**Abstract**

Credit card frauds are stress-free and friendly targets. Other online sites as E-commerce have improved the online payment modes, exposing online transactions to fraud. Researchers brainstormed and started using different machine learning methods to detect and analyses frauds in online transactions. The objective is to design and develop a novel fraud detection method for Streaming Transaction Data, to scrutinize the past transaction details of the customers and extract the behavioural patterns. Where cardholders are clustered into different groups based on their transaction amount. Using other strategies like sliding window, to cumulate the transactions made by the cardholders from different groups, so that the behavioural pattern of the groups can be extracted respectively. Classifiers are trained and tested over the groups separately, and then the classifier with better rating score can be chosen to be one of the best methods to predict frauds. Thus, followed by a feedback mechanism to solve the problem of concept drift. In this paper, European credit card fraud dataset is used.

1

**CHAPTER ONE**

**Introduction**

Credit card refers to a card that is assigned to the customer (cardholder), permitting them to procure goods and services within credit limit or withdraw cash in advance. Credit card offers the cardholder an advantage of the time, i.e. It allows customers to repay later in an approved time, carrying it to the next billing cycle. Credit card frauds are stress free targets. With little or no risks, a specific amount can be withdrawn without the consent of the owner, Fraudsters continuously try to make every fraudulent transaction legitimate, this makes fraud detection difficult and challenging task to detect. 2017, there were 1,579 data breaches and nearly 179 million records among which Credit card frauds were the most common form with 133,015 reports, employment or tax-related frauds with 82,051 reports, phone frauds with 55,045 reports followed by bank frauds with 50,517 reports from the statics released. With different frauds, mostly credit card frauds, often in the news for the past few years, frauds are in the top of mind for most the world’s population. Credit card dataset is highly imbalanced because there will be more legitimate transaction when compared with a fraudulent one. Advancement, banks are moving to EMV cards, which are smart cards that store their data on integrated circuits rather than on magnetic stripes, have made some on-card payments safer, but still leaving card-not-present frauds on higher rates. According to 2017, the US Payments Forum report, criminals have shifted their focus on activities related to CNP transactions as the security of chip cards were increased.

2

**CHAPTER TWO**

**Literature Survey**

Supervised and Unsupervised machine learning techniques are used for fraud detection. The objective is to overcome three main challenges with card frauds related dataset i.e., strong class imbalance, the inclusion of labelled and unlabeled samples, and to increase the ability to process a large number of transactions. Different Supervised machine learning algorithms like Decision Trees, Naive Bayes Classification, XGBOOST, Logistic Regression and SVM are used to detect fraudulent transactions in real-time datasets. Two methods under random forests are used to train the behavioural features of normal and abnormal transactions. Even though random forest obtains good results on small set data, there are still some problems in case of imbalanced data. The future work will focus on solving the above-mentioned problem. The algorithm of the random forest itself should be improved. Performance of Logistic Regression, **K-Nearest Neighbour**, and Naïve Bayes are analysed on highly skewed credit card fraud data where Research is carried out on examining meta-classifiers and meta-learning approaches in handling highly imbalanced credit card fraud data. Through supervised learning methods can be used there may fail at certain cases of detecting the fraud cases.

viding transaction details of a customer is considered to issue related to confidentiality, therefore most of the features in the dataset are transformed using principal component analysis (PCA). V1, V2, V3,..., V28 are PCA applied features and rest i.e., ‘time’, ‘amount’ and ‘class’ are non-PCA applied features, as shown in table 2.

Table 2: Attributes of European datasets.

Feature Description.

1. Time: Time in seconds to specify the elapses between the current transaction and first transaction.

2. Amount Transaction amount

3. Class: 0 implies -not fraud, 1 implies fraud

Fig. 3 shows the correlation matrix of the dataset. This matrix explains that attribute class is independent of both the amount and time of the transaction was made. It is even clear from the matrices, the class of the transaction is depending on PCA applied attributes.

3

**CHAPTER THREE**

**METHODOLOGY**

The study attempt to investigate “**Credit Card Fraud**”. The chapter describes the models to be used in the research in order to build a machine.

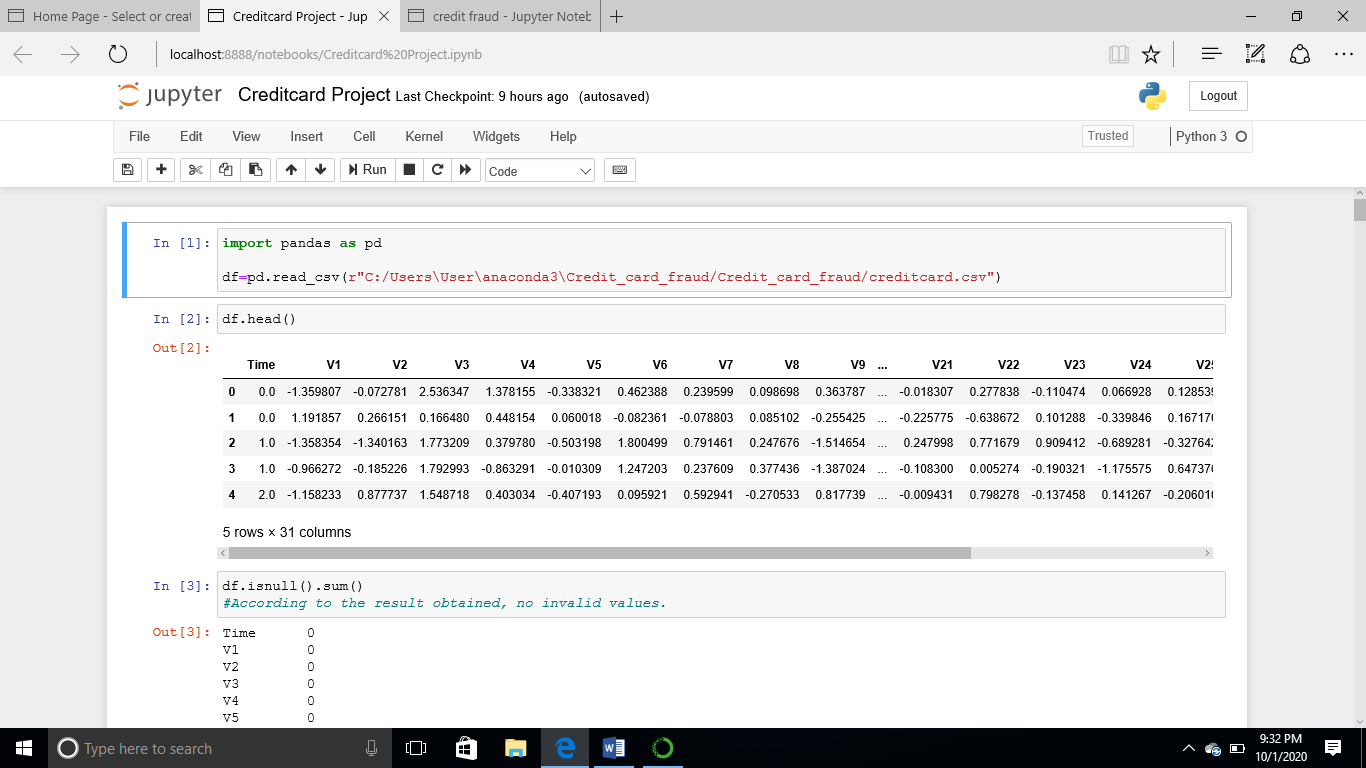
It is a classified machine learning. That predicts the label from the list of predefined classes.

**Challenges of this learning**:

* **Overfitting**: This is when the training performance is higher than the testing set performance.
* **Under fitting**: This occurs when the training performance is lower to that of the testing set result.
* **Generalization**: It occurs when the training and testing performance are equal.

**Machine Models and tools use in training**:

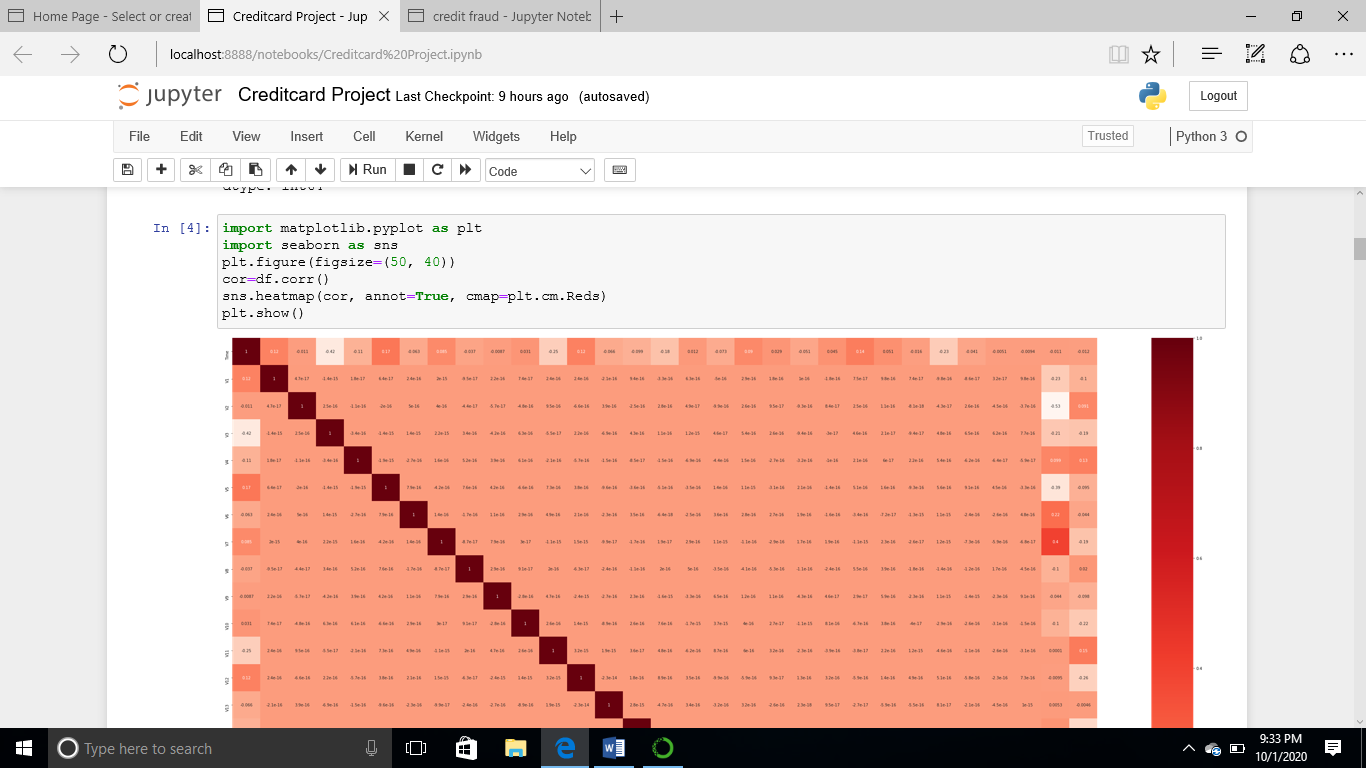
Pandas: A tool use in loading the dataset.

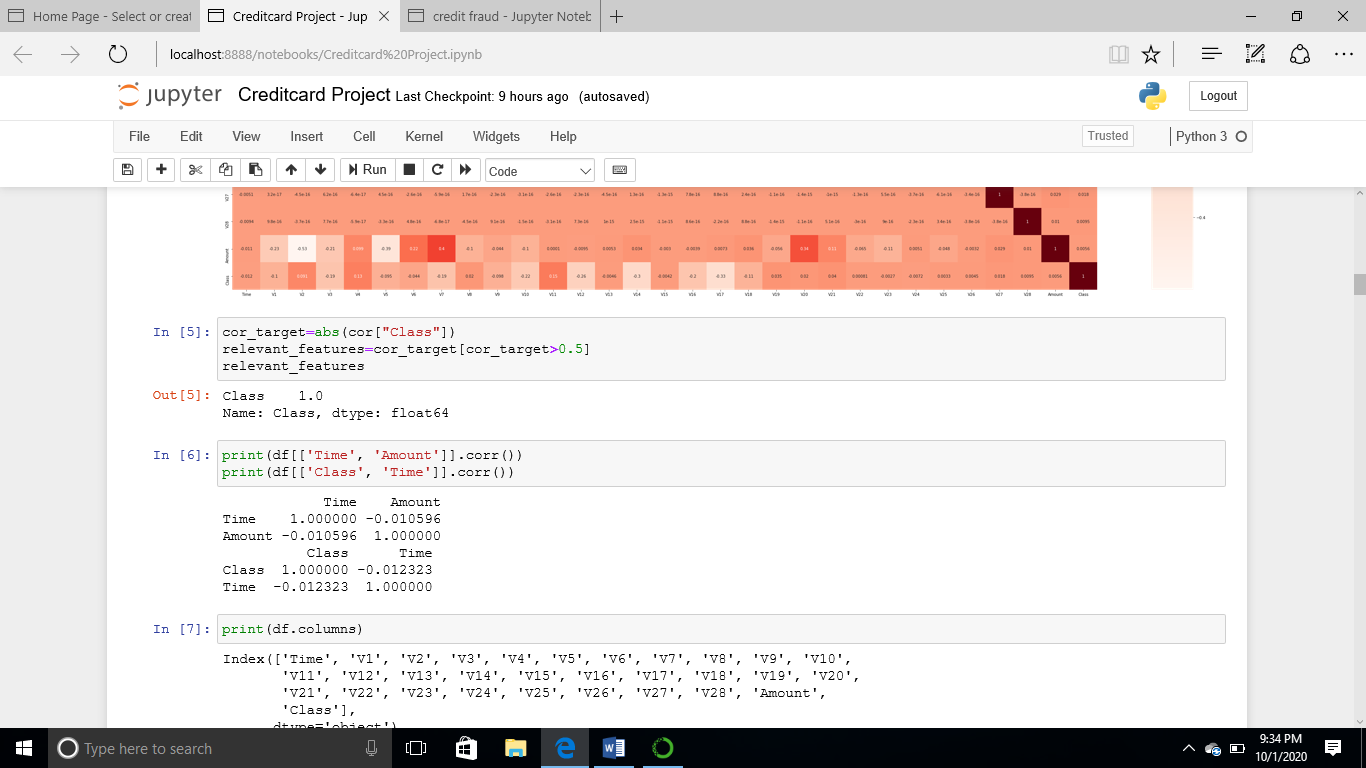


Use of (isnull) method to check for invalid or null values

Arithmetic operation such as sum, average, correlation, on the data. In order into ensure no null values and to check for the target variables and features.

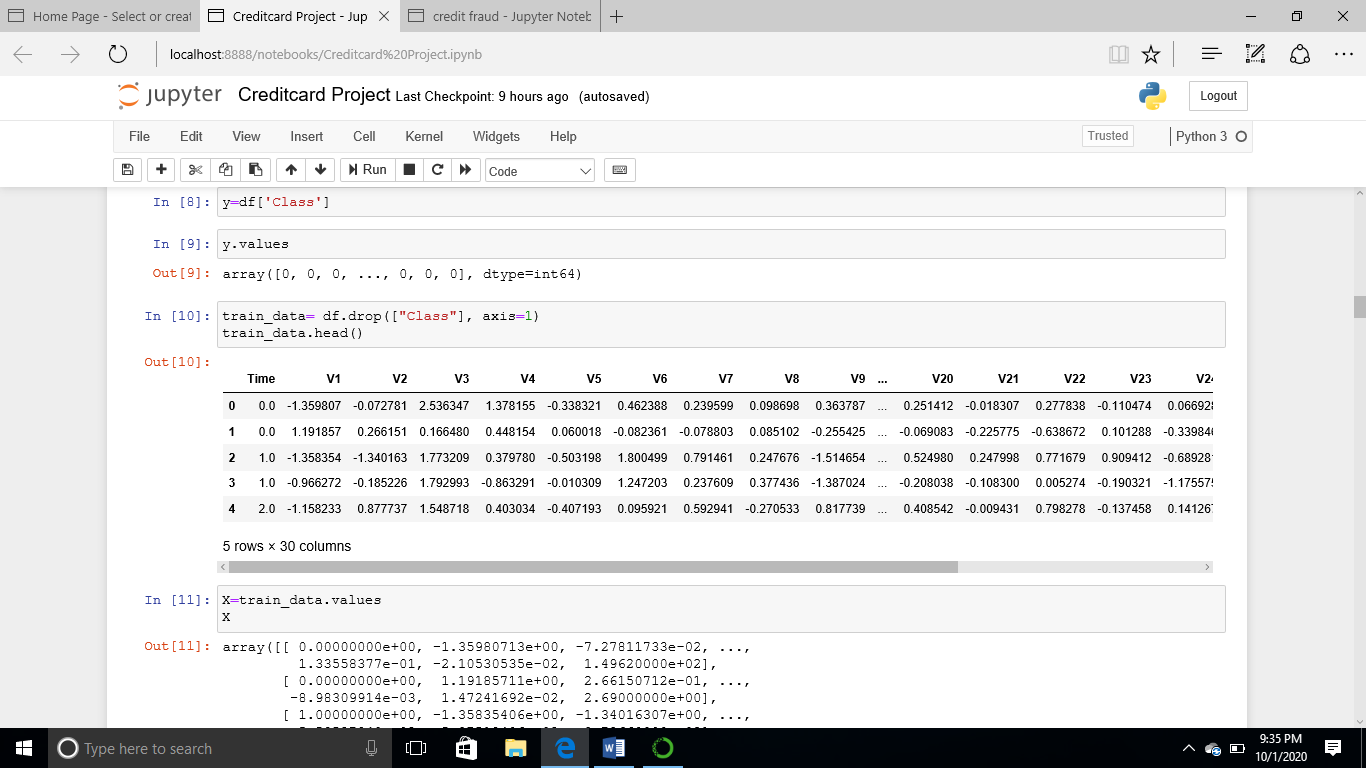
4



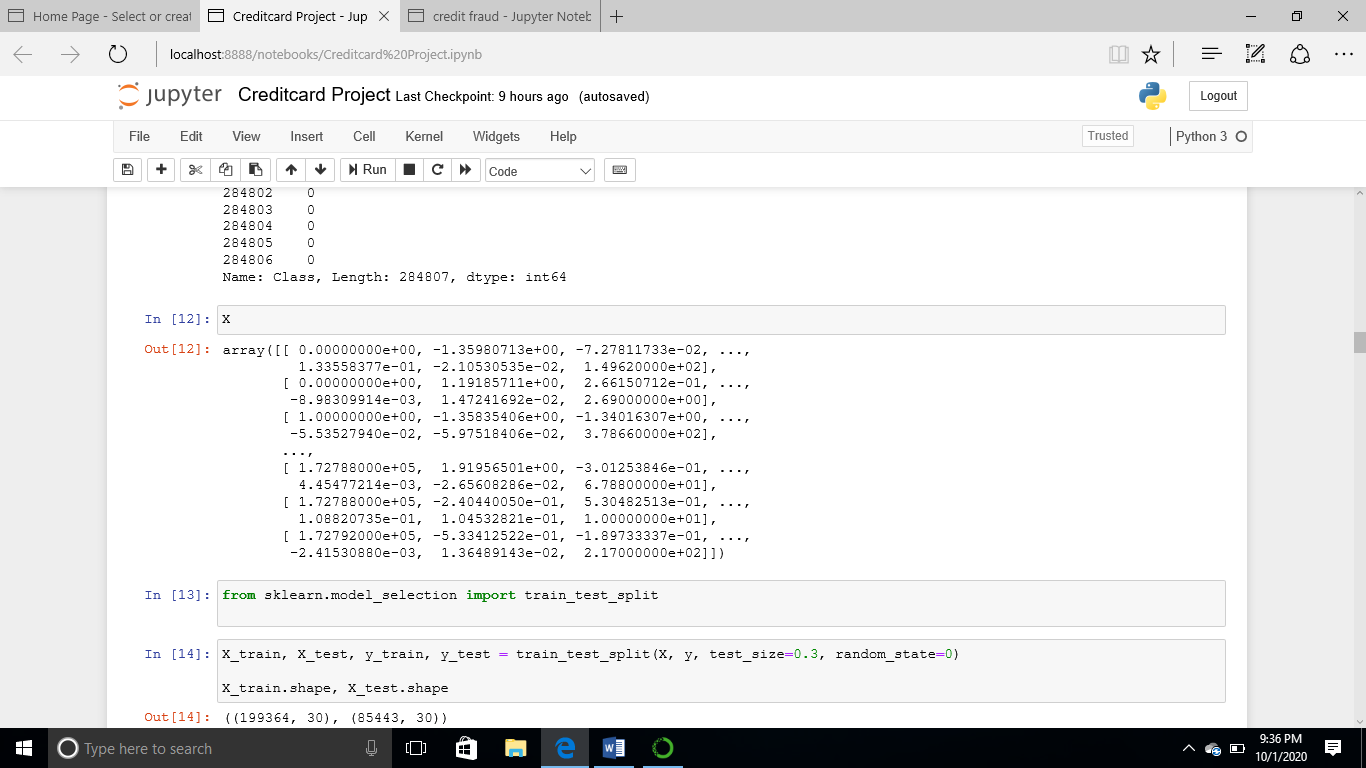


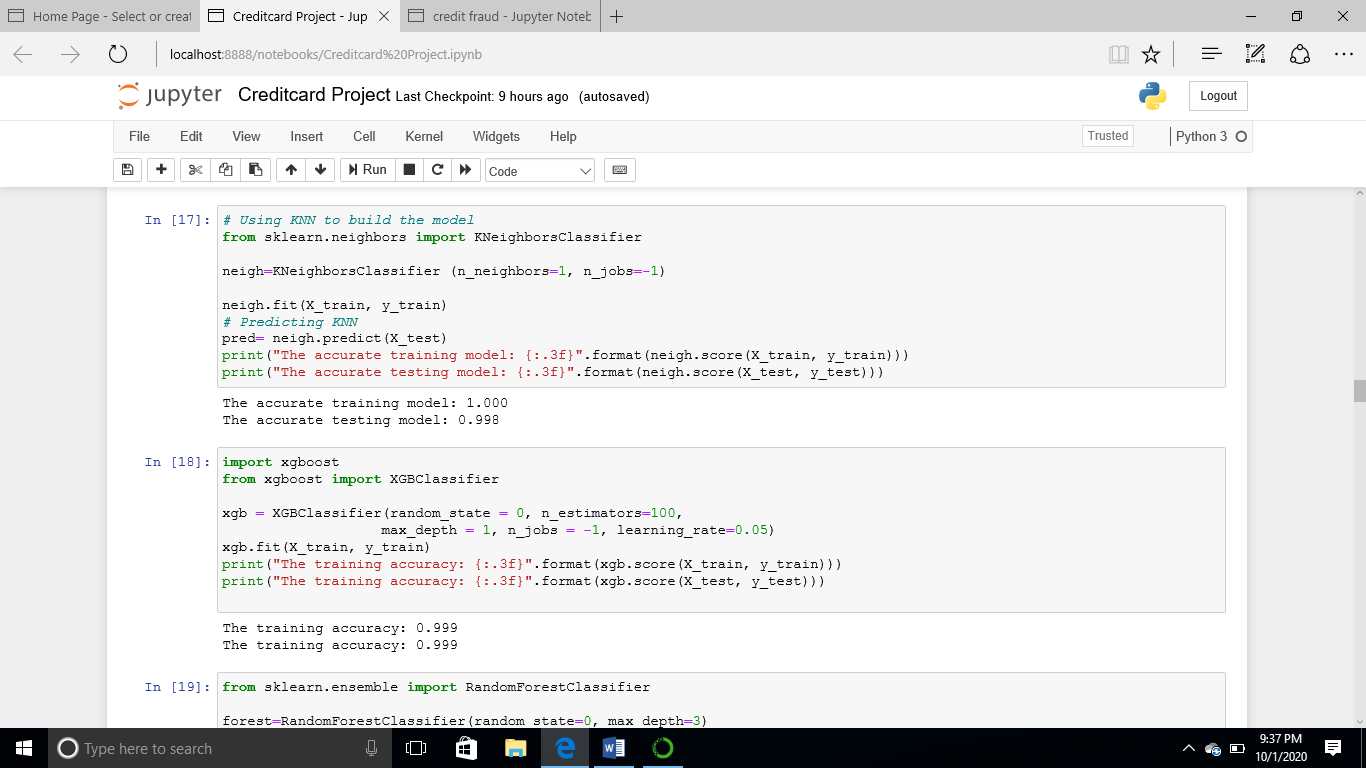
Converting the target variables to a numpy (y) and features as (x).

5

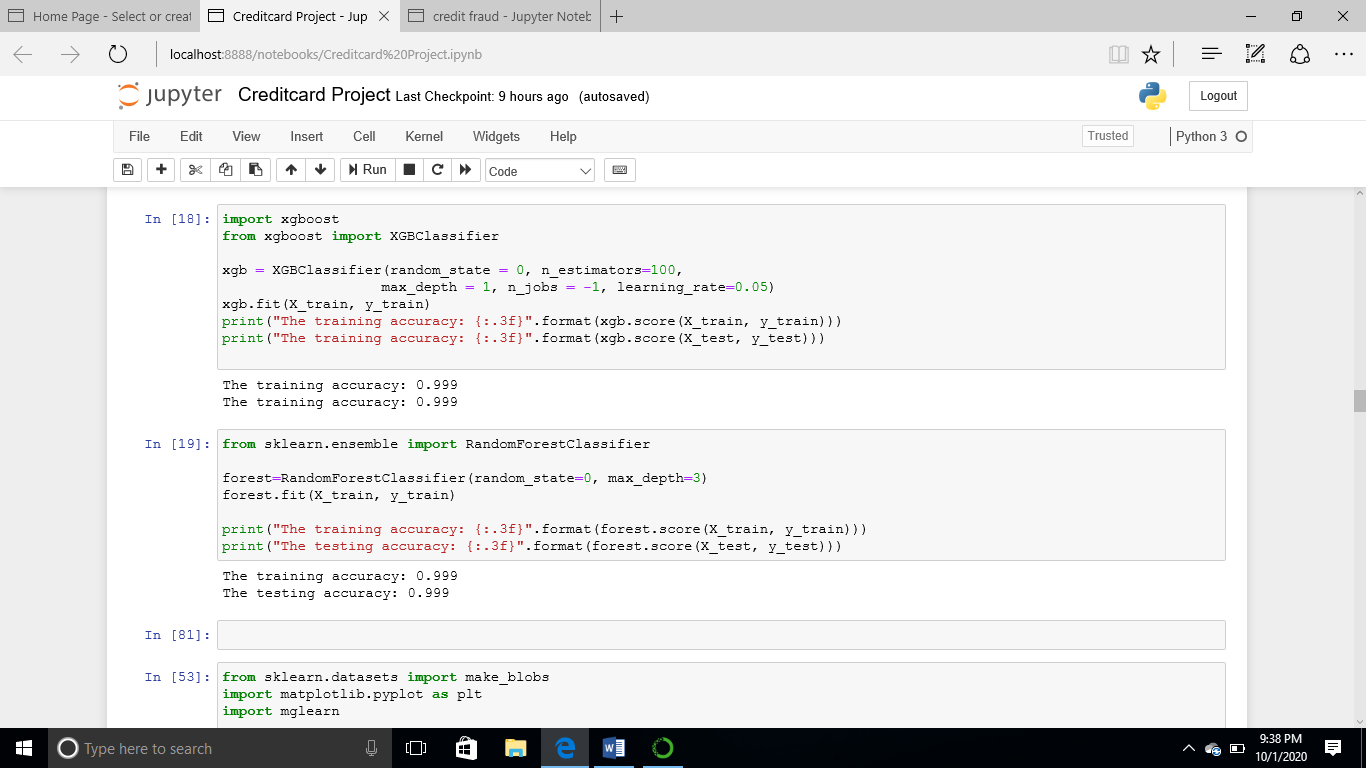


The use of model selection to split the data into x and y.



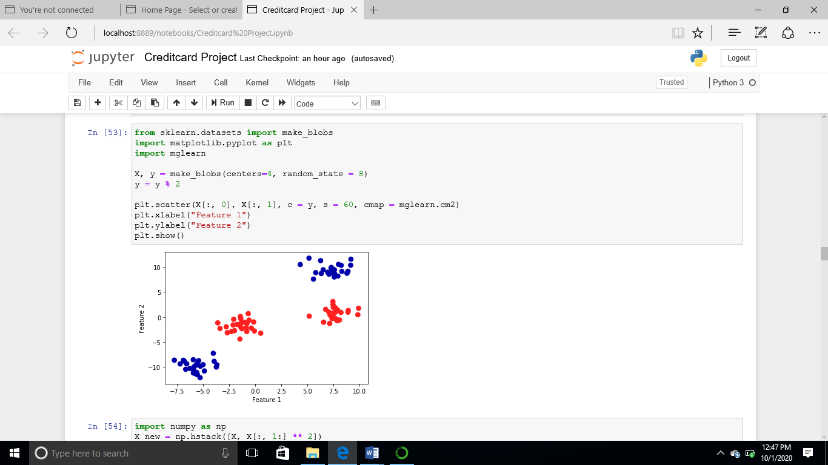
The use of **(sklearn.** **Neighbours importing KNeighbours Classifier)** to train and test the accuracy of the split data.

6

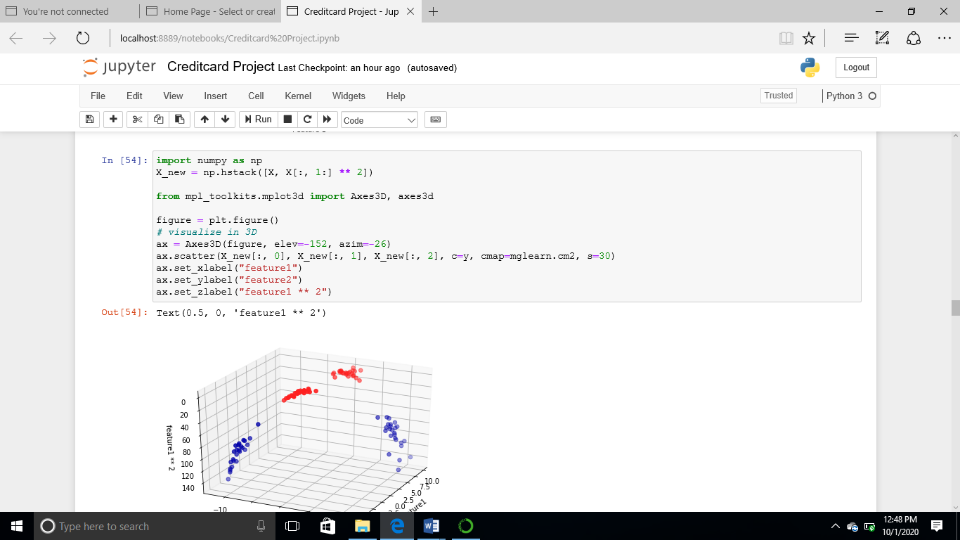
Explore **(Xgboost classifier)** in order to tune and correct the overfitting encountered while using **(KNN)**. And **Random Forest** which was used to check for better performance.

7

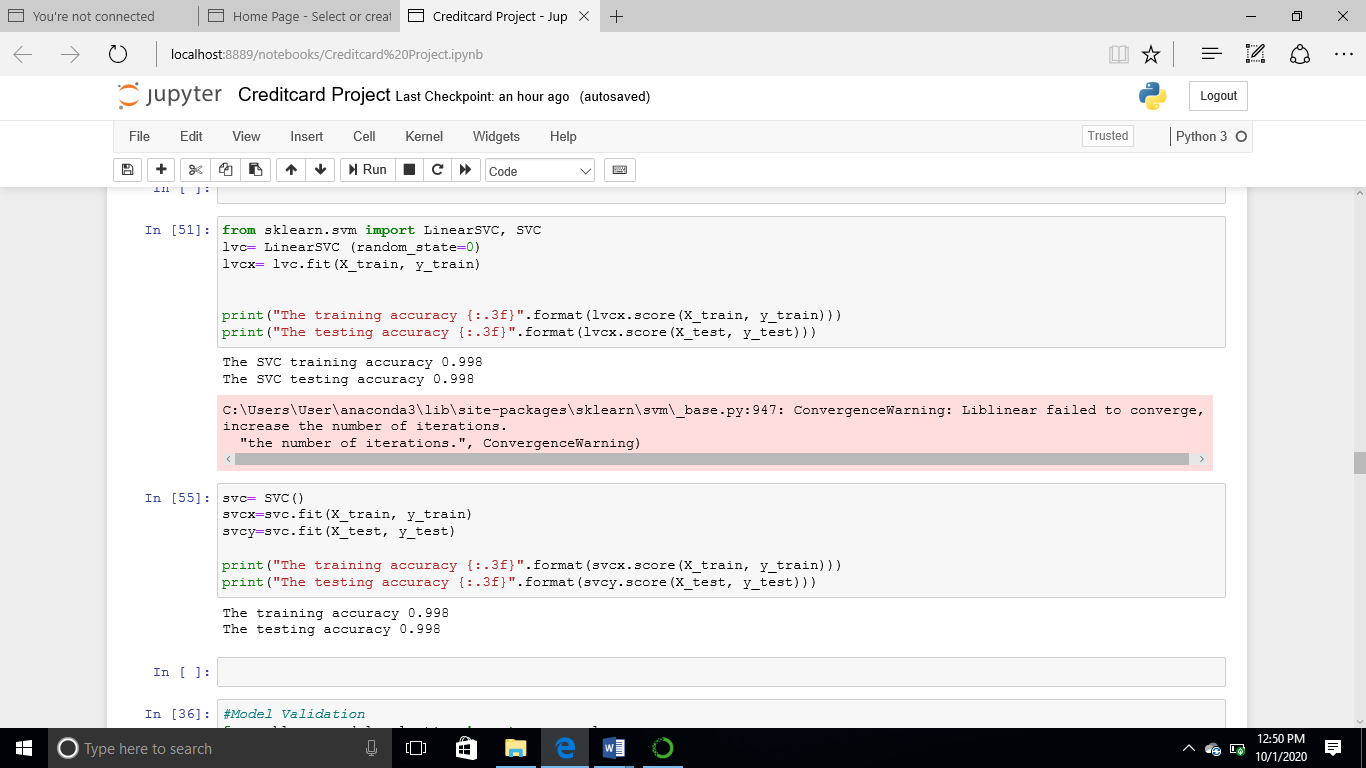
**Make blobs for visualization.**



**Numpy for visualizing the features.**

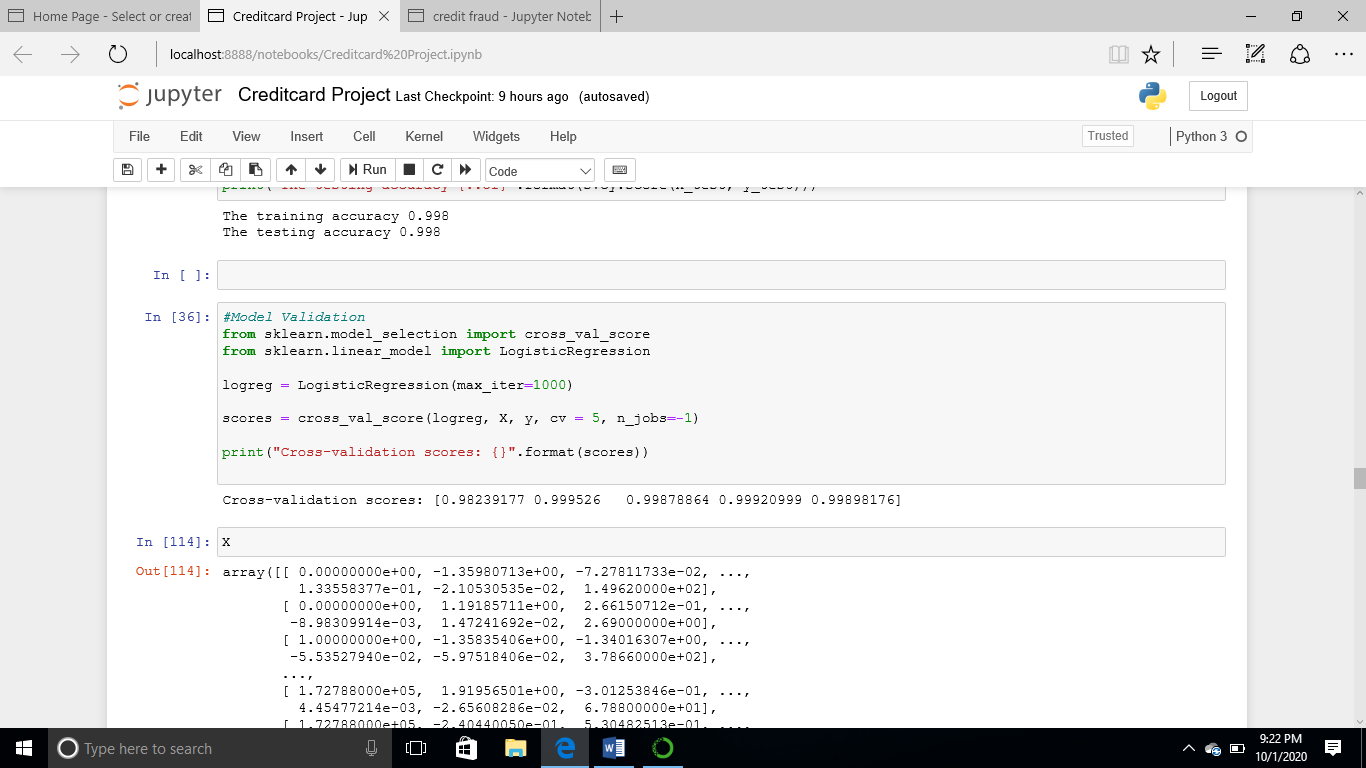


**Linear SVM/SVC models**

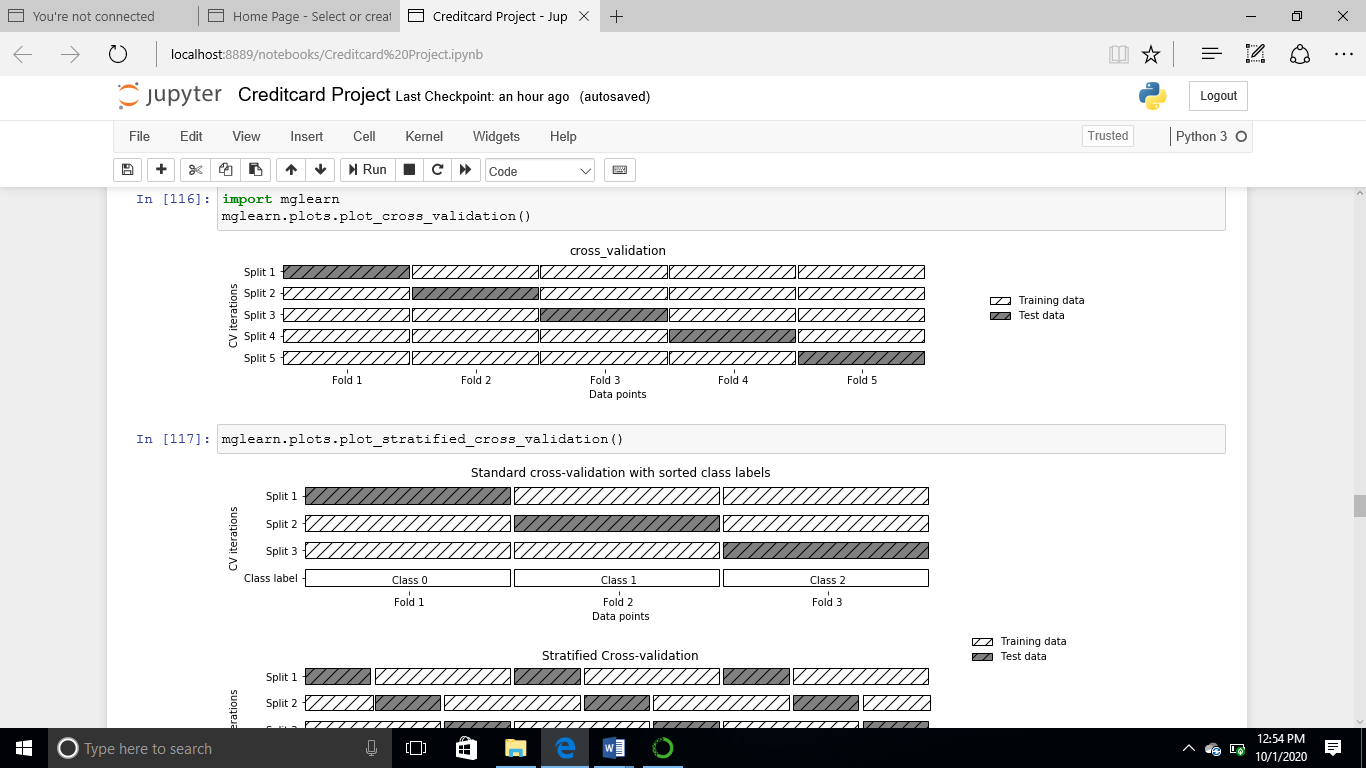


8

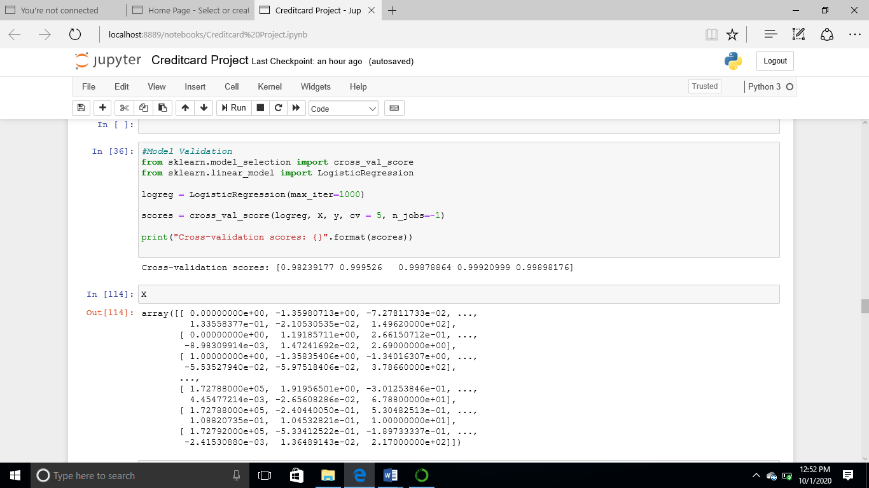
**Model Validation**



**Model Validation, Cross Validation Score and X values.**

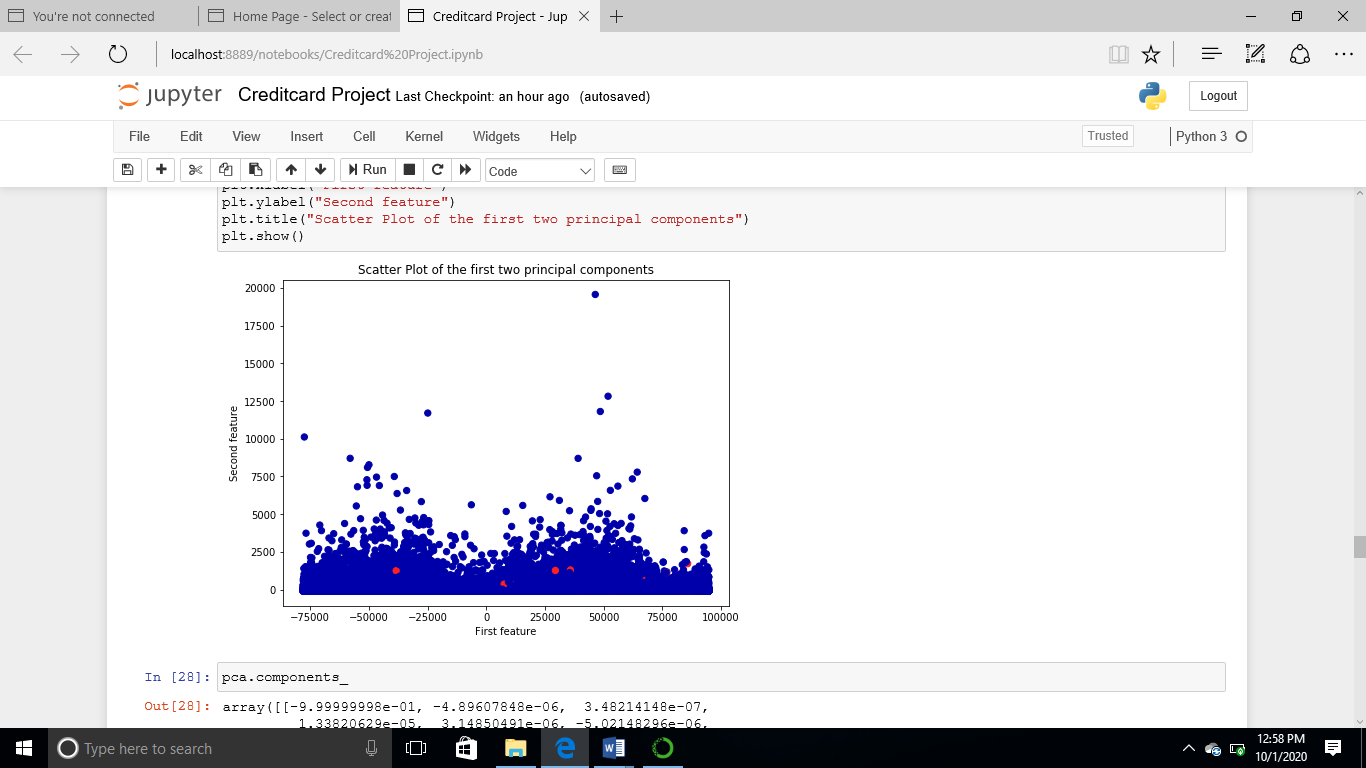


**Principal Component Authority**

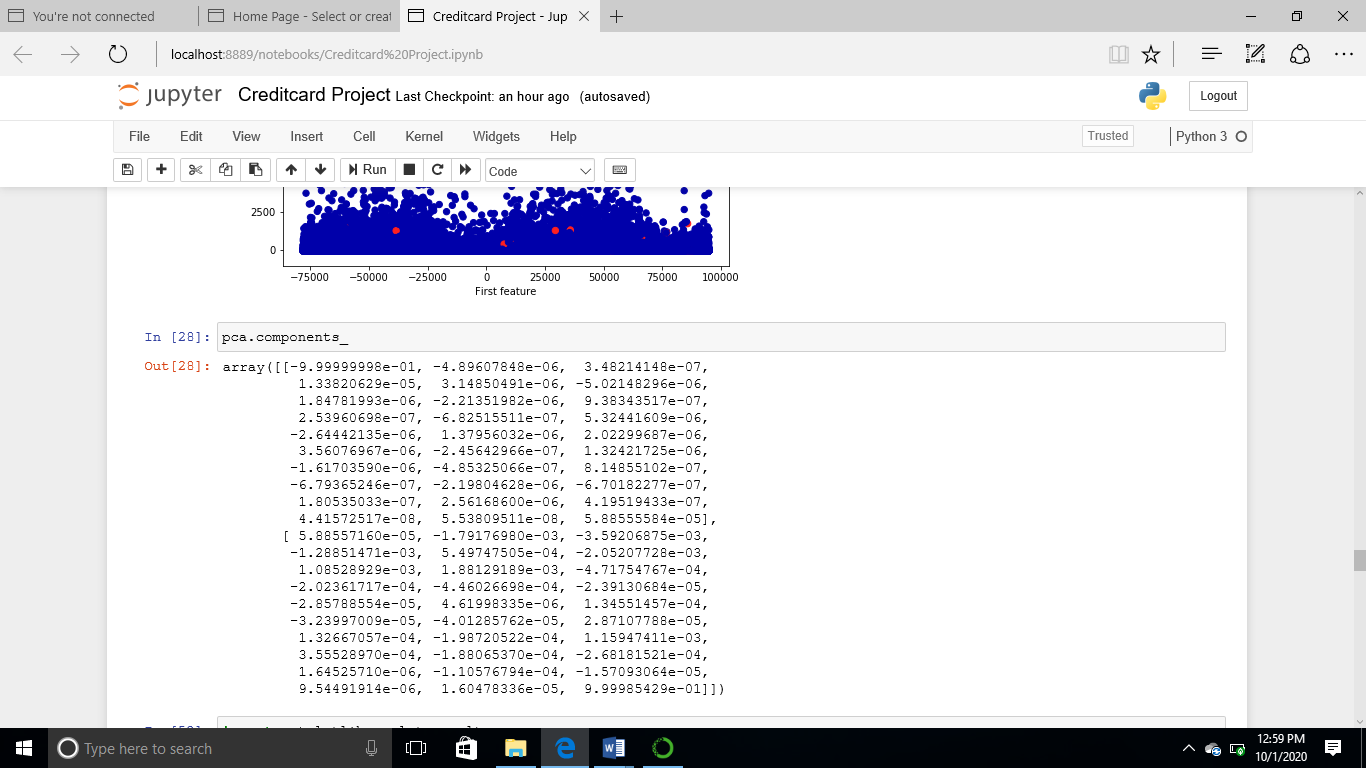


9

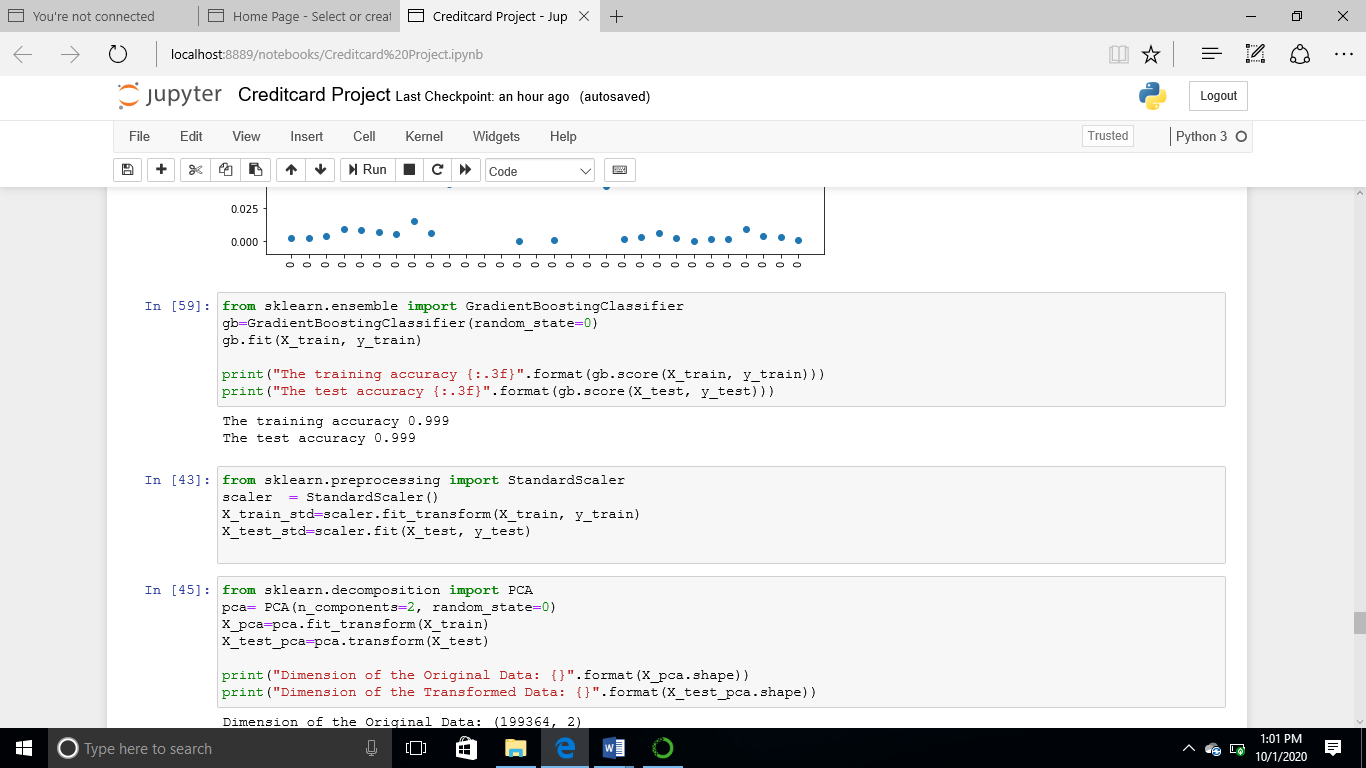
**Visualizing Principal Component Authority**



**PCA’s Components**

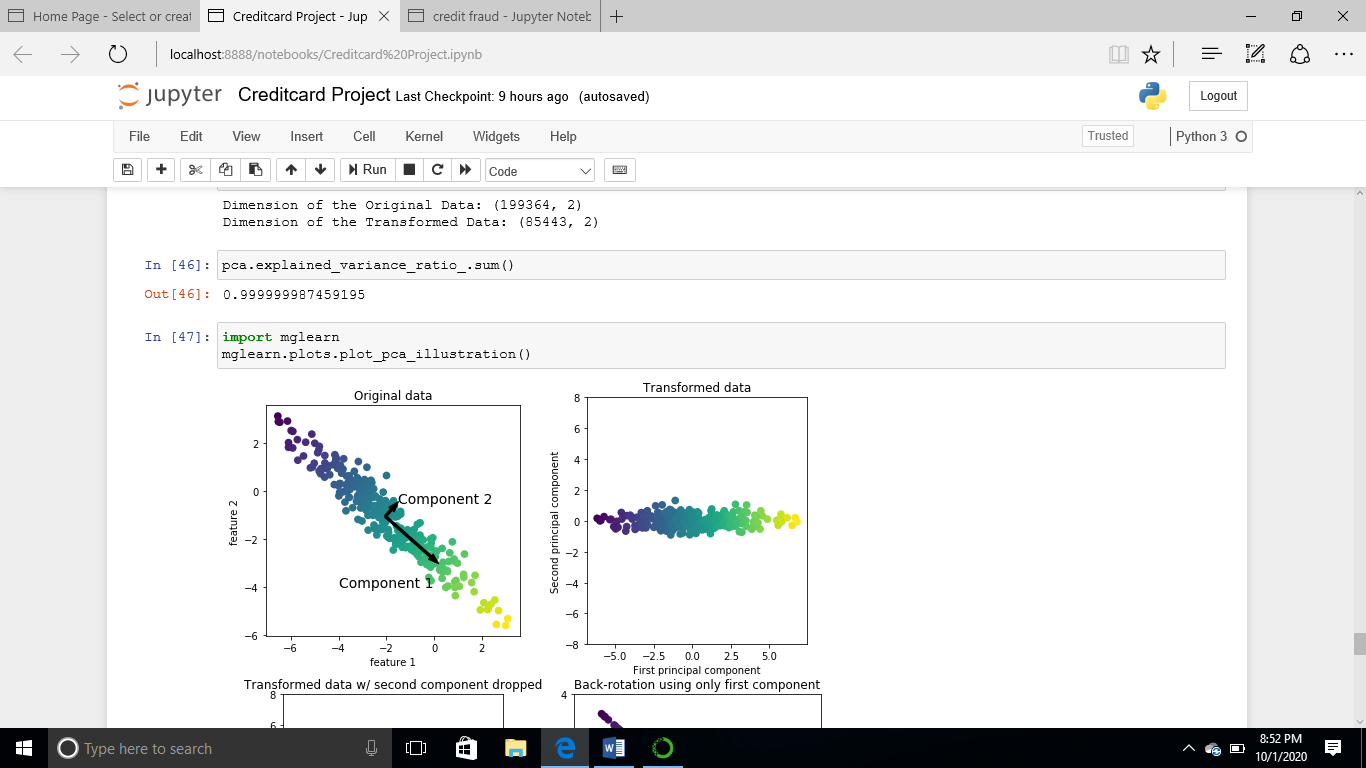
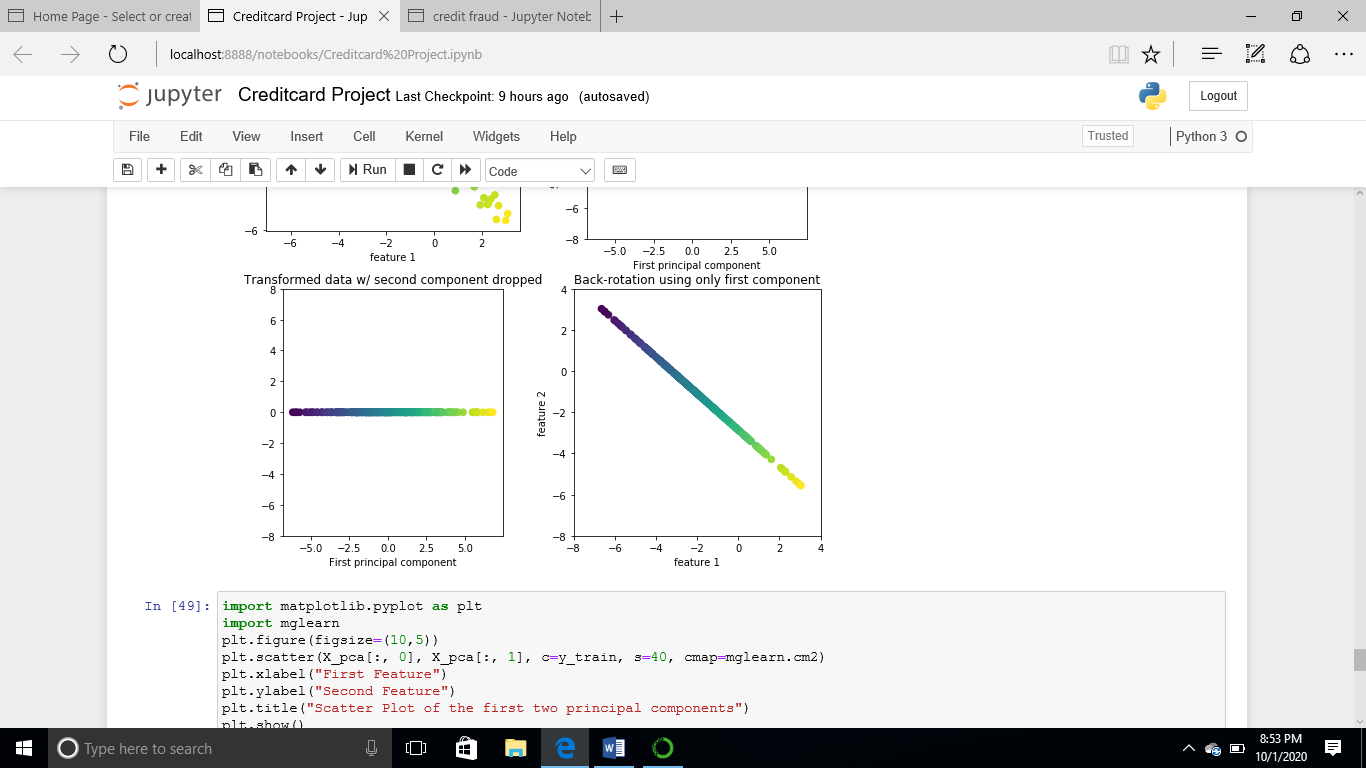


**Gradient Boosting Classifier and Standard Scaler**

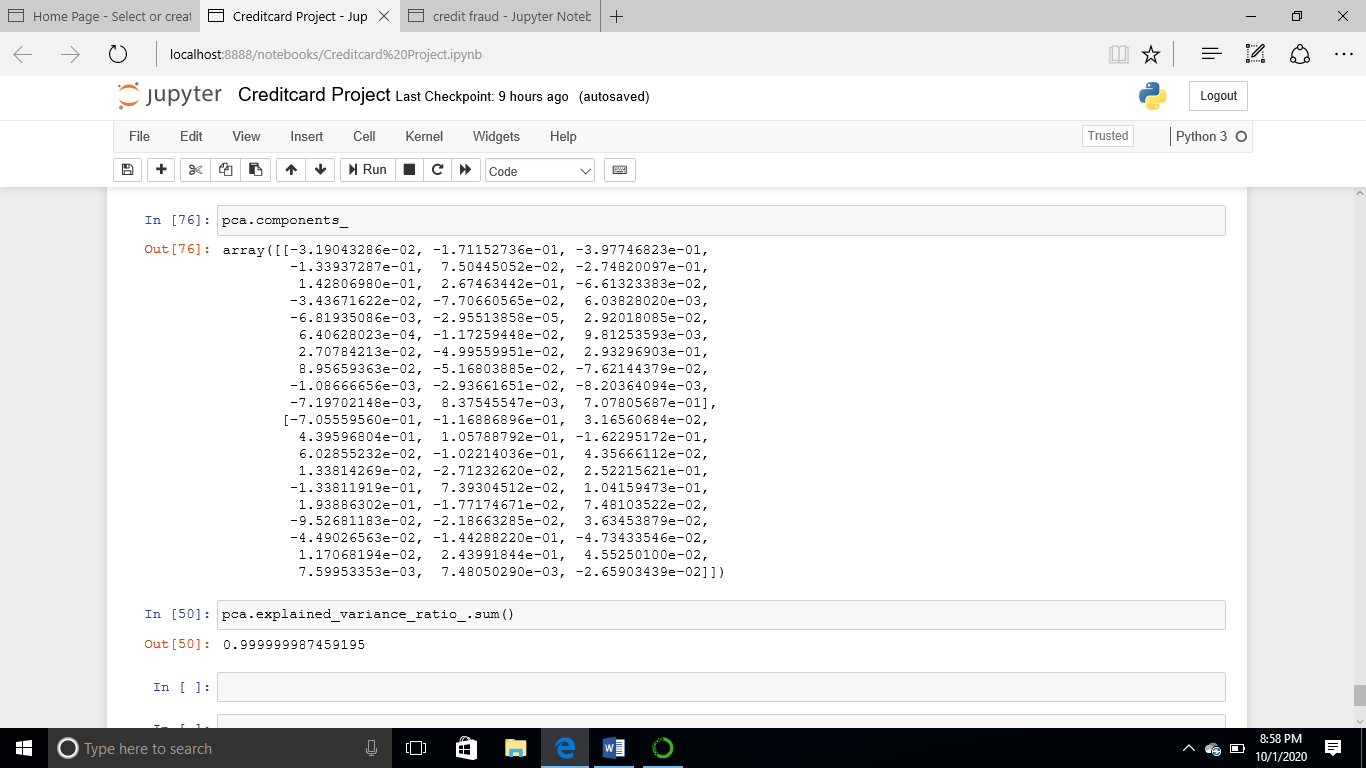
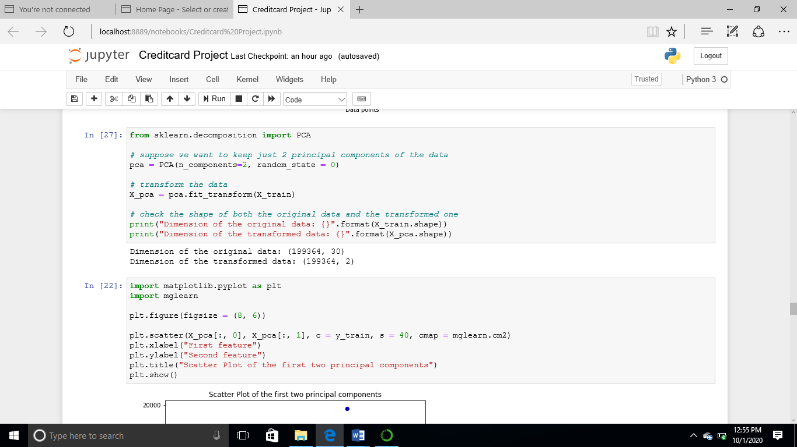


10

**PCA/Total value of Explained Variance and Ratio and Its Visualization**



**PCA. Components and Explained Variance Ratio**



11

**CHAPTER FOUR**

**Recommendation and Conclusions**

Of all the models used, XGBOOST, Standard Scaler, Random Forest and Principal Component Authority was able to generalize the learning. The training and testing set percentage is 99% while the PCA explained variance ratio is also 99%.

**In conclusion**

In order to get accurate result, it is recommended to use different models, pruning, standard scaler and tuning.

12