Home Credit Default Risk Prediction

Huiwen Wu

University of California, Irvine

huiwenw@uci.edu

September 26, 2019

Overview

1 Background

2 Models

3 Results

Background

- Many people struggle to get loans due to insufficient or non-existent credit histories.
- This population is often taken advantage of by untrustworthy lenders.
- A variety of alternative data – including telco and transactional information– is useful.
- Make use of additional data to provide the unbanked population a positive and safe borrowing experience.
- Data provided by Home Credit.

Mathmetical Problems

- Which problem? Regression or Classification?
- Input: Data of applications, bureau, credit card, installment payments, POS CASH and previous applications.
- Labels: 0 or 1.
- Eventually, there are only two statuses: fully paid and charged off.
- Output: Probability of clients' ability of loan repayment, a value from 0 to 1.

Data

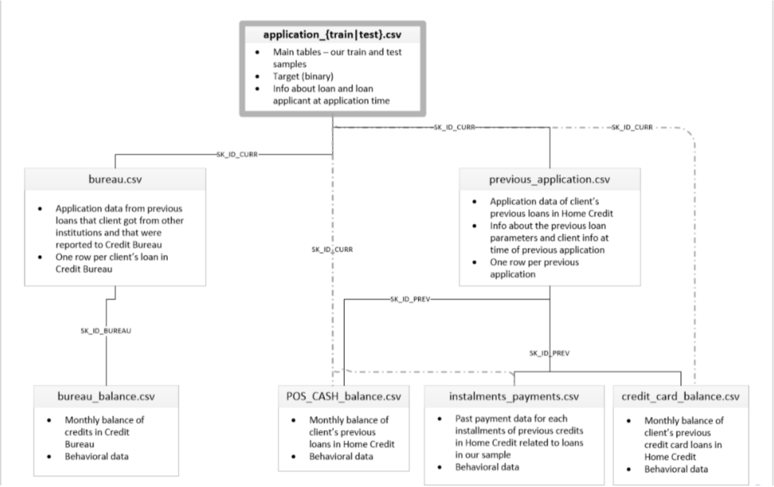
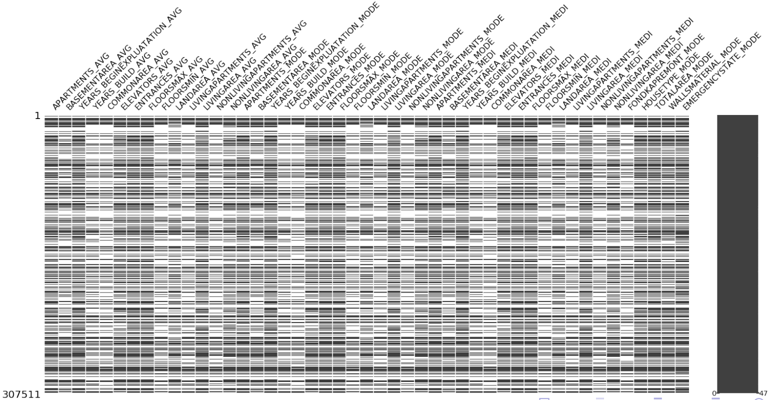


Figure: Home Credit Data

Data Characteristics

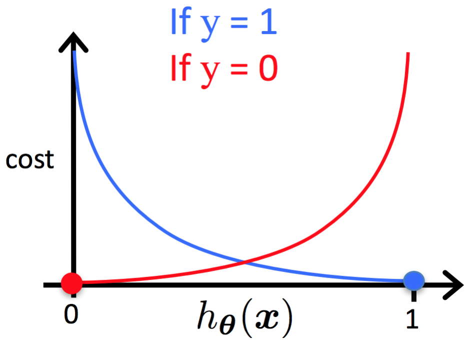
- Massive data. – 799 features.
- Redundant information.
- High sparsity features. – Cols 42-89, 94-114 have sparsity more than 50%.



Logistic Regression

Logistic Regression uses logistic function to estimate probability.

$$c(\theta) = \begin{cases} -\log(h_\theta(x)) & y = 1 \\ -\log(1 - h_\theta(x)) & y = 0. \end{cases}$$



Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decision and their possible consequences, including chance event outcomes, resource costs, and utility.

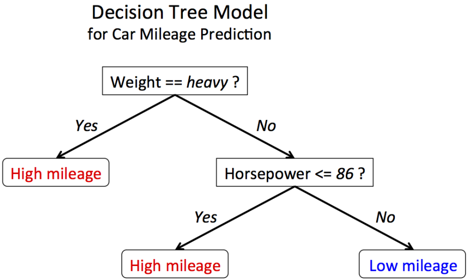


Figure: Decision Tree Example

Gradient Boosting Trees

Boosting is an ensemble technique in which the predictors are not made independently, but sequentially. This method tries to fit the new predictor to the residual errors made by the previous predictor.

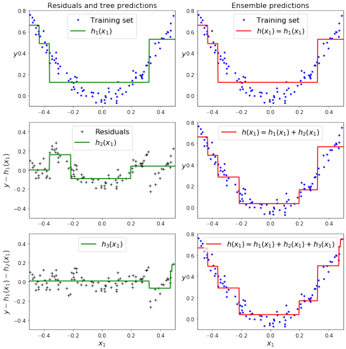


Figure: Gradient Boosting Tree Eaxmple

Models

Logistic Regression

(+) Easy to implement and very efficient to train.
($\pm$) Feature engineering plays an important roles.
(−) Produce linear decision boundary.

Decision Tree

(+) Less data cleaning is requires (NAN) and data type is not a constraint.
(+) Implicitly perform feature selections.
(−) Slower and overfitting.

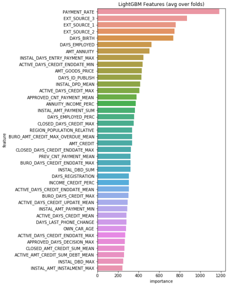
LightGBM

(+) Model is more powerful compared to decision tree.
(+) Fast compared to gradient boosting tree.

Pipeline

- Preprocess train and test data.
- Data augmentation.
- Feature Selection using LightGBM
- Logistic Regression, Decision Trees and LightGBM models.
- Bayesian optimization for LightGBM.
- Model Ensemble.

Feature Importance

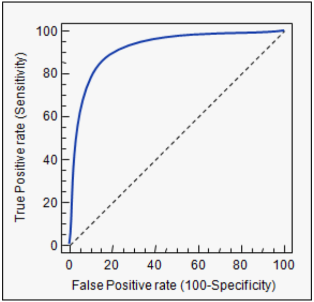


(a) LightGBM Importances

(b) Decision Tree Importances

Figure: Feature Importances

Results: AUC Scores



Logistic Regression	Decision Tree	LightGBM
0.671	0.678	0.787

Table: AUC Scores

Improvements

- Make more use of Exploratory Data Analysis.
- Use Neural Networks as one model.
- Ensemble various models.
- Light feature selections.

Thank you!