Enhancing Revenue Intelligence for BrokerChooser

Public Capstone Project Summary

By
Péter Bence Török

Submitted to
Central European University
Department of Economics

In partial fulfilment of the requirements for the degree of Master of Science in Business Analytics

Supervisor: Eduardo Ariño de la Rubia

Vienna, Austria
2025

AUTHOR'S DECLARATION

I, Péter Bence Török, the undersigned candidate for the MA/MSc degree in Business Analytics declare herewith that the present thesis titled "Enhancing Revenue Intelligence for BrokerChooser" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography.

I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 06 June 2025

Péter Bence Török

COPYRIGHT NOTICE

1

Capstone Project Summary

Introduction

This paper is a 3-page public summary of my Capstone project, that aimed to improve how the Client tracks and understands its revenue generation and performance. The Client was BrokerChooser, a Hungarian startup, that helps users to compare brokers on their website and earns commission through an affiliate model. Although the Client has a sophisticated data infrastructure, they lacked clear insight into which user behaviours actually lead to revenue. My goal was to build interactive dashboards to help monitor performance and to identify the key factors that influence revenue generation.

Methodology

Throughout project, the first phase focused on developing multiple dashboard concepts, from which the client selected two for implementation. In the second phase, these dashboards served as a foundation for further analysis. To investigate key performance patterns a high-level SQL analysis was implemented through BigQuery (Google Cloud, n.d.) while Python (Python Software Foundation, 2024) was used for more advanced methods, including a regression analysis to identify variables most associated with revenue generation and predictive modelling to evaluate how accurately revenue-generating sessions could be detected.

Project Outline

Dashboard Creation: After assessing and discussing the client's needs, four interactive dashboards were developed in Metabase (Metabase, n.d.) to give a view of key revenue metrics across both country and broker dimensions. For each dimension, two dashboard types were developed: an aggregated view offering a high-level summary of key revenue metrics and a monthly monitoring dashboard which shows recent performance trends. The dashboards were built on cleaned and processed user log data and filters allow users to slice the data by various attributes, that makes it easier for users to customize queries.

Metabase Analysis: The client also requested a high-level data analysis focusing on the metrics and dimensions shown in the dashboards. Accordingly, I carried out an SQL-based analysis examining key revenue performance indicators across major brokers and markets. The findings from this analysis were then used to support the final recommendations.

Based on the insights from the broker and country-level analysis, a set of actionable recommendations were shared with the client. These included educating users about underperforming brokers, sharing conversion-related learnings with partners and improving the onboarding process to boost engagement. Based on country level-insights, I suggested prioritizing top-performing markets due to concentrated revenue, monitoring emerging markets and leveraging the strong conversion rates observed in offshore jurisdictions (OECD, 2021).

Data Pipeline: To carry out the regression analysis and to build predictive models I also built a pipeline to prepare data for analysis. The data preparation began with SQL, where raw log-level data was transformed into a structured session-level dataset through custom queries and aggregation functions. The resulting dataset was then cleaned and formatted in Python, addressing missing values, filtering out internal sessions and standardizing categorical and time-based variables. Feature engineering was used to create new variables such as session duration, time-based flags and indicators for specific page visits to better describe user engagement.

Regression Analysis: A logistic regression analysis (Békés & Kézdi, 2021, pp. 307–308) was carried out to explore which session-level characteristics are associated with a higher or lower likelihood of revenue generation. The cleaned dataset was prepared for modelling by transforming categorical variables into dummy variables and the logistic regression model was applied through the statsmodels Python library (Statsmodels, n.d.).

The results highlighted that interactions with some broker-specific content, longer sessions and visits to the community site were most strongly linked to revenue, while visits to educational or research-oriented content were linked to lower likelihoods of conversion. However, the model showed some useful patterns, its overall explanatory power was limited and only a small portion of the variables showed statistically significant results. For the Client it was also highlighted that these findings offered directional insight but should be interpreted with caution, as they reflect associations rather than direct causation.

Predictive Modelling: To explore whether session-level data could be used to predict revenue generation, a set of classification models were developed using the same prepared dataset as the regression analysis. Since revenue-generating sessions were extremely underrepresented in the dataset, synthetic oversampling (Synthetic Minority Oversampling Technique - SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was applied to balance the data and improve model training. Five models were tested: standard logistic regression, Lasso, Random Forest

and Extreme Gradient Boosting (XGBoost) (Békés & Kézdi, 2021, pp. 407–409, 475–477; Chen & Guestrin, 2016). Among these, the Lasso model proved to be the most effective, which offered strong predictive accuracy while it uses fewer input variables. Besides, an asymmetric cost function evaluation was also used for classification, that penalized false negatives significantly more than false positives. This evaluation method also supported the model's selection. Overall, the analysis showed that it is feasible to identify high-value sessions with relatively high accuracy.

Benefits for the Client

Based on the Client's informal feedback the project deliverables provided real value to the company. The dashboards aimed to help recurring performance monitoring, while the deeper analysis gave statistical evidence what factors really influence user conversions. Besides, the Client also found the addressed recommendations useful and actionable, which may support future business decisions and strategy building.

Learning Outcomes

Over the course of the project, I had the chance to apply a wide range of skills I learnt throughout my master's program. These included both theoretical knowledge (e.g., prediction methodology, regression principals, data visualization guidelines etc.) and also technical tools (e.g., building data pipelines, SQL, Python, BI tools etc.). What I especially appreciated was that this project really gave me the full experience. I got to work through every stage myself, from the ideation phase and building dashboards to running the analysis and addressing recommendations. Going through the whole process helped me understand what a real-world analytics project actually looks like and how important it is to adjust project scope to the needs of the client's expectations along the way.

Bibliography

Békés, G., & Kézdi, G. (2021). Data analysis for business, economics and policy (1st ed.). Cambridge University Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Google Cloud. (n.d.). *BigQuery SQL reference*. Retrieved June 3, 2025, from https://cloud.google.com/bigquery/docs/reference/standard-sql

Metabase. (n.d.). *Metabase: The open-source business intelligence tool*. Retrieved June 3, 2025, from https://www.metabase.com

OECD. (2021). *Harmful tax practices – 2021 progress report on preferential regimes*. Organisation for Economic Co-operation and Development. Retrieved June 3, 2025, from https://www.oecd.org/tax/beps/harmful-tax-practices-2021-progress-report.pdf

Python Software Foundation. (2024). *Python (Version 3.12) [Computer software]*. Retrieved June 3, 2025, from https://www.python.org

Statsmodels. (n.d.). *Statsmodels: Statistical modeling and econometrics in Python*. Retrieved June 3, 2025, from https://www.statsmodels.org/stable/index.html