Entrega 3

No notebook `modelo_tempo_permanencia.ipynb`, desenvolvemos um modelo preditivo para estimar o tempo de permanência dos usuários na Ana Health. Este processo contou com a utilização de um dataset pré-processado e com feature engineering. Vale destacar que o pré-processamento foi realizado nos notebooks contidos na pasta 'pré-processamento' dentro da pasta 'notebooks', com o objetivo específico de lidar com valores nulos. Neste sentido, adotamos estratégias como descartar colunas com muitos valores nulos, consideradas irrelevantes para o modelo, ou preencher esses valores nulos com o número 0.

Além disso, a feature engineering foi realizada nos notebooks localizados na pasta 'notebooks/feature engineering'. Nessa fase, focamos em tratar colunas com valores categóricos, aplicando técnicas de encoding e outros métodos para transformar esses dados em um formato que pudesse ser efetivamente utilizado pelos algoritmos de aprendizado de máquina. Esse trabalho foi crucial para assegurar que os modelos pudessem interpretar e utilizar todas as informações disponíveis no dataset de forma otimizada.

O nosso dataset foi filtrado para incluir somente as linhas que representam assinaturas canceladas, em linha com o foco do nosso modelo. Dividimos o conjunto de dados em variáveis independentes e dependentes, sendo a 'stay_time' a variável dependente que indica o tempo em dias que um usuário permaneceu ativo na plataforma antes de cancelar a inscrição. Para assegurar a reprodutibilidade e a validade dos resultados, os dados foram divididos em conjuntos de treino e teste, com 25% destinados ao teste, utilizando uma seed aleatória.

Implementamos a técnica de Grid Search com diversos algoritmos e parâmetros para identificar o de melhor desempenho. Entre os métodos testados estava a Regressão Linear, usando features polinomiais e padronização dos dados. No nosso Grid Search, variamos os graus dessas features polinomiais, testando 1, 2 e 3. O desempenho do modelo foi avaliado principalmente através do Mean Squared Error (MSE) e do R^2 score.

Além da Regressão Linear, utilizamos o Random Forest Regressor e o Support Vector Regressor, cada um submetido a um Grid Search para determinar os melhores parâmetros e alcançar o melhor desempenho. Os resultados obtidos foram:

- **Random Forest Regressor**:
  - RMSE: 78.69
  - R2: 0.4634757948768473
  - EVS: 0.518854857312801
  - Erro percentual médio: 27.41%
- **Support Vector Regressor**:
  - RMSE: 85.16
  - R2: 0.4241861950076372
  - EVS: 0.4404396850224982
  - Erro percentual médio: 13.98%
- **Linear Regression**:

- RMSE: 166.92
- R2: -1.2121648332607773
- EVS: -1.1650483870435169
- Erro percentual médio: 13.51%

Utilizamos métricas como RMSE, R2, EVS e o erro percentual médio para avaliação. Comparando os modelos, concluímos que o Support Vector Regressor apresentou o melhor desempenho, com R2 e EVS mais próximos de 1, além dos menores valores de RMSE e erro percentual médio.

Este modelo com a target 'stay_time' é de grande valor para a Ana Health, especialmente no contexto de redução de churn. Ao prever o tempo de permanência dos usuários, a empresa pode identificar padrões e tendências que indicam um risco maior de cancelamento de assinatura. Com essas informações, a Ana Health pode implementar estratégias proativas para aumentar a retenção de usuários, oferecendo, por exemplo, incentivos ou melhorias personalizadas para aqueles com maior probabilidade de churn. Dessa forma, este modelo não apenas fornece insights valiosos para a tomada de decisão, mas também é uma ferramenta poderosa para melhorar a satisfação do cliente e fortalecer a fidelidade à marca.