

**C3879C Capstone Project**

Credit Card Fraud Detection System

Date of Submission: 31-JUL-2019

Submitted By:

|  |  |
| --- | --- |
| 18061866 | Kandasamy Subbaian Karthik |

# ACKNOWLEDGEMENTS

*I indeed thank Mr.Andy Lee, Mr.David Leong, Mr.Frank Ng, Mr.Lance Peng, Mr.Jimmy Goh, Mr.Tan Poh Keam and Mr.Wee Kah Huat, my family members and Republic Polytechnic, Singapore for their immense help and support throughout the course.*

TABLE OF CONTENTS

[ACKNOWLEDGEMENTS 2](#_Toc451850278)

[ABSTRACT 4](#_Toc451850279)

[1 Introduction 5](#_Toc451850280)

[2 Project Specification and Plan 6](#_Toc451850281)

[2.1 Project Overview 6](#_Toc451850282)

[2.2 Functional Requirements 6](#_Toc451850283)

[2.3 Project Plan 6](#_Toc451850284)

[3 Business Analysis 7](#_Toc451850285)

[3.1 Business Issues 7](#_Toc451850286)

[3.2 Market Analysis 7](#_Toc451850287)

[3.3 Business Solutions 7](#_Toc451850288)

[4 System Design and Implementation 8](#_Toc451850289)

[4.1 System Architecture 8](#_Toc451850290)

[4.2 Detailed System Design 8](#_Toc451850291)

[5 System Testing 19](#_Toc451850292)

[6 User and Technical Documentations 20](#_Toc451850293)

[6.1 User Documentation/Guide/Manual 20](#_Toc451850294)

[6.2 Technical Documentation (Installation guide/Manual) 22](#_Toc451850295)

[7 Conclusions 23](#_Toc451850296)

[References 24](#_Toc451850297)

[Appendices 25](#_Toc451850298)

[Project Poster 26](#_Toc451850299)

# ABSTRACT

*The project is about automation of “classifying the credit card transaction” such as Normal Transaction (or) Fraud Transaction. The Normal Transaction is considered as Negative (0) Transaction. The “Fraud Transaction” is considered as Positive (1) Transaction. It is used interchangeably throughout this report for easy understanding.*

*It describes the Machine Learning (ML) data-pipeline or workflow phases, which are being used in this capstone project for understanding the concepts in a more practical manner. It is implemented using classical machine learning classifer/ algorithms such as Logistic Regression, Naïve Bayes and Support Vector Machine (SVM).*

*The implementation is done with Python3, Numpy, Pandas and Scikit Learn framework. It runs on Windows10 operating System. The system helps the user to classify if the unseen data is Normal (or) Fraud Transaction using ROC-AUC scoring, Confusion Matrix, Recall, F1-score and Classification Report with the help of Machine Learning (ML) models.*

*The project is a kind of standalone project, that is, it is not tied to any source systems such as RDBMS / Distributed Database (Hadoop) and not interfaced with any other systems for inbound/outbound dataflow. In other words, it takes data in the form of CSV file from the local computer and outputs the result in a classical way.*

# Introduction

Utilisation of plastic money/credit card is invariably increasing due to ease

of use and safety. In parallel, the Financial Institutions are demanded, due to business need, to implement safety measurements to meet governance, compliances and increase customer satisfaction & experience.

As the credit card usage becomes prevailing among all kinds of people, the Financial Institutions needs to improve their legacy or rule-based safety measurements/systems, in order to safeguard their customer’s credit card usage. With the legacy systems, the Financial Institutions are not able to handle huge volume of transactions with intelligence and/or precisely. This is where the Machine Learning (ML) technology comes into picture and plays a vital role due to AI evolvement.

The scope of the system is to get dataset and train the model with the dataset and thereafter the model will generalise the unseen data. The system uses under-sampling technique to convert the imbalanced dataset into balanced dataset. It will help the model to generalise unseen data precisely. It helps the organisation to use the system by which we can identify if the transaction is Normal or Fraud Transaction.

The dataset is taken from [www.kaggle.com](http://www.kaggle.com) for training the model (hypothesis). The dataset has 28 anonymised variables, which are labelled from V1 thru V28. These 28 components in the dataset, which are in the form of PCA complaint. The remaining 3 non-anonymised variables such as Time, Amount and Class (Target). Both Time and Amount are not PCA complaint.

The solution is suitable for binary classification. If there is any change in the dataset features, then the model needs to be retrained accordingly, prior to predict. Otherwise, it will lead to erroneous result.

# Project Specification and Plan

## Project Overview

*The “Credit Card Fraud Detection” system is meant to generalise unseen data with the help of Classical (Supervised Learning) Algorithms. It is developed with Python Packages such as numpy, pandas and Scikit Learn frameworks.*

*Financial Institutions are very keen to leverage Artificial Intelligence to improve their operational efficiency in a cost-effective manner. It has led to develop this anomaly system using machine learning technology. It involves analysing the dataset, identifying accuracy methods such as ROC-AUC and tuning hyperparameters of the model for model’s better performance.*

## Functional Requirements

*Below are the system’s functional requirements:*

* *Analyse the given dataset and normalise it.*
* *Train a couple of relevant classification algorithms with the training dataset*
* *Identify the best model based on their performance with cross validation score.*
* *Tune hyperparameters of the selected model*
* *Predict unseen data with the learned model.*

## Project Plan

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Activities** | **Apr-19** | **May-19** | **Jun-19** | **Jul-19** |
| **Requirements Collection / Problem Statement** |  |  |  |  |
| **Data Collection** |  |  |  |  |
| **Analysis** |  |  |  |  |
| **Design** |  |  |  |  |
| **Development & SIT** |  |  |  |  |

# Business Analysis

## Business Issues

*Businesses have been trying to resolve their issues with the current rule-based enterprise systems but the business demands more advanced niche technologies to resolve their upcoming issues in a more vibrant way. As compared to the current legacy and/or rule-based systems, AI provides them more robustness to solve their business issues more precisely.*

*For example, with current systems, the business won’t be able to predict precisely. Besides, it does not have capability to learn on their own whereas, in AI, the system has capability to learn on their own and does not require coding/amendment to adopt the new changes.*

## Market Analysis

*The solution suits to wide-range of financial institutions, who have credit card business. Since the solution is very generic, Financial Institutions can customise the solution according to their business need.*

## Business Solutions

*Machine Learning is a subset of AI. It has 3 types of algorithms such as supervised, unsupervised and reinforcement learning algorithms. The Supervised Algorithms address regression and classification/anomaly findings problems. The unsupervised algorithms help to cluster unlabelled components. Reinforcement Learning helps the industry to address action-reward-punishment environment-based issues. It helps the model to learn and correct the error on its own.*

*The Credit Card Fraud Detection system uses supervised algorithms such as Logistic Regression, Naïve Bayes and Support Vector Machine (SVM) to identify the fraud transactions. The Anomaly detection is a kind of classification model, where the sensitivity is key to find the anomaly. In general, if the model threshold is greater than 0.5, then it will classify it as +positive case but here the sensitivity/recal/true positive rate is very important to classify +ve/-ve transactions.*

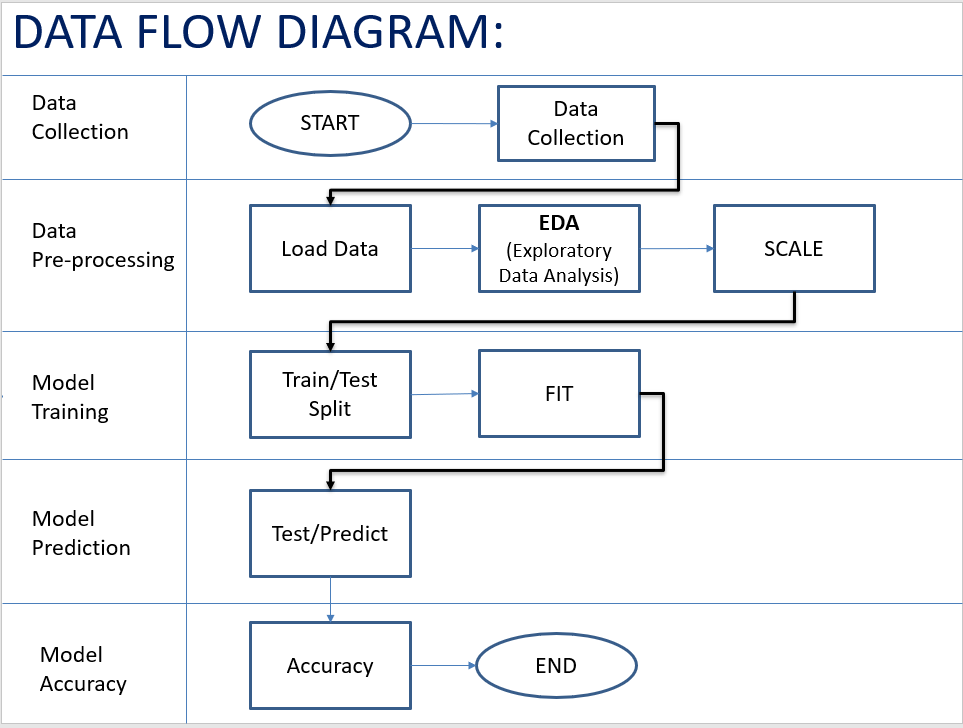
# System Design and Implementation

## System Architecture

*It is a simple standalone system, that is, it does not have any inbound/outbound interface to receive/send data from/to any other systems. In other words, it receives the data in the form CSV format from the local computer/server and process it further.*

## Detailed System Design

Below is the system’s data flow diagram.

**

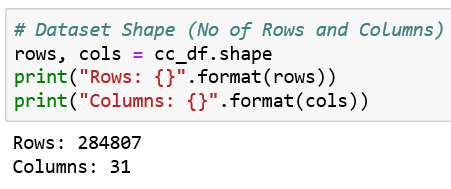
***Pseudo Code:***

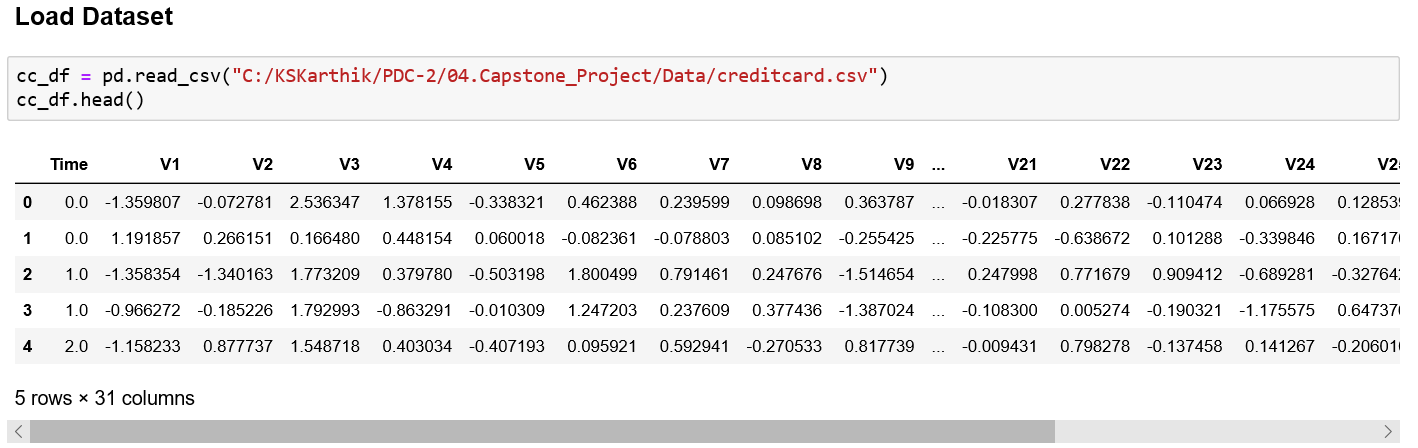
1. Load Dataset into the system.
2. Class Count function to check if the dataset is balanced or imbalanced
3. Scale Non-Anonymised features such as Time and Amount
4. Concatenate both Scaled Amount and Time with Actual Dataset/frame.
5. Drop old Amount and Time features
6. Random under-sampling function
   1. Split the scaled dataset into Train and Test
   2. Reset index of scaled train and test dataset
   3. Find no of fraud transactions in the (random) Training dataset
   4. Segregate normal and fraud transactions from the training data
   5. Randomly selecting the same no of normal transactions as fraud transactions
   6. Reset Index of both selected normal and fraud transactions
   7. Concatenate both selected normal and fraud transactions.
   8. Shuffle the subsample/final dataframe.
7. *Remove extreme outliners and create final dataset*
8. *Split final dataset into Train and Test*
9. *Spot-check a couple of Classification Algorithms (using cross validation)*
10. *Make Predictions on Test data*
11. *Tune the (best) model’s hyperparameters.*
12. *End.*

**Implementation:**

**Dataset:**

* It is taken from [www.kaggle.com](http://www.kaggle.com)
* It has 30 features and 1 target. Of 30 features, 28 features are labelled V1 to V28. The remaining 2 features are Time and Amount.
* The 28 features are in the form of PCA complaint.
* Both Amount and Time are not normalised, that is, not in line with other variables in terms of scale.

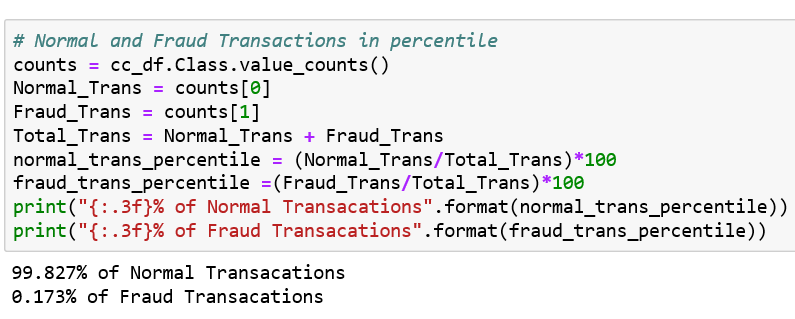
****

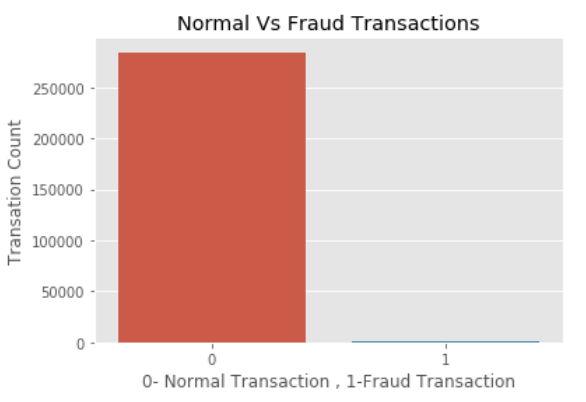
****

**Exploratory Data Analysis (EDA):**

**Imbalanced Dataset:**

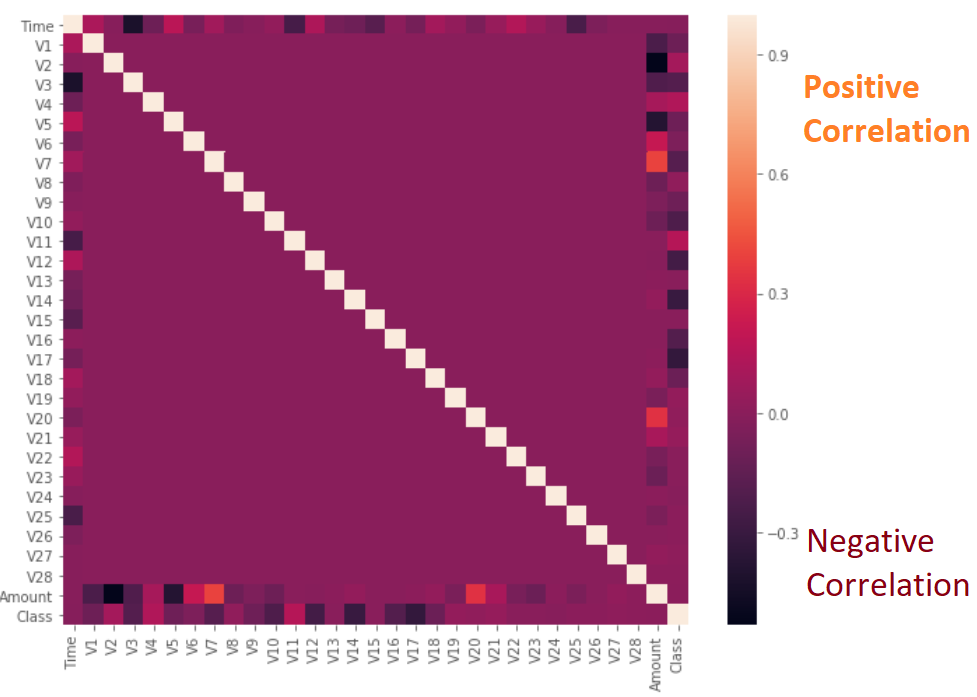
Based on Class counts, we realised that no of normal (negative) transactions are not equal to the no of fraud (positive) transactions. If we train the model with this imbalanced dataset, then it will lead the model to classify most (99%) of the transactions as normal.





**Feature Correlation (w.r.t Class):**

The correlation between all features w.r.t the Class variable is best illustrated with below heatmap. We do not find any high correlated variables with respect to the Class variable. If there is any, then we can remove since it will influence the models performance. In such case, we can remove highly correlated variables.



**Note**: If you refer to the diagonal, the correlation is 100% , that is, the correlation (that is., between the same variables) to itself is 100%. It is indeed very natural.

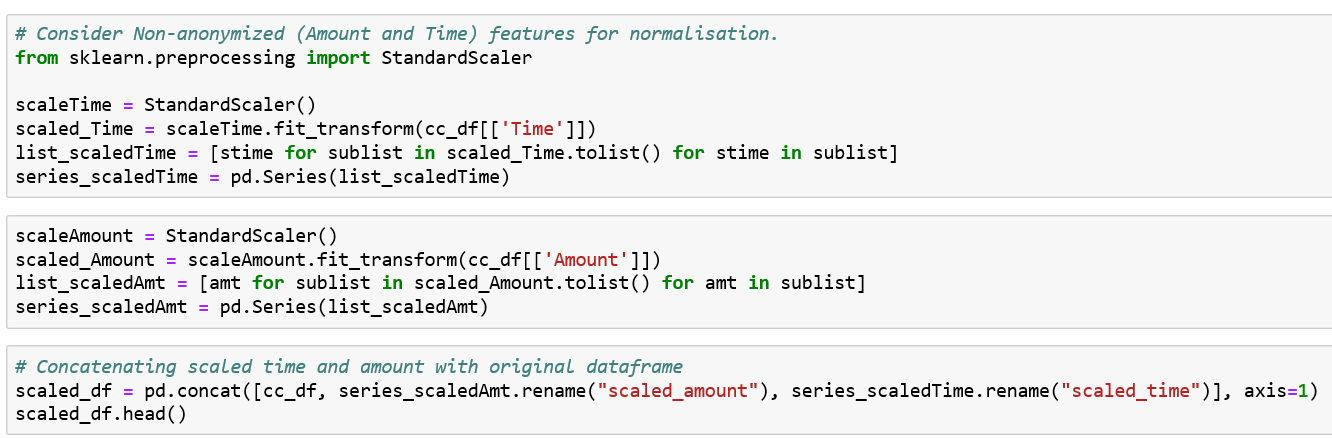
**Imputation of Missing Values:**

The dataset does not have any null values. In other words, all predictors are there in the PCA format with value of type int/float. In this case, it is easy for data processing. Besides, there is no need for data imputation of missing values.

|  |  |
| --- | --- |
|  |  |

**Scale Non-Anonymised Variables:**

Both Amount and Time are not in line with other variables in terms of scale. Hence, it needs to be scaled and added to the original dataframe for further processing.



**Create Dataset with Balanced Class Distribution:**

As the actual dataset is imbalanced, we can use below techniques to create balanced dataset.

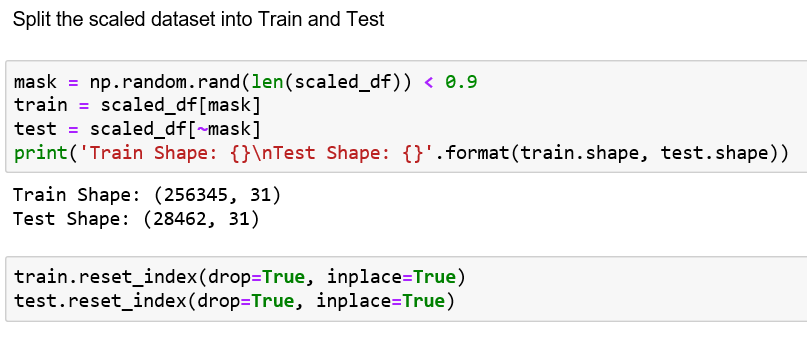
* Oversampling (minority class)
* Under-sampling (majority class)
* SMOTE

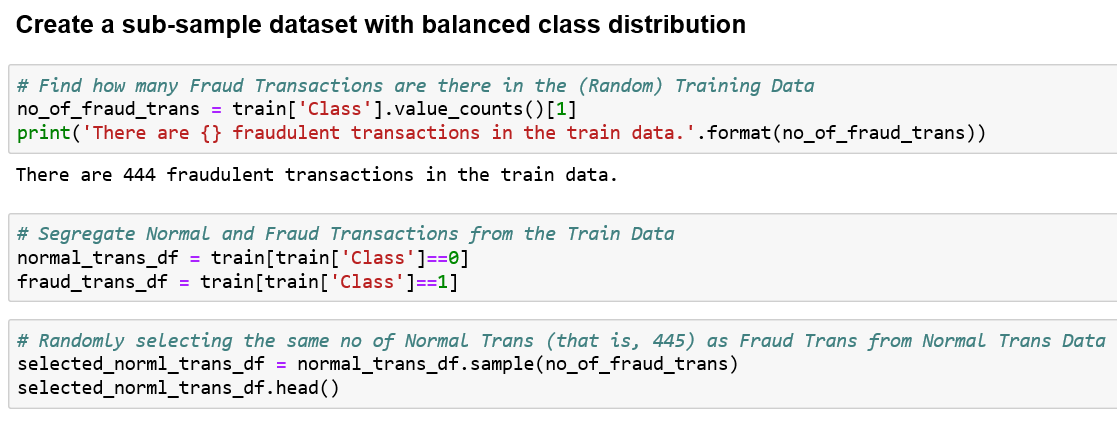
Each Technique has its own pros and cons. In this assignment, I opt for

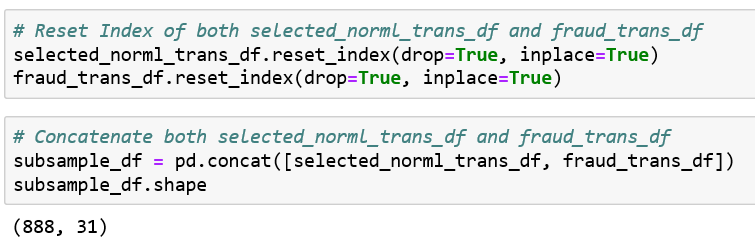
under-sampling technique, by which I create the balanced dataset for further data analysis.

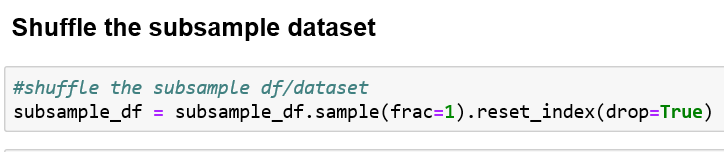
The purpose of creating balanced dataset is that if we use imbalanced dataset for this kind of problem, then the classifier tends to generalise most of the transactions as “Non-Fraud” but our objective is to identify the anomaly. The imbalanced dataset influences the classifier algorithms for generalisation.

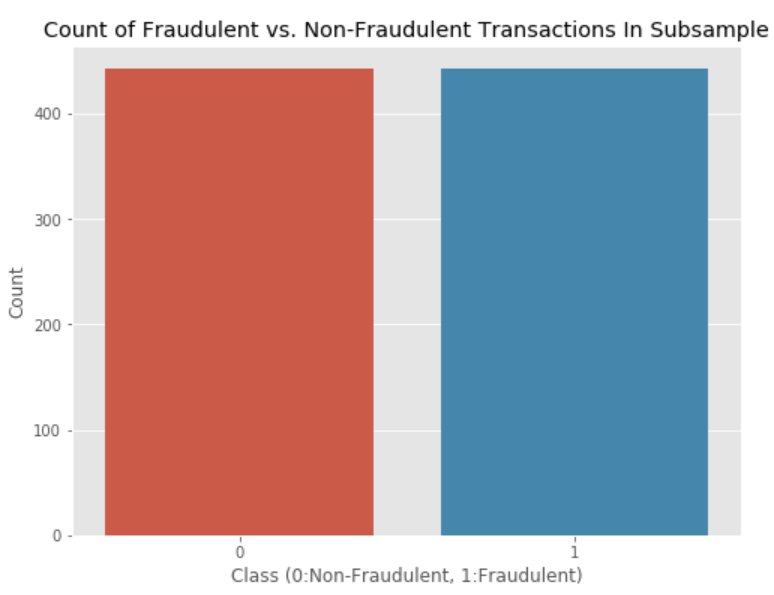
Before under-sampling, we need to split the imbalanced dataset into Train and Test dataset and thereafter we need to perform under-sampling with the training dataset. Always split into test and train sets BEFORE trying oversampling/Under-sampling techniques. Oversampling/Under-sampling before splitting the data can allow the exact same observations to be present in both the test and train sets. This can allow our model to simply memorize specific data points and cause overfitting and poor generalization to the test data.





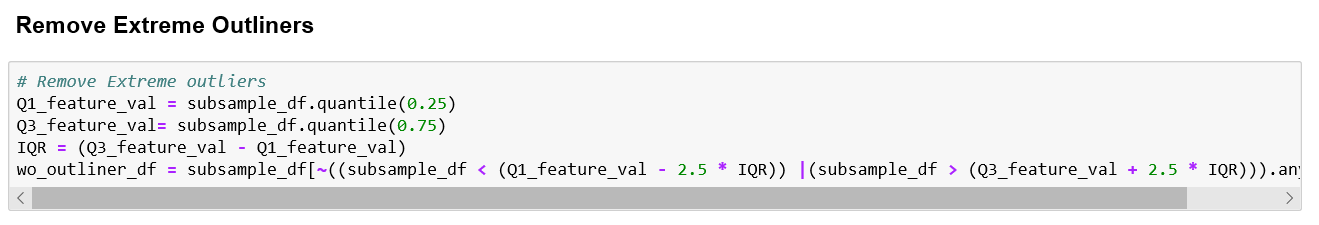


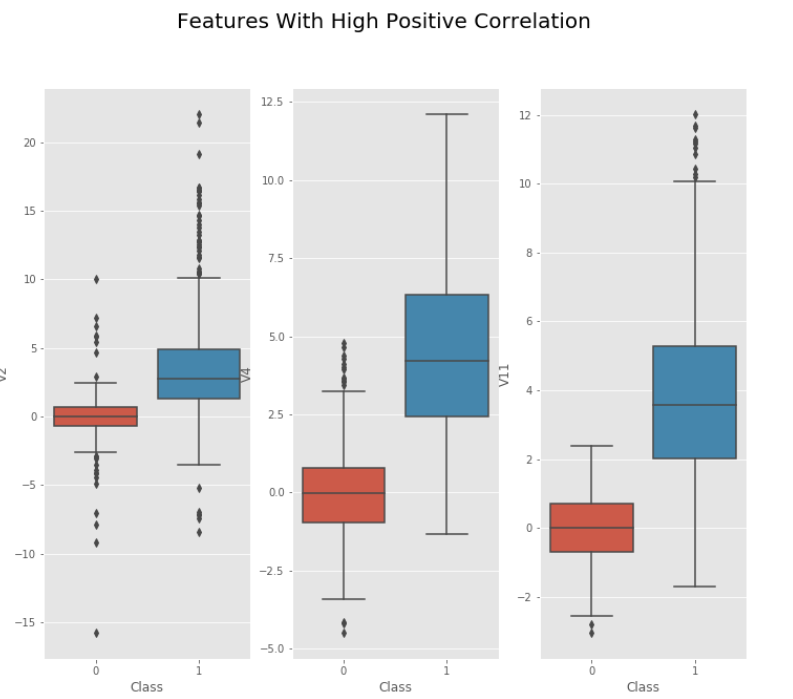




**Remove Outliners from the Balanced Dataset:**

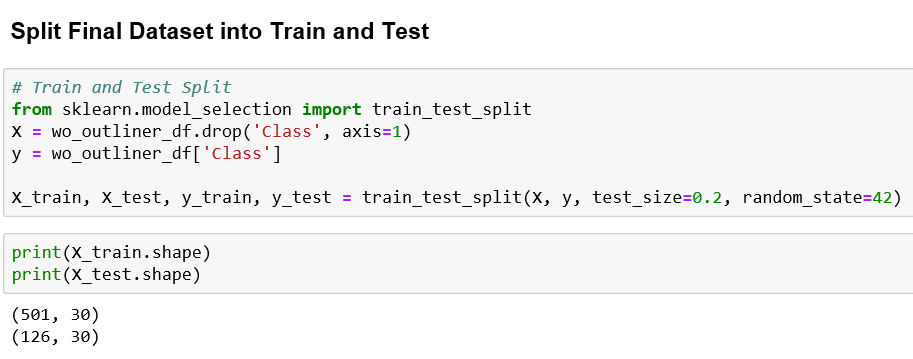
Outliners in the whole dataset can be visualized with the boxplot diagram, which is shown below. Below code snippet shows that observations below quantile 0.25 and above 0.75 will be removed from the balanced dataset.

****



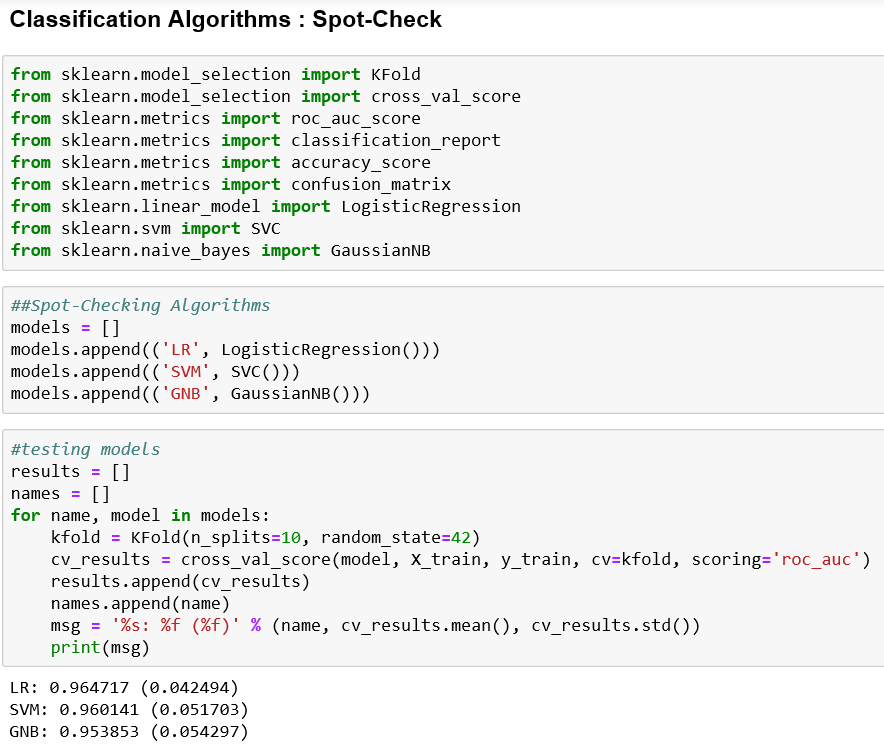
**Split Dataset into Train and Test:**

With scikit-learn’s model selection module, the dataset is split into Train and Test in the ration of 80:20 for learning purpose.



**Spot-Check Classifier Algorithms:**

In order for us to identify the best model for the given dataset, we take a couple of supervised classification algorithms and check which model best suits the dataset to generalise unseen data. For this, we use cross validation (cross\_val\_score function) to identity mean and spread (std). Based on the mean, we can identify the respective model’s performance score.

****

**Predictions on Test Data:**

In general, we use accuracy score for Classification problem but , as stated earlier, the dataset is imbalanced. In this case, the accuracy score is not the suitable metrics to scale the model’s performance. Hence, we need to consider ROC-AUC, Sensitivity/Recall/True Positive (TP) Rate, confusion matrix and classification report.

The ROC-AUC (Receiver Operator Characteristics -Area Under Curve) uses sensitivity to identify the area under curve (AUC). It is in the scale of 0 to 1. If it is less than 0.5, then the False Negative rate is high. If it is greater than 0.5, then True Positive rate is high.

Since it is an anomaly detection issue, it is good to have high sensitivity rate. If the sensitivity goes up, then specificity goes down.

**Precision:**

The number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier’s exactness. Low precision indicates a high number of false positives.

**Recall:**

The number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier’s completeness. Low recall indicates a high number of false negatives.

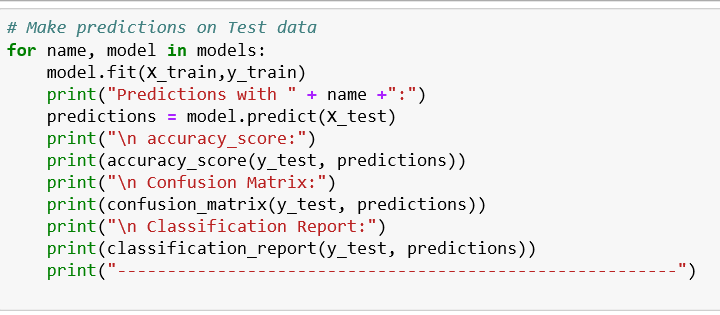
**F1-Score:**

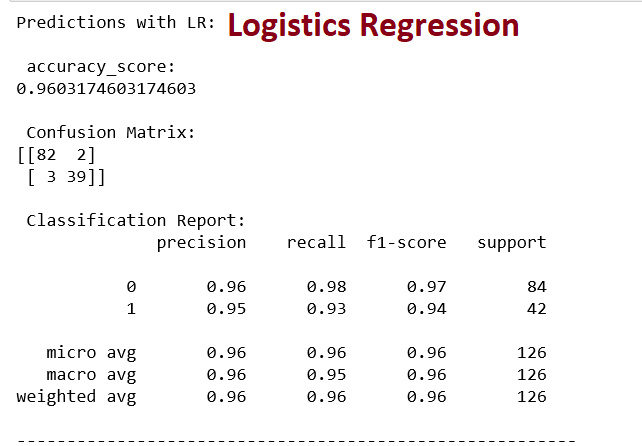
The weighted average of precision and recall.

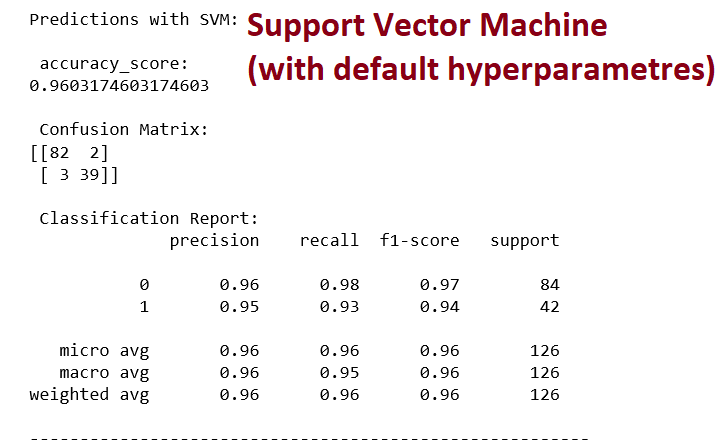
**Confusion Matrix:**

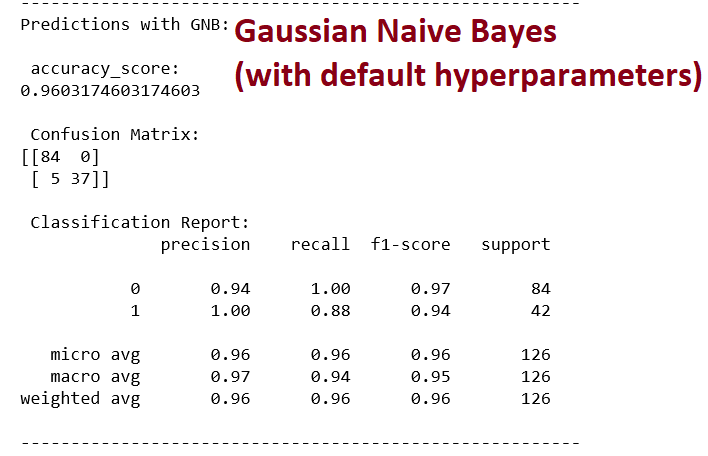
A table showing correct predictions and types of incorrect predictions.

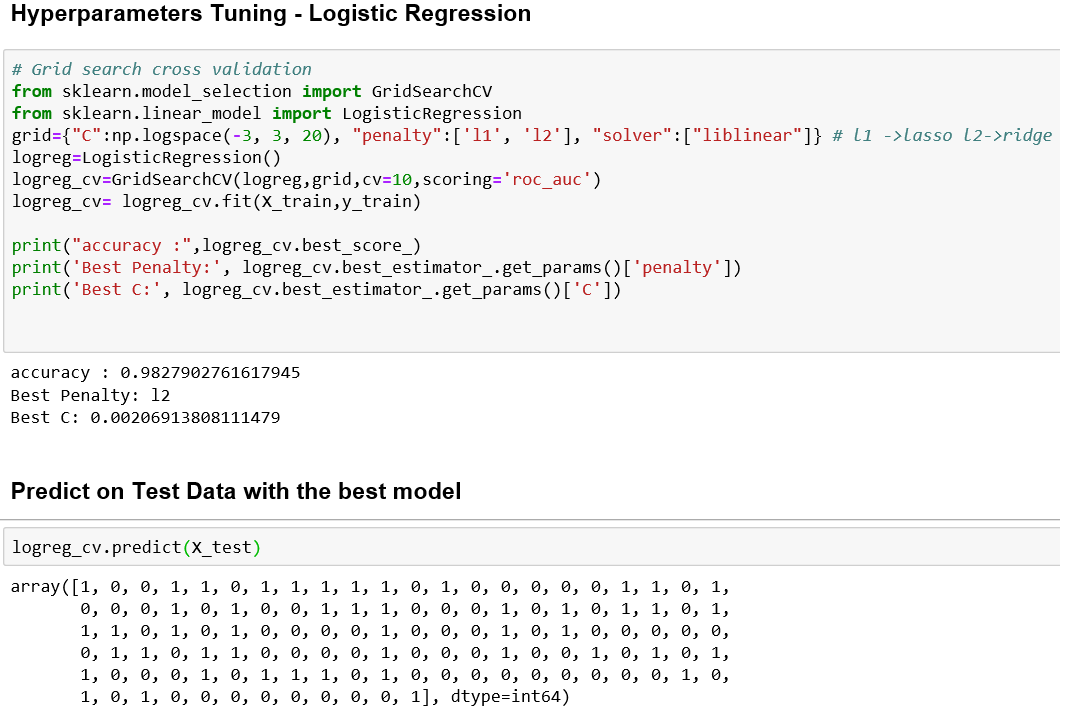
|  |  |  |
| --- | --- | --- |
|  | **Prediction** | |
| **Actual** | **0 (Negative)** | **1 (Positive)** |
| 0 (-ve) | TN (True Positive) | FP (False Positive) |
| 1 (+ve) | FN (False Negative) | TP (True Positive) |











# System Testing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Specification ID : TP\_CCFD\_01  Name of Tester : KS Karthik  Use Case ID : UC\_CCFD\_01  Date of Test : 27-Jul-2019 and 28-Jul-2019  Description of Test : It helps the tester to test if the system fulfils the requirements. | | | | |
| **S/No** | **Test Case** | **Expected Result** | **Pass/ Fail** | **Remarks** |
|  | * Prompts the user for dataset file path. | * + If we enter the file path, it checks if the file exists.   + Otherwise, it takes the default file path (on our local computer) | Pass |  |
|  | * Check if the dataset is balanced or imbalanced | * The class counts are not equal, which implies imbalanced dataset. | Pass |  |
|  | * Check the final dataset has the same no of normal cases as that of fraud transactions, after resampling. | * The final dataset, after resampling, has same no of normal and fraud transactions. | Pass |  |
|  | * Check if the classifier can identify the fraud/positive transactions on the Test Data with 90% performance rate. | * It can identify the fraud transactions with 90% performance rate. | Pass |  |
|  | * Check if the system generalise unseen data correctly. | * It generalises unseen data with 92 to 95% accuracy. | Pass |  |

# User and Technical Documentations

## User Documentation/Guide/Manual

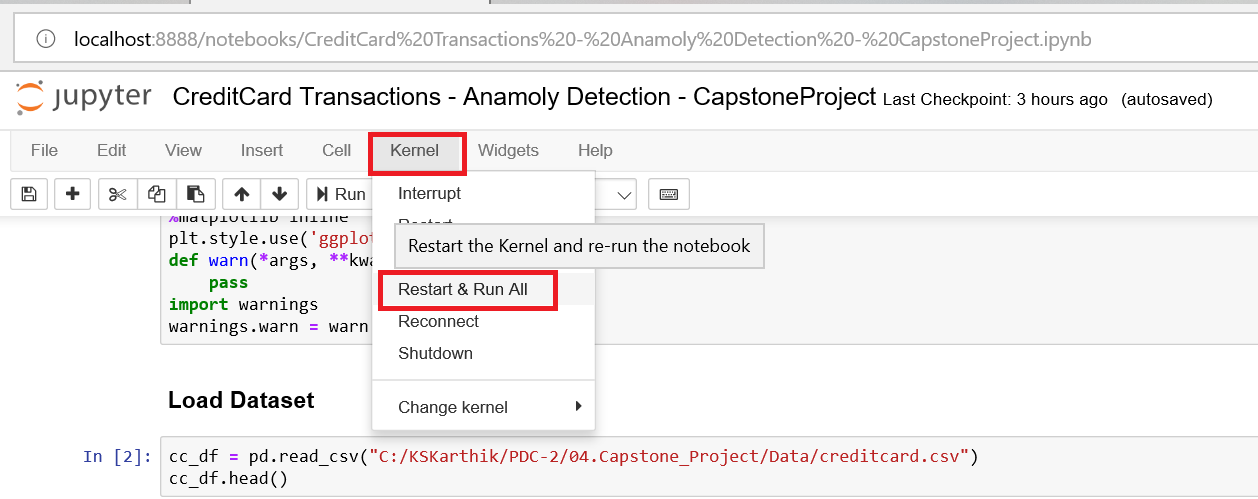
1. *It has two items* 
   1. *Jupyter Notebook file (****CreditCard\_Transactions\_Anamoly\_Detection\_CapstoneProject.ipynb****)*
   2. *Python Source Code file. (***CreditCardFraud\_Detection.py***)*

*CreditCard\_Transactions\_Anamoly\_Detection\_CapstoneProject.ipynb*

1. *To open the Jupyter Notebook file, you need to ensure if below items are installed on your Local computer.* 
   * + *Python 3.7*
     + *Numpy*
     + *Pandas*
     + *Scikitlearn*
2. *Open command prompt (cmd) and enter jupyter notebook to start the jupyter server.*
3. *Once you open the jupyter notebook console, upload the* ***CreditCard\_Transactions\_Anamoly\_Detection\_CapstoneProject.ipynb*** *file.*
4. *Change dataset file path in the jupyter notebook as outlined below.*

**

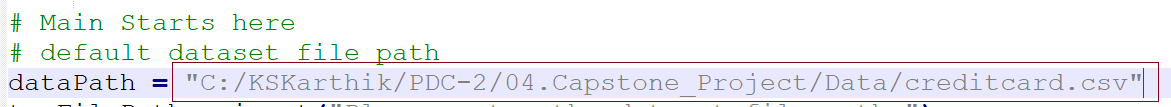
1. *Select “Kernel 🡪 Restart & Run Al”*

**

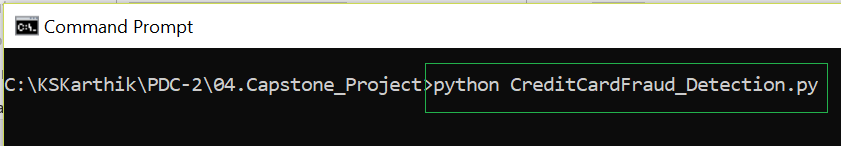
1. *It will enable the model to learn from the dataset and predict on test data.*

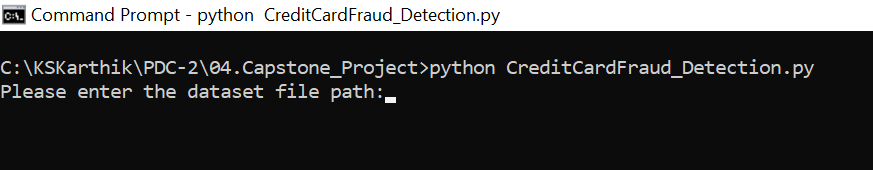
*To open and run the python source code (CreditCardFraud\_Detection.py), please follow the steps given below.*

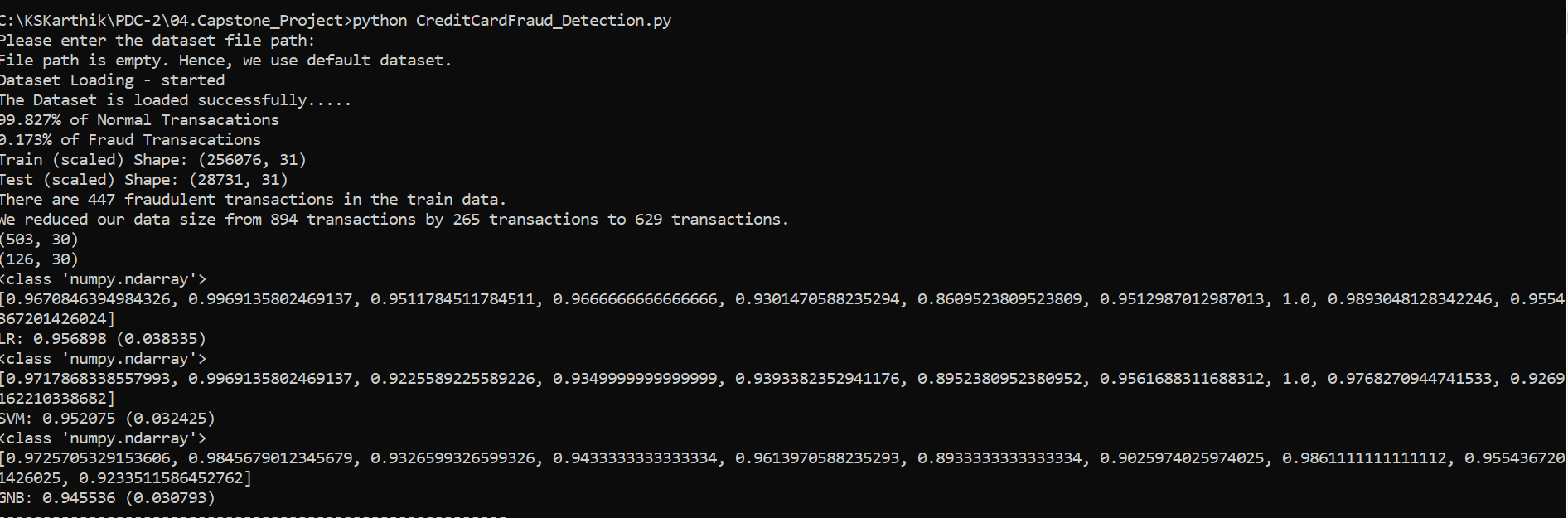
1. *Open the CreditCardFraud\_Detection.py file with Notepad / Notepad++ / IDE*
2. *Change Default file path as shown below.*

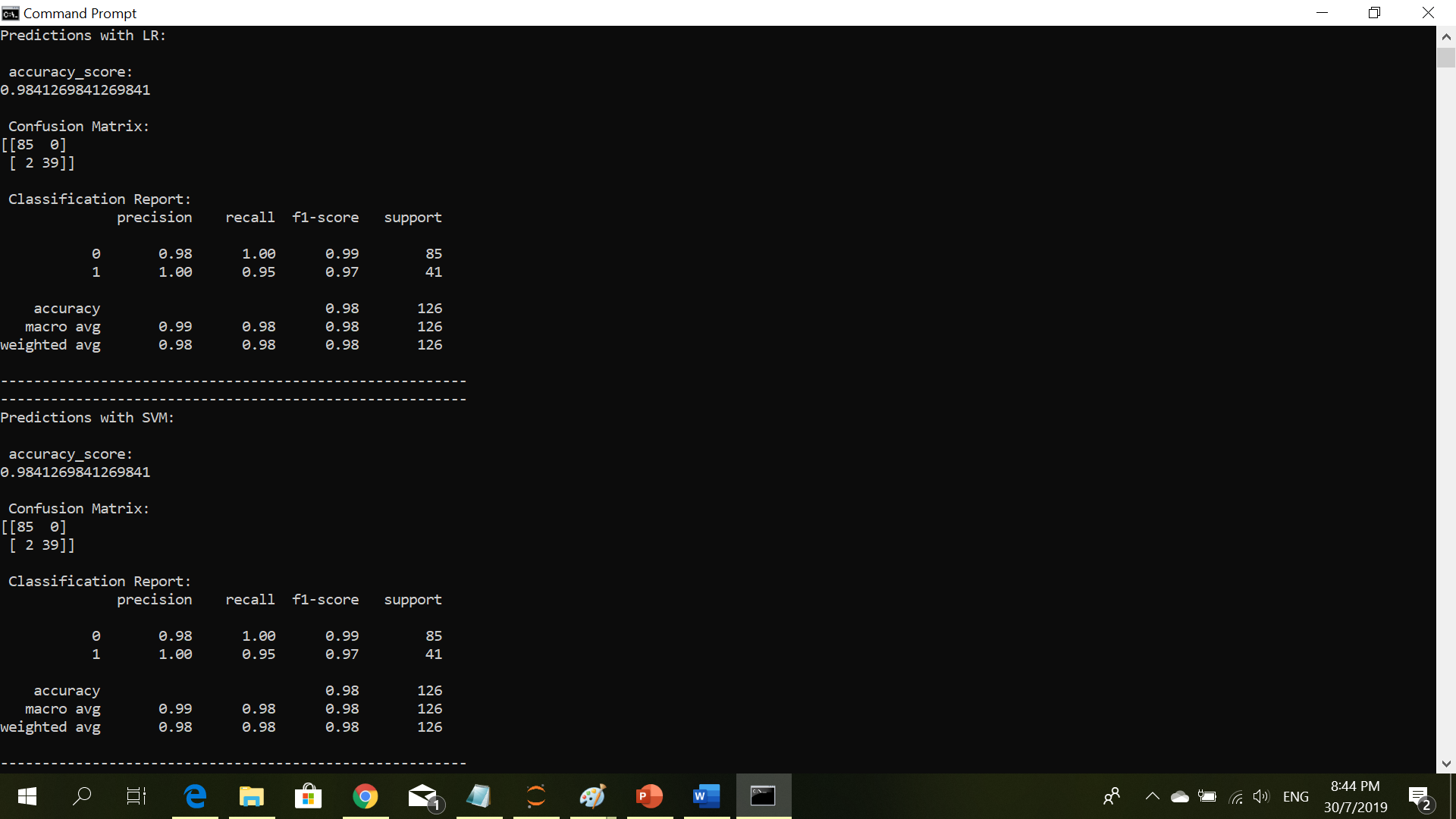
**

1. *Save & close the python file.*
2. *Open command prompt (cmd) in your local computer.*
3. *Run the program as shown below*

**

1. *Enter the dataset file path. Otherwise, leave it empty *
2. It will show below result.





## Technical Documentation (Installation guide/Manual)

**Python Package Installation:**

In order to use juputer notebook file and python source code, you need to ensure that you have Python3.7, Numpy, Pandas and Scikit-Learn Packages installed on your computer. Besides, the path are set accordingly. Use PIP package to install the above-mentioned python packages seamlessly.

For more info, please refer to

<https://www.python.org/downloads/>

<https://realpython.com/installing-python/>

https://scikit-learn.org

**Unpack Jupyter Notebook and Python Source Code:**

And thereafter you can create a folder (as you like) on your computer and copies these 2 files (CreditCard\_Transactions\_Anamoly\_Detection\_CapstoneProject.ipynb & CreditCardFraud\_Detection.py) into the folder for further exploration

For these 2 files, please refer to *Appendix A and Appendix B. You can download and save it on your computer as said above for further analysis.*

# Conclusions

*It gives me an immense knowledge on how to resolve a problem with the help of Machine Learning. Machine Learning has different kinds of learnings such as Supervised, Unsupervised and Reinforcement Learning. This project is accomplished with Supervised Learning (Classifier) algorithms.*

*With this project, we are able to identify/classify/predict the fraudulent credit card transactions. This project teaches me*

1. *How to get a high-level inference from the given dataset*
2. *Exploratory Data Analysis (EDA)*
3. *How to handle correlation between features*
4. *How to handle missing value (imputation)*
5. *Identify and remove outliners from the dataset*
6. *Resampling Techniques such as Oversampling, Under-sampling and SMOTE.*
7. *How to implement solution with the help of Numpy, Pandas and Scikit Learn frameworks*
8. *Tuning Hyperparameters of the model.*

# References

* https://scikit-learn.org
* https://towardsdatascience.com
* https://machinelearningmastery
* https://www.kaggle.com

# Appendices

***Appendix A:***

|  |  |
| --- | --- |
| Jupyter Notebook File |  |

***Appendix B:***

|  |  |
| --- | --- |
| Python Source Code |  |

# Project Poster

**