ME 781 Group 3

Team Members

Ayush Kumar Singh 22B2203

Nihar Mehta 22B0416

Shashwat Prakash 22B0678

Shrihari Wattamwar 19D100021

Siva Kishore Gollapalli 200050042

Meeting:

Meet 1 :
Sun 22 Oct, 23:00pm
1 hours appx

Members present :
Ayush Kumar Singh 22B2203
Nihar Mehta 22B0416
Shashwat Prakash 22B0678
Shrihari Wattamwar 19D100021
Siva Kishore Gollapalli 200050042

~Ayush Singh

Technical background:

- Nihar Mehta

Patents in the field:

- **IN100942A**: System and method for approving loans using machine learning (Flipkart Internet Private Limited, filed in 2017, granted in 2020)
  This patent describes a system for using machine learning to approve loans to customers of Flipkart. The system considers a variety of factors, including the customer's purchase history, payment history, and credit score. The system also considers the customer's social media activity and other online data.
- **IN104967A**: System and method for predicting loan defaults using machine learning and AI (IDFC FIRST Bank Limited, filed in 2019, granted in 2022)
  This patent describes a system for using machine learning and artificial intelligence to predict loan defaults. The system considers a variety of factors, including the customer's credit history, employment history, income, and debt-to-income ratio. The system also considers the customer's social media activity, online data, and other factors.

Publications in the field:

- **Machine Learning Models for Credit Risk Assessment: A Review (Applied Soft Computing)**
  This paper focuses on ML models for credit risk assessment, which is a key part of the loan approval process. Credit risk assessment is the process of evaluating the likelihood that a borrower will default on a loan. The paper reviews the different ML models that have been used for credit risk assessment, such as logistic regression, decision trees, random forests, and support vector machines.
- **A Review of Machine Learning Applications in the Credit Scoring Process (Expert Systems with Applications, 2020)**
  This paper reviews the different ML applications that have been used in the credit scoring process, which is a key part of the loan approval process. Credit scoring is a process of assigning a numerical score to a borrower based on their credit history and other factors. The paper discusses the benefits and drawbacks of using ML for credit scoring, such as the ability of ML models to learn complex patterns in data and the potential for bias in ML models.
- **Machine Learning for Loan Approval: A Case Study of a Large Commercial Bank (Journal of Financial Services Research):**
  This paper presents a case study of how a large commercial bank is using ML to approve loans. The bank's ML model is a random forest model that is trained on historical data of loans, such as the borrower's credit score, employment history, and debt-to-income ratio.
- **Explainable Machine Learning for Loan Approval: A Review of Techniques and Applications (IEEE Transactions on Knowledge):**
  This paper reviews different explainable ML techniques that can be used for loan approval. This is important for loan approval because borrowers need to understand why they were approved or denied for a loan. Some explainable ML techniques that are discussed in the paper include LIME, SHAP, and TreeExplainer.

Technical Resources

Open Source Libraries

- **Open CV:** Tools and libraries for the Biometric Authentication.
- **Ethereum:** Blockchain platform for the SDKs and the documentation
- **TensorFlow and Pytorch:** Deep Learning and the training the model
- **Scikit-Learn, Matplolib, Seaborn:** For ML tasks
- **CryptoGraphy:** For the security and the data encryption

Propietary Libraries:

- **Smart Contract** by Quantstamp or OpenZeppelin
- **Fabric** by the hyperledger
- **Solidity** by the Ethereum
- **AWS** blockchain (in case azure ... if needed)

Functions Applied

- **Classify function for training**
- **Logistic Regression**
- **DecisionTreeClassifier, RandomForestClassifier**
- **GradientBoostingClassifier**
- **Security and Encryption, Data Integration**

Source Code Resources

- **https://github.com/hyperledger/fabric**
- **https://github.com/ethereum/solidity**
- **https://ethereum.org/en/developers/**
- **And for the ML model functions we just googled**

~Ayush Singh

DATA

-Shashwat Prakash

Ai/ML model training data

- We plan to use the Loan Predicition dataset that belongs to Univ.AI . This dataset was used for a hackathon organised by Univ.AI. It is a renowned resource and is used for training many Loan Prediction ML models.
- Since the dataset is huge, and already divided into Training and Test datasets, we will not be using multiple datasets . However, if time permits we may use the Home Loan Prediction dataset of kaggle.

Size of data, column details and data type

- The dataset has 2.5 lakh entries , and 13 columns namely id, income, marital status, Job experience, age, House ownership, car ownership, profession, city , state, years in current job and years of current house.
- The dataset is divided into two files - training and test data. Training dataset has 2.5 lakh entries and test dataset has 28k entres.

Authenticity, error , bias and missing data

- Univ.AI data is generally considered authentic as it contains diverse data curated using survey.
- Error levels in the dataset can be minimal due to stringent data collection and verification, but some errors may exist.
- Missing data, if any, is generally low, as the dataset is well-structured and curated with a comprehensive set of face images

AI/ML Model Scheme

~Siva Kishore Gollapalli
~Shrihari Wattamwar

1. Which AI/ML models will be chosen?

- We'll use a Multi-Model approach. Considering Binary classification problem Logistic Regression will be first model.
- In order to make a decision on whether to provide a loan, we will also use Decision Tree and Random Forest.
- Our data is fairly imbalanced and has missing values as well and so we thought to use Gradient Boosting.

2. How would the training, validation and testing be done?

- We're splitting the dataset in 4:1 ratio for the training and testing purpose i.e. 80% for training & 20% for testing.
- After fitting the model we are using cross validation with 10 folds (cv=10) to evaluate the model's performance
- We are also incorporating the importance to the particular feature of the dataset and sorting the features by their importance scores in ascending order, so the least important features are at the beginning and thus, features with higher relative importance are more influential in the model's decision-making process.

3. Is your model too rigid or too flexible?

- To prevent much flexibility, we would be using a diverse dataset and using cross-validation with more folds to discard the possibility of overfitting of the model and reduce its sensitivity to noisy data.
- To prevent much rigidity, we would be testing our models on Random Forests, Decision Trees, and Logistic Regression and then choosing the one that gives us the least error which would therefore increase the flexibility.
- Overall our model ensures the flxibility in the feature selection and the giving importance index but at the same time it also ensures the rigid and precise decision of whether to provide loan or not.