**Projeto PBD**



**Trabalho realizado por:**

**Docentes:**

Adriano Martins Lopes

João Pedro Afonso Oliveira da Silva

Bernardo Arjones, Nº104787

Enderson Santos, Nº103954

**Índice**

[1. Introdução 3](#_Toc131944947)

[2. Business Understanding 4](#_Toc131944948)

[3. Data Understanding 4](#_Toc131944949)

[3.1. Data Explaining 4](#_Toc131944950)

[3.2. Exploratory Data Analysis 6](#_Toc131944951)

[4. Data Preparation 7](#_Toc131944952)

[4.1. Data Cleaning 7](#_Toc131944953)

[4.2. Data Preprocessing 8](#_Toc131944954)

[4.3. Feature Engineering 8](#_Toc131944955)

[5. Modeling and Evaluation 8](#_Toc131944956)

[6. Deployment 10](#_Toc131944957)

[7. Conclusão 10](#_Toc131944958)

# Introdução

Este trabalho foi realizado no âmbito da unidade curricular de PBD (Processamento de *Big Data*), cujo objetivo era a resolução de um problema, no âmbito da ciência dos dados, recorrendo às técnicas e metodologia de *machine learning*.

A metodologia utilizada neste projeto é conhecida como CRISP-DM, a qual consiste em:

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modeling*
5. *Evaluation*
6. *Deployment*

Para este projeto optámos pelo *dataset* relacionado com o Expedia *Hotels*, dos diversos disponibilizados pelos docentes da respetiva Unidade Curricular (UC). O principal objetivo do nosso projeto é criar um modelo de recomendação, assente nos atributos/características das pesquisas efetuadas por cada cliente (*user*).

No projeto tivemos oportunidade de colocar em prática todas as matérias lecionadas durante as aulas, tendo sido possível aplicar o conhecimento retido nas seguintes plataformas: Jupyter-Lab e AWS Academy. Adicionalmente, permitiu promover o desenvolvimento de *soft skills,* assim como *teamwork*, *communication*, *time management,* entre outras.

Os dados utilizados poderão ser localizados no seguinte *link*:

<https://www.kaggle.com/competitions/expedia-hotel-recommendations/data>.

A licença é concedida pela própria empresa, *Expedia Hotels*, para a utilização dos dados.

# Business Understanding

A Expedia *Hotels* disponibiliza registos do comportamento dos seus clientes, abrangendo os critérios de pesquisa, o modo como eles interagiram com os resultados obtidos (clique/reserva), e se o resultado da pesquisa era ou não um *pack* de viagem. Os dados resultam de uma seleção aleatória da empresa, não sendo representativos das estatísticas gerais.

A Expedia está interessada em prever qual grupo de hotéis que um cliente (*user*) irá reservar e possui algoritmos internos para formar *clusters* de hotéis, através do quais, hotéis semelhantes em determinada pesquisa são agrupados (com base no preço histórico, classificação por estrelas do cliente, localizações geográficas relativas ao centro da cidade, etc.). Esses *clusters* de hotéis servem como bons identificadores para a tipologia de hotéis que os potenciais clientes poderão vir a reservar, evitando valores discrepantes, como novos hotéis que não possuam dados históricos.

O objetivo principal é prever o resultado da reserva (grupo de hotéis) de um potencial cliente (*user*), com base na sua pesquisa e atributos associados. Os dados da pesquisa do utilizador serão então tratados, processados e posteriormente fornecidos ao nosso modelo que irá listar as unidades hoteleiras mais prováveis para esse utilizador. Essa lista será disponibilizada na página da web, por ordem de probabilidade.

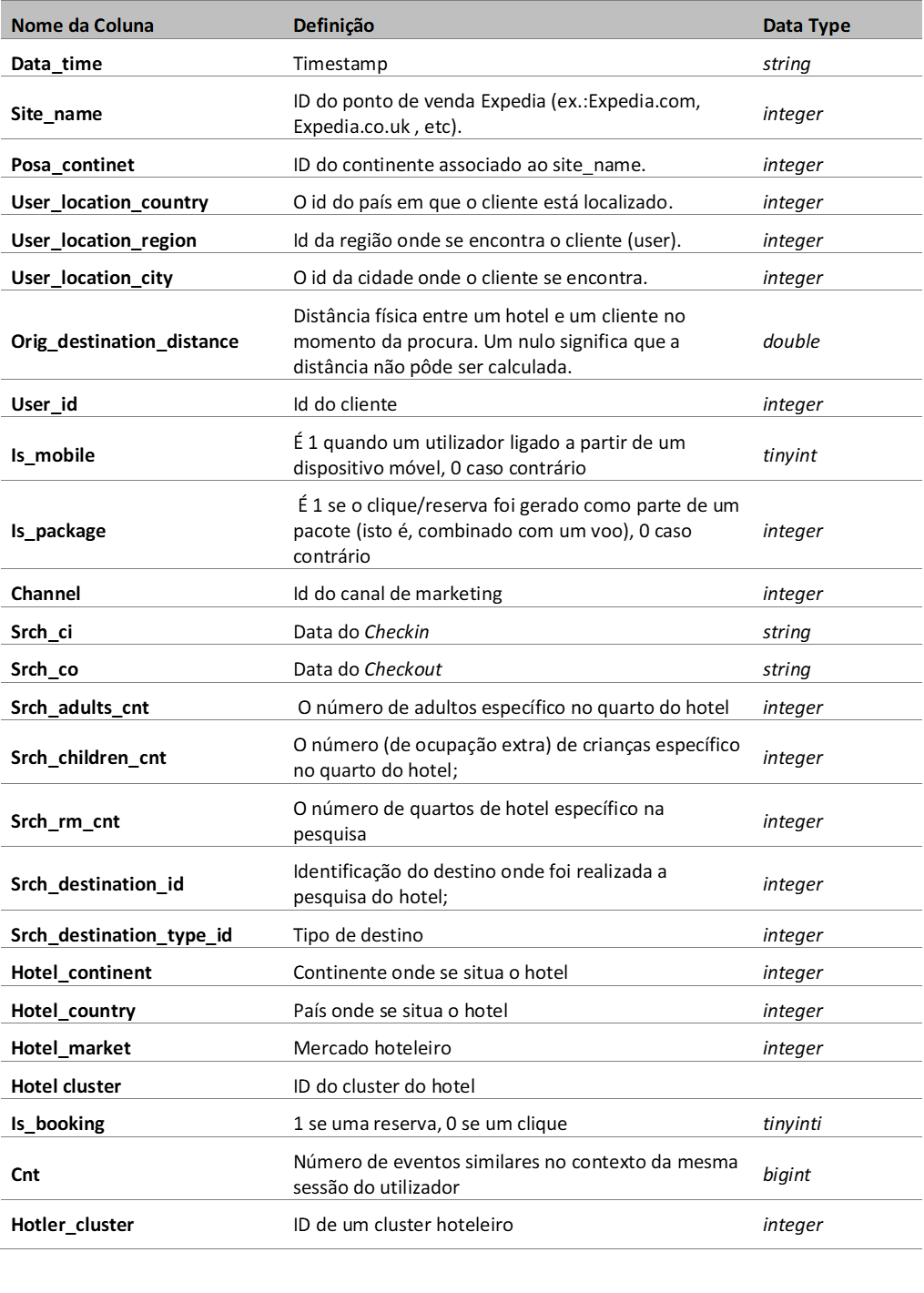
# Data Understanding

# Data Explaining

O *dataset* utilizado no projeto (train.csv) contém dados recolhidos entre 2013 e 2014. Optámos por utilizar este *dataset* de treino, tendo em conta que o *test dataset* tem um reduzido número de registos, o que poderia contribuir para o enviesamento da análise dos resultados, quando fosse implementado o modelo. Embora o *dataset* utilizado esteja desatualizado, optámos por aplicá-lo por estarmos a elaborar um trabalho de cariz académico. Caso se tratasse de um trabalho desenvolvido em contexto profissional, optaríamos por requisitar dados mais recentes, de forma que a previsão se aproximasse mais da atualidade.

Os conjuntos de dados de treino estão divididos com base temporal: dados de treino de 2013 e 2014. Os dados do placard público/privado também são divididos com base temporal. Os dados de treino incluem todos os utilizadores nos *logs*, abarcando os eventos de clique e eventos de reserva.

O *dataset* possui um total de 24 variáveis e aproximadamente 37 milhões de dados. Na tabela seguinte é possível consultar as variáveis, a sua definição e o respetivo *datatype*.



# Exploratory Data Analysis

Através da análise dos dados foi possível aferir que a variável *booking* não é equilibrada, uma vez que existem mais pesquisas sem reserva (consulta de mercado) do que o inverso.

Os dados mais pesquisados são:

**Número de noites** = 1, sendo o **nº de quartos** diretamente proporcional ao nº de noites.

**Nº de hotéis distintos** – 46.300

**Média de filhos** – 0,3

**Média de adultos** - 2

**Média de quartos** – 1,1

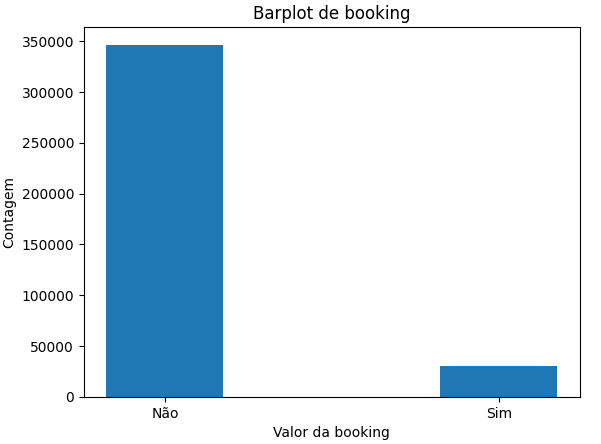
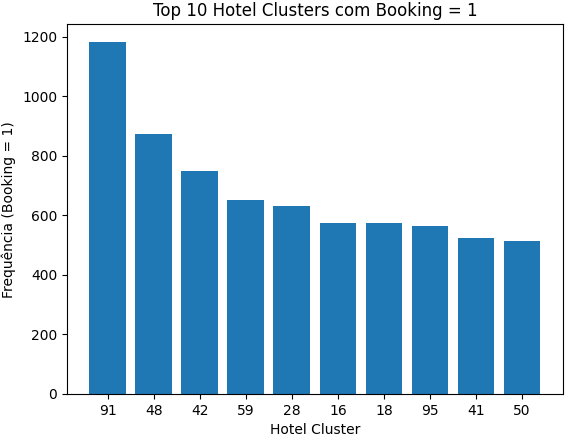
**Continente com mais pesquisas** - 3 (28.240.462)

**Site mais pesquisado** - 2 (23.790.351)

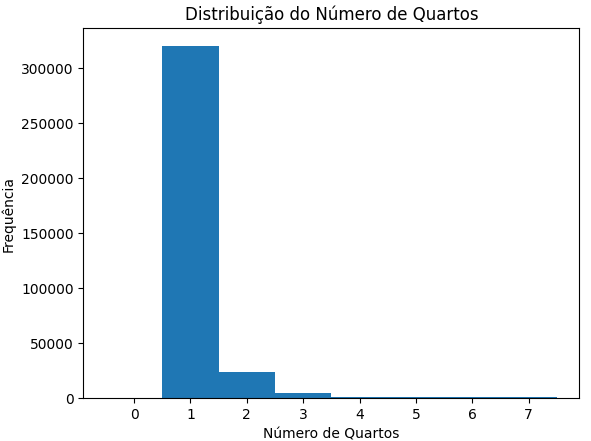
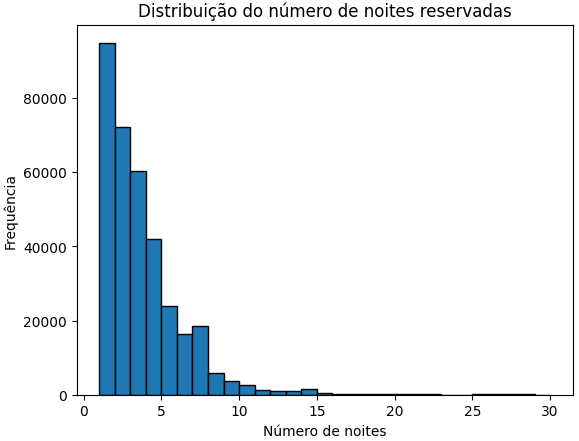
|  |  |
| --- | --- |
| ***Users* que fizeram pesquisa e reserva**  **(Booking = 1)** | ***Users* que fizeram apenas pesquisa (Booking = 0)** |
| 87% dos foram para fora do país | 92% dos *users* não fazem reserva |
| Mais de 50% das reservas são feitas para 2 pessoas | 13% dos clientes pesquisaram através de um dispositivo móvel |
| 33% das reservas são para uma única pessoa, sendo a média de 1,87 pessoas | A média de quartos por cada pesquisa é 1 quarto por cada pesquisa |
| 86% das reservas são feitas sem Package (= 0) | 13% das pesquisas são feitas para hotéis do mesmo país do *user* |
| 96% das reservas são para outros continentes | 90% das pesquisas com 0 crianças são para países fora do país do *user* que a efetua |

Nos gráficos abaixo, demonstram:

**Desequilíbrio da variável *booking* Top 10 de Hotéis com reserva**



**Resultados de pesquisa por nº de noites Resultados de pesquisa por nº de quartos**



# Data Preparation

# Data Cleaning

A variável orig\_destination\_distance apresentou 13.525.001 registos nulos, correspondendo a aproximadamente a 36% dos dados gerais. Como a maior parte dos hotéis não possui esta *feature*, excluímos a respetiva coluna.

Alterámos o nome das seguintes variáveis, para as seguintes descrições:

* is\_mobile **->** mobile
* is\_package **->** package
* srch\_adults\_cnt **->** num\_adults
* v**->** num\_children
* srch\_rm\_cnt **->** num\_room
* srch\_destination\_cnt **->** id\_destinantion
* srch\_destination\_type\_id **->** type\_destinantion
* cnt **->** similar\_srch
* is\_booking **->** booking

Para o aplicar o modelo na AWS retirámos da variável hotels\_col\_interest as colunas referentes às características dos hotéis (*cluster, market, country* e *continent*), visto que estas só foram úteis para a parte da análise dos dados.

Para finalizar, os dados foram guardados em dois ficheiros de formato parquet, por forma a tornar mais rápida a leitura dos dados. Um ficheiro mais pequeno, com cerca de 1% do *dataset* foi utilizado para estudar as probabilidades e correlações de dados. O segundo ficheiro representa a totalidade dos dados, com as variáveis úteis para implementação do modelo de *Recommendation Systems.*

Com o gráfico de correlação foi possível aferir que quanto maior o número de noites, maior é a probabilidade da reserva de um package. Quanto mais adultos estiverem incluídos nas pesquisas, maior é a probabilidade desta ter mais do que um quarto.

# Data Preprocessing

Na fase de pré processamento de dados, foram considerados os seguintes critérios:

**Variáveis Categóricas**: foi utilizado *one hot encoder* para fazer a codificação dessas variáveis previamente à alimentação do modelo.

**Variáveis Numéricas:** neste tipo de varáveis aplicámos uma standardização.

**Rating:** representado pela variável is\_booking, ao qual adicionámos a média desta variável para suavizar os resultados obtidos, de modo a que a pesquisa não apresente um *rating* igual a zero.

# Feature Engineering

Com o objetivo de potenciar o nosso modelo, foram criadas duas novas variáveis Id\_hotel e num\_nights. A primeira variável (Id\_hotel) foi criada com intuito de facilitar a identificação dos hóteis e é constituída pelo hotel\_cluster, hotel\_market, hotel\_country e hotel\_continent. A segunda variável, num\_nights, foi criada através das datas de *check-in* e *check-out,* servindo para demonstrar o número de noites que os utilizadores pretendem ficar no hotel. Estas variáveis continham campos a *Null,* pelo que optámos pela sua eliminação, uma vez que representavam uma quantidade muito reduzida face o *dataset* (0,13%).

# Modeling and Evaluation

Ao nível da modelagem e avaliação, dividimos o *dataset* em treino e teste, o que permitiu utilizar 80% dos dados para treino e 20% para teste. Foi também aplicada uma divisão estratificada na variável *booking,* uma vez que o *dataset* estava desequilibrado, evitando assim que todas as variáveis que continham booking igual a 1, acabassem no *dataset* treino por exemplo.

Foram utilizados os dados de treino para teste do modelo ALS, no qual foi considerado o *user\_id* como utilizadores, o *Id\_hotel,* como hotel e o *booking,* como o *rating*. Para encontrarmos *users* semelhantes foram considerados os seguintes atributos:

* num\_nights
* num\_rooms
* num\_adults
* num\_children
* channel
* mobile
* booking
* package
* user\_location\_country
* user\_location\_city
* user\_location\_region

Recorremos ao AWS para testar o modelo com a totalidade dos dados disponíveis, dada a elevada volumetria.

Para avaliar a performance analisámos as seguintes métricas:

|  |  |
| --- | --- |
| **Métrica** | **Valor** |
| Precision | 83.30% |
| F1 | 30.05% |
| Accuracy | 18.80% |
| Recall | 18.30% |

Com base nos dados da tabela acima, verificou-se que o modelo utilizado teve performances muito baixas, com exceção da precisão que aferiu o valor mais elevado (83,3%), significando que os utilizadores gostaram e efetuaram reserva com base no modelo aplicado. Porém, tivemos um *recall* baixo o que significa que perdemos a maioria dos *users* que fizeram a reserva.

No seguimento do modelo CRISP-DM, deveríamos retomar a fase do *business understanding* e *data understanding* para obtermos uma melhor compreensão sobre o motivo desta baixa performance; se devemos simplificar o modelo com menos dados; se é necessária a utilização de outros dados; ou ainda se deverá ser aplicado outro modelo.

# Deployment

A implantação do modelo passa por criar uma pipeline que efetue todo o pré-processamento e treino, finalizando com armazenamento do modelo em formato pickel. O modelo foi armazenado neste formato, de modo a ser possível a sua aplicação em contexto real, permitindo efetuar previsões mais recentes e atuais no *website* da empresa.

# Conclusão

Através da realização deste trabalho, resolvemos um problema, no âmbito da ciência dos dados, recorrendo às técnicas e metodologia de *machine learning*.

A metodologia aplicada neste projeto é conhecida como CRISP-DM, e está estruturada nas seguintes fases:

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modeling*
5. *Evaluation*
6. *Deployment*

A metodologia foi aplicada sobre o *dataset* selecionado, relacionando o Expedia Hotels, com o objetivo principal de criar um modelo de recomendação, assente nos atributos/características das pesquisas efetuadas por cada cliente (*user*).

Através da análise exploratória de dados, conseguimos identificar diversos *insights* que ajudaram a compreender melhor o modelo e a selecionar as *features* mais adequadas. Implementámos o modelo ALS, embora este não tenha obtido uma performance muito boa.

Os próximos passos seriam otimizar o modelo adicionando mais *features*, através da *feature engineering*, simplificando reduzindo *features* irrelevantes ou aplicando outros modelos. De qualquer das formas, optámos por salvar o modelo com pickel, para podermos utilizar ou até fazer um *deploy* numa plataforma de *cloud,* que permita aplicar o nosso modelo como a AWS.