LiBerTY – Store Sales Forcast with Machine Learning

Suhas Anand Balagar* Hardy Leung† Loukya Tammineni‡ Xichang Yu§

November 2022

Abstract

LiBerTY is an ensemble regression engine to predict the store sales given. It employed a variety of data engineering and machine learning techniques to most

Introduction[1]

The ability to predict the sales of a variety of stores is highly sought out in supply chain logistics, as it finds applications in increasing customer satisfaction and reducing food waste. We are proposing use of multiple Supervised Learning methods to predict the sales of stores based on time series dataset of Corporación Favorita, an Ecuador based grocery retailer. Ecuador is a country whose economy is strongly dependent on the oil and fluctuates with the price of oil.

We are planning to use dataset from an ongoing Kaggle competition [1], "Store Sales – Time Series Forecasting". The dataset includes multiple csv sheets of time series data. We will try to evaluate the different aspects that might impact the sales in a store like Holiday seasons, Oil prices and historical sales data from a variety of stores. The preprocessing of data will include checking for missing values and if found, imputing them so that no data is disregarded. We have also checked for frequency distribution of data elements as part of Data exploration. We have calculated and plotted correlation mapping to establish the correlations between different input and output parameters such as holiday events, oil prices, transactions per store type, dates of salary, natural calamities etc.

We plan on using different Supervised Learning models to predict the prices and then evaluate which of those models give out the best results. Supervised learning approach works best in this case, as we have huge time series data of both input and the output parameters mentioned above. We applied several transformation and optimization to improve the data quality in preparation for the optimization. To seek the best store sales prediction, We have evaluated several models including linear regression, Gradient Boost (XGBoost), Light GBM, Random Forest, Support Vector Regression (SVR), and Long Short-Term Memory (LSTM). We found XGBoost to be the best- performing individual method, and focused on hyper-parameter tuning via grid search. We further employed ensemble prediction to further improve our results, achieving a notable RMSLE score of 0.425 within our compressed project time-frame.

Related Work[2]

Cite reference to these models.

XGBoost [2] is a popular open-source software library which provides a gradient boosting framework for C++, Java, Python, R, Julia, Perl, and Scala.

LightGBM [3] is a gradient boosting decision tree (GBDT) implemented by Microsoft which focused on speeding up the training process of conventional GBDT by up to 20X while achieving almost the same accuracy. The huge runtime efficiency makes it a popular GBDT technique in recent years.

*San José State University, suhasAB@github
†San José State University, ksleung@github
‡San José State University, LoukyaTammineni@github
§San José State University, Codyyu36@github

[1] Suhas's section

[2] Hardy's section, and others

- Linear Regression
- Gradient Boost(XGBoost)
- Random Forest
- Support Vector Machine(SVM) /Support Vector Regression (SVR)
- Long Short-Term Memory (LSTM)

Data Preparation[3]

Experimental Setup[4]

Talk about the experimental setup, including how to

Part I[5]

- Linear regression
- XGBoost
- LightGBM
- Grid search technique to improve XGBoost performance

Part II[6]

- LSTM, architecture, slightly different flow

Part III[7]

- Ensemble approach. You can quote this [4]
- Final result
- Visualization

Discussion[8]

- Possible enhancements

[3] Hardy's section
[4] Hardy's section
[5] Loukya's section
[6] Cody's section
[7] Suhas's section
[8] TBD

Conclusions[9]

In this work, we have presented LiBerTY, an ensemble regression engine that successfully predicts the store sales, which successfully employed a variety of data engineering and machine learning techniques.

References

[1] "Store sales - time series forecasting," *Kaggle*. [Online]. Available: https://www.kaggle.com/competitions/store-sales-time-series-forecasting.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[3] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, 2017, vol. 30, [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[4] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review*, vol. 33, no. 1, pp. 1–39, 2010.

[9] Hardy