Titanic – Machine Learning from Disaster

Alexandre Araújo Pires Mourão
up201905967

Ruben Miguel Gomes Santos
up201905773

Abstract

A análise do naufrágio do Titanic é essencial para compreender os dados históricos. a relação entre as diferentes *features* dependentes e independentes foram observadas para determinar se estas tiveram ou não impacto na sobrevivência dos passageiros. Neste relatório exploramos 4 modelos incluindo *Random Forest*, *Logistic Regression*... que foram implementados para prever a taxa de sobrevivência. Foram considerados fatores como "*Age*", "*Gender*", "*Children*".

1 Introdução

O projeto foi realizado no âmbito da unidade curricular de "Aprendizagem Computacional". Algoritmos de aprendizagem computacional ou *Machine Learning* têm como objetivo analisados dados históricos.

Um dos desastres marítimos mais conhecidos na história foi o Titanic, que aconteceu a 15 de Abril, 1912. Este deveu-se à colisão com um *iceberg* que "rasgou" o casco do navio fazendo com que este afunda-se. Muitos fizeram várias especulações ao que levou à taxa de sobrevivência dos passageiros é exatamente isso que vamos abordar neste projeto.

2 Intrepertação dos dados

2.1 Ficheiros

São fornecidos três ficheiros:

- train.csv - Ficheiro com dados de treino para treino dos modelos. Estão aqui incluidas todas as *features*.
- test.csv - Ficheiro com dados de treino para o teste dos modelos. Estão aqui incluídas todas as *features*, excetuando a *feature survived*.
- gender_submission.csv - Ficheiro exemplo para demonstrar a estrutura do ficheiro de submissão.

2.2 Features

Foram analisadas 12 features diferentes:

- *PassengerId* - Identificação única de cada passageiro 1 - 891 no ficheiro train.csv e 892 - 1309 no ficheiro test.csv.
- *Survived* - 0 se o passageiro não sobreviveu, 1 se sobreviveu.
- *Pclass* - Pode ter o valor "1", "2" ou "3", representa a classe do bilhete comprado, sendo 1 a classe mais alta e 3 a mais baixa.
- *Name* - Nome de cada passageiro.
- *Sex* - Sexo de cada passageiro.
- *Age* - Idade de cada passageiro.
- *SibSp* - Número de irmãos e cônjuges a bordo do navio.
- *Parch* - Número de pais/filhos presentes no navio.
- *Ticket* - Número do bilhete de cada passageiro.
- *Fare* - Tarífa do passageiro.
- *Cabin* - Número da cabine.
- *Embarked* - Porto de embarque pode tomar os valores de "S" Southampton, "Q" Queenstown ou "C" Cherbourg.

3 Análise dos dados

3.1 *Cleaning Data*

Nesta parte do trabalho utilizou-se tanto o ficheiro *train.csv* como o ficheiro *test.csv* e após uma breve análise

reparou-se que as *features* e *Name*, *PassagerId* e *Ticket* não vão ser relevantes para a taxa de sobrevivência dos passageiros, por isso foi decidido não utilizar essas *features* retirando as respetivas colunas. Também foi retirada a *feature Cabin*. Após uma análise mais aprofundada desta variável, reparou-se que a maior parte dos valores recebidos são nulos e por isso decidiu-se não usar para a previsão final.

O código utilizado para tal foi o seguinte:

```
#tirar o nome, cabina e o ticket e o Id
data.drop({'Name', 'Cabin', 'Ticket',
'PassengerId'}, axis=1, inplace=True)
```

Para a submição no kaggle foi necessário guardar o *PassengerId*, do ficheiro *test.csv*, por isso utilizou-se o seguinte código:

```
##tirar o nome, cabina, ticket e o Id
IdTest=teste['PassengerId']#-->guadar o
id para submissão no kaggle
teste.drop({'Name', 'Cabin', 'Ticket', 'PassengerId'},
axis=1, inplace=True)
```

De seguida teve de se analisar os valores nulos (NaN) de cada *feature* tanto nos dados de teste como nos de treino, e eliminá-los ou substituí-los, neste caso, pela moda. Foram identificados dados nulos nas colunas da *Age*, *Fare* e *Embarked*.

No caso da coluna *Age*, os valores nulos foram substituídos pela moda tanto nos dados de treino como nos dados de teste:

```
#resolver os NaN para a idade de treino
mode=data['Age'].mode()[0]#usar a moda da idade para
preencher os restantes
data.fillna({'Age': mode}, inplace=True)

#resolver os NaN para a idade de teste
mode=teste['Age'].mode()[0]#usar a moda da idade para
preencher os restantes
teste.fillna({'Age': mode}, inplace=True)
```

No caso do *Embarked* nos dados de treino, como são apenas duas linhas nulas, dissidiu-se eliminá-las:

```
#resolver o embarked treino
#apagamos duas linhas do NaN
data.dropna(inplace=True)
```

No caso da *Fare* nos dados de teste, devido às restrições de submição do kaggle é necessário manter as linhas, e por isso substituiu-se as linhas nulas pela moda:

```
#resolver o fare teste
#temos de manter as linhas do NaN para entrega no kaggle
mode=data['Fare'].mode()[0]
teste.fillna({'Fare': mode}, inplace=True)
```

Foram também organizadas as *Features* de *Age* e *Fare* em vários *bins*:

Age: *Infant*: 0-5; *Kid*: 6-17; *Young-Adult*: 18-25; *Adult*: 26-50; *Old*: 51-80.

Fare: *Cheap*: [0;20]; *Medium*:]20;40]; *Medium-high*:]40;60]; *High*:]60;100]; *Expensive*:]100;512].

Age:

```
#organizar os valores de idade

#idade treino
intervalos=[0, 5, 17, 25, 50, data['Age'].max()]
tipo=['Infant', 'Kid', 'Young-Adult', 'Adult', 'Old']
data['Age'] = pd.cut(data['Age'], bins=intervalos,
labels=tipo)

#idade teste
intervalos=[0, 5, 17, 25, 50, teste['Age'].max()]
teste['Age'] = pd.cut(teste['Age'], bins=intervalos,
labels=tipo)
```

Fare:

```
#organizar fare

#fare treino
intervalos=[0, 20, 40, 60, 100, data['Fare'].max()]
tipo=['Cheap', 'Medium', 'Medium-High', 'High',
'Expensive']
data['Fare'] = pd.cut(data['Fare'], bins=intervalos,
labels=tipo)

#fare teste
intervalos=[0, 20, 30, 60, 100, teste['Fare'].max()]
teste['Fare'] = pd.cut(teste['Fare'], bins=intervalos,
labels=tipo)
```

3.2 Análise de dados de treino

A seguinte imagem mostra o número total de sobreviventes e mortos de todos as amostras presentes nos dados de treino. Daqui pode-se verificar que houve maior número de mortos do que de sobreviventes.

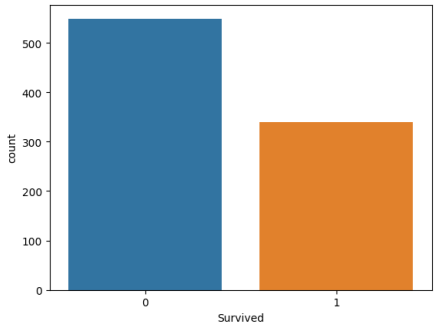

Figure 1: Número de sobreviventes (1) e de mortos (0).

Na seguinte imagem pode-se ver o número total de passageiros e as suas respetivas tarifas divididas entre *Barato*, *Médio*, *Médio-Alto*, *Alto* e *Caro*. Como é possível observar a tarifa mais regular é a *Barata*:

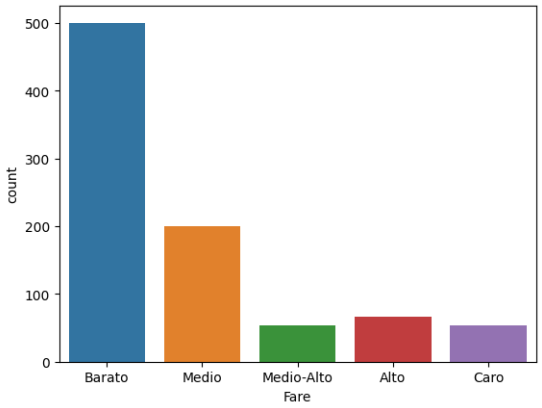

Figure 2: Tarífa de todos os passageiros dividido por bins

Como se pode observar na Figura 3, apesar de a tarifa mais recorrente ser a mais barata [0;20], a taxa de sobrevivência desta é a mais baixa, sendo que a mais cara,]100; 512], tem a taxa de sobrevivência mais alta das cinco.

Pode-se assim concluir que quanta maior a tarifa maior a taxa de sobrevivência.

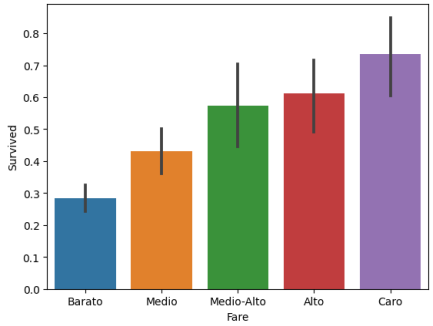

Figure 3: Taxa de sobrevivência por cada bin de tarífa.

Tal também se pode observar na classe dos bilhetes dos passageiros.

A figura 4 demonstra que a maior parte dos passageiros estava a navegar com bilhete de classe 3 (mais barato):

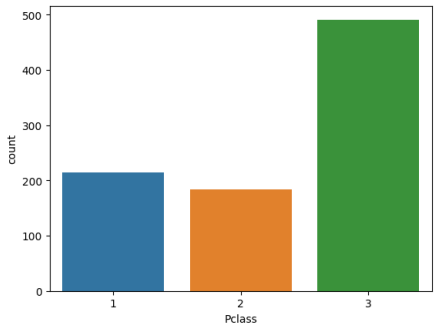

Figure 4: Bilhetes totais divididos por classes.

Porém a taxa de sobrevivência mais baixa é a da classe 3 (mais barata) e a mais alta é a da classe 1 (mais cara) como se pode observar na figura 5:

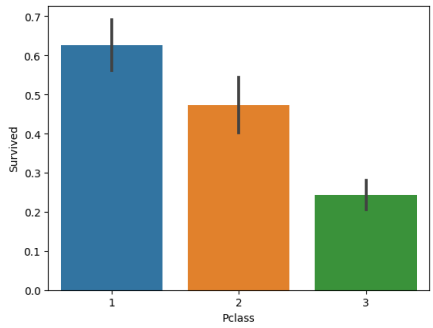

Figure 5: Taxa de sobrevivência por cada tarífa.

Na figura 6 verifica-se que a maior parte dos passageiros não tinha qualquer irmão/cônjuge a bordo:

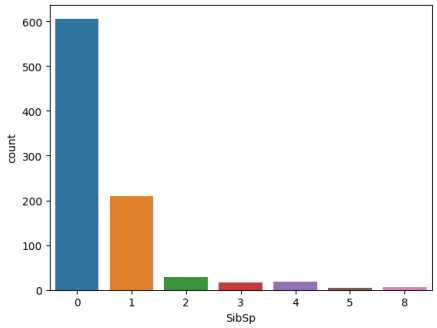

Figure 6: Número de irmãos/cônjuges dos passageiros.

No entanto pode-se verificar na figura 7 que a taxa de sobrevivência relaciona-se com esta *feature*, sendo que a maior taxa de sobrevivência acontece quando o número de irmãos/cônjuges é 1:

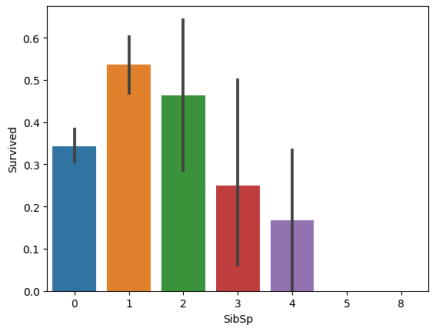

Figure 7: Taxa de sobrevivência por cada tarifa.

Na seguinte figura 8 pode-se observar o numero de pessoas que embarcaram em cada porto ("S" Southampton, "Q" Queenstown ou "C" Cherbourg.):

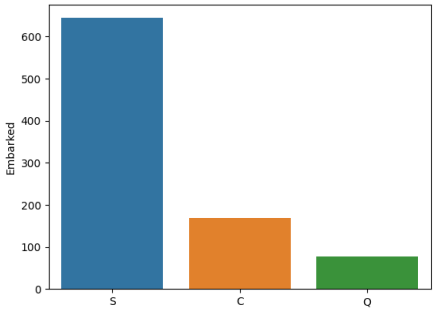

Figure 8: Total de passageiros que embarcaram em cada porto.

Na figura 9 verifica-se que as pessoas que embarcaram no porto de Cherbourg têm uma maior taxa de sobrevivência e as que embarcaram em Southampton tem uma menor hipótese de sobreviver.

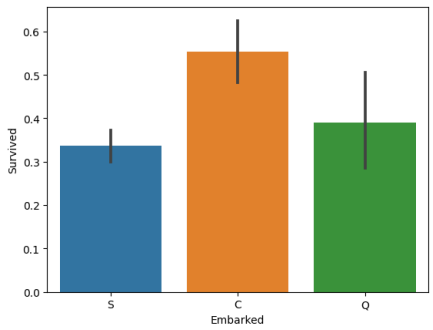

Figure 9: Taxa de sobrevivência de cada porto de embarque.

Quanto à *feature Age*, podemos ver que a idade mais recorrente está entre os 18 e os 25 anos.

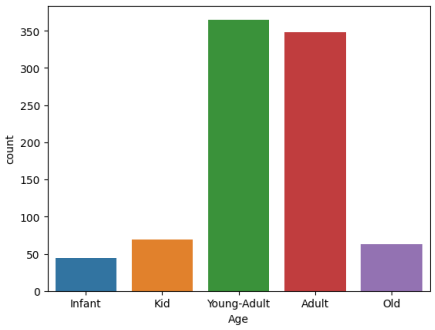

Figure 10: Taxa de sobrevivência de cada porto de embarque.

Pode-se observar na figura 11 que a maior taxa de sobrevivência é a dos mais novos (0-5 anos), como era de esperar.

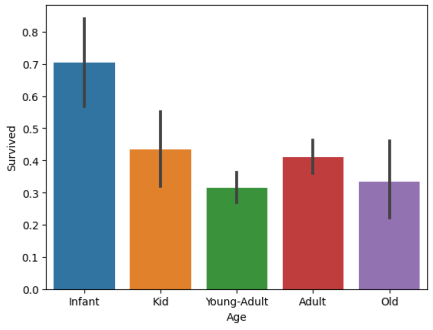

Figure 11: Taxa de sobrevivência de cada porto de embarque.

Na presente imagem 12 é possível observar que existem mais homens do que mulheres abordo do navio.

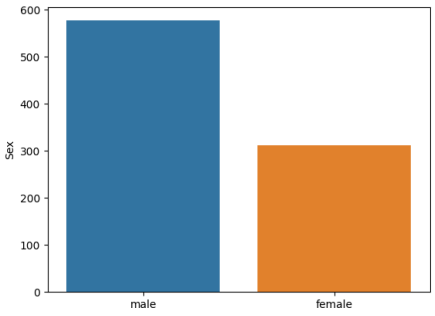

Figure 12: Número de homens e mulheres abordo.

Como era de esperar a taxa de sobrevivência é muito maior no caso das mulheres em relação à dos homens:

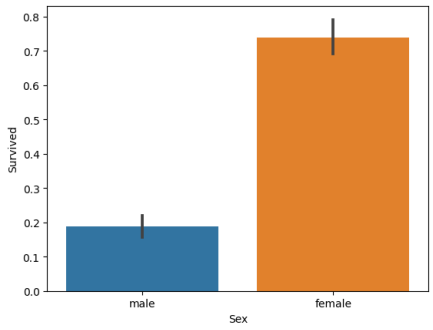

Figure 13: Taxa de sobrevivência entre gêneros.

3.3 Modelos de classificação

Antes da realização dos modelos dividiu-se os dados de treino em dtrain e dtest, de forma a poder-se treinar o modelo e testá-lo para se verificar o seu *accuracy score*.

Implementaram-se os modelos *Logistic Regression, Random Forest, Naibe Bayes, Gradient Descent e k-Nearest Nighbors* e estes foram os resultados:

Modelos de classificação	Precisão
Logistic Regression	80,2%
Random Forest	81,9%
Naibe Bayes	80,2%
Gradient Descent	81,3%
k.Nearest Neighbors	82,6%

Figure 14: Tabela com os valores de cada modelo.

Através da análise dos modelos foi possível verificar que o modelo com a melhor resolução do problema é o *k-Nearest Nighbors*. Porém usamos o modelo de *Logistic Regression* para submissão no kaggle.

3.4 Previsão de sobrevência

Uma vez que este projeto foi realizado através da plataforma Kaggle procedeu-se à submissão da previsão obtendo-se uma precisão de 77,03%

4 Conclusão

Neste projeto foram testados vários modelos de previsão abordados nas aulas de Aprendizagem Computacional. Foram também analisadas várias *features* que influenciaram a taxa de sobrevivência num dos maiores acidentes da história.

A realização deste trabalho permitiu ter uma visão sobre a importância desta cadeira de *machine learning*.

5 Referências

https://www.kaggle.com/competitions/titanic/overview