Relatório Técnico – Desafio Titanic (Kaggle)

Autora: Cariny Saldanha

Duração estimada: 1–2 horas

1. Introdução

Este relatório conta como foi o processo de resolver o desafio "Titanic: Machine Learning from Disaster", que está disponível na plataforma Kaggle. A ideia principal é prever quem sobreviveu ao naufrágio do Titanic com base em informações como idade, sexo, classe da cabine e outros dados dos passageiros. Além disso, o projeto serve como uma forma de praticar e aplicar conceitos importantes de ciência de dados, como análise exploratória, tratamento de dados faltantes, criação de modelos e avaliação de desempenho.

2. Metodologia

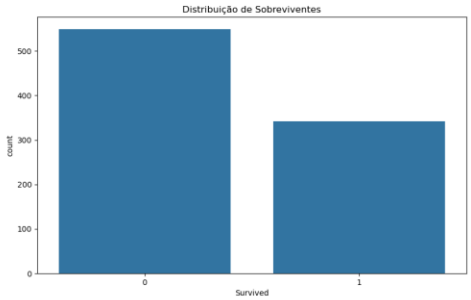
2.1 Análise Exploratória de Dados (EDA)

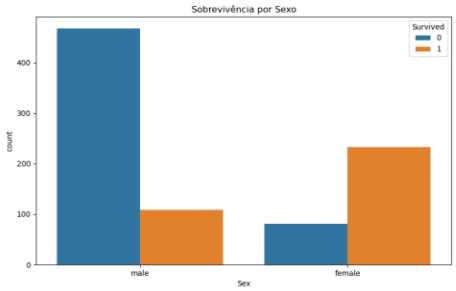
A análise exploratória foi feita com o objetivo de entender melhor quais fatores mais influenciaram na chance de sobrevivência dos passageiros. Alguns achados bem interessantes foram:

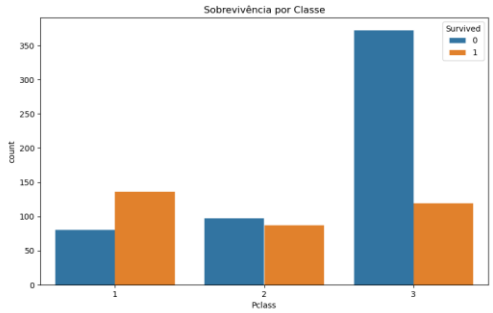
- As mulheres tiveram uma taxa de sobrevivência bem maior que os homens.
- Passageiros da 1ª classe também se saíram melhor, com uma proporção maior de sobreviventes.
- Crianças apresentaram uma taxa de sobrevivência mais alta em comparação com adultos.

2.2 Pré-processamento

Foram adotadas as seguintes etapas:
- Tratamento de valores ausentes: a idade foi imputada com a mediana, e a coluna 'Cabin', devido à alta quantidade de valores faltantes, foi descartada.
- Conversão de variáveis categóricas: 'Sex', 'Embarked' e 'Pclass' foram codificadas com LabelEncoder e get_dummies.
- Criação de novas features: foi criada a variável 'FamilySize' e ajustadas variáveis como 'Title', extraída do nome dos passageiros.







2.3 Modelo Escolhido

O modelo final utilizado foi o Random Forest Classifier, escolhido após testes com diferentes algoritmos (Logistic Regression, Decision Tree, SVM). O modelo foi ajustado com GridSearchCV para otimizar hiperparâmetros como:
- n_estimators: 100
- max_depth: 5
- min_samples_split: 4

2.4 Submissão no Kaggle

Após o treinamento, o modelo foi usado para fazer previsões no conjunto de teste fornecido pela competição e o arquivo CSV foi submetido na plataforma.

3. Resultados

3.1 Validação Cruzada

Foi utilizada validação cruzada com 5 folds.
Acurácia média: 0.82

3.2 Pontuação no Kaggle (Leaderboard Público)

Score: 0.7751

3.3 Insights dos Dados

Feature	Importância (Random Forest)
Sex	Alta
Pclass	Alta
Age	Média
Fare	Média
Embarked	Baixa

4. Discussão

4.1 Desafios Enfrentados

- Alto número de valores ausentes em variáveis como 'Cabin' e 'Age'.
- Desequilíbrio de classes em variáveis categóricas.
- Escolha dos hiperparâmetros ideais sem overfitting.

4.2 Limitações

- Algumas variáveis foram descartadas por dificuldade de imputação (ex.: 'Cabin').
- Títulos dos nomes foram considerados, mas poderiam ser refinados com agrupamentos semânticos melhores.
- Modelos mais complexos (como XGBoost) não foram utilizados por limitação de tempo.

4.3 Possíveis Melhorias

- Utilizar ensemble de modelos (stacking/blending).
- Aplicar feature engineering mais aprofundado (ex.: criar interações entre variáveis).
- Usar pipelines com Pipeline e ColumnTransformer para automação e reprodutibilidade.

5. Conclusão

Participar desse desafio foi uma ótima oportunidade para colocar em prática as etapas principais de um projeto , desde a análise exploratória, passando pelo pré-processamento e modelagem, até a avaliação e submissão do modelo. Além disso, ajudou a ganhar mais familiaridade com a plataforma Kaggle e com as métricas que ela usa para avaliar os resultados. O titanic criado teve um desempenho competitivo para uma primeira tentativa, e ainda tem bastante espaço para melhorias no futuro.