

EMPREGOS EM CYBER SECURITY E DATA SCIENCE

Dados na Ciência, Gestão e Sociedade

2022/2023

João Dias nº 110305
David Franco nº 110733
Felipe Licas nº 110861
Samuel Ricardo nº 110884
Rafael Cerqueira nº 110860

Introdução	3
Business/Domain Understanding	3
➤ Cyber Security	3
Como estão as tendências na área de Cyber Security?	4
➤ Data Science	5
Como estão as tendências na área de Data Science?	5
➤ Mercado Financeiro	6
Data understanding	6
➤ Análise Exploratória de Dados	9
Data preparation	12
Modeling	15
Evaluation	16
Principais Conclusões	16
Webgrafia	17

Introdução

O presente relatório foi realizado no âmbito da unidade curricular Dados na Ciência, Gestão e Sociedade e tem como objetivo analisar um conjunto de dados reais sobre os empregos em Ciência de Dados e Cibersegurança, de acesso livre (open source data).

As áreas de Ciência de Dados e Cibersegurança têm apresentado um crescimento significativo de ano para ano. No decorrer deste relatório vamos aprofundar a análise dos dois datasets do ai-jobs.net que contêm os salários e algumas características dos profissionais que trabalham nestas áreas. Esta análise tem como objetivo ajudar-nos a perceber se os salários praticados nestas áreas acompanham o crescimento verificado no setor e se a inflação e a guerra (temas que marcaram o ano de 2022) impactaram negativamente os salários praticados no setor tecnológico.

A par da análise descritiva das variáveis, vamos desenvolver um modelo de regressão para prever os salários dos profissionais de Ciência de Dados e de Cibersegurança.

O relatório seguirá a metodologia de Cross Industry Standard Process for Data Mining (CRISP-DM), processo base utilizado em múltiplos projetos na área de Data Science, que é constituído por 6 fases: Business/Domain Understanding, Data Understanding, Data Preparation, Modeling, Evaluation e Deployment. Como o objeto deste trabalho não passa por desenvolver um modelo para colocar em produção, optou-se por não aprofundar a última fase desta metodologia (deployment).

Business/Domain Understanding

A 1ª fase do CRISP-DM passa precisamente pela compreensão da área e do problema. Por esse mesmo motivo, estabelecemos algumas questões de partida, que vamos procurar responder com o desenvolvimento do relatório, nomeadamente:

- Como se pode prever o salário bruto dos profissionais nas áreas de Data Science & Cyber Security?
- Como está o mercado de trabalho relativamente a estas áreas? Haverão mais vagas? Com que condições?
- Quais são os países que apostam mais nestas duas áreas?

Cyber Security

Comecemos pela cibersegurança. Esta área tem sido cada vez mais procurada e melhor remunerada, uma vez que existem cada vez mais dados computacionais nas empresas e uma maior necessidade de os proteger de potenciais ciberataques.

De acordo com a TechTarget, o número de ciberataques cresceu 63% após o início da pandemia de Covid-19 em 2020 e em Setembro de 2021 registou um aumento de 17% em relação ao ano anterior. Um dos exemplos mais recentes foi o ataque à Vodafone, em Portugal, em fevereiro de 2022, onde a empresa viu os seus serviços suspensos até conseguir reparar os estragos.

Além disso, segundo a IBM o valor médio das empresas cresceu de 4.24 milhões para 4.35 milhões de 2021 para 2022. O aumento do número de ataques aliado ao aumento do valor das empresas leva a um

reforço nas áreas de proteção das mesmas, sendo que, de acordo com o Fortinet, 60% das empresas têm dificuldade em recrutar profissionais para a área de cibersegurança e 52% mostra-se com dificuldades em mantê-los.

Segundo o Fortune, os empregos de cibersegurança vão crescer em 33% na próxima década, ou seja, mais do que quatro vezes mais rápido que a média de empregos.

E como estão as tendências na área de Cyber Security?

A digitalização não é novidade e ao pé dela seguem-se as preocupações com a cibersegurança. Da mesma forma que os ataques se tornaram cada vez mais complexos e avançados em termos tecnológicos, o mesmo deve ser feito em relação às estratégias de segurança.

Estas estratégias traduzem as novas tendências em cibersegurança.

- **Plataformas de dados e informações em crescimento**

De uma forma geral, as empresas estão a recolher cada vez mais dados e a armazená-los na cloud, dando acesso a outras pessoas e organizações (como os fornecedores, por exemplo). Muitos ataques visam este acesso elevado de dados.

Para responder a estes ataques são necessários quatro recursos de cibersegurança essenciais, que devem ter-se em conta:

- a) Arquitetura de confiança zero (ZTA)
- b) Análise comportamental
- c) Monitorização de Elastic log para grandes conjuntos de dados
- d) Criptografia homomórfica

- **Os hackers estão a utilizar IA, machine learning e outras tecnologias para lançar ataques mais sofisticados**

Os hackers já não trabalham sozinhos. Hoje, o cyber hacking é parecido a uma empresa de grande escala, com diferentes níveis hierárquicos e com investimentos em pesquisa e desenvolvimento. Os hackers utilizam mecanismos avançados, como a IA, machine learning e diferentes automações. Os ataques são cada vez mais eficazes e ágeis.

A resposta a esta tendência passa pela utilização de mecanismos de automação capazes de combater estes ataques mais sofisticados, devendo concentrar esforços em respostas defensivas, como contramedidas do centro de operações de segurança (SOC) e atividades de trabalho intensivo. A IA e machine learning podem ser utilizados para acompanhar as variações nos padrões de ataque.

- **Há uma regulação cada vez maior, mas existem lacunas em recursos, conhecimento e talento, que superam a segurança cibernética**

A falta de profissionais de cibersegurança e a falta de conhecimento e experiência na área é um obstáculo quando se fala de segurança cibernética. A gestão de riscos cibernéticos não

acompanhou a crescente transformação digital, fazendo com que empresas não saibam como gerir os riscos digitais.

A incorporação de segurança nos recursos tecnológicos é uma forma de combater o problema, quer seja através do desenvolvimento de software seguro, quer seja através da codificação da estrutura e dos processos de engenharia de controlo.

Data Science

E o mercado de Data Science, como está?

Todos os dias são gerados grandes volumes de dados. Por esse motivo, há uma necessidade de haver pessoas que possam compreender e tirar conclusões desses dados.

A Data Science pode acrescentar valor a qualquer empresa, uma vez que permite a compreensão de uma elevada quantidade de dados provenientes de várias fontes com o objetivo de obter informações importantes que possam ajudar na tomada de decisões, de forma mais inteligente.

A procura por profissionais de Data Science está a crescer de dia para dia, um dos grandes exemplos são as empresas do setor financeiro e de retalho que pretendem aumentar a eficiência, reduzir custos e ficar um passo à frente da concorrência.

No entanto, a escassez de profissionais desta área é definitivamente um problema. O que acontece é que há muita procura por parte das empresas, mas pouca oferta deste tipo de profissionais. A oferta é escassa devido a esta carreira ser relativamente recente.

Para mostrar a relevância deste setor, consideremos a título de exemplo, o setor desportivo. Como pode a Data Science revolucionar as análises desportivas? Uma equipa da empresa Booz Allen especializada em Data Science e Machine Learning desenvolveu uma aplicação para os treinadores da MLB (liga americana de Baseball) capaz de prever qualquer lançamento de um pitcher com até 75% de precisão.

Este modelo tem como base várias estatísticas e situações de jogo para além da observação de todos os pitchers que fizeram mais de 1.000 lançamentos, o que proporciona a um treinador não só uma capacidade extraordinária de executar táticas realmente efetivas contra a equipa adversária, como também permite mudar a forma como o jogo acontece.

Este é um apenas um dos exemplos que realça a elevada importância de profissionais de Data Science. Se transitarmos entre setores, as opções são infinitas.

E como estão as tendências na área de Data Science?

Mesmo que nos últimos anos, as empresas já tenham aderido a tecnologias data-driven para orientar os seus negócios, as tendências relacionadas com a data science não param de chegar e transformar o mercado, seja na agilidade ou na eficiência de qualquer processo.

Os baixos custos associados ao armazenamento de dados e a eficiência no que diz respeito à recolha e extração de dados fazem da data science uma ferramenta altamente lucrativa para as empresas.

Existem inúmeras tendências relacionadas com Data Science, mas podemos destacar algumas delas que são e serão o “futuro” dos próximos anos:

- **Small Data**

As empresas estão a trabalhar cada vez mais com dados fáceis de gerir e com formatos acessíveis. Há cada vez mais empresas a aderir a mecanismos automáticos, utilizando dados em volumes menores para dar início às próximas estratégias data-driven.

- **Cloud Computing**

A adoção da cloud por parte das empresas é cada vez mais representativa, quer seja para melhorar a participação em quotas de mercado ou para melhorar a experiência dos clientes. Além disso, há dados que indicam que mais de 1000 milhões de toneladas de carbono podem ser evitadas com a adoção da computação em cloud de 2021 até 2024, de acordo com a IDC. Outros fatores como a velocidade, o peso e a interface da computação em cloud fazem disto uma boa opção tanto para clientes, como para empresas.

Mercado Financeiro

As áreas de Cyber Security e Data Science são muito apelativas e em ambos os casos, há tendências que se verificam e que apontam para um crescimento notável nos próximos anos, mas como está a economia global? Este é um ponto fundamental de perceber, pelo facto do objeto deste trabalho não poder ser dissociado da economia. A oferta de emprego está estreitamente ligada ao estado dos mercados financeiros. Por isso, passamos à análise da economia global.

Atualmente, a economia global está a sofrer uma grande recessão muito por culpa da invasão russa e da persistente e prolongada pandemia da COVID-19, fazendo com que as projeções financeiras e económicas sejam cada vez menos otimistas. A previsão do crescimento económico global aponta para um abrandamento de 2.7% em 2023. Trata-se da previsão mais baixa desde 2001, tirando a crise financeira global de 2009 e a fase mais grave da pandemia. Está previsto também que, um terço das economias mundiais sofram um retrocesso quando os Estados Unidos, a China e a União Europeia estiverem perto da estagnação.

Outro fator que está a influenciar o crescimento económico global é a inflação, que está no seu valor mais alto dos últimos anos, e que está a tornar o custo de vida cada vez mais caro e insustentável para grande parte das pessoas. Este valor subiu de 4.7% em 2021 para 8.8% em 2022, mas prevê-se que em 2023 desça para 6.5% e para 4.1% em 2024. As perspectivas para o mercado são efetivamente positivas, mas para já aparecem condicionadas, em grande parte, pela guerra entre a Rússia e a Ucrânia.

Data Understanding

Tendo já uma noção clara do estado do mercado e das perspectivas para as áreas de Data Science e Cyber Security, importa agora perceber os dados presentes nos datasets a ser analisados, para aprofundar a análise no que diz respeito ao tipo de ofertas, salários, entre outros, que são praticados nestas duas áreas.

A compreensão dos dados que figuram os datasets fornecidos é uma fase de extrema importância para o CRISP-DM. Nesta etapa vamos procurar explorar o significado das diferentes variáveis. Ao mesmo tempo, vamos analisar os dados descritivos destas features e respectivas correlações para um maior

entendimento sobre o tema. A tabela abaixo mostra o nome das variáveis nos respectivos datasets, bem como o seu significado.

Dataset Data Science	Dataset Cyber Security	Significado
<i>Working Year</i>	<i>work_year</i>	A variável indica o ano em que o salário foi pago e é extremamente relevante para o desenvolvimento do modelo de regressão, uma vez que mostra as diferentes variações dos salários desde 2020 até 2022.
<i>Designation</i>	<i>job_title</i>	Feature que indica o cargo desempenhado, que não tem impacto para efeitos de modelação, uma vez que foi criada uma nova variável que agrupa todos os títulos de trabalho relacionados com Data Science e todos os títulos de trabalho relacionados com Cyber Security. Por esse mesmo motivo, não será considerada na seleção de features para treinar o modelo.
<i>Experience</i>	<i>Experience_level</i>	Nesta variável categórica, podemos encontrar 4 diferentes tipos de observações: EN que significa Entry-level/Junior; MI que significa Mid-level/Intermediate; SE que significa Senior-level/Expert e, por fim, EX que significa Executive-level/Director. Como o salário varia sempre consoante a experiência dos funcionários, esta variável é importante para treinar o modelo.
<i>Employment Status</i>	<i>employment_type</i>	Esta variável apresenta os diferentes regimes de trabalho e não é importante para efeitos de uma modelação uma vez que a maior parte das observações correspondem a FT (trabalhadores a full time). Esta feature apresenta como possíveis valores PT (trabalhadores a part time), CT (trabalhador a contrato) e FL (freelancer).
<i>Salary In Rupees</i>	<i>salary_in_usd</i>	Corresponde à remuneração dos trabalhadores em rúpias indianas e dólares americanos, respectivamente. Estas features foram retiradas dos datasets, uma vez que a partir delas foi criada uma nova variável com o salário em euros, que corresponde precisamente à variável target.
<i>Employee Location</i>	<i>employee_residence</i>	Corresponde ao país de residência do trabalhador e é apresentada no formato ISO 3166-1 alpha-2. Trata-se de uma variável muito relevante para efeito de modelação, uma vez que

		os salários variam de país para país.
<i>Company Location</i>	<i>company_location</i>	Corresponde ao país da empresa e é apresentada no formato ISO 3166-1 alpha-2. Trata-se de uma variável muito relevante para efeito de modelação, uma vez que tal como descrito acima os salários variam de país para país.
<i>Company Size</i>	<i>company_size</i>	Esta variável apresenta a dimensão das empresas e é muito relevante para treinar o modelo, uma vez que empresas maiores têm uma maior capacidade de conseguir oferecer melhores salários. Esta feature apresenta 3 opções: <ol style="list-style-type: none"> 1. S (Small) - Menos de 50 trabalhadores 2. M (Medium) - Entre 50 e 250 trabalhadores 3. L (Large) - Mais de 250 trabalhadores
<i>Remote Working Ratio</i>	<i>remote_ratio</i>	Corresponde à quantidade de trabalho não presencial efetuada pelos colaboradores e é uma variável relevante para treinar o modelo, uma vez que o trabalho presencial implica mais custos para as empresas do que o trabalho remoto, o que pode refletir-se a nível salarial. A feature apresenta 3 categorias: <ol style="list-style-type: none"> 1. 0 - Trabalhadores que trabalham remotamente menos de 20% do seu tempo. 2. 50 - Trabalhadores que trabalham metade do tempo de forma remota. 3. 100 - Trabalhadores que desenvolvem mais de 80% do seu trabalho de forma remota.
	<i>Salary</i>	Variável desconsiderada para efeitos de modelação, uma vez que foi criada uma nova variável com o salário em euros.
	<i>Currency</i>	Variável desconsiderada para efeitos de modelação, uma vez que foi criada uma nova variável com o salário em euros.

Tabela 1 - Descrição das variáveis presentes nos dois datasets

A tabela abaixo mostra a descrição das novas variáveis que foram criadas.

Merged Dataset	Significado
Salary_Eur	Esta nova variável representa o valor do salário bruto, em euros.
Type	<p>Esta nova variável representa a área de trabalho desempenhado. Tem duas opções:</p> <ol style="list-style-type: none"> 1. Data Science (empregos relacionados com Ciência de Dados) 2. Cyber Security (empregos relacionados com Cibersegurança)

Tabela 2 - Descrição das novas variáveis criadas no dataset final (Merged Dataset)

Análise Exploratória de Dados

A análise exploratória de dados permitiu conhecer de forma mais aprofundada os dados contidos nos datasets. Para facilitar a apresentação de algumas conclusões, procedeu-se a algum pré-processamento de dados, que permitiu juntar os dois datasets num só. Todas as transformações necessárias para tal são explicadas na secção seguinte, na fase de Data Preparation.

Conforme mostram as imagens abaixo, é possível perceber que algumas das features requerem algum tratamento. A variável relacionada com o salário em euros apresenta alguns outliers, uma vez que o valor máximo presente nas observações é de 938 321 euros por ano. Este valor não é considerado realista, uma vez que algumas das mais conhecidas plataformas de emprego como o Glassdoor e o Indeed apresentam valores máximos de salário em empregos nas áreas de Data Science e Cyber Security na ordem dos 250 000 euros.

Além disso, conseguimos perceber também que as variáveis Employee_Location e Company_Location apresentam inúmeras categorias, o que poderá dificultar o treino do modelo.

Através do widget Feature Statistics conseguimos perceber que nenhuma das variáveis apresenta missing values.

Algumas das principais conclusões retiradas da análise das distribuições permitiu saber que:

- a) Tem havido uma tendência crescente nas áreas de Data Science e Cyber Security, uma vez que de 2020 até 2022, tem havido cada vez mais registos de salários a ser efetuados a profissionais destas áreas. Com base na informação apresentada pelo dataset, passou-se de 256 salários efetuados em 2020 para 897 salários efetuados em 2022.
- b) O dataset apresenta bastante mais profissionais de Cyber Security do que Data Science.

- c) A maior parte dos registos de salários são de profissionais mid-level e seniores, o que pode justificar salários mais elevados, dada a experiência dos profissionais com este tipo de experiência.
- d) 1813 das 1854 observações são de funcionários que se encontram a trabalhar a full-time.
- e) A maior parte dos trabalhadores trabalham mais de 80% do seu tempo de forma remota. Só 360 dos 1854 funcionários trabalham menos de 20% do seu tempo de forma remota.
- f) Há pouca representatividade de empresas com menos de 50 trabalhadores e uma grande representatividade de empresas grandes, com mais de 250 trabalhadores, o que pode justificar também mais capacidade destas empresas para adotar metodologias de trabalho remotas e praticar salários mais elevados.

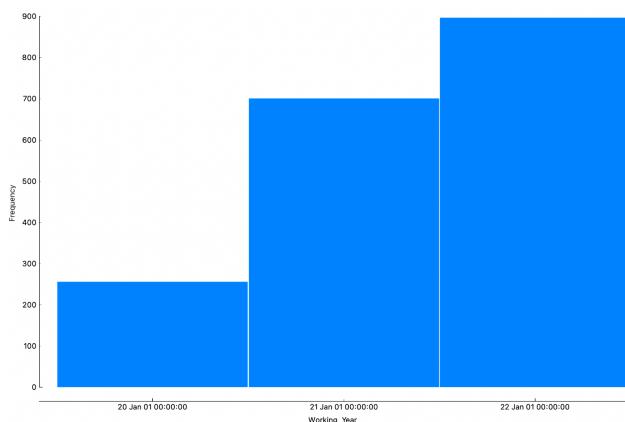


Fig. 1 - Distribuição da variável Working_Year

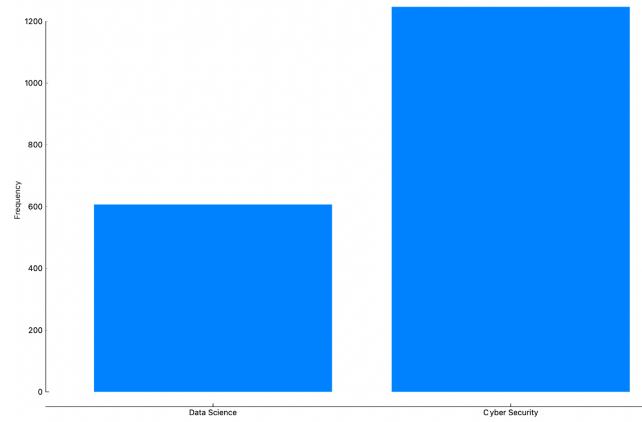


Fig. 2 - Distribuição da variável Type

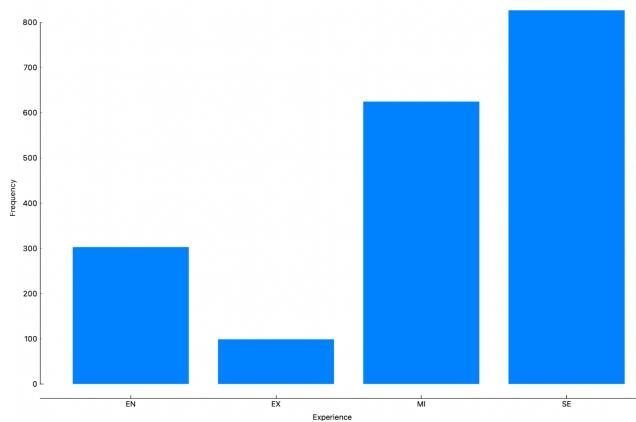


Fig. 3 - Distribuição da variável Experience

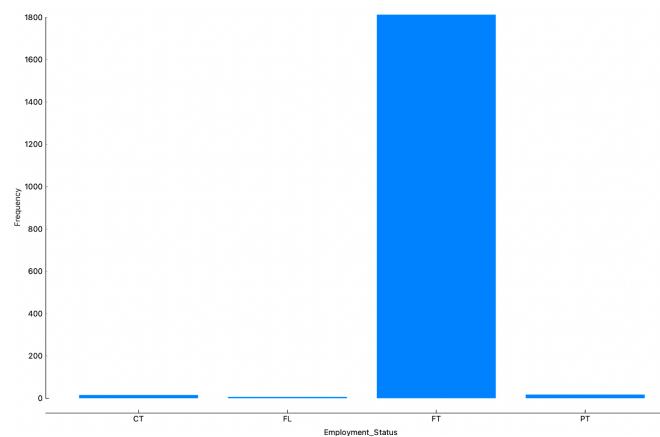


Fig. 4 - Distribuição da variável Employment_Status

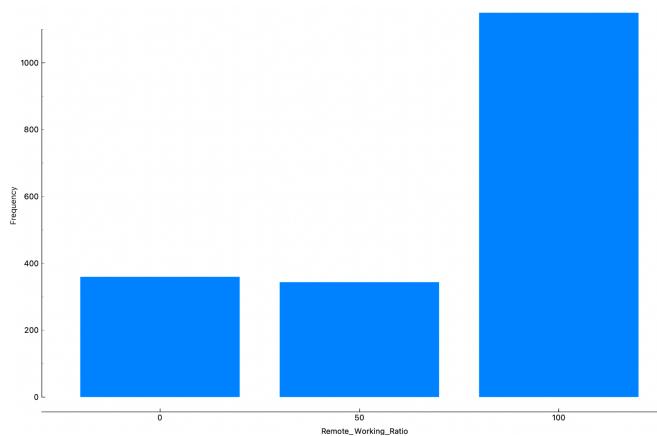


Fig. 5 - Distribuição da variável Remote_Working_Ratio

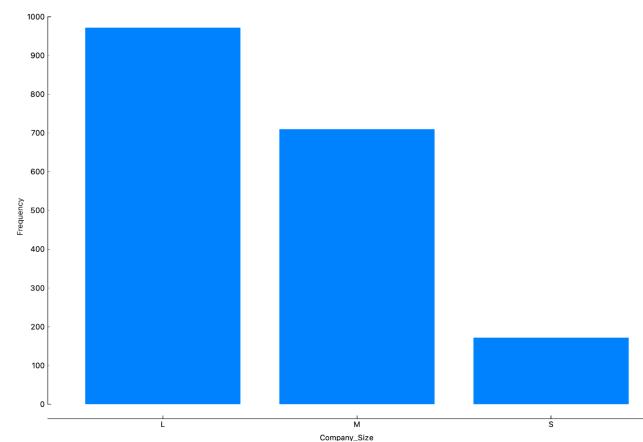


Fig. 6 - Distribuição da variável Company_Size

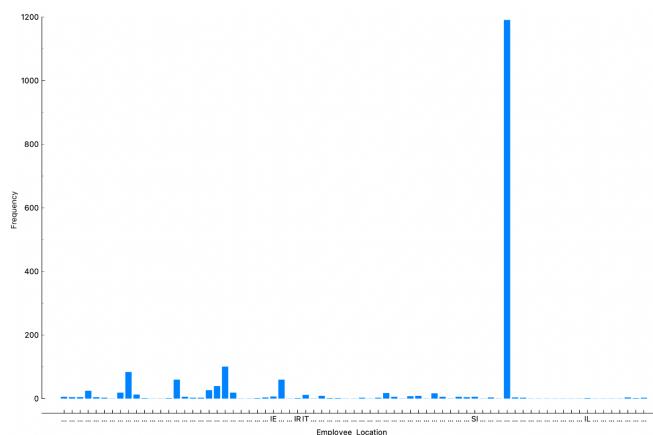


Fig. 7 - Distribuição da variável Employee_Location

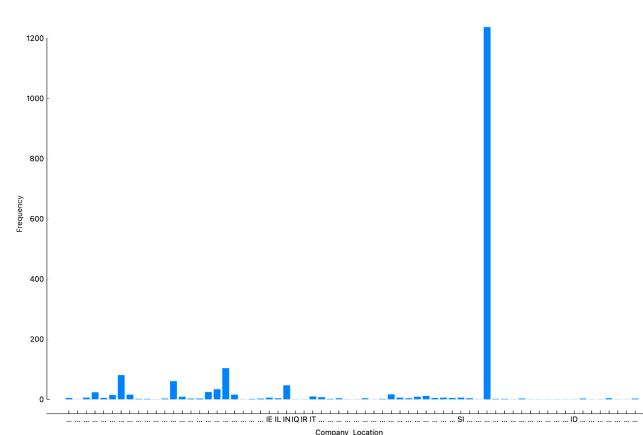


Fig. 8 - Distribuição da variável Company_Location

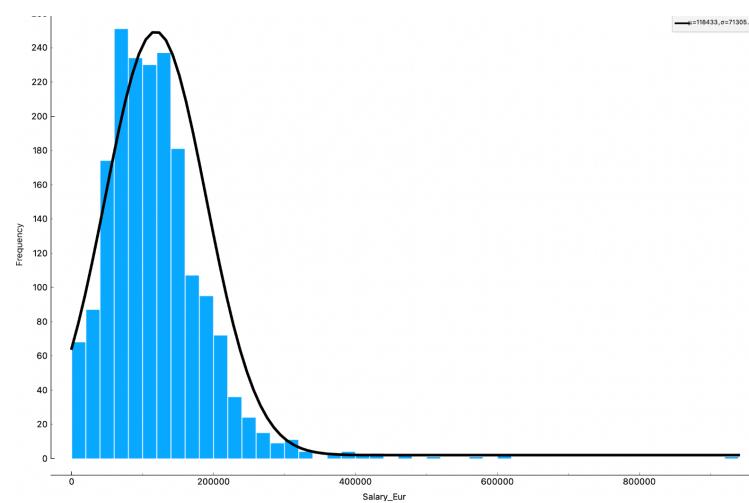
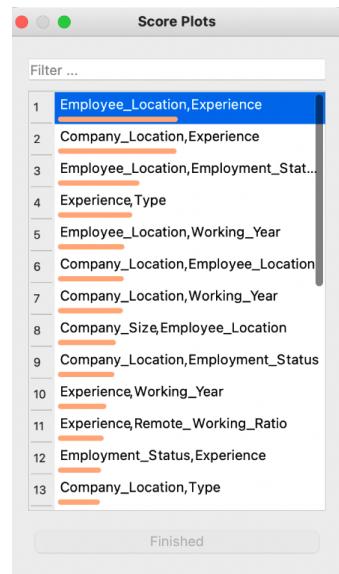


Fig. 9 - Distribuição da variável Salary_Eur

A imagem ao lado mostra-nos algumas das principais correlações entre as variáveis. As cinco maiores correlações relacionam:

- a) A localização dos trabalhadores com o nível de experiência;
- b) A localização da empresa com o nível de experiência;
- c) A experiência e a área de trabalho (Data Science ou Cyber Security);
- d) A localização dos trabalhadores com o ano em que foi pago o salário;
- e) A localização da empresa e a localização dos funcionários.



Data Preparation

Depois de conhecer de forma aprofundada as áreas de Data Science e de Cyber Security e depois de conhecer com mais detalhe os datasets e as variáveis neles contidas, começou-se a preparar os dados para a modelação. Esta preparação é uma fase extremamente importante de acordo com o CRISP-DM. Como os dois datasets continham observações diferentes, mas uma estrutura semelhante ao nível das variáveis, tal como já tínhamos mencionado anteriormente, foi feito algum pré-processamento para que pudessem ser juntos num só dataset. As features de ambos os datasets eram todas iguais, à exceção das que estavam relacionadas com o salário, que no dataset de Data Science aparecia em rúpias indianas, e no dataset de Cyber Security aparecia em dólares americanos. Para uniformizar esta feature, aplicámos uma transformação através do widget *feature constructor* que nos permitiu aplicar a taxa de câmbio a todos os valores das colunas nos respectivos datasets com a respetiva taxa de câmbio (à data de 20/10/22) e criar uma nova feature com o salário em euros (Salary_Eur) em cada um deles.

Depois de concluído este passo, eliminámos as colunas com os valores dos salários que não estivessem em euros, por serem redundantes. Importa destacar que as variáveis salary e currency do dataset de Cyber Security foram ignoradas ao fazer-se a importação inicial do ficheiro, uma vez que este dataset já disponibilizava o salário em dólares americanos.

De seguida, criou-se uma nova variável com a área de trabalho (Type), nos respectivos datasets, que agregava as inúmeras designações de cada uma destas áreas.

Por fim, transformámos os nomes das features do dataset de Cyber Security para corresponderem aos nomes das features do dataset de Data Science, para que depois pudéssemos aplicar o widget Concatenate, tratando as variáveis com o mesmo nome como a mesma variável. O dataset com estas transformações foi gravado e importado novamente para o Orange - programa utilizado para a visualização e preparação dos dados para modelação - para continuar a análise com o dataset final (Merged Dataset). A descrição das novas variáveis criadas, descritas acima, pode ver-se na secção anterior, na tabela 2.

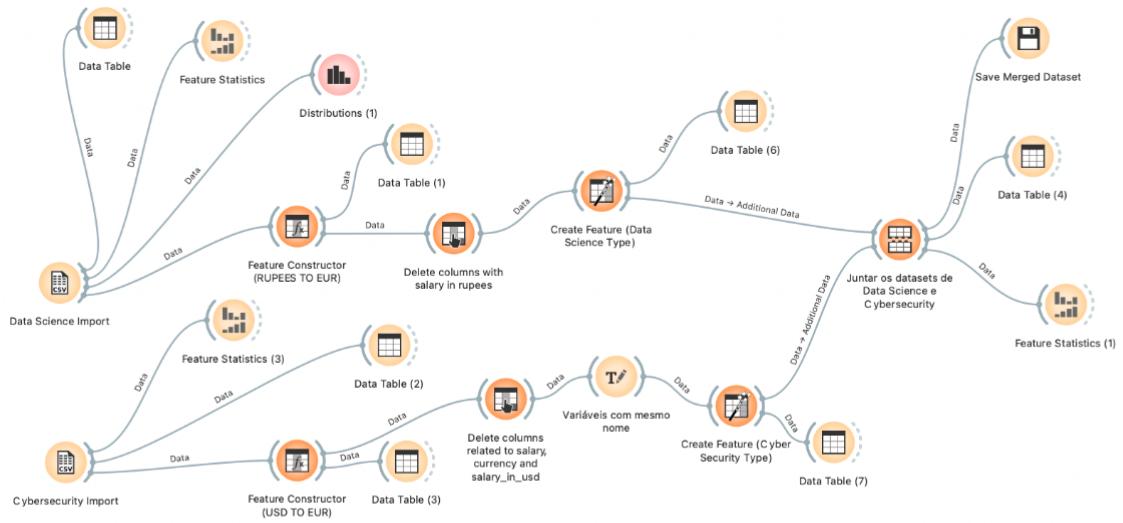


Fig. 10 - Pré-processamento de dados no Orange para juntar os datasets

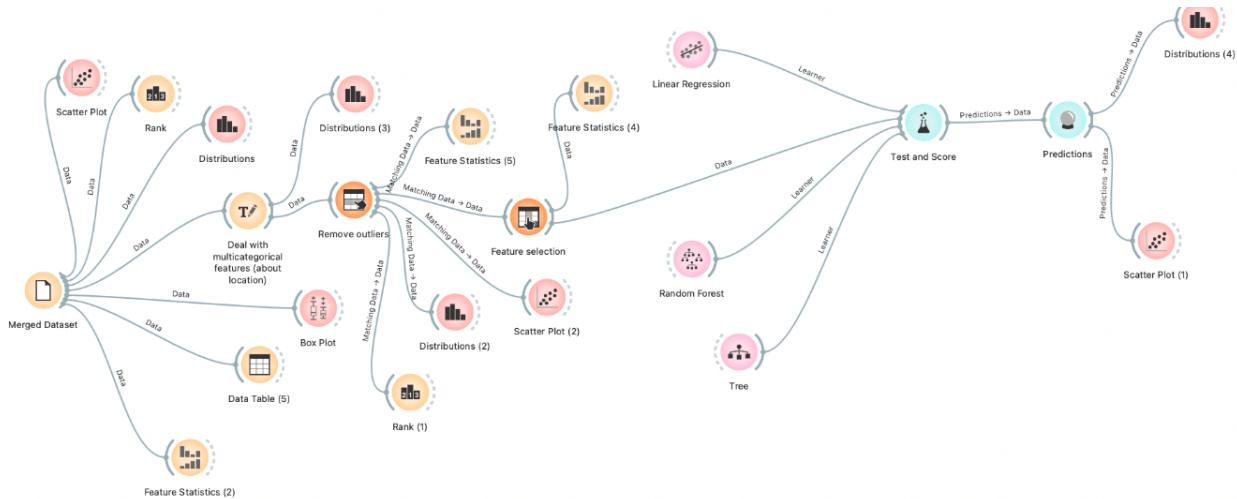
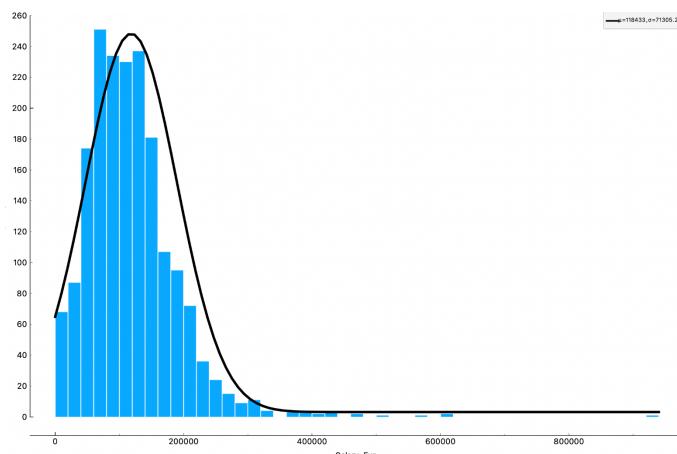


Fig. 11 - Preparação, modelação e avaliação do dataset final no Orange

Tal como mencionado na fase de Data Understanding, a variável `Salary_Eur` apresentava alguns outliers. Nesse sentido e tal como já foi mencionado anteriormente, de acordo com algumas das plataformas de emprego mais conhecidas como o Glassdoor e o Indeed, o salário mais alto de um Cientista de Dados e de um Especialista em Cibersegurança ronda os 250 000 euros por ano. Por isso, para



normalizar a distribuição da variável, decidimos remover as observações que se encontravam acima desse valor.

Além disso, nas variáveis Employee_Location e Company_Location podiam ver-se 73 e 66 diferentes categorias, respetivamente. Nesse sentido, aplicámos uma transformação, através do widget Edit Domain que nos permitiu manter as cinco categorias mais comuns em cada uma das variáveis e agrupar todas as outras como “Other”. Depois disto, tal como mostram as distribuições nos histogramas abaixo, é possível perceber que a maior parte dos registos deste dataset pertencem a funcionários dos Estados Unidos da América. O Reino Unido, o Canadá, a Alemanha e a Índia são os países imediatamente a seguir que têm maior peso no dataset analisado.

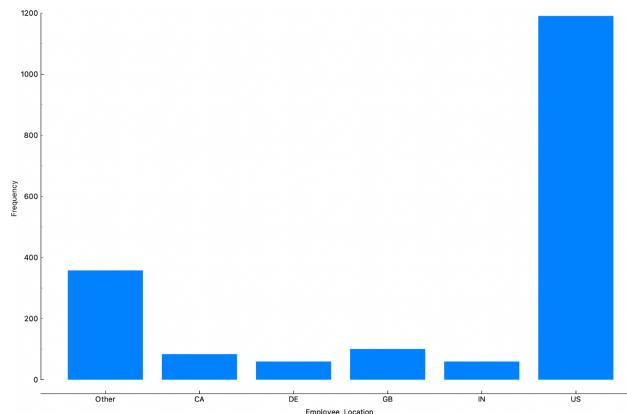


Fig. 12 - Distribuição da variável Employee_Location

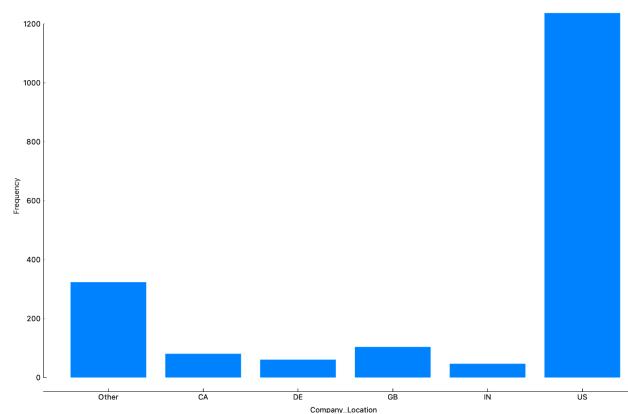


Fig. 13 - Distribuição da variável Company_Location

A imagem abaixo mostra de forma resumida as estatísticas dos dados no fim da etapa de Data Preparation.

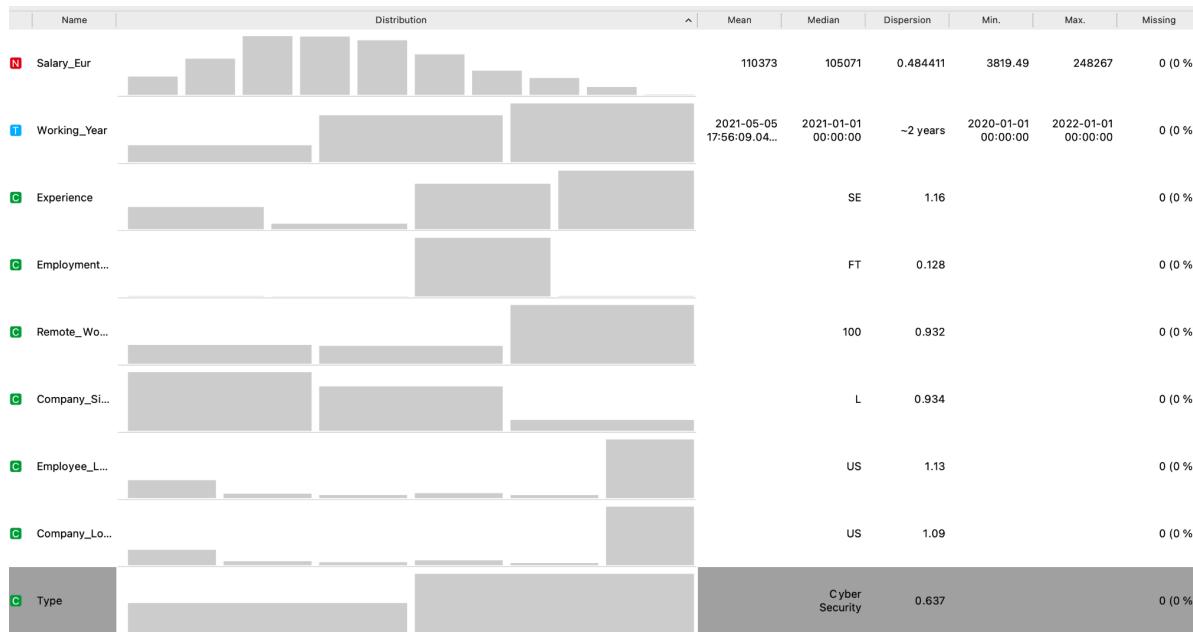


Fig. 14 - Feature statistics dos dados após preparação

Modeling

Face aos dados descritos, e uma vez que o objetivo passava por prever o salário bruto dos trabalhadores nas áreas de Data Science e de Cyber Security, definimos a variável Salary_Eur como o target. Por outras palavras, esta é a feature que vamos prever.

Para efeitos de modelação, não foram consideradas algumas features como:

- Designation (uma vez que foi feito algum feature engineering para criar a nova feature Type, que agrupa os diferentes cargos presentes na feature Designation);
- Employment_Status (uma vez que a maior parte das observações correspondem a Full Time e, tal como se pode ver na imagem ao lado, não tem relevância para a seleção de features utilizadas para treinar o modelo).

		#	RReliefF
1	C Employee_Location	6	0.237
2	C Type	2	0.159
3	C Remote_Working_Ratio	3	0.138
4	T Working_Year		0.137
5	C Company_Size	3	0.114
6	C Experience	4	0.102
7	C Company_Location	6	0.099
8	C Employment_Status	4	0.032

Tratando-se de um problema de regressão, optámos por selecionar três técnicas de modelação - regressão linear, decision tree e random forest - que têm pressupostos diferentes e que serão descritos abaixo de forma resumida.

- Para a regressão linear foi necessário assumir que:
 - Há uma relação linear entre as features e o target;
 - Não existe multicolinearidade entre as variáveis independentes (ou features);
 - Tem-se o pressuposto da homocedasticidade - o error term é o constante em todos os valores das variáveis independentes (ou features);
 - Há uma distribuição normal dos residuais.

Além disso, de acordo com a documentação do programa, o Orange faz algum pré-processamento de forma automática para implementar o widget da Linear Regression, transformando as variáveis categóricas presentes no dataset em variáveis contínuas, através de One-Hot-Encoding. Este é um passo essencial, que não foi feito através do widget Preprocess, por ser automático ao selecionar-se o widget da Linear Regression.

- Ao contrário da regressão linear, tanto a Decision Tree como a Random Forest não requerem um grande investimento na preparação dos dados. Esta é uma das principais vantagens de utilizar estes modelos em particular.

Evaluation

Depois da aplicação dos três modelos (Decision Tree, Random Forest e Linear Regression), procedemos à análise das métricas de avaliação da performance dos modelos, recorrendo ao widget Test and Score. De uma forma geral, a regressão linear e a random forest apresentaram um desempenho muito similar, com um coeficiente de determinação próximo de 0,5, conforme mostra a imagem abaixo, reforçando a correlação linear verificada entre as features e o target. Este R2 mede a variância da variável dependente (target) que é explicada pelas variáveis independentes. O modelo que apresentou o pior desempenho foi o da Decision Tree.

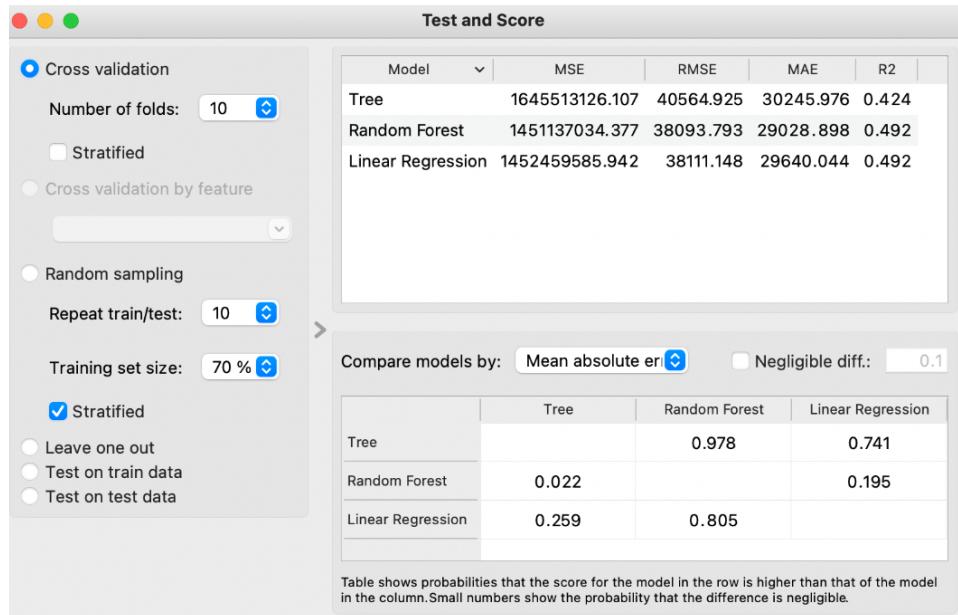


Fig. 15 - Métricas de avaliação da performance dos modelos

Principais Conclusões

A partir do Business Understanding e da Data Understanding foi possível concluir que tem havido um aumento da procura de profissionais nas áreas de Data Science e Cyber Security. Os dados analisados mostram que há cada vez mais vagas nestas áreas, mas faltam profissionais e talento para poder desempenhar este tipo de funções. Apesar da guerra ter vindo a prejudicar o estado de crescimento do mercado, prevê-se que até 2024 o valor da inflação volte a baixar, trazendo alguma estabilidade para a contratação deste tipo de profissionais.

Há espaço para crescer e as tendências retratam bem este facto. Contudo, importa reforçar que na análise deste dataset vimos grande parte das observações corresponderem a profissionais com um nível de experiência intermédio e sénior, levando-nos a questionar se podemos considerar existirem profissionais seniores em áreas tão recentes como Data Science e Cyber Security?

Ainda na análise do dataset, vimos uma aposta nestas áreas a ser feita, na maioria, por empresas de grande dimensão. A maior parte dos casos analisados retratava observações de funcionários de

empresas com mais de 250 trabalhadores, com uma maior expressão de posições em Cyber Security, face às existentes em Data Science.

Nestas áreas, os salários são efetivamente elevados e o trabalho remoto é uma realidade na maior parte dos casos analisados. Já no que diz respeito aos países que mais apostam nestas áreas, podemos destacar os Estados Unidos da América, seguindo-se do Canadá, Reino Unido, Alemanha e Índia.

Para reforçar a análise, foram desenvolvidos alguns modelos que tinham como objetivo prever o salário bruto dos profissionais destas áreas. De uma forma geral, os resultados dos modelos mostraram a existência de uma correlação linear entre as features presentes neste dataset e o respetivo target. Contudo, importa reforçar que podem ser obtidos valores maiores nas métricas de performance dos modelos, se acrescentarmos observações para treinar os modelos, ou alternativamente, se recorrermos ao desenvolvimento de novas features, com mais informações de outros datasets que possam acrescentar informação e trazer mais variáveis relevantes para a modelação.

Webgrafia

[Top 8 in-demand cybersecurity jobs for 2022 and beyond](#) (consultado a 19/10/2022)

[Eight Cybersecurity Skills in Highest Demand | Harvard Extension School](#) (consultado a 24/10/2022)

[Vodafone Portugal alvo de ciberataque](#) (consultado a 24/10/2022)

[Is Data Scientist Still the Sexiest Job of the 21st Century?](#) (consultado a 24/10/2022)

[Ronald Van Loon Discusses the Future of Data Science: Career Outlook for 2020](#) (consultado a 19/10/2022)

[The Data Science Talent Gap: Why It Exists And What Businesses Can Do About It](#) (consultado a 19/10/2022)

[How Data Science Drives Growth and Improves Customer Experiences in FinServ and Retail](#) (consultado a 19/10/2022)

[Data science jobs are a top pick for Gen Z, per new Glassdoor report | Fortune](#) (consultado a 19/10/2022)

[Tendências em cibersegurança para 2022](#) (consultado a 19/10/2022)

[6 grandes tendências em cibersegurança apontadas por especialistas | Ciber Segurança | Valor Econômico](#) (consultado a 21/10/2022)

[6 Data Science Trends That Can Shape 2022 | by Rashi Desai](#) (consultado a 21/10/2022)

[Cybersecurity trends: Looking over the horizon | McKinsey](#) (consultado a 23/10/2022)

[World Economic Outlook, October 2022: Countering the Cost-of-Living Crisis](#) (consultado a 24/10/2022)

[Policymakers Need Steady Hand as Storm Clouds Gather Over Global Economy](#) (consultado a 24/10/2022)