A Three-Year Sales Forecasting for a Leading Retail Store

Capstone Summary by Ali Hasnain Khan Sial

Introduction

This document encompasses a summary of a time-series analysis performed on a retail store chain in Hungary. The prime objective of the project is to predict 3 years sales of each store the client owns. Data sources, approaches and models used in this project will be briefly explained in this document. The second objective of this project required research to find external variables that can be used to help improve existing or new sales forecasting models. Information regarding the discovery of the new variable have also been included in this report. Data cleaning and modeling was carried out on Azure Databricks and the main coding languages used were Apache Spark and Python Programming.

Business Problem

The client focuses greatly on one of its key values, which is to improve performance and eradicate wastage to grow and counter modern day challenges that hinder success. Thus, the nature of this project is less of a direct problem, but rather research and development that can help the organization identify hidden predictors for sales forecasting. The main deliverable of the project is to design a robust monthly sales prediction model that can be reproducible across all stores in Hungary. The second objective, as stated above, deals with exploring and identifying exogenous variables that tend to have an impact on sales.

Why am I trying to solve this problem?

In the past couple of years, we have witnessed a huge transformation in the global economy and in the purchasing pattern of customers. The features that used to have an impact on the buying behaviour of the customers might not perform well in the future, especially in this ever changing and fast paced economic environment. Thus, for a business that has a large-scale operational setup, the need to regularly update its processes is great. Moreover, sales forecasting is an important cornerstone of revenue operations for any business. It assists the finance team with streamlining cash flows, operations, preempting credit and financing requirements along with site planning and creating logistical strategies.

Data Resources and Preparation

The main data source for this project was the client's in-house database. The database has various tables, but the two tables primarily used in the analysis are daily sales data and calendar data. A new variable was discovered during the research phase which was directly obtained from an external source, Eurostat (it is the statistical office of the European Union). The daily sales data became the foundation of the entire analysis which was aggregated to calculate monthly sales value for each store. The calendar data had to go through similar a transformation to create monthly calendar variables. Using the calendar data, the variables that were created are seasons dummies and 9 numerical counts variables (e.g., total number of event days in a month, total number of days the store was closed or total number of national holidays in the month). The external variable that was discovered during the research was also a monthly time series data. It is an economic indicator that gives comparable measures of inflation and the change over time in prices of consumer goods and services. For confidentiality purposes the variable will be referred as 'X' in this document. Two additional variables were also computed using the variable X. The additional variables recorded the first month change in value and the twelve months change in value of the original variable.

Time Series Modeling and Evaluation Metric

To begin modeling, the monthly sales data was filtered by selecting one store owned by the client. There are a lot of open-source time series techniques available that can be used to build a sales prediction model. Eventually, the approaches selected and finalised were the ones that could be tailored to perform best for the type of data being used in this analysis. In total, 3 overarching time series approaches were used for prediction which are, Seasonal Auto Regressive Integrated Moving Average Model (SARIMA), Facebook Prophet Model and Extreme Gradient Boosting Method (XGBoost). Using these approaches, various univariate and multivariate models were built. The main evaluation metric that was used to measure the performance of a model was Mean Absolute Percentage Error (MAPE), which highlights the overall error in the predictions. When building these models, for every approach, the most basic model was a univariate model and then gradually external regressors were added to check whether it improves the performance of the model. The idea behind this modeling technique was to ensure that, alongside building a robust sales prediction model at store a level, external variables that are important and tend to explain some degree of variation in sales could be identified.

Model Selection and Predicting the Future

Multivariate models such as the Prophet Model (with all versions of variable X as external regressor) and XGBoost (with all 28 exogenous variables), performed slightly better than the best performing univariate model. Due to the limited scope of this project, it was decided to use Univariate Prophet Model to predict future sales. Even though, it was a univariate model, there are few built-in capabilities that can be tuned to perform as a multivariate model. To ensure the success of this model it was used to predict 3 years sales (till Dec 2025) for various stores of the client across Hungary. The model performed well for the stores that had more observations compared to stores that were new. Overall, the MAPE for this model was between 3% to 10%.

Limitations

The major limitation of this project was researching for an external data source, which was not only time-consuming but difficult. Finding economic variables that are available on a monthly basis was not an easy task, most of these variables are only available on a yearly basis. Moreover, having limited number of observations in the dataset, which is a monthly time series, was also not very helpful in improving the performance of the models.

Conclusion

The primary purpose of this project was to develop a robust time series monthly sales prediction model that can be reproduced for all the stores of the client in Hungary. As stated above, a model that best suited the client's requirement was used to predict future sales. The second objective of the project was more inclined towards research to find new variables that can be used to improve the model's performance. The discovery of variable X was probably the most important finding of this project. It is a new variable that had not been previously used by the client's data science team. Therefore, it is not wrong to say that the analysis performed was able meet the and exceed the expectations the client had from this project.

Recommendations

The main recommendation made to the client was to use multivariate prediction models incorporating future values of important external variables which was beyond the scope of this project. Secondly, use daily sales data to develop forecasting models and then compare the results with monthly forecasting models used in this analysis. This might offer a new perspective and more flexibility in terms of calendar variables in comparison to monthly data.