LEAD SCORING CASE STUDY

RAGHAVENDRA RAO M S

RAHUL SUPOLIA

CONTENTS

PROBLEM STATEMENT

- We have been given real world data of applications and bureau as shared by Home Credit, to practice the end-to-end process of model development in Credit Risk for Banks, Financial Institutions and NBFCs. We are required to build a bank's internal end-to-end scoring mechanism, based on the application information, clubbed with the raw bureau information.
- The primary objective of this study is to assist Home Credit in deciding which loan applications should be disbursed, and which should be rejected, based on the applicant's past behaviour and application information.

BUSINESS OBJECTIVES

- We are supposed to first gather the information and clean it to make it usable.
- The bureau information is at trade level, each individual trade level information is provided. We need to apply ‘Feature Engineering’ techniques to roll up the information at applicant level, and thereby create manual features for model building
- Build a classification model to differentiate applicants between approves and rejects.
- Once model is built, how to translate the model output into strategies and business insights for the bank?

DATA SET

The dataset provided has 3 files which are explained below:

- *'applications_base.csv'* contains all the information of the client at the time of application. TARGET is the dependent variable for our classification problem (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases). SK_ID_CURR is the unique identifier for the applications table.

- *'bureau.csv'* contains data at trade level and has a total of 17 features. Since this data is at trade level, we would be required to apply Feature Engineering and roll these variables up at SK_ID_CURR level, so that it can be combined with the applications data for model building.

- *'Dataset Description.csv'* is data dictionary which describes the meaning of the variables.

SOLUTION METHODOLOGY

Below steps were used:

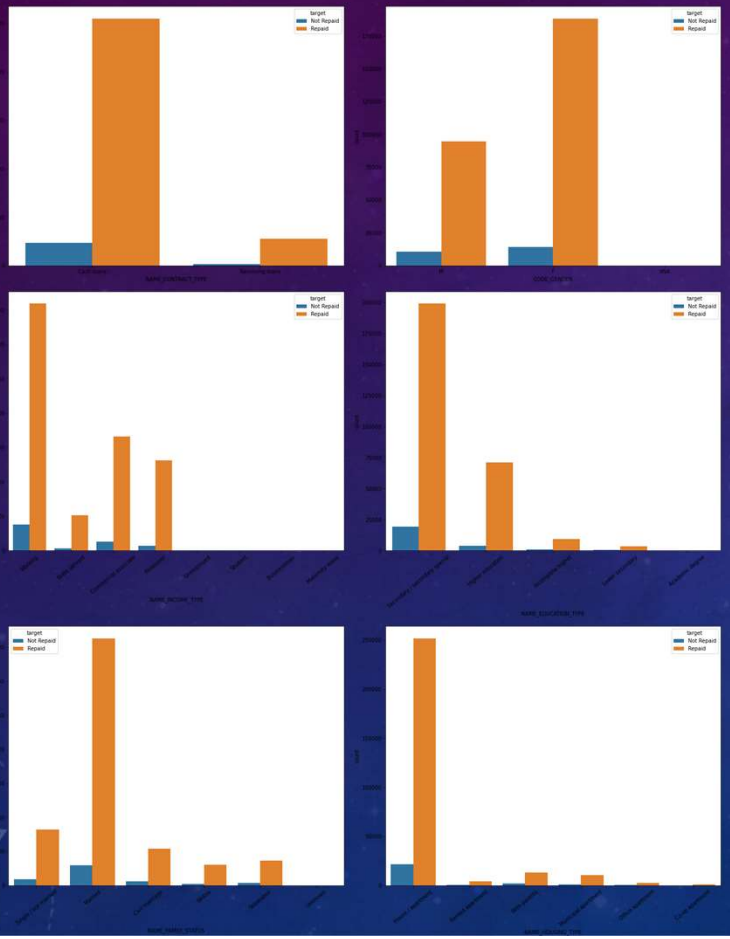
1. Data Cleaning and EDA
2. Feature Engineering
3. Data Preparation
4. Train- Test split
5. Model Building
6. Model Evaluation
7. Performance Metrics
8. Conclusion

DATA PREPARATION AND EDA

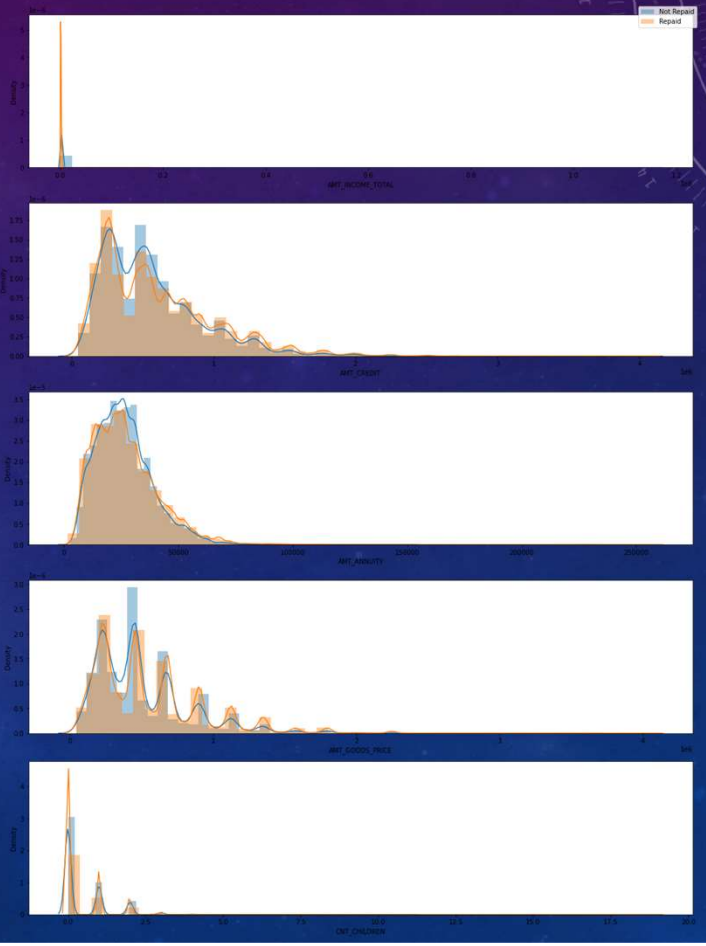
- Removing missing values
- Checking for outliers
- Analysing the categorical and numerical features

EDA

Analysis of Categorical Features in Application data



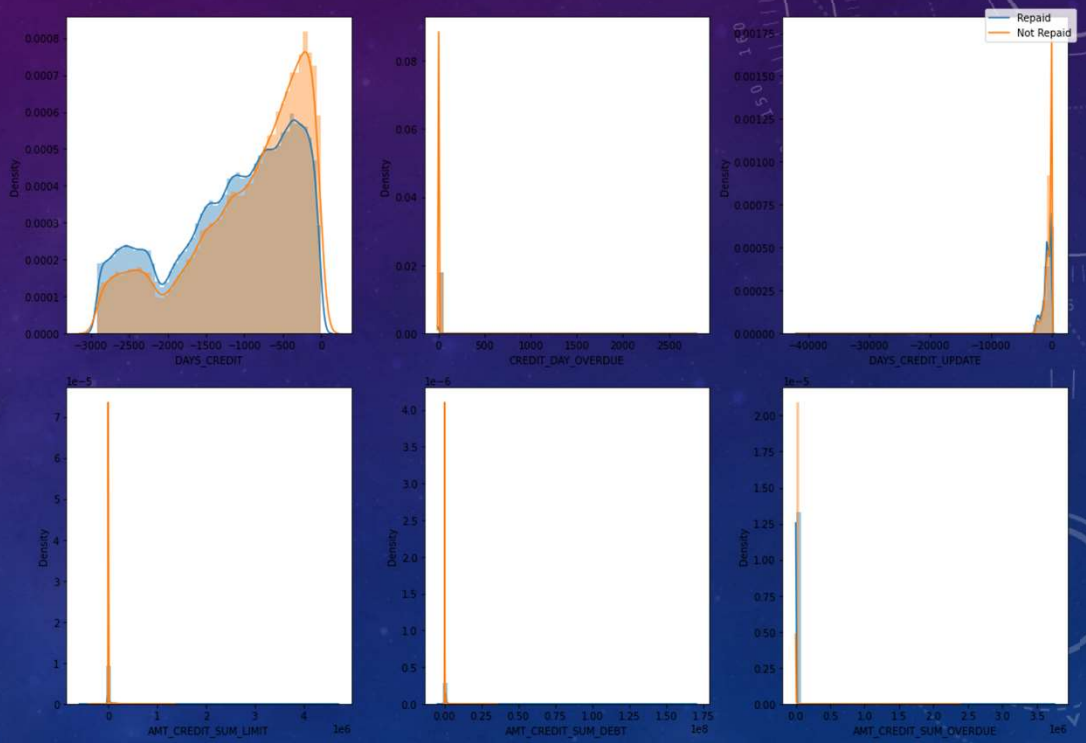
Analysis of Numerical Features in Application data



EDA

Plotting Categorical Features in Bureau data

Plotting Numerical Features in Bureau data



FEATURE ENGINEERING

1. Correlation of features with respect to target variable.
2. Filling outlier with median value for amt income total feature.
3. Replacing the outlier with nan value for days employed feature.
4. Label encoding for application data.
5. One hot encoding for application data

DATA PREPARATION

1. Copying and storing the application data in variable for model.
2. Merging the bureau with application data.
3. Filling missing values with zero.
4. Splitting data into input and output variable.
5. Scaling by Standard scaler.
6. Defining training and testing data using train-test split

MODEL BUILDING

- Splitting dataset into training and testing sets.
- The split was done at 80% and 20% for train and test data respectively.
- Based on the results, we chose the LightGBM & Logistic Regression on which we tested the other metrics to see in depth performance of these 2 models based on several different metrics to choose the best model for our analysis.
- Predictions on test dataset.
- Fitting and Transforming the data to reduce dimensions in to 100
- Overall accuracy comes to 91%

PERFORMANCE METRICS

Confusion Matrix for LGBM

[[42728 13826]

[1752 3197]]

Confusion Matrix for Logistic

[[56498 56]

[4899 50]]

Here, we can see that Logistic Model is failed to predict Defaulters classification, But, In our Case we are more interested in Defaulters classification. So LGBM is more better than Logistic here.

PERFORMANCE METRICS

Since the data available to us is an Imbalanced Dataset, we cannot simply use Accuracy as a metric for evaluating the performance of the model. There are some metrics that work well with imbalanced datasets, of which we will use the below-mentioned metrics

Precision and Recall for LGBM

Precision Score = 0.1878

Recall Score = 0.6459

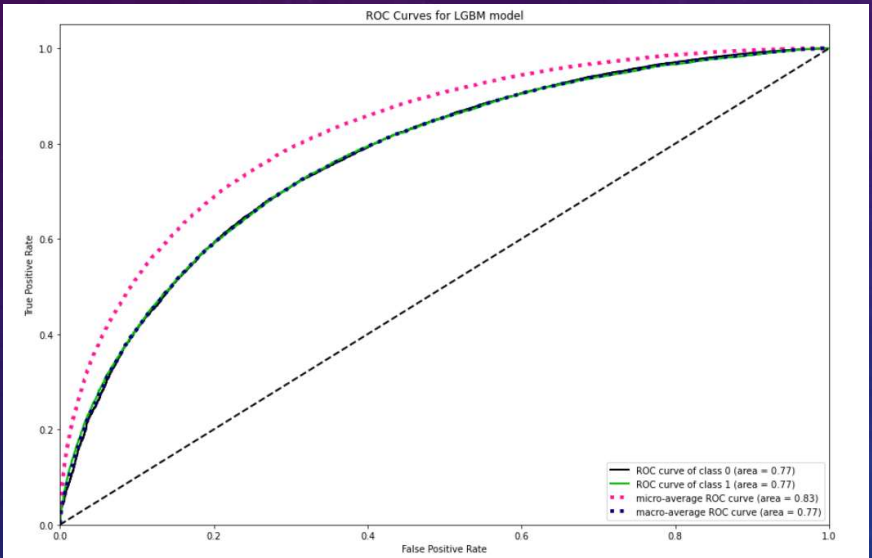
Precision and Recall for Logistic

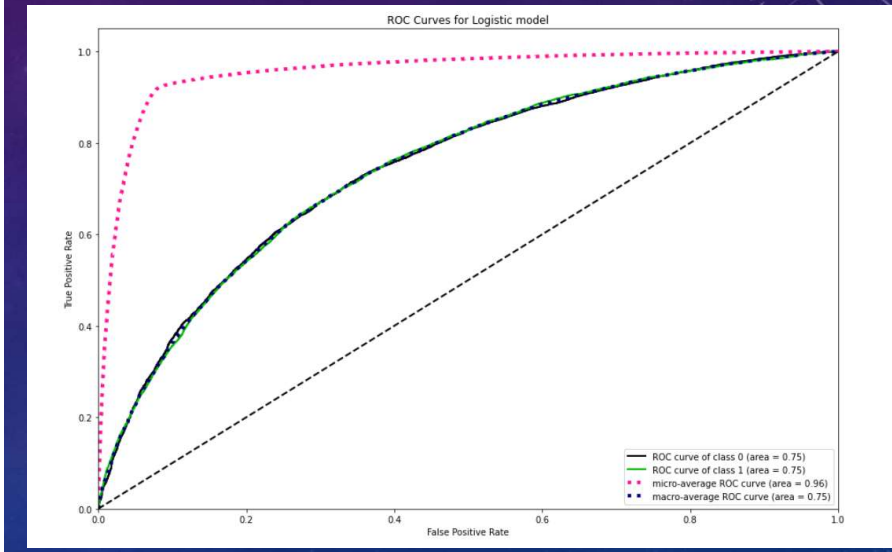
Precision Score = 0.4716

Recall Score = 0.0101

We want more Recall Score even though precision score is less, This is because we care more about minimizing the False Negatives, i.e. the people who were predicted as Non-Defaulters by the model but were actually Defaulters. We do not want to miss out on any Defaulter as being classified as Non-Defaulter

PLOTTING ROC/AUC CURVE





From above ROC plot, looking at AUC values we can say that LGBM model is good at performace of classification.

CONCLUSION

- By Analyzing the data and building LGBM and Logistic Model, considering Recall Score and AUC, We came to know that LGBM is giving better results