Classification of Waze data

Project Summary

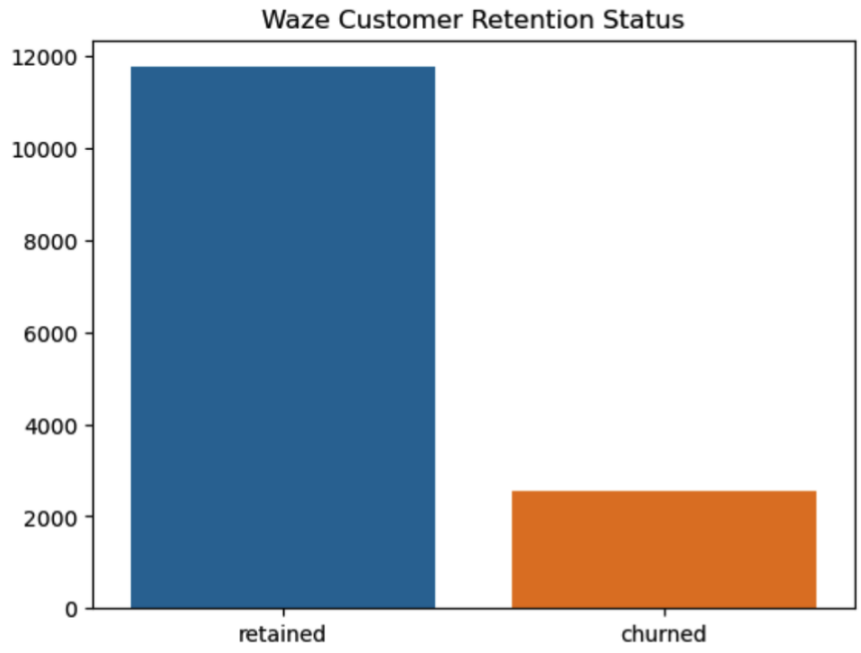
The ultimate goal for this project was to develop a machine learning (ML) model that predicts user churn. For the purposes of this project, churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. The dataset used for this project contains synthetic data created by Google in partnership with Waze. Engineered features accounted for six of the top 10 features including km per hour, percent of sessions in last month, total sessions per day, percent of drives to favorite. The XGBoost model fit the data better than the random forest and logistic regression model created to fit the data. Its recall score (17%) is nearly double the score from the logistic regression model while still maintaining a similar accuracy and precision scores of around 90% and 40%.

Business Context

This project is considered to be part of a larger effort at Waze to increase growth. Typically, high retention rates indicate satisfied users who repeatedly use the Waze app over time. Developing a churn prediction model will help prevent churn, improve user retention, and grow Waze's business.

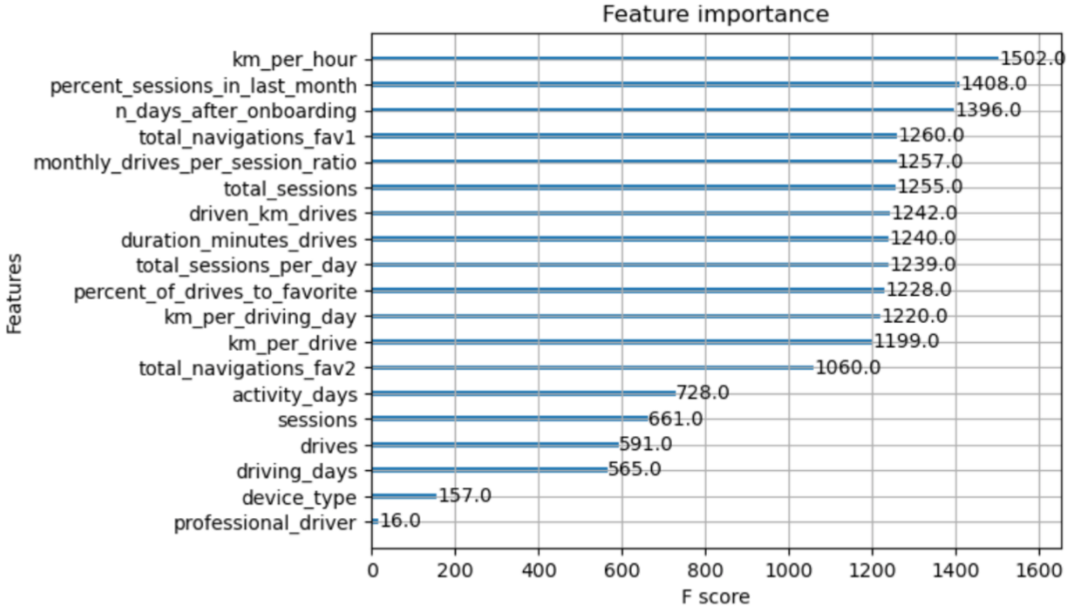
Data Insights

This project uses a waze dataset, which is a synthetic dataset created by Waze. The dataset contains information about 13 features of almost 15000 of its customers. As indicated by the figure below, about 15% of Waze users churn which is more than 2000.



Model Development and Results

Engineered features accounted for six of the top 10 features include km_per_hour, percent_sessions_in_last_month, total_sessions_per_day, percent_of_drives_to_favorite, km_per_drive, km_per_driving_day. The XGBoost model fit the data better than the random forest model. Additionally, it's important to call out that the recall score (17%) is nearly double the score from the previous logistic regression model built, while still maintaining a similar accuracy and precision score. The ensembles of tree-based models e are more valuable than a singular logistic regression model because they achieve higher scores across all evaluation metrics and require less preprocessing of the data. However, it is more difficult to understand how they make their predictions.



Conclusion

The findings provide insights for taxi drivers on the likelihood of receiving significant tips. Further analysis, such as a parametric model, could refine understanding of how different variables affect the tipping amount. Future improvements could include incorporating data on individual riders' previous tipping habits to better support the financial well-being of taxi drivers. The extensive analysis of Waze datasets yields significant results for the Waze business. Although consequential business decisions cannot be made based on these models due to poor recall score, they can be very helpful to guide further exploration. New features could be engineered to try to generate better predictive signal.