**LOAN DEFAULT ANALYTICS REPORT BY OSEDEME KENOSE**

**CONTENT**

1. Introduction
2. Data
3. Methodology
4. Findings and Recommendation
5. **Introduction**

In this case study, we are given a Loan Default dataset that contains 31 columns and 888,000 rows. The objective of this project is to predict the target variable; “Loan\_No\_Loan” which contains 1(Loan) and 0(No Loan). This simply means that given the parameters in the train dataset, we should be able to predict whether the customer should be given a loan or not.

For the purpose of this case study, I have adapted a Problem Statement that will help guide us in the exploration and analytics of this data.

*“A US Loan company Xerloan has contacted you as a technology consultant to aid in the development of a system that can help in determining whether individuals who apply for loan should be granted or not, based on their personal data”.*

1. **Data**

The historical dataset is in form of Excel worksheet. It contains columns like ID, US states, experience, validation, yearly income, home status, if individual has already defaulted, Job type, etc.

By analyzing variables that describe loans and the financial situations of their borrowers, we may determine key relationships between default rates and a few other variables.

1. **Methodology**

* **Data Preparation**- The original dataset had 888K entries and 31 columns but some of the observations were not related to our business goal. Using SQL in Microsoft SQL Server views were created to specify the columns needed and then imported into Power BI through the import connectivity mode and specifying SELECT \* FROM *created view.* The resulting view had just 20 selected columns.
* **Data Exploration**- After preparing the response variable based on the values of the status variable and dealing with missing values the next step is to first look at the distributions of and between some characteristics of the loan or the borrower of the loan. Graphs were made to explore the relationships between the predictors and loan status. For instance, a side-by-side boxplot of a quantitative variable to see if the variable is distributed differently for Loan and No loans. After all this the data is split into 20% test and 80% train.
* **Building the Logistic Regression Model** -The next step is to create a logistic regression model, using the training data, that uses all of the remaining predictors to predict loan status using the default threshold of 0.5.

Using a Confusion Matrix and Statistics;

**Prediction**  0 1

0 12193 2427

1 290 524

**Accuracy**: 0.824, **Confidence Interval**: 95%

* **Evaluate the Model** -By varying the classification threshold from 0.5 the accuracy can be fine-tuned. The threshold should be being optimized for accuracy. The final accuracy of the model is 82.4%.

1. **Finding and Recommendations-**

Some of the key findings from the visualizations are:

* The individuals with mortgage house types have the highest unpaid loans, followed by those in rented houses
* California is the State in the US with the highest account open as well as loans followed by New York then Texas.
* California also has the highest yearly income and this also helps them to pay back their loans easily
* School Teachers have the highest amount lent according to our data. Perhaps because of their low income.

As a further recommendation, Xerloan should definitely give priority to individuals in higher income states such as California, New York, Texas, Florida, NJ, and so on. Because they stand a higher chance of paying back their loan. Also, individuals in the Air Force, AMC Mechanical International Companies as well as some managers have the lowest amounts lent. Finally, from the prediction results we can see that about 90% of people will not be granted loan (No Loan)