Universidade do Minho
Escola de Engenharia

Perfil de Machine Learning: Fundamentos e Aplicações

Sistemas Baseados em Semelhança

1º/2º Ano, 1º Semestre

Ano letivo 2022/2023

Enunciado Prático nº 3

Tema: *A Data Science Perspective*

Joana Mota PG45528

Tarefas:

T1.

a e b) No que diz respeito à *web activity*, esta apresenta uma média de 0.99 horas, ou seja, aproximadamente, a 1h de atividade na plataforma, mas as horas mais utilizadas pelos utilizadores são de meia hora. O desvio padrão é de 1.5 horas, o que significa que os dados estão dispersos, pois quanto maior o desvio de padrão, maior é a variação dos dados.

Quanto ao *sentimental rating*, a avaliação do sentimento está composta entre 0 e 5, sendo que a média deste é de 1.85, aproximadamente, 2. No entanto, a moda fica entre o rating de 0 a 0.5. Já o desvio padrão é 1.62, o que indica que os dados não se encontram concentrados em torno da média.

Em relação ao salário anual estimado, a média é de 57 718 euros e a moda destes localizam-se entre os 50 mil e os 70 mil euros. Já o desvio padrão é de 32 mil euros, o que mostra alguma dispersão dos dados.

A nível de contratos, o mínimo de contratos é de 0 e o máximo é de 4 por utilizador. A média do número de contratos por utilizador é de 1 contrato, já a moda é de 2 contratos por cliente. O desvio padrão é de 1.14, os dados estão mais concentrados em torno da média.

A média de idades dos utilizadores é de 48 anos, sendo que a moda se encontra entre os 36 e os 43 anos. Já o desvio padrão é de 11 contratos o que mostra alguma dispersão dos dados. Significa que existem idades muito diferentes.

c) As *features* que têm maior correlação positiva são *sentimental rating* com a idade e o salário anual estimado com *call activity*. As *features* com maior correlação negativa são *sentimental rating* com *birthday* e *birthday* com *age*. A *feature* género é a que não possui nenhuma correlação.

T2. Nesta tarefa foi criado uma espécie de *dashboard* de forma a visualizar alguns *plots*. Assim, consegue-se visualizar o número de registos por análise de sentimento, sendo que a "very negative" foi a mais registada e a "very positive" a menos, o que significa que a maioria dos utilizadores se sentem de forma negativa em relação à plataforma. De seguida, visualiza-se a percentagem de utilizadores por produtos, mostrando que o investimento privado é o mais utilizado. Também se visualiza o salário anual estimado para cada utilizador, sendo que este fica entre 10 000 euros e 130 000 euros, no entanto apresenta algumas exceções (*outliers*) acima dos 130 000 euros. Verifica-se que 50% dos utilizadores têm salários entre os 10 000 euros e os

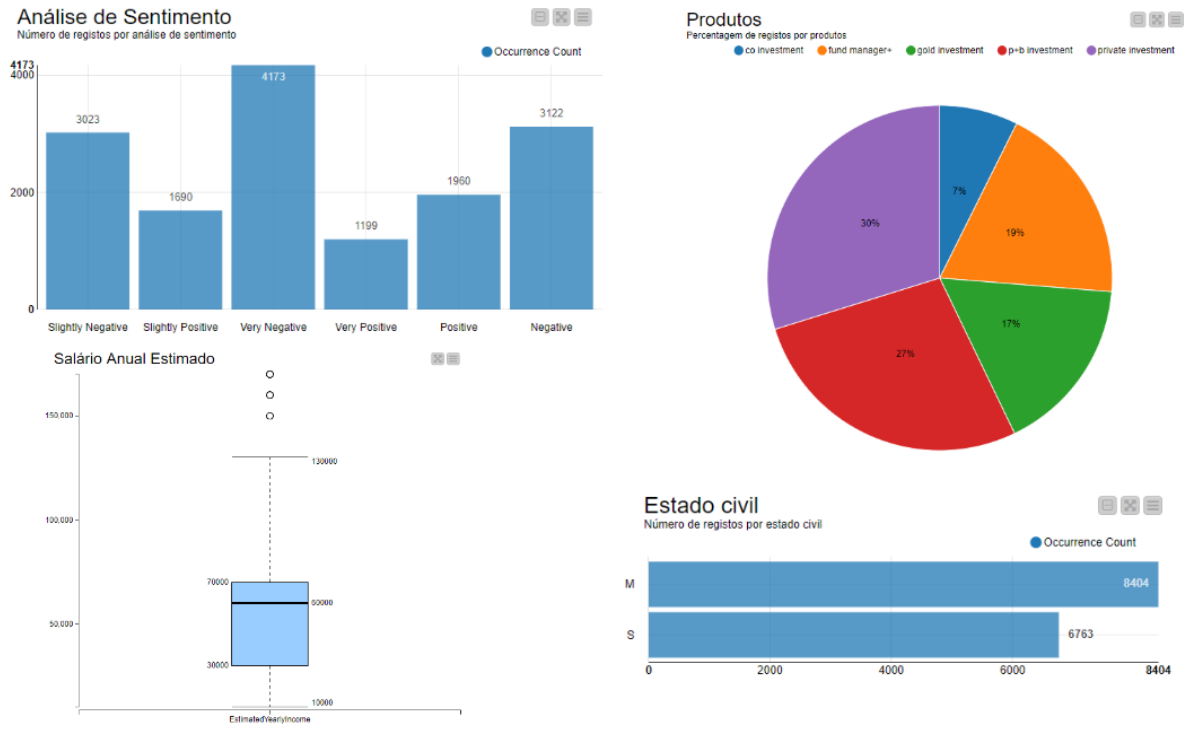

Figura 1. Plots

60 000 euros. Por fim, é apresentado o número de casados e de solteiros, sendo que existem mais utilizadores casados, do que solteiros na plataforma.

T3. De modo a tratar os dados, excluiu-se, primeiramente, as colunas do tipo *double*, utilizando o nodo *column filter*. De seguida, utilizando o nodo *missing values*, tratou-se dos registos em falta, neste caso, para valores de tipo *string* substituiu-se por "desconhecido" e para valores de tipo *numeric* substituiu-se por "-1". Isto servirá, mais tarde, para o modelo verificar que está em posse de dados que são diferentes, visto que estes são nulos. Posteriormente, removeu-se registos duplicados utilizando o nodo *duplicate row filter* e, seguidamente, criou-se três bins de igual frequência para a *feature* idade recorrendo ao nodo *auto-binner*. A seguir, passou-se a *feature birthday* para *string* de modo a poder extrair ano, mês e dia da semana, utilizando os nodos *string to date&time* e *extract date&time fields*, respetivamente. Depois, recorreu-se ao nodo *rule-based row filter*, para excluir utilizadores que tivessem uma atividade na plataforma inferior a 1 hora e que tivessem mais de 70 anos. Por último, excluiu-se os registos que tivessem a sub-*string* "co" no produto.

T4. Após o tratamento de dados, foram realizadas algumas agregações através do nodo *group by*. Primeiramente, agregou-se por género, obtendo o número e a percentagem de registos, bem como a o mínimo, o máximo e a média da idade dos utilizadores e a média da atividade da plataforma. De seguida, agregou-se por género e atividade na plataforma, de modo a obter a moda da análise

de sentimento dos utilizadores e média da avaliação do sentimento. Por último, agregou-se por análise de sentimento, de forma a obter o número de utilizadores, a média da estimativa do salário anual, o somatório deste e a média do número de contratos.

T5. Uma vez que as agregações foram feitas sequencialmente, isto é, depois do tratamento dos dados, os resultados serão diferentes do *dataset* inicial. Deste modo, quanto à agregação feita na alínea a) na T4, verifica-se que há mais utilizadores do sexo feminino do que masculino, sendo por isso o sexo feminino que apresenta, em média, uma maior utilização de atividade na plataforma. A idade mínima é de 30 anos e a máxima é de 70 anos, tendo uma idade média de 48 anos em ambos os sexos.

Relativamente à agregação da alínea b), verifica-se que quanto mais tempo de atividade na plataforma, mais o sentimento em relação à plataforma aumenta.

Por último, a agregação da alínea c) mostra que o sentimento que apresenta mais registos é "slightly negative". No entanto, os utilizadores que apresentam, em média, um maior salário são aqueles que sentem de forma muito negativa em relação à plataforma ("very negative"), que por sua vez é aquele que tem menos registos. Os utilizadores que possuem, em média, mais contratos são aqueles que sentem de forma bastante positiva em relação à plataforma ("very positive").

Resumidamente, através destas agregações conclui-se que a plataforma é ligeiramente mais utilizada pelas mulheres, das idades entre os 48 anos e quanto mais tempo passam na plataforma, cada vez mais têm um sentimento positivo em relação à mesma. Além disto, os utilizadores que têm maiores salários são aqueles que utilizam menos a plataforma e quando utilizam têm um sentimento negativo.

T6. Para este *dataset*, começou-se por explorar o top-5 de jogadores que marcaram mais golos, utilizando um nodo de *group by* de forma a poder agregar jogador por golo. Assim, verifica-se que os 5 jogadores que mais golos marcaram são: Mohamed Salah, Harry Kane, Sergio Agaçeiro, Jamie Vardy e Raheem Sterling. De modo a perceber quais as equipas respetivas a estes jogadores, utilizou-se um *joiner* e verificou-se que pertencem a Liverpool, Tottenham, Manchester City, Leicester City e Manchester City, respetivamente.

De seguida, analisou-se as equipas com mais golos e verificou-se que a equipa Manchester City foi a que marcou mais golos nesta edição. O que significa também que o jogador com mais golos não pertence à equipa que mais golos marcou. Também se analisou a média de idades nas equipas, sendo que a equipa West Bromwich é a que apresenta uma média de idades mais alta, com 28

anos, e a equipa com a média mais baixa de idades é o Liverpool, com 24 anos. Por último, explorou-se quais as posições de jogo que marcaram mais golos, neste caso, foram os avançados e os médios.