Relatório sobre Avaliação dos Modelos (CRISP-DM - Fase 5)

Dataset: Gender Gap in Spanish WP

Grupo:
Nilo Bemfica Mineiro Campos Drumond *(nbmcd)*
Pedro Didier Maranhão *(pdm)*
Pedro Tenório Lemos *(ptl)*

Neste projeto, usamos modelos de machine learning para classificar os usuários da Wikipedia em espanhol como 'homem' ou 'mulher' com base em suas atividades e informações fornecidas. A aplicação potencial de um modelo de classificação preciso é vasta, desde a personalização de conteúdo até análises demográficas mais profundas.

Nesse sentido, nosso objetivo de negócio principal era entender a distribuição de gênero entre os usuários da Wikipedia em espanhol para melhorar as estratégias de engajamento, personalização e segmentação. Grande parte dessa avaliação e desse entendimento vieram já mesmo na etapa de EDA, onde verificamos a distribuição de gênero entre os usuários e comportamentos sazonais de homens comparados a mulheres.

Já na etapa de modelagem, os modelos como RandomForest e GradientBoostingTrees, mostraram um desempenho promissor com métricas de avaliação como AUC e precisão. No entanto, foi observado um desafio na predição precisa da classe 'mulher', que é a classe minoritária, esperado devido a sua baixíssima representatividade.

A capacidade de identificar corretamente o gênero dos usuários pode ter implicações significativas:

- **Personalização de Conteúdo**: Saber o gênero do usuário pode ajudar na personalização do conteúdo, tornando-o mais relevante para grupos específicos. Por exemplo, destacar tópicos ou artigos que se mostraram populares entre um gênero específico.
- **Estratégias de Engajamento**: Com uma compreensão clara da demografia de gênero, estratégias de marketing e engajamento podem ser moldadas para atrair e reter mais usuários de um gênero específico. Nesse caso, atrair mais mulheres para editarem postagens da Wikipedia em espanhol.
- **Segmentação Avançada**: O modelo pode ser até mesmo usado internamente dentro das análises da Wikipedia para criar uma 'feature' ou característica de gênero mais sólida, dado que entendemos que vários usuários ‘unknown’ na verdade se encaixam como homem ou mulher. Essa feature pode ser combinada com outras características dos usuários, como atividade, preferências ou histórico de navegação, para segmentar os usuários em grupos mais refinados.

Apesar dessas vantagens, alguns modelos mostraram sinais de overfit, o que pode comprometer sua performance em dados não vistos. Será crucial monitorar a performance do modelo em produção e ajustá-lo conforme necessário. Isso se deve principalmente à classe 'mulher' estar sub-representada, o que levou a predições menos precisas para essa classe.

Ao personalizar conteúdo com base no gênero, é crucial garantir que não se perpetuem estereótipos ou preconceitos. Reforçamos isso bastante ao longo do projeto pois acreditamos que pode ser bastante tênue a linha entre o direcionamento de conteúdo e o preconceito. Além disso, a privacidade dos usuários deve ser sempre respeitada, e eles devem estar cientes de como seus dados estão sendo usados.

Enfim, o projeto fornece uma ferramenta poderosa para entender a demografia de gênero dos usuários da Wikipedia em espanhol. Com as devidas considerações e ajustes, o modelo tem o potencial de trazer insights valiosos e melhorar a experiência do usuário na plataforma. Contudo, é fundamental reavaliar e recalibrar o modelo periodicamente para garantir sua relevância e precisão contínuas se imaginarmos ele em um contexto real.