Loan Credit Risk and Evaluation of Different Machine Learning Architectures on Tabular Data

Muhammad Ismaeel - MSDS19029

Loan Risk Problem

- Large financial institutions receive hundreds of loan applications everyday. By automating the process of loan request approval or decline, thousands of man hours could be saved.

Problem Statement

- Large financial institutions receive hundreds of loan applications everyday. By automating the process of loan request approval or decline, thousands of man hours could be saved.
- Our objective is to develop a system which based on features present in the loan application can determine whether loan should be approved or declined.

Dataset

We used an open source dataset provided by Home Credit Group which has issued more than 200 Million loans so far.

The dataset was chosen because of its richness as it contains more than 200 features and its realistic nature as it contained class imbalance and sparse features which are expected to be present in any data in financial domains.

Dataset

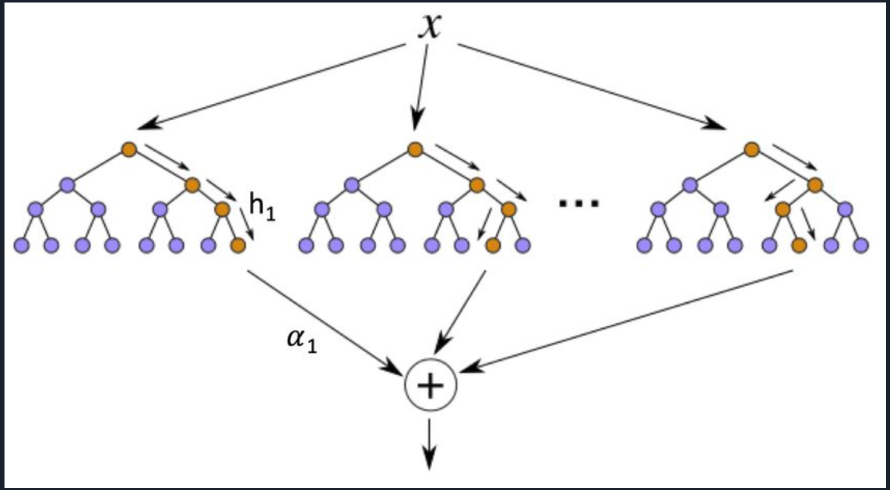
- We used an open source dataset provided by Home Credit Group which has issued more than 200 Million loans so far.
- The dataset was chosen because of its richness as it contains more than 200 features and its realistic nature as it contained class imbalance and sparse features which are expected to be present in any data in financial domains.

Challenges

- Class Imbalance: As large number of loan applications get rejected, hence, dataset had a class imbalance. Positive examples were only 10% of the total examples in the dataset.
- Sparsity Of Data: Large number of examples in dataset were missing entries for one or more features. This sparsity is a huge challenge for ML based methods

Methods - Gradient Boosting Trees

- Gradient Boosting Trees uses an ensemble of decision trees to attain a prediction on any given example.
- The ensemble is trained by using famous *boosting* algorithm which enables it to find non linear patterns in the data.



Methods - Neural Network

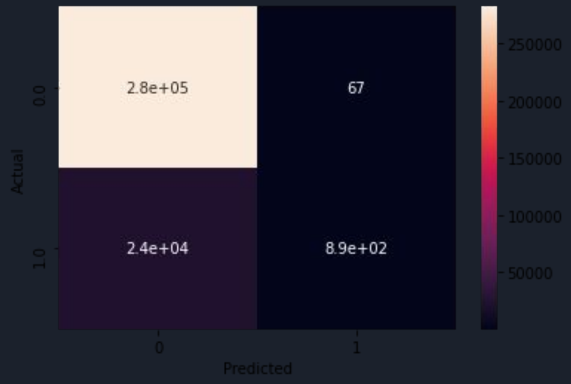
- Deep Neural Nets are very bad at handling discrete data and expect inputs to be continuous features. Embedding layers have therefore been developed which map discrete data into low dimensional continuous spaces.
- Linear methods, on the other hand, thrive on sparse discrete features.

Methods

- TabNet
- Node - Neural Oblivious Decisions Ensembles
- AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks

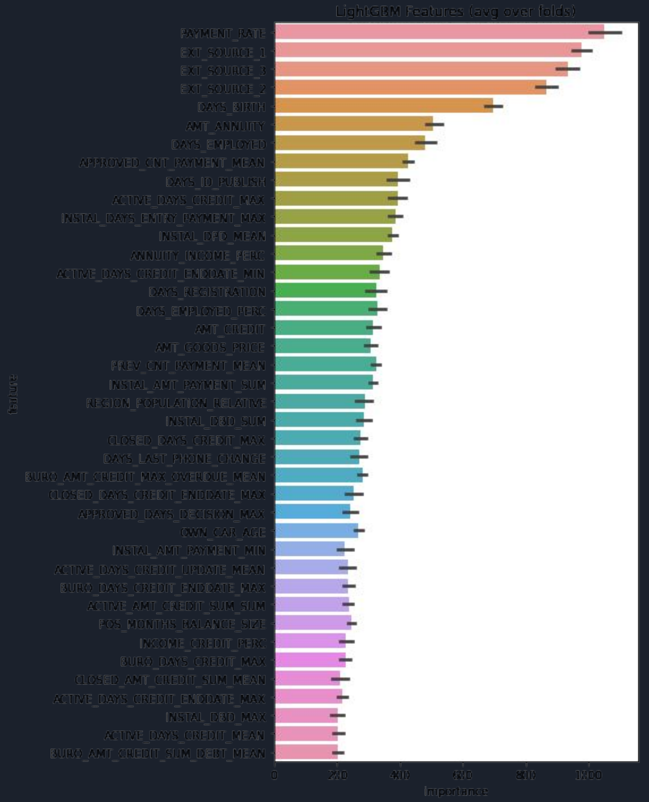
Results - Gradient Boosting Trees

- With gradient boosting trees, we were able to achieve AUC of 0.792 which is quite close to the state of the art.
- To the best of our knowledge the state of the art public result on this dataset has AUC of 0.805.



Feature Importance

- By interpreting the model, we see that model gives high importance to feature such as rate of payments, external information sources on applicant by credit bureau, customer's age, amount of loan, employment information etc.
- These features are also considered important by humans hence verifying that model has indeed captured meaningful patterns present in the data.



Thank You.