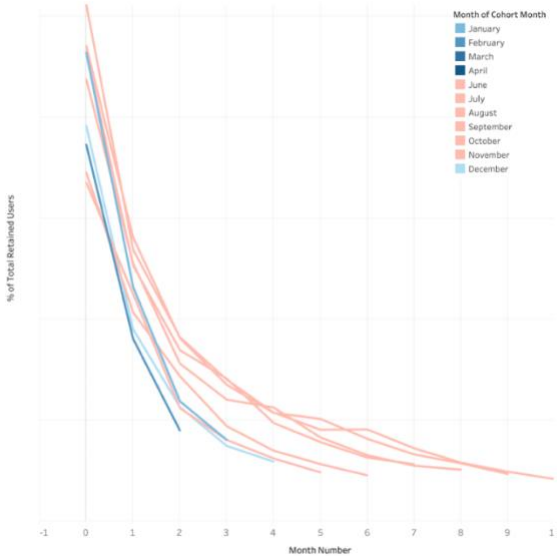
Mauricio Moreno - BrainStation Capstone Report
Elenas Churn prediction Model

Problem & context

Elenas is a social commerce marketplace that operates in Mexico and Colombia. The company has empowered tens of thousands of women, mostly in rural areas of both countries, by enabling them to launch online businesses and generate additional income through e-commerce platforms or social media sales. Elenas takes care of end-client management, product sourcing, delivery, and payment collection, thereby eliminating inventory risk. The company offers a portfolio of wholesale products, including categories such as beauty, home goods, electronics, clothing, and more, on a B2B2C model where the sellers don't have to spend nor invest anything for their businesses while Elenas focuses on integrating with product providers, third-party logistics, and building tools for the sellers who define their earnings for each order.

Monthly cohort of seller retention



(seller retention starting June 2022, y values removed)

In recent months, the company has shifted its focus towards profitability and has significantly reduced acquisition and retention incentives to be more cost-efficient. This has led to a decrease in the number of new monthly sellers.

Elenas' main revenue stream comes from a commission charged to the provider when a transaction is performed, so their most important input metrics are number of active sellers (retention), average order price and orders per seller per month. But due to the new priorities, the retention has not improved in recent months.

This model consists of a churn predictor algorithm that segments sellers who are likely to churn (i.e., not place any orders) each month. This was performed in two iterations, the first one analyzing sellers created after June 2021 and the second one, for sellers created after October 2022. For this document I will focus only on the second one, which is most relevant for the company.

About the data

Elenas provided a dump containing the dataset (10GB) comprising several tables. Through feature engineering / domain expertise, I selected 28 variables that fall into transactional, operational, and behavioral categories. The features were chosen to strike a balance between expected outcomes - average earnings per seller, and exploratory insights for future learning, such as seller focus on new or existing products.

The target variable, a binary column defined as 0 (not churn) for the sellers with >= 1 order for the last month of data was imbalanced, with 82% churn (1) and 18% not churn.

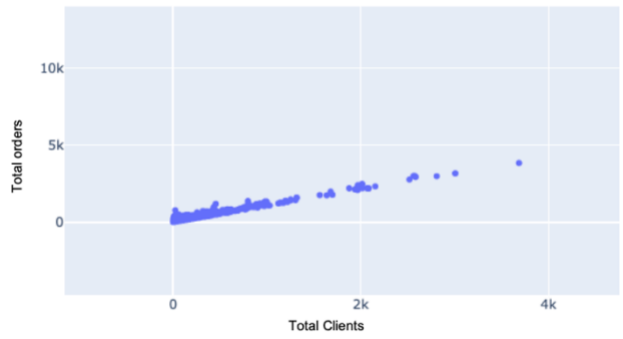
Variables were modified, including logistic transformation to normalize the seller earnings distribution, one-hot encoding for the seller creation time of month or converting seller creation date to days since seller creation, among other operations. These modifications derived from insights of the first iteration.

EDA

Some of the most relevant insights from the Exploratory Data Analysis are as follows:

i) There is an almost perfect correlation between the number of clients and total orders, indicating that sellers tend to focus on acquiring multiple clients rather than building relationships with existing ones. As a result, the repurchase order rate is low.

Relationship between total orders and clients



(Total orders vs total clients scatterplot)

ii) The average client price for orders from non-churned sellers is 13% higher compared to churned sellers. However, the notable contrast lies in the earnings from these orders, which are on average 88% higher for non-churned sellers.

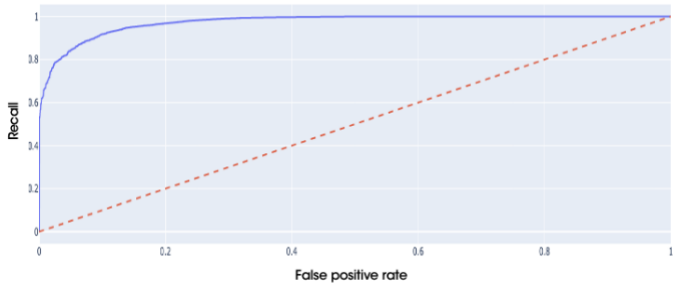
iii) The distribution of seller earnings follows a power user pattern. Considering the decrease in the number of new monthly sellers, it is important to note that the average ticket per seller may be skewed and not accurately represent the real scenario of sellers. To avoid misinterpretation, it is recommended to calculate this metric as a weighted average through segmentation.

Model

Given the distribution of the target variable, it is deemed relevant for a model to achieve an accuracy of over 82%. However, considering the objective of optimizing retention in a cost-efficient manner, additional evaluation metrics were prioritized. These metrics include recall and specificity for the negative class (not churned) and computing power, for example, a distance based KNN yielded good results but proved to be significantly expensive to run.

The top-performing models were Random Forest and XGBoost. While they had similar performance overall, XGBoost was chosen as the preferred model due to its computational efficiency (less depth layers) and an improvement of 200 basis points in specificity for the negative class.

XGBooster Model ROC AUC Score

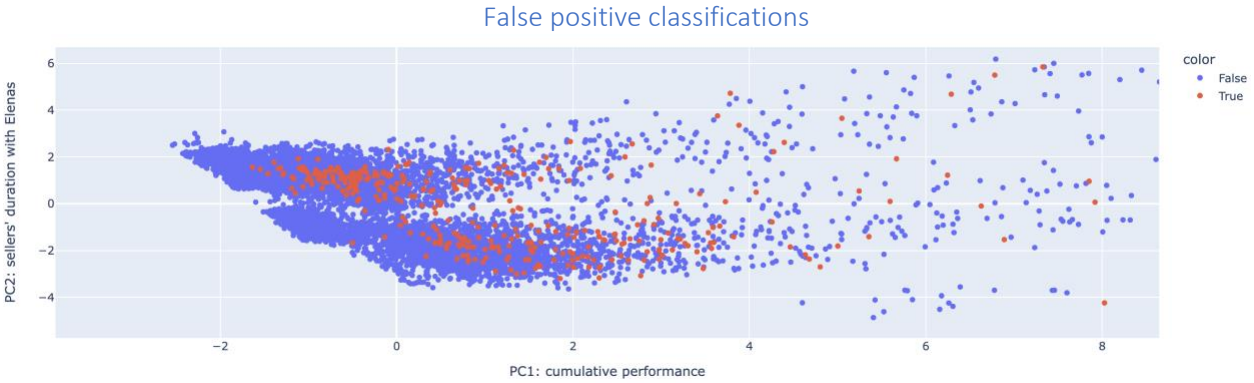


(XGB ROC AUC: 0.97)

The final model achieved an accuracy of 94%, recall of 98%, 94% precision. For the negative class, the model achieved a true negative ratio of 75% and a negative predicted value of 91%. These results indicate that the model effectively classifies instances, with a high rate of correctly identifying positive cases and a satisfactory rate of avoiding false negatives.

Various analyses were conducted to gain insights into the misclassifications labeled by the model. One approach involved hypothesis testing for different groups of misclassified sellers, because there was a pattern of misclassifications. It was observed that certain metrics, such as the effective order ratio, had significantly higher average scores among the

misclassified sellers. However, since this metric is a positive indicator, it should not be considered indicative of churn. This suggests that in future iterations, additional app-related engagement features should be incorporated to better understand these patterns.



(Scatterplot of the false positive classifications, pattern observed in PC1 range -1.05X – 0.08X)

On the other hand, another type of misclassification was identified, where the average total effective earnings were notably higher for the false positive cases. This is an area that should have been addressed by the model to avoid misclassifying sellers with higher earnings as churned. This finding highlights the need to refine the model's performance in accurately identifying such cases.

During the SHAP analysis visualizations, additional interesting patterns were uncovered. One notable finding was the significant impact of the number of vouchers given to a seller, which proved to be highly indicative of churn. A hypothesis that could explain this phenomenon is that the vouchers are provided in response to returns or cancellations, suggesting negative experiences for the sellers. However, further analysis is required to validate and explore this hypothesis in more depth.

Overall, these analyses shed light on specific areas where the model can be improved in future iterations to enhance its predictive capabilities and reduce misclassifications.

Next steps

This iteration of the model is ready for production use by the growth team to prioritize retention efforts. According to the results, it should capture 98% of the sellers that will churn for the following month and the precision of its classification is of 94%. However, there are several potential improvements for future iterations, such as:

1. Including app-behavior features that were not available in the initial dataset, as these features could explain many false-positive cases.
2. Replicating the model for the Mexico to understand if the patterns and insights are consistent.
3. Testing referral-related variables that are relevant to the business model, as they could provide valuable insights into customer acquisition and retention strategies.
4. Conducting regular model monitoring and evaluation to ensure its ongoing effectiveness.

These iterations have the potential to further enhance the model's performance and provide a more comprehensive understanding of the factors influencing churn.