Home Credit Default Risk Prediction

Huiwen Wu

University of California, Irvine

huiwenw@uci.edu

September 25, 2019

Overview

1 First Section
- Subsection Example

2 Models

3 Second Section

Background

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustyworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience. A variety of alternative data – including telco and transactional information – are used to predict clients' repayment abilities.

Mathmetical Problems

- Input: Data of applications, bureau, credit card, installments payments, POS CASH and previous applications.
- Output: Probability of clients' ability of loan repayment, a value from 0 to 1.
- Probelm: Regression or Classification?

Data

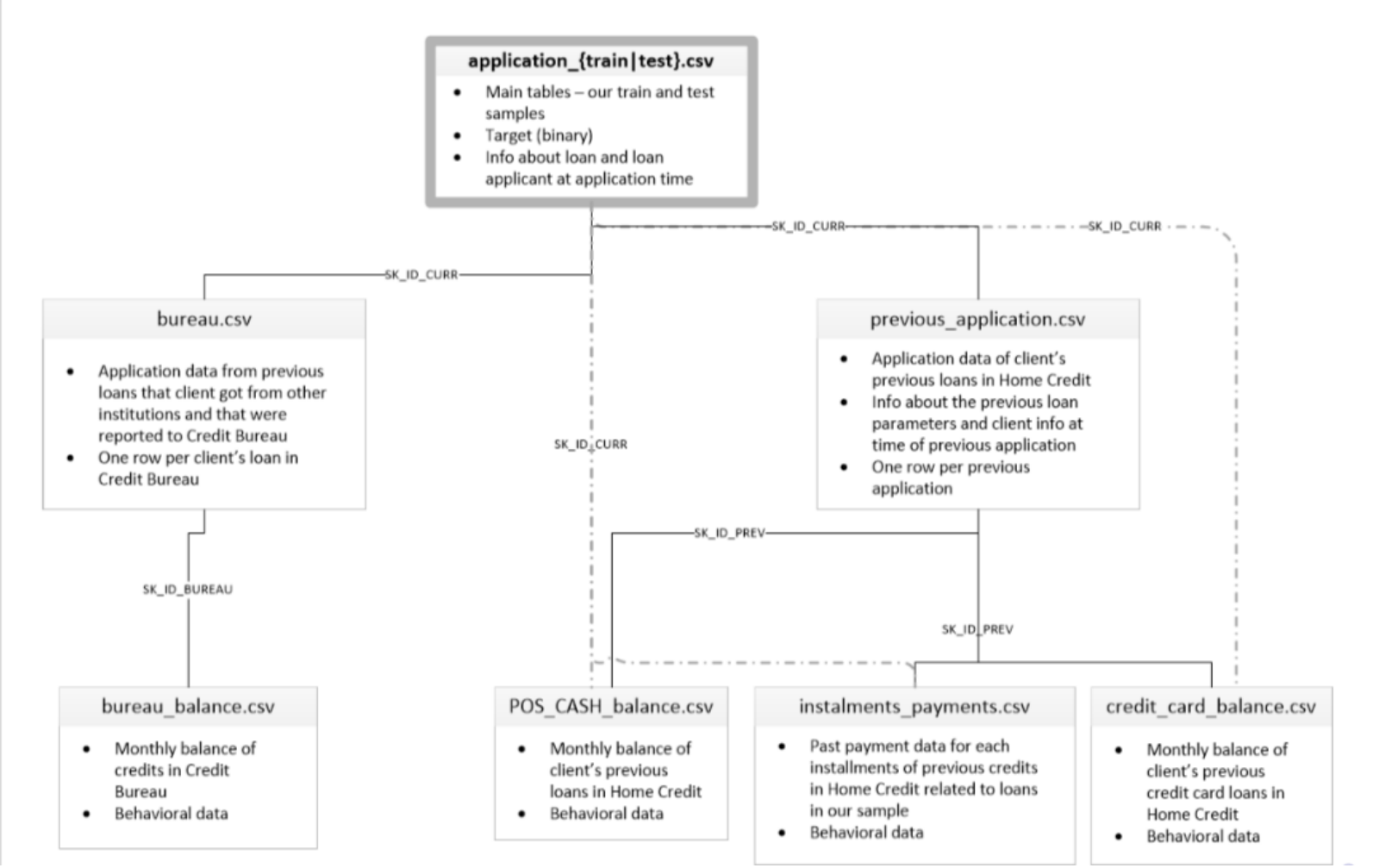


Figure: Home Credit Data

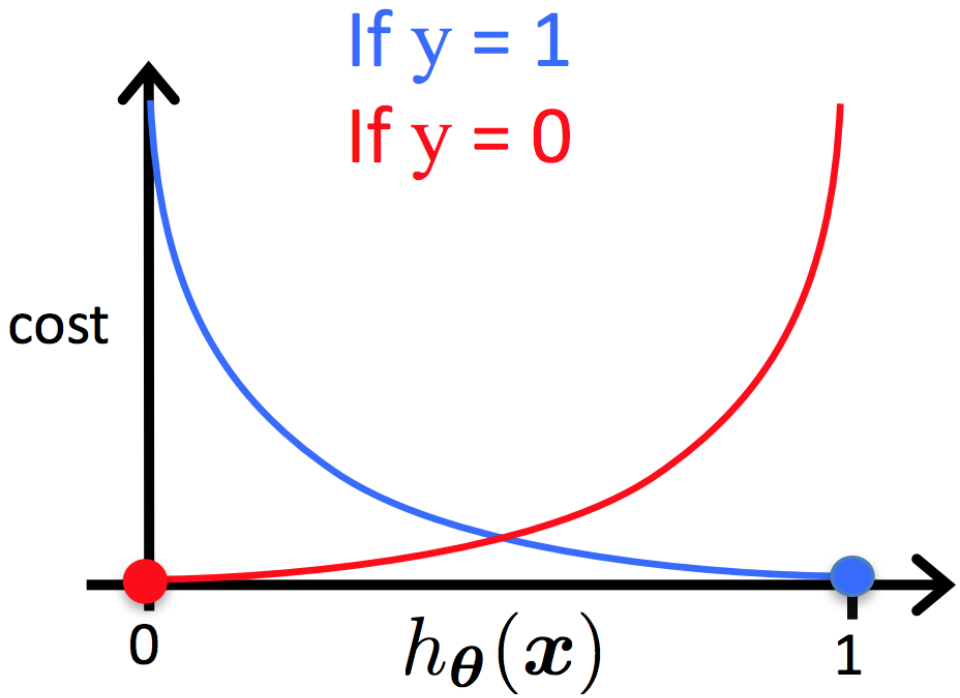
Data Characteristics

- Massive data and a lot of features 799.
- Redundant information.
- A lot of sparse features.
- High correlation between some features – number of children, house type, ages.

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.

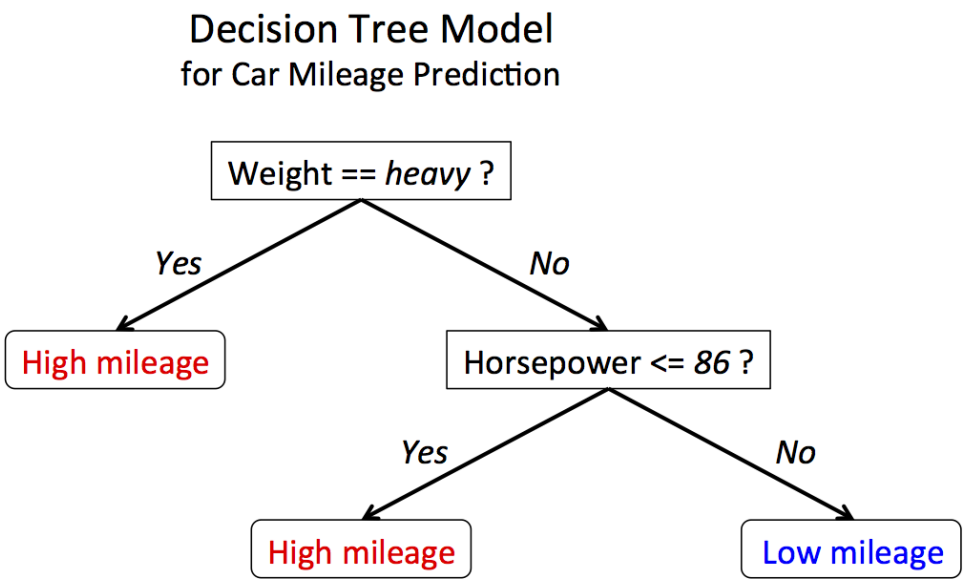


Figure: Decision Tree Example

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 redictor to the residual errors made by the previous predictor.

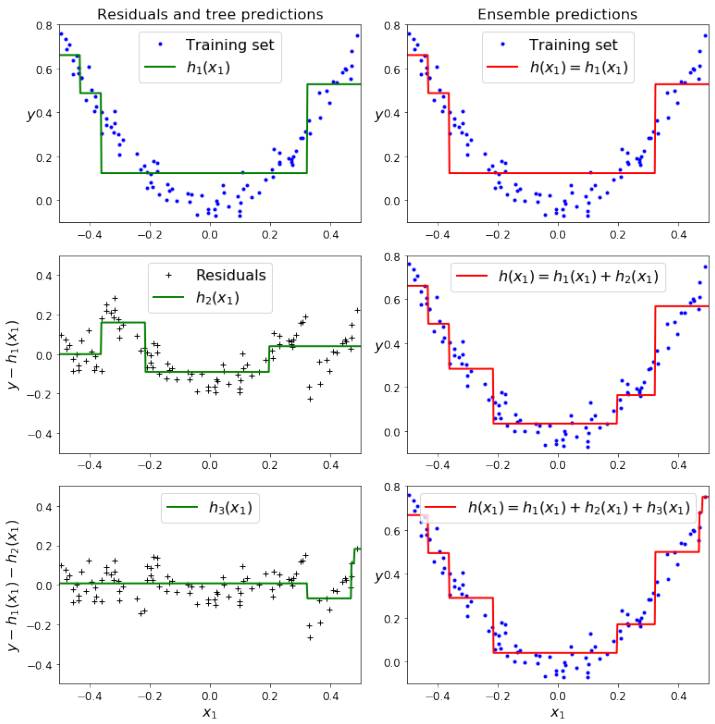


Figure: Gradient Boosting Tree Eaxmple

Models

Logistic Regression

Easy to implement and very efficient to train.
Feature engineering plays an important roles.
Only produce linear decision boundary.
Data needs to be well preprocessed.

Decision Tree

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

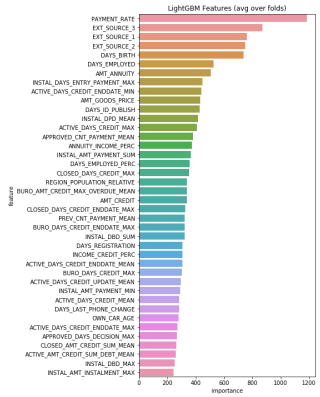
LightGBM

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

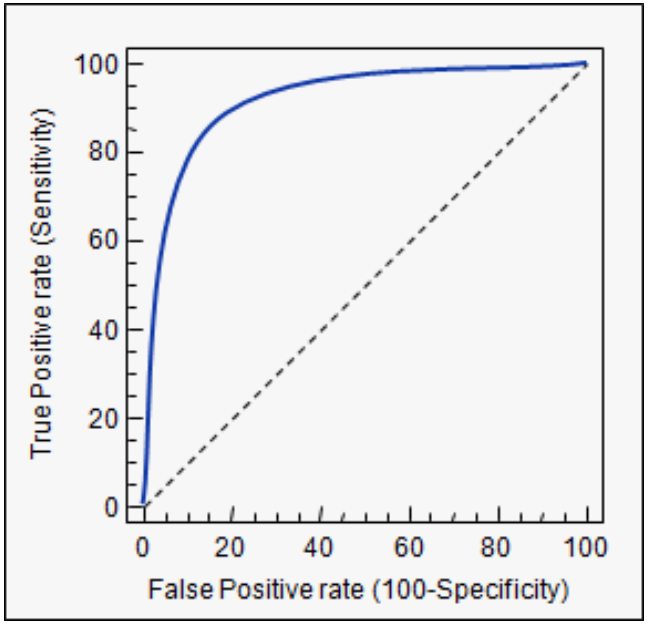


(a) LightGBM Importances

(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 EDA.
- Use Neural Networks as one model.
- Ensemble various models.
- Light feature selections.

Thank you!