CREDIT RISK CLASSIFICATION USING ML

VENKATA SAI KRISHNA ABBARAJU

UNIVERSITY OF MINNESOTA
Driven to Discover℠

INTRODUCTION

- Credit risk classification is a crucial task for financial institutions to assess the likelihood of loan default.
- Accurate credit risk assessment enables lenders to make informed decisions and mitigate potential losses.
- The objective of this presentation is to evaluate the performance of machine learning models in classifying credit risk based on the provided dataset.

- Research question: “Given a new customer, what is the risk associated if they were to be given a loan?” aka "How well does ML models perform in predicting credit risk on unseen data?"

Overview of the dataset

- The dataset contains information about loan applicants, including demographic, financial, and credit history attributes.
- The dataset consists of 50k samples and 88 features, with 4 classes representing different levels of credit risk.

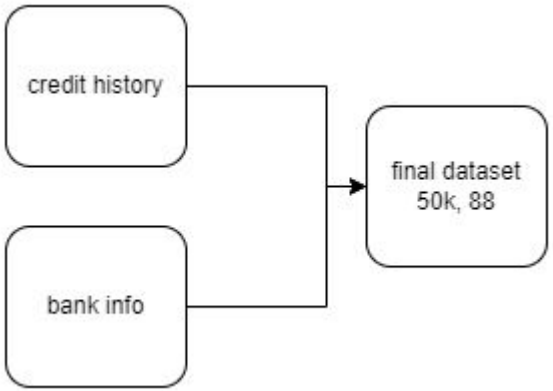


Fig 1. Dataset construction

Credit History - score, Home loan, Age of oldest loan, tot active loans, closed_12M

Bank info - Age, Marital status, time since last default, default_12M, enquiries_6M

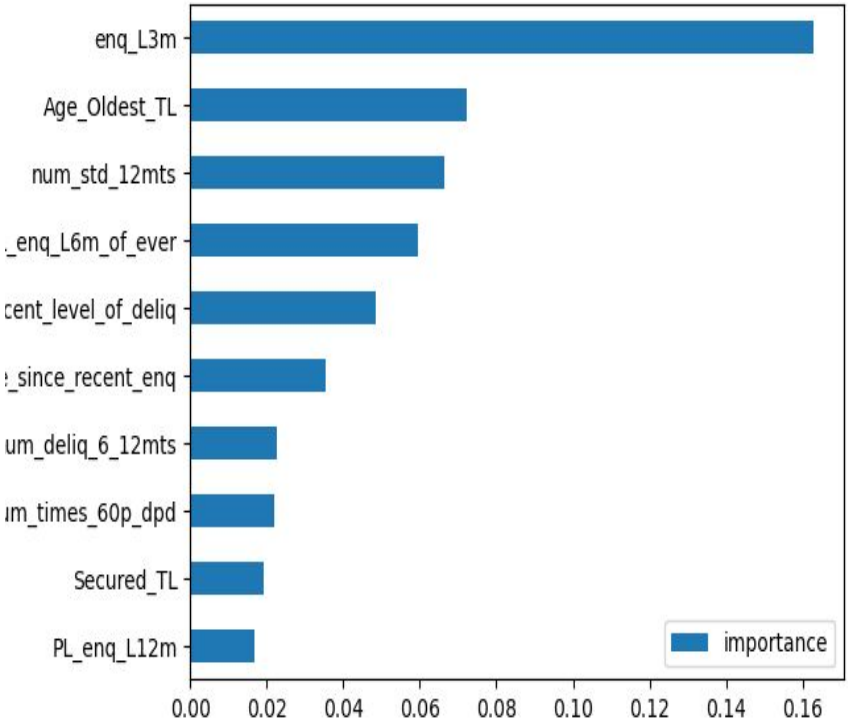
Methodology

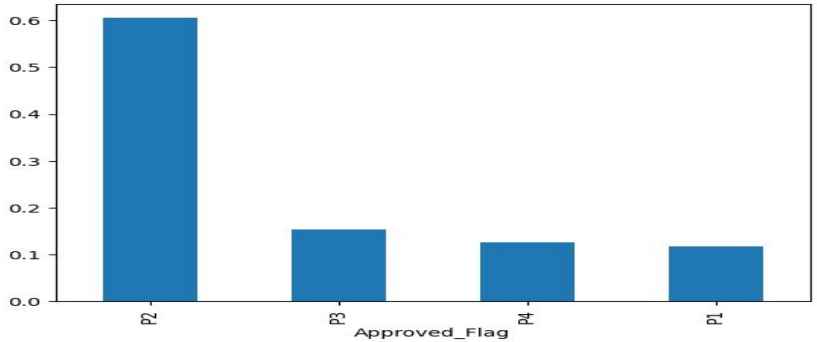
- Data preprocessing
  1. Handled missing values by dropping columns if >= 20% missing values (dropped 4 features)
  2. After dropping null rows, able to retain more than 80% of the data. So finally the dataset is 40k rows and 84 columns
- Feature selection
  1. Applied Chi-squared test on categorical features and target.
  2. Dropped numerical columns that showed multicollinearity (vif > 5). Then applied ANOVA on numerical columns and target.
  3. Finally, there were 34 features

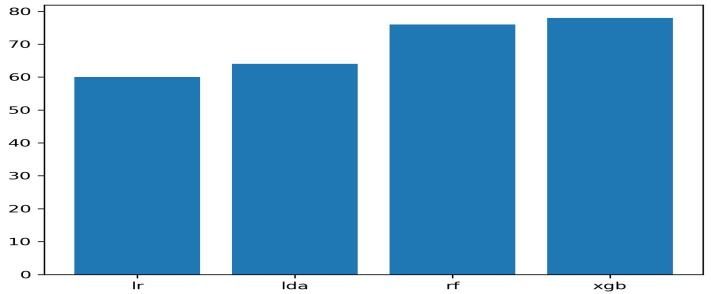
METHODOLOGY

- Feature engineering
  1. Label encoding - Education
  2. One hot encoding - Gender
- Model building using CV
  3. Used 20% as unseen data and performed 5-CV on 80%
  4. Tried out Logistic Regression, LDA, Random Forest, and XGBoost
  5. Used cross-entropy as the loss function and F1 as the scoring measure in CV
  6. Metrics reported F1, Accuracy, Precision, Recall

Feature Importance

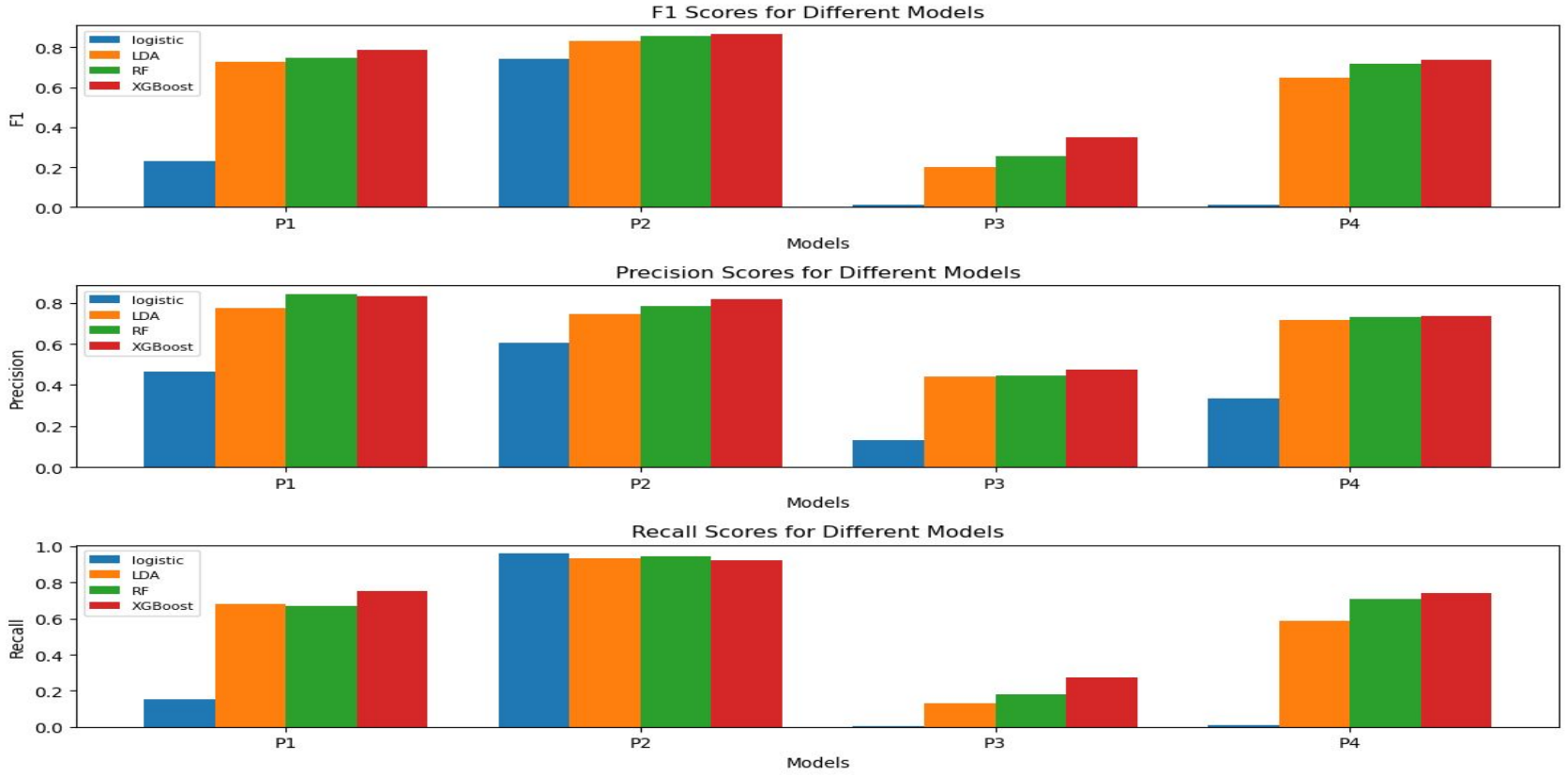






Test accuracies

Results on Unseen set



Interpretation of results

- A lower recall indicates that a lot of high risk applicants were classified as low risk, which means the bank might lend out loan to a customer who is more likely to default.
- A lower precision tells us that a high risk applicant is actually not a high risk applicant. This means that, the bank might not lend out loan to a customer who might not default.
- Observed that, P3 has lower recall and precision than other classes. A lot of P3 were classified as P2. This could be because of abundance of P2 or due to probability thresholds.

Next Steps

1) Try out assigning class weights
2) SMOTE
3) Adjusting class thresholds
4) One vs Rest

Learnings

1) Domain knowledge of credit lending
2) Applying Ensemble models
3) Interpreting results
4) Figuring out what to do next

THANK YOU