Escola de Engenharia
Universidade do Minho

Mestrado Integrado em Engenharia de
Gestão e Sistemas de Informação
2023/2024

4º Ano

1º Semestre

Aprendizagem Automática em Sistemas de Informação

Francisco Miguel Pinheiro Cardoso

A79570

Índice

1. Introdução

O presente relatório realiza-se no âmbito da unidade curricular "Aprendizagem Automática em Sistemas Empresariais", lecionada no 1º ano do Mestrado em Engenharia e Gestão de Sistemas de Informação, tendo como intuito a aplicação da metodologia CRISP-DM "Cross Industry Standard Process for Data Mining" para compreensão dos conceitos, princípios e recursos associadas ao Data Mining.

A partir da base de dados facultada pelo docente - Used Car Price Prediction Dataset -, a qual têm informação relacionada com a indústria automóvel (tendências, preferências do consumidor, etc.), procurei encontrar soluções que permitem aumentar o volume de negócio.

De acordo com a metodologia CRISP-DM, a análise a ser efetuada divide-se em 6 fases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evalution e Deployment.

O presente relatório apenas abrange o Business Understanding, centrado em entender os objetivos, requisitos e critérios de sucesso do negócio, a fim de desenvolver um plano inicial que permita satisfazer os clientes. Assim sendo, subdividi o relatório em quatro subseções. Na primeira subseção, enquadra-se os objetivos do negócio e refere-se qual o critério de sucesso. Na segunda subseção, expõe-se quais os recursos disponíveis para a elaboração do presente projeto, os requisitos que têm de ser cumpridos e os pressupostos que permitem validar os resultados e as restrições que podem ocorrer no desenvolvimento do projeto. Na terceira subseção, apresenta-se os objetivos de data mining e o critério de sucesso do mesmo. Por último, clarifica-se o plano de projeto e quais as ferramentas a serem utilizadas para o seu desenvolvimento.

2. Business Understanding

A Compreensão do Negócio é a fase inicial da metodologia CRISP-DM, que se foca em compreender os objetivos do projeto e requisitos de uma perspetiva de negócio, para mais tarde converter este conhecimento numa definição de um problema de data mining e um plano preliminar desenhado para alcançar os objetivos.

2.1 Determine Business Objetives

2.1.1 Background

Cars.com é líder no mercado digital e fornecedor de soluções para a indústria automóvel que interliga compradores e vendedores de automóveis. A empresa fornece aos compradores dados, recursos e ferramentas digitais necessárias para tomarem decisões de compras informadas e estabelecerem uma ligação com os retalhistas. Num mercado em constante mudança, a Cars.com dispõe de soluções técnicas inovadoras e de informações baseadas em dados para conseguirem alcançar e influenciar as pessoas prontas a comprar, aumentarem a circulação de inventário e ganharem quota de mercado.

Em 2018, a Cars.com adquiriu Dealer Inspire, uma empresa de tecnologia inovadora que desenvolve soluções que preparam os concessionários para o futuro com operações mais eficientes, um processo de compra de automóveis mais rápido e fácil, assim como experiências digitais conectadas que vendem e fazem manutenção de mais veículos.

Cars.com inventou a pesquisa de automóveis. O seu website e as soluções inovadoras fazem o contacto entre o comprador e o vendedor. A empresa conta com colaboradores espalhados pelos Estados Unidos da América. Ao fim de muitos anos, continuam com uma cultura de start-up com inovação e paixão pelos colaboradores no centro do negócio.

A Cars.com é uma marca premiada, uma equipa de liderança que conta com os melhores e mais brilhantes funcionários da indústria. Foram considerados um dos melhores locais para trabalhar peço The Chicago Tribune, Built in Chicago e Chicago Innovation.

2.1.2 Business Objetives

De forma a atingir a solução pretendida, é necessário definir objetivos de negócio, para tal, defini os seguintes:

- Melhorar a experiência no website, tanto do vendedor como do comprador;
- Expandir os serviços oferecidos relacionados com o negócio automóvel;
- Apostar na sustentabilidade e promover o negócio com foco nos veículos elétricos;
- Aumentar a base de clientes, utilizar técnicas para adquirir novos clientes e manter os existentes.

2.1.3 Business Success Criteria

Com o intuito de aumentar os resultados da empresa na perspetiva do negócio, é necessário definir critérios de sucesso. Com os dados fornecidos foi possível definir um critério de sucesso:

- Aumentar as vendas dos carros mais antigos no website.

Para tal, utilizarei os dados fornecidos para saber quais os carros com melhores condições, melhores preços, para que possam ser mais facilmente promovidos no website.

2.2 Assess Situation

2.2.1 Inventory of Resources

Os recursos disponíveis para a realização deste projeto, incluem:

- Pessoal: A realização do projeto dispõe de um aluno, do Mestrado Integrado em Engenharia de Gestão de Sistemas de Informação, com alguma experiência em data mining, análise de negócio e dados.
- Dados: Os dados são fornecidos no website kaggle, onde contêm as características sobre cada viatura disponível no website, num ficheiro em formato .csv.
- Hardware: O aluno possui dois computadores para a análise e tratamento de dados.
- Software: O aluno tem disponível para utilizar neste projeto, ferramentas como Jupyter para programação em Python e RapidMiner para a realização de datamining. Para o processamento de dados tem disponível o Talend, para a modelação dos modelos e dashboards o Tableau e por fim, o Microsoft Word e Excel para documentação e arquivo de dados respetivamente.

2.2.2 Requirements, Assumptions, and Constrains

Neste projeto, existem requisitos que têm de ser cumpridos, assim como pressupostos que permitem validar os resultados e por fim, restrições que podem ocorrer no desenvolvimento do projeto.

Restrições:

- Realizar as entregas nos prazos estipulados;
- Apresentar resultados compreensíveis e com qualidade;
- Utilizar a metodologia CRISP-DM;
- Utilizar as devidas ferramentas projetadas.

Pressupostos:

- Os datasets têm de apresentar dados reais;
- Os dados não podem apresentar erros;
- Os datasets têm de ser suficientes para responder aos requisitos do projeto.

Restrições:

- Pouca experiência em Data Mining;
- Sobrecarga de trabalho de grupo, sendo um só aluno a realizar;
- Pouca experiência na metodologia CRIPS-DM;
- Pouca experiência nas ferramentas a ser utilizadas.

2.2.3 Risks and Contingencies

Riscos	**Consequência**	**Impacto**	**Contingência**
Inexperiência na utilização das ferramentas	Fraca evolução no desenvolvimento do projeto	4	Pesquisa e prática na utilização das ferramentas; solicitar apoio ao docente
Elevada sobrecarga de trabalho	Fraca demonstração de resultados	4	Boa gestão e organização de tempo
Inexperiência da metodologia CRISP-DM	Incorreto desenvolvimento do projeto	4	Revisão constante da metodologia e solicitar apoio ao docente

2.2.4 Terminology

O projeto dispõe de terminologias relevantes compostas por duas componentes, entre elas de negócio e data mining.

- CRISP-DM: "Cross Industry Standard Process for Data Mining" é uma metodologia utilizada para estruturar projetos de data mining. Fornece uma estrutura abrangente para planear, implementar e avaliar o processo de data mining, constituído por seis fases.
- Data Mining: Processo de descobrir informações, padrões e conhecimentos em grandes conjuntos de dados. Envolve a utilização de técnicas computacionais para analisar dados e extrair informação significativa. Utilizada para tomadas de decisões, identificação de tendências, previsões e otimizações.
- DataSet: Conjunto de dados estruturados por colunas e linhas, onde cada coluna representa uma variável e cada linha corresponde a um determinado conjunto de dados.

2.2.5 Costs and Benefits

Este projeto não dispõe de custos, pois encontra-se associado a uma unidade curricular de nível académico. A nível de benefícios, encontra-se o aproveitamento da utilização da metodologia CRIPS-DM, assim como a utilização de ferramentas tecnológicas.

Caso fosse necessário, avaliar os custos no caso de projeto ser real, seriam identificados custos de recursos humanos, custos de software e hardware, assim como infraestruturas necessárias, entre outros. Já os benefícios seriam aplicados à empresa, já que a mesma disponibilizaria do output de projeto, numa maior transação de vendas associadas ao website.

2.3 Determine Data Mining Goals

2.3.1 Data Mining Goals

A fim de determinar os objetivos de data mining, é preciso realizar a análise e tratamento dos dados, entender e compreender de que forma é possível explorar as características de cada veículo de forma que os entusiastas, compradores e investigadores interessados em análises, tomem decisões de compras informadas a nível da indústria automóvel e preferências do consumidor.

2.3.2 Data Mining Success Criteria

Para definir o critério de sucesso data mining, defini um valor mínimo (valor desejável) para um critério baseado na fórmula que irá ser utilizada para a avaliação conhecida como Root Mean Squared Error (RMSE). Esta fórmula representa uma medida que calcula a raiz quadrática média dos erros entre valores e reais e possíveis.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

O critério definido é:

- Valor de RMSE de < 0.2, na atribuição dos preços desejados para os veículos.

2.4 Produce Project Plan

1.4.1 Project Plan

Business Understanding

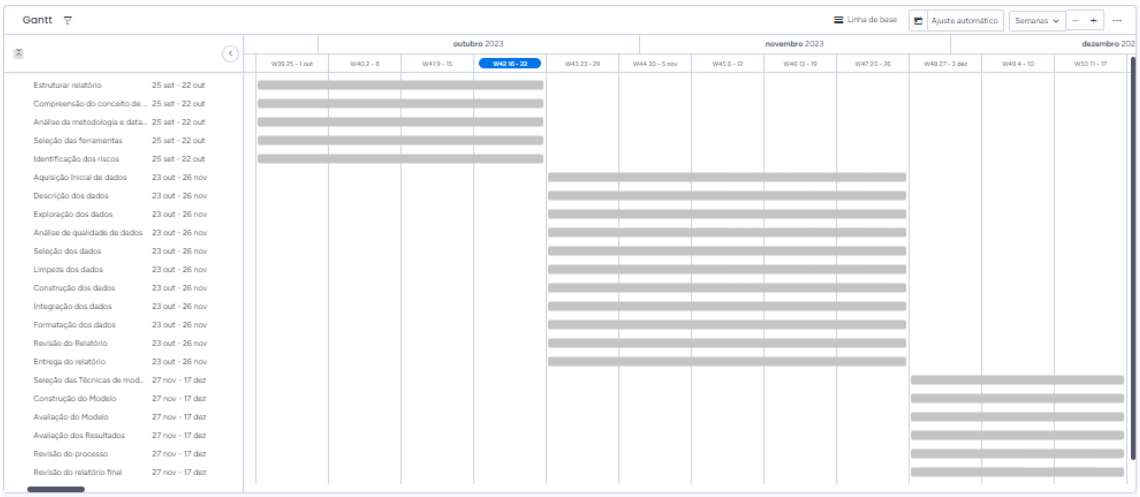
Tarefa	Data	Responsável
Estruturar relatório	25 set - 22 out	Francisco Cardoso
Compreensão do conceit...	25 set - 22 out	Francisco Cardoso
Análise da metodologia e...	25 set - 22 out	Francisco Cardoso
Seleção das ferramentas	25 set - 22 out	Francisco Cardoso
Identificação dos riscos	25 set - 22 out	Francisco Cardoso
Revisão do relatório	25 set - 22 out	Francisco Cardoso
Entrega do relatório	25 set - 22 out	Francisco Cardoso

Data Understanding + Data Preparation

Tarefa	Data	Responsável
Aquisição Inicial de dados	23 out - 26 nov	Francisco Cardoso
Descrição dos dados	23 out - 26 nov	Francisco Cardoso
Exploração dos dados	23 out - 26 nov	Francisco Cardoso
Análise de qualidade de d...	23 out - 26 nov	Francisco Cardoso
Seleção dos dados	23 out - 26 nov	Francisco Cardoso
Limpeza dos dados	23 out - 26 nov	Francisco Cardoso
Construção dos dados	23 out - 26 nov	Francisco Cardoso
Integração dos dados	23 out - 26 nov	Francisco Cardoso
Formatação dos dados	23 out - 26 nov	Francisco Cardoso
Revisão do Relatório	23 out - 26 nov	Francisco Cardoso
Entrega do relatório	23 out - 26 nov	Francisco Cardoso

Modeling + Evaluation

	Tarefa	Data	Responsável
☐	Seleção das Técnicas de ...	27 nov - 17 dez	Francisco Cardoso
☐	Construção do Modelo	27 nov - 17 dez	Francisco Cardoso
☐	Avaliação do Modelo	27 nov - 17 dez	Francisco Cardoso
☐	Avaliação dos Resultados	27 nov - 17 dez	Francisco Cardoso
☐	Revisão do processo	27 nov - 17 dez	Francisco Cardoso
☐	Revisão do relatório final	27 nov - 17 dez	Francisco Cardoso
☐	Entrega do relatório final	27 nov - 17 dez	Francisco Cardoso



1.4.2 Initial Assessment of Tools and Techniques

Neste projeto, irei utilizar diferentes ferramentas para o seu desenvolvimento.
Na tabela a seguir, estão representadas as mesmas consoante as suas funcionalidades.

Ferramenta	Funcionalidade
Microsoft Word	Elaboração dos relatórios.
Microsoft Excel	Visualização dos dados utilizados no projeto
Talend	Análise e tratamento de dados.
Rapid Miner	Criar modelos de dados para os requisitos de negócio.
Jupyter Notebook	Executar scripts em Python, para manipulação de dados.
Tableau	Visualização de modelos e dashboards.

3. Conclusão

Durante a primeira etapa de Business Understanding baseada na metodologia CRISP-DM, concluo o sucesso não só da sua realização, como da importância na idealização de todo o negócio para a dinâmica e estudo da unidade curricular. Foi possível identificar os objetivos do negócio e de Data Mining, apesar de algumas dificuldades. Nos critérios de sucesso de data mining tive dificuldade na definição dos mesmos, assim como na produção do plano de projeto, pois recorri a uma alternativa ao MS Project para a realização do Diagrama de Gantt.

Nas próximas etapas, irei basear-me nos objetivos definidos para ajudar a Cars.com a definir com maior precisão as tendências, preferências, entre outros aspetos do consumidor para aumentar o volume de negócio da empresa.