India ML Hiring Hackathon 2019

1. Problem Statement

Given the information like mortgage details, borrowers related details and payment details, our objective is to identify the delinquency status of loans for the next month given the delinquency status for the previous 12 months (in number of months). Delinquency occurs when a borrower misses a payment against his/her loan.

2. Data Dictionary

Feature name	Description
loan_id	Unique loan ID
source	Loan origination channel
financial_institution	Name of the bank
interest_rate	Loan interest rate
unpaid_principal_bal	Loan unpaid principal balance (described as initial loan amount)
loan_term	Loan term (in days)
origination_date	Loan origination date (YYYY-MM-DD)
loan_to_value	Ratio of loan amount to asset value against which loan was granted
number_of_borrowers	Number of borrowers
debt_to_income_ratio	Debt-to-income ratio (debt includes all loans borrower has taken)
borrower_credit_score	Borrower credit score
loan_purpose	Loan purpose
insurance_percent	Loan Amount percent covered by insurance
co-borrower_credit_score	Co-borrower credit score
insurance_type	0 - Premium paid by borrower, 1 - Premium paid by Lender
m1 to m12	Month-wise loan performance (delinquency in months)
m13 (target)	**loan delinquency status (0 = non delinquent, 1 = delinquent)**

3. Evaluation Metric

- **F1-Score** between the predicted and observed class.

4. Data Exploration

★ 2 Target Classes - 0 and 1
★ **Imbalanced Dataset**- Ratio of 0:1 is **99.45 : 0.55**
★ No duplicates
★ No missing data
★ 'loan_id' was dropped since it can not be used for making predictions.
★ 5 categorical features
(source,financial_institution,origination_date,first_payment_date,loan_purpose)
★ 23 Numerical features
★ Label Encoding on categorical features was performed since it gave better F1 score on the public leaderboard compared to one-hot encoding.

5. Feature Engineering

Several new features were created using the existing ones.

- 'origination_date' had 3 unique dates (2012-01-01, 2012-02-01, 2012-03-01), so a new feature '**origination_month**' was created using this and 'origination_date' was dropped.
- Similarly 'first_payment_date' had 4 unique dates (02/2012, 03/2012, 04/2012, 05/2012), so a new feature '**firstpayment_month'** was created and 'first_payment_date' was dropped.
- A new feature '**orig_pay_gap'** was created which equaled to the difference between the number of months of 'firstpayment_month' and 'origination_month'.
- 'number_of_borrowers' and 'insurance_type' were cast into integer type from float type in order to save memory.
- A new feature '**total_credit_score**' was made using the sum of 'borrower_credit_score' and 'co-borrower_credit_score'.
- A new feature '**loan_months**' that shows loan term in number of months was created using 'loan_term'.

6. Model Building and Selection

- To compare the performance of different models **5-Fold Cross-Validation** Scheme was used which also at the same time ensured that there was no underfitting or overfitting in the model.

- Different models ranging from simple techniques like Logistic Regression, KNN, SVM, Decision Trees to ensemble techniques like Random Forest, AdaBoost, BaggingClassifier, LightGBM, Gradient Boosting, CatBoost, XGBoost were used.

- Amongst all models, **XGBoost** gave the best performance, so it was chosen to further improve the results.

7. Feature Selection and Hyper-Parameter Tuning

- Recursive Feature Elimination (RFE) along with 5- Fold Cross-Validation was done to obtain the optimal set of features.

- **9 features** were obtained after using RFE, which were-
  (m1, m4, m5, m8, m9, m10, m11, m12, borrower_credit_score)

- After selecting the best features, Hyper-Parameter tuning was done using GridSearchCV to obtain the best hyperparameters for the XGBoost model.

- Finally, the predictions were made and these were the results.
  - ★ **Public LeaderBoard Score- 0.316326530612245**
  - ★ **Public LeaderBoard Rank- 396 out of 3740 participants**
  - ★ **Private LeaderBoard score- 0.566037735849057**

- Although RFE improved the Public LeaderBoard score but submissions involving all the features had a better F1 score on the Private LeaderBoard.