9/17/2019

Simran Chhabra

for Edwisor.com

Santander Customer Transaction Prediction

Contents

1. **Introduction**
   1. Problem Statement
   2. Data
2. **Methodology**
   1. Pre-Processing
      1. Missing values
      2. Outlier Analysis
      3. Feature Selection
   2. Modelling
      1. Visualization
      2. Logistic Regression
      3. Naïve Bayes
      4. Decision Tree Classifier
      5. Random Forest Classifier
3. **Conclusion**
   1. Model Evaluation
   2. Model Selection

**Appendix A – Python Code**

**Appendix B – R Code**

**References**

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

*“At Santander, mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals.”*

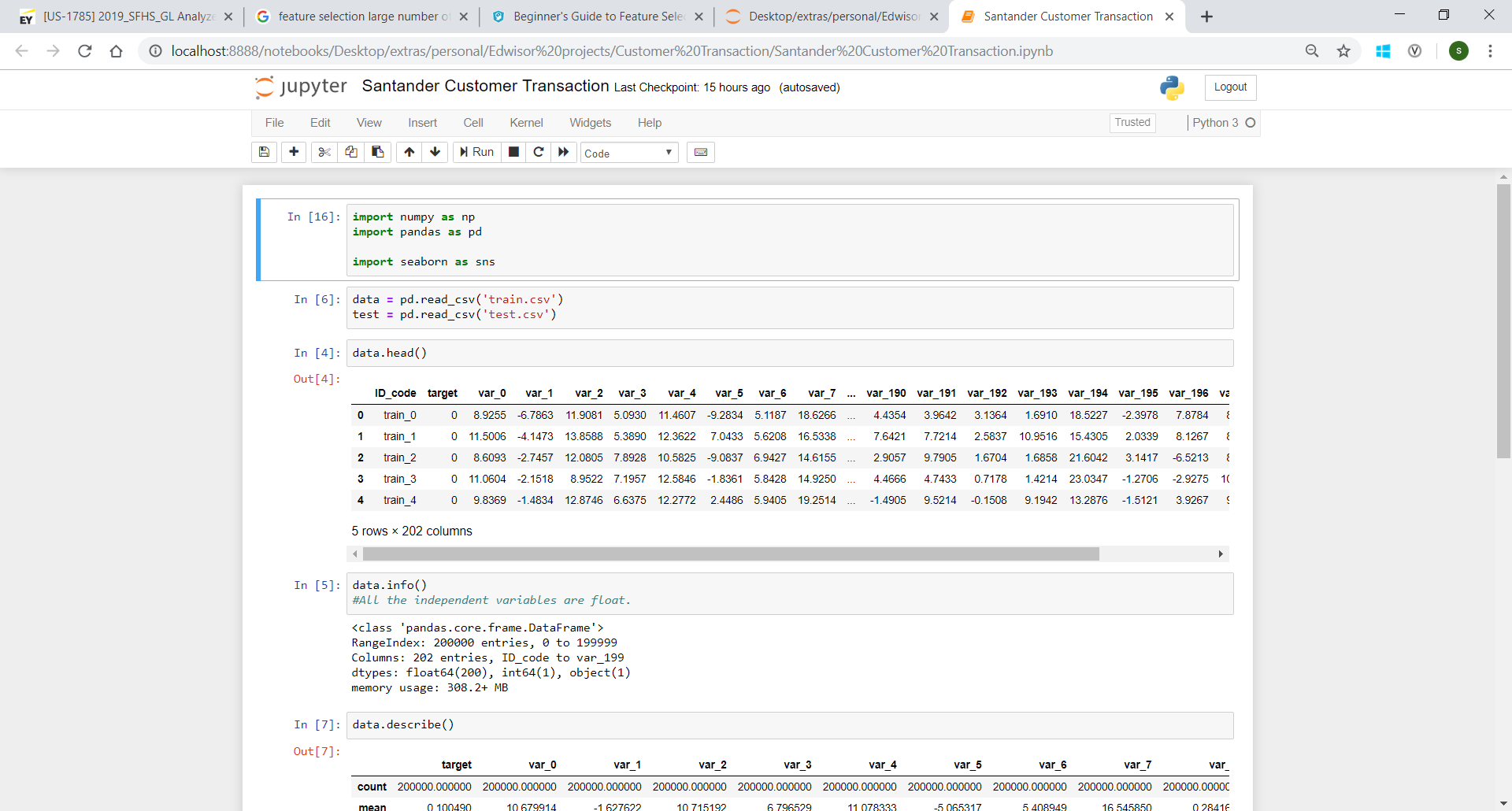
In this challenge, we need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

**1.2 Data**

1) train.csv; 200k rows

2) test.csv; 200k rows

We are provided with an anonymized dataset containing 200 numeric feature variables, a binary target column, and a string ID\_code column. The task is to predict the value of target column in the test set.



**Chapter 2**

**Methodology**

* 1. **Pre-Processing**

Any predictive modelling requires that we look at the data before we start modelling. However, in Data Mining terms *looking at data* means so much more than looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start with, we first detect anomalies in data like missing values, etc. On observing the data on a high level, it was observed that the variable “ID\_code” is identifying each row uniquely. It, in no way helps in predicting the target variables thus it can be dropped.

**2.1.1 Missing values**

Ideally if a variable has less than 30% missing values then they are filled using appropriate metrics like mean, median, mode, K-nearest neighbours etc. This will preserve the data and there will be no loss of information. We observed that the provided data was complete as there were no missing values.

**2.1.2 Outlier Analysis**

IQR score method was used to detect outlier in each variable.

*The****interquartile range****(****IQR****), also called the****midspread****or****middle 50%****, or technically****H-spread****, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR =*Q*3 −*Q*1.*

*In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data.*

*It is a measure of the dispersion similar to standard deviation or variance but is much more robust against outliers.*

The detected outliers were corrected by first replacing them with nulls and then filling these null values with mean. Standard deviation was used to check if the process carried out was right.

* + 1. **Feature Selection**

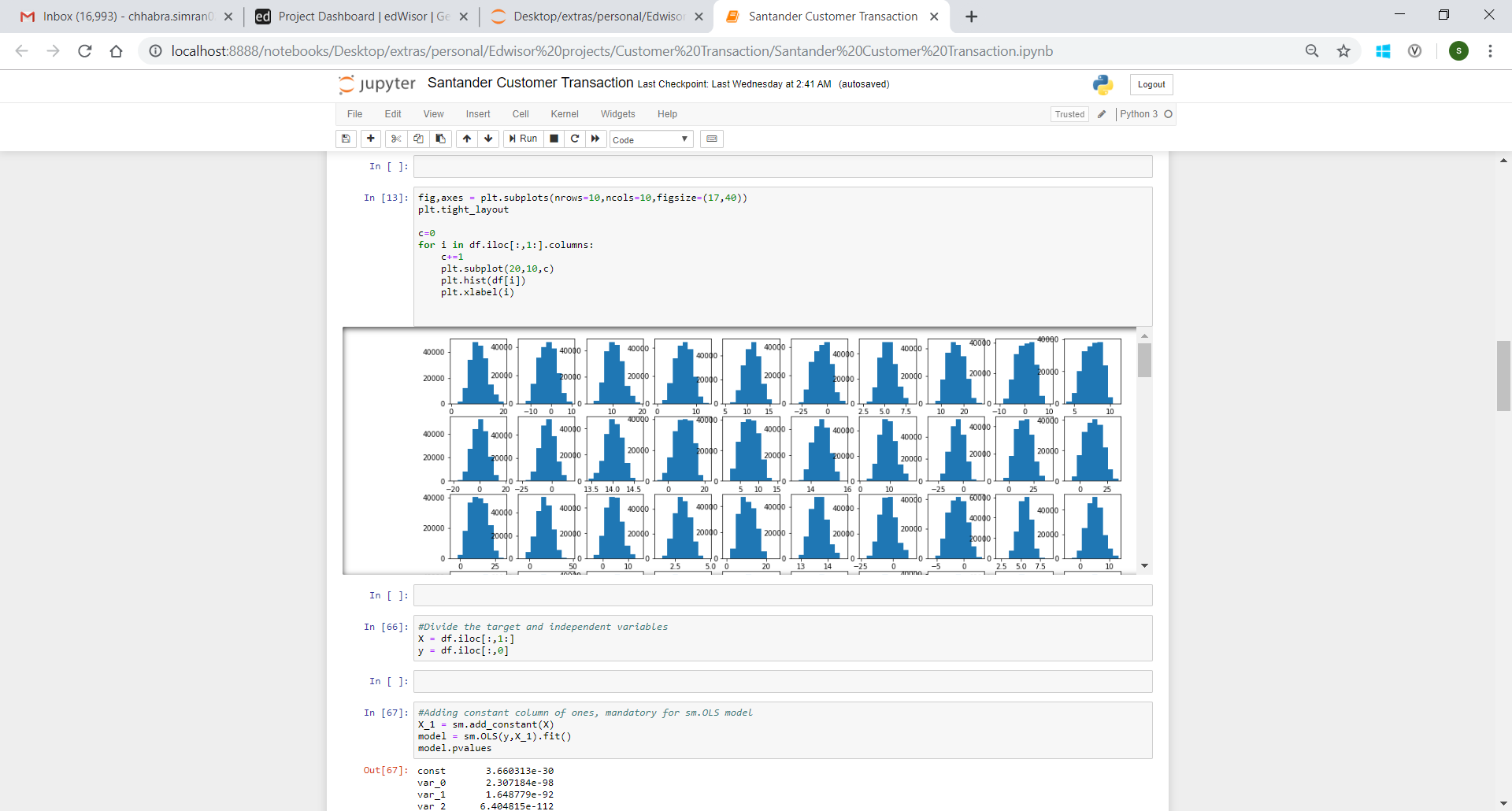
There could be variables in the data which could not be used directly in the analysis. We try to retrieve helpful information from such data. Then all the non-required fields are removed from the data.

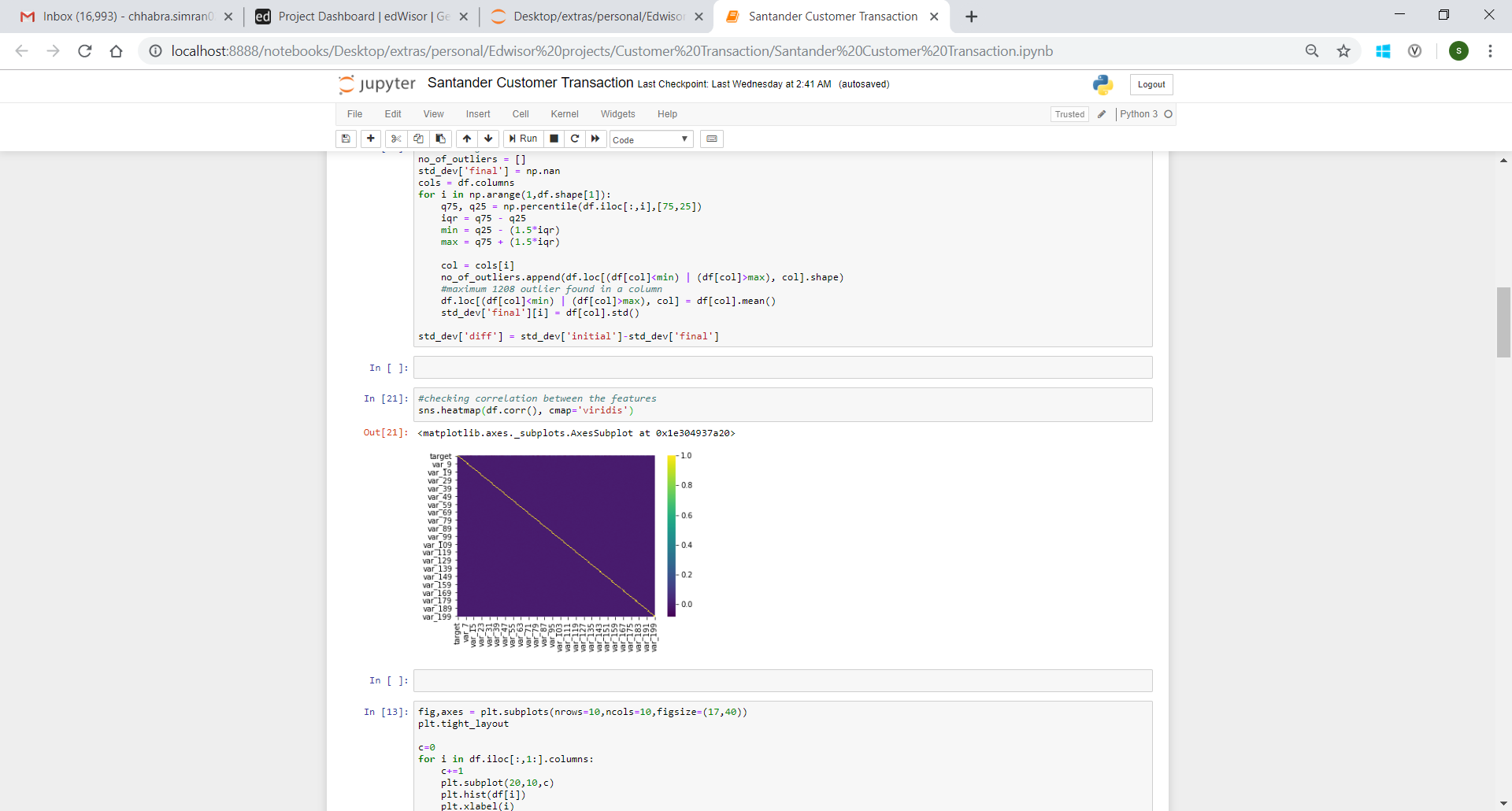
1. To check the inter-dependence of features on high level, we first created heatmap plot on the correlation values. Correlation values seemed to be around 0 for most values.
2. To drop out irrelevant features from our data, Backward elimination method was used. For doing this, pvalues were calculated as the performance metric to evaluate feature performance. If the pvalue of a feature was above 0.05 then the feature was removed, else it was kept. 179 features were left after performing this exercise.
   1. **Modelling**

As the data is cleansed now, next step is to apply different models on it and find the most suitable one to be used for further prediction.

* + 1. **Visualization**

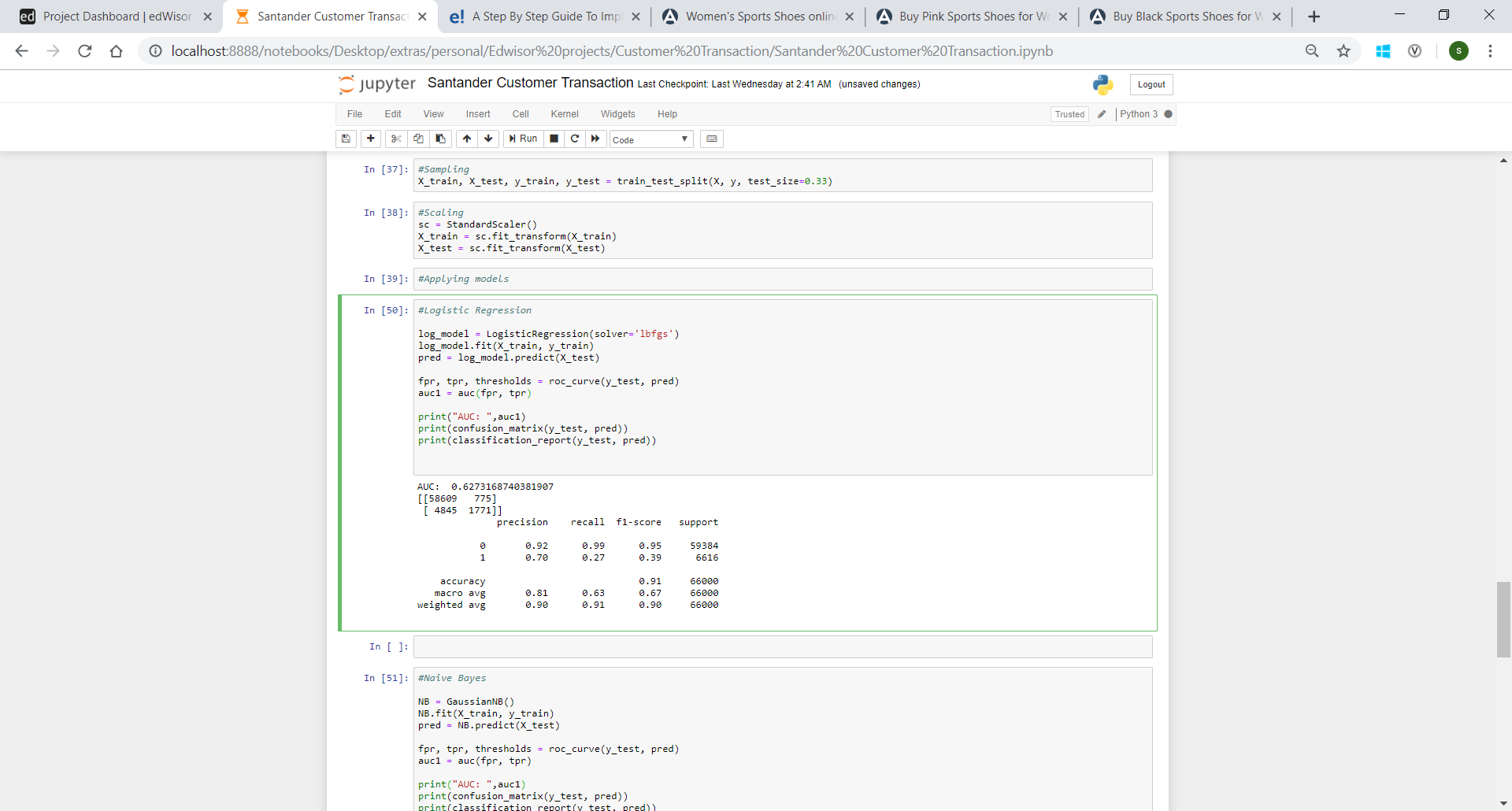
Visualization tools are used to make sure that the data is normalized, there is no correlation between the independent variables etc.





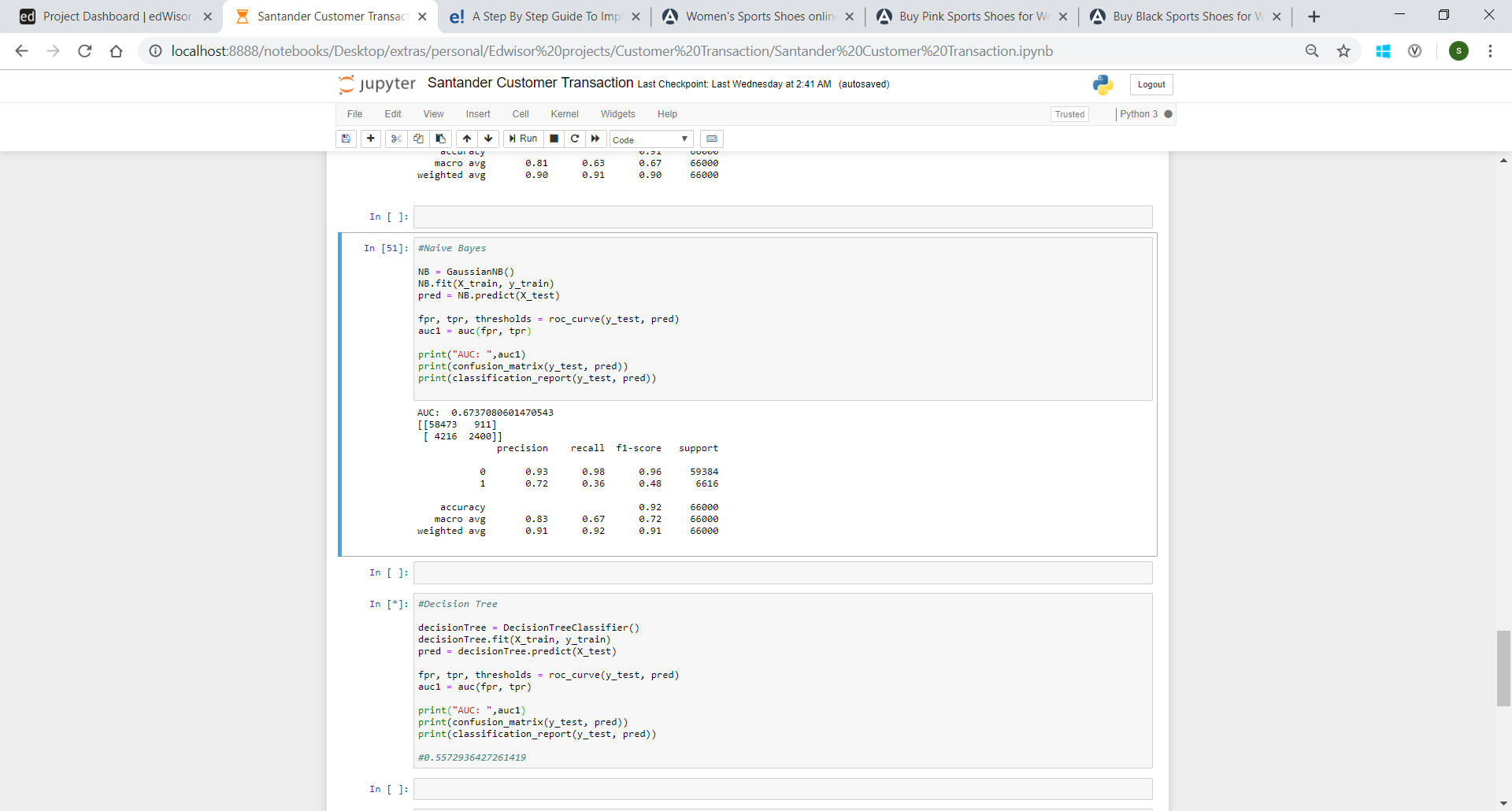
* + 1. **Logistic Regression**

First model used is Logistic regression. We have checked in previous section that the data is normalised, and we have also made it free of correlation. We get following confusion matrix and classification report on performing logistic regression:



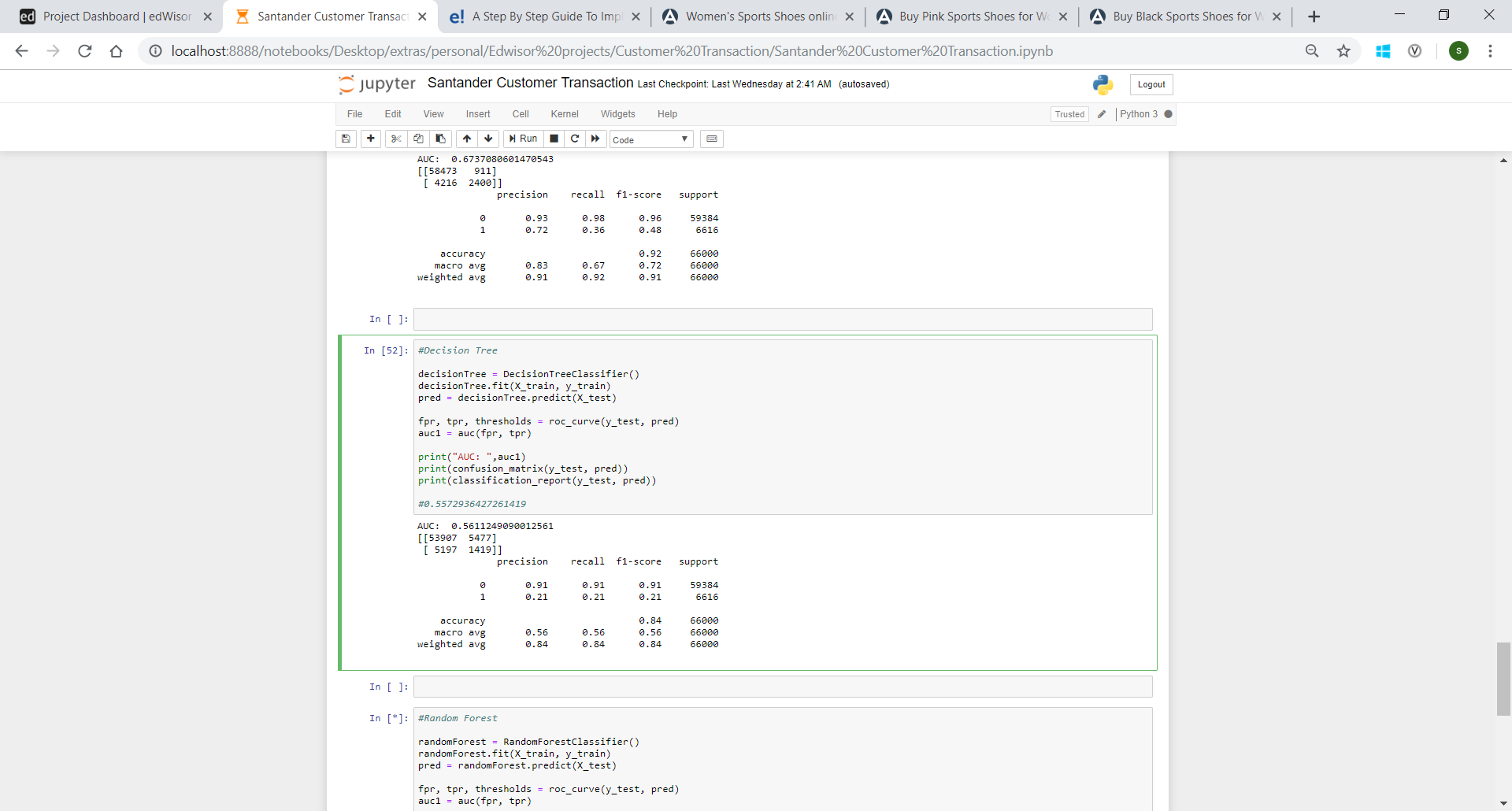
* + 1. **Naïve Bayes**

Naïve Bayes model, another classification model is then used to predict the target variable. We get the following results:



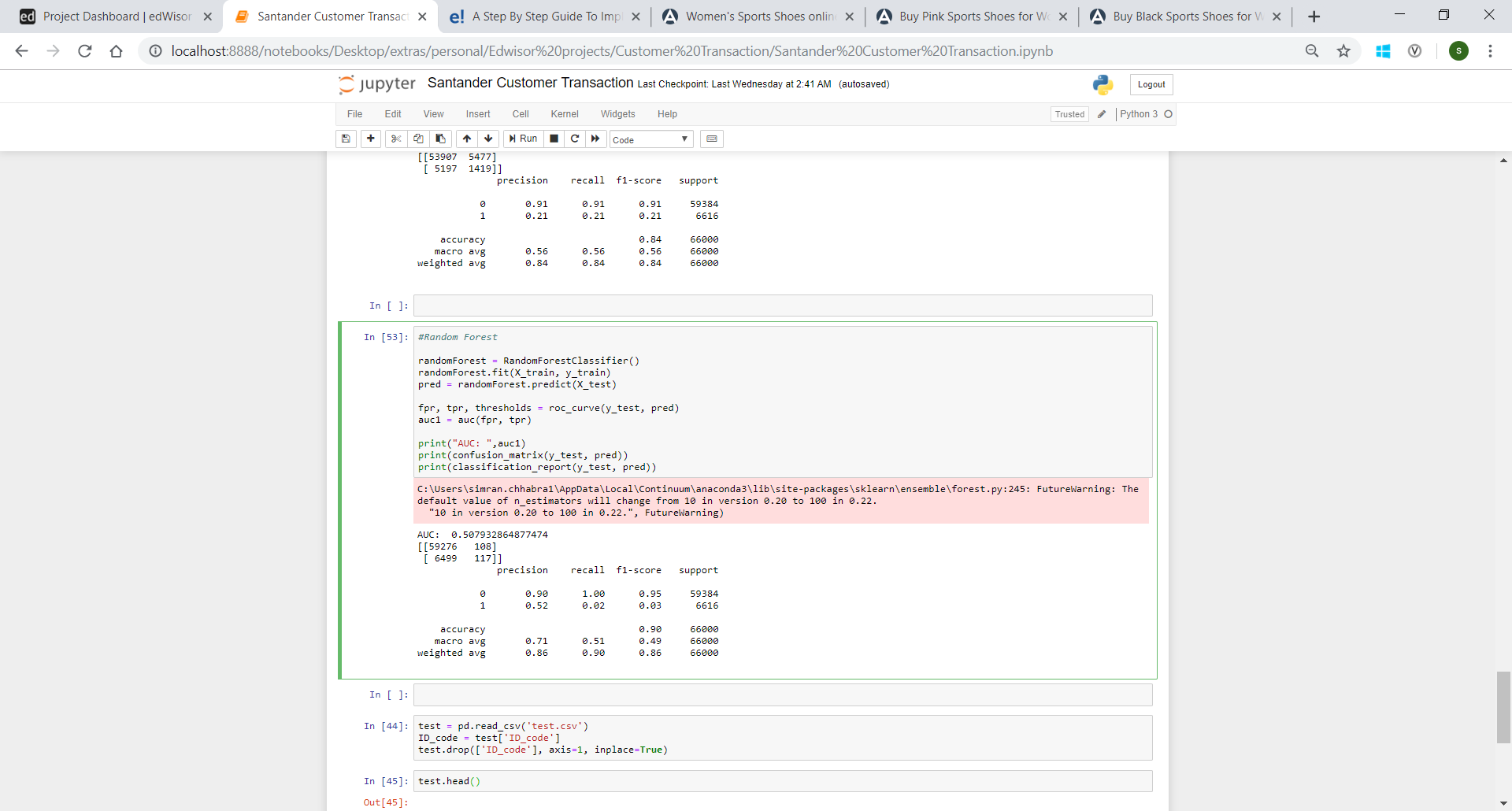
* + 1. **Decision Tree Classifier**

Now we try and use a different classification model to predict our target variable which is Decision Tree classifier. We get following confusion matrix and classification report on performing it:



* + 1. **Random Forest Classifier**

Next, test values are predicted using Random Forest classifier and the following results are observed:



**Chapter 3**

**Conclusion**

* 1. **Model Evaluation**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. We have prepared classification report and confusion matrix for each model. AUC (Area Under Curve) which gives the rate of successful classification by the logistic model is also used as a metric.

Confusion matrix is something like below:



It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

Classification report gives us a summary of precision, accuracy, support etc. from the model.

**3.2 Model Selection**

More the accuracy and precision, better a model. In some cases where wrong classification of the target variable could be dangerous, false positives and negatives from confusion matrix are critical values to check.

In our case, Naïve Bayes gives highest accuracy and AUC score of all three models thus it is selected.

**Appendix A**

Please find attached python notebook “Santander Customer Transaction.ipynb” file.

**Appendix B**

Please find attached “Santander\_R\_code.R” file.

**References**

<https://learning.edwisor.com/>