Title	**Personal Loan defaulter prediction**
Student name:	**Ashish Sasanapuri**
Supervisor name:	**Jize Yan**

Aims/research question and Objectives

Banking industry are one of the most important aspects of a nation's economy. It has in many ways helped in contributing to the economic household of the citizens. Among those is the loan business which has been established by the banks for the consumers to run their households. The loan lending system is a risky business that has to be carried out by the banks in order to balance their revenue. However, many banks follow a rigorous procedure of analyzing and running legal background checks on the customer to lend a loan. Customers borrow loans for various reasons, be it for a business, education, or house. The loan lending process can be a part of the traditional bank industry as well as the internet finance industry. The major problem faced by banks that affect their economy is loan defaults. There is a huge risk of customers not being able to pay their debt on time or in some cases avoiding paying entirely.

Loan default analysis has become a necessity in risk assessment to avoid bankruptcy. The banking industry requires to follow strict guidelines when assessing a customer's background. It involves many factors such as loan history, customer assets, customer income, etc. The aim of this project is to design a prediction system using machine learning algorithms to predict whether an individual who plans to borrow a loan from a bank will be able to pay the debt on time. The risk involved in loan defaults has an impact at an organizational level which falls upon the bank to be liable for the actions. Implementing this project will help the banking industry avoid relying on 3rd party agencies and identify the factors impacting the loan defaulters. An early prediction will help minimize the risk associated with loan disbursements to individuals.

Summary of proposed research and analysis methodology

The main objective of this project is to design a personal loan defaulter prediction system that provides an outcome of whether a customer who is willing to borrow a loan is susceptible to loan defaulting or not. This involves the use of a machine learning approach to build a model which decides the mentioned output. Following are the objectives of the project:

1. Research/ Study - A substantial amount of research and study is required to understand the various factors leading to bank loan defaults. This covers all the methods and approaches adopted by banks or organizations to provide risk assessment solutions. Many machine learning models which have been adopted for this purpose have to be studied. Various research papers can be obtained from sources like IEEE Xplore, arXiv, Google Scholar, etc., for this purpose.

2. Data collection – A plethora of dataset repositories are available publicly such as Kaggle, Lending club dataset, etc. While collecting data, we need to make sure relevant parameters are available for training.

3. Cleaning the data – All datasets are not easily understood. Most datasets contain additional data which might be insignificant or data that might be in an unreadable format. There may be errors in the data or outliers which might need to be handled via some pre-processing techniques. Apart from this, the data points might not be in standard form. Hence, scaling of the data is vital.

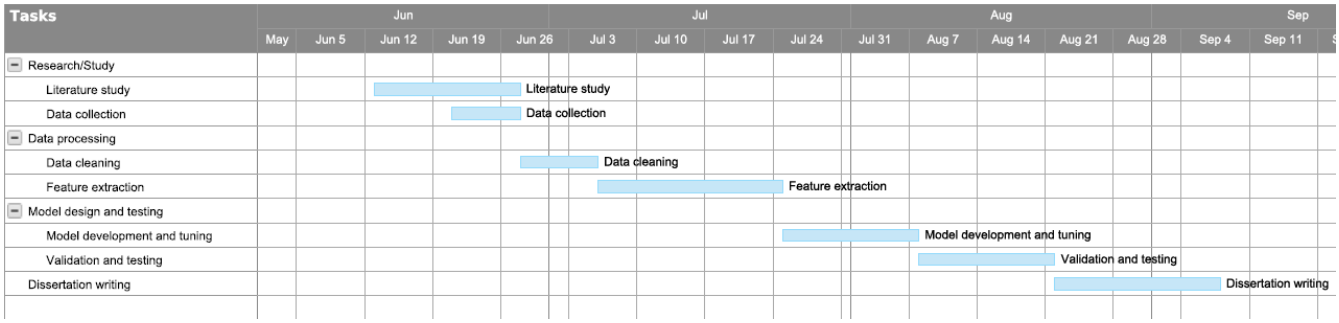
4. Understand data for feature extraction and tuning – This phase involves exploring the data using different feature extraction methods such as EDA, PCA or ICA, etc. This step plays an important role in understanding the data and the parameters involved. The mentioned procedures help in extracting features that impact the performance of machine learning models.

5. Selecting Machine learning model – There are many Machine learning models which help in training the dataset and predict the outcome as per requirement. These usually involve statistical models, neural networks, or ensemble techniques. Also, deep learning will be explored for better performance.

6. Performance metrics – Comparing and evaluating the machine learning models involves the use of different performance metrics such as accuracy, precision, recall, AUC, ROC, etc. With these metrics, the better-performing model can be decided.

7. Dissertation – This is the last step which involves documenting all the above phases and the procedure of the project.

Research plan – Gantt chart or Pert chart



1. Research/study:
    a. Literature study – Studying research papers and machine learning models
    b. Data collection – Collecting relevant bank customer data
2. Data processing:
    a. Data cleaning – Handle missing data, errors, and outliers
    b. Feature extraction – Applying EDA, PCA, ICA techniques
3. Model design and testing:
    a. Model development and tuning – Designing the machine learning models and tuning the parameters using grid search method
    b. Validation and testing – Evaluating the models using performance metrics and testing the model on test data for performance improvement
4. Dissertation: Final documentation of the project procedure

Ethical Statement and Data Management Plan

The purpose of this project is to design a machine learning model to predict early bank loan defaulters. The data collected might include human information as the project involves working on bank customer data. However, any personal information regarding a person will not be included such as name, address, birthday, or contact details. Hence, the data collected will in no manner be able to help in tracking purposes. Moreover, the data that is going to be collected is obtained from online public repositories which don't have any legal obligations.

This project will not have any concerns regarding health and safety as this will be implemented on a personal device such as a laptop or computer which may not harm any individual. Working on this project will only involve the consumption of energy during training complex machine learning models. However, it will not cause any impact on environmental aspects.

For the duration of the project, the data collected from public repositories will be stored on a personal computer or laptop for immediate access. Upon completion of the project, any data collected for experimental purposes will be removed from the device.

Ethical aspects

As the project is going to be developed on a personal computer such as a laptop, this will require the consumption of energy. However, the process of training the machine learning models will be computationally inexpensive. Also, the project will be implemented in python and no physical product is being developed. Hence, there is no concern or harm related to the environmental aspect.

Some research papers mention that the data collected might have a bias with respect to gender or age. For example, in the case of gender, for a specific region, the data might have a bias towards females having the majority of defaulters. In such scenarios, the proposed method will maintain transparency, and such variables showing bias will be dropped from the dataset. As the project focuses on individual banking data regardless of gender, the ethical aspect will be maintained. Also, ethical aspects will be maintained in regards to application in a specific region. Though the dataset may contain data pointing to a specific region, the model can be adopted in any region with any regional bias.

Commercial aspects

The proposed ideology is developed keeping in mind the intent of the project which primarily focuses on the banking industry. Assessing and analyzing the individual loan customer requires manpower and time consumption which can be reduced with the help of this project. The banking industry can profit from obtaining the simple to use and easy-to-understand software which helps them save resources in terms of the financial aspects. Replacing the existing traditional approaches can be advantageous to the banking industry.

This will help in modeling the banking industry positively with respect to the growth of financial markets and reduce the efforts of additional manpower and hiring of 3rd party agencies. The growth in the advancement of technology in the financial industry will be significant and improve efficiency by assisting customers in day-to-day processes. In addition, the time management involved will improve customer experience and increase loan approvals with the minimal risk associated.

Although, the proposed methodology will not entirely replace human assistance as this will only be a guide for further approvals. However, many financial institutions adopting machine learning technologies over the past few years have research and studies being carried out to improve the models to obtain better results.

Legal aspects

The procedures adopted in developing this project will be transparent to any legal issues. Data collection involves the use of public repositories which already have already passed legal approvals as they are available for research and learning purposes. However, every institutional organization adopting this technology will have to train the model with appropriate data as different regions around the world have their own rules and regulations.

One other legal aspect that might come into effect is the complete use of AI technology replacing the traditional approach. The advancement of machine learning models does not completely replace human assistance. Human supervision is still required to carry out ethical or legal decisions when selecting an individual for loan disbursement. Automating this technology will for loan disbursement is not advisable. Financial experts can use these models for effective decision-making with consideration of terms and conditions. Finally, it is up to the banking authority to approve or reject a loan.