**Exploratory Data Analysis and Model construction**

**Introduction**

This project aims to use 23 customers case default payments to predict fraudulent transaction. The

problem statement is based on default payments and compares the predictive accuracy of the

probability of default among data mining methods. From the perspective of risk management, the

result of the predictive accuracy of the estimated probability of default will be more valuable than

the binary result of classification.

Following the literature review where I reviewed relevant research and determine how they compare and differ from my capstone project; I proceeded to explore the data and begin model construction. The purpose of this paper is to summarize the exploratory data analysis and Model construction activities.

**Exploratory Data Analysis**

The purpose of exploratory data analysis it to perform pattern discovery analysis on data using summary statistics and graphical representations. Here, I investigated the data by checking for data unbalancing, perform correlation analysis, visualized the features and understanding the relationship between different features, and prepare the data for model construction. For this capstone project, I discovered some outliers in the features so I have to remove 1% of the features which were used to build the algorithms. Also, I plotted a correlation matrix to show features with the highest absolute correlation. Additionally, I ran Logistic Regression, Random Forest, Decision Tree, KNeighbors and Gaussians algorithms both on standardized and unstandardized data. Random Forest algorithm seems to perform better among others.

**Model Construction Methodology**

* Split dataset into training and test sets.
* Run Logistic Regression, Random Forest, Decision Tree, KNeighbors and Gaussians algorithms.
* Test model.
* Check accuracy.

**Model Observation**

The algorithms was ran on both standardized and unstandardized data. Among those, the

Random Forest of the unstandardized seems to perform better compared to others producing

test accuracy and f1-score. Since my focus is to catch more of fraudulent transaction (1) than

non-fraudulent transaction (0). Below are the breakdown results from the classification tables

and precision-recall curves plot:

# Accuracy – 0.82

# Precision 0 – 0.85

# 1 – 0.65

# Recall 0 – 0.94

# 1 – 0.38

# F1 score 0 – 0.89

# 1 – 0.48

# 

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

**Next Steps**

* Finalize analysis by doing some tuning on how to improve the model.
* Re-run machine learning algorithm and use confusion matrix to test accuracy of true positives.