Relatório Técnico: Predição de Daily Active Users (DAU)

Vitor Salla

11 de julho de 2025

Processo de Coleta e Tratamento dos Dados

O primeiro passo do projeto consistiu na conexão com o banco de dados e na identificação das tabelas disponíveis. As seguintes tabelas foram encontradas:

- `daumau`
- `desinstalacoes`
- `installs`
- `ratings_reviews`

Dentre essas, a tabela `daumau` foi identificada como a principal fonte de dados para o objetivo de predição, pois contém as informações de **usuários ativos diários (DAU)** por aplicativo. Dado o tempo limitado para execução do projeto, optou-se por focar exclusivamente nesta tabela.

Durante a análise inicial da `daumau`, foram identificados alguns problemas de qualidade dos dados, como:

- Presença de registros duplicados;
- Aplicativos com elevada proporção de valores ausentes;
- Datas inconsistentes ou fora do intervalo esperado.

A Tabela a seguir apresenta estatísticas descritivas sobre os valores ausentes para alguns dos aplicativos com maior taxa de dados faltantes:

appId	Taxa de valores faltantes	Maior lacuna (dias)	Total de registros
com.app.36433	0,996	272	273
com.app.96037	0,996	272	273
com.app.28498	0,993	271	273
com.app.91706	0,960	262	273
com.app.97548	0,960	262	273
com.app.34279	0,956	261	273
com.app.35190	0,952	260	273
com.app.67629	0,949	259	273
com.app.66309	0,949	259	273
com.app.64667	0,945	258	273
com.app.60934	0,941	257	273
com.app.20103	0,934	255	273

Além disso, a Figura a seguir ilustra a distribuição temporal dos pontos disponíveis para os aplicativos analisados:

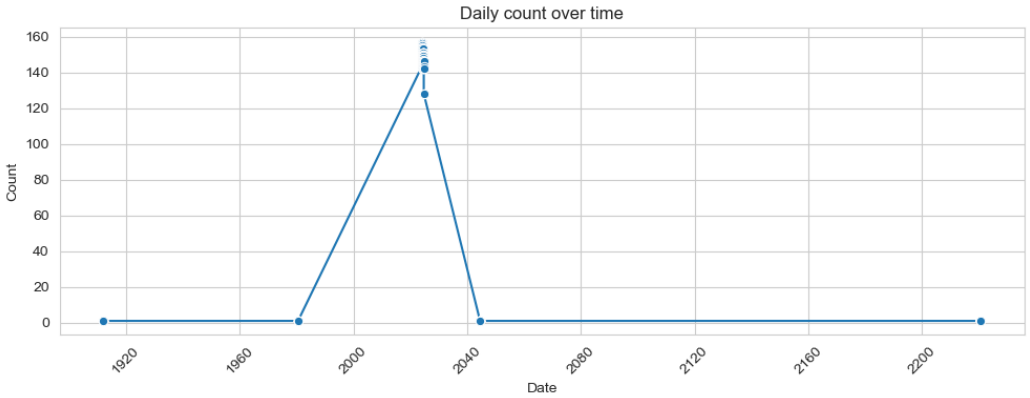


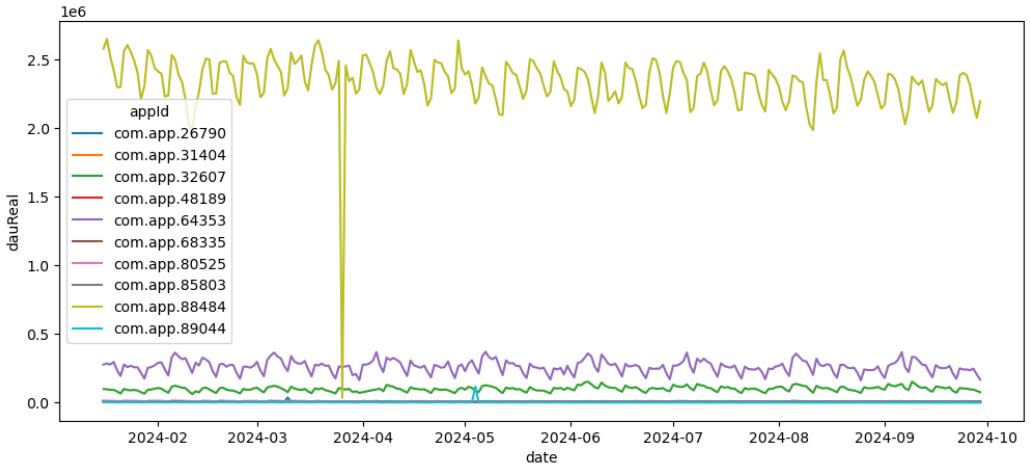
Figura 1: Distribuição temporal de registros por data

Com base na análise exploratória, definiu-se um intervalo temporal consistente para a análise: de **01/01/2024** a **29/09/2024**. Além disso, com o intuito de garantir a qualidade dos dados utilizados na modelagem, foram aplicados os seguintes critérios de filtragem para seleção dos aplicativos:

- **Taxa máxima de dias faltantes:** 30%;
- **Maior lacuna de dias consecutivos sem registro:** 7 dias.

Análise Exploratória, modelagem e treinamento

Após o pré-processamento dos dados, foi realizada uma análise exploratória do número de usuários ativos diários (DAU) por aplicativo. Identificaram-se alguns valores atípicos (outliers) em determinadas séries, conforme ilustrado na Figura a seguir:



Para mitigar a influência desses outliers no modelo, foi aplicada uma técnica de substituição baseada no escore-z (z-score). Considerou-se como outlier qualquer valor com escore-z superior a 3 em módulo, o que, assumindo uma distribuição normal, corresponde a valores fora de 99,73% da distribuição. Esses pontos foram substituídos pela média do DAU do respectivo aplicativo.

Em seguida, foram criadas variáveis derivadas com o objetivo de capturar padrões temporais e melhorar a performance preditiva do modelo. Optou-se pela utilização de um modelo de regressão, em vez de um modelo específico para séries temporais. Essa decisão baseia-se na maior simplicidade e escalabilidade do modelo de regressão, já que permite o treinamento conjunto de todos os aplicativos — bastando incluir o identificador do aplicativo (*appId*) como variável de entrada. Em contraste, um modelo de séries temporais exigiria o treinamento individualizado para cada aplicativo.

Para viabilizar a modelagem com regressão, foram criadas as seguintes variáveis:

- **Dia da semana**: categoria representando o dia da semana da observação.

- **Lags temporais**: valores de DAU dos dias anteriores à data de previsão. Os lags utilizados foram de 1, 2, 3, 4, 5, 6, 7, 10 e 15 dias.
- **Médias móveis**: médias dos DAUs dos últimos 3 e 15 dias.
- **Desvios padrões móveis**: desvios padrão dos DAUs dos últimos 3 e 15 dias.

O modelo escolhido foi o `XGBRegressor`, devido à sua robustez, capacidade de lidar com dados tabulares com variáveis numéricas e categóricas, e também devido a minha familiaridade com ele e com as técnicas de comitê de aprendizes. Os principais hiperparâmetros configurados para o modelo `XGBRegressor` foram:

- `objective` = reg:squarederror
- `learning_rate` = 0.1
- `max_depth` = 6
- `n_estimators` = 500
- `tree_method` = hist

Por motivos de tempo escasso, não foi possível fazer de forma efetiva ajuste nos parâmetros ou testar mais modelos.

Validação

Para a etapa de validação, foi reservado um conjunto de dados correspondente aos últimos 30 dias, o qual não foi utilizado durante o treinamento do modelo. Com esse conjunto, foram gerados gráficos ilustrando as previsões e calculadas métricas de desempenho para avaliação quantitativa, incluindo **MAE** (Mean Absolute Error), **RMSE** (Root Mean Squared Error) e **MAPE** (Mean Absolute Percentage Error).

É importante destacar que as previsões foram realizadas de forma **iterativa**, utilizando as saídas anteriores como entradas futuras, de forma a simular o cenário real de previsão e evitar vazamento de dados (*data leakage*).

A seguir, são apresentadas algumas visualizações da previsão de DAU para aplicativos específicos:

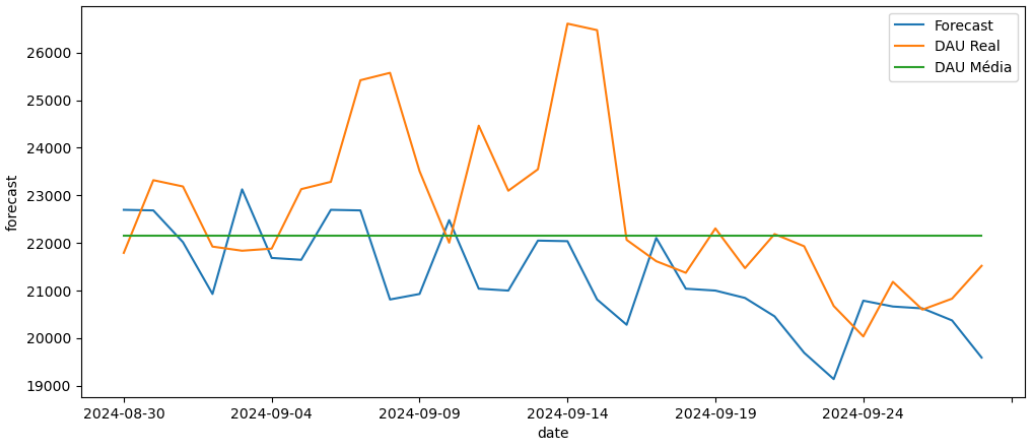


Figura 2: Previsão de DAU para o App 1

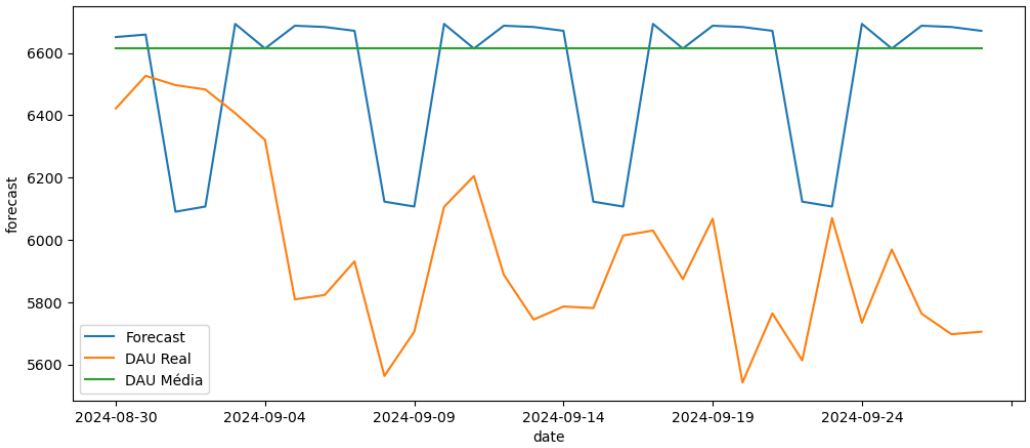


Figura 3: Previsão de DAU para o App 2

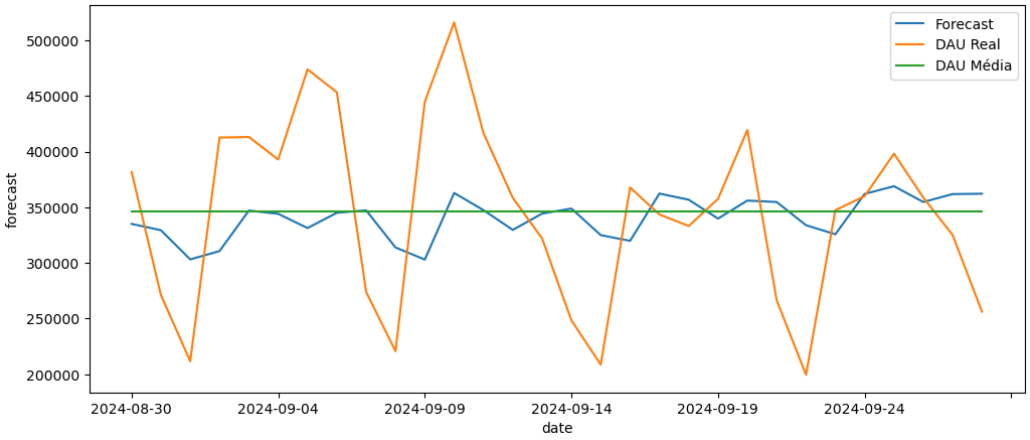


Figura 4: Previsão de DAU para o App 3

As métricas obtidas pelo modelo `XGBRegressor` foram:

- **MAE**: 55.652
- **RMSE**: 203.204
- **MAPE**: 18,3%

Como baseline, foi utilizado um modelo **dummy**, que prevê a média histórica de DAU de cada aplicativo. Os resultados para esse modelo foram:

- **MAE**: 54.168
- **RMSE**: 184.475
- **MAPE**: 23,5%

Observa-se que, embora o erro absoluto médio (MAE) do modelo preditivo seja levemente superior ao do modelo de média, ele apresenta uma pequena melhora no erro percentual (MAPE). Portanto, observa-se resultados pré-liminares e onde teria que realizar muitas melhorias para que o modelo de tornasse útil na prática.

Sugestões e Melhorias

A seguir, são apresentadas algumas sugestões que poderiam ser exploradas para aprimorar o desempenho do modelo desenvolvido:

Modelagem

Uma possibilidade de melhoria está na utilização de outras tabelas disponíveis no banco de dados, como `desinstalacoes` e `installs`. A partir delas, seria possível calcular a diferença entre instalações e desinstalações diárias, gerando um indicador da variação da base de usuários — um possível reflexo do engajamento e da atratividade do aplicativo.

Além disso, informações provenientes da tabela `ratings_reviews`, especialmente as avaliações mais recentes, poderiam ser utilizadas como proxy de satisfação ou insatisfação dos usuários, complementando os sinais utilizados na predição de DAU.

Outra sugestão seria enriquecer as variáveis relacionadas à data, incluindo feriados nacionais ou regionais, eventos sazonais e outras datas especiais, que podem influenciar diretamente o comportamento de uso dos aplicativos.

Treinamento

No que se refere à etapa de treinamento, uma melhoria importante seria a utilização de técnicas de otimização de hiperparâmetros, como *Grid Search*, *Random Search* ou *Bayesian Optimization*, a fim de encontrar as melhores configurações para o modelo.

Além disso, seria relevante comparar o desempenho do `XGBRegressor` com outros algoritmos, como o `LGBMRegressor`, que possui características semelhantes, ou ainda com modelos específicos para séries temporais, como *ARIMA*, *Prophet* ou modelos baseados em redes neurais (e.g., LSTM).

Infelizmente, devido a restrições de tempo, essas abordagens não puderam ser exploradas.