Sales Quantity Predictions on Item Level for a Retail Store

Project Summary for CEU Business Analytics MSc

Nawal Zehra Hasan

June 2022

Contents

Introduction

Historically, forecasts have been done using different statistical methods and experience. However, with the advances in machine learning this task has become more automated that ever before. It can enable business to get rid of having multiple files to store and process data. The models can provide consolidated information available and with the help of visualizations tools the results can be easily understood. The aim of my thesis was to utilize data available from 2014 till 2022, aggregated on a monthly level and investigate which machine learning model could forecast sales quantities the best. The models include XGBoost, SARIMAX, LSTM, and Facebook Prophet and Naïve model. Overall, XGBoost and FBProphet performed the best and had lower Root Mean Squared Error as compared to the rest. The final table with the predictions for all the items in the dataset was used to create a visualization using SQL dashboard in Databricks.

Research Objective

My client is a promiment player in ther retail industry running supermarket chains in many countries. They were given this project as a request by their Goods Logistics Department whose role is to secure quantities of goods in the warehouse and dispatch accordingly to the

stores. The primary objective of this project was to design a data product for the client which allows the user to follow the actual state of delivered quantity of the effected items and give predictions for the next 12 months. The result of the project being able to automate the process which was previously done using several sources including excel files and manual processes. These items are ordered for the next year and manufactured by the client as well.

Data Preprocessing

As this is a time series prediction problem, I used historical data available on Databricks from 2014 till data. The data was receipts data for daily sales, and included several details about the items and the sales made. I also added items data using which I choose the given list of items to make the predictions on.

The receipt data is the main data source for this project. The data shows sales of all the items sold by the business in each store within Hungary. It includes data from 2014 and updated daily. This includes several columns. However, not all columns are relevant for this project. I also had to ensure to not include duplicate scanned items to avoid over counting sales quantities. Certain items were also returned by customers, for reasons other than duplicate counting.The items are either sold by weights or by pieces. This needed to be sorted and cleaned. I aggregated monthly sales using the daily sales and sorted them by the date. At the end I selected the dates. Since I have monthly sales quantities and the data available is from 2014, I included every month starting 2014 till May 2022. The maximum number of observations I have for one item would be 101 which includes all the previous months up until May 2022.

The items data was another important source of information. I was given a list of 74 items which included ice cream and chocolates only. Using the items data, I filtered all the information for these items which included the date from which these items were available from, the food category they belonged to, the item number and names. After this I merged this data which the receipts data.

Prediction Models Evaluated

The chosen models include Naïve model, LSTM, SARIMAX, Facebook Prophet, Holt Winter and XGBoost. Initially, I did not plan on using XGBoost as I was unfamiliar with it but as suggested by the client for improved model performance, I did incorporate it and it did indeed result in better predictions. The model evaluation metric chosen was Root Mean Squared Error due to its easer of interpretation and a sound method to compare models and choose. However, one drawback that I must mention here that it does not truly reflect how good a model is on its own without comparing it to other models because it is merely a number.

Each of the models were evaluated on the test set which was 29 observations. XGBoost resulted in the lowest RMSE of all the models followed by FBProphet. Hence, I went ahead

and used this model to carry out predictions for all the items in the data. A possible reason for imrpoved performance of FBProphet could be related to the model incoporating seasonality effects, in our case yearly seasonality. XGBoost perfomed slightly better than FBPorphet perhaps due to the non-linearlity of the model where the model could capture non-linear relationships within the data which could be relevant as retails sales tend to fluctuate heavily due to the multiple complex variables that affect sales patterns.

Conclusion

This thesis impleted five supervised maching learning models on a specific case for a client from the retail industry. Given this case and the data at hand there were no major differences in the predictive power of these models. However, XGBoost and FBProphet showed promising results. The expections of the clients were met insofar as consolidated information about items is available to them. With some fine tuning of the hyperparameters the results can certainly be improved and the predictions can be even more accurate. The Naïve model performed better as compared to the supervised machine learning models which may be an indication that simpler models perform better and there are other exogenous factors that affect this list of items sales such as weather. However, it must be noted that the Naïve model is independent of time while an experience-based forecasting take into consideration time as a factor.

Further Studies

The client can benefit using these prediction models not just for this list of items but any items that they would like to check. Using prediction models they can determine their best and worst selling items and perhaps make changes to their marketing and sales techniques including price, packaging and availability for these items. The models can further be improved by adding weather in Hungary. Furthermore, the predictions could also be improved by checking why the models are unable to capture the peak sales. My understanding was that there are other factors that were not accounted for and perhaps as a retail business, the data science team has more information regarding this. To prove this, I ran my models for some randomly selected item to see how the model performs for items such as coke and water. These are also items with some seasonality and are sold more in summers as compared to winters.