Predicting the Number of Open Cashiers on an Hourly Rhythm: A Time-series Forecasting Analysis

Capstone Project Summary for CEU MSc Business Analytics 2022

By Haaris Afzal Cheema

Contents

Description of the project

This analysis was done for a major company that operates in the retail industry. The company store managers needed a forecasting tool that would help them know about the expected number of customers in the store on an hourly basis, as well as the required number of open cashiers. This was done primarily to schedule employee tasks based on low requirement hours and hence avoid disruption in their tasks. Secondly, this was also essential for them to avoid long customer queues and dissatisfaction. Therefore, this project aimed to formulate time-series forecasting models which would help them with the points mentioned above.

Summary of the work done

The data was accessed and analyzed on a cloud-based tool known as Databricks. Firstly, due to the sheer volume of the data, several data engineering tasks needed to be performed before I could proceed with the analysis. The data was loaded as a spark data frame and a lot of filtering and aggregation had to be done via SQL. This was a step that was not anticipated since I had not been exposed to such a large volume of data. Several levels of aggregation had to be done until the spark data frame finally reached a manageable size. After that, the spark data frame was converted to a pandas data frame and further data cleaning and munging tasks were done, such as checking for missing values, converting the date variable to a date-time format, and extracting relevant metrics such as hour of the day, day of the week and so on. Next, exploratory data analysis was done where the distributions for the number of customers and the number of open cashiers were analyzed based on different units of dates and times. Store opening and closing times were looked at, and any unusual values were filtered and analyzed to see if they were erroneous or whether they were simply unexpectedly high or low values. Once I got well versed with the data, I proceeded with formulating the prediction models.

Before the model building stage, the final step taken was to check for the stationarity and trends in the data. For this, an augmented Dickey-Fuller test was carried out for both relevant variables and it was concluded that customer distribution, as well as the distribution for the open cashiers, were stationary. Because I wanted to test out a multivariate time-series regression model too, several models were first built for the number of customers. Models of varying levels of complexity were tested, ranging from simple moving averages to Holt Winter's seasonal exponential smoothing method, Seasonal ARIMA models, and Facebook's open-source prophet algorithm. Similar models were tested for the number of customers as well as the number of cashiers as their respective distributions seemed to be very similar. Based on the visual representations I was able to establish that in both cases, the moving average models outperformed the more complex models. Due to the similar results for both our variables, the final model which was tested was a multivariate forecasting method called vector autoregression. The models were evaluated based on the root mean square errors between the actual values and the predicted values. It was found that the 2-point moving averages had the lowest RMSE in both cases. The SARIMA models and Fbprophet did not work out well due to unequal daily store hours during the week, hence causing difficulty in specifying the seasonality in the models. The multivariate regression had reasonable results in the case of predicting customers, but for the main problem of predicting the required number of open cashiers, it performed better.

Key outcomes

It was found that customer distributions, as well as open cashier distributions, had two peaks daily, like two normal curves. The increases as well as the decreases generally were not severe but were rather gradual. Moreover, it was found that during the starting hour of the day, as well as the last hour of the day, the number of customers and the required open cashiers was fairly low. This was why the 2-point moving averages performed very well in this prediction exercise. Even for predicting the unusually low value for the start of each day, the last value of the previous day would help to lower the prediction, keeping the error at a minimum.

Another key outcome of this analysis was that due to the nature of the data at hand, the more contemporary forecasting methods did not perform well. To elaborate, the selected store had the same operating hours for six days of the week whereas the seventh had two lesser hours. As our data had daily seasonality, this became a problem in the modeling stage. Algorithms like the Seasonal Arima and the Holt Winter's seasonal exponential smoothing require you to specify the seasonal periods in the data, which in essence indicate the frequency with which the seasonality is observed. Ideally, if the store timings were the same each day, this could be set equal to the number of operating hours. Since this was not the case, sub-optimal alternatives had to be looked at such as grid-search tools and some manual trial and testing. Due to this, these methods did not produce the high-precision forecasts that were expected initially. However, since the simpler models had very low RMSEs, the performance of the more sophisticated methods did not matter much.

Benefit to the client

The store managers could obtain some very valuable insights into their objective of scheduling tasks at their respective stores. The data explorations done on different time metrics, coupled with the prediction models would allow them to gauge the number of customers as well as the required number of open cashiers with a high degree of accuracy. Moreover, while this analysis was done by subsetting for a specific store, they could easily reproduce the same analysis for another store, by simply changing the store ID and the opening and closing hours if needed. As the notebook was integrated as part of Databricks, the client could easily access the script and in the future, if they need to tweak these models further, they could easily do that as per their requirement. In case the client decides that equal operating hours can be looked at for these tasks, simply by adjusting the hours, they can obtain even higher precision models which will further aid them in scheduling their tasks and optimizing their store processes.

Learning and experience

This project was a very fruitful experience in terms of learning. Firstly, a very detailed introduction to the business processes and the organization were given along with a couple of field visits. These helped me to understand the context of the problem at hand and were very useful in identifying relevant variables for the analysis. I also learnt a lot specifically about the retail industry, the extent to which planning and coordination are required in the daily operations of the store, and how various variables tend to impact each other in a store setting. Secondly, from an analytics perspective, this was also a very rewarding experience. This was my first exposure to doing a project using a cloud-based tool, and I was able to get hands-on with many aspects of it. Moreover, it was also the first time I worked with big data. Initially, it was a major struggle and I constantly had to restructure my data processing and aggregations to optimize the time it took to run the script. This project also enabled me to learn a lot of time-series forecasting methods and how this type of modeling is different from cross-sectional modeling. I was also able to strengthen my grip over the Python programming language as well. As this was a full-fledged analytics project, tasks from data cleaning, munging, explorations, and visualizations to statistical modeling and evaluation, all were done in Python. From a data science perspective, a key learning for me was that I did not need to use the most sophisticated methods to obtain the most accurate predictions. Initially, I wasted a lot of time looking for complex time-series forecasting algorithms and in the end, the most basic model captured my data the best and produced the best forecasts.