Relatório Técnico – Predição de DAU

1. Introdução

- O projeto tem como objetivo a previsão do valor de DAU (Daily Active Users) de diversos apps.
- A previsão teve foco em dois pontos principais para cada app:
    - **Daureal d+1**: prever o 'daureal' de amanhã,
    - **Daureal d+7 soma móvel:** nos próximos 7 dias, qual é a soma dos DAU?

---

2. Coleta e Tratamento de Dados

2.1 Conexão com o Banco de Dados

- O banco de dados foi acessado através do sqlalchemy
- As tabelas foram baixadas e salvas em csv no diretório ./data/raw
- Tamanho inicial da base: 53657

2.3 Limpeza e Preparação dos Dados

- As tabelas no geral não apresentaram inconsistências, com excessão de:
    - **Daumau**: datas estranhas, como anos 2220 e 1912. Esses anos foram ignorados por falta de evidências de erro de digitação. (Era possivel que fosse 2024, como todo o resto da base)
    - **Installs + Desinstalações**: valores faltantes apareceram depois do merge nas colunas "date" e "appid"
    - **Rating reviews**: daily ratings e daily reviews negativos em alguns dias, podem ser remoções de usuários ou da plataforma. Foram mantidos.

- Remoção de NaNs:
    - Contagem de NaNs por apps: apps com muitos NaNs foram retirados.
    - Se 30% ou mais das colunas disponíveis forem NaNs, o app é ignorado.
    - Se 30% ou mais de "daureal" é NaN, o app é ignorado

- Preenchimento de NaNs:
    - Os valores NaN restantes foram preenchidos com a mediana dos valores presentes de cada app.

- Outliers:
    - Foram tratados posteriormente juntamente com Scaling na modelagem.

2.4 Exportação dos Dados Tratados

- O arquivo final gerado foi a base tratada no caminho "./data/base_tratada.csv"

- Tamanho final da base tratada: 34292

- Possui as seguintes colunas:
    - appId: str
    - date: datetime
    - newinstalls: int
    - predictionloss: int
    - daureal: int
    - maureal: int
    - category: str
    - ratings: int
    - daily_ratings: int
    - reviews: int
    - daily_reviews: int

---

3. Análise Exploratória e Modelagem

3.1 Análise Exploratória de Dados (EDA)

- Correlações de Pearson:
    - Observação: número de desinstalações passadas não tem correlação com DAU futuro. Diff "instalações – desinstalações" tem correlação negativa com DAU.
      Bastante fora do esperado. Pode ser explicado com campanhas de recuperação após queda do DAU, usuários pouco engajados que demoram pra entrar no app ou churn imediato.

- Foram feitas 2 correlações numéricas:

    - **correlação total**: correlação feita com todos os dados juntos
    - **correlação por appid (melhor versão):** correlação individial para cada appid, e depois feito a média geral das correlações

- **Correlação categórica**:
    - Dia_semana e target: 0.31
    - Método usado: Mutual Information

- **Principais variáveis de correlação**: daureal, maureal, daily_reviews, target_lag, target_lag7, daureal_maurea, dia_semana

- **Correlação média com o target**: 0.43

3.2 Outliers e Scaling

- As features e target foram escalados considerando o seguinte:
    - Sem valores negativos: normalização logarítmica
    - Com valores negativos: normalização winsorize

3.3 Escolha e Justificativa do Modelo

- Seleção de features:
    - As features foram selecionadas com base em correlações com a variável target e a feature_importance de modelos de regressão.

- Modelos utilizados:
    - LME: Modelo misto com foco em usar dia_semana como efeito aleatório. Além do efeito, também é transparente. Inicialmente foi considerado usar appId como efeito aleatório também, mas GroupKFolding pareceu uma opção melhor.

    - GLM: Usando a família Poisson ou Negative Binomial, o GLM modela diretamente distribuilção de contagem (ótimo para o target DAU). Efeitos de dia da semana e categoria entram de forma natural, os coeficientes são facilmente interpretados.

3.4 Treinamento e Tuning

- Estratégia de treino:
    - Grids de hiperparâmetros aplicada aos 2 modelos
    - MLFlow para acompanhar métricas de desempenho
    - O melhor modelo foi salvo em "./models/mlruns"

- Validação cruzada: GroupKFolding, usando appid como grupo
- Foi separado um conjunto de appids para validação (~6000)

---

4. Validação e Avaliação do Modelo

4.1 Coleta de Dados para Validação

- Nova consulta SQL e tratamento dos dados de validação
- Carregamento dos modelos MLFlow armazenados localmente

4.3 Métricas de Avaliação

- Métricas utilizadas: MedAPE e RMSE
    - MedAPE: Mesma escala do DAU, absoluta, reflete desempenho global.
    - RMSE: Percentual e adimensional, usa mediana (pouco influenciada por valores extremos), relativa.

- Resultados target d+1:
    - Durante a validação, o LME se mostrou pior que a baseline. Por outro lado, o GLM perdeu apenas 0.5% de MedAPE.

    - Baseline treino:
        - RMSE = 269.267
        - MedAPE = 10.43%

    - LME treino:
        - RMSE = 189.014
        - MedAPE = 9.62%

    - GLM treino
        - RMSE = 147.561
        - MedAPE = 5.97%

    - LME validação:
        - RMSE = 333.411
        - MedAPE = 11.22%

    - GLM validação
        - RMSE = 139.689
        - MedAPE = 6.55%

- Resultados target d+7:
    - O treino do target d+7 superou muito as expectativas, e novamente o GLM apresentou resultados ótimos de 4.75% MedAPE, até melhores do que a previsão d+1.

    - Baseline treino:
        - RMSE = 8.826.680
        - MedAPE = 85.42%

    - LME treino:
        - RMSE = 1.557.383
        - MedAPE = 9.28%

    - GLM treino
        - RMSE = 724.486
        - MedAPE = 3.65%

    - LME validação:
        - RMSE = 1.244.893
        - MedAPE = 6.43%

    - GLM validação
        - RMSE = 1.415.330
        - MedAPE = 4.75%

4.4 Visualizações de Performance

- Os 12 gráficos plotados estão no notebook "models_validation.ipynb". São os seguintes gráficos:
    - Resíduos vs Fitted
    - Real vs Predito
    - Histograma de Resíduos

- Os gráficos descrevem os modelos LME e GLM, com os targets d+1 e d+7;

- Insights Resíduos vs Fitted:
    - Resíduos centrados em 0, o que é bom
    - Maior variância de resíduos nas pontas: incerteza em valores extremos (esperado)

    - GLM tem comportamento diferente: a variância é mais constante do que no LME

- Insights Real vc Predito:
    - Grande dispersão em todos os modelos, indicando erro na previsão
    - Modelo parece subestimar valores maiores

- Histograma de resíduos:
    - Muito concentrada em torno do 0 (ótimo)
    - Assimetria leve e caudas longas: outliers e previsões ruins nos extremos

---

5. Conclusões e Recomendações

5.1 Principais Conclusões

- O melhor modelo testado foi o GLM:
    - Target d+1: MedAPE 6.55%
    - Target d+7: MedAPE 4.75%

- No geral, o padrão notado foi: o modelo **generaliza bem**, mas **falha em valores extremos** e outliers de previsão.

5.2 Possíveis Melhorias

- Melhorar a categorização de appids: apps mais segmentados podem ter previsões mais precisas.
- Testagem de novos modelos: modelos de previsão de tendência, como ARIMA ou Prophet com variáveis exógenas, como feriados.
- Investigação aprofundada nas correlações estranhas (install-uninstall)
