

Project Report – Tural Azimov

The Personalized Loan Pricing Prediction project aims to develop a machine learning model that accurately predicts loan pricing based on customer financial and demographic information. By leveraging historical loan data and various factors, the project provides a tool for estimating personalized loan pricing. Through data analysis, model development, evaluation, optimization, deployment, and documentation, the project enables financial institutions to make informed decisions and improve risk management strategies.

The project utilizes the "Loan Prediction Problem Dataset" from Kaggle that was a sample dataset for this task. This dataset contains customer profiles, loan details, and pricing information. It offers a comprehensive and diverse collection of data, including credit history, income level, employment status, loan amount, and duration. This dataset serves as the foundation for training and evaluating the loan pricing prediction model.

After collecting the "Loan Prediction Problem Dataset," the next step is data preprocessing. This involves cleaning the data, handling missing values, outliers, and inconsistencies. By applying appropriate techniques, we ensure the data's quality and integrity. Exploratory data analysis (EDA) is performed to gain insights into the dataset, identify patterns, and understand the relationship between variables. This step sets the stage for feature engineering and model development.

In the feature engineering phase, we extract and create relevant features from the preprocessed dataset that are likely to impact loan pricing. This involves transforming and combining existing variables to provide more meaningful and predictive information. We also perform feature selection using chi2 function in order to get more relevant features.

For model development first we select base model and for this model there were many options such as logistic regression, svm ,K-nn but random forest classifier was chosen because it had relatively higher accuracy. Then we used standart scaler from sklearn preprocessing library to improve this model. After that to further optimize our model we fine-tuned the model by optimizing hyperparameters through grid search. Then we preformed model evaluation on our optimized model with cross validation and calculated addditional evaluational values such as mean absolute error and root mean square error.

At the end we saved our improved model to a file so that we can create an interface that allow stakeholders to input customer information and get a prediction whether customer is eligible for loan or not. In order to do this Streamlit wich is an open source app framework.

To further improve this project we can expand our dataset and add even more data that contains distinct values. We also can further improve our model to get better accuracy. For example we can apply Bayesian optimization. And also we can improve our interface to make it even more user- friendly.

In conclusion in this project we tried to build a model that predicts whether a customer should get a loan or not and we achieved 79.7 % accuracy. And we created user friendly interface for ease of use. Also there are some space for improvements if they are needed.