Daily Sales Quantity Forecasting for Employee Performance Planning in Retail Industry

Capstone Project Summary

This project was conducted to generate a three-week, daily store-level sales quantity forecast for an international discount retailer chain that has several stores across Europe and the United States.

The purpose of this project is to enhance the quality of end to end employee performance planning and resource optimization in each store separately. An accurate estimate of the sales quantity for the next three weeks will improve the performance management process for the retailer. An effective performance management process will produce high-performing workforce that can offer better customer service and improve performance at retail stores.

Currently, the performance of the stores is based upon historical quantities of items sold by the retailer in each store. To increase revenue and remove wastage of resources, a system is needed that incorporates customer behaviours and habits for accurate forecasting. This will allow the client to make informed decisions and properly manage their inventory and human resources. Since current systems use historical data for performance measurement, they can not incorporate the shopping habits of the consumer and hence a lacuna exists that does not allow optimal management of inventory and labour. Due to a crunch of skilled labour, we need a system that manages them when the rush of customers is high and suggests a decrease in workforce when not many consumers are expected to enter the stores. An accurate estimate of the quantitative sales, along with optimized workforce and labour can result in increased revenue and less wastage of resources.

The outcome of the project is a dashboard which allows the store managers to view and download the forecast in the form of a CSV. The aggregation is on the level of a store. The user of the dashboard is given the option to filter on date-interval, and individual stores. The output is in the form of a notebook that is rerun daily or as per the requirement of the client.

The project was developed and deployed on the Databricks platform for automated cluster management, and big data processing. The Databricks platform is used as an end-to-end data pipeline in this project. Pyspark and Python are used as the programming languages. Data is

read from the delta lakes and the spark file store. Several packages and libraries are used throughout the project, predominantly Pandas, Pyspark, NumPy, Scikitlearn, Matplotlib, and Statsmodel.

In this project, three supervised machine learning methods are used to model the daily sales quantity for the next three weeks for each retail store. The client provided historical receipt data for all the stores present in the HU(Hungary) region. Three and a half years of the most recent data was chosen to perform the modelling. The dataset consisted of past receipt data of all stores in Hungary from January 2019, and calendar data which contained the national events and their respective dates on which the stores remain closed. Next, data cleaning and transformation was performed, and it was aggregated to obtain the total daily sales quantity of each store. The time series data was analysed to discover trends, seasonality, stationarity, noise and hidden patterns to help fit a model which captures its trend accurately.

In the modeling phase of the project, three models namely Seasonal Auto-regressive Integrated Moving Average with eXogenous variable (SARIMAX), Facebook Prophet (FBProphet), and eXtreme Gradient Boosting (XGBoost) were fitted on the data and their hyperparameters were tuned to minimize the improve the forecasting performance. The modeling and tuning of the models were performed in the following order: FBProphet, followed by SARIMAX and XGBoost. Facebook Prophet was the easiest to model to implement in terms of complexity since it didn't require any feature engineering and extensive tuning. It is fast and provides completely automated forecasts. It is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with our time series data because it has strong seasonal effects and several seasons of historical data. FBProphet proved to be robust to missing data and shifts in the trend, and typically handled outliers well. It automatically found seasonal trends and its parameters were intuitive and easy to tune. It performed the second best with a mean absolute percentage error.

Sarimax was the most time consuming to model. After determining the stationarity of the time series, the Auto Arima function was called to provide a starting point for choosing the input parameters to the model. However, the best p, d, and q values were chosen by trial and error. Holidays were modelled separately as exogenous variables in SARIMAX. This model did not perform well comparatively.

XGBoost showed quite promising results on our time series data. It trained gradient boosted decision trees. Various features were created manually such as calendar, holidays, near holiday dates, and day of the week as input to the model which improved the accuracy of the model significantly. XGBoost performed better than both SARIMAX and FBProphet, with the lowest mean absolute percentage error.

A comparison was made between the model outputs and their accuracy on the test dataset was visualized. The comparison metric was the Mean Absolute Percentage Error (MAPE) and the Mean Absolute Error (MAE). The XGBoost model performed the best amongst all the tested models. It had the lowest mean absolute percentage error in the range of 3 to 10 percent for maximum retail stores. Its error was above 15 percent for only 4 stores out of the 197 input stores. The cause of the incorrect forecasts was investigated, and it was found that the 4 stores had been closed permanently and thus do not have the most recent receipt data. FBProphet performs the second best with a mean absolute percentage error in the range of 8 to 25 percent. SARIMAX performs the worst with the mean absolute percentage error in the range of 14 to 30 percent. It was concluded that XGBoost performed the best amongst all models because it fit the data best for all the 197 stores. The 21-day future demand forecasts were written in a database table and then exported to a SQL dashboard with a store selector drop-down feature within the Databricks environment.

Lastly, more features can be added to the modelling which affect the purchase pattern of the customer to further improve the performance of the model. Incorporating more features like marketing campaigns, discounts, days following the holidays can greatly improve the performance of the model.

Most of the errors in the prediction are made around the holiday dates. This time series data can also be modelled without including the holidays. The holiday feature can be excluded from the modelling which can improve the performance of the model. The store managers know about the holiday in advance and thus do not necessarily require a sales quantity value on the dashboard for holiday.

The results were presented to the client along with explanations of choices made in the project and the client gave positive feedback and agreed that all the project requirements are met aptly.