Project Summary

In this credit risk modeling project, the goal is to develop a predictive model that assesses the risk associated with lending money to individuals or businesses. The project involves using historical loan data to train a model that can predict whether a new loan applicant is likely to default on their loan payments or not.

Data Preprocessing

1. **Loading Data:** The project begins with loading the loan data from a CSV file into a Pandas DataFrame. The data includes various features such as loan amount, interest rate, employment length, and loan status.
2. **Data Exploration and Cleaning:** Initial exploration of the dataset includes checking for missing values, understanding the distribution of variables, and handling missing data by either imputing mean values or filling with zeros.
3. **Feature Engineering:** Several new features are created from existing ones to extract more information, such as converting categorical variables into dummy variables and transforming dates into meaningful durations (e.g., months since the earliest credit line).

Probability of Default (PD) Model Data Preparation

1. **Preprocessing Discrete Variables:** Discrete variables like loan grade, home ownership, and verification status are processed using Weight of Evidence (WoE) encoding. This involves grouping similar categories, calculating WoE and Information Value (IV) to assess their predictive power, and then creating dummy variables.
2. **Preprocessing Continuous Variables:** Continuous variables like interest rate, employment length, and credit history are binned into intervals. WoE and IV are calculated for each interval to transform these variables into categorical ones.
3. **Splitting Data:** The dataset is split into training and test sets using the **train_test_split** function from Scikit-learn. This ensures that the model's performance can be evaluated on unseen data.

Model Development (Not Included)

While the notebook covers extensive data preparation steps, it does not include the actual model development phase. In a typical credit risk modeling project, after preprocessing the data, one would proceed to select a suitable modeling technique (e.g., logistic regression, decision trees, or ensemble methods), train the model on the training data, and evaluate its performance on the test data. This will be done in the part II of the project which will also be posted in this repository when completed.

Conclusion

This notebook demonstrates the crucial steps involved in preparing data for credit risk modeling, including preprocessing discrete and continuous variables, handling missing values, and splitting

the data for model training and evaluation. Further steps would involve model development, validation, and deployment for real-world application.