Capstone Project Summary

Traffic Forecasting of a Hungarian startup

Thesis by Attila Serfozo

MS in Business Analytics

Supervisor: Agoston Reguly

June 2021

In this project, I am working closely together with a Hungarian startup (hereinafter referred to as the "Company"), which operates in the financial markets industry. The aim of this collaboration is (i) to build a model capable of forecasting the traffic of the Company 3 months in advance, (ii) search for the top stock market related variables which movements affect the traffic of their website and (iii) investigate the connection between the revenue and the traffic of the Company.

The Company decided to look for possible modeling solutions, because their currently applied approach tends to be inaccurate. Furthermore, during the Covid-19 Pandemic the website of the Company became quickly popular due to people's increased interest in their service. As a result, it got even more indispensable to forecast the traffic of the website more accurately, in order to adjust business decisions and support further growth according to the expected site traffic. An accurate forecast could not only help the general business intelligence of the Company, but also the Content and User Acquisition team could receive useful information about when to increase site visibility and launch new campaigns.

To complete all goals, I needed to prepare the available data ready for analysis. I had 3 data sources, the website related data from the Google Analytics platform of the Company, the monthly revenue data received in an excel file, and the data of the stock market variables collected automatically by code via the tidyquant R package. In the end of the data preparation exercise, I finished with two datasets. The first one includes the monthly website traffic and site technical indicators data in addition to the changes and volatility of market variables. The second consists of the revenue and website traffic of the Company. Before I conducted the modeling exercises, I also checked the stationarity of the variables. Based on the Phillips-

Perron unit-root test, both the traffic and the revenue variables should be converted to the first level of integratedness (changes of traffic and changes of revenue) for modeling purposes.

In the Variable Importance section, I created two different models to search for the top 5 most important variables affecting the website traffic. I created models instead of checking the pure correlations as I wanted to receive a deeper understanding of the connections. The first model was a lasso model. Based on the coefficients, the variables which have the largest effect on the traffic changes of the website are the Nasdaq, Dow Jones, Crude Oil, DAX and FTSE100 indices volatilities. I also created a random forest model to check whether it presents the same outcome. Surprisingly, the top variables according to the random forest model are the Gold volatility and the changes of DAX, S&P 500, Dow Jones and Nasdaq indices. Even though, the two models result different outputs, it is an interesting outcome that the DAX and Nasdaq indices appeared among the top variables by both models. This indicates that they are important predictors of the website traffic.

In the Traffic forecasting section, I created 7 different ARIMA models to forecast the traffic changes of the website. For model selection I used AIC and forecasting RMSE, but I did not use cross-validation due to time constrains. The final model decision pointed to Model 6, an ARIMA(0,1,3)(2,1,2) model, regressing the website traffic changes on the Covid-19 binary variable and the Nasdaq, Dow Jones, Crude Oil, DAX and FTSE volatilities. The selected model had a relatively high, 187,000 forecast RMSE, meaning that it has missed the traffic in the forecasted period on average by 187,000 session numbers on a monthly base. However, it is important to highlight that the period (between February 2021 and April 2021 used for the holdout set) was one of the most volatile in the website history. The predicted time horizon not only included the effect of the GameStop scandal, but also the third Covid-19 wave has started in March.

Finally, during the Revenue analysis I examined the connection between the changes of revenue and the changes of traffic. Based on the results, traffic changes have a significant effect on revenue changes of the Company, and on average every third visitor of their website generates 1 EUR revenue. Moreover, the models also highlighted that the extreme events of March 2020 and March 2021 impacted not only the traffic changes variable, but they also had an additional effect on the revenue changes. As a result, they caused large one-month extreme values in the website traffic and revenue.

The final regression models indicated possible characteristics of the user journey of the website. The 1st, 4th and 6th month lags of traffic changes signaled strong correlation with the revenue changes. These values may mean that in most cases users turn into revenue by the end of the upcoming month. However, in some cases the user journey can last 4 to 6 months before the user occurs in the revenue. The final model expands the previous one by taking into account the cumulated effect of traffic changes. According to the results, on average every 7th visitor of the past 12 months generates an additional 1 EUR revenue in the current month for the Company.

I gained a lot of valuable experience during the project. I had the opportunity to get an insight in the life of a consultant and practice my communication skills during the informal meetings with the client. I led weekly catch-up calls with the client which taught me how to prepare for, write notes during and lead business meetings. Furthermore, I also gained professional experience in the data field, as I needed to create a product in R concentrating on reproducibility and transparency. I learnt how to conduct an analysis step-by-step to achieve the key outcomes. I also had the opportunity to practice time series forecasting using ARIMA models, which knowledge was missing from my professional portfolio so far. Finally, before this project I lacked the overall picture about when the analyst has to use unit-root and serial correlation tests and did not feel confident about interpret statistical outputs. As a result of this project, now I have the basic understanding of how to conduct such exercises and how to interpret them.