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

Muhammad Ismaeel
Information Technology University Lahore, Punjab, Pakistan
msds19029@itu.edu.pk

Abstract

Every large financial institution in the world receives hundreds of application for loan/leasing. By automating the process of loan/lease approval or decline, thousand of man hours can be saved with fast application processing. For this problem we are using home credit dataset from Kaggle competition.

Besides this we are also performing comparison between traditional machine learning algorithms to deep learning architecture to evaluate performance of these models on tabular data. We looked into the problems we face with tabular data in deep learning architectures.

1. Introduction

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.

1.1. 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.

1.2. Challenges

Class Imbalance: As large number of loan applications get rejected, hence, dataset had a class imbalance. Positive examples were only —

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

2. Related Work

To be added later

2.1. Methods

To be added later

3. Proposed Methodology

To be added later

3.1. Gradient Boosting models

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.

3.2. TabNet

To be added later

3.3. 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.

3.4. Neural Oblivious Decision Ensemblers

To be added later

4. Experiments

To be added later

5. Results

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.

6. Conclusion

To be added later

References