**NAME: WAIRIUKO SAMUEL**

**TITLE: CLASSIFICATION USING LOGISTIC REGRESSION**

**TITTLE: CREDIT CARD DEFAULT PREDICTION**

**1.0 INTRODUCTION**

This project is about building a prediction model to predict Credit Card default. I got the dataset from [www.kaggle.com](http://www.kaggle.com) and build a successful classification model using Logistic regression with accuracy above 82%. The dataset had 25 variables and 30000 observations. The 25 variables include the target variable Default.

**1.1 PROPERTIES OF DATASET;**

* ID
* LIMIT BALANCE; means amount of credit given
* SEX; means gender 1 is male gender 2 is female
* EDUCATION; means 1 is university,2 is college,3 is high school,4 is others
* MARRIAGE; 1 means is married, 2 is single, 3 is others
* AGE
* PAY; historical payments from September to April 2005 i.e. sept, Aug., Jul., Jun, may, April. This show whether the customer pay credit on time.eg if he missed to pay the credit on time for two times during this period of 6 months the target variable predict default

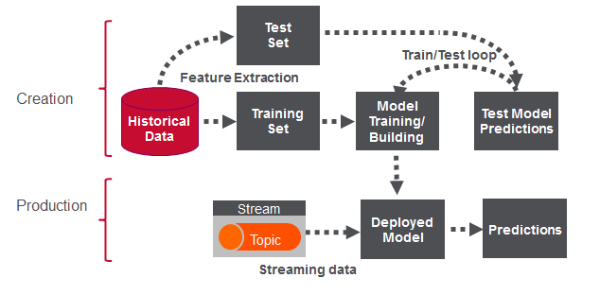
0 is repayment status in sep,1 is repayment status Aug.……6 is repayment status for April. -1 is prepayment 1 month and -2 is prepayment 2 months

* BILL AMOUNT divided into 6 columns; historical bills paid from September to April 2005 i.e. sept, Aug., Jul, Jun, may, April. This show whether the customer pay bill on time.eg if he missed to pay the bill on time for two times during this period of 6 months the target variable predict default
* PAY AMOUNT; amount of credit loan paid
* PREDICTION NEXT MONTH; Y predict;1 default,0 no default

**OBJECTIVES**

* To build a precise model for predicting profits.
* To analyze the dataset and give conclusion and insights to the management.

**2.0 DATA PREPARATION, CLEANING & ANALYSIS**

[](https://www.google.ca/url?sa=i&url=https%3A%2F%2Fmapr.com%2Fblog%2Freal-time-credit-card-fraud-detection-apache-spark-and-event-streaming%2F&psig=AOvVaw2ghL1uMDaSbPIwXvJi_6jG&ust=1582556729755000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPC02cb55-cCFQAAAAAdAAAAABAN)

I followed the following steps;

* Imported the Libraries
* Imported the dataset
* The data was checked for duplicates and missing values and irrelevant variables dropped.
* Correlation analysis
* Selected the appropriate model and import from sklearn
* Divided the dataset into training and set.
* Trained the model on training set
* Evaluated the model on the test set

**3.0 DATA ANALYSIS & SUMMARY FINDINGS**

3.1 Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| CONFUSION MATRIX | | |  |
|  | 1  (PREDICTIONS) | 0  (PREDICTIONS) |  |
| 1  (ACTUAL) | TRUE POSITIVE  6897 | FALSE NEGATIVE    163 | 7060 |
| 0  (ACTUAL) | FALSE POSITIVE  1492 | TRUE NEGATIVE    448 | 1940 |
| TOTAL | 8398 | 602 | N=9000 |

True Positives; A total of 6901 customers defaulted and were correctly identified by the algorithm

True Negatives ;443 customers didn’t default and were correctly identified by the algorithm

False Negatives;159 customers defaulted while the algorithm predicted they didn’t

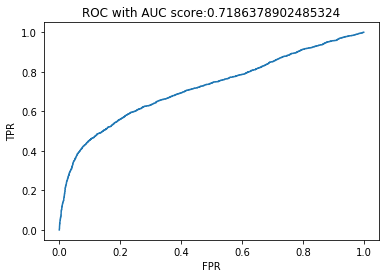
False positives-the algorithm misclassified that 1497 customers won’t default while they did.

3.2 Classification Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CLASSIFICATION REPORT | | | | |
|  | PRECISION | RECALL | F1-SCORE | SUPPORT |
| 0 | 0.82 | 0.98 | 0.89 | 7060 |
| 1 | 0.74 | 0.23 | 0.35 | 1940 |
| ACCURACY |  | | 0.82 | 9000 |
| MACRO AVG | 0.78 | 0.60 | 0.62 | 9000 |
| WEIGHTED AVG | 0.80 | 0.82 | 0.78 | 9000 |

Precision refers to percentage of relevant results,recall refers to total percentage of total rerevant results correctly classified by the algorithim.

3.3 Area Under the Curve (ROC) WITH Receiver Operating Characteristics (ROC) SCORE



The Area Under the Curve (AUC) and Receiver Operating Characteristics (ROC) curve shows how well the model is capable of distinguishing between True Positive Rate and False Positive Return

**4.0. CONCLUSIONS & RECOMMENDATIONS**

**4.1 RECOMMENDATIONS**

False Positives should be minimized to increase the accuracy level of the algorithm through increase of the test size.

**4.2 CONCLUSIONS**

* The model test score is 82% compared to cross-validation score of 81%.
* The model is good but not the best.
* The management should ensure there are several modeling algorithms for accuracy comparison.

**5.REFFERENCES**

* www.kaggle.com {open data set}
* Python Project Codes(attached)