Forecasting Real Estate Market Volume

a Capstone Project Summary

by John Joyce

Submitted to

Department of Economics and Business at Central European University

In partial fulfilment of the requirements for the degree of Masters of Science

Budapest, Hungary

2021

Table of Contents

Forecasting Real Estate Market Volume

by John Joyce

About the Client

The client operates a marketplace platform for real estate and related services in Hungary. Most revenue comes from advertising fees associated with listings on the platform. The client's business is impacted by seasonal fluctuations as well as broader changes in real estate market trends. Because the client is already a dominant player in this space, the key to boosting profit comes from greater operational efficiency rather than expanded market share.

About the Project

The client would like a model to forecast monthly transaction volume in the Hungarian real estate market. Due to issues with the quality of official data sources, the project requires the development of a novel data source which can be used to model transaction volume. The client wants the model to exist in an easy-to-use software tool, so that any member of the organization can access the forecast.

Acquiring the Data

Official data sources regarding real estate transactions are not suitable for this project. Unfortunately, the official registration of home and apartment sales is quite slow, which means that this information only becomes available far past the window in which it could be useful for short-term forecasting. Additionally, historical data provided by the Hungarian Central Statistical Office is aggregated at the quarterly level. Because the client is interested in monthly projections, the predictive model requires training data that exists at the monthly level or smaller.

This project identified an alternative, quantitative indicator which is highly correlated with real estate transactions. On a quarterly basis, this alternative indicator shares a 0.99 correlation with the number of properties transacted. Information related to this indicator is updated frequently online.

The relevant information required to track this indicator is not available in a format which is simple to access. The website which hosts the information is designed to support one query at a time, and each query requires the users to enter specific information into text boxes. In order to build a full historical dataset related to the indicator of interest, a user would need to make more than one million individual queries.

To acquire the data, this project created a web scraping script which constructs queries and retrieves information automatically. For initial data collection, the script was only run during nighttime hours to avoid interfering with normal website traffic. (A large number of requests sent in rapid succession has the potential to slow down access to the site for other users.) Due to the large number of records to be retrieved, the process of collecting all relevant historical information lasted several weeks. In the future, the client may run the scraping script on a monthly basis to acquire new data. Monthly data collection will likely requires less than one hour and requires no human supervision.

Modeling the Data

The forecasting model is a univariate time series analysis. This project considers the alternative indicator as a proxy for real estate transactions and uses patterns identified in past time periods to predict future values.

Data was aggregated at a monthly level and used to train multiple time series forecasting models. This project considered ARIMA, ETS, and the Prophet forecasting tool by Facebook. To evaluate the performance of the models, this project used a rolling cross validation process. The model is given an initial training set and evaluated by forecast errors for a test set of future values. Then, the model is given incrementally larger (more recent) training sets paired with test sets pushed slightly further into the future. This process simulates the historical performance of the model on a period-by-period basis. The model with the lowest errors in rolling cross validation was the ARIMA model; therefore it is used to generate forecasts for the client.

Creating a Tool for Users

This project makes forecasts available to users by hosting the model and predictions in a Shiny application. Users can see a plot of historical and forecasted values, as well precise numeric predictions available in a table. Shiny is the ideal software tool for this task because

Shiny applications are simple for any user to interact with, and because the client organization is already familiar with the process of deploying a Shiny application on internal servers.

Project Outcomes

The client has gained the following capabilities as a result of the project:

- The client can automatically collect recent information about market volume by scraping data from the web.
- The client can create accurate forecasts based on recent, month-level data rather than outdated quarterly data.
- The client can share these forecasts with anyone in the organization by hosting a simple Shiny application.

Access to an improved market volume forecast has the potential to boost the efficiency of any facet of the client's operations. The client can allocate resources more strategically in response to the shifts in market volume which affect the client's core business.

Lessons Learned

As a student, this project was an incredibly valuable experience. I had the opportunity to exercise a broad range of skills from my academic program in a real-world setting. I list some reflections below:

- Identifying a new data source is one of the most valuable contributions a data scientist can make.
- Acquiring and cleaning data are the most time-intensive components of a typical project.
- When putting a data science project into production, it pays to use tools with which the organization is already familiar.
- Clients appreciate solutions that are simple to understand and use. This applies to the documentation as well!

This project delivered value to the client in the form of business intelligence, and it delivered value to me in the form of a real-world consulting experience.