Relatório de Análise de Dados e Modelagem Preditiva: O Caso do Titanic

Edson Pimenta de Almeida

Junho de 2025

Sumário

1 Relatório do Processo de Análise

Este relatório descreve o fluxo de trabalho completo seguido para a análise do conjunto de dados do Titanic, desde a preparação inicial até a aplicação de múltiplos algoritmos de aprendizado de máquina. O objetivo foi extrair insights e construir modelos para que os fatores que influenciaram a sobrevivência dos passageiros fossem compreendidos.

1.1 Importação e Configuração do Ambiente

Para a realização deste estudo, o ambiente foi preparado com as bibliotecas essenciais do Python para ciência de dados:

- **Pandas & NumPy:** Utilizadas para manipulação e operações numéricas com os dados.
- **Matplotlib & Seaborn:** Empregadas para a criação de visualizações estáticas e informativas.
- **Missingno:** Usado para a visualização de padrões de dados ausentes.
- **Scikit-learn:** Aplicado para pré-processamento, modelagem e avaliação.

Configurações iniciais foram realizadas para melhorar a visualização dos gráficos e dataframes no ambiente do notebook.

1.2 Carregamento e Inspeção Inicial dos Dados

Os conjuntos de dados `train.csv` e `test.csv` foram carregados. Uma inspeção inicial revelou:

- **Dimensões:** O conjunto de treino possui 891 registros e 12 colunas; o de teste possui 418 registros e 11 colunas.
- **Tipos de Dados:** Uma mistura de tipos numéricos (inteiros e flutuantes) e categóricos (objetos/strings).
- **Valores Ausentes:** Valores nulos foram preliminarmente identificados nas colunas `Age`, `Cabin` e `Embarked`.

1.3 Análise Exploratória de Dados (EDA)

Uma Análise Exploratória de Dados (EDA) foi conduzida para que padrões e insights iniciais fossem descobertos.

- **Análise da Variável Alvo:** A análise da variável `Survived` mostrou um desbalanceamento, com aproximadamente 61.6% dos passageiros no conjunto de treino não sobrevivendo.
- **Análise Univariada:** Observou-se que a maioria dos passageiros viajava na 3ª classe, era do sexo masculino e embarcou em Southampton. A distribuição de idade concentrava-se entre 20 e 40 anos, enquanto a de tarifas era fortemente assimétrica à direita.

- **Análise Bivariada:** Correlações fortes foram identificadas entre a sobrevivência e outras variáveis:
  - **Sexo:** Mulheres tiveram uma taxa de sobrevivência drasticamente maior ($\approx$ 74%) que homens ($\approx$ 19%).
  - **Classe Social (Pclass):** A sobrevivência era diretamente proporcional à classe (1ª classe $\approx$ 63%, 2ª $\approx$ 47%, 3ª $\approx$ 24%).
  - **Idade e Tarifa:** Crianças pequenas e passageiros que pagaram tarifas mais altas tiveram maiores chances de sobreviver.
- **Dados Ausentes:** A análise aprofundada confirmou que `Cabin` tinha $\approx$ 77% de dados ausentes. `Age` tinha $\approx$ 20% de ausência, necessitando de uma estratégia de imputação robusta.

1.4 Pré-processamento e Engenharia de Atributos

Esta etapa foi crucial na preparação dos dados para os modelos.

- **Tratamento de Valores Ausentes:**
  - `Embarked` e `Fare` foram preenchidos com a moda e a mediana (agrupada por classe), respectivamente.
  - `Age` foi imputado de forma mais elaborada, utilizando-se a mediana da idade agrupada por `Title` e `Pclass`.
  - `Cabin` foi descartada devido à alta quantidade de dados nulos.
- **Engenharia de Atributos:**
  - **Title:** Títulos como "Mr.", "Miss.", etc., foram extraídos da coluna `Name`, agrupados e usados como um novo atributo.
  - **FamilySize e IsAlone:** As colunas `SibSp` e `Parch` foram combinadas para criar um atributo de tamanho da família, e uma flag binária foi criada para identificar passageiros que viajavam sozinhos.
- **Transformação de Variáveis:**
  - Variáveis categóricas como `Sex`, `Embarked` e `Title` foram convertidas para formato numérico usando mapeamento direto e One-Hot Encoding.
  - Atributos numéricos contínuos (`Age`, `Fare`, `FamilySize`) foram padronizados com `StandardScaler`.
- **Limpeza Final:** Colunas originais que se tornaram redundantes (`Name`, `Ticket`, `SibSp`, `Parch`) foram descartadas.

1.5 Modelagem e Avaliação

Quatro tipos de algoritmos foram aplicados para que diferentes perspectivas sobre os dados fossem obtidas. Com o aprendizado supervisionado para classificação, o objetivo foi treinar um modelo capaz de prever uma categoria ou classe discreta. No caso do Titanic, a classe a ser prevista era `Survived`. Com o aprendizado não supervisionado, que é usado quando não há rótulos pré-definidos, o objetivo foi explorar a estrutura intrínseca dos dados.

1.5.1 Modelos de Classificação (Aprendizado Supervisionado)

Random Forest (Floresta Aleatória)

Como Funciona: O Random Forest é um algoritmo de *ensemble learning* no qual múltiplas árvores de decisão são construídas durante o treinamento e suas previsões são combinadas.

1. **Bootstrap Aggregating (Bagging):** Diversas amostras de treinamento são criadas com reposição.
2. **Construção de Árvores Independentes:** Para cada amostra, uma árvore de decisão é construída. Em cada nó, apenas um subconjunto aleatório de atributos é considerado para a divisão.
3. **Agregação das Previsões:** A classe final prevista é aquela com mais "votos"entre todas as árvores.

Vantagens:

- Alta precisão e robustez a overfitting.
- Lida bem com dados faltantes e outliers (até certo ponto).
- Fornece medida da importância dos atributos.
- Versátil para classificação e regressão.

Desvantagens:

- Interpretabilidade reduzida comparada a uma única árvore.
- Pode ser computacionalmente mais custoso.
- Pode ser tendencioso com atributos categóricos com muitos níveis.

Na análise do Titanic: Uma acurácia de **82%** foi atingida pelo modelo Random Forest treinado no conjunto de validação. A análise da importância dos atributos revelou que `Sex`, `Fare`, `Age` e o título `Mr` foram os preditores mais influentes.

Rede Neural (Perceptron de Múltiplas Camadas - MLP)

Como Funciona: O MLP é uma rede neural artificial de alimentação direta com múltiplas camadas de neurônios.

1. **Camada de Entrada:** Os valores dos atributos são recebidos.
2. **Camadas Ocultas:** Uma ou mais camadas aplicam transformações não lineares (funções de ativação) aos sinais ponderados da camada anterior, permitindo o aprendizado de representações complexas.
3. **Camada de Saída:** A previsão final é produzida (ex: probabilidade de sobrevivência via função sigmoide).
4. **Treinamento com Backpropagation:** Os pesos das conexões são ajustados iterativamente para minimizar o erro entre a previsão e o valor real, usando-se a retropropagação do erro e um algoritmo de otimização.

Vantagens:

- Capacidade de modelar relações não lineares e interações complexas.
- Alta flexibilidade na arquitetura.
- Bom desempenho em grandes conjuntos de dados.

Desvantagens:

- Considerada uma "caixa-preta"em termos de interpretabilidade.
- Requer ajuste fino de hiperparâmetros e pode ser sensível à inicialização.
- Custo computacional elevado para treinamento.
- Propensa a overfitting sem regularização adequada.
- Sensível à escala dos atributos (requer padronização/normalização).

Na análise do Titanic: A MLP implementada atingiu uma acurácia de **83%** na validação. A análise de importância por permutação também destacou `Sex`, `Age` e `Fare` como os atributos mais decisivos.

1.5.2 Modelos de Descoberta de Padrões (Aprendizado Não Supervisionado)

K-Means (Agrupamento/Clustering)

Como Funciona: O K-Means particiona observações em `k` clusters, onde cada observação pertence ao cluster com a média (centroide) mais próxima.

1. **Inicialização:** `k` é escolhido e os centroides iniciais são definidos.
2. **Atribuição:** Cada ponto de dado é atribuído ao cluster do centroide mais próximo.
3. **Atualização dos Centroides:** Os centroides são recalculados como a média dos pontos em seu cluster.
4. **Iteração:** Os passos 2 e 3 são repetidos até a convergência.

Vantagens:

- Simples e eficiente para grandes conjuntos de dados.
- Relativamente fácil de interpretar os clusters formados.

Desvantagens:

- Necessidade de especificar `k` a priori.
- Sensível à inicialização dos centroides e a outliers.
- Assume clusters esféricos e de tamanho similar.
- Requer que os dados sejam escalonados.

Na análise do Titanic: Com `k=3`, o K-Means auxiliou na identificação de perfis distintos de passageiros. A análise da taxa de `Survived` (um rótulo externo usado *após* o agrupamento) dentro de cada cluster forneceu insights sobre como esses perfis se relacionavam com a sobrevivência.

Apriori (Mineração de Regras de Associação)

Como Funciona: O Apriori encontra itemsets frequentes em dados transacionais e gera regras de associação "se-então".

1. **Geração de Itemsets Frequentes:** Conjuntos de itens que aparecem juntos com frequência acima de um *suporte mínimo* são identificados, usando-se o Princípio Apriori para otimizar a busca.
2. **Geração de Regras de Associação:** A partir dos itemsets frequentes, são geradas regras que satisfazem uma *confiança mínima*. O *lift* mede o quão mais provável é o consequente ocorrer dado o antecedente.

Vantagens:

- Descoberta de relações e padrões entre itens.
- Regras geradas são geralmente interpretáveis.

Desvantagens:

- Custo computacional pode ser alto com muitos itens ou suporte baixo.
- Pode gerar um número excessivo de regras, necessitando filtragem.
- Requer dados em formato transacional (binarizado).

Na análise do Titanic: Associações conhecidas, como `{Sex_female, Pclass_1} -> {Survived_True}`, foram quantificadas, mostrando não apenas que a relação existe, mas quão forte ela é (via confiança e lift).

2 Conclusão Final e Comparação dos Modelos

A análise abrangente do conjunto de dados do Titanic, na qual uma combinação de técnicas supervisionadas e não supervisionadas foi utilizada, permitiu não apenas a construção de modelos preditivos precisos, mas também a extração de insights profundos e consistentes sobre os fatores que determinaram a sobrevivência.

2.1 Comparação das Abordagens

- **Random Forest vs. Rede Neural:** Os modelos de classificação treinados performaram de forma muito similar, com acurácias em torno de 82-83%. O Random Forest ofereceu maior interpretabilidade e robustez com menos necessidade de ajuste fino. A Rede Neural, embora ligeiramente superior em performance, confirmou que os mesmos atributos eram decisivos.

- **Classificação vs. Descoberta de Padrões:** A principal diferença reside no objetivo. Os modelos de classificação (RF e MLP) foram otimizados para **prever** a sobrevivência. Os modelos não supervisionados (K-Means e Apriori) foram usados para **entender** a estrutura dos dados.

  O K-Means aplicado agrupou os passageiros em "personas"distintas, e a análise da taxa de sobrevivência média desses grupos validou as hipóteses da EDA. O Apriori forneceu regras quantificáveis que verbalizam os padrões encontrados.

2.2 Síntese dos Insights Obtidos

Todos os modelos desenvolvidos, independentemente da abordagem, convergiram para a mesma conclusão central: **a sobrevivência no Titanic não foi um evento aleatório, mas sim fortemente estratificada por fatores socioeconômicos e de gênero.**

1. **O Privilégio Salvou Vidas:** Ser de uma classe social mais alta (`Pclass=1`) foi o indicador mais forte de sobrevivência depois do gênero, conforme evidenciado nos resultados.

2. **"Mulheres e Crianças Primeiro":** A política de evacuação é claramente visível nos dados analisados. O atributo `Sex` foi o preditor mais importante nos modelos de classificação treinados.

3. **Viajar Sozinho Era Perigoso:** A engenharia de atributos realizada, criando `FamilySize` e `IsAlone`, revelou que passageiros completamente sozinhos tiveram uma taxa de sobrevivência menor.

Em suma, este projeto demonstra o poder de um fluxo de trabalho de ciência de dados completo. A análise exploratória conduzida levantou as hipóteses, o pré-processamento efetuado preparou o terreno, e a aplicação sinérgica de modelos supervisionados e não supervisionados permitiu não só a validação dessas hipóteses com rigor estatístico, mas também a construção de um entendimento robusto e multifacetado do trágico evento.