tiendanube nuvemshop

Previsão de Quality Leads (QL)

Murilo Mazzotti Silvestrini

https://github.com/muriloms/nuvemshop-prediccion-ql

Sumário

Visão Geral e Objetivos

Problema

- Quando uma pessoa registra sua loja na plataforma, ela tem um período de teste de 30 dias nos quais pode decidir se continua ou não usando o serviço. Durante o período de teste, a loja é um teste. Após os 30 dias, a loja pode optar por continuar usando o serviço, mas pagando por ele, nesse caso, ela se torna um pagamento.

Objetivo

- Desenvolver um modelo preditivo para identificar quais lojas, atualmente em período de teste de 30 dias, têm maior probabilidade de se tornarem assinantes pagantes. O modelo ajudará a priorizar o atendimento e o suporte, focando nos clientes em potencial que demonstram maior interesse em continuar utilizando o serviço após o período de teste gratuito. Com isso, otimizar os recursos e garantir a eficiência do atendimento, frente ao crescente volume de novos registros diários na plataforma

Tarefas

- Desenvolver um ou mais modelos que permitam prever com maior precisão os testes que se converterão em pagamentos.
- Explicar brevemente qual foi o critério pelo qual o modelo foi escolhido e por que essa métrica foi selecionada.
- Obter pelo menos 3 insights do modelo que possam ser úteis para o negócio

Etapas

1 Preparação dos dados

- Revisão dos tipos de dados presentes no dataset.
- Identificação de necessidades de conversões ou ajustes para compatibilizar os tipos de dados com os algoritmos de machine learning
- Detecção e quantificação de valores ausentes nos dados.
- Avaliação das melhores abordagens para tratar valores ausentes
- Realização de contagem das classes presentes nas variáveis categóricas para identificar desequilíbrios.
- Consideração de técnicas de balanceamento de classes

2 Exploração dos dados

- Utilização de gráficos de barras para análise da frequência das variáveis.
- Emprego de histogramas para visualizar a distribuição dos dados.
- Uso de box plots
- Testes estatísticos para identificar padrões
- Análise de variabilidade, médias e desvios padrão.
- Análise de correlação para entender como as variáveis estão inter-relacionadas.
- Utilização das informações coletadas nas visualizações e análises estatísticas para selecionar as variáveis mais relevantes para os modelos preditivos

3 Machine Learning

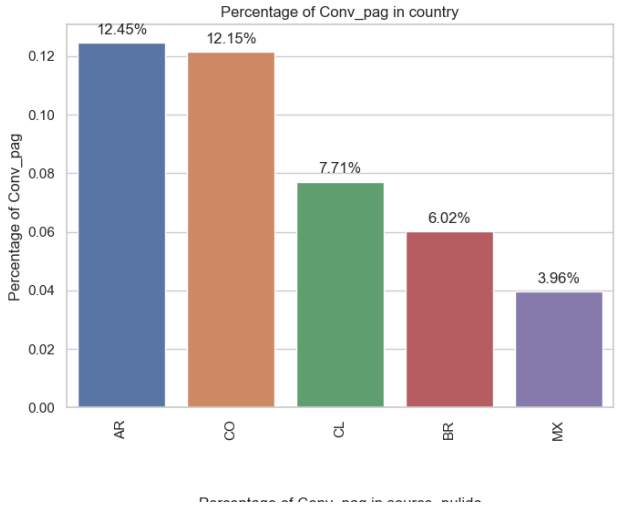
- Preparação dos Dados para Modelagem
- Divisão dos Dados em Treino e Teste
- Realização de seleção de features para identificar as variáveis mais impactantes
- Aplicação e Avaliação Inicial de Vários Modelos
- Métricas como acurácia e a área sob a curva ROC (AUC) são utilizadas para avaliar e comparar os modelos
- Realização de ajustes nos hiperparâmetros dos modelos selecionados para otimizar ainda mais seu desempenho
- Exportação dos Modelos Ajustados e Artefatos Relevantes

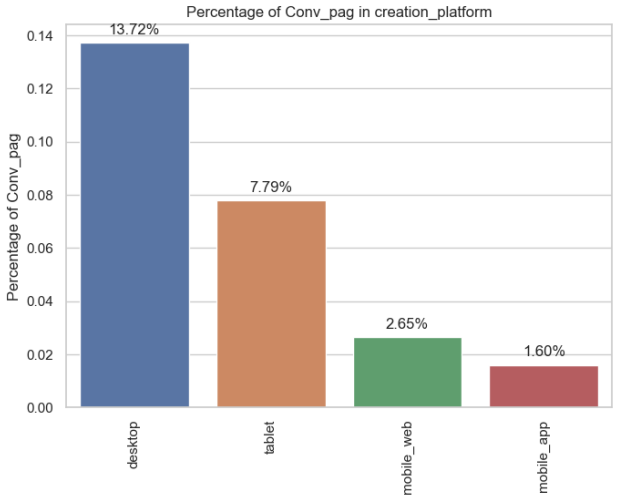
4 Deploy

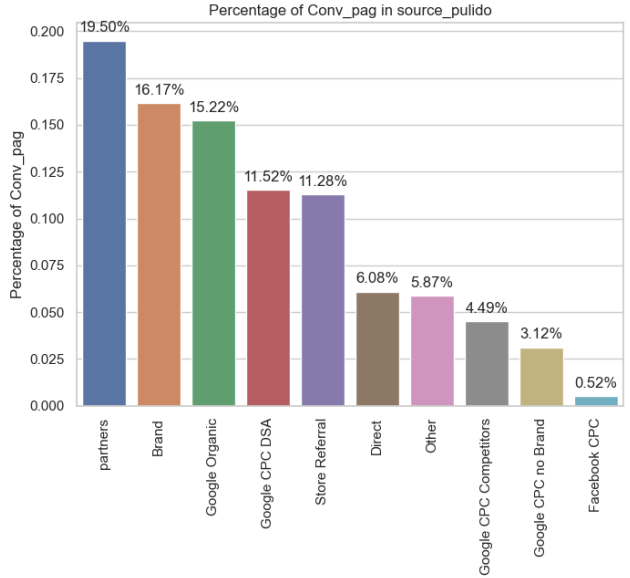
- Preparação dos dados novos ou de teste seguindo o mesmo procedimento utilizado durante o treinamento dos modelos.
- Aplicação de encoders nas variáveis categóricas e normalização das variáveis numéricas para assegurar a consistência e a precisão nas previsões
- Realização de Previsões com os Modelos Selecionados
- Análise de como cada modelo responde a novos dados, observando a precisão e a relevância das previsões feitas

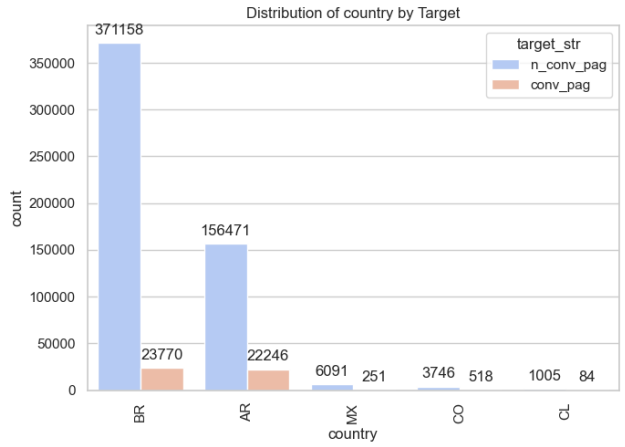
Análise Exploratória

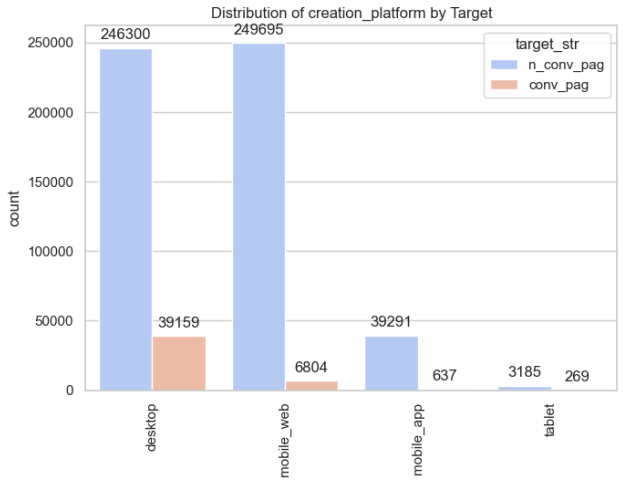
Variáveis categóricas

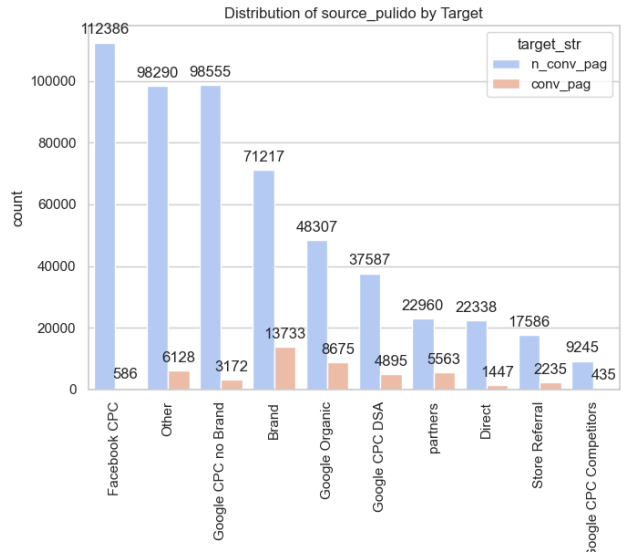






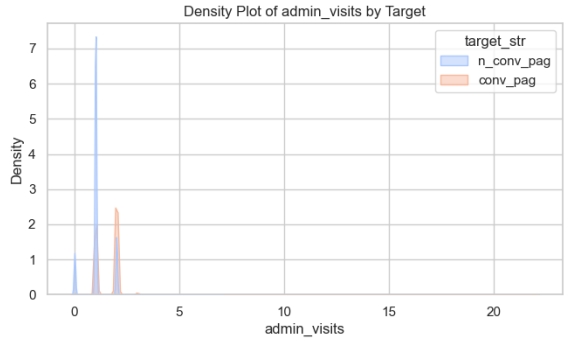


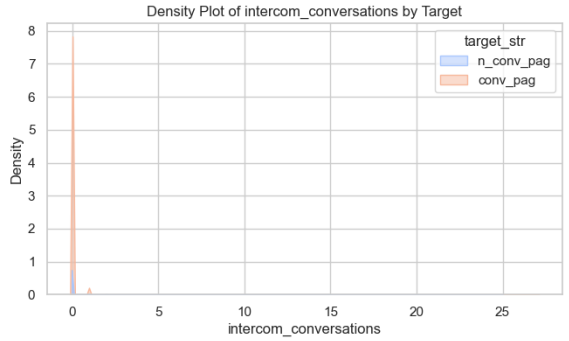


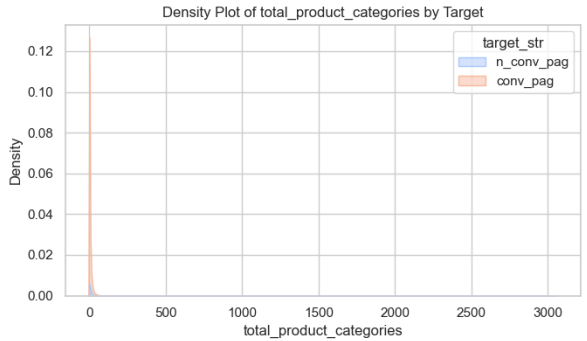


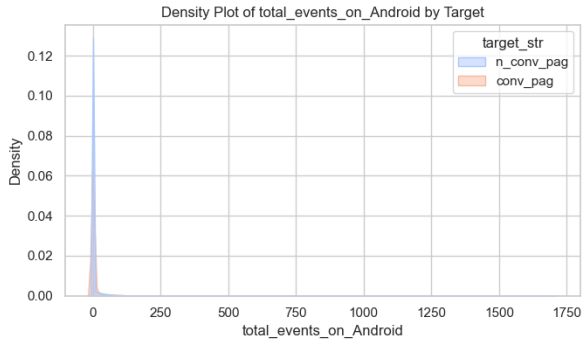
Análise Exploratória

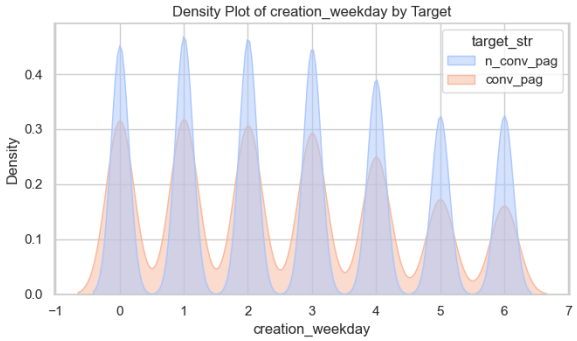
Variáveis numéricas

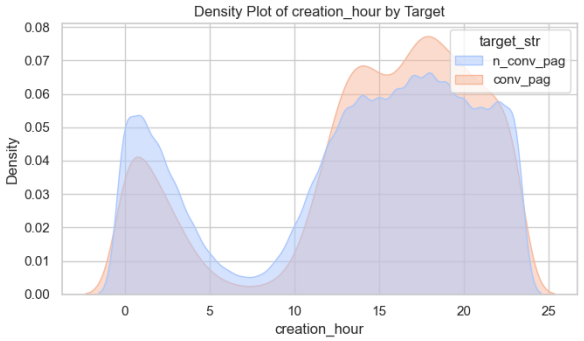


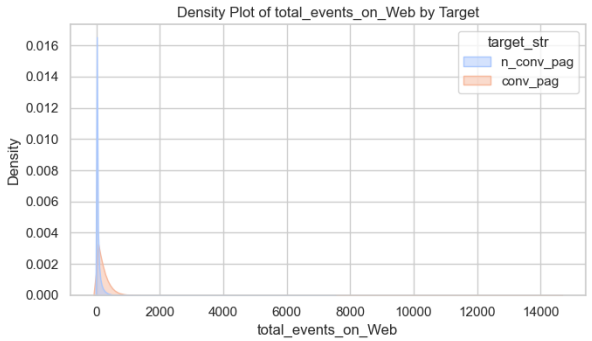


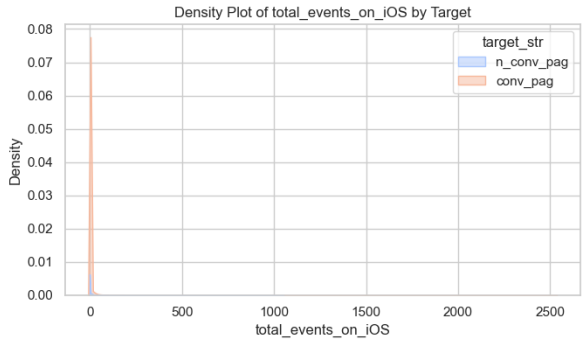


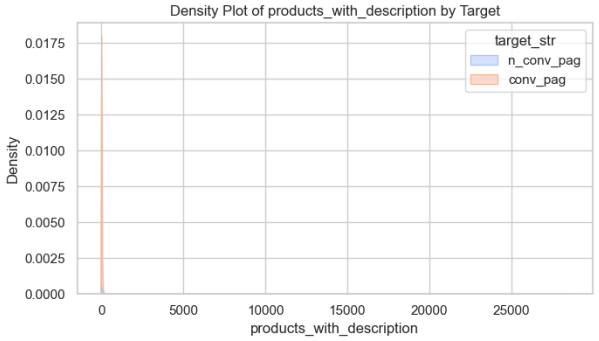


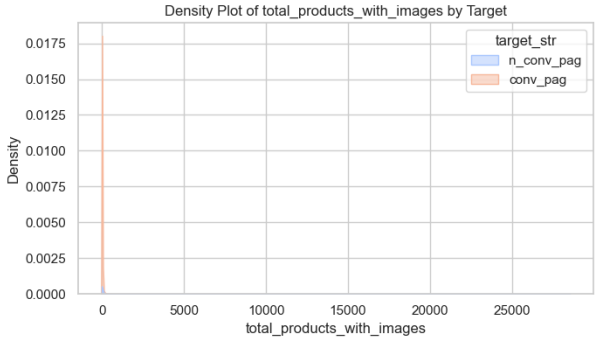






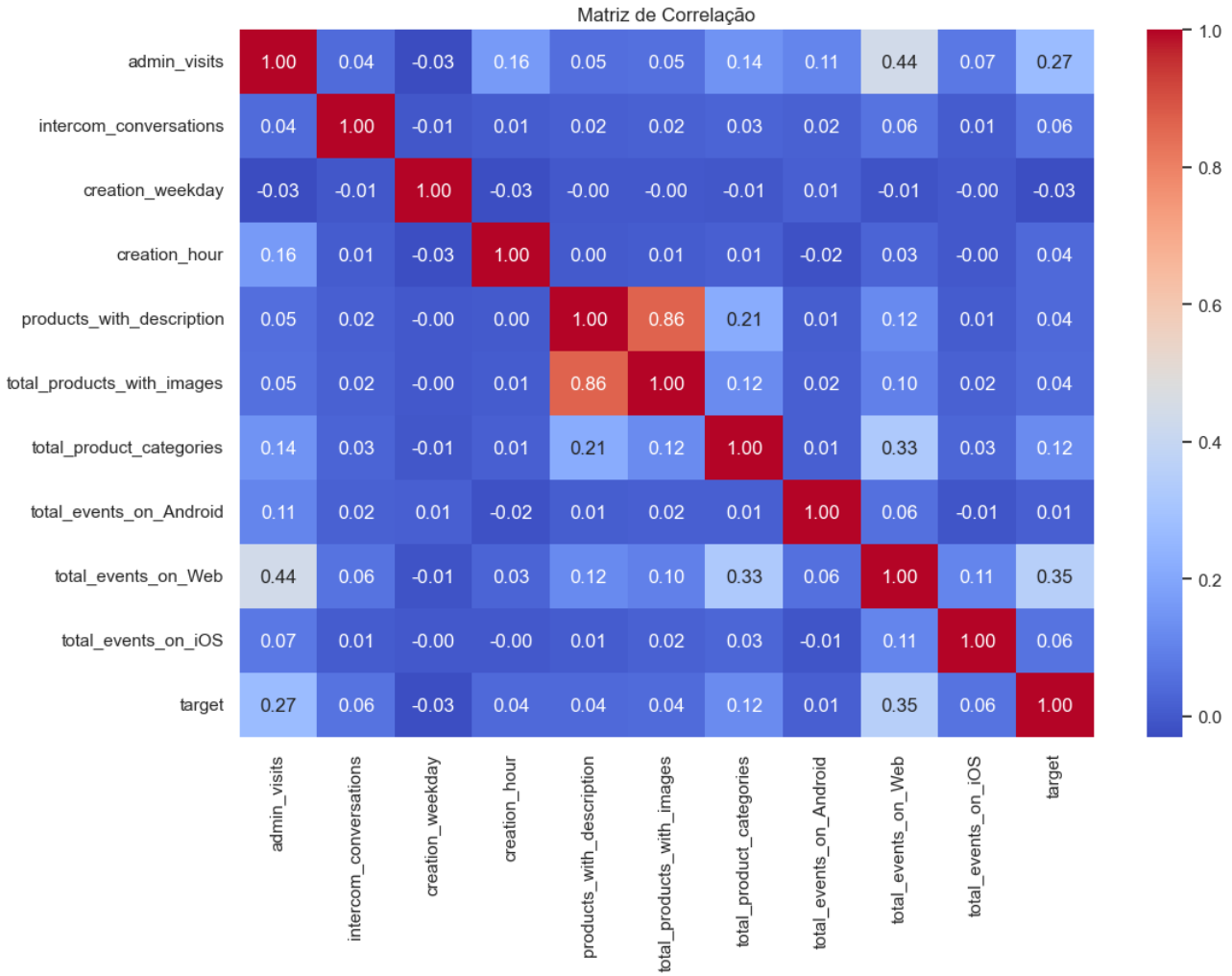




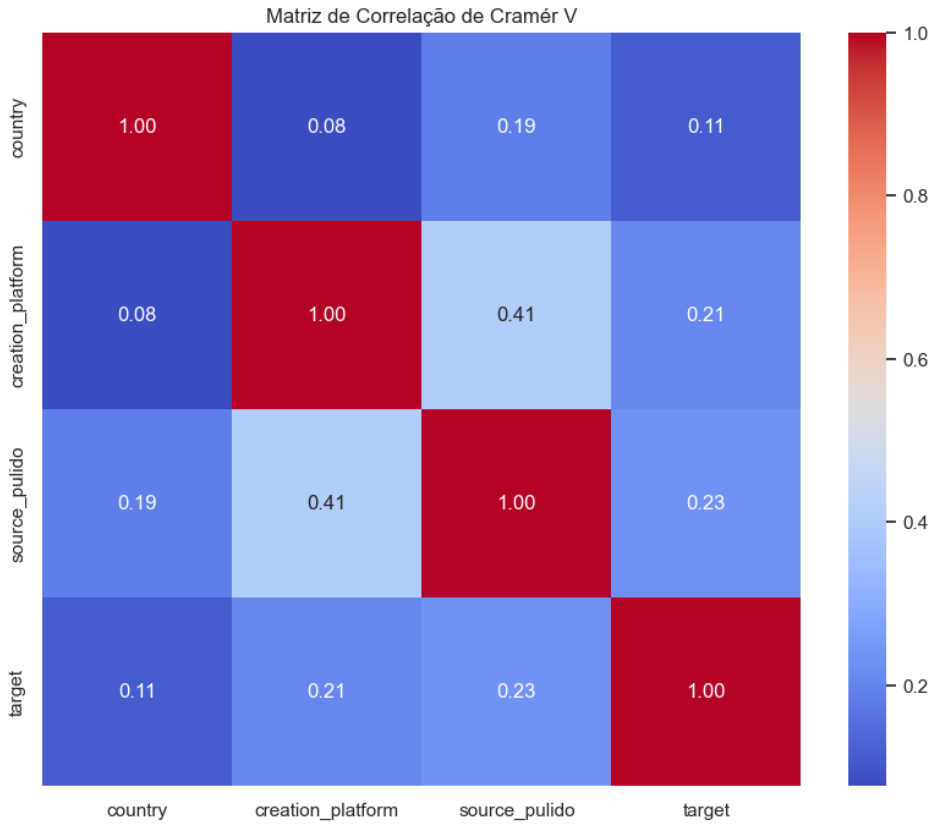


Análise de Correlação

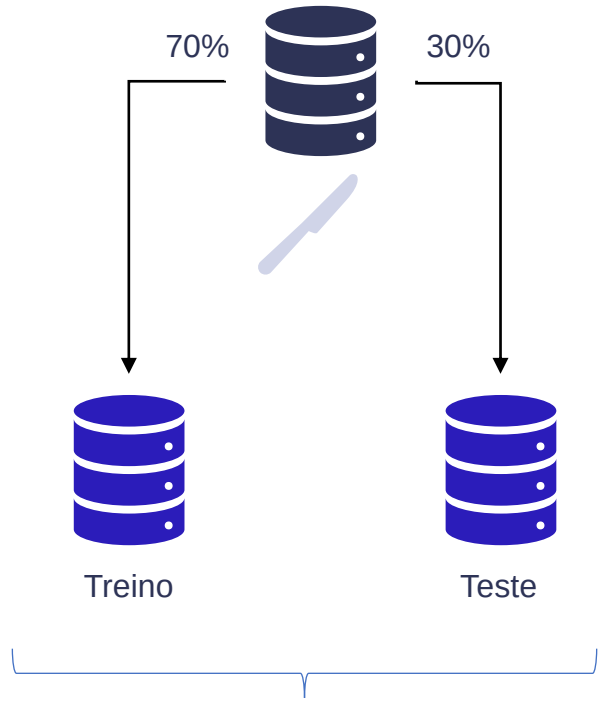
Variáveis numéricas



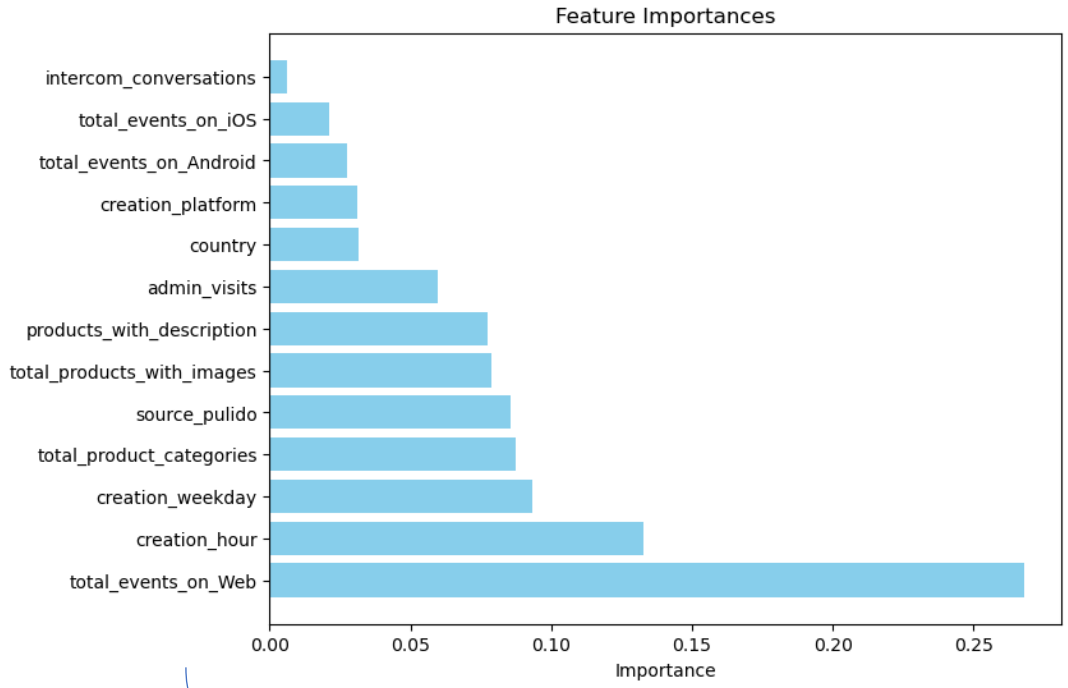
Variáveis categóricas



Preparação dos dados



Feature selection

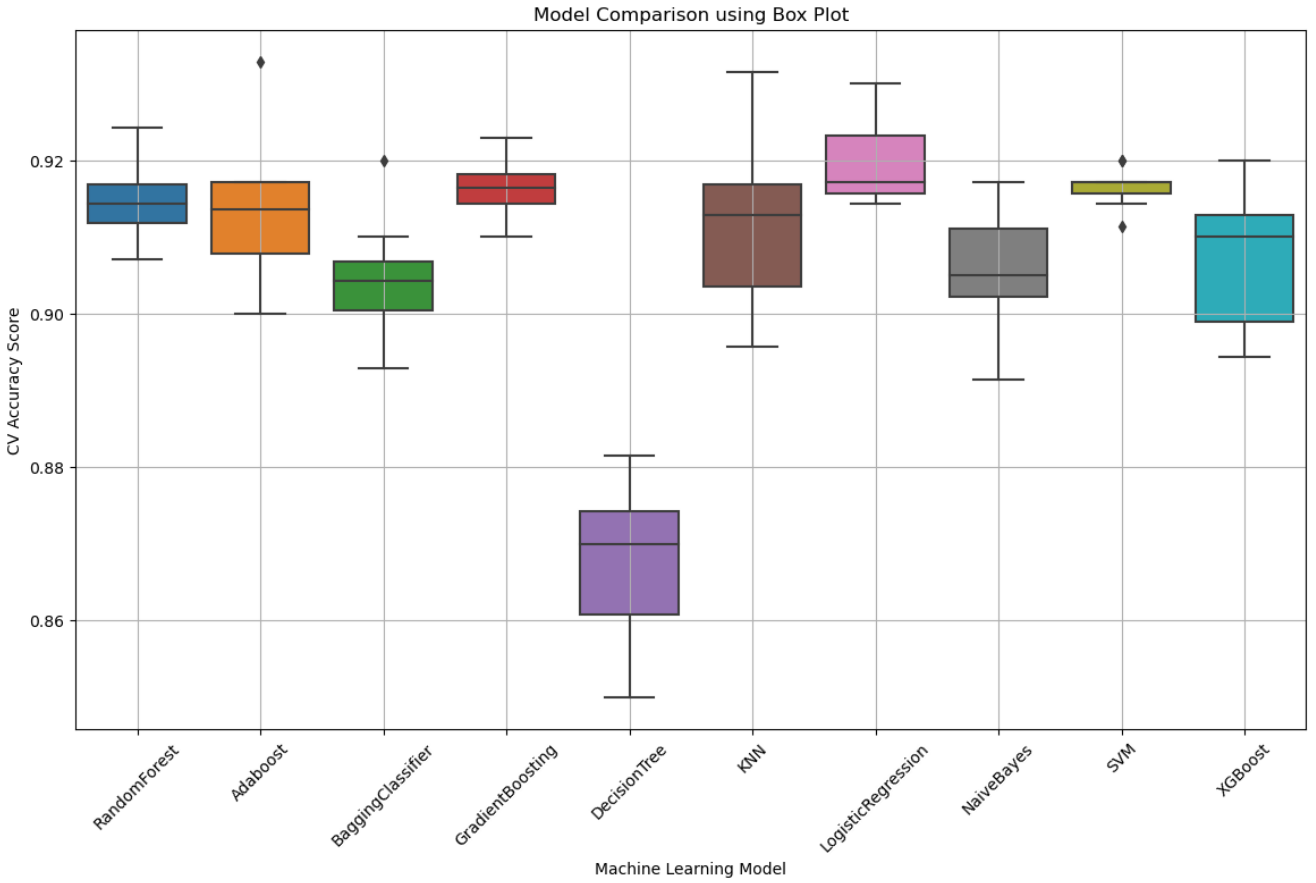


- codificação de variáveis categóricas
- normalização de variáveis numéricas

- total_events_on_Web
- creation_hour
- creation_weekday
- total_product_categories
- source_pulido
- total_products_with_images
- products_with_description
- admin_visits

Seleção dos Modelos

Model	
LogisticRegression	0.919714
GradientBoosting	0.916571
SVM	0.916571
RandomForest	0.914429
Adaboost	0.913286
KNN	0.911714
XGBoost	0.907143
NaiveBayes	0.905286
BaggingClassifier	0.904571
DecisionTree	0.868143



Seleção dos Modelos

1. Capacidade de Lidar com Dados Desbalanceados

RandomForest e Adaboost são conhecidos por lidarem relativamente bem com dados desbalanceados. O RandomForest, por exemplo, pode minimizar o viés introduzido por classes desbalanceadas através da construção de árvores de decisão que se concentram em classificar corretamente os casos minoritários. Adaboost ajusta os pesos das instâncias de treinamento, dando mais peso aos casos mal classificados, o que pode ser benéfico em situações de desbalanceamento. SVM também pode ser eficaz para dados desbalanceados, embora possa exigir mais ajustes e experimentação para obter resultados ótimos em comparação com métodos baseados em árvores.

2. Performance em Grandes Conjuntos de Dados

RandomForest e Adaboost são algoritmos que se escalam relativamente bem com grandes conjuntos de dados. SVM pode ser menos escalável para grandes volumes de dados devido ao seu custo computacional elevado. No entanto, a sua inclusão pode ser justificada pela qualidade das fronteiras de decisão que consegue produzir em problemas complexos.

3. Generalização para Dados Não Vistos

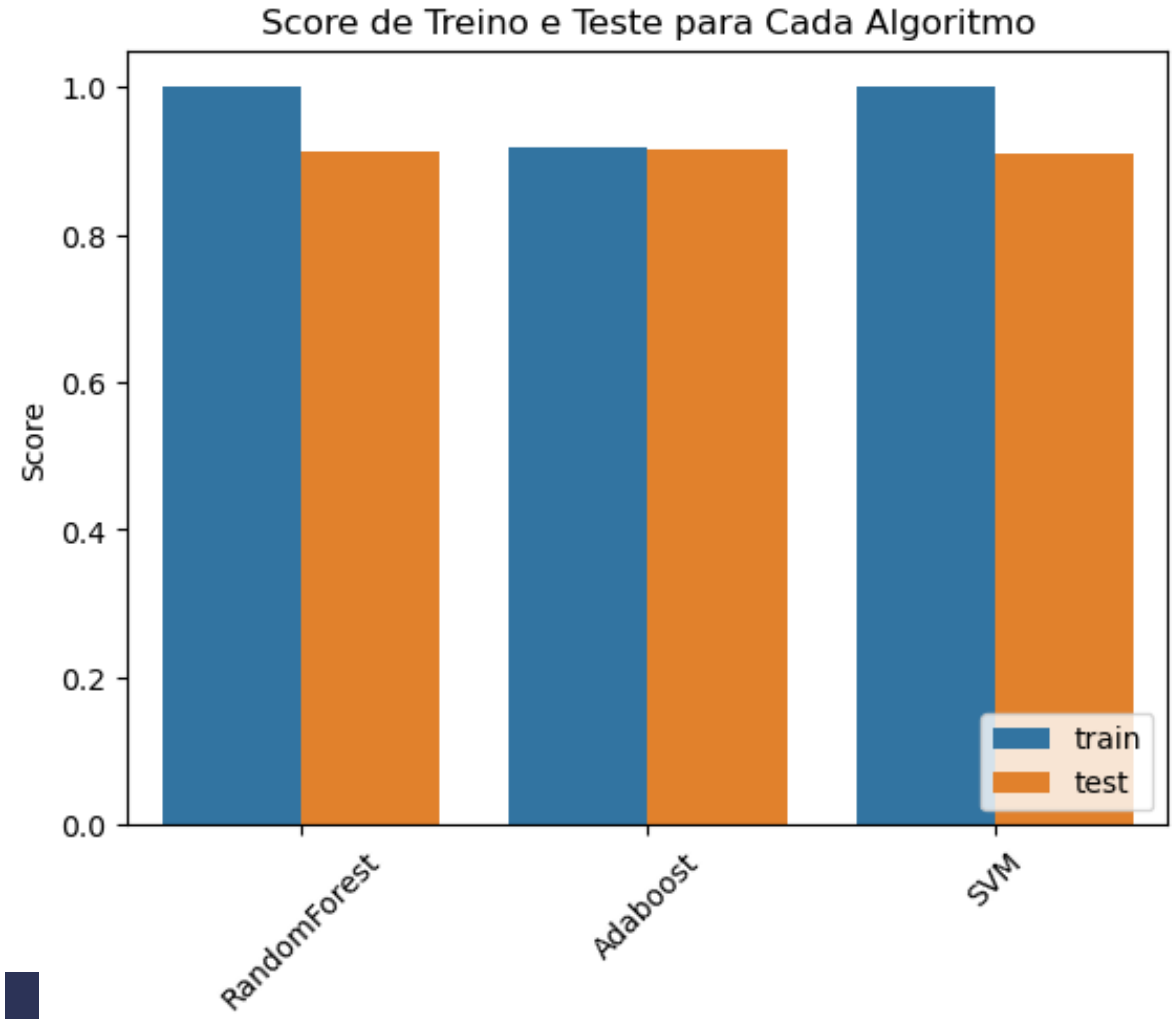
Algoritmos como RandomForest e Adaboost tendem a generalizar melhor para dados não vistos comparativamente a modelos como Logistic Regression e Gradient Boosting. RandomForest, em particular, por usar múltiplas árvores de decisão para fazer suas previsões, reduz o risco de overfitting. Adaboost ajusta iterativamente os pesos aplicados aos classificadores e dados, o que ajuda a melhorar a capacidade de generalização do modelo.

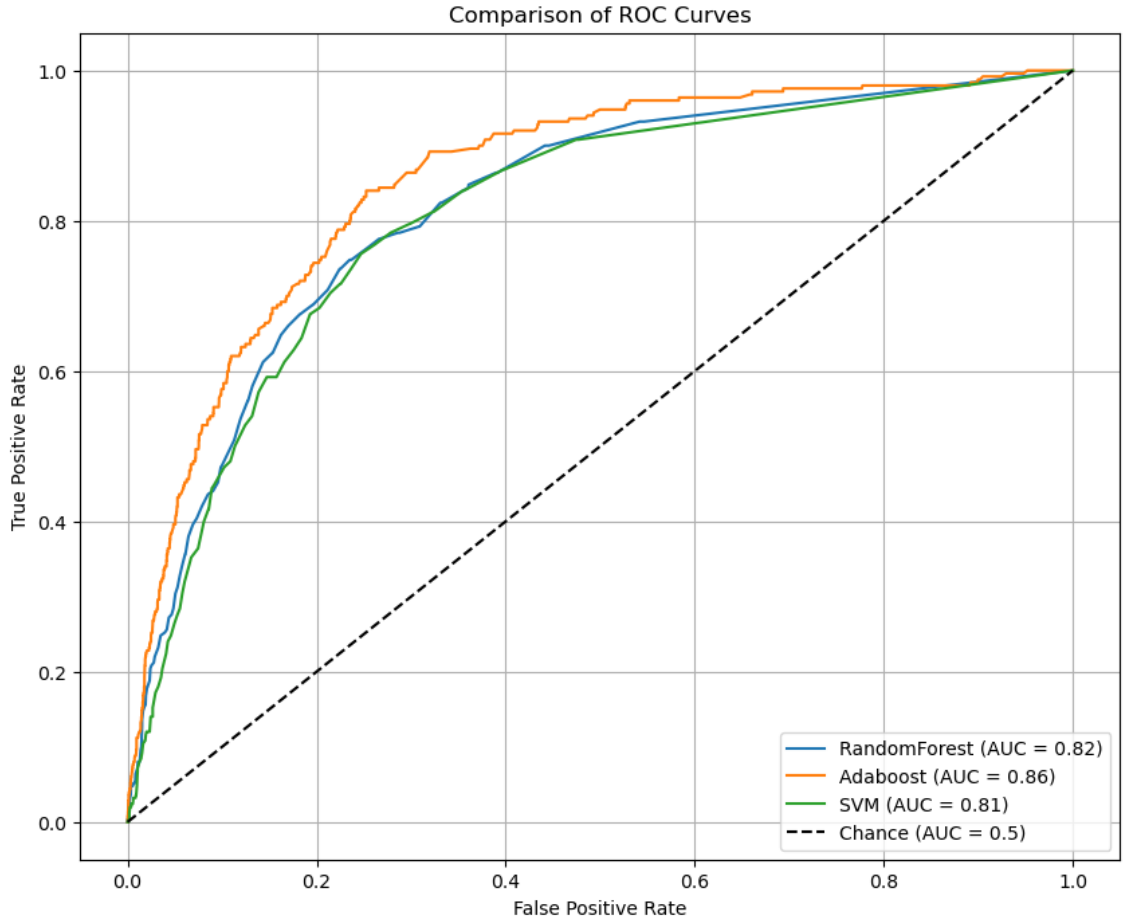
4. Robustez a Ruídos e Outliers

RandomForest é bastante robusto a ruídos e outliers devido ao método de bagging (bootstrap aggregating), que reduz a variância sem aumentar o viés. Isso o torna uma escolha atraente em dados complexos e desordenados.

Embora modelos como Logistic Regression e Gradient Boosting possam ter apresentado desempenhos ligeiramente superiores em termos de métricas específicas como acurácia, a escolha do RandomForest, Adaboost e SVM pode ser direcionada por suas características de robustez, capacidade de lidar com desbalanceamento e escalabilidade. Além disso, a diferença de desempenho não é substancialmente grande, o que permite optar por modelos que podem oferecer outras vantagens estratégicas e operacionais no contexto do problema específico enfrentado.

Análise dos modelos selecionados





Tuning de hiperparâmetros

Random Forest Classifier

```
Best Score: 0.916
Best Parameters: {'classifier__n_estimators': 100, 'classifier__max_depth': 10, 'classifier': RandomForestClassifier(max_depth=
10, random_state=42)}
```

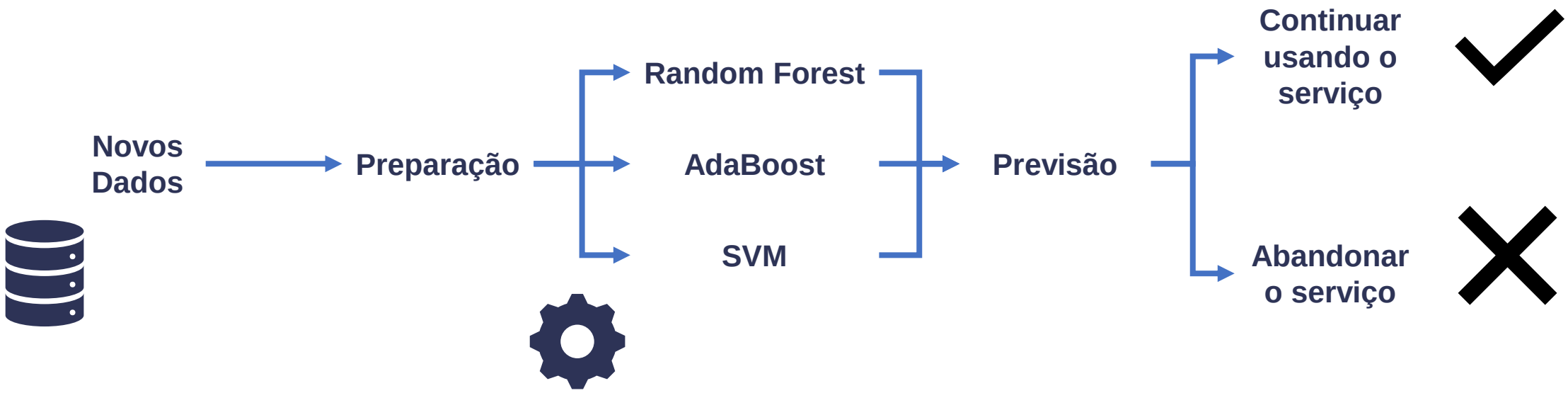
AdaBoost Classifier

```
Best Score: 0.9174285714285715
Best Parameters: {'classifier__n_estimators': 100, 'classifier__learning_rate': 0.1, 'classifier': AdaBoostClassifier(learning_
rate=0.1, n_estimators=100, random_state=42)}
```

Suport Vector Machine Classifier (SVM)

```
Best Score: 0.9167142857142856
Best Parameters: {'classifier__kernel': 'linear', 'classifier__C': 1, 'classifier': SVC(C=1, kernel='linear', probability=True,
random_state=42)}
```

Implementação



Implementação

Streamlit

https://streamlit.io/

1. Instalar o streamlit

pip install streamlit

2. Abrir prompt na pasta:

...\nuvemshop-prediccion-ql\codes\streamlit

3. Executar o comando:

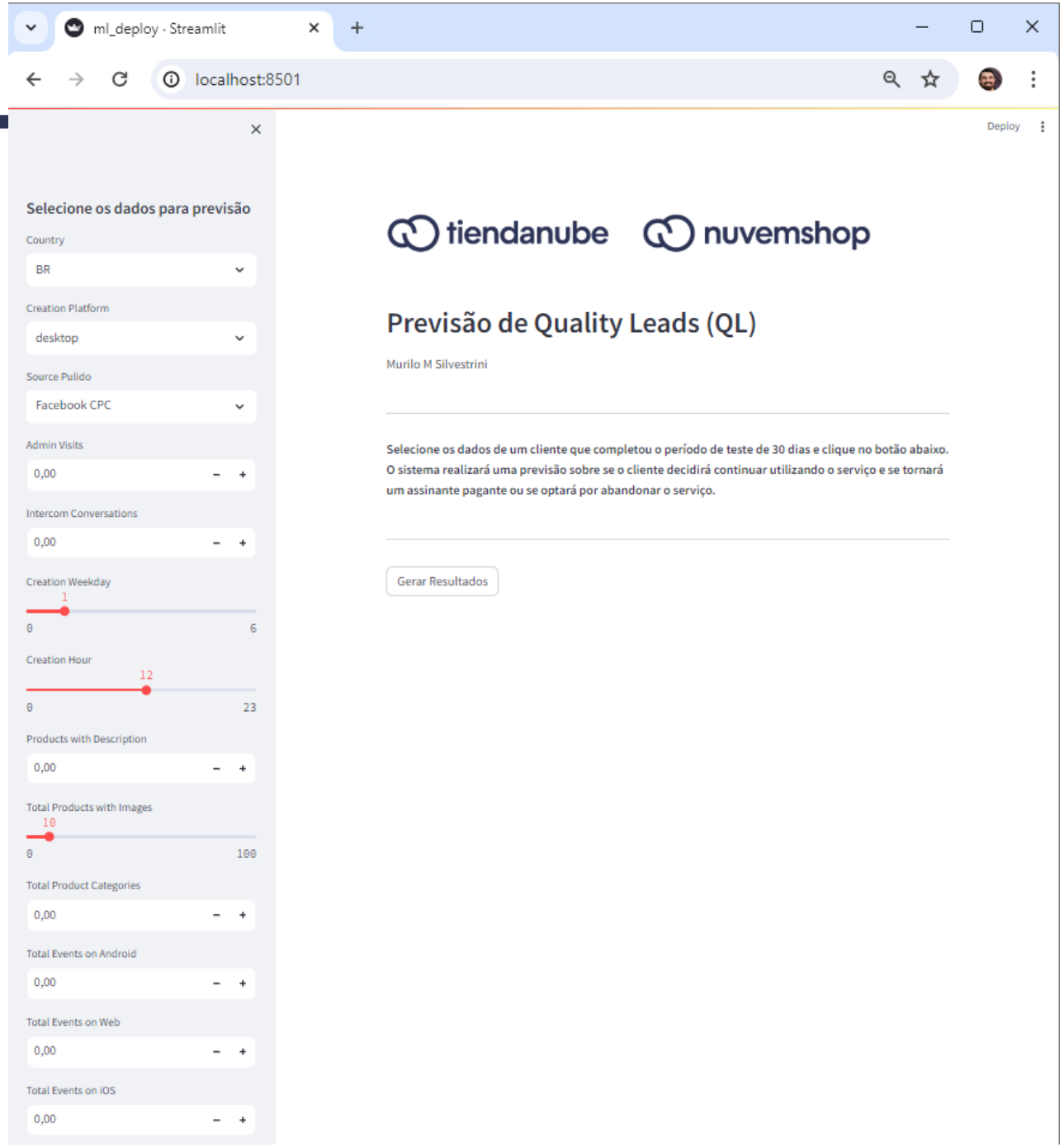
streamlit run ml_deploy.py

ou:

python –m streamlit run ml_deploy.py

4. Abri no navegador:

http://localhost:8501



Estratégias

- **Desconto de incentivo:** Ofereça um desconto exclusivo para esses clientes como um incentivo para que eles formalizem a assinatura após o período de teste.
- **Teste A/B de campanhas de engajamento:** Divida esses clientes em dois grupos. Um grupo recebe uma campanha promocional específica, enquanto o outro não. Analise a eficácia da campanha em aumentar a taxa de conversão para assinantes pagantes.

- **Intervenção personalizada:** Contate esses clientes com mensagens personalizadas destacando os benefícios e recursos que podem não ter sido totalmente explorados durante o período de teste.
- **Teste A/B de retenção:** Similarmente aos clientes que continuarão, divida os clientes em risco de abandono em dois grupos. Aplique estratégias de retenção diferentes, como suporte proativo ou ofertas especiais, e compare os resultados para identificar as abordagens mais eficazes.
- **Pesquisa de feedback:** Envie uma pesquisa para entender as razões do possível abandono e use essas informações para melhorar os serviços ou ajustar as comunicações futuras.

Insights

Matriz de Correlação

Visitas de administração e eventos na web: Há uma correlação positiva entre as visitas de administração e os eventos totais na Web, o que pode indicar que uma interação mais frequente com a interface de administração está associada a uma maior atividade no site.

- Insight: Incentivar o engajamento ativo na plataforma através da interface de administração pode aumentar a retenção do cliente.

Produtos com descrição e com imagens: A alta correlação entre produtos com descrição e produtos com imagens sugere que os usuários que dedicam tempo para adicionar descrições tendem também a incluir imagens.

- Insight: A plataforma pode enfatizar a importância de completar as informações do produto, oferecendo tutoriais ou ferramentas de preenchimento fácil, para melhorar a apresentação e potencialmente aumentar as conversões.

Importância dos Recursos

Eventos na web como recurso mais importante: A variável total de eventos na Web é destacada como o fator mais significativo.

- Insight: Melhorar a experiência do usuário na Web e fornecer mais funcionalidades que possam aumentar a interação do usuário na plataforma podem ser chave para conversão de clientes em teste para pagantes.

Importância menor do dia da semana e da hora de criação: Estes fatores têm importância menor na previsão, sugerindo que o momento da inscrição é menos relevante.

- Insight: As campanhas de marketing e suporte ao cliente não precisam ser tão específicas no tempo, mas devem se concentrar mais em como os usuários interagem com a plataforma.

Gráfico de distribuição

Plataforma de criação: Há diferenças na taxa de conversão com base na plataforma de criação, com desktop liderando sobre mobile.

- Insight: Refinar a experiência do usuário em plataformas móveis ou entender por que usuários de desktop têm maior conversão pode ajudar a ajustar estratégias de produto e marketing.

Fonte de tráfego: Algumas fontes de tráfego, como o Google Orgânico, têm maiores taxas de conversão.

- Insight: Focar esforços e recursos em otimização para mecanismos de busca (SEO) e outras fontes de alto desempenho pode aumentar as taxas de conversão.

Próximos passos

- Explorar os dados com modelos estatístico para aprofundar a análise da relação entre as variáveis e o seu comportamento
- Testar novas estrutura de modelos e realizar a avaliação comparativa
- Criar novas variáveis baseada nas informações das variáveis coletadas (eature engineering)
- Realizar o balanceamento dos dados com técnica como o Smote
  - https://medium.com/data-hackers/como-lidar-com-dados-desbalanceados-em-problemas-de-classifica%C3%A7%C3%A3o-17c4d4357ef9

tiendanube nuvemshop

OBRIGADO!