Prediction of the Applicant's Ability to Repay the Loan

Home Credit Default Risk

Yu Chun Peng, Chien-Chu Hsu, Chia-Yen Ho, Devansh Bhasin

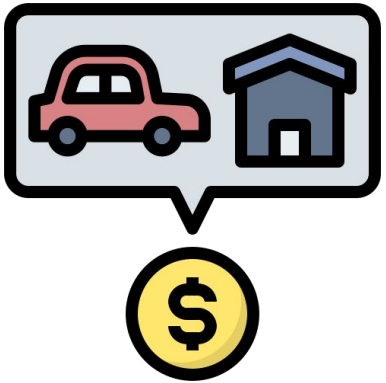
Abstract

- An international non-bank financial institution
- Focuses on responsible lending to unbanked population with little or no credit history

People with insufficient credit history face…

Declined by bank

Exorbitant interest



Collateral required

Challenges for Home Credit

High risk due to lack of credit history

Time-consuming because of the scattered records

Objective

Find out the loan applicants who are capable of repaying a loan, given financial information from Home Credit and other sources.

There are 8 different sources of data

01

Application Train

The main table includes information about each loan application. Use this to train our models.

02

Application Test

The main table excludes information about each loan application. Use this to test our models.

03

Bureau Data

All applicant's previous credits provided by other financial institutions from Credit Bureau.

04

Bureau Balance Data

The monthly balances of previous credits in the Credit Bureau.

There are 8 different sources of data

05

Previous Application

Applicant's previous loan applications with Home Credit.

06

Cash Balance

Point of sales and cash loans that the applicant had with Home Credit.

07

Installments Payments

Repayment data for each installments of credit with Home Credit.

08

Credit Card Balance

Previous credit cards loans that the client applicant with Home Credit.

Data Preprocessing

Create features

The process of creating features based on our domain knowledge

Encoding the Categorical Variable

The process of converting categorical data into integer format

Imputation of Missing Value

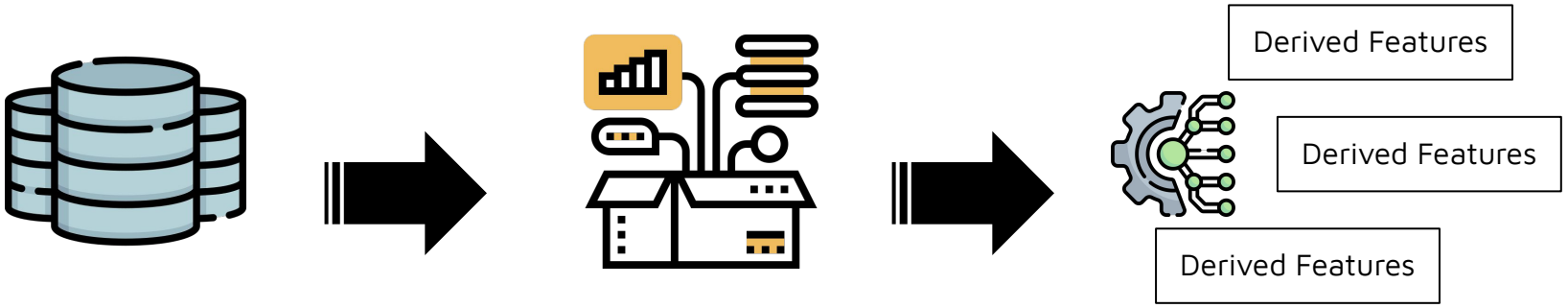
The process of replacing missing data with substituted values

Aggregation

The process of aggregating data sets together using min, max, mean, and standard deviation

Feature Engineering

New features were created based on our estimate of impact of default-rate:



Raw Data

220
Original Variables

618
Total Variables

Methodology

Logistic Regression	is easy to implement yet provides great training efficiency in some cases. It makes no assumptions about distributions of classes in feature space.
XGBoost	is an optimized library for regularized distributed gradient boosting designed to be highly efficient, flexible and portable.
LightGBM	is a fast, distributed, high-performance gradient boosting framework that uses a tree-based learning algorithm.
Catboost	implements symmetric trees, thus helping in decreasing prediction time.
Stacking	is a model-centric ensemble model that can combine the predictions from multiple machine learning models.

Model Results

Models	AUC Score
Logistic Regression	**0.71174**
XGBoost	**0.79111**
LightGBM	**0.77594**
Catboost	**0.78778**
Stacking (XGBoost+LightGBM+Catboost)	**0.79316**

We use AUC Score as the metric

The predictive performance of a model can be quantified in terms of the area under the ROC curve (AUC), which lies in the range [0,1]. And a higher value of it means better accuracy.

Model Results

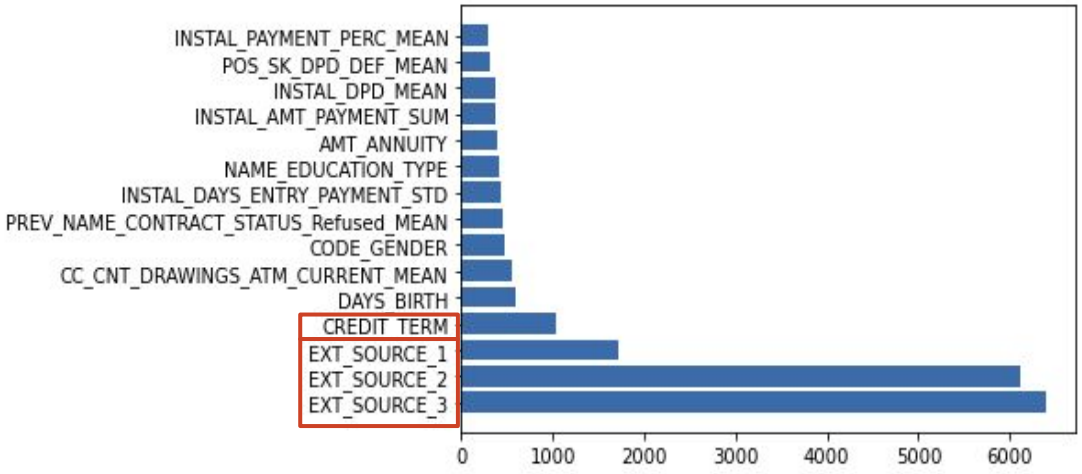
Models	AUC Score
Logistic Regression	**0.71174**
XGBoost	**0.79111**
LightGBM	**0.77594**
Catboost	**0.78778**
Stacking (XGBoost+LightGBM+Catboost)	**0.79316**

We use AUC Score as the metric

The predictive performance of a model can be quantified in terms of the area under the ROC curve (AUC), which lies in the range [0,1]. And a higher value of it means better accuracy.

Feature Importance

- **External Resources** are top 3 key contributors
- **CREDIT_TERM** also stands out. This is a feature we created using Loan annuity divided by Credit amount of the loan



Limitations

Limited Time

Didn’t have enough time to explore all the variables completely before we fully understand the business domain knowledge

Missing Values

Some features used for building our model contained over 60% missing records

Future Work

1. Do better at feature engineering, imputation of missing values, and handling of extreme values by **developing business understanding**

2. Implement **advanced techniques**, such as Neural Network

3. Consider the **cost of error** when evaluating the performance of models

Thank You