Capstone Public Summary:

Survival Analysis for a Hungarian E-commerce Firm

István Jancó

MS in Business Analytics

Budapest 2022

Project Summary

The project was created for a B2C e-commerce company, residing in Hungary. The Firm (aka. Client) is operating an online marketplace for various types of handmade and vintage products. The conducted analysis was aimed at providing the Firm an insight to expected seller attrition and the variables that might affect it. At The Firm, seller accounts become inactive in case, there are no purchasing offers from clients. The Firm's management was interested in the probability distribution of the sellers' activity rate and lifetime. The created distribution and variable analysis would serve as a base for the Firm to understand the possible reasons behind attrition. While one of the main goals for the Firm is to scale up its operation and enter to new markers, seller attrition rate remains one of the most significant constraints.

One of the main goals was to discover those seller attributes, that might contribute most to seller's departure from the platform, so that going forward, management can mitigate, using preventive tactics (e.g., reminders). As the sellers are the main source of revenue, it is crucial to understand what factors might drive them away from using the Firm's marketplace and possibly to a competitor. To have a stronger base for analysis, a database of approx. ten thousand sellers was assembled with additional data including sales records, product groups and buyers. Any previous analysis on seller survival at the Firm is missing. The ability to forecast changes in seller base more accurately enables the company to establish a proactive resource allocation strategy as well as adjust its business model accordingly. The project was intended to serve as a base for further retention or failure analysis.

The provided data was mainly related to vendors' marketplace accounts. One of the most important variables is the seller's registration time and the time when the last posting expired. These variables allowed to establish the time boundaries of the analysis and provide insight to the state of vendors at different times.
Apart from the time-related data, additional covariate data was derived. A pool of possible explanatory variables was created and boiled down, using various methods (e.g., Recursive Feature Elimination, etc.).

Once the most predictive features were determined, a method called Survival Analysis was used to estimate the distribution of survival probabilities of vendors against the time frame.

Survival Analysis is the analysis of data involving times to some event of interest. In the context of this project the event of interest was the sellers' departure from the Firm's marketplace. Since there is no way of knowing the exact time of vendor deciding to leave the following logic was used to establish a proxy for the time of departure. If a seller has not made a sale in the past six month, the time of departure equates to the date of the last made sale. This logic established the accounts, which could be considered "dead". Accounts for which the above logic is not true, are considered censored. In context of Survival analysis, censoring means that a subject did not experience during the timeframe of the study, however it still might sometime in the future. The fact of censoring in the data added some unique challenges to the analysis.

Various statistical approaches were used to validate the outcomes of the model. During the development the data was split to train and test samples to avoid overfitting. Log-rank test and concordance indexes were used to validate the performance and predictions of the model. Furthermore, the created model was tested against a Random Survival Forrest to discover any possible shortcomings of the developed algorithm. The performance of the created model was proven to be equitable to the Random Survival Forrest.

Main Challenges and Learning Points

Since Survival Analysis was a new research area, a variety of question and challenges arose during the course of the project. One of the main such questions was whether Survival Analysis should be the primary approach for the project. This depended heavily on the goal of the analysis. If the client wished to focus on factors that influence seller attrition, it might've been more suitable to use an approach that prioritizes interpretability over model performance and perhaps re-classify the study as a retention analysis. Ultimately, the Client confirmed, that prediction of survival probabilities should be the main focus of the project.

Once the expectation was clarified by the Firm, a new question arose. Which type of modelling is the most appropriate to use? Given the range of the available methodologies to estimate survival, the chosen approach had to maintain the balance between performance, interpretability, and reproducibility. Since the project deliverable is intended to be used in a real business environment, the meaning of the coefficients and their effect on the survival probabilities must be transparent enough to be used by management as actionable feedback.

Consequently, the decision to use Cox Proportional Hazards models was made. This regression model measures the effect on survival in hazard rates. It acts as an extension of a Logistic Regression, that predicts hazard in a given point of time. Although, less sophisticated than a Random Survival Forrest or a Gradient Boosting Model, Cox provides a more straightforward interpretation of the variables.

Since the provided data base contained clientele-specific variables as well, an issue of project scope came to light. It was unclear whether the project should be limited to analyzing only shops-specific attributes or exploration of the client base would be expected. On one hand it might have been beneficial to see if there are distinct client groups that favor a particular group of products, which might affect the survival of some shops. Eventually, the buyer-related variable analysis was descoped form the project due to time and capacity limitations.

Another matter which was discussed with the Client was the form of the final deliverable. Since the outcome of the analysis is intended to be used on future data, several options were considered. A dashboarding solution was examined. It would enable the user to feed in new data and reproduce the analysis instantly. However, it would also somewhat limit the flexibility to adjust the analysis in terms of covariates. At the end the Firm indicated a preference towards keeping the analysis in a Jupyter Notebook, since the code can be easily adjusted or modified if needed. To enhance the ease of use and keep some degree of automation the practical implementation of Survival analysis was packaged to several functions and stored in a separate notebook.

The final issue that emerged during the project was the reproducibility. Since new shops are registering every day at the Firm's platform, it was crucial to make sure that the analysis can be repeated using future data. To achieve a significant level of reproducibility, the data cleaning, and Survival Analysis logic, used in the theoretical part, were packaged to several functions in Jupyter Notebook. This allows the Client to re-run the whole sequence on new data and adjust variables like the time to after which a shop is considered dead on the fly.

Overall, the project provided plenty opportunities to enhance both theoretical and technical knowledge. On the theoretical side it provided a deep insight to differences in assumptions between Survival analysis and other modeling methods. In addition, since it is mostly used in the medical studies, the project also provided the chance to learn about exercises like drug

trials. From technical perspective, there was a steep learning curve related to usage of Python and Jupyter Notebook for data analysis.
Finally, perhaps the main skill that was affected by the project was communication. The ability to understand the Client's needs was proven to be essential. It allowed the analysis to be more tailored and thus provide more business value.

Key Outcomes

Several milestones were accomplished during the project. In the first part, the data cleaning and enrichment processes were established and mostly automated, increasing the reproducibility of the analysis for new data. Subsequently, the features with most predictive potential were determined, these included several binary indicators as well as continuous variables. Using statistical methods and modeling (e.g., Logistic Regression, etc.) the pool of features was narrowed down. The features were then used as covariates for developing a model that can predict both survival and hazard for one or multiple shops. The study concluded one of the most predictive features was a binary variable which indicate whether a shop is also registered as a business enterprise. The variable seems to have a positive effect on the survival probability of shops.

The established theoretical background served as a basis for the analysis automation. A separate Jupyter Notebook was created, the objective of which was to automate the Survival Analysis and make it easily reproducible and scalable. This was achieved by defining several functions that can conduct the exercise based on user-selected inputs and criteria.