Home Credit Group Loan Defaulter Prediction

Md Saimoom Ferdous, PhD
Springboard Data Science Career Track, March 2020 Cohort

Mentored by
Rahul Sagrolikar
Senior Data Scientist, Amazon

Table of Contents

Section 1: Problem Statement

Project Flow

	Problem	Solution
Business	With little information hard to decide on loan sanction	Know the important features to look at
Loan defaulter prediction	Imbalance data set weighs in for majority class	Important features extraction from the trained models

Project Flow - Details

Problem

Business

1. Home Credit Group – needs to disburse loans to most vulnerable group with little to no information
2. Need to make balance, real defaulters don't get loans (FN) and non-defaulters are not barred from loans (FP)

Loan defaulter prediction

1. Minority class always has very less population than majory class which leads to bias
2. Trained model can have better separability yet having high FN/FP ratio

Solution

1. Under-sampling is better option for making balanced dataset
2. Model training and Explainability can help identify and quantify important features

1. Knowing the most important features to look at
2. Minimize FN/FP ratio

Section 2:
Data Collection and Wrangling

[GitHub Link]

Data

Kaggle data
Sourced from Home Credit Group Inc.

Application

Contains 307.5 K clients information with 121 features and a target variable

We will limit to this file only

6 Other Files

Contains information about ‘bureau’, ‘previous credit/cash balance’, ‘installments’ information

Data Wrangling

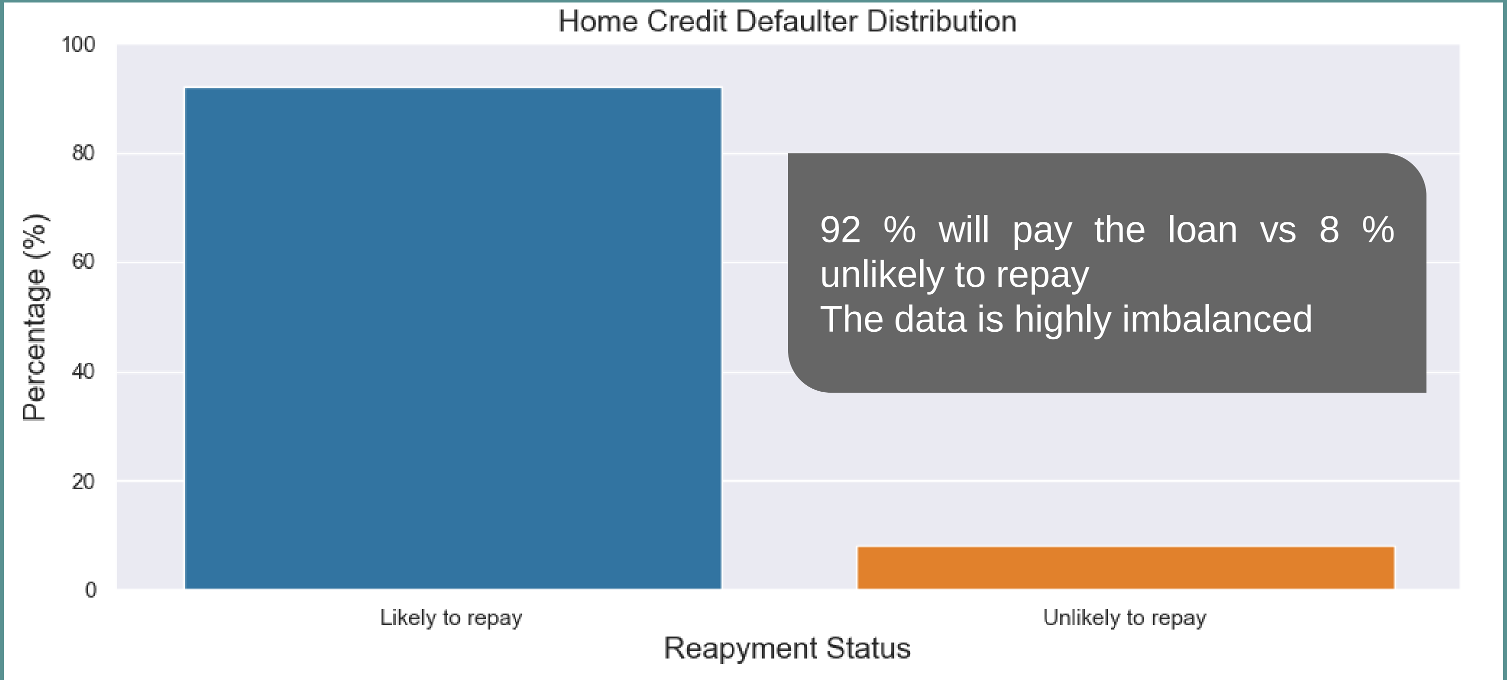
Missing values fixed and duplicate checked

Saved cleaned data for EDA stage

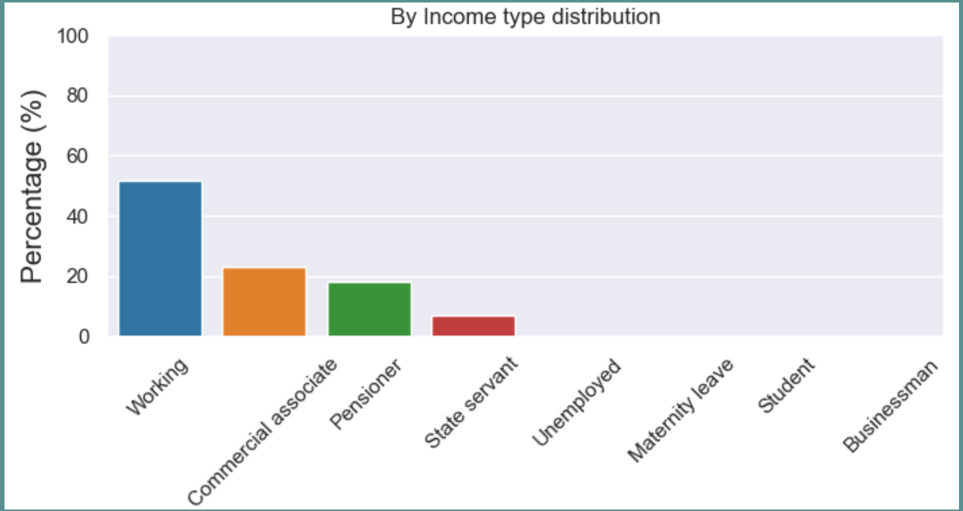
Section 3: Exploratory Data Analysis

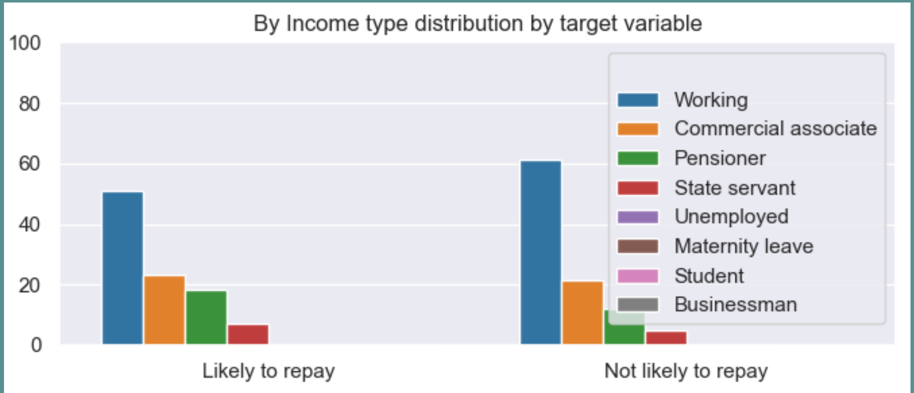
GitHub Link

How Unbalnced the Target is?

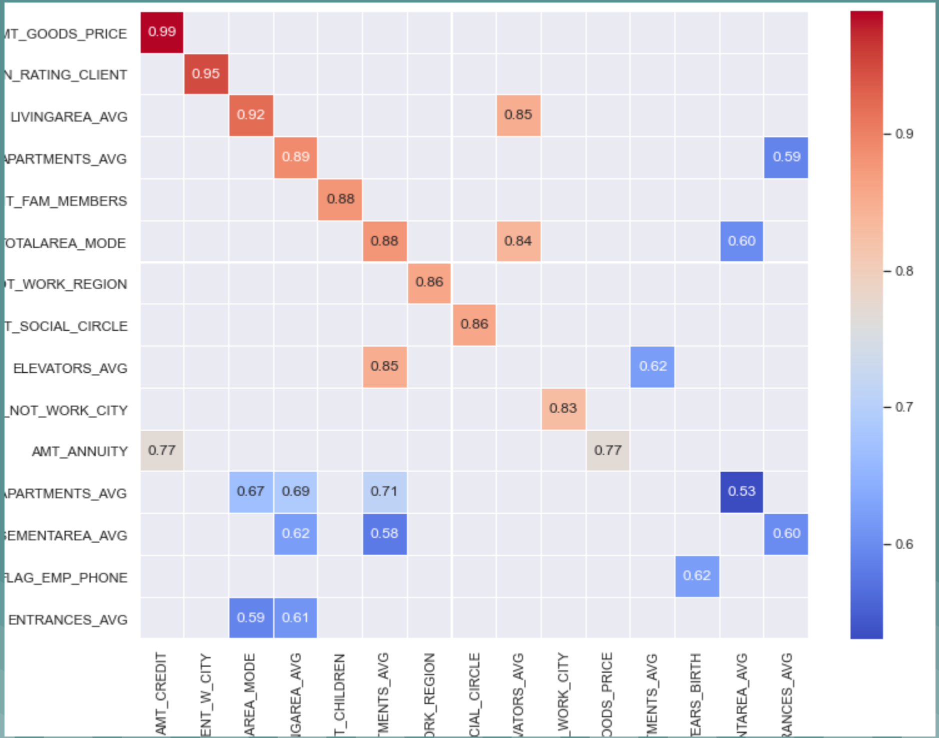


What Income Group the Clients Come from?





Highly Correlated Feature Removal



Variables		Correlations
GOODS_PRICE	CREDIT	0.99
REGION_RATING	REGION_RATING_CLIENT	0.95
LIVINGAREA_AVG	LIVING_AREA_MODE	0.92

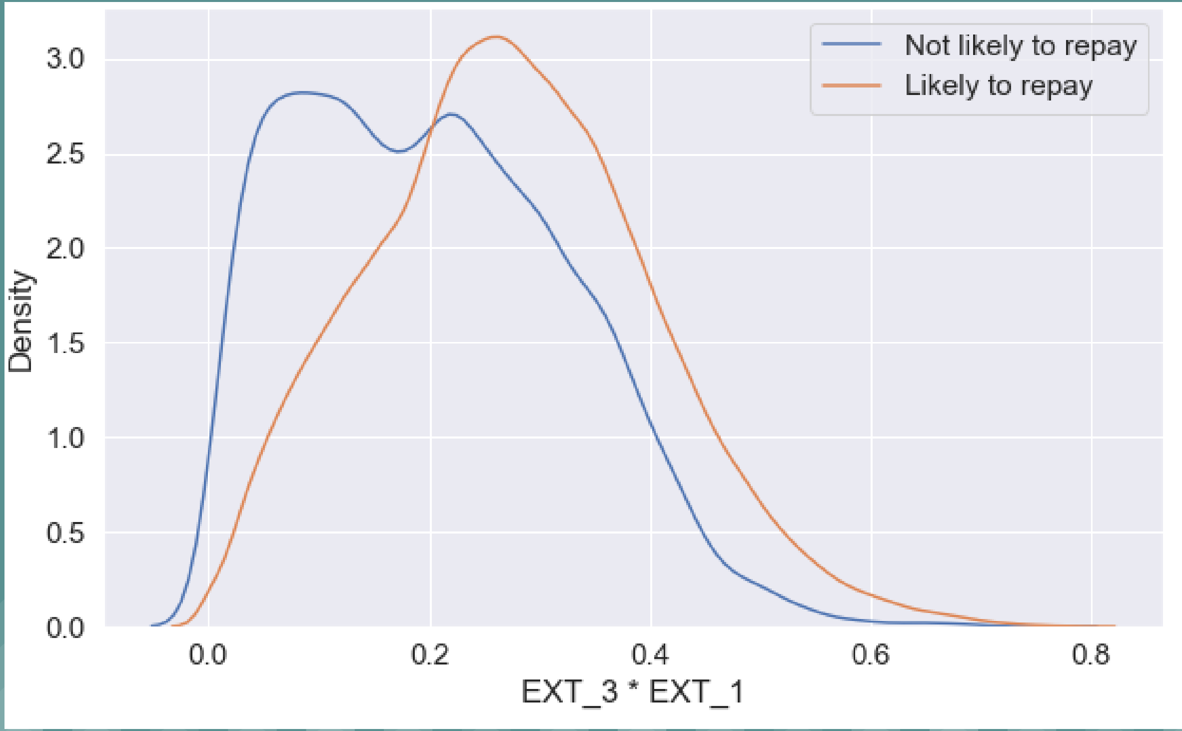
Highly correlated variables (>0.80) were dropped to avoid data redundancy

Feature Creation

14 additional features were created from anomalous features, observations and multiplicative terms

EXT_ features showed maximum correlation with 'target'

Distribution of (EXT_3 * EXT_1) are quite distinct for 'loan repayment' vs 'unlikely to repay'



Section 4: Machine Learning Modelling

Machine Learning

Random Forest

GBM

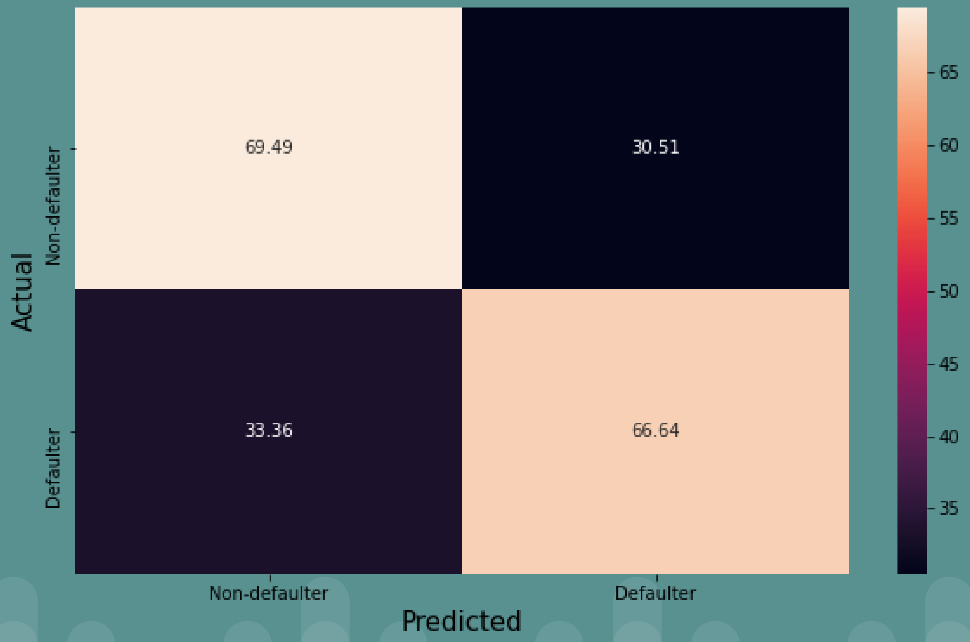
Deep Neural Network (TensorFLow 2.0)

The scientific study of algorithms and statistical models to effectively predict future outcomes based on previously known patterns

XGBoost

Voting Classifier

Deep Neural Network (TensorFlow 2.0)

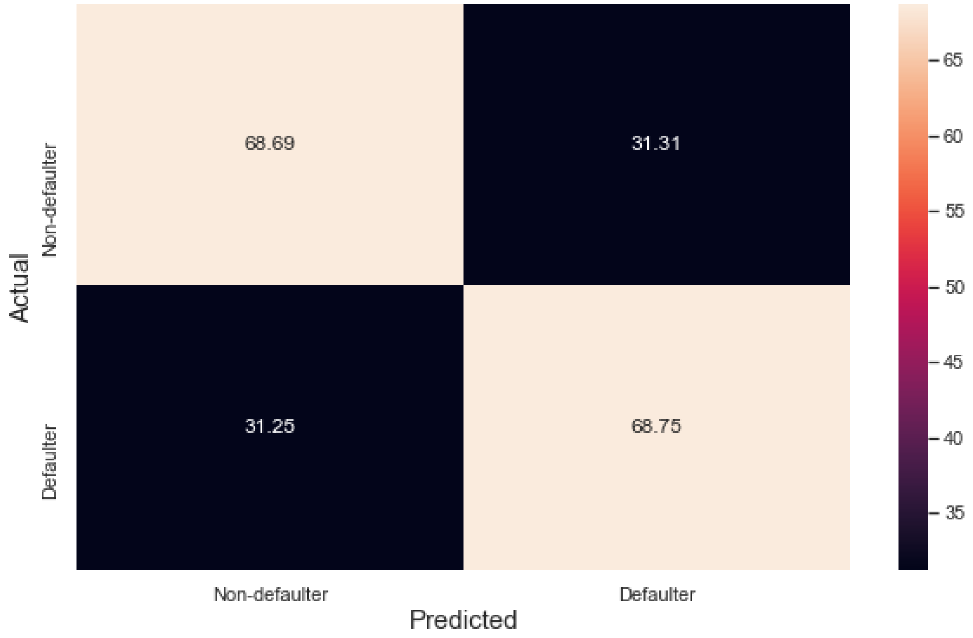


Hyperparameter optimized

False Negative/False Positive = 33/30 %

AUC = 0.7393
Accuracy = 67.85%

GBM



Hyperparameter optimized:

False Negative/False Positive = 31/31 %

AUC = 0.7541
Accuracy = 68.72%

Important Features:
EXT_3_2, EXT_3_1, EXT_2_1

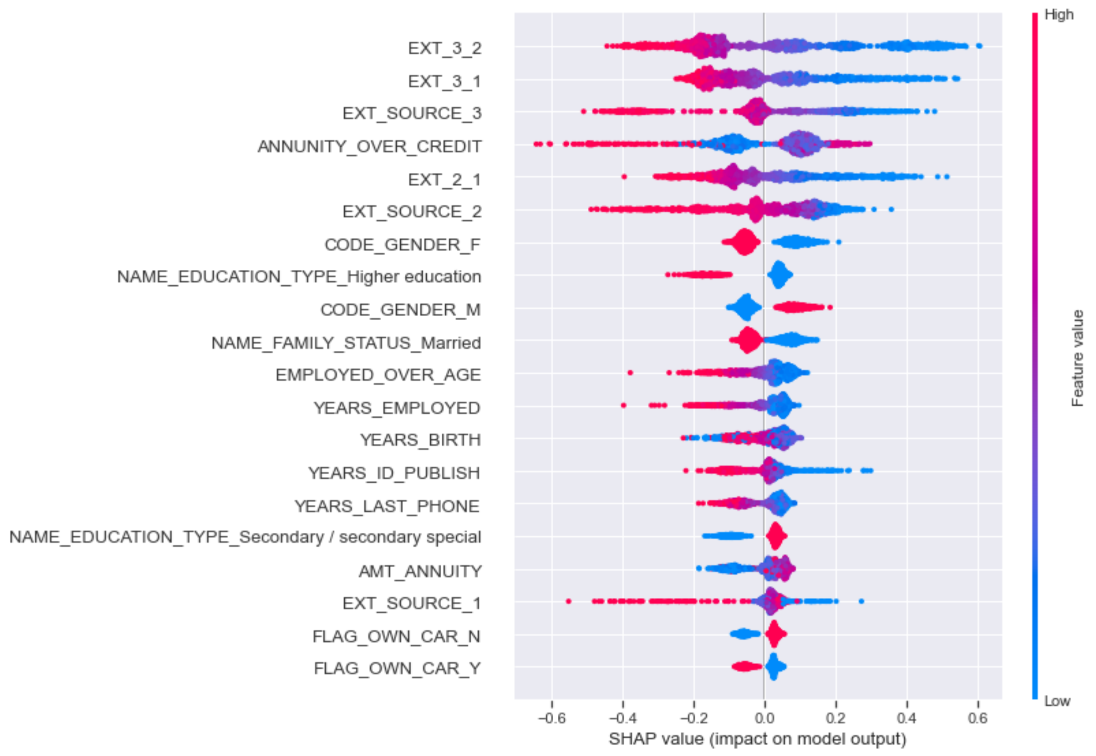
Voting Classifier

Hyperparameter optimized:

False Negative/False Positive = 31/31 %

AUC = 0.7535
Accuracy = 69.04%

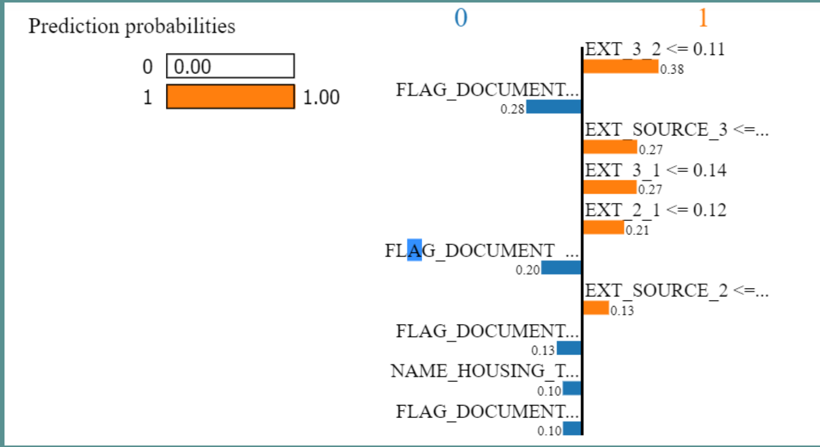
Model Explainability: SHAP Value with GBM Model



Top features with positive/negative correlativity with target variable

The magnitude of individual observation's contribution is also shown

Model Explainability: LIME Coefficients with GBM Model



Feature	Value
EXT_3_2	0.00
FLAG_DOCUMENT_17	0.00
EXT_SOURCE_3	0.00
EXT_3_1	0.00
EXT_2_1	0.00
FLAG_DOCUMENT_7	0.00
EXT_SOURCE_2	0.00
FLAG_DOCUMENT_19	0.00
NAME_HOUSING_TYPE_Co-op apartment	0.00

- Lower value of EXT_3_2 positively correlates with target variable
- Lower value of FLAG_DOCUMENT_17 negatively impacts target variable
- Similar explanation applies to other variables

Machine Learning Results

Metric	
Model	**AUC**
TensorFLow 2.0	0.7393
Random Forest	**0.7450**
GBM	0.7541
XGBoost	**0.7537**

Top 3 Important Features from the Models

Models	Negative Correlation	Positive Correlation
Random Forest	EXT_3_2, EXT_3_1, EXT_2_1	-
GBM	EXT_3_2, EXT_3_1, EXT_SOURCE_3	-
XGBoost	EXT_3_2, ANNUINITY_OVER_CREDIT,	ANNUINITY_OVER_CREDIT, CODE_GENDER_M

NOTE:

- GBM model got the best AUC score
- Low values of EXT_3_2, EXT_3_1 scores in the male population are important to scrutiy for loan approval

[GitHub Link for loan defaulter classification]

Section 5: Conclusion

Conclusion

EDA

Age, gender, demography, socio-economic distribution for loan repayment vs defaulter has been shown

Unknown variables (EXT_X, X=3, 2, 1) are highly correlated with 'target' variable

Modelling

Hyperparameter optimized for Deep Neural net, Random Forest, GBM, XGBoost and Voting Classifier models

Feature importance and explainability was determined for tree based models

Results

GBM yielded best AUC score of 0.7541 which is 2.6% improvement over base model

Low threshold of EXT_3_2, EXT_3_1, EXT_2_1 in the male clients are prone to becoming loan defaulters

Looking Forward

	Data Balancing Method	Feature Engineering	Optimization	Online App Deployment
Current	Currently not done	Manual feature engineering	Step by step hyperparameter optimization for TF model	Currently not done
Future	Over sampling and synthetic data creation	Automatic feature engineering with additional dataset	Keras tuner can be used	This model can be integrated into production step

Thank You

Md Saimoom Ferdous
Email: saimoom_026@yahoo.com
LinkedIn: https://www.linkedin.com/in/saimoom-ferdous/
GitHub: https://github.com/saimoom026
Project Details: https://github.com/saimoom026/Springboard/tree/student-branch/springboard/Capstone%20Three