12º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2021

ANÁLISE DE DADOS USANDO A COMPETIÇÃO TITANIC DO KAGGLE

BÁRBARA A. N. MARQUES[1], MARCOS WILLIAN DA S. OLIVEIRA[2], ROSEMEIRE A.R. OLIVEIRA[2]

1 Graduando em Engenharia Mecânica, Bolsista PIBIFSP, IFSP, Câmpus São José dos Campos, barbara.marques@aluno.ifsp.edu.br.

2 Docente IFSP, Câmpus São josé dos Campos

Área de conhecimento (Tabela CNPq): 1.02.02.08-0 - Análise de Dados

RESUMO: Este estudo do conjunto de dados da competição "Titanic - Machine Learning from Disaster", presente na plataforma Kaggle, baseia-se na análise de gráficos, que demostram relações entre as variáveis e a sobrevivência dos passageiros, e de códigos de previsão de dados utilizando os modelos de classificação *Random Forest* e *Decision Tree*. Tal estudo mostrou as tendências para que um passageiro pudesse ter mais chances de ser salvo durante o naufrágio, mas também evidenciou o que pode ser interpretado como o reflexo da sociedade da época. Essas conclusões levaram ao questionamento do quanto os modelos de previsão podem ser replicáveis e imparciais considerando que são construídos baseados em ações humanas.

PALAVRAS-CHAVE: modelos de previsão; previsões tendenciosas.

GRAPHIC DATA ANALYSIS USING KAGGLE'S TITANIC COMPETITION

ABSTRACT: This study of the dataset in competition "Titanic - Machine Learning from Disaster", presented on the Kaggle platform, is based on graphics analysis, that demonstrate relationships between the variables and the survival of passengers, and of data prediction codes using Random Forest and Decision Tree classification models. This study showed the trends for a passenger to be saved during the wreck, but it also showed what can be interpreted as a reflection of society at the time. These conclusions led to the questioning of how much predictive models can be replicable and impartial considering that they are built based on human actions.

KEYWORDS: predict models; biased forecasts.

INTRODUÇÃO

Na competição da plataforma Kaggle[1], a história do naufrágio do navio Titanic ganha forma em um conjunto de dados de informações sobre os passageiros a bordo. A competição tem como objetivo construir um modelo de previsão que consiga identificar quais pessoas sobreviveriam ao naufrágio a partir apenas de suas características registradas na base de dados. Não há evidências de que os dados disponibilizados na plataforma sejam reais e nem de como teriam sido construídos.

Através de Machine Learning (ML), faz-se a previsão de quais passageiros sobreviveriam ao naufrágio. Isso se dá pela construção de modelos [2,3], que são os códigos que analisam o conjunto de dados. Esses modelos fazem comparações entre as informações da base de dados e determinam quais características mais se repetem ou se aproximam, considerando-as determinantes para o resultado final. Neste caso, a sobrevivência dos passageiros. Essas características são as relações que foram estudadas de forma explícita através dos gráficos gerados com auxílio da biblioteca Seaborn [6], buscando entender mais profundamente algumas das possíveis interações que os modelos de ML poderiam encontrar.

[1] *Competição "Titanic - Machine Learning from Disaster":* https://www.kaggle.com/c/titanic

Sendo assim, foram construídos gráficos para explorar melhor as relações entre as variáveis e a sobrevivência das pessoas no acidente. Na sequência, construiu-se modelos de previsão diferentes para especular qual teria um desempenho melhor ao trabalhar com os dados, verificando sua precisão ao realizar as previsões. Obteve-se que as características como Name (nome), PassangerID, Cabin (cabine) e Ticket não apresentaram relações com a sobrevivência, mas as demais informações exibiram essa relação através dos gráficos. Além disso, verificou-se que ambos os modelos estudados (sendo eles o *Random Forest* e *Decisio Tree*) possuem um bom desempenho ao serem utilizados para análise dessa base de dados.

MATERIAL E MÉTODOS

Para análise do conjunto de dados "Titanic - Machine Learning from Disaster", foram construídos códigos em linguagem Python, com o uso de modelos de Árvore de Decisão e Random Forest [4], bem como foram testados modelos com a aplicação de Cross Validation [3]. Também foram usadas a biblioteca Pandas [1] para manipulação dos dados, a biblioteca Seaborn [6] para explicitar as possíveis relações que os modelos estariam encontrando ao trabalhar com a base de dados através da plotagem de gráficos, e a biblioteca Scikit-Learn [5] para determinar o erro médio quadrado [5] de cada modelo com relação as suas previsões *(mean absolute error)*.

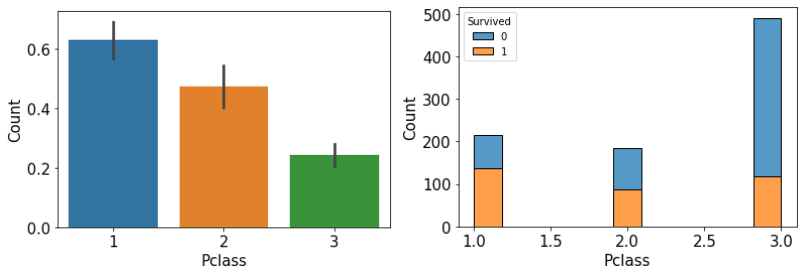
O conjunto de dados é composto pelas características: Name (nome), PassangerID (número do passageiro na lista), Cabin (cabine), Fare (valor pago pela passagem), Ticket (código da passagem), Age (idade), Sex (sexo), Embarked (local de embarque), Pclass (classe), SibSp (número de irmaos e cônjuges), Parch (número de pais e crianças) e Survived (confirmação ou não de sobrevivência). Portanto, o conjunto de dados consiste de uma matriz com 12 colunas e 891 linhas ou instâncias.

As relações entre as séries de dados foram exploradas através da biblioteca Seaborn, realizando-se a construção de gráficos que pudessem evidenciar tais relações com a sobrevivência ou não dos passageiros.

GRÁFICOS E DISCUSSÕES

Com todas as colunas da matriz foram feitas essas relações em busca de evidências que indicassem a interação entre características dos passageiros, determinando se foram salvos ou não. Entre as colunas Pclass, Sex, Age, SibSp, Parch, Fare e Embarked foram encontradas relações com a sobrevivência dos passageiros, evidenciadas através dos gráficos estudados. Abaixo é possível encontrar os principais gráficos usados para plotar tais informações, sendo eles: gráfico de barras, de dispersão categórico e de barras empilhadas.

A característica Pclass se trata da classe em que o passageiro se encontrava na embarcação, sendo elas 1ª, 2ª e 3ª classe. Não se encontrou um mapa do navio, não sendo possível verificar a probabilidade de sobrevivência a partir da localização dos quartos. No gráfico de barras na Figura 1(a), é possível perceber que a porcentagem de pessoas da 1ª classe que sobreviveram ao naufrágio é maior do que nas outras classes. Já no gráfico de barras empilhadas na Figura 1(b), observa-se que a maior parte das pessoas estavam na 3ª classe, e complementa a informação do gráfico na Figura 1(a). Pode-se inferir que há desigualdade entre as classes, uma vez que mesmo com um número menor de pessoas na 1ª classe, ela foi a que teve um percentual maior de sobrevivência, e a 3ª classe teve um percentual de pessoas mais baixo mesmo tendo um maior número de integrantes. Da relação entre Fare e Survived pode-se inferir que quanto mais cara a passagem que uma pessoa pagou para a viagem, maiores as chances dela sobreviver, uma vez que os maiores números de não sobreviventes aparecem na região de preços entre 0 e 100.

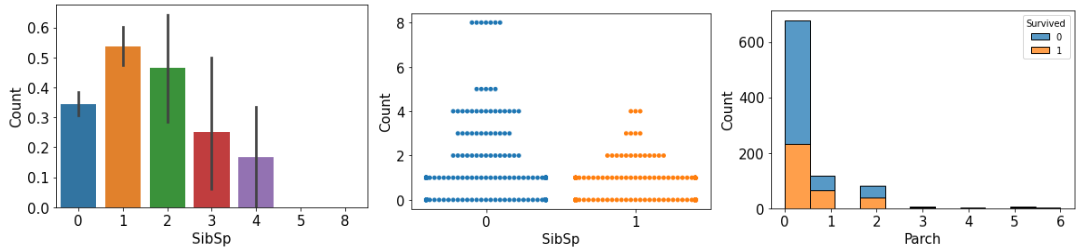


(a) (b)

FIGURA 1. Gráficos (a) e (b) referentes a relação entre Pclass e Survived

A série SibSp contém o número de irmãos e/ou cônjuges que o passageiro possuía na embarcação. O gráfico de barras na Figura 2(a), mostra que os sobreviventes possuíam menos irmãos/cônjuges. No gráfico de dispersão categórico na Figura 2(b), demonstra que a maior parte das pessoas que possuíam muitas relações desse tipo não sobreviveram ao naufrágio. Além disso, esse gráfico mostra também que pessoas com até duas relações do tipo irmãos/cônjuges não há índices muitos diferentes entre os sobreviventes ou não sobreviventes. Porém a partir de mais de três relações desse tipo há diferença nos dados, indicando que quanto mais relações, maiores as chances de não sobrevivência.

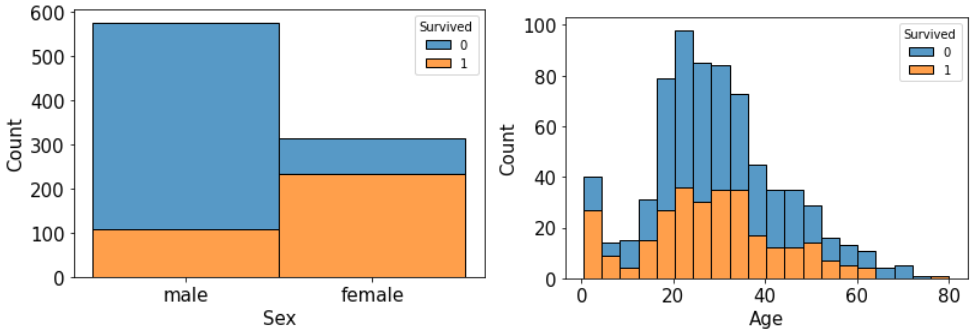
A série Parch apresenta que pessoas com vínculos a crianças ou pais (mais velhas que o passageiro) também possuíam privilégio no salvamento. Por conta provavelmente da prioridade que essas pessoas as quais estava vinculado possuíam tal privilégio - uma criança ou pessoa idosa não poderia ser embarcada em um bote sem alguém responsável por ela. O gráfico da Figura 2(c) mostra que o número de pessoas sem vínculo do tipo Parch que não sobreviveu é muito maior do que a porcentagem de pessoas que possuíam ao menos um vínculo.



(a) (b) (c)

FIGURA 2. Gráficos (a) e (b) referentes a relação SibSp com Survived, e gráfico (c) referente a relação Parch e Survived

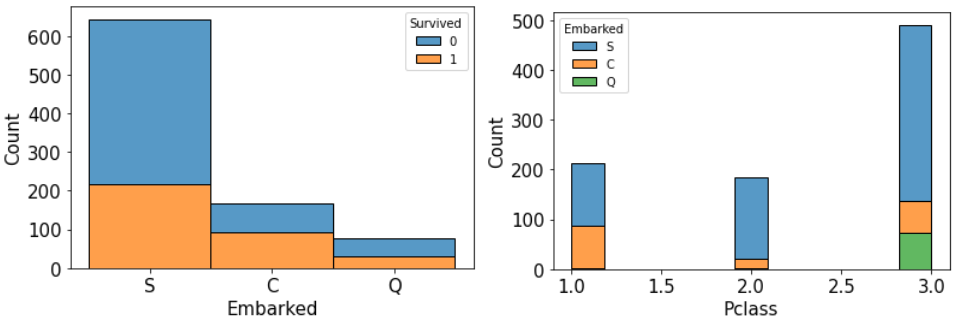
A partir das característica Sex e Age, apresentadas na Figura 3(a) e 3(b) respectivamente, percebe-se uma ideia típica das ações no momento do salvamento. Infere-se que foram privilegiadas mulheres, crianças e idosos, sendo essas as categorias que tiveram maior índice de sobrevivência apresentado nos gráficos, em comparação com a faixa etária dos 20 aos 40 anos.



(a) (b)

FIGURA 3. Gráfico (a) referente a relação Sex e Survived, e gráfico (b) referente a relação Age e Survived

Por último, na Figura 4, Embarked apresenta relação com Survived na Figura 4(a), indicando que a maior parte dos sobreviventes havia embarcado em Southampton e também era de lá que grande parte das pessoas da 1ª classe vieram, como mostrado no gráfico da Figura 4(b), da relação entre Embarked e Pclass. Ao mesmo tempo, percebe-se que as pessoas que não sobreviveram haviam embarcado em Queenstown, e não pagaram caro por suas passagens, como mostrado na relação de preços em Fare e do local onde embarcaram.



(a) (b)

FIGURA 4. Gráfico referente a relação entre Embarked e Survived (a) e Pclass (b)

EXPERIMENTOS DE PREVISÃO

Os modelos usados para construção dos códigos de previsão foram Árvore de Decisão e *Random Forest*, também estudando como a aplicação de *Cross Validation*, utilizando a configuração 5-fold, valida os modelos através de testes o desempenho dos modelos. Dessa forma, calculou-se os tempos que cada modelo demorou para rodar e qual a pontuação que cada um deles alcançou (sendo calculada através da função *mean absolute error* [5] e *neg mean absolute error* - usado no Cross Validation).

No tratamento de dados faltantes, foi utilizada imputação simples, na qual adicionou-se valores às colunas que possuíam elementos numéricos de dados contínuos do tipo "int64" ou "float64" com células vazias ("NaN"), no caso apenas a variável Age foi submetida ao processo de imputação. Já as colunas com strings e/ou com elementos categóricos foram submetidas ao processo de categorização dos elementos através do One Hot Encolder (OHE) [2]. Eram elas as variáveis Sex, Embarked, Pclass, SibSp e Parch. Essas formas de ajuste da base de dados foram escolhidas após estudo com outras formas de ajuste, como a retirada de variáveis com células vazias ou strings, pois foram as formas que obtiveram melhor resultado quando aplicadas aos modelos de previsão.

O Cross validation utilizado para validação dos modelos foi do tipo 5-fold, com separação da base de dados em cinco grupos, utilizando uma parte para teste e as outras quatros partes restantes para treinamento do modelo, a cada repetição da validação cruzada. Na sequência, baseada em tempos para 15 repetições de cada código, tem-se a Tabela 1 apresentando os desvios padrões e as médias dos tempos, seguida da Tabela 2 com os valores das pontuações alcançadas por cada modelo.

TABELA 1. Tempos dos códigos dos modelos de previsão

Tempos	Decision Tree	Random Forest
Média	0,1019803286	0,9539995193
Desvio Padrão	0,008691274735	0,01466644501

TABELA 2. Mean Absolute Error para os modelos

Decision Tree			Random Forest		
Apenas Imputação	Imputação + OHE	Cross validation	Apenas Imputação	Imputação + OHE	Cross validation
0.4578572822198606	0.3347760146000535	0.31062265851324355	0.4647071148385778	0.3346562765083999	0.31165351787865025

A Tabela 1 mostra a média dos tempos em que cada modelo foi executado, e observasse que o modelo *Random Forest* demorou alguns milésimos de segundos a mais para rodar do que a árvore de decisão. Dessa forma é possível inferir que tal diferença de tempo é ínfima para determinar com certeza qual modelo é o melhor para ser aplicado.

Analisando a Tabela 2, onde obtêm-se os valores de erro médio absoluto em cada etapa de ajuste de dados (Imputação e OHE) e também o erro médio da aplicação do modelo em *Cross validation*, pode-se dizer que ambos os modelos são muito próximos em termo de análise e previsão dos dados, podendo ser utilizado qualquer um dos dois modelos apresentados para predição de acidentes como o Titanic.

CONCLUSÕES

A partir da base de dados da competição TITANIC da plataforma Kaggle, realizou-se um estudo de como essas séries de dados poderiam ser analisados por modelos de ML, sendo eles *Random Forest* e *Decision Tree*, não apenas construindo os códigos de programação, mas também buscando através da análise gráfica uma forma de expor quais séries possuem maior relevância para os modelos. Esses estudos mostraram que as características Age, Embarked, Fare, Parch, Pclass, Sex e SibSp se destacam por possuirem dados relevantes que indicam, além do impacto sobre os modelos de previsão, uma sociedade que privilegia determinadas classes sociais.

Com as análises dos gráficos, indagou-se sobre a precisão real dos modelos. Mesmo estes que possuem um valor baixo para *mean absolute error*, não se trata da precisão com base nos dados inicialmente considerados, mas o quão justo esses modelos conseguiriam ser quando fossem aplicados a novos dados, uma vez que são modelos construídos a partir de uma base que não é imparcial.

Após a análise dos gráficos, deixa-se a reflexão sobre o impacto do uso de modelos que são baseados em ações humanas injustas, reproduzindo-as em previsões futuras. Diante dos fatos, pode-se admitir que o uso de dados tendenciosos para a construção de modelos de previsão é ruim para a justiça e precisão para com os novos dados, provenientes de novas situações em uma sociedade em constante mudança e crescimento.

REFERÊNCIAS

[1] BILOGUR, Aleksey. Pandas. Kaggle. Disponível em:<https://www.kaggle.com/learn/pandas>. Acesso em: 08 de março de 2021.

[2] COOK, Alexis. Intermediate Machine Learning. Kaggle. Disponível em:<https://www.kaggle.com/learn/intermediate-machine-learning>. Acesso em: 29 de março de 2021.

[3] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. The Elements of Statistical Learning. New York, NY, USA, Springer New York, 2001.

[4] PINTO, PEDRO C. P. Missing Data Analysis in Classification and Regression Problems. 2016. 93 f. Dissertação - UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Rio de Janeiro, 2016.

[5] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research Vol 2., p. 2825-2830, 2011. Disponível :<https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error>. Acesso em: 22 de março de 2021.

[6] WASKOM, Michael L. Seaborn: Statistical Data Visualization. Journal of Open Source Software, Vol. 6, nº 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>. Acesso em: 10 de Junho de 2021.