Credit Card Default Prediction using Machine learning techniques

**Project Description**

**Business Context**

Banks are primarily known for the money lending business. The more money they lend to people whom they can get good interest with timely repayment, the more revenue is for the banks. This not only save banks money from having bad loans but also improves image in the public figure and among the regulatory bodies.

The better the banks can identify people who are likely to miss their repayment charges, the more in advance they can take purposeful actions whether to remind them in person or take some strict action to avoid delinquency.

In cases where a borrower is not paying monthly charges when credit is issued against some monetary thing, two terms are frequently used which are delinquent and default.

Delinquent in general is a slightly mild term where a borrower is not repaying charges and is behind by certain months whereas Default is a term where a borrower has not been able to pay charges and is behind for a long period of months and is unlikely to repay the charges.

This case study is about identifying the borrowers who are likely to default in the next two years with serious delinquency of having delinquent more than 3 months.

**Objective**

Building a model using the inputs/attributes which are general profile and historical records of a borrower to predict whether one is likely to have serious delinquency in the next 2 years

We will be using Python as a tool to perform all kind of operations in this credit score prediction machine learning project.

**Dataset**

In this credit scoring system project, we will use a dataset containing two files- training data and test data. We have a general profile about the borrower such as age, Monthly Income, Dependents, and the historical data such as what is the Debt Ratio, what ratio of the amount is owed with respect to the credit limit, and the no of times defaulted in the past one, two, three months.

We will be using all these features to predict whether the borrower is likely to default in the next 2 years or not having a delinquency of more than 3 months.

**Main Libraries used**

Pandas for data manipulation, aggregation

Matplotlib and Seaborn for visualization and behavior with respect to the target variable

NumPy for computationally efficient operations

Scikit Learn for model training, model optimization, and metrics calculation

Imblearn for tackling class imbalance problem

Shap and LIME for model interpretability

Keras for Neural Network(Deep Learning architecture)

**Approach for Credit Card Default Prediction in Python**

1. **Data Cleaning**

Data cleaning is the process of organizing and correcting data that is badly structured, incomplete, duplicate or, otherwise messy. It involves eliminating inconsistencies in data, as well as reorganizing data to make it much easier to use. Standardization of dates and addresses, ensuring consistent field values (e.g., "data cleaning" and "Data Cleaning"), parsing area codes from phone numbers, etc., are all instances of data cleaning.

In this project, we will treat outliers, resolve some accounting errors, and treat missing value values.

1. **Feature Engineering**

When developing a prediction model using machine learning or statistical modeling, feature engineering refers to the method of selecting and transforming the most significant variables from actual data using industry knowledge. The purpose of feature engineering and selection is to boost machine learning algorithms' efficiency. This credit score prediction project entails applying feature engineering techniques to the training and test dataset. It also involves scaling features with Box-Cox transformation, standardization, upsampling, downsampling, and SMOTE.

**Deep Learning Algorithms**

Deep Learning is a set of algorithms driven by the human brain's data-processing and pattern-creation capabilities, which are advancing and developing on the idea of a single model architecture termed Artificial Neural Network. Deep learning is a part of Machine Learning that does data processing and calculations on a large quantity of data using numerous layers of neural networks. In this credit scoring system project, we have built a neural network model and fitted it on Box-Cox transformed credit score dataset, Standardized credit score dataset, etc. For this credit scoring system project, we have a number of deep learning algorithms (Logistic regression, Random Forest, XGBoost, etc.) being applied to the prediction model.

1. **ROC AUC Curve**

The Receiver Operating Characteristic curve, or the ROC curve, is a graph of the false positive rate (x-axis) vs. the true positive rate (y-axis) for a variety of candidate threshold values ranging from 0.0 to 1.0. The roc\_auc\_score() function computes the area under the ROC curve. The project involves plotting ROC AUC plots for each of the machine learning algorithms and for each transformed dataset.

1. **MLFoundry**

TrueFoundry's MLFoundry experiment tracking and model monitoring system combines the strengths of open-source tools such as MLFlow, Whylogs, and others. It comes with a shareable dashboard where you can keep track of your tests and model, among other things.  We create a client for the MLFoundry repository and assign a project name. To make experiment tracking easier, we assign different names for different experiments as well as different runs.

**FAQs**

**Q1. What are the types of credit scoring models?**

FICO and VantageScore are the two different types of credit scoring models.

**Q2. What is the most common credit scoring system?**

The FICO scoring system is the most commonly used and reliable scoring system due to its proven track record.