Mini Project

Loan EMI Default Payment Prediction

Introduction

- We have all heard about the term EMI, short for Equated Monthly Installments. Most of the things we buy can be paid using the EMI facility.

- EMI facilities are being extensively used these days, it is very common for banks to offer EMI to its customers. These can come at an interest or it could come at no cost.

- Regular payments of EMIs is very crucial for customers as this may directly affect the credit score.

- In this project, the main main aim is to predict default EMI payments of customers.

Steps in the project

- This project is completed in 5 major steps:

1. Data Cleaning
2. Exploratory Data Analysis (EDA)
3. Feature Engineering
4. Data Preparation
5. Model Building and Tuning

The Data

- The data here consists of 209593 rows and 37 columns.
- It contains records of customers for both 30 days and 90 days.
- The columns in the dataset are as follows:

label	Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan
msisdn	mobile number of user
aon	age on cellular network in days
daily_decr30	Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
daily_decr90	Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
rental30	Average main account balance over last 30 days
rental90	Average main account balance over last 90 days
last_rech_date_ma	Number of days till last recharge of main account
last_rech_date_da	Number of days till last recharge of data account
last_rech_amt_ma	Amount of last recharge of main account (in Indonesian Rupiah)
cnt_ma_rech30	Number of times main account got recharged in last 30 days
fr_ma_rech30	Frequency of main account recharged in last 30 days
sumamnt_ma_rech30	Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)

medianamnt_ma_rech3(	Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)
medianmarechprebal30	Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)
cnt_ma_rech90	Number of times main account got recharged in last 90 days
fr_ma_rech90	Frequency of main account recharged in last 90 days
sumamnt_ma_rech90	Total amount of recharge in main account over last 90 days (in Indonesian Rupiah)
medianamnt_ma_rech9(	Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah)
medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level (in Indonesian Rupiah)
cnt_da_rech30	Number of times data account got recharged in last 30 days
fr_da_rech30	Frequency of data account recharged in last 30 days
cnt_da_rech90	Number of times data account got recharged in last 90 days
fr_da_rech90	Frequency of data account recharged in last 90 days
cnt_loans30	Number of loans taken by user in last 30 days
amnt_loans30	Total amount of loans taken by user in last 30 days
maxamnt_loans30	maximum amount of loan taken by the user in last 30 days
medianamnt_loans30	Median of amounts of loan taken by the user in last 30 days
cnt_loans90	Number of loans taken by user in last 90 days
amnt_loans90	Total amount of loans taken by user in last 90 days
maxamnt_loans90	maximum amount of loan taken by the user in last 90 days
medianamnt_loans90	Median of amounts of loan taken by the user in last 90 days
payback30	Average payback time in days over last 30 days
payback90	Average payback time in days over last 90 days
pcircle	telecom circle
pdate	date

Data Cleaning

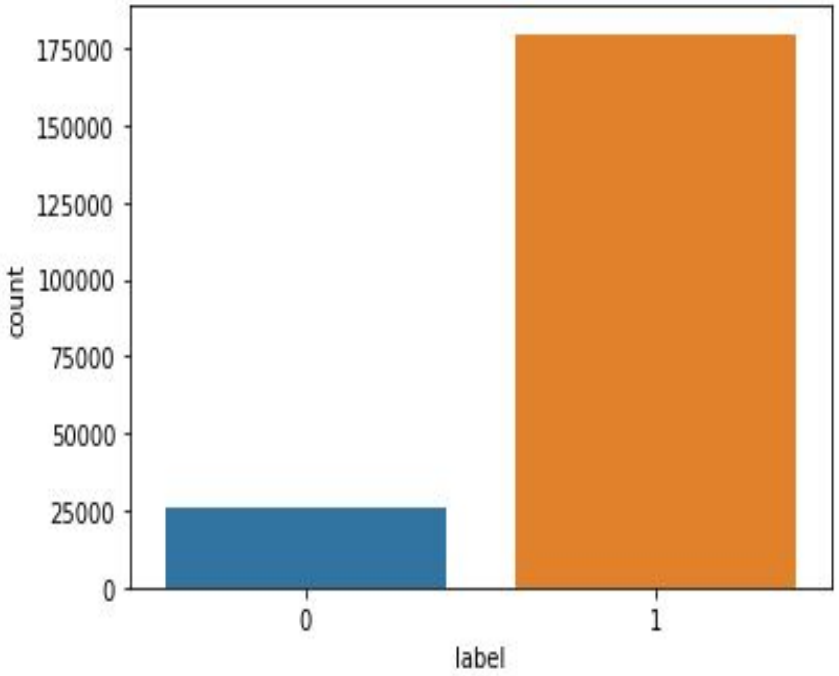
- The first process that is completed in this project is of data cleaning.
- The data has to be cleaned to ensure appropriate analysis and modelling.
- There is an ‘unnamed’ column which will be dropped.
- The ‘msisdn’ contains the IDs of customers which will not be used for the analysis.
- Columns such as ‘pcircle’ and ‘pdate’ have also been dropped as they do not aid analysis.

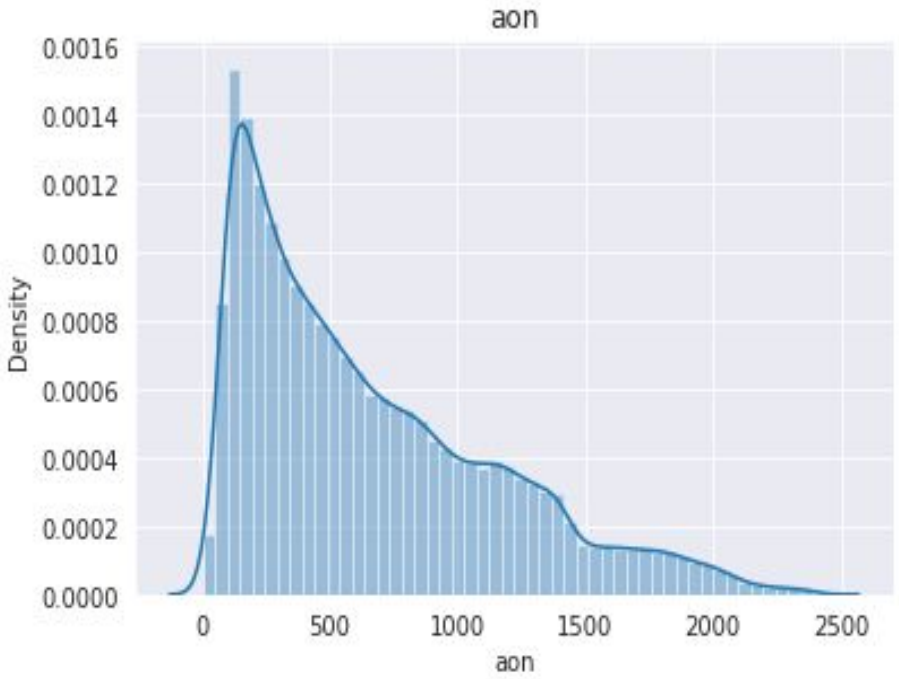
Data Cleaning

- Various negative values are found in columns such as ‘aon’, ‘daily_decr90’, ‘last_rech_date_ma’ and ‘last_rech_date_da’. These values cannot be negative as they represent days or amount spent. These values are converted to positive values.
- As there are records for both 30 and 90 days, records of 30 days are chosen for faster prediction of default EMI payments. This is chosen after comparing values in both records for 30 and 90 days.
- In columns such as aon, last_rech_date_ma and last_rech_date_da there exists some absurdly high values. These values are removed by capping the values to a max of 10 years.

Exploratory Data Analysis

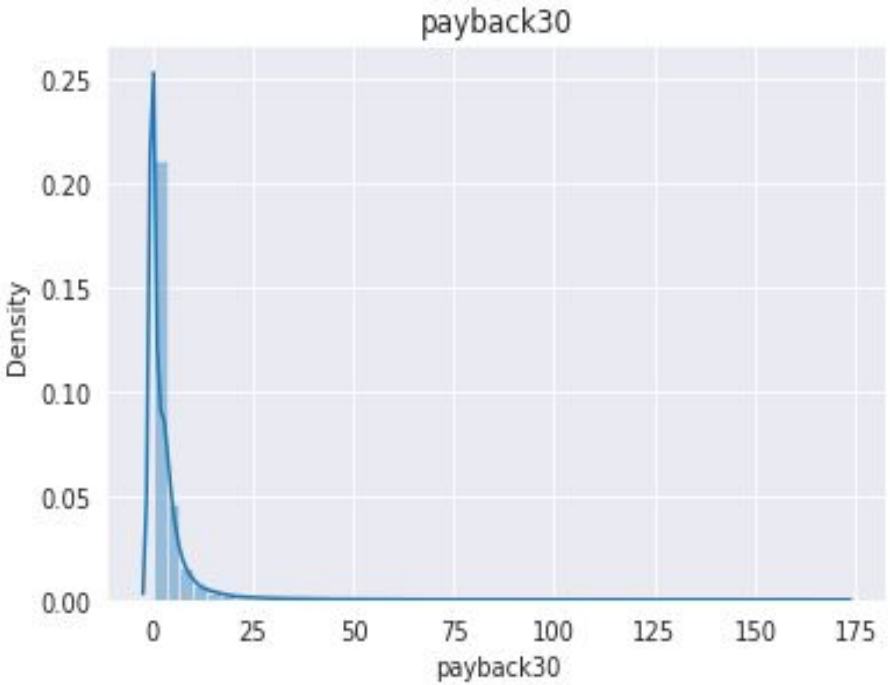
- Univariate Analysis

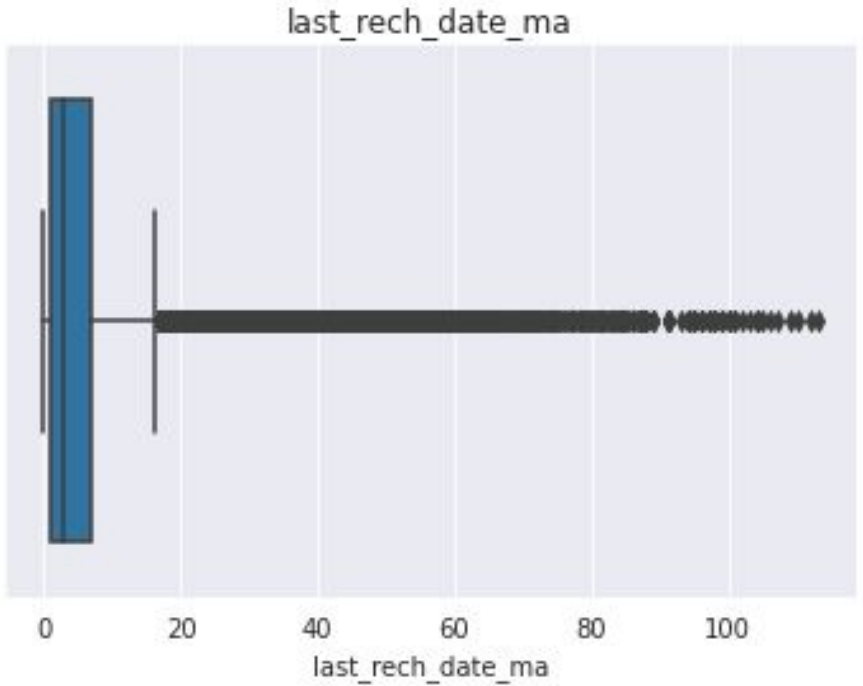




Exploratory Data Analysis

- Univariate Analysis



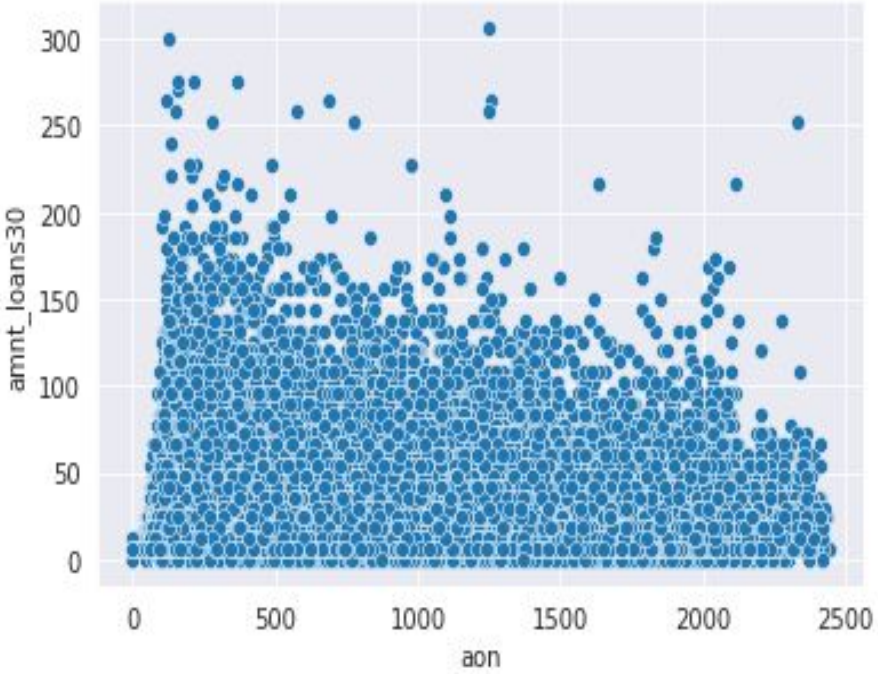


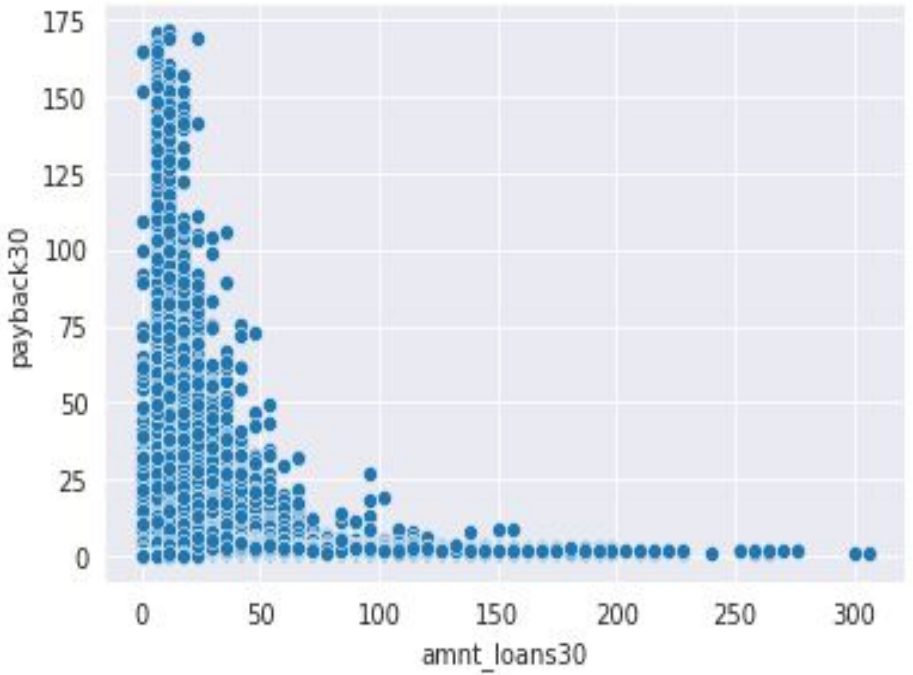
Findings in the univariate analysis

- There is a clear imbalance in the target variable.
- Most of the customers have been on the network from 1 - 5 years.
- Most of the customers have recharged their main account in the last 20 days.
- Most customers have recharged their data accounts in the last 2-3 days. This shows high data usage.
- Most of the recharge amounts are less than 10,000 Indonesian Rupiah.
- Average payback time in the last 30 days has been mostly between 10-15 days.

Exploratory Data Analysis

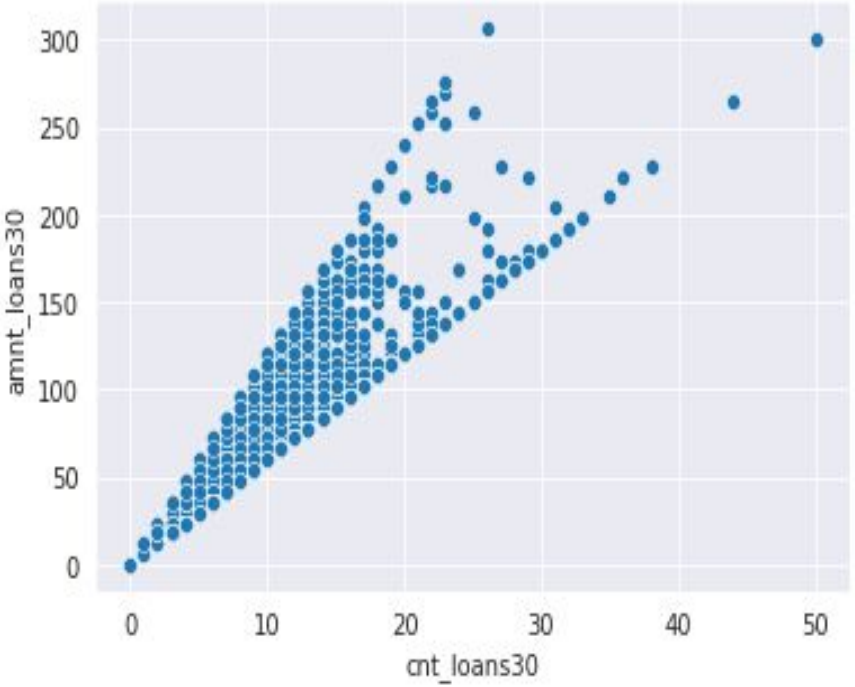
- Bivariate Analysis

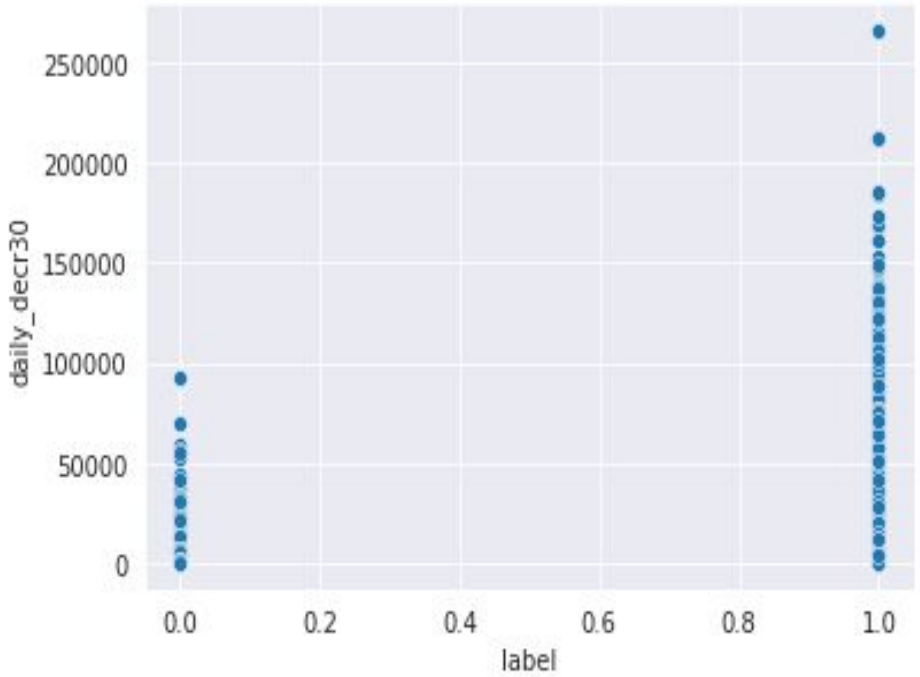




Exploratory Data Analysis

- Bivariate Analysis





Findings in the bivariate analysis

- The amount of loans taken reduces as the age on network increases.
- Customers with higher age on network have lower negative balances than the customers who are new to the network.
- Columns such as ‘cnt_loans30’, ‘amnt_loans30’ and ‘cnt_ma_rech30’ have a higher correlation with the target variable.
- Columns such as ‘maxamnt_loans 30’, ‘fr_ma_rech30’ have lower correlation with the target variable.
- ‘Amnt_loans30’ and ‘cnt_loans30’ are highly correlated to each other. Which makes sense as the amount and number of loans are aligned.

Some of the hypothesis tested

- Does age on network effect average daily amount spent?
- Does age on network effect payment times?
- How does age of the customer on network impact recharge recency, or longer recharges?
- Is there a relation between last recharge amount and last recharge date?
- Does total amount of recharges increase with number of recharges?
- Does larger amount of loan mean longer payback time?
- Do people with larger amounts of loan default more?

All they hypothesis have been tested using visual inspection.

Feature Engineering

- The 3 groups are : < 1 year, between 1 year and 5 years, more than 5 years.
- Feature of average amount per loan was engineered using amnt_loans30 and cnt_loans30.
- Feature of average recharge amount was engineered by dividing the total amount of recharges in 30 days and number of recharges in 30 days.
- Total recharge frequency feature was engineered by combining the recharge frequencies of main account and data account.

Data Preparation

- Data preparation is an essential part of any project. The data that is used for modeling has to be accurate and precise.
- The data in all the columns are of different ranges, hence, the data will be scaled using a min max scaler.
- Imbalance is present in the target variable. Such an imbalance can cause bias in the predictions.
- The imbalance in the target variable is removed by using SMOTE(Synthetic Minority Oversampling Technique).

Model Building and Tuning

- The 2 models that are experimented for this project are:
  a. Random Forest Classifier
  b. XGBoost Classifier
- The main metric used for this project is Recall.
- Both the baseline models perform well, the Random Forest Model overfits but performs better than the XGB Classifier.
- The Random Forest classifier achieves 93% recall on test data, whereas the XGBoost Classifier achieves 81%.

Model Building and Tuning

- As the Random Forest Classifier base model performed well, this model will be chosen for hyperparameter tuning.
- Randomized Search CV is used for the purpose of hyperparameter tuning.
- Parameters such as n_estimators, max_depth, min_samples_split, min_samples_leaf and criterion is used for hyperparameter tuning.
- The overfitting is reduced after hyperparameter tuning and the results on test data have been maintained at 90% Recall.

Model Building and Tuning

Classification report on training data after hyperparameter tuning

	precision	recall	f1-score	support
0	0.91	0.95	0.93	143761
1	0.95	0.91	0.93	143917
accuracy			0.93	287678
macro avg	0.93	0.93	0.93	287678
weighted avg	0.93	0.93	0.93	287678

Classification report on test data after hyperparameter tuning

	precision	recall	f1-score	support
0	0.89	0.90	0.90	36038
1	0.90	0.89	0.89	35882
accuracy			0.90	71920
macro avg	0.90	0.90	0.90	71920
weighted avg	0.90	0.90	0.90	71920

Conclusion

- The main aim of this project was to predict default EMI payments of customers of a telecom company in Indonesia.
- Selected features related to 30 days for faster prediction. Engineered an important features such as customer group, average amount per loan, average recharge amount and total frequency of recharge.
- Used SMOTE to eliminate imbalance in the dataset.
- Built a Random Forest Classifier to predict the default EMI payments.
- Achieved a Recall score of 90% after hyperparameter tuning.

THANK YOU!