**REPORT ON THE FINANCIAL INCLUSION IN ARICA ML PROJECT**

**Project Overview:** This project aims to develop a machine learning model to predict customers would default on a loan or not. Customer defaulting or not defaulting on a loan is critical in the financial sector since it can cause financial loss.

**Data Description:** The dataset used in this project contains information about customer demographics, the lender, amount the borrower receives, duration, and total amount to be paid back by the borrower.

**Data Preprocessing**

**Handling Missing Values:** I checked for the availability of missing values but there seems to be none in all the columns.

**Handling Duplicated Values**: Ther were 5499 duplicated values in the dataset. I dropped all of them because they can cause trouble to my model.

**Feature Engineering:** I removed columns which would not contribute to the prediction process such as customer and lender ID

**Encoding Categorical Variables:** Categorical variables were converted into numerical representations using one-hot encoding.

**Resampling**: I made sure the output values were balanced. I did so by using the random sampling method which helped me balance both equally (50,50)

**Splitting data:** I split the data into test and train data so the test data can be used to test the model since it has not seen it before.

**Outlier detection and removal:** I detected the presence of outliers by using box plot visual. I transformed it by using the log transform which helped adjust the skew of each column.

**Model Selection and Training**

The model I used was Logistic Regression.

**Evaluation Metrics:**

Model score = 0.93

Confusion matrix = [963 42]

[ 98 910]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Support |
| 0 | 0.91 | 0.96 | 0.93 | 1005 |
| 1 | 0.96 | 0.90 | 0.93 | 1008 |
| Accuracy |  |  | 0.93 | 2013 |
| Macro avg | 0.93 | 0.93 | 0.93 | 2013 |
| Weighted avg | 0.93 | 0.93 | 0.93 | 2013 |
|  |  |  |  |  |

**Conclusion**

The model used was logistic regression because the data has columns which are categorical. Per my observations of the performance of my model, it is due to the resampling of the data which helped reduced biased predictions.