

Projeto de Retenção de Usuários

Resultados do modelo de ML



Preparado para: Equipe de Liderança doWaze

➤ Visão Geral

A equipe de dados do Waze está desenvolvendo um projeto que visa aumentar o crescimento geral da empresa, evitando a perda mensal de usuários. Como parte do esforço para aprimorar a retenção, o Waze deseja entender melhor o comportamento dos usuários. O objetivo final deste projeto é desenvolver um modelo de aprendizado de máquina (ML) que preveja a perda de usuários. Este relatório oferece detalhes e insights da Etapa 6, que podem impactar o desenvolvimento futuro do projeto, caso mais trabalho seja realizado.

➤ Impacto

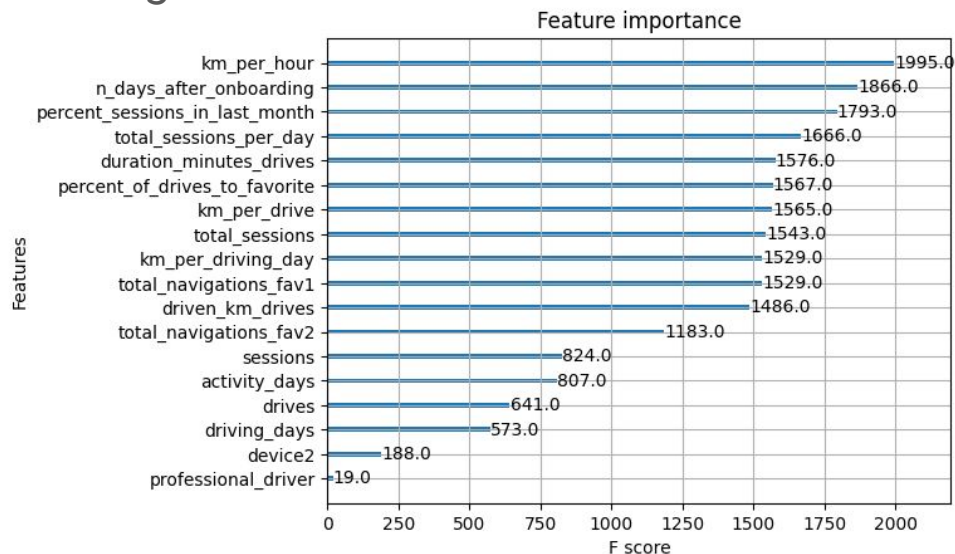
- ➔ Os modelos de aprendizado de máquina desenvolvidos para a Etapa 6 demonstram uma **necessidade crítica de dados adicionais** para prever com mais precisão a perda de usuários.
- ➔ O **esforço de modelagem destaca que os dados atuais são insuficientes** para prever consistentemente a perda de usuários. Sugere-se obter informações de viagem detalhadas e dados mais granulares sobre a interação dos usuários com o aplicativo.
- ➔ Como a **engenharia de features** é comprovadamente valiosa para melhorar o desempenho dos modelos de aprendizado de máquina, a **equipe do Waze recomenda uma segunda iteração do Projeto de Retenção de Usuários.**

➤ Descrição

- Para obter um modelo com o maior poder preditivo, a equipe de dados do Waze desenvolveu dois modelos diferentes para comparar resultados: **Random Forest e XGBoost.**

- Os dados foram divididos em conjuntos de treinamento, validação e teste. Embora a divisão em três conjuntos reduza a quantidade de dados para treinar, permite a seleção do melhor modelo usando um conjunto de validação separado, resultando em uma melhor estimativa do desempenho futuro em comparação com uma divisão em dois conjuntos.

➤ Insights



- **Características criadas representaram seis das dez principais:** km_per_hour, percent_sessions_in_last_month, total_sessions_per_day, percent_of_drives_to_favorite, km_per_drive, km_per_driving_day.

- **O modelo XGBoost se ajustou melhor aos dados do que o modelo de random forest.** Além disso, é importante destacar que a pontuação de recall (17%) é quase o dobro da pontuação do modelo de regressão logística anterior construído na Etapa 5, mantendo uma precisão e acurácia semelhantes.

- **Os conjuntos de modelos baseados em árvores nesta etapa do projeto superam um único modelo de regressão logística** em termos de desempenho, exigindo menos pré-processamento de dados, embora a interpretação de suas previsões seja mais complexa.