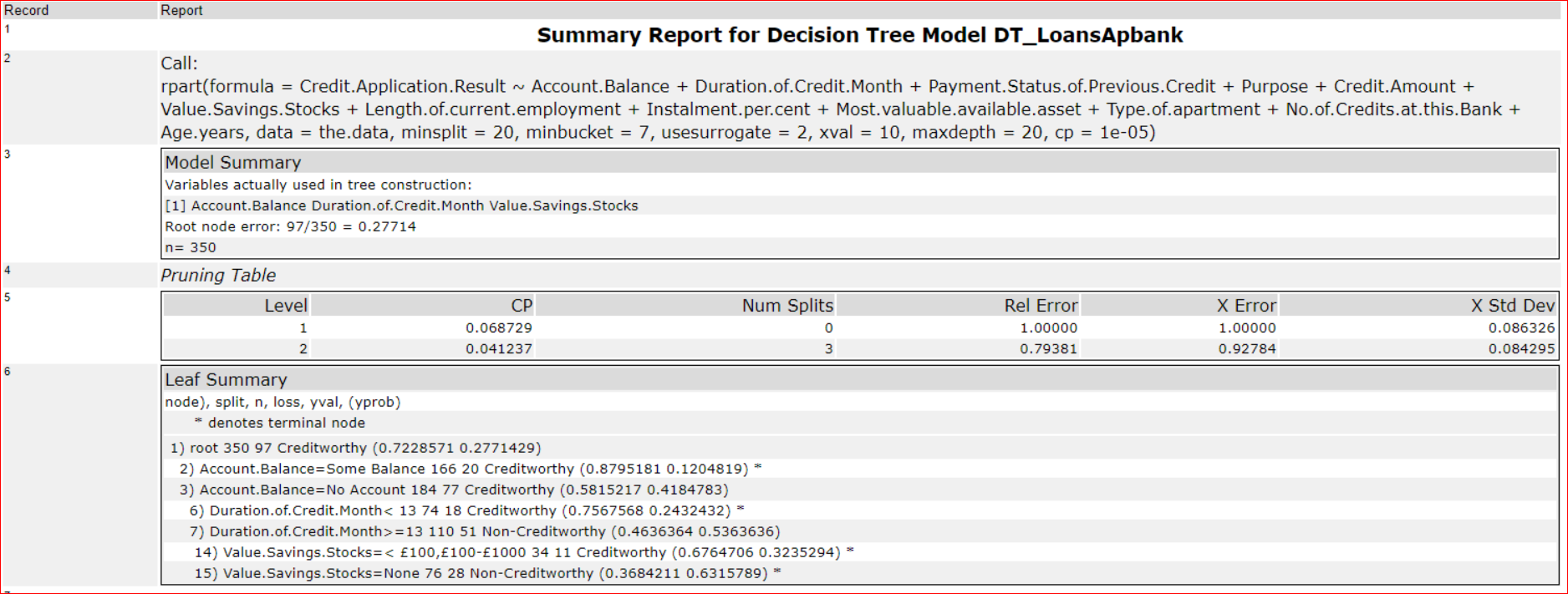
accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 22/ (13+22) = 0.6286, 62.86%

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

3-Decision Tree

Decision Tree can help predict classification of categorical or continuous variablesin 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 datain 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%.

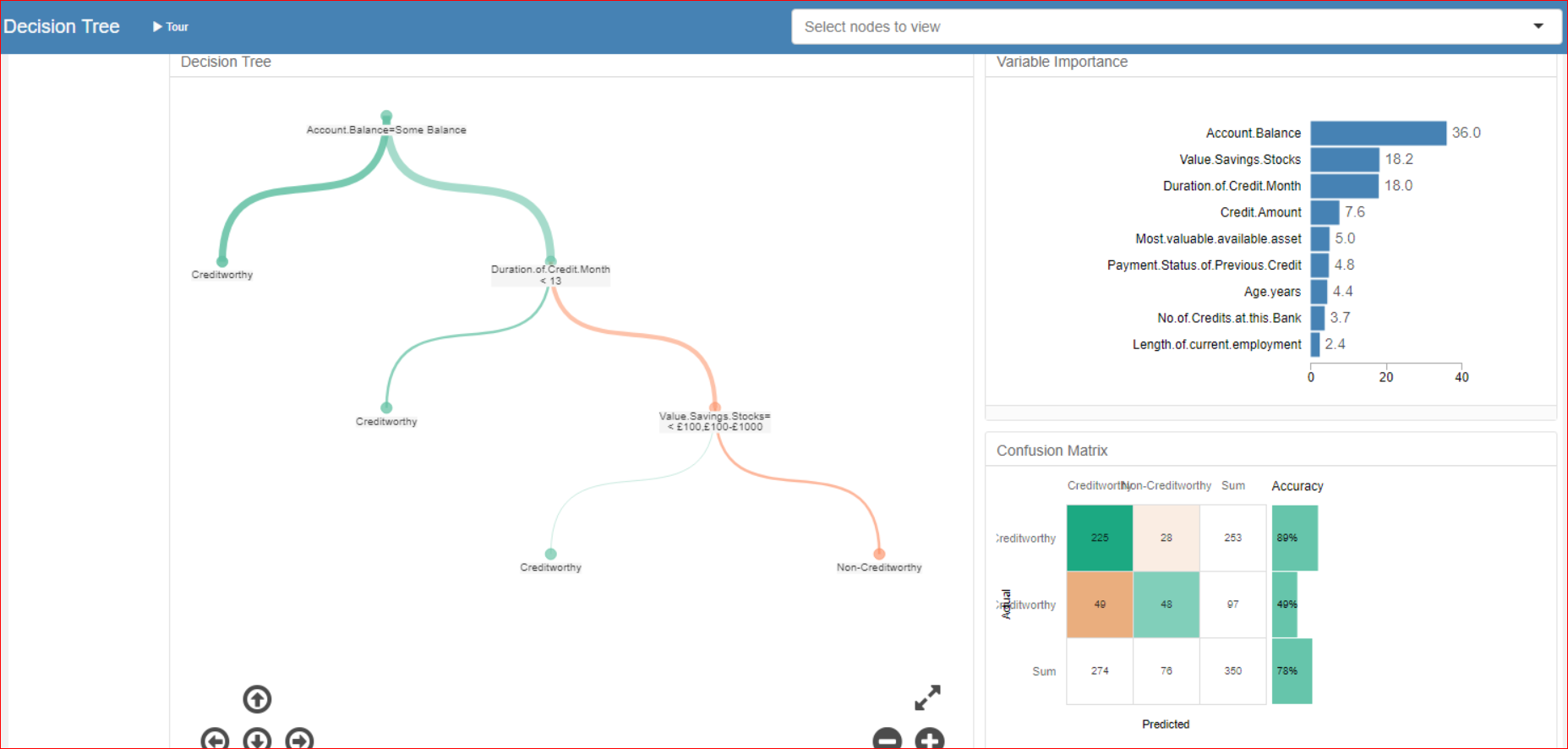


Using the confusion matrix,

**accuracy for creditworthy** = actual creditworthy / (predicted creditworthy) = 91/ (91+24) = 0.7913, 79.13%

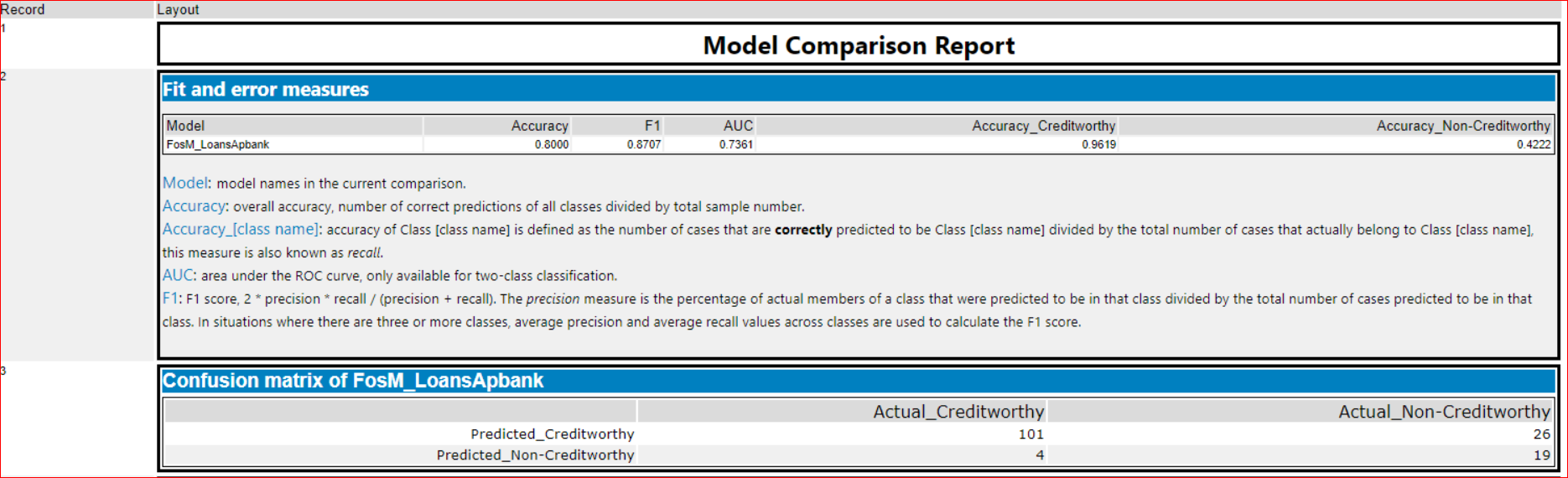
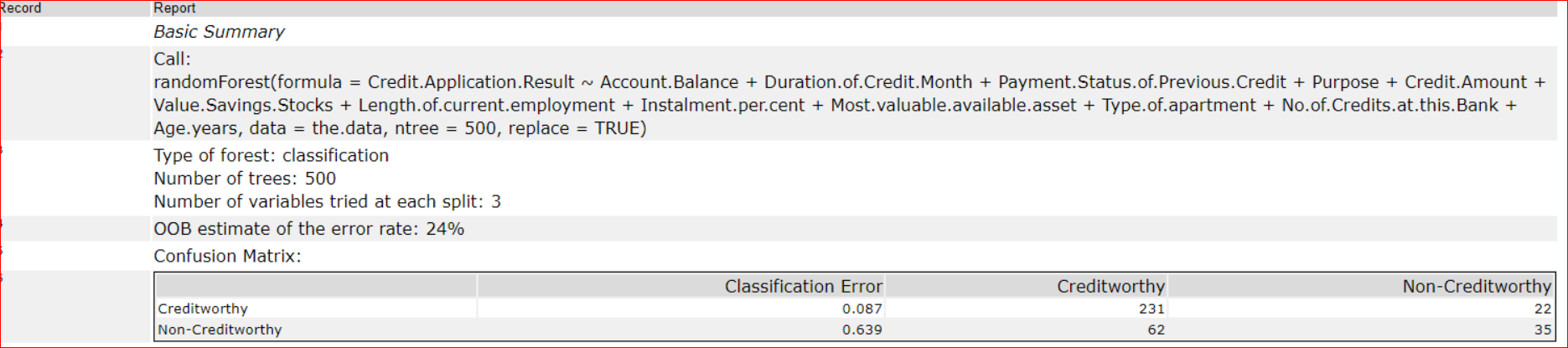
**accuracy for non-creditworthy** = actual non-creditworthy / (predicted non-creditworthy) = 21/ (14+21) = 0.6, 60%

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



4-Forest Models

Forest models can help predict classification of categorical or continuous variablesin 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 datain 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%.

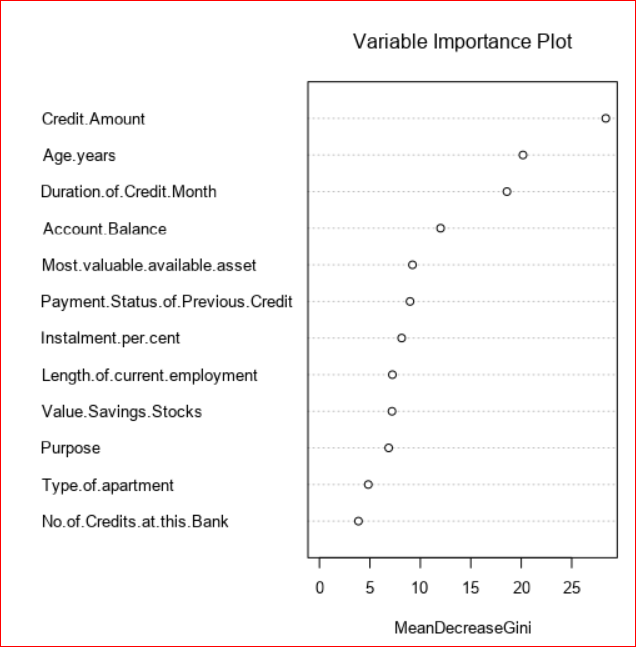
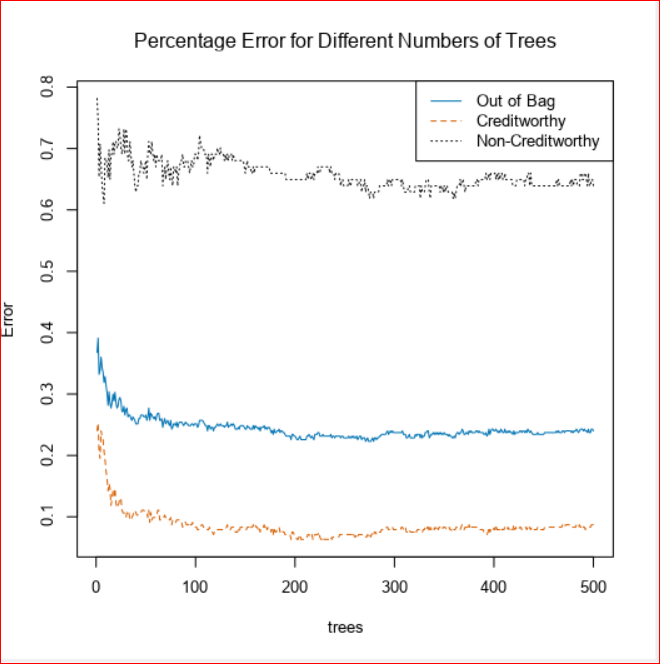


Using the confusion matrix,

**accuracy for creditworthy** = actual creditworthy / (predicted creditworthy) = 101/ (101+26) = 0.7952, 79.52%

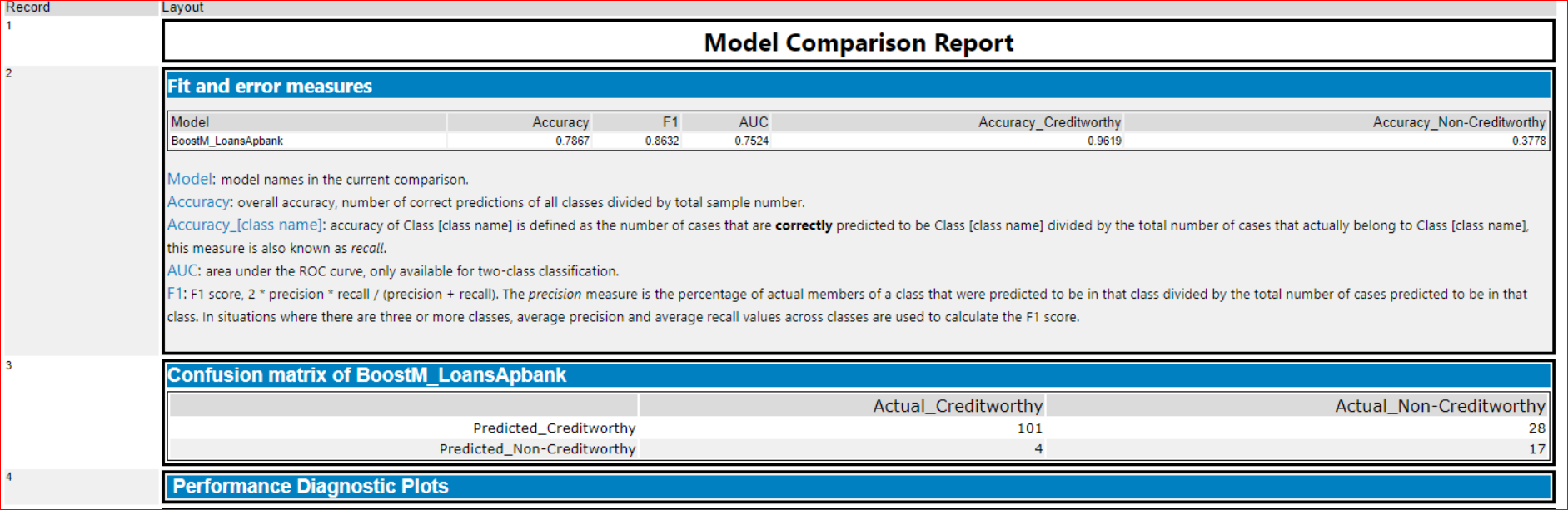
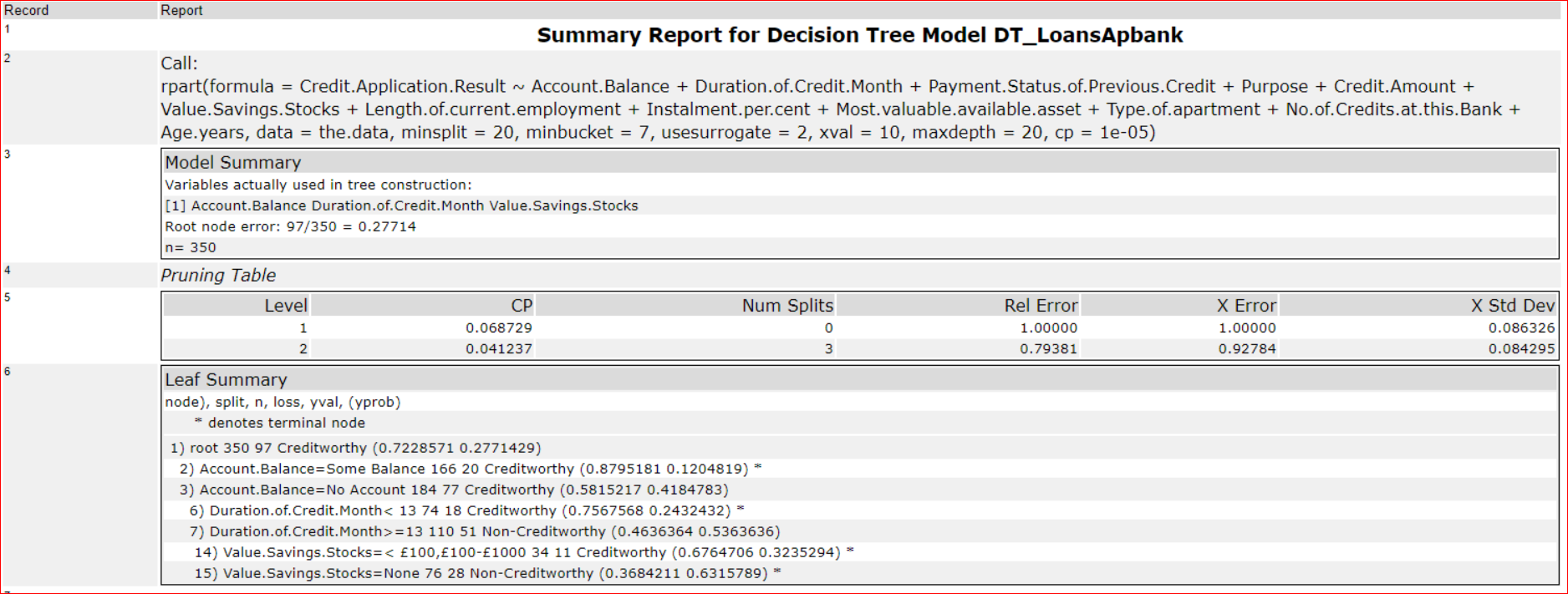
**accuracy for non-creditworthy** = actual non-creditworthy / (predicted non-creditworthy) = 19/ (19+4) = 0.8260, 82.60%

Since accuracies for creditworthy and non-creditworthy are comparable 79.52% and 86.37% respectively, this model isn’t biased



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%.



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.



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

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)*

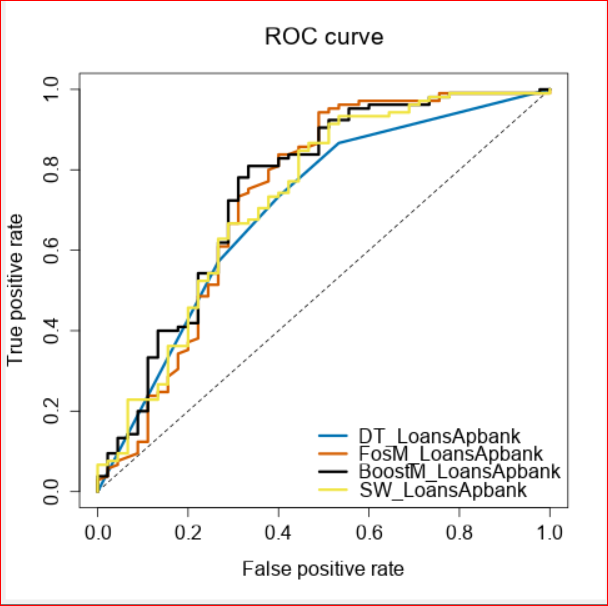
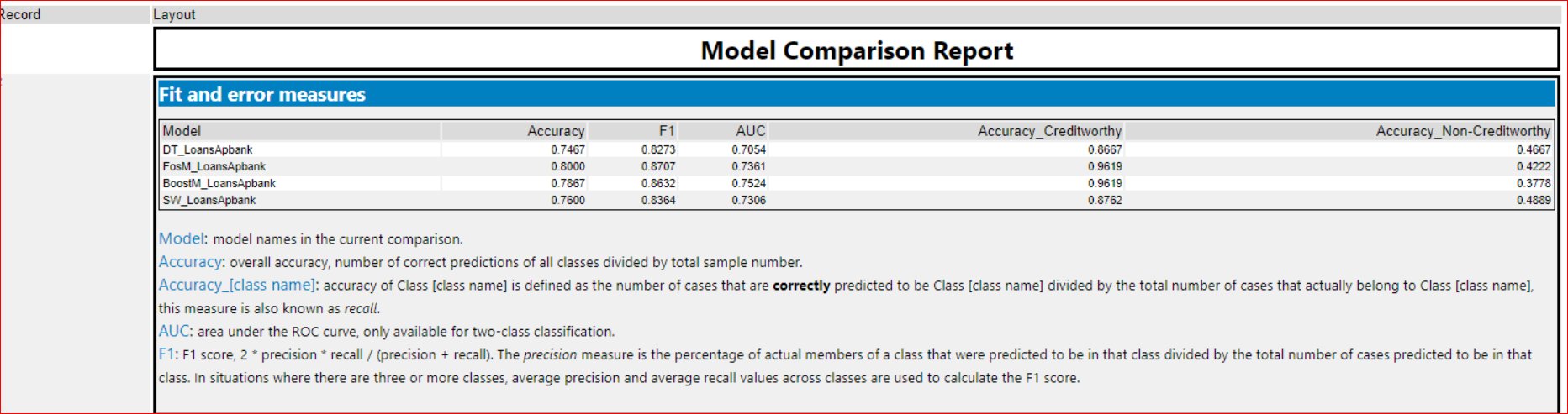
*Answer these questions:*

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.

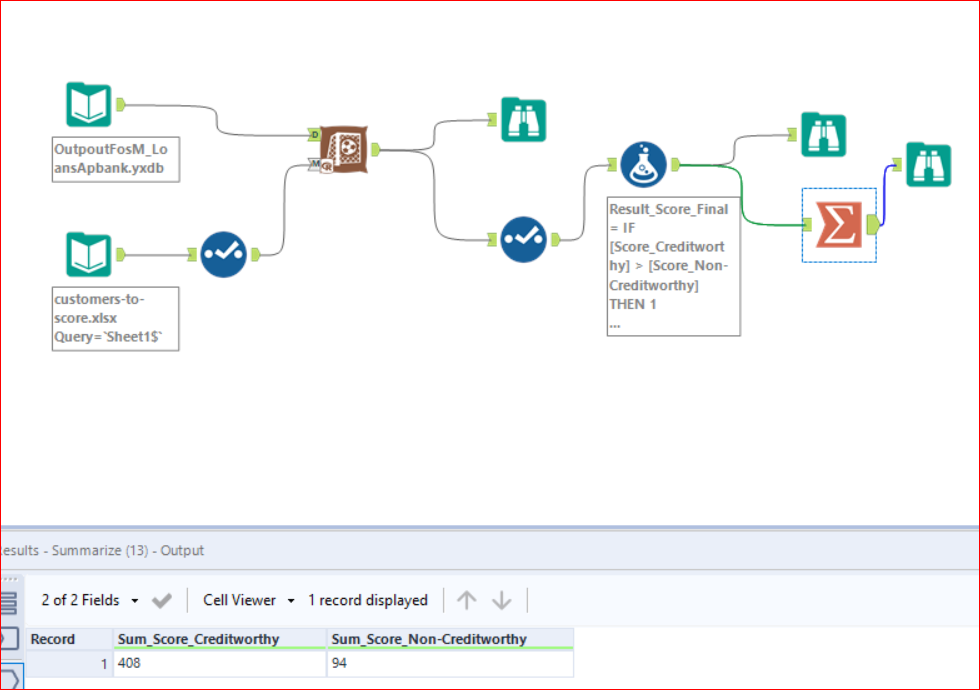
Forest Model has been chosen since it has the highest accuracy of 80% among all four classification models. Also, accuracies for creditworthy and non-creditworthy are among the highest of all



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

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.



*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.**

M*arwan Saeed* *Alsharabbi*