Data Mining for Business Analysis Final Project

# Loan Default Prediction

By: Anshu Mathur, Minkyoung Kim Bryant, Swati Dixit and Yash Shah

# Table of Contents



# Problem Statement

- Loan repayment is critical. It provides a financial institution with the ability to issue credit (reject or approve loans), set principals, maturity, and repayment calendars.
- When conducting credit  analysis, investors, banks and analysts use a variety of tools (ratio analysis, cash flow analysis, trend analysis) to determine the default risk of a customer.
- Loan default analysis is being used to evaluate a customer's ability to honor its debt obligations.
- The lender evaluates a business to determine if it generates adequate cash flows to meet the debt service, i.e., principal and interest payments.
- Loan default analysis will show a risk rating to an individual customer, based on each customer’s level of risk and the estimated amount of losses that the lender will suffer in the event of default.

***GOAL: to predict clients' repayment abilities using predictive analytics to increase the accuracy of targeting customers types and mitigate risk for the banks.***

# Roadmap


1
Merging the Datasets
2
Median Imputation
3
Feature Selection
4
Modeling
5
Business Implications

01 The Problem Statement
A brief overview and a snapshot of the dataset

**02 Data Cleaning and Feature Selection**
Using Imputation methods and selection of important features

03 Modeling
Review of models and comparing results

04 Business Insights
Findings and Business Implications

# Data

Three datasets were given, previous app, credit card balance and application train

### Previous_Application.csv

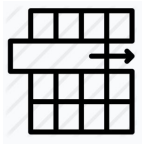
All previous application for Home Credit Loans of clients who have loans in the sample data. Includes features like contract type, loan purpose, payment type, reject reason, etc.

1,670,214 rows

37 columns

### Application_Train._S20.csv

The main table – contains static data for all applications. Each row represents on loan in this datasets. Incudes features like age, gender, employment details of the customer, etc.



292,140 rows

122 columns

### Credit_Card_Balance.csv

Monthly balance snapshots of previous credit cards that the applicant has with Home Credit. Includes features like credit limit, month balance, balance amount, drawing amount, etc.

3,840,312 rows

23 columns

# Overview

Brief Insights into the data

**50k**

total customers

**2K**

seek cash loans

**36k**

are married

**11k**

possess an academic degree

**10k**

default in timely payments

01 The Problem Statement
A brief overview and a snapshot of the dataset

02 Feature Selection
Using imputation methods and selections of important features

**03** Modeling
Review of models and feature of importance

04 Business Insights
Findings and Business Implications

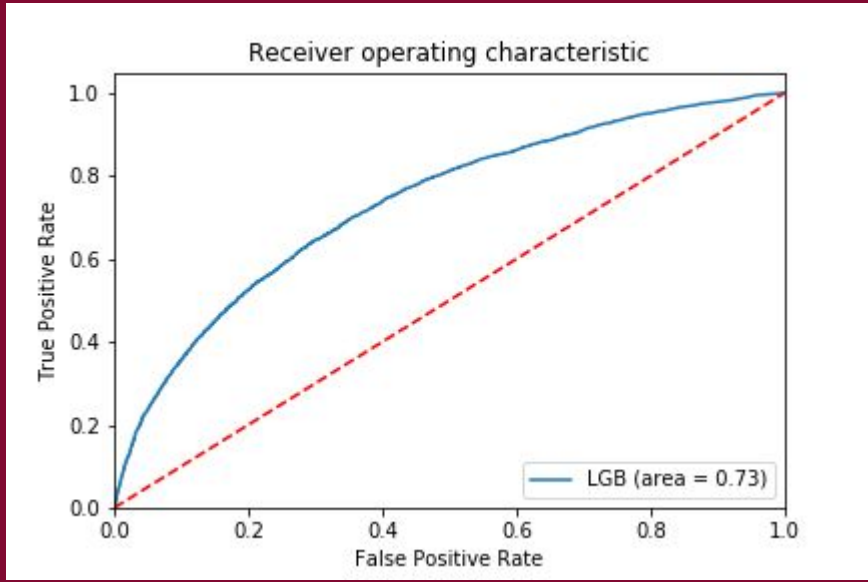
# Models

Six kinds of models were tuned and tested – Random Forest, XGB, LGB, Naive Bayes, Logistic Regression and Decision Tree

| | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| **Random Forest** | 0.6636 | 0.505 | 0.4557 | 79.58% |
| **XGBoost** | 0.6867 | 0.5861 | 0.5988 | 80.33% |
| **LGB** | 0.8264 | 0.5008 | 0.4449 | 80.46% |
| **Naive Bayes** | 0.6478 | 0.5003 | 0.4439 | 79.55% |
| **Logistic Regression** | 0.7102 | 0.5654 | 0.5694 | 80.65% |
| **Decision Tree** | 0.5668 | 0.5671 | 0.567 | 71.77% |

# LGBoosting Machine Model

LGB was selected as the most apt for case




```
Confusion Matrix LGB:
[[12816   310]
 [ 2913   461]]
Accuracy LGB:  80.46666666666667 %
```

**Light Gradient Boosting Machine Model**

**Specifications:**

1000 trees
Cutoff 0.5
101 predictors out of which 98 were non zero predictors

The top 5 features are:
Organization Type
Ext Source 2
Ext Source 3
Ext Source 1
Occupation Type

Accuracy
80.464%

F score
0.4449

01 The Problem Statement
A brief overview and a snapshot of the dataset

02 Feature Selection
Using imputation methods and selections of important features

03 Modeling
Review of models and comparing results

**04 Business Implications**
Findings and business implications

## Business Implications

### Data Accuracy

Loan default analysis shows a risk rating to an individual customer's past history, therefore there is always high risk to predict default rate accurately. The bank should ensure the client information on file is updated and accurate at all times. Since age, organization type, education level etc. are important predictors. External credit scores should be updated frequently as well.

### Income v/s Occupation Type

Occupation type is a more important influencer than income: low-skill labors and drivers have high payment difficulty rate (16.77% and 11.35%) in comparison to accountants and IT staff.

### Lending Policy

Credit utilization, types of credits and changes in spending patterns can be considered attributes of factors for default for the future analysis. Clients with the highest probability to not default could be offered more benefits such as lower interest rates and a higher credit amount. On the contrary potential defaulters should be given conservative rates.

# Thank You!

# APPENDIX

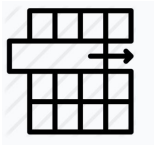
| TOPICS | CONTRIBUTING MEMBERS |
|---|---|
| - Business Understanding | Minkyoung Kim Bryant |
| - Data Understanding | Yash Shah, Swati Dixit, Anshu Mathur |
| - Data Preparation | Yash Shah, Swati Dixit, Anshu Mathur |
| - Modeling | Yash Shah, Swati Dixit, Anshu Mathur |
| - Evaluation | Yash Shah, Swati Dixit, Anshu Mathur |
| - Deployment | Minkyoung Kim Bryant |

# Data Preprocessing

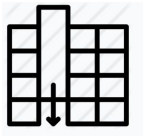
Left joins were used to join the two datasets since every row from the main dataset was required.

**merged.csv**

An amalgamation of the two datasets

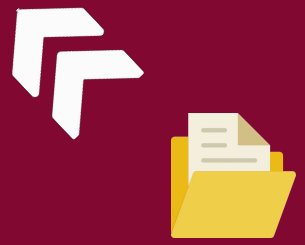


292139 rows



143 columns

**Application_Train._S20.csv**



**Credit_Card_Balance.csv**

| | Confusion Matrices |
|---|---|
| **Random Forest** | Confusion Matrix Random Forest:<br>[[13087 39]<br>[ 3330 44]] |
| **XGBoost** | Confusion Matrix XGB:<br>[[12516 610]<br>[ 2636 738]] |
| **LGB** | Confusion Matrix LGB:<br>[[12816 310]<br>[ 2913 461]] |
| **Naive Bayes** | Confusion Matrix NB:<br>[[13123 3]<br>[ 3371 3]] |
| **Logistic Regression** | Confusion Matrix LR:<br>[[12775 351]<br>[ 2842 532]] |
| **Decision Tree** | Confusion Matrix DT:<br>[[10706 2420]<br>[ 2339 1035]] |