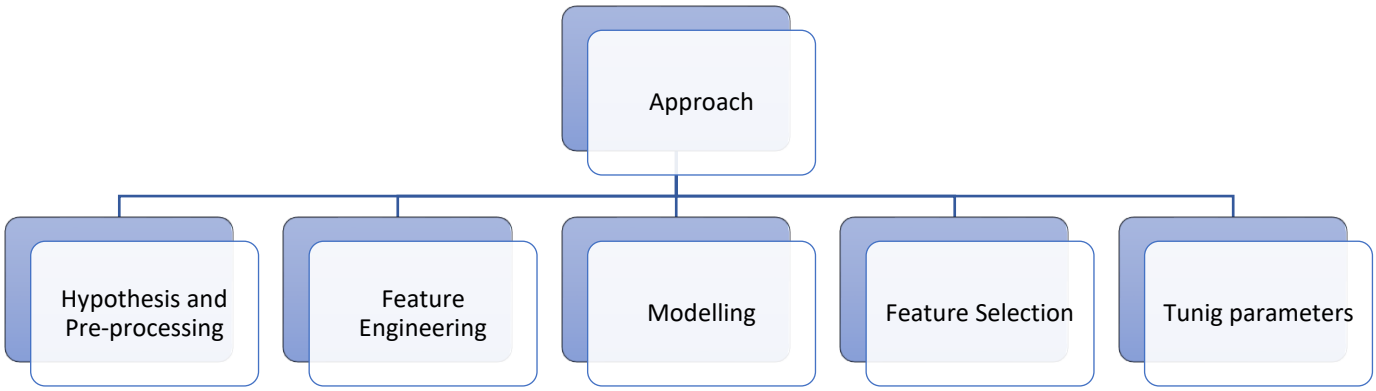
EPSILION 2.0

Name	College	Contact No.	Email ID
Shankar Lal	IIT Madras	7822025458	ershankariitian@gmail.com

Problem statement: - The task is to estimate the likelihood of the bank loan default on account of a borrower based on multiple factors like Loan term, Credit Score, Gender, Type of loan, Employment status, and other such characteristics metrics.

Data Information: - Training data has 804 samples and test data has 203 samples. There are 16 independent features and 1 target binary variable (**Default**).

Target Variable Distribution: - Out of 804 samples 568 samples (**71%**) are non-default and 235 samples (**29%**) are default which is clearly a skewed distribution of target variable (**Imbalanced dataset**).



Hypothesis and Pre-processing: -

- It is imbalanced dataset with 71% non-fraudulent and 21% fraudulent samples
- Train and test dataset have come from same population distribution
- One sample in training data set had missing value in Default variable. So we removed this samples.
- Missing values has varied distribution among all the variables. So I imputed the values based on the similarity samples of the missing values.

For example, take one sample which has ID (1183), has missing values in Checking_amount, Home_loan, Education_loan, and Emp_duration. To fill the value the values in these columns, I took other columns information for this particular customer and search for similar customers and filled the mode value in categorical features and mean value in numerical feature.

- Overall 59% Fraud claims are by male customer where 41% in female customers.
- 59% Fraud customers are single and 41% are married
- Customers who don't have any personal loan, are 69% defaulters and only 31% are non-defaulters who have personal loan
- Customers who don't have any home loan more prone to make frauds, which are nearly 98% from total frauds
- For term period 22 and higher have more percentage to make frauds.
- For term period less than 22 have more percentage to non-fraudulent customer. Even for customers who have took loan for term period 9, are 100% non-fraudulent

- Customers who are between 18 and 24 age, are 100% fraud case whereas customer between 35 to 42 age, are 100% non-fraudulent
- Customers who have credit score less than 593 are 100% fraudulent customer and greater than 811 credit score customers are more non-fraudulent.
- Binary encoding used in the all the categorical variable

Feature Engineering: -

- Since there were some customers which have negative checking balance. So, I have created a variable called **bal_status** which is a binary variable (1 for negative checking balance otherwise 0)
- Created a variable **total_bal** which is total balance of customer (total_bal = checking_balance+saving_balance)
- Most important factor which most of the banking use is **the loan value ratio.** I have created two variables on this factor which are as follows
    - **Loan_to_tot_bal = Amount/total_bal**
    - **Loan_to_saving = Amount/saving_amount**
    - **Loan_to_score = Amount/credit_score**
- There multiple loans that a customer can take. For example , Education loan, personal loan, car loan, home loan etc. I created a variable which is **total_loan**, is total number of loan that customer has taken
- Each customer has different no of credit account. So to find balance in each account and loan amount in each account, I have created two variables which are as follows:
    - **Tot_bal_per_acc = total_bal/no_of_credit_acc**
    - **Loan_per_acc = Amount/no_of_credit_acc**
- Loan taken by customer has some specific term period (Assumed term period is given month). I have created a variable called **loan_per_term** which is basically the loan amount in each month (Amount/term)
- Bins variables-
    - **Credit_score_bins =** divided the credit scores in three levels (**L- Low, M-Medium, H-High**)
    - **Age_bins =** divide the age in three levels (**A-Adult, Y-Young, S-Senior**)
- Some of the customers are employed and some are unemployed. I have created a variable called **term_diff_emp_duration** which is difference between loan term period and customer's employment duration
- **Tot_bal_diff_loan_bal** is difference between total balance in customer's account and the amount of loan taken by customer
- I also created some aggregates features, by grouping the Age variable-
    - **Age_count –** number of customers in particular age
    - **Age_tot_loan –** total loan taken by customer in that particular age
    - **Age_mean_loan –** average loan taken by customer in that particular age
    - **Age_tot_saving_bal –** total saving balance in that particular age
    - **Age_mean_saving_bal –** average saving balance in that particular age group
    - **Age_flag_mean –** average number of defaults in this particular age group
- Reconstruction error by using autoencoder is calculated and used as an independent variables

Total number of variables after feature engineering were **32** (excluding ID and Default variables).

Modelling: - I have used several classifier models as baseline model with 5 fold cross validation. These classifier models are as follows Logistic Regression, Random Forest, Lightgbm classifier, XGBoost classifier, CatBoost Classifier. The score from baseline models are tabulated below:

Models	Accuracy	F1 Score (weighted)	Precision Score	Recall Score	AUC Score
Logistic Regression	0.8905	0.8899	0.89087	0.89055	0.98809

Random Forest	0.92106	0.92039	0.9229	0.92106	0.97421
LGBM Classifier	0.9252	0.9247	0.9268	0.9252	0.9807
XGB Classifier	0.9321	0.93169	0.93312	0.9321	0.9811
Catboost Classifier	0.9349	0.93409	0.9356	0.9349	0.9795

From above table is clearly seen that XGBoost Classifier perform better compare to other models in all evaluation metric. But I selected three models final modelling which LGBM, XGBoost, Catboost classifier.

After feature engineering using all the 32 independent features, the evaluation score from LGBM Classifier and XGB Classifier are as follows:

Model	Accuracy	F1 Score (weighted)	Precision Score	Recall Score	AUC Score
LGBM Classifier	0.9252	0.9247	0.9268	0.9252	0.9806
XGB Classifier	0.9321	0.9317	0.9331	0.9321	0.9811
Catboost Classifier	0.934961	0.93401	0.9356	0.9349	0.9795

Feature Selection: - Every machine learning performs better when we input best and relevant features to the models. For that purpose I have tried feature selection through four steps which are as follows:

1) **Raw feature importance: -** In this I have calculated gini score each feature and then sort the feature based on their gini score. Selected features which have gini score more than 0.3.
2) **LGBM Feature importance: -** Created a simple lgbm model with all 32 independent features and find out the feature importance score for each feature and sorted them in descending order and stored.
3) **Forward Selection LGBM Wrapper:** - LGBM classifier model is used for forward feature selection. Started with top 5 features from baseline lgbm model and calculated the baseline score. After adding one more feature to this baseline features if baseline score increases then the feature will be added otherwise it is not. Like that this process went up to all 32 features. Total 14 features selected out of 32 features.
4) **Backward Selection LGBM Wrapper: -** In this also LGBM Classifier model is used as baseline. It started with all the 32 features and started removing one by one with all permutation and score calculated. If the score after elimination feature increases from baseline model then that features is kept. Total 18 features were selected through backward selection from 32 features.

After feature selection two models were created on two different subset of features, one from forward selection and another from backward selection. The scores from both the process are as follows:

a) **From Forward Selection LGBM Wrapper and construction error features-** total 20 features

Model	Accuracy	F1 Score (weighted)	Precision Score	Recall Score	AUC Score
LGBM Classifier	0.936506	0.93613	0.93698	0.936056	0.98354
XGB Classifier	0.93652	0.93633	0.93689	0.936514	0.98033
Catboost Classifier	0.93272	0.932101	0.93317	0.932725	0.97951

b) **From Backword Selection LGBM Wrapper features: -**total 18 features

Model	Accuracy	F1 Score (weighted)	Precision Score	Recall Score	AUC Score
LGBM Classifier	0.919068	0.91875	0.91920	0.91906	0.9769
XGB Classifier	0.92775	0.92745	0.92778	0.92775	0.9758
Catboost Classifier	0.92022	0.91976	0.92075	0.92028	0.9751

The final prediction was mode from all three models in two subset features, It means mode of total six predictions.

Tuning Parameters: - I have used Bayesian Optimization to tune some parameters like min_data_in_leaf, bagging_fraction, feature_fraction, num_leaves, lambda_l2.

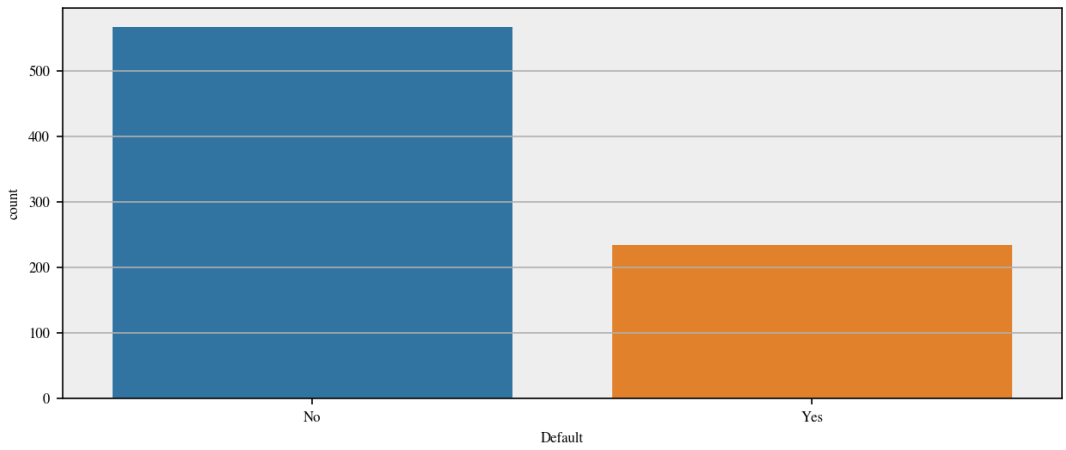
The optimal parameters which were used for final model is as below:

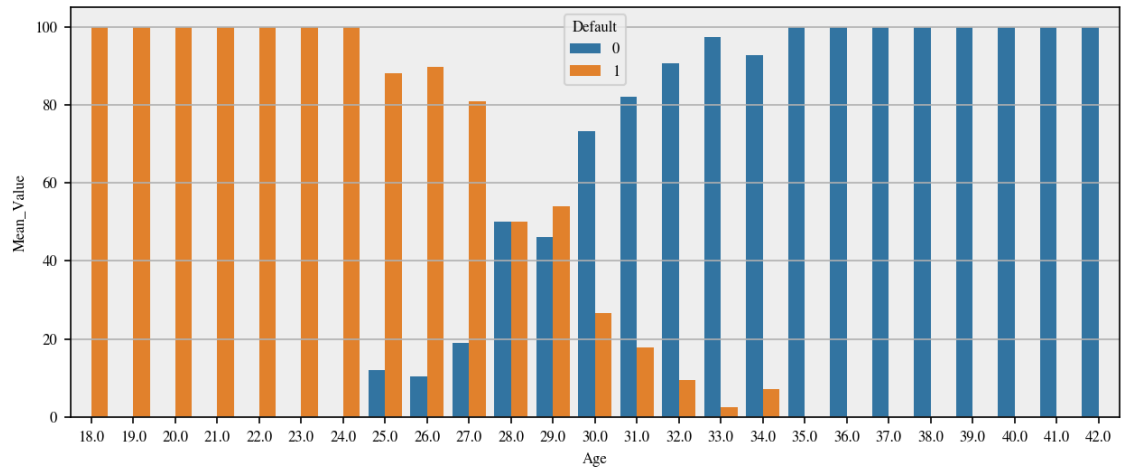
N_estimators	Learning_rate	Colsample_bytree	Scale_pos_weight	Bagging_fraction	Feature_fraction	Lambda_l2	Min_data_in_leaf	Num_leaves
1000	0.03	0.2	1	0.7265	0.5369	0.05784	5	4

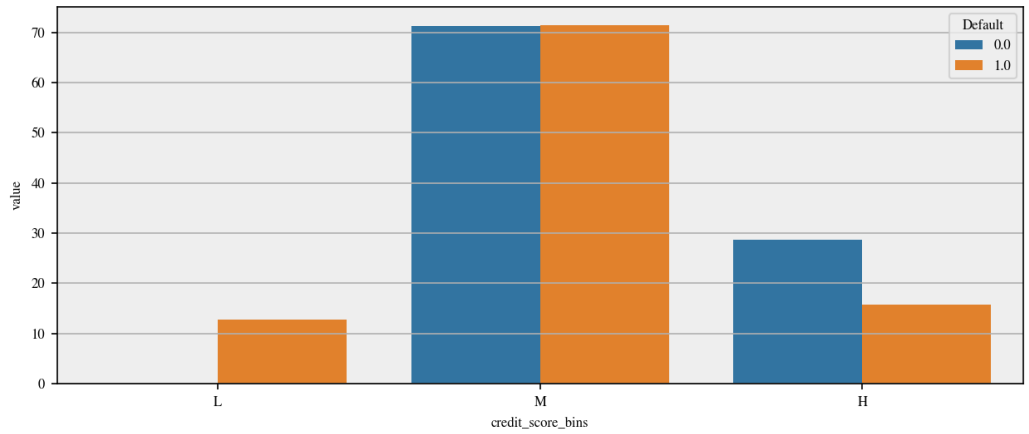
Things that I didn't try because of time constraints are: -

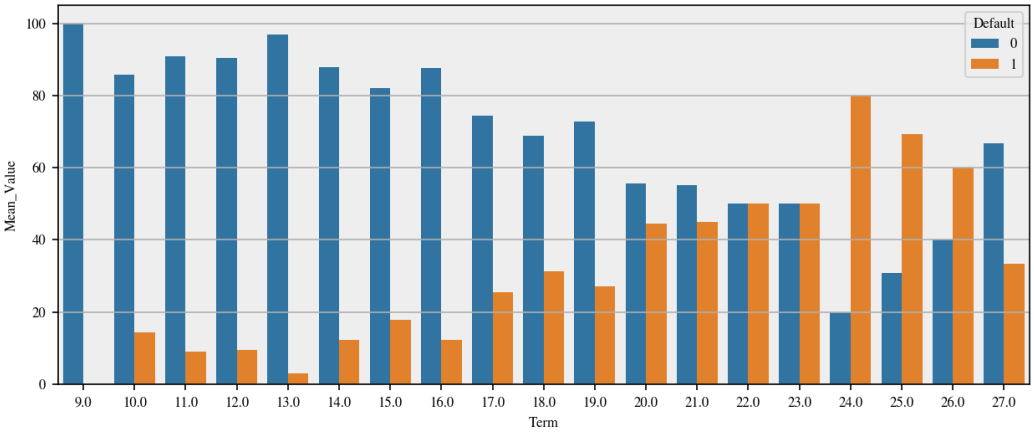
- This dataset was imbalanced. So, to handle this skewed distribution one thing I can do to tune the scale_pos_weight parameters or another way is to use either undersampling or SMOTE sampling techniques.
- Create more aggregate and statistics-based features
- Try Neural network and automl h2o models.

These are the following some plots: -

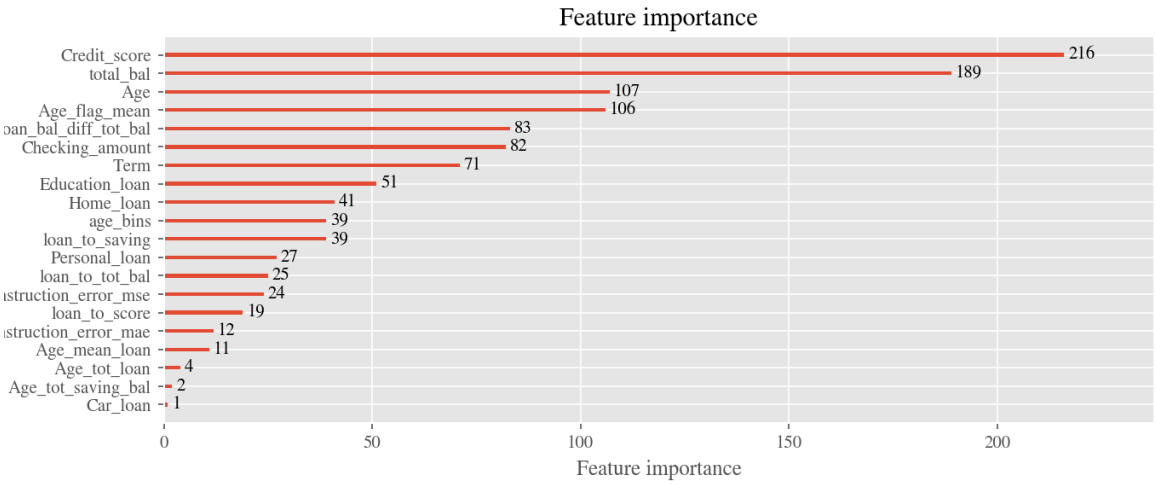




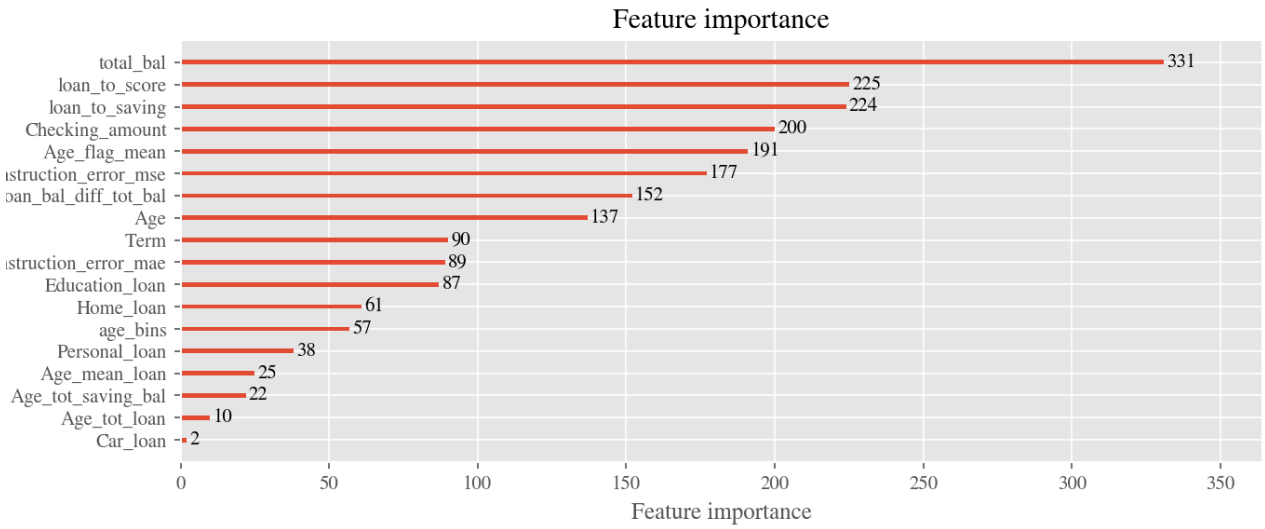




Through Forward Feature Selection importance plot



Through Backward Feature Selection importance plot


Feature importance
total_bal
loan_to_score
loan_to_saving
Checking_amount
Age_flag_mean
ıstruction_error_mse
oan_bal_diff_tot_bal
Age
Term
ıstruction_error_mae
Education_loan
Home_loan
age_bins
Personal_loan
Age_mean_loan
Age_tot_saving_bal
Age_tot_loan
Car_loan
331
225
224
200
191
177
152
137
90
89
87
61
57
38
25
22
10
2
0
50
100
150
200
250
300
350
Feature importance