**PROBLEM STATEMENT -**

**Loan Defaulter Prediction by Application Information**

1. **Define and explain the problem. Why is this problem important to solve and what is the significance of this problem to the organization/society/people? What are the goals that the organization wants to achieve with your solution?**

In today's day and age, banks and financial organizations are swamped with a plethora of applications and submissions for loans. The loan approval process, which relies on several factors, entails exhaustive documentation and requires a lot of time, which the banks can utilize in devising profitable schemes for their customers. On the other hand, consumers must bear the brunt of tedious loan application processes that involve waiting for the application status for long periods.

Addressing this predicament of the consumers and the banks is crucial as it is instrumental in speeding up the loan approval process. Based on the customer's background, **identifying whether a specific customer will default a particular loan** is the **crux** of the problem. Consumer background includes assessing the consumer profile through information pertaining to their employment status, income, marital status, to name a few. Additionally, banks are working diligently with the credit department and investing many resources to eventually identify borrowers posing a credit risk to steer clear of losses.

We are implementing a machine learning approach to identify and understand the credit risk of a loan applicant. As we will be predicting the applicant's capability to repay a loan by unleashing all aspects of data, our approach would be all-encompassing. An educated machine learning solution would save the time of both consumers and banks drastically. It will also enable banks or financial institutions will save on monetary and time-wise resources, pinpoint their probable defaulters, and rid themselves of any losses.

1. **Explain your solution and why your proposed solution should solve the problem. What kind of value will your solution bring to the organization, society, or lives of the people?**

Developing a machine learning model to obtain loan prediction based on customer behavior. Through this model we can predict who the possible defaulters are for the customer loan products. This model will help the organization to monitor its non-performing assets, differentiate between defaulter and non-defaulter customers. The organization can control its non-performing assets by predicting the probable defaulters. The organization will be able to use its resources wisely to offer better services to its customers.

1. **How does the organization currently work to solve the problem (if any)? How will your solution fit into their current process and help them improve the outcomes?**

Currently the organization entirely relies on their Due-Diligence team for manual portfolio and background checks to ascertain the dispatch of loans to legible candidates and thus screen the possible loan defaulters. Several factors are weighed in on an optimum benchmark before finalizing the portfolios of eligible customers. The complete financial profile of a customer is studied, that includes parameters like credit score, credit history, income, assets, history of repayment and the amount of debt currently in the market under the candidate’s name. The credit / FICO score plays the most important role in determining the probability of sanctioning a loan. Hence, the whole process is currently running manually in the organization and is consuming a huge number of work hours, thus reducing the overall efficacy of the organization.

A machine learning prediction model will be trained and prepared by the end of this project, using the public training and test data obtained. This model will be aimed at taking in the candidate's financial profile and other relevant data and based on the training dataset, it will predict whether a candidate is likely to default his repayment or not. This model will not only reduce the manual working hours by manifolds but will also improve the accuracy of loan defaulter prediction as the model will be capable of taking more factors into consideration compared to the current process. This will also eliminate any human error thus increasing accuracy of prediction thus increasing the profitability of the organization altogether.

1. **How will you acquire data? Will the organization provide, or will you find public data, or will you need both? What will be the likely size of the data?**

* We plan to acquire the data through public Loan prediction data sets. As defined by fundamental concepts of model building, we follow the practice of bifurcating test datasets and training datasets. In our case, we plan to apply the training datasets for model building.
* As discussed in point (a), we plan to acquire public data for now. However, if we receive organization specific data in future which is like the data gathered from public datasets, we will evaluate the possibility to apply that as well.
* Regarding the likely size of the data, we plan on 12 features and a defaulter with an approximate data row count of 3276013.

1. **What data analytic components will your solution include? E.g., data acquisition/collection, data pre-processing, exploratory analysis, predictive modeling, evaluation, etc.**

1. **Data Acquisition/Collection**: - We have acquired a public dataset that will serve as our primary dataset to address the loan defaulter problem. After developing our final model, the banks/financial institutions can incorporate it at the preliminary stage of the loan application process. For this purpose, we would be replicating their datasets in a manner analogous to the dataset we have used to construct the model.
2. **Data Pre-Processing: -** The pre-processing of data mainly involves identifying the missing or null values after analyzing the distribution. We may also need to perform **One-Hot encoding** of the relevant categorical variables. To tackle any anomalies due to scaling, we would standardize all such features that contribute to this anomaly. These are further explained later in the document.
3. **Exploratory Data Analysis: -** One of the best ways of performing the EDA would be by visualizing the features through a box plot or a histogram. Visualization will depict the data distribution and aid in identifying essential features by using the mean of the data for consumers who are defaulters as against those who aren't.
4. **Predictive Modeling: -** We would be exploring the behavior of our data with different models such as **Logistic Regression, Random Forest Gradient Boosting, XGboost**, and **Neural Network**. Active testing of the performance of models will be done too to select the one which fits our requirement eventually.
5. **Evaluation: -** A detailed study of the result of the confusion matrix for each of the models would guide the assessment.

1. **How will you process your data, in real-time or in batches? Give an overview of the data, e.g., data types, variables, labels, etc. What kind of preprocessing techniques will you apply and why?**

To train and test a model for predictive analysis, a huge data set is required, in order to study the trend and rule out outliers. In this case we have around 250 million records. Hence, for training the model, data is taken in batches. However, once the model is ready for prediction, real-time or batch data can be accepted as input for loan defaulter prediction. A complete overview of the data is given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| * **Data types -** String, Integer | | | * **Label –** risk\_flag | | | |
| **Column** | **Description** | | **Type** | **Feature Type** |
| income | Income of the user | | int | Numerical |
| age | Age of the user | | int | Numerical |
| experience | Professional experience of the user in years | | int | Numerical |
| profession | Profession | | string | Categorical |
| married | Whether married or single | | string | Categorical |
| house\_ownership | Owned or rented or neither | | string | Categorical |
| car\_ownership | Does the person own a car | | string | Categorical |
| risk\_flag | Defaulted on a loan | | string | Categorical |
| current\_job\_years | Years of experience in the current job | | int | Numerical |
| current\_house\_years | Number of years in the current residence | | int | Numerical |
| city | City of residence | | string | Categorical |
| state | State of residence | | string | Categorical |

**Pre-processing Techniques –**

1. **Data Cleaning/Cleansing:** In this process, we will be identifying the missing values and replacing them with suitable values that would smooth out the dataset thus reducing redundancies and, in the process, will identify the outliers and rectify the inconsistencies.
2. **Data Transformation:** For exploratory data analysis and visualization, the data is classified into Numerical and Categorical data. In this process, the categorical data will be encoded and transformed into numerical data and the complexity will be reduced.
3. **Numerosity Reduction:** Finally, the cleaned and compact data set will be furthermore estimated by alternative and smaller data representations such as parametric models (which store only the model parameters instead of the actual data, for example, Regression and Log-Linear Models or non-parametric methods like Clustering, Sampling, and the use of histograms.

Real world data is generally noisy, incomplete, and inconsistent. This leads to a poor quality of collected data and further to a low quality of models built on such data. Sudden missing data can lead to faulty analysis and thus incoherent predictions. To eliminate such inconsistencies a data mining technique that involves transforming raw data into an understandable format called data preprocessing is undertaken. This process organizes the data into a proper form for better understanding and further wrangling.

1. **What predictive analytics/modeling solutions will you propose to address the problem and why this approach is a good solution?**

We will propose a machine learning model solution that works on data that the bank is already collecting from the loan applicants. It will be faster and more accurate than the current heuristic approach. We will use logistic regression to achieve baseline model accuracies and after that, we will use more advanced models such as random forest, gradient boosting trees to achieve more accurate models.

1. **What will be your evaluation strategy? What metrics will you use to evaluate your predictive model and why? How will you evaluate the improvement that your solution provides to the outcomes of the organization? E.g., how will KPI change after your solution is integrated into a company’s ecosystem?**

We will split our training data into 80:20 for training and testing data respectively. From this we can calculate the performance of our model on the testing data set. Because our data is highly skewed, accuracy is not a good measure for the data. We want to reduce false negatives, that is if the person is going to be a defaulter and we predict that he is not going to be a defaulter. To reduce that important matrix for us will be **Recall** which is = **True Positive / (True Positive + False Negative)**. We can also look at the **F1 score** of the model to get a holistic performance of the model

1. **How do you expect your predictive model to be maintained over time, in terms of longevity and adaptability to the changing conditions and information in data? What would be your maintenance strategy?\**

* **Data Longevity:** Banking and finance institutions have dedicated clusters available for storing customer data with measures in place to ensure data security and integrity.  Therefore, we do not foresee an immediate issue with data longevity.
* **Adaptability:** Considering that we have a structured training dataset, we do not foresee the need to train very frequently. We will analyze the changing data trends to foresee the frequency for adaptability.  For instance, if we are to encounter too many missing values or skewed data in a particular duration, we might need to adapt the model sooner than necessary.
* We will maintain the data on a secure cluster or cloud with unique access.

1. **How will the organization be expected to utilize your solution in their workflow? What is the capacity requirement for the company to utilize your solution?**

We can get all the loan applicant's data on daily basis and feed it to our model to get the default probability of the specific person.  The model will be lightweight, so it will be easier for the organization to adopt it. Normal EC2 instance / Computer can be used for utilization.