Case Técnico - Cientista de Dados

Comentários e Resultados

Candidato: Eric Henrique da Silva Passos

Obs. Algumas tarefas importantes (técnicas e não técnicas) foram deixadas de lado por causa do pouco tempo disponível para o desenvolvimento (o meu tempo disponível).

1 – Primeira Fase:

A tarefa da primeira fase consistiu em analisar as tabelas, identificar a melhor forma de junta-las e prepara-las para um modelo de ML, tratar esses dados (limpeza e exclusão de variáveis visualmente inúteis), e analisar a relação e distribuição dos seus dados (verificação de outliers, etc)

Sobre a variável alvo:

- Segundo a descrição que consta no repositório, a variável alvo seria a "Daily Active Users (DAU)", entretanto não existe uma variável com esse nome em nenhuma dessas tabelas. Considerando que não foi fornecido um dicionário de dados, identifiquei que a variável com o nome mais aproximado seria “dauReal” que está na tabela “daumau”.
- Por essa razão, decidi que a variável alvo seria “dauReal”.

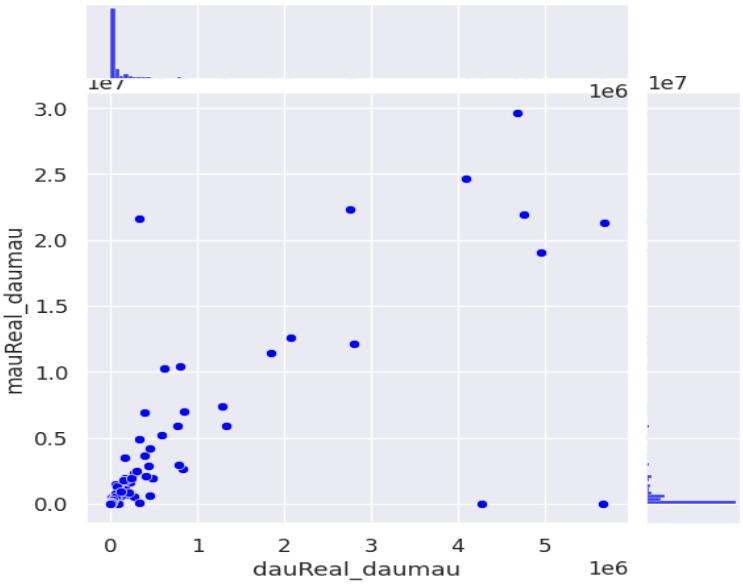
Sobre a estruturação e o relacionamento das tabelas:

- A parte mais desafiadora foi e que demandou mais tempo.
- Poderia ter sido feita de várias maneiras; dependendo do tratamento aplicado, os dados aceitariam modelos de séries temporais ou não temporais.
- O tratamento realizado por mim foi voltado para modelos não temporais. Não houve tempo suficiente para desenvolver um tratamento para modelos de séries temporais, pois isso exigiria muito mais tempo em termos de “engenharia de features”, pivotagem etc.
- Foram feitos dois tratamentos diferentes: o primeiro busquei entender se havia correspondência entre a variável “appid” e a variável “date” em todas as tabelas. Criei um subset apenas com as observações das tabelas em que “appid” tivesse as mesmas datas na coluna “date”. Verifiquei que, apesar do nome “appid”, essa variável poderia não representar IDs, mas sim uma coluna que agregasse grupos ou alguma agregação feita anteriormente.
- No final, isso não foi possível, e o que restou foram dois DataFrames (com as tabelas ratings_reviews e daumau), um unido pela coluna de data, com aproximadamente 300 linhas. A tabela ratings_reviews foi escolhida porque tinha mais datas iguais às da tabela daumau.
- A segunda tentativa consistiu em agrupar pela variável “appid” e combinar todas as informações em uma única linha. Foi essa que eu utilizei para gerar o modelo e assim por diante. Obs. Não é a melhor abordagem, pois estou perdendo informações (dados). Foi essa que eu subi no repositório.
- Obs.: Talvez a melhor maneira de lidar com este projeto fosse construir um modelo de séries temporais. Para isso, o tratamento dos dados envolveria muita pivotagem das

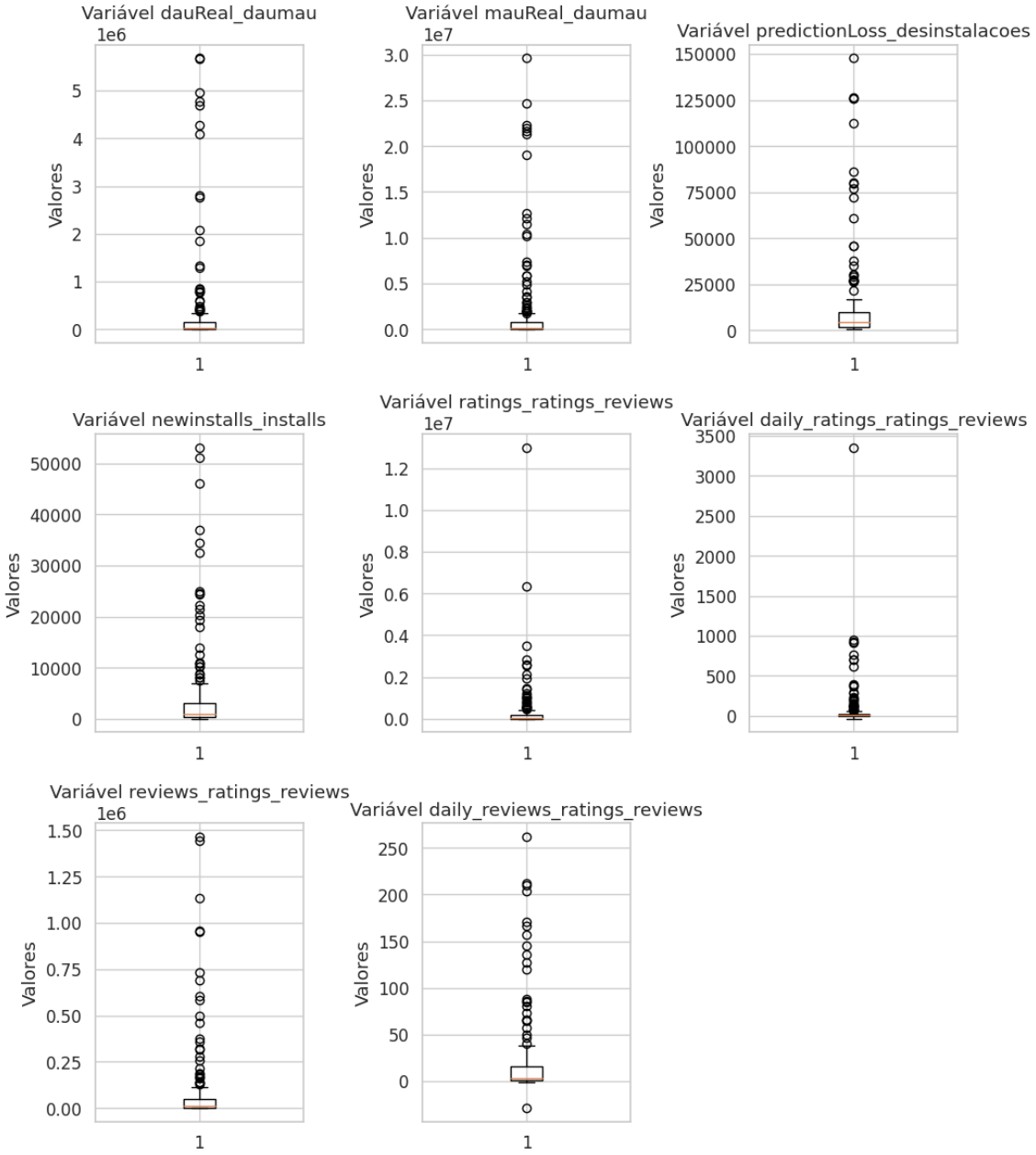
tabelas e a criação de vários subsets, até "encaixar" os dados de maneira a minimizar a geração de valores NaN e perda de informações.

Sobre o tratamento dos dados:

- As colunas lang_desinstalacoes e country_desinstalacoes foram excluídas pois tinham apenas a categoria “br”
- Foi analisada a distribuição da variável alvo “dauReal_daumau”, e através do gráfico ficou claro que não tem uma distribuição normal (o que dificulta um pouco) e que tem correlação com a variável “mauReal_daumau”



- Como a ideia é não trabalhar com séries temporais, todas as colunas de datas foram excluídas.
- A coluna Appid também foi excluída, pois, como mencionei acima, para facilitar o desenvolvimento, ela está sendo considerada como um ID. Observação: em um projeto, eu não poderia fazer isso, pois provavelmente distorceria os dados (precisaria confirma o seu real significado).
- A variável "category_ratings_reviews" foi "dumizada".
- Foi analisada a multicolinearidade entre as variáveis preditoras.
- Algumas variáveis possuem uma colinearidade de 0,8 com outras, o que é considerado um pouco alto. Entretanto, vou manter essas variáveis e verificar se elas não serão eliminadas na fase de "feature importance".
- Outliers: todas as variáveis possuem valores outliers:



- A melhor opção é sempre fazer uma análise desses outliers. Para esse caso, após testes, a exclusão dos outlier perderia muitas informações (linhas) que já foram reduzidas após a transformação. Além disso, todas as variáveis (com exceção das categóricas) incluindo o target, possuem valores outliers e por essas razões optei por manter os outliers.
- Obs. O legal seria ter um modelo com e sem os outliers, além disso, como a outra base citada (base final tratada) possui mais linhas, talvez fosse possível excluir esses outliers.

2 – Modelo Preditivo

- Os dados foram separados em treino e teste (optei por não separar também em validação, pois a quantidade de linhas diminuiu muito)
- Para esse conjunto de dados, talvez a padronização não seria necessária, pois os dados estão praticamente na mesma escala. No entanto, segui as premissas do modelo de

regressão e padronizei. O padronizador foi treinado apenas com os dados de treino. (decidi isso antes de decidir usar o AutoML e resolvi manter já que não impacta negativamente)

- O correto seria desenvolver cada modelo isoladamente, analisando as suas suposições, fazendo alguns tratamentos para lidar com a distribuição não normal do target, modelos que lidam melhor com outliers e assim por diante. Mas levaria muito tempo e por isso acabei optando por utilizar o H2o AutoML.
- Modelos Treinados:

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
GBM_3_AutoML_3_20241025_84705	1.09448e+06	1.19788e+12	572647	nan	1.19788e+12
GLM_1_AutoML_3_20241025_84705	1.09649e+06	1.2023e+12	579654	3.20975	1.2023e+12
StackedEnsemble_BestOfFamily_1_AutoML_3_20241025_84705	1.09693e+06	1.20326e+12	586360	3.23044	1.20326e+12
GBM_2_AutoML_3_20241025_84705	1.10105e+06	1.21231e+12	582346	nan	1.21231e+12
GBM_grid_1_AutoML_3_20241025_84705_model_3	1.1021e+06	1.21462e+12	564231	nan	1.21462e+12
GBM_4_AutoML_3_20241025_84705	1.103e+06	1.21661e+12	569842	nan	1.21661e+12
StackedEnsemble_AllModels_1_AutoML_3_20241025_84705	1.10463e+06	1.2202e+12	575250	3.1923	1.2202e+12
DeepLearning_grid_3_AutoML_3_20241025_84705_model_1	1.11443e+06	1.24196e+12	514554	nan	1.24196e+12
GBM_grid_1_AutoML_3_20241025_84705_model_1	1.11545e+06	1.24422e+12	578019	nan	1.24422e+12
DeepLearning_grid_1_AutoML_3_20241025_84705_model_1	1.13289e+06	1.28345e+12	613464	nan	1.28345e+12

- Melhor modelo:

```
Model Details
=============
H2OGradientBoostingEstimator : Gradient Boosting Machine
Model Key: GBM_3_AutoML_3_20241025_84705


Model Summary:
    number_of_trees    number_of_internal_trees    model_size_in_bytes    min_depth    max_depth    mean_depth    min_leaves    max_leaves    mean_leaves
--  -----------------  --------------------------  ---------------------  -----------  -----------  ------------  ------------  ------------  -------------
    23                 23                          3153                   3            6            4.08696       5             8             6.30435

ModelMetricsRegression: gbm
** Reported on train data. **

MSE: 970018373876.9517
RMSE: 984895.1080581889
MAE: 506234.71788194444
RMSLE: 3.01032495077279
Mean Residual Deviance: 970018373876.9517

ModelMetricsRegression: gbm
** Reported on cross-validation data. **

MSE: 1197880285571.0645
RMSE: 1094477.174531778
MAE: 572647.022143943
RMSLE: NaN
Mean Residual Deviance: 1197880285571.0645
```

- Principais métricas:

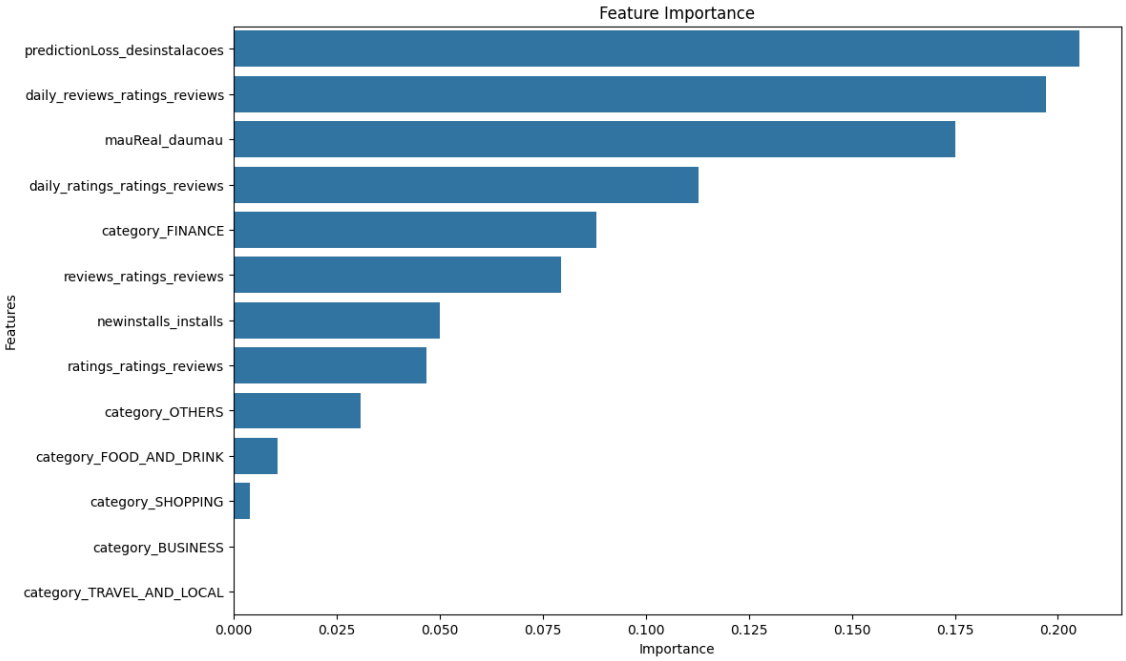
RMSE: 984895.1080581889

MAE: 506234.71788194444

MSE: 970018373876.9517

R^2: 0.188

- Feature Importance - É importante entender quais variável mais contribuem para explicar os resultados:



3- Considerações sobre os Resultados:

Olhando apenas para o R^2 de 0.1888 podemos concluir que: um R^2 de 0.1888 é relativamente baixo. Isso significa que apenas cerca de 18.88% da variabilidade dos dados pode ser explicada pelo modelo.

Em geral, um R^2 abaixo de 0.2 é considerado um sinal de que o modelo não está capturando bem a relação entre as variáveis preditoras e a variável alvo. Isso pode indicar que o modelo não é adequado ou que faltam variáveis importantes.