TMDB Box Office Prediction with Machine Learning Methods

Ali GUVEN [1] **Muhammet Bugra TORUSDAG** [2] **Furkan KOC** [3]

Abstract

This document handles the predictions of the revenues of the competition. The focus points on applying the methods, which are Extreme Gradient Boosting (XGB), Light Extreme Gradient Boosting (LXGB) and Support Vector Machine (SVM), are the features and selection of them, and k-fold parameter tuning. Feature selection method is used for eliminating the features that have bias. The error is reduced by using the selected features and using k - fold cross validation. The main idea of this document is that comparing the methods with each other and measuring the performance by Root Mean Square Error (RMSE).

1. Introduction

The data set is available on Kaggle web site at:

https://www.kaggle.com/c/tmdb-box-office-prediction/data/

XGB and LXGB are ensemble models and they perform very good in general. They are newer models than SVM. In contrast with SVM, they are very fast. SVM is a method that performs well in classification problems. XGB and LXGB are performs well in both regression and classification problems. Hyper parameter tuning is the key point of good performance of the methods. For SVM, there are two important hyper parameter, which are cost (C) and gamma. For XGB, there are two significant hyper parameter, maximum depth and minimum child weight to avoid to over fit. For LXGB, there are two significant hyper parameter, maximum depth and learning rate, to avoid to over fit. It can be taken the same hyper parameters for both XGB and LXGB, but the most important ones are selected to improve the model. The data set has few features with bias. Hence, at first it needs to be prepared. The first step before training is that feature extraction and selection.

1.1. Feature Extraction

The features that have not a numbers are extracted to having not a number or having number or string value. The features that have specific string values like language, country or etc. extracted into binary values. If they contain the string the feature has 1 else 0. The budget feature and the revenue feature have big weight rather than the others. So that, they are normalized with logarithmic function. Some features have no importance to the model, but they are aggregated with the important ones. Also, adding the new unusual features that are mean of samples, variance of samples in form of logarithmic function and sum of the features of each sample improves the model. Hence, they are used for the model to predict the revenue. Some of the feature extractions are taken from the user[1]. Also using the imdb film ids obtains the imdb ratings and votes. It is obtained from the data set taken from the web site at:

https://datasets.imdbws.com/

1.2. Feature Selection

The features are selected by backward step wise feature selection method. The method is applied by calculating the p-values (under 0.05).

2. Methods

SVM, XGB and LXGB are the methods.

2.1. Support Vector Machine (SVM)

The library in python is used to predict the revenue of the films in the test data. The model is regression model so that it performs bad. The data set has high bias and missing values. Hence, SVM does not converge the data very well.

*Equal contribution [1]Department of Electrical and Electronic Engineering, University of Economy and Technology of TOBB, Turkey [2]Department of Computer Engineering, University of Economy and Technology of TOBB, Turkey [3]Department of Computer Engineering, University of Economy and Technology of TOBB, Turkey. Correspondence to: Ali GUVEN <aliguven@etu.edu.tr>, Bugra TORUSDAG <bugratorusdag@gmail.com>, Furkan KOC <fkoc@etu.edu.tr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Because of its sensitivity of hyper parameters the model created by SVM has high bias and high variance. After optimization of the parameters the model still has high bias and high bias.

2.2. Extreme Gradient Boosting (XGB)

It is an ensemble method including trees and regularization parameters. When creating the model, regularization parameter for L1 is used. As every ensemble model, the avoidance of over fitting parameter is about the complexity of the model itself. Hence, one of the tuning parameter is maximum depth.

2.3. Light Extreme Gradient Boosting (LXGB)

It is an ensemble method including trees and regularization parameters. When creating the model, regularization parameter for L1 is used. As every ensemble model, the avoidance of over fitting parameter is about the complexity of the model itself. Hence, one of the tuning parameter is maximum depth. For difference of XGB, it has trees to ensemble them, but the regularization progress is different from XGB.

3. Experimental Results

The final submission for the competition is obtained by using XGB model. XGB model converges more the data than the others. Kaggle shows the results in terms of RMSE for the comparison of the methods and sorting the users. The importance of the features:

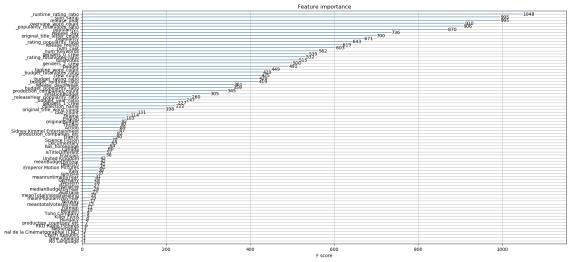

Figure 1. Importance of the features that modeled with XGB

The LXGB model has similar results with the XGB. However, it is made for a competition, small changes or small error decreases are so important. Hence, the importance of the features for LXGB are illustrated in Figure 2.

The RMSE error figures of the best models, XGB and LXGB respectively, are showed in Figure 3 and Figure 4 respectively.

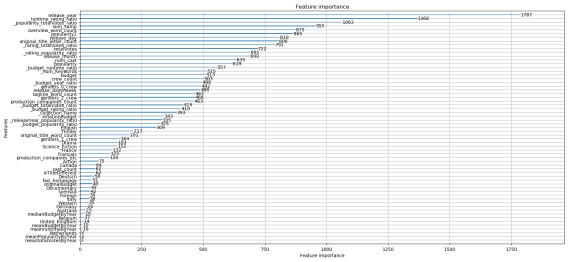

Figure 2. Importance of the features that modeled with LXGB

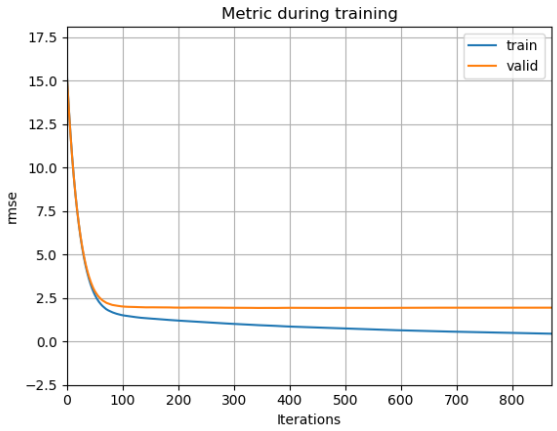

Figure 3. RMSE for XGB

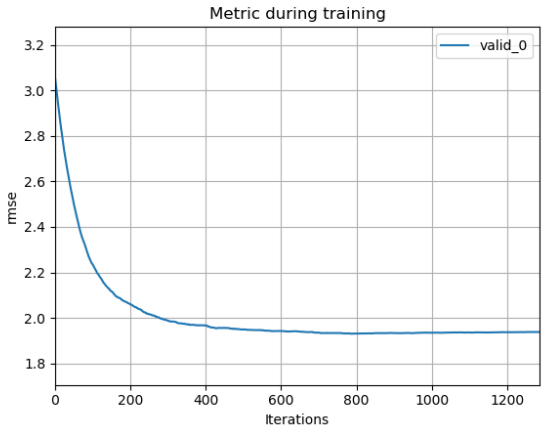

Figure 4. RMSE for LXGB

The RMSE results show that XGB model has lowest RMSE and lowest bias. Variance of XGB model is small same as LXGB.

Now SVM regression results:

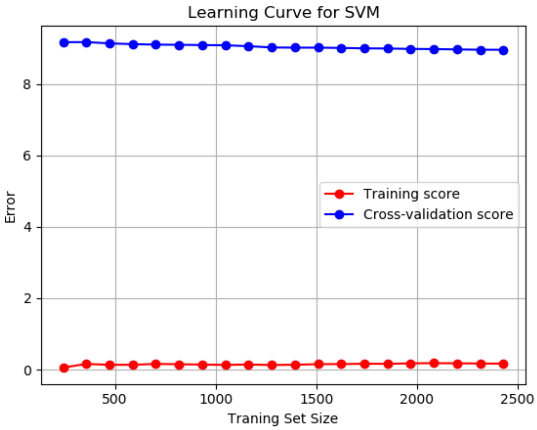

Figure 5. Learning Curve for SVM Regression

In Figure 5, SVM regression model has high variance and high bias although using hyper parameter tuning.

3.1. Tables

As it is seen in Table 1, XGB has lowest RMSE and it is consistent with the training and validation scores. The validation and training sets are separated by k-fold cross validation.

Table 1. Comparison among to methods.

METHODS	RMSE	BETTER?
XGB	1.76128	√
SVM	4.73	×
LXGB	1.763	√

Table contains real results taken from Kaggle.

Conclusion

XGB is the best method among the methods used in this document. SVM regression normally works well in some problems, but the data sets has missing values or not a numbers. LXGB also work fine to converge the data. The model has bias because of the data is not sufficiently large.

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

https://www.kaggle.com/kamalchhirang/eda-feature-engineering-lgb-xgb-cat, Kamal Chhirang