Predict Future Sales

Anshul Shaive, Arun Kumar, Aaryaman Sharma, Adithya Abhishek

Abstract

Predicting future sales is one of the most important aspect of the business world, trillions of dollars are transacted every single day all over the world. Since a lot of money is involved, we need systems to get maximum profit from products sales. Machine learning models are helpful in predicting future sales to get maximum profits. Multiple models are applied among which, we choose the one which gives the lowest RMSE Score. *Random Forest* was the one among many which gave the lowest RMSE value. LSTM is a deep learning model which was also implemented but RMSE score was poor. The dataset is relatively smaller for neural network to perform well.

I. Introduction

THIS era of finance world is still growing because every day new technology is emerging out which gives better accuracy. Since reducing loss is the main aim in finance i.e. max profit sales, everyday new algorithm and model are implemented which improves the overall prediction.

Common challenge we face are the missing values and outliers which are taken care of by different regularization technique. There are many Machine / Deep Learning models which can be applied here based upon the problem statement. We applied Random Forest Regressor, LSTM, XGBoost, neural network out of which random forest regressor outshined them all with lowest RMSE score of 0.901.

XGBoost was second best with the RMSE score of 0.921 then neural network with 0.964 and LSTM with 1.03.

This paper was submitted on 21 of June'19. This work was supported in part by the Bennett University under the supervision of LeadingIndia.ai.

Anshul Shaive in Laxmi Narain college of tech, Bhopal, (e-mail: Anshulshaive321@gmail.com)

Aaryaman Sharma in Laxmi Niwas Mittal institute of information technology (e-mail: Aaryaman09@gmail.com).

Arun Kumar in National Institute of Technology, Jamshedpur (e-mail: Arunsuryan250619@gmail.com)

Adithya Abhishek from St. Joseph's College of Engineering (e-mail: Adithya.abi1123@gmail.com)

Starting with neural network, training it with the train dataset took a lot of time for a small dataset containing few thousand records, as time is a crucial factor, we just can't ignore it. If we scale it up to million shops in few countries, it won't be feasible.

Since data is non-sequential, we did transformation of dataset and later implanted LSTM on dataset. After running on test dataset, LSTM was not giving lower RMSE score hence the model was dropped. XGBoost is a powerful model used with different sets of hyperparameters. Altering hyperparameters was not significant as RMSE score had reached a saturation value.

Random Forest regressor gave the lowest RMSE score. We altered the parameters and have specified the best in methodology.

II. Models

Neural Network is deep learning model in which we specify number of layers and nodes for each layer. We can apply activation and optimizers on layers. Relu activation is applied on layers and Adam optimizer is used with drop out to remove chances of over fitting.

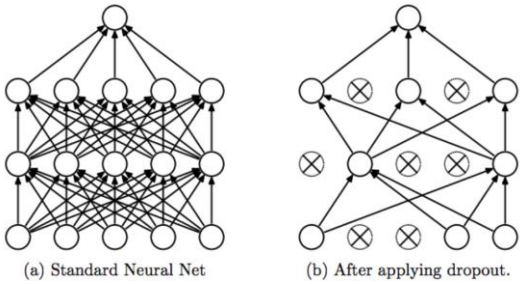

Figure: 01: Neural Network

LSTM works upon sequential data in which it forms sequential interconnected blocks in which every block is dependent upon last previous block.

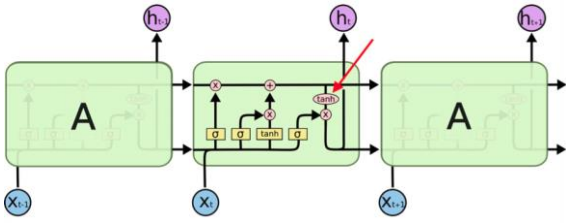

Figure: 02: LSTM

XGBoost is a gradient boosting algorithm. It creates its model by forming multiple weak prediction trees based upon earlier tree and later combing them to form a powerful model.

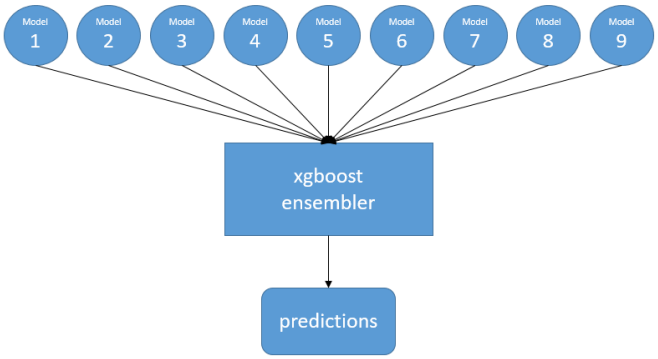

Figure: 03: XGBoost Ensembler

Random Forrest regressor is a machine learning model in which it forms a tree based upon different weights and features, after utilizing all the conditions according to the problem, it provides prediction.

Tree regressor model is efficient and faster but there is not too much modification you can do with it to increase accuracy. Since the dataset is formed upon Russian city shops over periods of years, we can train model upon dataset of other country and language.

III. METHODOLOGY

The different models used are described below:

The number of features selected are 38.

1. Neural Network:

The architecture of the network is as follows-

Layer 1: Dense layer with 20 units
Activation: Relu,
Batch Normalization Layer
Layer 2: Dense layer with 15 units
Activation: Relu
Dropout Layer with rate=0.20
Layer 3: Dense layer with 10 units
Activation: Relu,
Batch Normalization Layer
Layer 4: Dense layer with 10 units
Activation: Relu
Layer 5: Dense layer with 1 output unit.
Activation: Linear

Optimizer: Adam
Loss and Metric: Mean squared error
Number of epochs: 100

2. XGBoost:

XGBoost is an optimized distributed gradient boosting library designed to be highly **efficient**, **flexible** and **portable**. It implements machine learning algorithms under the Gradient Boosting framework.

XGBRegressor is used for the predictive model with the following hyperparameters:

max_depth=12,
gamma=2,
n_estimators=5000,
min_child_weight=1,
objective = 'reg:linear',
colsample_bytree=0.5,
subsample=0.8,
reg_alpha=1,
eta=0.05,
seed=42

Evaluation Metric: Root mean squared error
Early stopping rounds: 10

3. ExtraTreesRegressor:

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

The following parameters are used:

max_depth=20,
random_state=42,
n_estimators=200,
n_jobs=-1

4. Long short-term memory (LSTM):

LSTM is an artificial Recurrent Neural Network (RNN) architecture used in the field of Deep Learning. Unlike standard Feedforward Neural Networks, LSTM has feedback connections that make it a general purpose computer

The architecture is described below:

Firstly, the input is converted to three dimensional so that it can be passed through the lstm layer.

Layer 1: CuDNNLSTM with 8 units and shape (1,38)
Layer 2: Dense with 12 units and relu activation
Layer 3: Dense with 12 units and relu activation
Dropout layer with a rate of 0.20
Layer 4: Dense with 8 units and relu activation
Dropout layer with a rate of 0.10
Layer 4: Dense with 5 units and relu activation
Layer 5: Dense with 5 units and relu activation
Layer 6: Dense with 1 output unit and linear activation

IV. Related work

Ensemble of Regressors:

The models can be stacked together for improving the overall RMSE value.

Other approaches have calculated the average of predictions of XGBoostRegressor, ExtraTreesRegressor, LightGBM but the result had not improved much, and the inference time have been increased significantly.

Different techniques such as cross-validation can be used alongside XGBoost to further reduce the RMSE value.

Some approaches have combined the models such as CatBoost and LightGBM, but the results were not as good as using XGBoost or ExtraTreesRegressor alone.

Some features with lower correlation with the variable to be predicted can be dropped to lower the inference time and reduce over-fitting.

The data can be converted to sequential time-series data and then different models such as UCM (Unobserved Components Models) can be used.

V. Experimental Results

Dataset can be found here: https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data

RMSE is the evaluation criteria used here, with standard equation:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}$$

The rank was determined from 35% of test data for public leaderboard.

38 Features were created from the given CSV datasets. Multiple models were trained on the data.csv file which was extracted by feature extraction from data.pkl.

Models	RMSE
Neural Networks	0.964
XGBoost Regressor	0.926
ExtraTreesRegressor	0.901
LSTM	1.030

Figure 04: Tabulation result of models.

VI. Conclusion

It was found under rigorous testing that Random forest regressor outshined all the other machine learning modes with a RMSE score of 0.901. Since the dataset was small for neural network to be trained and it couldn't give accurate predictions.

Further improvement can be made by altering the hyperparameters. We learned that feature engineering is one of the most important aspect of improving model accuracy.

Important features: -

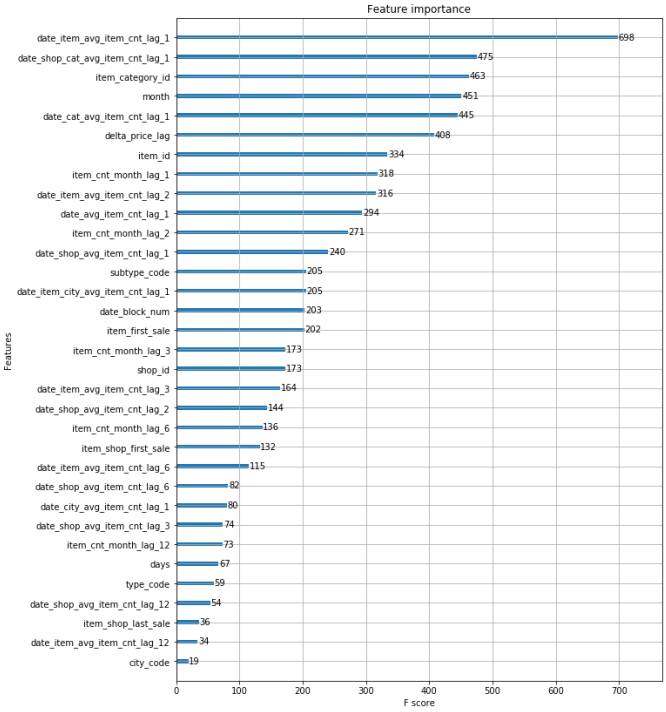

Figure: 05: Feature with significance ranking

VII. Acknowledgment

We thank LeadingIndia.AI for providing an opportunity to work with such challenging supervised problems.

We also thank Bennett University for providing the facilities, and in the end, we thank our mentor Mr. Nithin Prince John for providing his undivided attention toward this project.

VIII. References

[1] https://www.kaggle.com/c/competitive-data-science-predict-future-sales/rules

[2] http://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting

[3] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning

[4] https://towardsdatascience.com/how-to-practice-python-with-google-colab-45fc6b7d118b