**Projeto**

**US Airline Flight Routes and Fares 1993 – 2024**

**Tópico 1: Previsão de Tarifas Aéreas** aborda bem todas essas instruções. Aqui está como ele atende a cada ponto:

* **Entendimento do problema**: O problema seria prever o valor da tarifa com base nas características da rota e da companhia aérea, o que está alinhado com a ideia de identificar fatores que influenciam o preço.
* **Compreensão do dataset**: O dataset oferece diversas variáveis relacionadas às rotas e às tarifas (distância, número de passageiros, participação de mercado das companhias, etc.), o que permite uma análise robusta.
* **Preparação dos dados**: O projeto exigiria um pré-processamento, como lidar com dados ausentes e ajustar tipos de variáveis, especialmente nas colunas de tarifas, distâncias e companhias.
* **Construção e avaliação de modelos**: Modelos de regressão (linear, árvore de decisão, etc.) seriam utilizados para prever as tarifas, com avaliações baseadas em métricas como erro quadrático médio (RMSE).
* **Resumo e conclusões**: Após a análise e o desenvolvimento do modelo, o projeto permitiria tirar conclusões sobre os fatores que mais influenciam os preços das passagens, oferecendo insights sobre o mercado de aviação.

**Previsão de Tarifas Aéreas**

Problema: Prever o valor da tarifa aérea (coluna fare) com base em diferentes características das rotas, como a distância entre as cidades, a companhia aérea e o número de passageiros. Isso ajudaria a entender os fatores que influenciam as variações de tarifas e poderia ser usado por companhias aéreas ou plataformas de viagens para ajustar preços.

Técnicas de Data Science:

* Regressão Linear: Usada para modelar a relação entre a tarifa (variável dependente) e outras variáveis independentes, como a distância da rota (nsmiles), o número de passageiros (passengers), e a companhia aérea.
* Árvores de Decisão ou Random Forest: Técnicas mais complexas que podem capturar interações não-lineares entre variáveis. Ideal se as relações entre os fatores e o preço forem mais complexas.
* Gradient Boosting (XGBoost, LightGBM): Para melhorar a precisão em relação à Random Forest, especialmente em datasets grandes.

Recursos disponíveis:

* Distância entre as cidades: A coluna nsmiles (número de milhas).
* Número de passageiros: Coluna passengers, que pode indicar a demanda e ajudar a prever a tarifa.
* Companhias aéreas: As colunas carrier\_lg (companhia aérea de maior market share) e carrier\_low (companhia aérea de menor market share) podem ser úteis para ver a influência das diferentes companhias nos preços.
* Participação de mercado: Colunas large\_ms e lf\_ms para medir o domínio de mercado das companhias em uma rota específica.

**Netflix**

**1. Predição de Gêneros Populares ao Longo do Tempo**

- Problema: Prever quais gêneros de filmes e séries serão mais populares em determinado ano com base em dados históricos.

- Técnica: Modelos de séries temporais, como ARIMA ou Prophet, ou classificação com algoritmos de aprendizado supervisionado, como Random Forest ou XGBoost.

- Recursos: Analisar a popularidade por gêneros ao longo dos anos e tentar prever o comportamento futuro.

**2. Prever a Classificação Indicativa (Rating)**

- Problema: Prever a classificação indicativa (TV-MA, PG-13, etc.) de novos títulos com base em atributos como país, duração, elenco, gênero e ano de lançamento.

- Técnica: Classificação supervisionada usando algoritmos como KNN, SVM ou redes neurais.

- Recursos: Utilizar variáveis como país, elenco, gênero e outros para prever a classificação.

**3. Recomendação de Títulos com Base no Gênero e Preferências**

- Problema: Criar um sistema de recomendação de filmes e séries com base no gênero e histórico de visualização (se houvesse dados de usuários).

- Técnica: Sistemas de recomendação, como Collaborative Filtering (filtragem colaborativa) ou Matrix Factorization.

- Recursos: Embora o dataset não inclua dados de usuários, você poderia simular dados ou usar métodos de filtragem de conteúdo com base nos gêneros e categorias.

**4. Previsão de Adições Mensais de Conteúdo**

- Problema: Prever quantos títulos serão adicionados à plataforma em um mês específico, com base em tendências passadas.

- Técnica: Modelos de séries temporais, como SARIMA ou Prophet.

- Recursos: A data de adição pode ser usada para criar uma série temporal e prever futuras adições de títulos.

**5. Análise de Sucesso de Títulos por País**

- Problema: Determinar quais países estão produzindo títulos que têm maior sucesso ou popularidade, e prever a probabilidade de sucesso com base no país de origem.

- Técnica: Modelos de classificação ou regressão com aprendizado supervisionado.

- Recursos: Utilizar o país, gênero, diretor e elenco como fatores para prever o sucesso de novos títulos.

**6. Predição de Duração Baseada em Gênero e Ano**

- Problema: Prever a duração média de um filme ou série com base no gênero e no ano de lançamento.

- Técnica: Regressão linear ou modelos mais avançados como Gradient Boosting.

- Recursos: Usar a duração como variável dependente e prever sua média com base em atributos como gênero e ano de lançamento.

**7. Classificação de Títulos Baseada em Descrições**

- Problema: Prever o gênero ou tipo de conteúdo (filme ou série) com base na descrição textual.

- Técnica: Processamento de linguagem natural (NLP) e classificação de textos com algoritmos como Naive Bayes, TF-IDF com Logistic Regression, ou redes neurais.

- Recursos: Utilizar a coluna de descrição para extrair palavras-chave e prever o gênero ou tipo de conteúdo.

**8. Clustering de Conteúdos**

- Problema: Agrupar títulos de filmes e séries em clusters com base em atributos como elenco, diretor, país, gênero, etc.

- Técnica: Algoritmos de clustering como K-Means ou DBSCAN.

- Recursos: Identificar padrões em grupos de títulos similares e categorizar o conteúdo de forma não supervisionada.

**9. Previsão de Lançamento de Novos Títulos por País**

- Problema: Prever a frequência de novos lançamentos por país com base em tendências anteriores.

- Técnica: Modelos de séries temporais, regressão ou machine learning.

- Recursos: Utilizar o país e o ano de lançamento para prever a produção de novos títulos.

**10. Análise Sentimental com Base nas Descrições**

- Problema: Realizar análise de sentimentos nas descrições dos títulos para classificar o conteúdo como positivo, neutro ou negativo, e prever o impacto sobre a popularidade.

- Técnica: Processamento de linguagem natural (NLP) com análise de sentimentos usando técnicas como VADER ou TextBlob.

- Recursos: Analisar a descrição textual para prever o sentimento geral do conteúdo e relacioná-lo ao sucesso de títulos.