INSTITUTO UNIVERSITÁRIO DE LISBOA

Unidade Curricular de Processamento de Big Data

Projeto Final

Elaborado por:

André Simões Novo, n.º 93343

Sebastião Manuel Inácio Rosalino, n.º 98437

Licenciatura de Ciência de Dados - 2º ano - Turma CD

Ano Letivo 2021/2022 - 2º Semestre

Professores:

Adriano Lopes

João Pedro Oliveira

Data de entrega: 09 abril de 2022

Índice

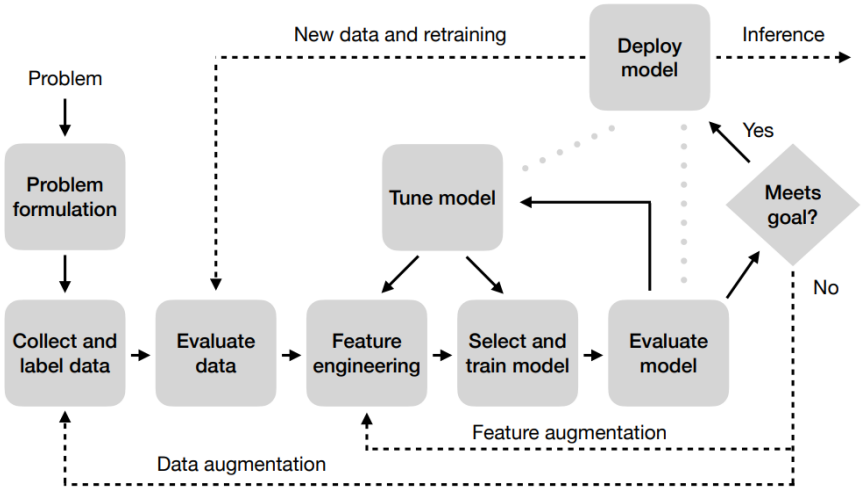
1. Objetivo e plano de trabalho

O presente trabalho tem por objetivo implementar uma solução computacional para estudo e análise de um problema com dados em larga escala, envolvendo a construção de um modelo de aprendizagem automática. Os dados tratados são de grande dimensão e foram retirados da Open Data da AWS (https://registry.opendata.aws/).

No desenvolvimento do presente trabalho, para além da capacidade computacional que os seus autores dispunham, foram utilizados serviços em ambiente *cloud* da Amazon para contexto académico: o AWS Academy.

Em termos de ferramentas, o projeto foi implementado recorrendo a funcionalidades disponibilizadas pelo Apache Spark e à linguagem de programação Python.

A implementação da solução teve por base a metodologia ML Pipeline, sistematizada na figura abaixo (retirada dos slides da UC):



Procurar-se-á, no presente relatório, descrever resumidamente as fases de realização do trabalho, a forma de implementação do algoritmo e as principais conclusões retiradas da sua utilização.

2. Identificação do domínio de dados e formulação do problema

a) Domínio dos dados

Como primeira tarefa, procedeu-se à pesquisa de um *dataset* (na Open Data da AWS) que contivesse, no mínimo, 3 GB de tamanho.

Para além desta condição, procurou-se escolher uma base de dados que contivesse informação interessante e, ao mesmo tempo, pertinente para o desenvolvimento das tarefas do projeto proposto.

O *dataset* escolhido integra dados de viagens realizadas por táxis e outros veículos de aluguer na cidade de Nova Iorque entre 2009 e 2021 (*New York City Taxi and Limousine Commission (TLC) Trip Record Data*), cujo link de acesso é o seguinte:

https://registry.opendata.aws/nyc-tlc-trip-records-pds/

De modo a cumprir o requisito de tamanho mínimo do conjunto de dados a estudar, optou-se por trabalhar os dados dos meses de julho e agosto de 2014, cujo tamanho ascende a 4.1 GB. O *dataset* inclui 25 795 242 registos.

O conjunto dos dados descreve, no essencial, todos os elementos relevantes de uma viagem efetuada por táxi ou por outro veículo de aluguer, desde o número de passageiros transportados, ao valor total pago por viagem, à rota percorrida, etc. (ver na seção 3 toda a estrutura de dados).

b) Formulação do problema

Os dados em análise traduzem um histórico de viagens na cidade de Nova York efetuadas por viaturas de aluguer, recolhendo informação dos operadores do setor registados na NYC Taxi & Limousine Commission.

Esta informação é de grande importância para as empresas do setor, no sentido em que lhes pode obter *insights* sobre, entre outros aspetos, custos médios por viagem, número de passageiros por viagem, percursos mais procurados, adequabilidade da frota, distâncias percorridas, custos com portagens, impostos, entre outros elementos.

No presente trabalho, com base num histórico de viagens de 2 meses procurar-se-á explicar/prever o valor (*total_amount*) que um cliente irá pagar numa futura viagem através da aplicação de um modelo de aprendizagem supervisionada, usando um algoritmo de regressão linear e utilizando as variáveis do *dataset* que apresentação uma maior correlação com a varável a explicar (*total_amount*).

A resolução deste problema pode gerar valor aos clientes e às empresas, porquanto pode disponibilizar o custo aproximado da viagem previamente à sua realização; adequar a escolha e distância dos percursos ao valor da despesa prevista; adequar a viatura e o serviço às características da viagem (tempo, n.º de passageiros, rota, etc.).

Neste quadro, o objetivo do presente trabalho passa por prever, com base em dados históricos padronizados, o custo que uma viagem deverá ter, conhecendo as suas características especificas (preditores).

3. Análises aos dados e testes realizados

Formulado o problema e conhecido o objetivo a prosseguir, iniciou-se o processo de tratamento de dados, tendo sido percorridas as seguintes etapas:

1ª) - Extração dos dados do *dataset*, recolhendo a informação dos meses de julho e agosto de 2014 (cumprindo os critérios do trabalho).

Existem 12.484.250 registos com pelo menos um valor omisso, o que correspondia a 48.4 % do número total de linhas.

Optou-se por não eliminar os registos com *nulls* já que a sua eliminação significaria a perda de quase metade do *dataset*. Verificou-se, ainda, que a sua manutenção não criava dificuldades acrescidas à realização do trabalho.

2ª) - Análise à estrutura das variáveis e respetiva classificação

Nome da variável	Definição	Classificação
Vendor_id	Código do prestador do serviço	String
pickup_datetime	Inicio da viagem	String
dropoff_datetime	Fim da viagem	String
passenger_count	Número de passageiros	Integer
trip_distance	Distância percorrida	Double
pickup_longitude	Longitude do sítio inicial	Double
pickup_latitude	Latitude do sítio inicial	Double
rate_code	Tarifa da rota	Integer
store_and_fwd_flag	Registo guardado ou não	String
dropoff_longitude	Longitude do sítio final	Double
dropoff_latitude	Latitude do sítio final	Double
payment_type	Tipo de pagamento	String
fare_amount	Valor das tarifas da viagem	Double
surcharge	Valor das sobretaxas	Double
mta_tax	Imposto sobre o serviço	Double
tip_amount	Valor da gorjeta	Double
tolls_amount	Portagens da viagem	Double
total_amount	Total pago	Double

3ª) - Constituiu-se uma subamostra para realização de tarefas intensivas e frequentes, composta por 10% dos dados [1.269.809 de registos]. Esta subamostra permitiu uma análise detalhada da informação, a sua melhor compreensão e a produção de alguns indicadores estatísticos. Foi criada, neste âmbito, uma nova coluna (variável) correspondente à duração da viagem – Trip_Duration (fim - início). Para isso foi necessário converter as variáveis *pickup_datetime* e *dropoff_datetime* (que estavam previamente em string) em datas.

4ª) - A análise, compreensão e limpeza de informação incidiu, no âmbito desta sub-amostra, sobre as seguintes variáveis:

i) rate_code (tarifa da rota), concluindo-se que existiam 7 "tarifas de rota" diferentes (0 a 6). Foram eliminados todos os registos que tinham uma "tarifa de rota" superior a 6, por se concluir que eram outliers (pe.: uma única viagem) ou valores incorretos;

ii) payment_type (tipo de pagamento), concluindo-se que os tipos de pagamentos frequentes são: Credit card; Cash; No charge; Dispute; Unknown;

iii) Passenger_count (número de passageiros), concluindo-se que existiam entre 0 a 9 passageiros. Considerou-se que o "0" poderia corresponder a apenas transporte de bagagem;

iv) Store_and_fwd_flag (registo guardado ou não), concluindo-se que a maioria dos motoristas não guardam os dados logo que após o final da viagem, por razões de ligação.

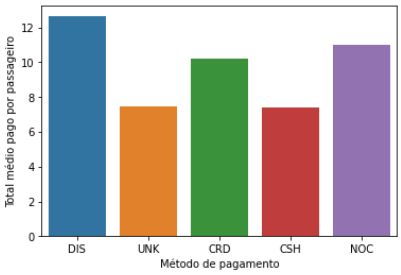
Todas as demais variáveis não suscitaram dúvidas o seu âmbito de aplicação.

5ª) - Após a compreensão e limpeza de dados, procedeu-se à análise estatística e descritiva dos dados, sendo de relevar os seguintes aspetos para a resposta ao problema formulado: A distância máxima da viagem foi de: 90.4 (milhas) e o valor pago pela viagem varia entre: min [2.5] e max [550.0].

Análise gráfica:

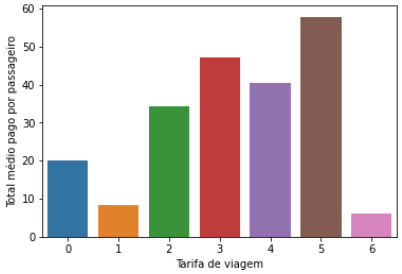
a) Cálculo e representação gráfica do valor médio pago por tipo de pagamento

	Método de pagamento	Total médio pago por passageiro
0	DIS	12.209143
1	UNK	7.260346
2	CRD	10.231366
3	CSH	7.439904
4	NOC	11.081927

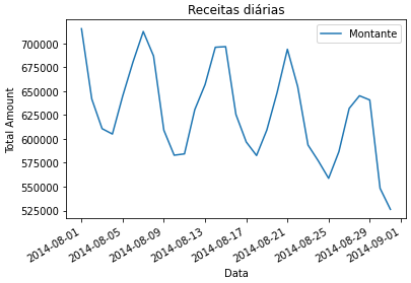


b) Cálculo e representação gráfica do valor médio pago por tarifa de viagem

	Códigos de viagem	Total médio pago por passageiro
0	0	11.234800
1	1	8.255433
2	2	33.997671
3	3	44.937373
4	4	42.054536
5	5	57.188112
6	6	6.528000



c) Representação gráfica da evolução das receitas no período em análise



6ª) Escolha do conjunto de variáveis a usar na implementação do algoritmo

O critério para escolha das variáveis a usar atendeu à correlação entre os preditores numéricos e a variável alvo.

Os valores de correlação podem ser observados na imagem seguinte:

```
passenger_count :  0.01569829163999021

trip_distance :  0.9236638149761347

pickup_longitude :  -0.0038727737744397306

pickup_latitude :  0.004279754854484198

rate_code :  0.5653735408337559

store_and_fwd_flag :  0.014237776734215224

dropoff_longitude :  -0.003437077639327101

dropoff_latitude :  0.003650084376583162

fare_amount :  0.9827981937840934

surcharge :  -0.05109163183034506

mta_tax :  -0.2990729548469016

tip_amount :  0.6492654979008966

tolls_amount :  0.6735357837188547

trip_duration :  0.8151073023644785
[Stage 278:=========================================>
total_amount :  1.0
```

Após análise das correlações, escolheu-se como preditores promissores os seguintes: 'trip_distance', ' rate_code', ' fare_amount', ' tip_amount', ' tolls_amount', ' trip_duration'.

4. Algoritmo implementado

Foi utilizado um modelo de aprendizagem supervisionada, baseado em regressão linear.

[Ver código da solução aplicacional nos notebooks remetidos]

O algoritmo de regressão opera através da divisão da subamostra em conjuntos de treino (70%) e teste (30%).

A aprendizagem algorítmica sobre os preditores é efetuada no conjunto de treino e a previsão do *total_amount* é feita no conjunto de teste.

Os resultados obtidos foram:

a) `R quadrado sobre o conjunto de teste = 0.994238`

O que significa que 99.4% (0,994238x100) da variância do total gasto por cliente é explicada pela variação dos preditores.

b) `Root Mean Squared Error (RMSE) no conjunto de treino = 0.965403`

O que significa que o afastamento das previsões obtidas pela regressão face ao real observado foi 0.9654 %.

5. AWS

Passou-se o dataset para a AWS de modo a usufruir dos serviços cloud de modo a permitir utilizar o dataset na sua integra (meses de julho e agosto). Exemplo de medidas descritivas sobre a totalidade dos dados:

summary	total_amount
count	25795207
mean	15.35619476207311
stddev	12.842169780742521
min	2.5
max	597.0

summary	trip_duration
count	25795207
mean	778.4846959359543
stddev	6614.14062636973
min	-7595670
max	607380

6. Conclusão

Face aos bons resultados do modelo concluímos que os preditores escolhidos [*'trip_distance', ' rate_code', ' fare_amount', ' tip_amount', ' tolls_amount', ' trip_duration']* são adequados para explicar a despesa total que será efetuada numa determinada viagem, o que já era esperado face às altas correlações dos preditores com a variável alvo.

Com esta informação, as empresas podem planear melhor a sua atividade, designadamente na definição do número de carros a ter em frota, do número de motoristas a contratar, das receitas potenciais, dos percursos mais rentáveis, das tarifas a aplicar, etc.

Cumpriu-se, assim, o objetivo proposto com o presente trabalho: prever, com elevada fiabilidade, o custo total que uma viagem deverá poderá ter, conhecendo as suas características especificas.

Nota final:

Conjuntamente com o presente relatório, enviam-se no ficheiro .zip os *notebooks* com a solução computacional desenvolvida, incluindo o notebook realizado na AWS.

Bibliografia consultada

- Building A Linear Regression with PySpark and MLlib – consultado em https://towardsdatascience.com/building-a-linear-regression-with-pyspark-and-mllib-d065c3ba246a

- Open Data da AWS - consultado em https://registry.opendata.aws/

- New York City Taxi and Limousine Commission (TLC) Trip Record Data) - consultado em https://registry.opendata.aws/nyc-tlc-trip-records-pds/

- Big Data: Algorithms, Analytics, and Applications. Kuan-Ching Li et al., 2015, Chapman and Hall/CRC

- Learning Spark: lightning-fast big data analysis. H. Karau, A. Konwinski, P. Wendell & M. Zaharia, 2015, O'Reilly Media, Inc.

- Spark: The definitive guide: Big data processing made simple. B. Chambers, M. Zaharia, 2018, O'Reilly Media, Inc.