Arnaud Ajamian
BrainStation – Data Science
04/04/2022

Capstone Project – Final Report

Introduction:

Since the introduction of retail e-commerce, companies have been vying for customer loyalty to allow for the acquisition of greater market share. As such, customer retention is paramount to retain relevance in the hyper-competitive e-commerce landscape. From this, the principle aim of the project was to develop a series of models, whose predictive power could be utilized to help identify customers that are susceptible to churning (ie. customers that may be prone to no longer make us of a company's platform). This would enable user experience and/or customer service professionals to proactively address their customers' concerns and in doing so, could gain a better understanding of their customers priorities.

Customer churn is inherently a lagging indicator of poor customer experience, whereas customer churn prediction may enable companies from being reactive to proactive. The result of which would increase the customer retention rate, thereby reducing the rate of customer churn. Therefore, the deployment of such models could prove a competitive advantage to help companies retain and gain market share.

About the Dataset:

With regards to the origin of the dataset, although it had been made available via Kaggle.com; its true source cannot be readily identified. Though the information contained within provides insights with respect to customer demographics, their respective purchasing behaviors, and order history at a particular moment in time; one particular field of interest is not supplied. Although, information is provided with regards to which customer has churned, the dataset does not provide any insight as to why a customer has in fact decided to leave. This crucial piece of information could have further elevated the resourcefulness of the predictive model, in that not only could customer churn be predicted, but also why they may leave the service. If, at present, the reason for having lost a customer is not recorded, then additional forms of customer outreach need be put in place to help target the root cause of dissatisfaction and to enhance the customer experience going forward. Equipped with this knowledge, the user experience or customer service professional would have been able to leverage the model's predictive power to know in advance what might be the cause of customer dissatisfaction and would therefore be better equipped to resolve the issue in a timely manner.

Additionally, it is important to note that certain variables only represent information for the previous month. For example, according to the data dictionary, the 'complain' field only captures whether a complaint had been raised in the last month. This is problematic given that the 'churn' flag is not time bound. As such, it may be the case that a customer who has churned may appear as not having filed a complaint although they may have 2+ months prior. As such, additional information (such as an aggregation of historical information) would need to be made available to construct an accurate customer profile.

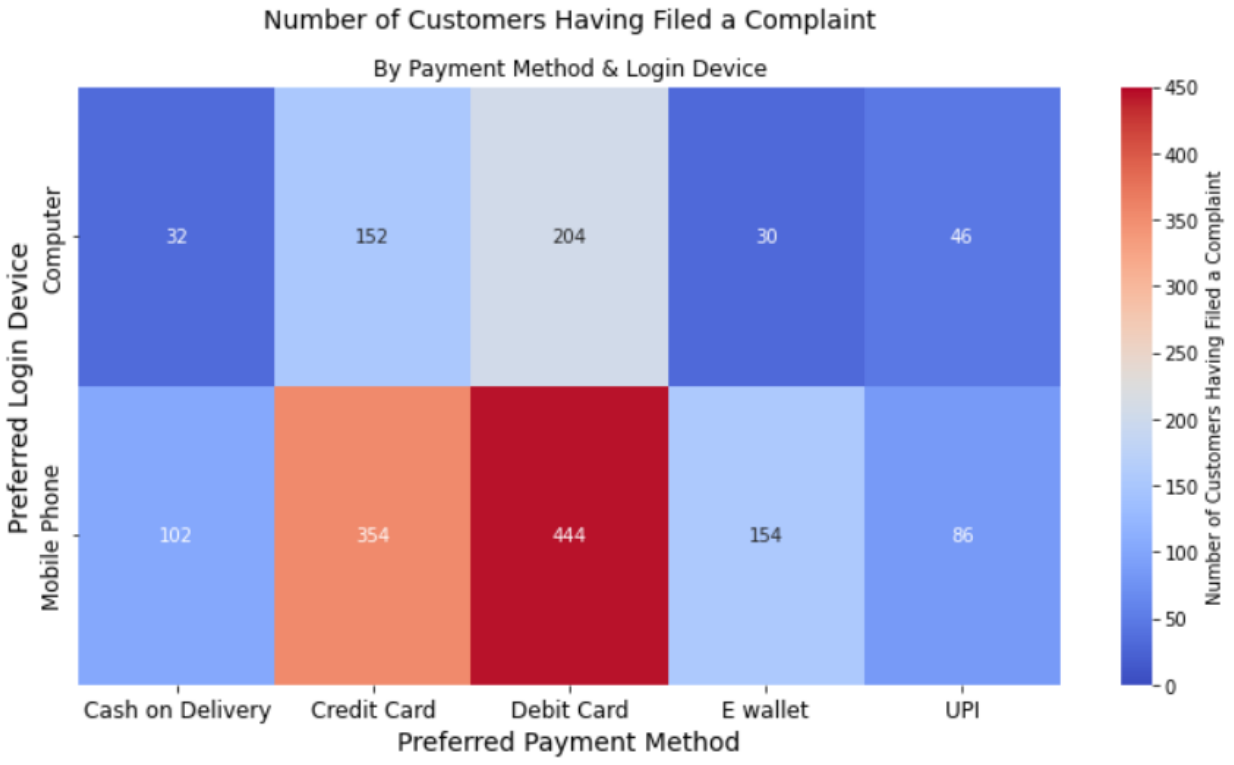
Furthermore, the dataset does not provide any demographic information as it relates to region, country, or state to help identify which demographic is represented. This would have

allowed for the integration of external data sources to supplement that of the dataset itself. With no additional information provided, one is unable to gather supplementary information concerning economic or market specific indicators applicable to the demographic represented in the dataset.

Exploratory Data Analysis (EDA):

With respect to EDA, I sought to identify which factor(s) appeared to have the most influence on the target variable 'churn'. In attempting to do so, it was first necessary to understand, under which circumstances, the rate of customer churn appeared to be the highest. In essence, the objective was to try and identify the root cause for customer churn. Of the EDA conducted, supporting variables such as 'satisfaction score' and 'complaints filed' were of central focus as it was presumed that customers who had churned had either filed a complaint or provided a low satisfaction score (defined as less than 4, on a scale of 1 to 5).

During the EDA process, a notable find had been made as it relates to the number of customers who had filed a complaint (refer to the figure below). Based on the preferred login device and payment method, it had been found that the most complaints had been filed by customers whose preferred login device was via mobile phone and whose preferred payment method was either by credit or debit card. This suggests that customers may be confronted with a hindrance when attempting to provide their payment information to complete their purchase. This is of particular significance because such an occurrence may lead to a higher rate of shopping cart abandonment. Whereby customers may decide to exit the checkout queue and seek to make their purchase elsewhere.



Modeling and Evaluation:

In seeking to develop a predictive model capable of identifying customers who may churn; an array of models were deployed to address this binary classification problem. Such models included Logistic Regression, K Nearest Neighbors, Decision Trees, Random Forests, Support Vector Machines, Boosting (Adaptive Boosting, Gradient Boosting and XG Boost) as well as the utilization of a Neural Network. Where applicable, additional methodologies were applied towards the aforementioned models such as Ensemble Learning and Synthetic Minority Oversampling Technique (SMOTE) with the intent of increasing the model's predictive power.

Though each model was initially evaluated based on accuracy, the introduction of SMOTE shifted the model performance metric toward model recall. In support of the value proposition of developing models that could be leveraged to identify cases of customer churn; optimizing for recall meant that the models were evaluated based on how well they succeeded in identifying all the positive cases within the data. However, it is important to note that in optimizing a model's performance towards recall; this may lead to instances whereby the model has incorrectly classified non-churn customers as belonging to the churn class (ie. false positives).

Although this may result in a company spending resources on customers that aren't ultimately going to leave, it can be argued that this approach helps fulfill the value proposition of increasing customer retention and reducing the rate of customer churn. This can be attributed to the fact that a company will be able to gain a better understanding of its customer base through customer outreach and may be able to improve its user experience, which may ultimately assist in gaining additional market share from competitors who do not proactively take the initiative to address the concerns of their customers.

Moreover, it can be stated that although following up with customers who were never intent on leaving is resource intensive; it may translate to greater customer satisfaction and retention. The result of which may also help to attract the customers of other retail e-commerce competitors given the adoption of a customer centric business model.

Conclusion:

Once the business proposal had been concretized, the order of operation entailed data acquisition and preprocessing, exploratory data analysis and modeling. In seeking to develop a model whose predictive power would be capable of correctly identifying all instances of customer churn; it can be concluded that XG Boost and Neural Networks resulted in the most powerful models[1] with a score of approximately 90%. Which is to say that such algorithms are capable of accurately identifying 90% of true positive observations as belonging to class 1.

[1] Powerful: Define as that which attained the highest recall score