Ailin Chu

Weihao He

Xiao Qin

MSAI 349-Final Project Proposal

24 October 2024

Final Project Proposal

Q1:
The task we will address is predicting whether an applicant is approved for a loan based on various features such as income, credit history, and employment details. This task is interesting because it reflects a real-world financial decision-making problem, helping participants develop skills in classification models while understanding factors that influence loan approval. It also highlights the practical impact of machine learning in automating and improving lending processes and can serve as a topic of discussion during our job hunting interviews for machine learning engineer positions.

Q2:
For our project, we will acquire datasets from reliable public sources. After exploring potential sources, we have identified the following datasets that align with our project needs:

1. **Kaggle - Loan Prediction Problem Dataset**
    - **URL**: Loan Prediction DatasetLoan-Approval-Prediction-Dataset
    - **Description**: This dataset includes loan application information and whether a loan was approved or denied. It contains valuable attributes such as applicant income, credit history, and loan amount, which are useful for our analysis.
2. **UC Irvine Machine Learning Repository - Credit Approval Dataset**
    - **URL**: Credit Approval Dataset
    - **Description**: This dataset contains anonymized information about credit card applications and their outcomes. However, due to the anonymization of feature names, it may be challenging to interpret and utilize effectively. As a result, we may rely more on other datasets unless specific features become clear during preprocessing. We can consider training a model using this dataset separately and applying an ensemble method during the prediction phase.
3. **Kaggle - Home Credit Default Risk**
    - **URL**: Home Credit Default Risk
    - **Description**: This competition dataset focuses on predicting clients' ability to repay loans, providing detailed information about customers' financial and credit histories.
4. **GitHub Datasets**
    - After review, we found that most GitHub repositories mirror data from Kaggle. So, we will not rely on it.
5. **Government Data**
    - We searched through World Bank Open Data and ConsumerFinance.gov but did not find any datasets that meet our requirements.

Given our analysis, we plan to primarily use the **Kaggle Loan Prediction Dataset** and **Home Credit Default Risk Dataset** due to their rich feature sets and accessibility. Both datasets are well-documented, publicly available, and suitable for predictive modeling tasks. If needed, we can use the **UC Irvine Credit Approval Dataset** as a supplementary source, though its anonymized feature names may limit its usability.

With access to these datasets, our project is **feasible** without needing to collect new data.

Q3.

For our loan approval prediction task, we will use the following key features from the datasets:

1. **Applicant's Income:** This provides an indication of the applicant's ability to repay the loan. Higher income often correlates with a higher likelihood of loan approval.
2. **Co-applicant's Income:** For joint loans, this gives additional information about the financial stability of the applicant.
3. **Loan Amount:** The size of the loan requested, which is a critical factor in the approval decision.
4. **Loan Amount Term:** The duration over which the loan is to be repaid. Longer terms may affect the risk and approval probability.
5. **Credit History:** An important feature reflecting the applicant's past borrowing and repayment behavior. A good credit history is a strong indicator for loan approval.
6. **Employment Details:** Employment status, such as whether the applicant is self-employed or employed by an organization. This may impact the stability of the application's income and the approval decision.
7. **Marital Status:** Marital status may correlate with financial stability, which can impact loan approval.
8. **Education Level:** Higher education levels may be associated with better job prospects and income, influencing loan approval chances.
9. **Property Area:** Whether the applicant lives in a rural, urban, or semi-urban area, which may influence loan approval depending on regional risk profiles.
10. **Dependents:** The number of dependents supported by the applicant. More dependents may indicate a higher financial burden, which could impact loan approval.

We will also consider using additional features related to loan purpose, savings, and other demographic attributes if available and relevant to the task.

These features are selected based on their relevance to financial decision-making and their availability in the chosen datasets. We will perform feature selection based on correlation and model interpretability during the analysis phase.

Q4.

(a) **Data preprocessing**:

We will **handle missing data** by filling missing values with the median for numerical columns and possibly the most frequent for categorical columns if necessary. Additionally, we'll apply **normalization** using MinMaxScaler to scale numerical features within the range of [0,1]. This improves LightGBM's performance since the model benefits from normalized features, especially when there are wide ranges in the data.

Moreover, categorical variables will be encoded, either using **label encoding or one-hot encoding**, but LightGBM also natively handles categorical variables through built-in support. This ensures that the machine learning model can effectively process both types of data.

(b) **Data visualization**

We'll create a set of visualizations to help understand the relationships between important features. This could include:

1. **Histograms** to examine the distribution of key numerical features.
2. **Correlation Heatmaps** to visualize relationships between numerical variables and the target (loan status).
3. **Boxplots** to explore the spread and outliers for features like loan amount, income, and credit history.

These visualizations will give us insight into the important drivers of loan approval, such as credit history or income, and whether any transformations might be useful for skewed features.

(c) **Machine learning model**

We will use **LightGBM** as our primary model for classification. LightGBM is a powerful boosting algorithm that builds multiple decision trees in a sequence, where each new tree tries to correct the errors of the previous ones. It also supports early stopping to prevent overfitting. We will fine-tune hyperparameters such as: learning_rate，n_estimators，max_depth and num_leaves.

Additionally, **L1 and L2 regularization** will be applied to reduce overfitting by penalizing overly complex models.

(d) **Evaluation the outcome**

The model's performance will be evaluated using the Area Under the Receiver Operating Characteristic Curve (**AUC-ROC**), which is especially useful for binary classification problems like loan approval. AUC-ROC helps measure the model's ability to distinguish between the positive (loan approved) and negative (loan not approved) classes. **K-fold cross-validation** will also be used to ensure the model generalizes well to unseen data and doesn't overfit to the training set.