RankMyApp

Predição de Daily Active Users (DAU)

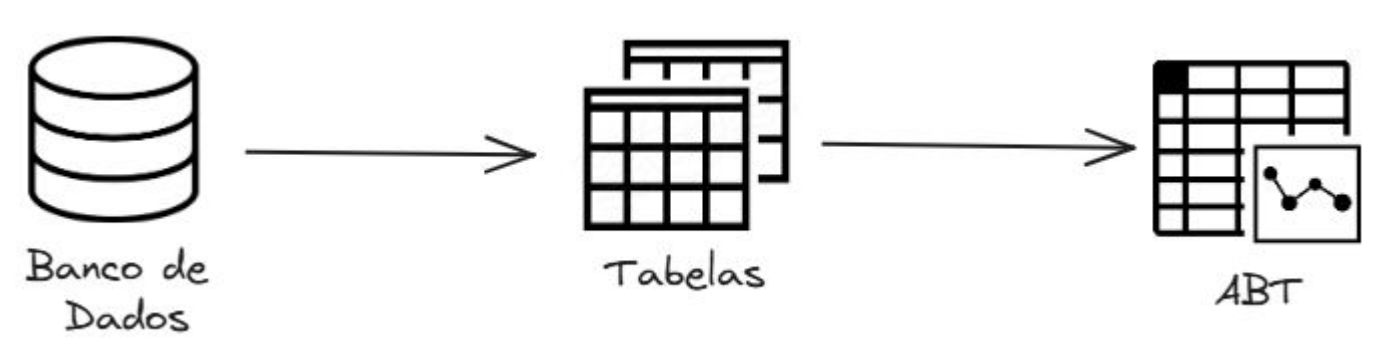
Rodrigo Marques

Extração e Tratamento dos dados

- Esta etapa está registrada nos notebooks **extract_data.ipynb** e **create_abt.ipynb.** O objetivo foi extrair os dados do banco de dados e criar um pequeno book de variáveis que são variáveis criadas com diferentes combinações matemáticas para aproveitar 100% dos dados do Data Lake, uma técnica muito utilizada por bancos.

- Por fim salvei os dados em csv no arquivo chamado abt

Tratamento dos dados



- No processo de geração da ABT eu tratei os dados duplicados, removi os dados que estavam nulos no target

Modelagem

- Dividi os dados de desenvolvimento e teste com corte nas datas, deixando o último mês para teste.

- Treino:
  - De Janeiro de 2024 até Agosto de 2024
- Teste:
  - Mês de Setembro de 2024

Modelagem

- Uma das coisas que eu não tratei antes de criar a ABT foi os valores nulos. Com os dados de treino e de teste divididos eu peguei a média de todas as variáveis de treino para imputar no lugar dos valores nulos.

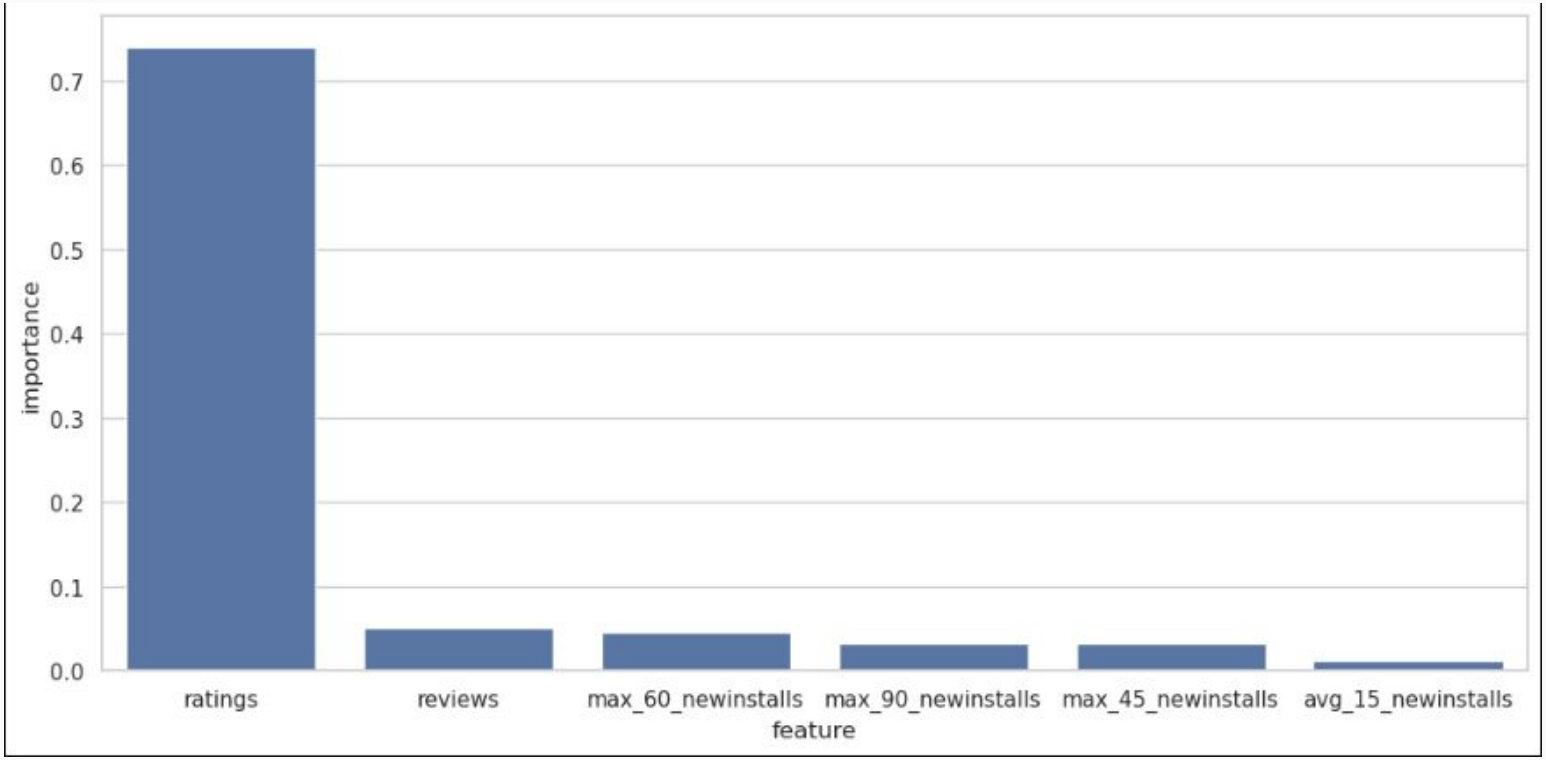
- Eu salvei essas informações na pasta de artefacts, pois como depois do deploy em processos streaming não é possível fazer uma média para 1 app apenas, essas informações precisam ficar registradas para serem usadas em produção.

Seleção de Variáveis

- Depois de limpar os valores nulos eu rodei um modelo de Random Forest Regressor nos dados de treinamento com as 100 variáveis e então utilizei o valor do feature importance que o modelo gera para pegar as 20 melhores

- Uma das variáveis estava com uma importância muito alta em relação as demais, então eu removi ela pois em alguns casos quando isso acontece o modelo tende ao overfitting.

Gráfico Importância das Variáveis



Treinamento

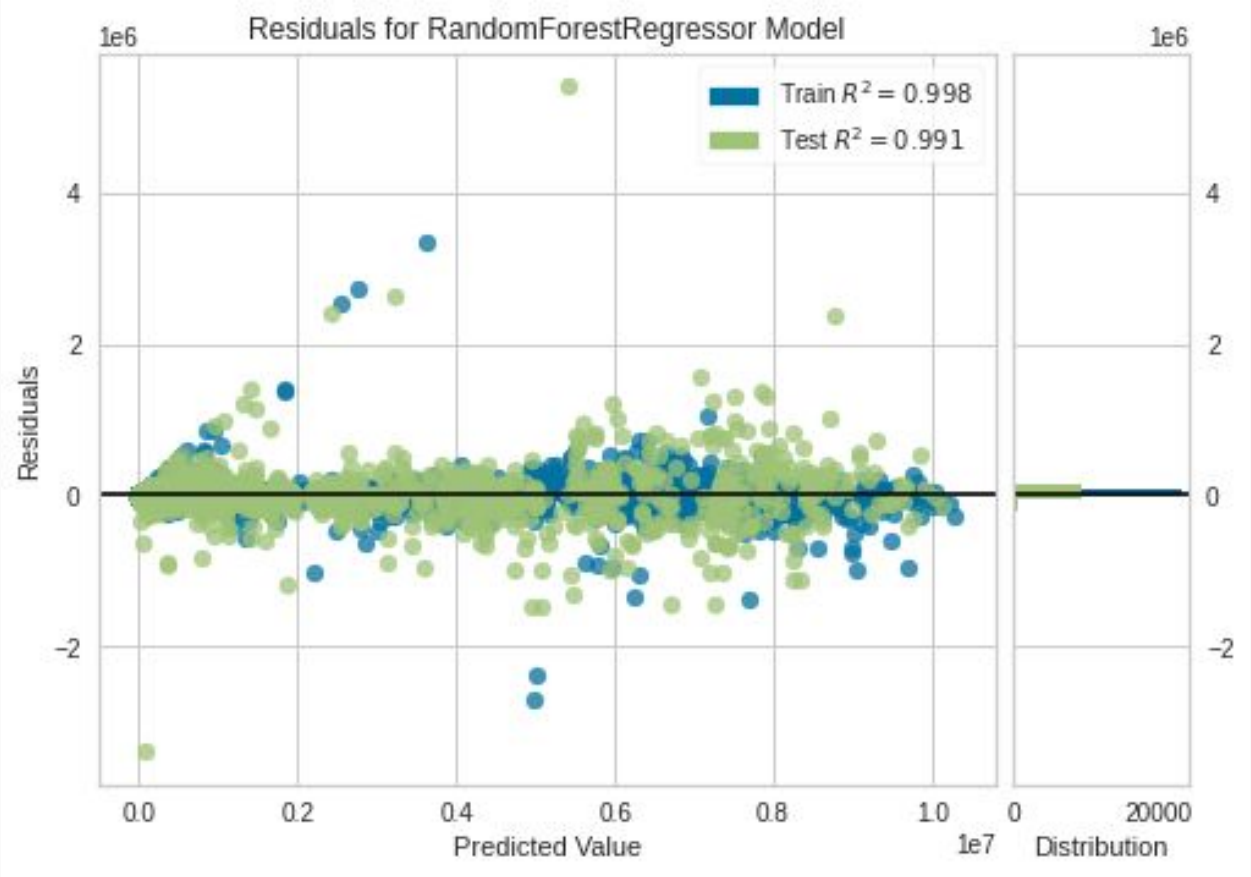
- Optei por utilizar o pycaret apenas para fazer o treinamento e o tuning dos modelos, dentre os modelos treinados o Random Forest Regressor teve o melhor desempenho.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	32858.5359	28598307712.9995	163739.7510	0.9846	0.2783	0.5985	1.7500
gbr	Gradient Boosting Regressor	75558.0736	41626062282.7261	199144.5596	0.9776	1.1279	2.3957	1.8260
ridge	Ridge Regression	227584.6596	249139078230.7788	496740.7671	0.8640	1.8053	9.5711	0.0230
lasso	Lasso Regression	227514.8180	249891744513.9298	497498.2892	0.8636	1.8077	9.5123	0.1020

Outliers

- Optei por não remover os outliers da base (mesmo acreditando que eu poderia ganhar um pouco de performance) pois em produção dados como outliers são inevitáveis, então mantive eles para habituar o modelo a um cenário mais próximo dos dados reais de produção.

Gráfico de residuos do modelo após o tuning de hiperparâmetros



Métricas

- Nesta imagem mostra a comparação da avaliação do modelo nos dados de treino e nos dados de teste.

Treino

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	17558.9668	7908807610.8809	88931.4771	0.9957	0.2143	0.2855

Teste

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	52575.8134	27606001230.3848	166150.5379	0.9859	0.5904	0.9385

Food and Drink

- Uma análise interessante é ver as métricas do modelo agrupado pelas categorias. As imagens acima são resultados do treino e na debaixo nos dados de teste.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	16655.1420	4728991238.6787	68767.6613	0.9674	0.1388	0.0765

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	39221.6939	4177723238.6871	64635.3095	0.9634	0.1351	0.1175

Finance

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	24601.3739	12732725832.6447	112839.3807	0.9961	0.1998	0.2201

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	82845.2160	58941421654.1429	242778.5445	0.9833	0.3658	0.2867

Business

- A categoria Business foi uma das mais afetadas negativamente no teste, pode ser pelo perfil do público ou algum outlier. Um teste interessante que eu faria se tivesse mais tempo seria treinar um modelo apenas para essa categoria, se for devido a público pode funcionar bem.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	8770.8388	769705950.6897	27743.5749	0.9947	0.1826	0.1096

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	32874.2519	6976587558.4222	83525.9694	0.9552	0.5873	0.9244