Credit Risk Assessment

Areas to be covered

- Understanding what is credit risk?
- How to define label using Roll Rate Analysis
- Understanding Data
    - Features
    - EDA
    - New Feature Development
- Model Development
    - Categorical Feature Treatment (Target Encoding)
    - Feature Selection
    - Model Setting
    - Lightgbm Implementation
    - Hyperparameter Tuning using Hyperopt
- Model Selection & Evaluation
    - ROC AUC, PR AUC
    - Score Distribution
    - Class Rate Curve
- Feature Importance
    - Split & Gain
    - SHAP
- How to choose right cutoff according to business target?

Structure

- Data - *credit_risk_data.csv*
- Utility Function and Classes - *utils.py*
- Relevant Libraries and Version - *requirements.txt*
- Model - *model.ipynb*
- Project Document - *project_document.pdf*

Problem Definition

Credit - A sum of money/equivalent any financial institution, business or individual lends to the borrower (like Cash Loan from Bank, Product purchase on EMI, Credit Card etc.)
Risk - Condition of not receiving money back

Credit Risk is the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations

In Mathematical terms Probability of Default $-$ $P(default)$

Structure of Credit

- A Fixed amount lended at $T = 0$ at some interest rate I
- Customer has to payback the total Principal + Interest in 6 equal monthly instalments
- Customer has to pay instalment by a particular date of month, failure to do so will tag loan as X dpd (Days past dues)

Label Definition

Now the question is what do we call default?

Default = Customer achieve dpd X in First Y Months
And we have to define X and Y using Roll Rate Analysis

Following things take in consideration while defining X & Y

- Customer is less likely to pay after achieving dpd X
- It is less likely that customer will default after paying Y installments

DPD Movement

DPD Attained	Customers	% Customers
0	86,250	100.00%
30	22,073	25.59%
60	8,099	9.39%
90+	7,926	9.19%

From here we can see % of people attained dpd 60 are going beyond 90+

So answer to Right X = dpd 60
Now we analyze dpd60 customers, in which emi they are reaching dpd60

Month Movement

EMI Number	Customers	% Customers
1	6,546	80.82%
2	1,084	13.38%
3	421	5.20%
4	22	0.27%
5	18	0.22%
6	8	0.10%

As we can see from this table that most of the defaulters have defaulted in first 3 EMIs, so it is very likely that people who have paid first 3 EMIs will pay next 3 EMIs

Label = dpd60 in First 3 EMIs

Understanding Feature Dimensions

Financial Institutions tries to judge their customers according to

- Past Credits (sometimes called Credit Bureau variables)
- Customers Employment (Work Experience, Employment Type, Company types, Income etc.)
- Customer Demographics (like Marital Status, House is rented or owned, Age etc.)

So our features revolves around these dimensions

Understanding Data

Variables - Numeric, Categorical Variables
Univariate - % Null, 0 Variance, General Stats
Label - Label Distribution

Model Setup

Features (X) - Customer Related, Past Credit Behaviour
Label (Y) - DPD 60 in First 3 Repayments
Loss Function - LogLoss
Algorithm - Boosting
Implementation - Lightgbm
Feature Selection - Random Forest, Decision Tree
Categorical Features Treatment - Target Encoding
Null Value Treatment - Auto treatment by Lightgbm
Hypeparameter Tuning - Hyperopt
Primary Evalution Metric - ROC AUC
Other Evalution Metrics - PR AUC, Class Rate Curve, Score Distribution etc.
Feature Importance - Split & Gain on Training Data, SHAP Importance

$$LogLoss = \frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

Target Encoding

Library - https://contrib.scikit-learn.org/category_encoders/

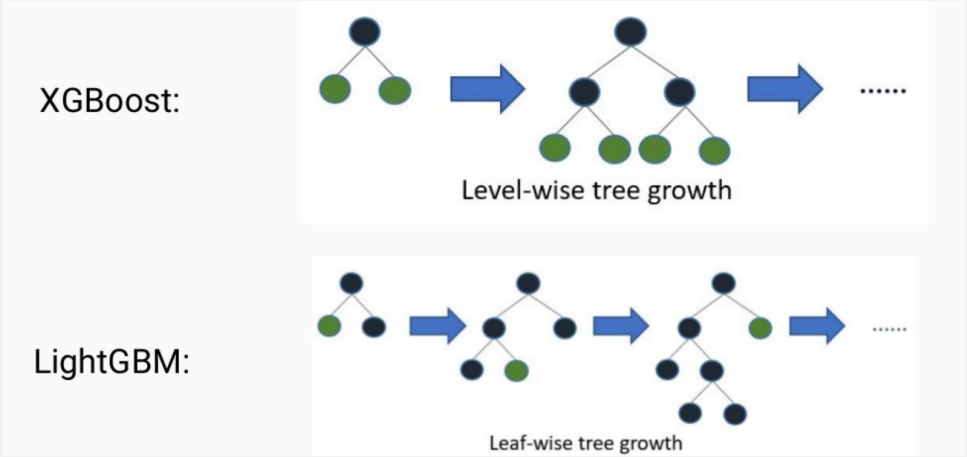
Target Encoding

workclass	target
State-gov	0
Self-emp-not-inc	1
Private	0
Private	0
Private	1

workclass	target mean
State-gov	0
Self-emp-not-inc	1
Private	1/3

workclass
0
1
1/3
1/3
1/3

Lightgbm



Other Properties of Lightgbm

- It is relatively faster than Xgboost
- Lightgbm can handle Categorical and Null Values

Paper - https://www.microsoft.com/en-us/research/wp-content/uploads/2017/11/lightgbm.pdf

Library - https://lightgbm.readthedocs.io/en/v3.3.2/

Hyperopt

Hyperopt is a Python library for Sequential Model-Based Optimization (SMBO) Hyperopt uses results of previous runs to find best suited hyperparameters for next iteration

Paper - https://conference.scipy.org/proceedings/scipy2013/pdfs/bergstra_hyperopt.pdf
Library - http://hyperopt.github.io/hyperopt/

SHAP

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.

Paper - https://github.com/slundberg/shap#citations
Library - https://shap.readthedocs.io/en/latest/index.html

Right Score Threshold

Step 1 - Fix Cummulative NPA cutoff
Cummulative NPA = % of defaulter we are ok with like if i lend to 100 people i am ok to bear loss of 2 people that means Cummulative NPA = 2%

Step 2 - Sort validation users score in ascending manner and create X equivolume binns (X depends on how large is validation data)

Step 3 - Start from bucket 1 and calculate NPA above that bucket
No. of Defaulter in all bucket above that / Total number of users in that bucket

Step 4 - Point where your calculated NPA touches or come very close the decided cumulative NPA that is your threshold point
Highest Score of that bucket is the threshold point

For further refining, further segregation of threshold bucket can be done repeat process to find more granular threshold