Exploiting LightGBM Ensemble Method for Stock Prediction

Vatsal Mitesh Tailor

Abstract— This paper leverages the LightGBM Ensemble Method to predict stock prices. First, the time features are from the dates and these generated features are used to build a regression model. Experiments are performed on the Tesla and the Coca Cola stock historical data to show the effectiveness of the method in predicting stock prices

Index Terms— Stock Prediction; Ensemble Models; Machine Learning; Time Series; Data Mining; Regression; Feature Generation

1 Introduction

Ensemble methods techniques have yielded better accuracy and results than machine learning algorithms in numerorus domains. These models train rapidly as they combine many weak learners and the technique of boosting is used to yield much accurate results. This paper shows the application of one such ensemble learning method, LightGBM to the task of stock price prediction.

The task of stock prediction requires non-linear learning and this application of the model is chosen to see whether such an ensemble model could generate a promising stock price predictor.

Section 2 introduces the ensemble model used. Section 3 showcases the feature generation and the data description part. Section 4 shows the results achieved by the model, and with the proposal of further research, the paper is concluded.

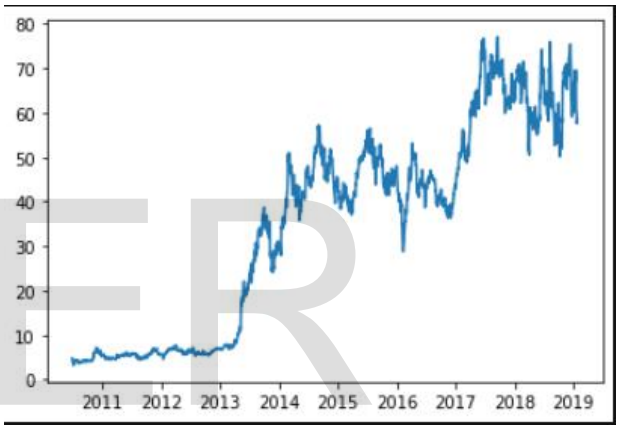

Fig. 1. Daily Close Price Stock data of Tesla (TSLA)

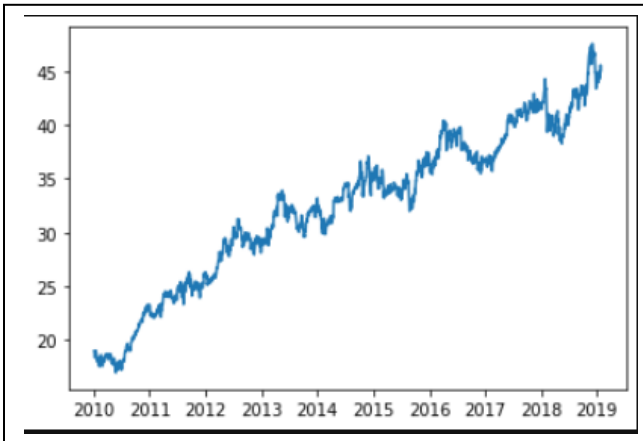

Fig. 2. Daily Close Price Stock data of Coca Cola (KO)

2 Model Description

LightGBM regressor with GOSS as the boosting type is used for the task of stock prediction. Most of the winners of Kaggle competitions are now winning using ensemble models, and XGBoost is one of the most widely-used ensemble models. But LightGBM is fast gaining ground as one of the most popular ensemble models due to its advantages over the XGBoost model such as better predictive performance for the same running time. Thus the LightGBM model was preferred for our task here.

The model uses decision trees as the weak learners and numerous such weak learners are used. The boosting technique weights output from all the weak learners, giving more weight to the better learner, thus minimizing the overall error. And the employment of weak learners ensures much faster training times.

3 Data Description and Feature generation

The daily close price data of the two stocks: Coca Cola (KO) and Tesla (TSLA) from 2010 to 2018. Refer to the Fig.1 and Fig, 2 for the visualized data of both the stocks.

No additional data except the closing stock price is used.

From the date data, the following date features were generated:

- Year
- Date
- Month
- Quarter
- Day of the Week
- Is weekend or not

These 6 features were used as input to train the LightGBM model. The dataset consisted of closing stock price from 2158 days. It was split such that 1956 days were training data and the last 200 days were testing data.

4 Results

The LightGBM model was trained on the above features and the hyper-parameters of the model were fine-tuned to minimize the Mean Absolute Error (MAE) metric.

TESLA STOCK

One can refer to Fig 3 to see the fitting of the model on the training data+training data. Fig 4 focuses solely on the model predictions of the test data.

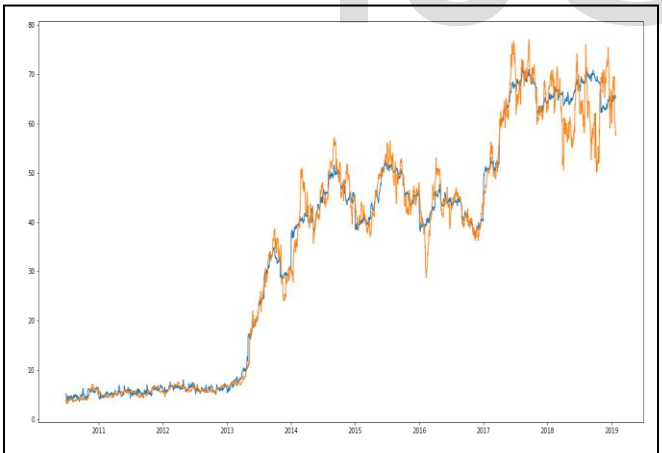

Fig. 3.Overall fit of model output– Tesla Closing Price stock

For the test data, an MAE of 6.583 was achieved, and taking into account the average stock price, one can safely approximate the errors of the prediction to be near to 10%. Refer to Fig. 3 for the visualization of original data (orange line) vs the predicted data (blue line).

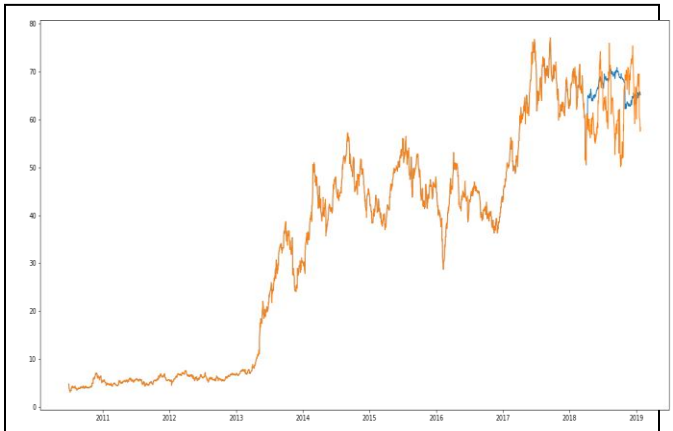

Fig. 4. Prediction comparison on the test data – Tesla Closing Price stock

COCA COLA STOCK

Refer to Fig 5 for the overall fit of the model with the test+training data. Refer to Fig 6 to focus solely on the model's fit for test data.

Fig. 5. Overall fit of the model output – Coca Cola Closing Price stock

For the test data, an MAE of 1.785 was achieved, and taking into account the average stock price, one can safely approximate the errors of the prediction to be less to 5%. Refer to Fig. 3 for the visualization of original data (orange line) vs the predicted data (blue line).

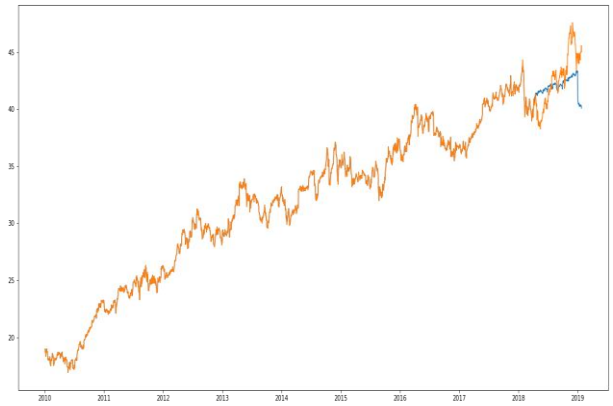

Fig. 6. – Prediction comparison on the test data – Coca Cola Closing Price stock

The above results show that while some immediate peaks cannot be predicted in the test data, the error that was achieved by the model solely from the historical data is very less.

5 Further Research

While the model takes into account only the historical data, models that take into account news and other factors that might change stock prices might be incorporated in the future, which might increase the performance. Further, training of a hybrid LSTM – LightGBM model is to be undertaken in the future for better regression results.

6 Conclusion

In spite of taking into consideration only the historical stock data, the error that was achieved by the model was very low. One must take into account that numerous other factors such as news, traders' instinct affect the stock price significantly. Thus, the results that were achieved by our regressor are great and can be combined with other models that incorporate other data that affects stocks.

References

[1] Y. Freund, R.E. Schapie,"A short introduction to boosting",Journal of Japanese Society for Artificial Intelligence, 14(5):771-780, September, 1999

[2] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen,W. Ma, Q. Ye., T. Liu,"LightGBM: A Highly Efficient Gradient Boosting Decision Tree", 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[3] Ketikakurita, "LightGBM and XGBoost explained", https://mlexplained.com/2018/01/05/lightgbm-and-xgboost-explained/. 2018