RUGSTAND

BUSINESS UNDERSTANDING

Projeto 1 - 1ª Entrega

Aprendizagem Automática em Sistemas Empresariais

MEGSI-ESI
2023/2024

Membros da Equipa

Luís Miguel Bernardes André Silva

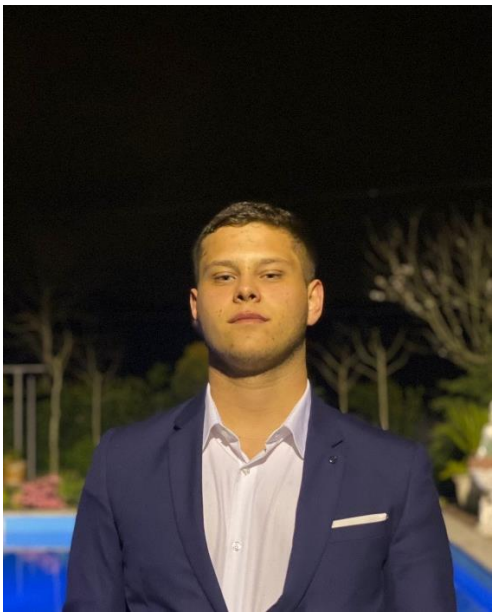
PG54014

Ricardo Barbosa Gonçalves Pereira

PG54178

Marco Meira Pires

PG54032



Miguel Malheiro Ferreira

PG54102

Índice

1. Introdução

Neste trabalho proposto pelos docentes da Unidade Curricular de Aprendizagem Automática de Sistemas Empresariais, iremos estudar um dataset sobre o preço de carros usados e com ferramentas de datamining iremos prever se é viável ao nosso cliente adquirir o veículo para revenda.

Na era da informação, os dados tornaram-se muito valiosos, especialmente em setores como o de revenda de automóveis. Este trabalho explora a aplicação do processo de data mining, seguindo a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining), para otimizar as operações de um stand de revenda de carros. A crescente concorrência e a necessidade de compreender melhor o mercado automóvel fazem da mineração de dados uma ferramenta crucial para a tomada de decisões informadas, previsões de procura, identificação de tendências de mercado e melhorias na gestão de stock. Neste estudo, abordaremos as várias etapas do CRISP-DM, adaptando-as ao contexto da revenda de carros, com o objetivo de demonstrar como a mineração de dados pode impulsionar o sucesso deste negócio num ambiente altamente competitivo.

2. Compreensão do Negócio

2.1. Objetivos do Negócio

Background

A informação obtida acerca da situação de negócio da organização engloba um mercado digital que se baseia na compra e venda de veículos, anúncios de carros novos e usados, comparação de veículos e informações detalhadas e conexões com concessionárias e vendedores particulares.

No que toca ao dataset, a informação que nos é dada contém características sobre os carros, como marca e modelo, ano do modelo, quilometragem, tipo de combustível, tipo de motor, transmissão, cor do interior e exterior, histórico de acidentes, título limpo e preço.

Objetivos de Negócio

O objetivo de negócio deste projeto, é desenvolver um sistema de previsão de compra que avalie se a aquisição de um carro para revenda é uma decisão lucrativa. O sistema deve ajudar a identificar carros que têm maior probabilidade de gerar lucro após a compra, otimizando assim as decisões de aquisição de veículos usados para revenda.

Critérios de Sucesso de Negócio

O sucesso do nosso projeto pode ser determinado através da definição de diversos motivos, tais como:

Margem de lucro aumentada: O aumento na margem de lucro é fundamental, pois é um dos principais indicadores de sucesso financeiro do negócio.

Redução de prejuízos: Evitar prejuízos financeiros é essencial, uma vez que a compra de carros que resultam em perdas pode impactar negativamente os resultados.

Taxa de revenda bem-sucedida: O sucesso na revenda dos carros adquiridos com base nas previsões é um indicador direto de eficácia no mercado.

Eficiência operacional: Melhorar a eficiência operacional pode levar a economizar tempo e recursos, o que é essencial para o sucesso a longo prazo.

Taxa de retorno do investimento: Avaliar e determinar se os benefícios financeiros superam os custos do projeto, o que é fundamental para justificar os investimentos.

2.2. Avaliação da Situação

A Avaliação da Situação é uma atividade fundamental presente no CRIPS-DM que envolve analisar o ambiente atual do negócio e compreender o contexto no qual o projeto de mineração de dados será realizado. É onde detalhamos os recursos, requisitos, restrições, riscos, terminologia, custos e benefícios, que influenciam o projeto.

Inventário de Recursos

Neste ponto do CRISP-DM vamos enumerar os recursos disponíveis para o desenvolvimento do projeto de análise de dados.

Pessoas	A nossa equipa é constituída por 4 pessoas do 1º de MEGSI-ESI, que com o suporte do docente irá desenvolver este projeto.
Dados	A nossa equipa trabalha com os dados fornecidos pelo enunciado, disponíveis na plataforma Kaggle em formato CSV sobre carros usados
Hardware	Temos disponíveis 4 máquinas para o desenvolvimento do projeto e processamento de dados.
Software	O software utilizado pela nossa equipa é o Microsoft Word, o Microsoft Excel

Requisitos, suposições e restrições

Requisitos

A equipa deve ter 4 elementos do TP1 de AASE;
Cumprir a deadline é a 16/12/2023;
A equipa deve ter acesso ao Kaggle;
Manter o trabalho utilizado;
Utilizar a metodologia adotada nas aulas (CRISP-DM);
Utilizar ferramentas de Data Mining;
Utilizar os datasets fornecidos.

Pressupostos

Não vai existir fabricação de dados pois os fornecidos são suficientes para a realização do projeto;
Os datasets estão relacionados com o negócio da empresa;
A equipa docente apoia a nossa equipa durante a realização deste processo.

Restrições

Pouca experiência em ferramentas Datamining;
Pouca experiência em ferramentas de visualização de dados;
Pouca experiência em metodologia CRISP-DM;
Deadline do projeto;
Complexidade na integração de dados.

Riscos e Contingências

Riscos	Contingências
Dificuldade na divisão justa de tarefas	Entreajuda onde os colegas com menos tarefas ajudam os colegas com mais tarefas. Reatribuir tarefas. Registar o número de horas utilizado por cada membro por semana e reajustar na semana seguinte.
Comunicação entre a equipa	Criar meios de comunicação, como chats e reuniões semanais, onde os colegas possam atualizar o estado das tarefas e possíveis dificuldades nas tarefas atribuídas.
Mau tratamento dos dados fornecidos	Analisar os dados fornecidos em equipa com espírito crítico e atenção, de modo que o tratamento de dados seja bem-sucedido e os dados fornecidos não tenham erros.
Dificuldade na interpretação dos dados	O negócio deve ser bem percebido pela equipa de modo que os dados sejam mais facilmente interpretados pela nossa equipa.
Dificuldade na compreensão e utilização da metodologia CRISP-DM	Os membros da equipa devem utilizar o material fornecido pelo docente para perceber a metodologia. Caso não o consigam fazer devem pesquisar informação online. Em último caso os alunos devem consultar o docente para esclarecer possíveis dúvidas.
Dificuldade na utilização das ferramentas	Os membros da equipa devem pesquisar na internet tutoriais e documentação das ferramentas utilizadas no trabalho. Se a dificuldade persistir devemos consultar o nosso docente para solicitar ajuda. Em último caso devemos mudar de ferramenta.

Terminologia

Na tabela seguinte temos os termos mais específicos usados durante o relatório de modo que os stakeholders obtenham contexto sobre os estes termos.

CRISP-DM	Cross Industry Standard Process for Data Mining. Metodologia criada em 1997 para guiar projetos em data mining.
Data Mining	É o processo de analisar grandes conjuntos de dados em busca de padrões e informações relevantes. É essencial em áreas como negócios, ciência e saúde, proporcionando insights valiosos para a tomada de decisões informadas e a descoberta de conhecimento oculto nos dados.
Dataset	Um dataset é uma coleção organizada de dados, geralmente em formato tabular, utilizado para análises, estudos ou treinamento de modelos de machine learning. Cada linha representa uma observação e cada coluna, uma variável.

Custos e Benefícios

Os custos associados a este projeto serão nulos, uma vez que, será realizado no decorrer da Unidade Curricular de Aprendizagem Automática em Sistemas Empresariais. No entanto num caso real, os custos no nosso projeto estariam associados a obtenção de ferramentas licenciadas e respetivas formações, obtenção e tratamento de dados, custos de implementação e pagamento dos profissionais.

Os benefícios associados à elaboração deste projeto, é desenvolvimento das nossas capacidades com ferramentas e metodologias de data mining, habilidades em trabalho de grupo. Ao nível do negócio os benefícios serão o aumento do lucro do negócio, através de um algoritmo que auxilia na escolha de veículos para revenda no stand de automóveis do cliente.

2.3 Determinar metas de Data Mining

Objetivos de Data Mining

A definição dos objetivos da análise de dados é um passo crucial para avaliar a viabilidade do negócio de uma concessionária. Para isso, é necessário identificar padrões ou tendências nos dados disponíveis através de técnicas de extração de dados. O tipo de problema de Data Mining abordado é uma regressão, pois estamos a lidar com valores numéricos. O objetivo passa por criar um modelo que possa fazer previsões precisas com base nos dados disponíveis de modo a perceber se o negócio é vantajoso ou não para a concessionária.

Critérios de sucesso de Data Mining

Métrica	Justificação	Fórmulas	Valores
Raiz do Erro Médio Quadrático (RMSE)	É a métrica mais comum para problemas de regressão. Ele mede a média dos quadrados dos erros entre os valores previstos e os valores reais. Quanto menor o valor do MSE, melhor o desempenho do modelo.	$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$	<=10000
Erro Médio Absoluto (MAE)	É a média dos valores absolutos dos erros entre os valores previstos e os valores reais.	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - p_i \rvert$	<=5000
Erro Percentual Médio (MAPE)	É útil quando você deseja avaliar o erro em termos percentuais. Ele calcula o erro médio como uma percentagem dos valores reais.	$MAPE = \frac{1}{n}\sum_{i=1}^{n} \lvert \frac{y_i - p_i}{y_i} \rvert$	=<25%
Coeficiente de Determinação (R^2)	Mede a proporção da variabilidade dos dados que é explicada pelo modelo.	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i \quad \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	>=0.8

2.4 Produzir o Plano de Projeto

Planeamento de Projeto

Como iremos realizar este projeto com base na metodologia CRIPS-DM identificamos todas as tarefas da metodologia assim como a estimativa do tempo necessário para realizar cada tarefa, e os elementos do grupo que a têm de realizar.

Dito isto, abaixo encontra-se o planeamento realizado.

Projeto 1	84 dias	**20-09-2023 8:00**	**15-01-2024 17:00**	
Análise da metodoligia CRISP-DM	0,25 dias	21-09-2023 8:00	21-09-2023 10:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Leitura e análise do enunciado	0,25 dias	21-09-2023 8:00	21-09-2023 10:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Estudo do Negócio	19 dias	**20-09-2023 8:00**	**16-10-2023 17:00**	
Reunião de Equipa	1 dia	20-09-2023 8:00	20-09-2023 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Determinar Objetivos do Negócio	4 dias	**21-09-2023 8:00**	**26-09-2023 17:00**	
Recolha de Informação	1 dia	21-09-2023 8:00	21-09-2023 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Descrever Objetivos de Negócio	2 dias	22-09-2023 8:00	25-09-2023 17:00	Marco Pires;Miguel Ferreira
Criterios de Sucesso do Negócio	1 dia	26-09-2023 8:00	26-09-2023 17:00	Ricardo Pereira;Luís Silva
Descrição Situação	7 dias	**27-09-2023 8:00**	**05-10-2023 17:00**	
Descrição dos Recursos Disponíveis	1 dia	27-09-2023 8:00	27-09-2023 17:00	Marco Pires;Luís Silva
Requisitos, Pressupostos e Restrições	1 dia	28-09-2023 8:00	28-09-2023 17:00	Miguel Ferreira;Ricardo Pereira
Riscos e Contigências	1 dia	29-09-2023 8:00	29-09-2023 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira
Terminologia	2 dias	02-10-2023 8:00	03-10-2023 17:00	Ricardo Pereira
Custos e Benefícios	2 dias	04-10-2023 8:00	05-10-2023 17:00	Marco Pires;Miguel Ferreira
Determinar Objetivos do Data Mining	4 dias	**06-10-2023 8:00**	**11-10-2023 17:00**	
Objetivos do Data Mining	2 dias	06-10-2023 8:00	09-10-2023 17:00	Miguel Ferreira;Ricardo Pereira
Criterios de Sucesso do Data Mining	2 dias	10-10-2023 8:00	11-10-2023 17:00	Marco Pires;Luís Silva
Produzir o Plano de Projeto	3 dias	**12-10-2023 8:00**	**16-10-2023 17:00**	
Plano do Projeto	2 dias	12-10-2023 8:00	13-10-2023 17:00	Luís Silva;Ricardo Pereira
Pressupostos Iniciais das Ferramentas e Técnicas	1 dia	16-10-2023 8:00	16-10-2023 17:00	Marco Pires;Miguel Ferreira
Compreensão dos Dados	10 dias	**17-10-2023 8:00**	**30-10-2023 17:00**	
Reunião de Equipa	3 dias	17-10-2023 8:00	19-10-2023 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Recolha Inicial dos Dados	3 dias	**20-10-2023 8:00**	**24-10-2023 17:00**	
Relatório da Recolha Inicial dos Dados	3 dias	20-10-2023 8:00	24-10-2023 17:00	Marco Pires;Ricardo Pereira
Descrever os Dados	1 dia	**25-10-2023 8:00**	**25-10-2023 17:00**	
Relatório com a Descrição dos Dados	1 dia	25-10-2023 8:00	25-10-2023 17:00	Miguel Ferreira;Luís Silva
Explorar os Dados	1 dia	**26-10-2023 8:00**	**26-10-2023 17:00**	
Relatório da Exploração dos Dados	1 dia	26-10-2023 8:00	26-10-2023 17:00	Marco Pires;Ricardo Pereira
Verificar a Qualidade dos Dados	2 dias	**27-10-2023 8:00**	**30-10-2023 17:00**	
Relatório da Qualidade dos Dados	2 dias	27-10-2023 8:00	30-10-2023 17:00	Miguel Ferreira;Luís Silva
Preparação de Dados	14,5 dias	**31-10-2023 8:00**	**20-11-2023 13:00**	
Reunião de Equipa	1 dia	31-10-2023 8:00	31-10-2023 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Selecionar os Dados	3 dias	**01-11-2023 8:00**	**03-11-2023 17:00**	
Relarório dos Dados Selecionados e Eliminados	3 dias	01-11-2023 8:00	03-11-2023 17:00	
Limpeza de Dados	0,5 dias	**06-11-2023 8:00**	**06-11-2023 13:00**	
Relatório da Limpeza de Dados	0,5 dias	06-11-2023 8:00	06-11-2023 13:00	Marco Pires;Luís Silva
Construção dos Dados	4 dias	**06-11-2023 13:00**	**10-11-2023 13:00**	
Derivar novos Atrivutos	3 dias	06-11-2023 13:00	09-11-2023 13:00	Ricardo Pereira;Luís Silva
Gerar novos Registos	1 dia	09-11-2023 13:00	10-11-2023 13:00	Marco Pires;Miguel Ferreira
Integração dos Dados	3 dias	**10-11-2023 13:00**	**15-11-2023 13:00**	
Juntar dados	3 dias	10-11-2023 13:00	15-11-2023 13:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Formatação dos Dados	3 dias	**15-11-2023 13:00**	**20-11-2023 13:00**	
Dados Reformatados	3 dias	15-11-2023 13:00	20-11-2023 13:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Modelação	9 dias	**20-11-2023 13:00**	**01-12-2023 13:00**	
Reunião de Equipa	2 dias	20-11-2023 13:00	22-11-2023 13:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Seleção de Técnicas de Modelação	1 dia	**22-11-2023 13:00**	**23-11-2023 13:00**	
Técnicas de Modelação	0,5 dias	22-11-2023 13:00	22-11-2023 17:00	Marco Pires;Ricardo Pereira
Pressupostos de Modelação	0,5 dias	23-11-2023 8:00	23-11-2023 13:00	Miguel Ferreira;Luís Silva
Criar uma concepção de Teste	1 dia	**23-11-2023 13:00**	**24-11-2023 13:00**	
Concepção de Teste	1 dia	23-11-2023 13:00	24-11-2023 13:00	
Construir o Modelo	3 dias	**24-11-2023 13:00**	**29-11-2023 13:00**	
Ajuste de Parametros	1 dia	24-11-2023 13:00	27-11-2023 13:00	Marco Pires
Modelo Gerado	1 dia	27-11-2023 13:00	28-11-2023 13:00	Miguel Ferreira
Descrição do Modelo	1 dia	28-11-2023 13:00	29-11-2023 13:00	Ricardo Pereira;Luís Silva
Rever o Modelo	2 dias	**29-11-2023 13:00**	**01-12-2023 13:00**	
Revisão do Modelo	1 dia	29-11-2023 13:00	30-11-2023 13:00	Marco Pires;Miguel Ferreira
Revisão dos Parametros Usados	1 dia	30-11-2023 13:00	01-12-2023 13:00	Ricardo Pereira;Luís Silva
Avaliação	6,5 dias	**01-12-2023 13:00**	**11-12-2023 17:00**	
Reunião de Equipa	2 dias	01-12-2023 13:00	05-12-2023 13:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Avaliar os Resultados	2 dias	**05-12-2023 13:00**	**07-12-2023 13:00**	
Validação dos Objetivos de Data Mining	1 dia	05-12-2023 13:00	06-12-2023 13:00	Marco Pires;Miguel Ferreira
Aprovação do Modelo	1 dia	06-12-2023 13:00	07-12-2023 13:00	Ricardo Pereira;Luís Silva
Rever o Processo	1 dia	**07-12-2023 13:00**	**08-12-2023 13:00**	
Revisão do Processo	1 dia	07-12-2023 13:00	08-12-2023 13:00	Ricardo Pereira
Determinar os Proximos Passos	1,5 dias	**08-12-2023 13:00**	**11-12-2023 17:00**	
Lista das Possíveis Ações	1 dia	08-12-2023 13:00	11-12-2023 13:00	Miguel Ferreira;Ricardo Pereira
Decisões	0,5 dias	11-12-2023 13:00	11-12-2023 17:00	Luís Silva;Marco Pires

Implementação	24 dias	**12-12-2023 8:00**	**12-01-2024 17:00**	
Reunião de Equipa	8 dias	12-12-2023 8:00	21-12-2023 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Planear Avaliação de Resultados	7 dias	22-12-2023 8:00	01-01-2024 17:00	Ricardo Pereira;Luís Silva
Resultados da Avaliação	1 dia	02-01-2024 8:00	02-01-2024 17:00	Marco Pires;Miguel Ferreira
Planear, Monotorizar e Manutenção	1 dia	03-01-2024 8:00	03-01-2024 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Produzir o Relatório Final	6 dias	**04-01-2024 8:00**	**11-01-2024 17:00**	
Relatório Final	3 dias	04-01-2024 8:00	08-01-2024 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Apresentação Final	3 dias	09-01-2024 8:00	11-01-2024 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Rever o Projeto	1 dia	**12-01-2024 8:00**	**12-01-2024 17:00**	
Documentação do Estudo	1 dia	12-01-2024 8:00	12-01-2024 17:00	Marco Pires;Miguel Ferreira;Ricardo Pereira;Luís Silva
Entrega Final	1 dia	15-01-2024 8:00	15-01-2024 17:00	

Avaliação de Ferramentas e Técnicas

De seguida apresentamos uma tabela com as ferramentas que utilizamos para alcançar os objetivos deste projeto.

Ícones	Nome	Funcionalidades
	Microsoft Office Word	Plataforma utilizada para a realização de todos os artefactos utilizados neste documento.
	Microsoft Office Excel	Plataforma que auxilia na visualização dos dados.
	Discord	Utilizado para a realização de reuniões e para compartilhar documentos importantes para a realização deste projeto.
	WhatsApp	Plataforma utilizada para a interação dos elementos do grupo.
	ProjectLibre	Plataforma utilizada para realizar o planeamento do projeto
	Tableau	Plataforma utilizada para analisar os dados com mais detalhe.
	Python	Plataforma utilizada para desenvolver o processo de análise de qualidade de dados, de limpeza dos dados e desenvolvimento dos modelos de Data Mining.

3. Compreensão dos Dados

3.1. Relatório de Aquisição Inicial dos Dados

Na segunda fase do nosso projeto, realizamos a recolha, descrição e avaliação dos dados. Este processo não só visa reunir informações pertinentes, mas também assegurar que esses dados estejam alinhados com os objetivos definidos para o projeto. Posteriormente, será analisada a qualidade dos dados, garantindo que estejam aptos a sustentar análises precisas e conclusões fundamentadas.

Background dos Dados

Relativamente às fontes dos dados, foram-nos disponibilizados dois datasets que se encontram na pasta 2324aaesetp1. Os dois datasets incluem 12 atributos, sendo que 11 destes são comuns a ambos. Esses atributos englobam informações como a marca, o modelo, o ano do modelo, o número de quilómetros percorridos, o tipo de combustível, o tipo de motor, a transmissão, a cor exterior, a cor interior, se o veículo já esteve envolvido em algum acidente e o clean title.

Pasta	Dataset
2324aaesetp1	test
	train

O dataset "test " tem 802 linhas e 12 colunas, incluindo os atributos mencionados anteriormente, além de um identificador (ID). Já o dataset "train" possui 3207 linhas e 12 colunas, com os mesmos atributos mencionados anteriormente, mas incluindo também o preço dos veículos.

Após uma análise detalhada de ambos os conjuntos de dados, a equipa optou por utilizar o dataset "Data_Train". Esta escolha fundamenta-se na consideração de que este é o único dataset que fornece informações suficientes para atingir os objetivos do projeto, uma vez que, além de conter outros atributos relevantes, também inclui o preço dos veículos. Esta característica adicional no dataset é crucial para as análises e conclusões desejadas no contexto do projeto.

3.2. Relatório de Descrição dos Dados

Descrição dos Dados

O dataset utilizado para desenvolver este projeto foi o “Used Car Price Prediction Dataset” através do qual foi desenvolvida a seguinte tabela que contém informações sobre os atributos do dataset, assim como a sua descrição, formato, tipo de atributo e exemplo.

Atributo	Descrição	Formato	Tipo de atributo	Exemplo
Marca e modelo	Identificação da marca e do modelo específico de cada veículo.	String	Categórico	“Mazda” “Dodge” “E-Class E 350” “GX 460 Base”
Ano do modelo	Informação sobre o ano de produção dos veículos.	Int	Numérico	“2014” “2015” “2020”
Quilometragem	Indicação da quilometragem de cada veículo.	Float	Numérico	“103,726 mi.” “3,158 mi.”
Tipo de combustível	Especificação do tipo de combustível utilizado.	String	Categórico	“Gasoline” “Diesel” “Hybrid”
Tipo de motor	Detalhes sobre o motor.	String	Categórico	“3.0 Liter Turbo” “5.0L V8 32V PDI DOHC”
Transmissão	Informação sobre o tipo de transmissão.	String	Categórico	“Automatic” “4-Speed A/T”
Cores exteriores e interiores	Descrição das opções de cores do exterior e interior do veículo.	String	Categórico	“Red” “Gray” “Silver” “Black”

Histórico de acidentes	Informação sobre o histórico de acidentes ou danos anteriores.	Boolean	Categórico	“At least 1 accident or damage reported” “None reported”
Título limpo	Avaliação da disponibilidade de um título limpo.	Boolean	Categórico	“Yes”
Preço	Valor de cada veículo.	Int	Numérico	“15000” “45950”

3.3. Relatório de Exploração

Análises Gerais

Vírgulas, Pontos e Travessões

Na coluna “milage” a nossa equipa apercebeu-se do valor “71,000 mi.”, que possui um “,” e “mi.” e que esta tendência se repetia em todos os valores da coluna. Por ser desnecessário e dificultar a análise de dados pois não podemos representar esta coluna por Integer a nossa equipa decidiu evidenciar este facto.

```
for index,text in enumerate(df['milage'][69:74]):
    print('Carro %d:\n'%(index+1),text)
```

✓ 0.0s

```
Carro 1:
 50,648 mi.
Carro 2:
 2,000 mi.
Carro 3:
 24,280 mi.
Carro 4:
 179,700 mi.
Carro 5:
 87,500 mi.
```

Em relação à identificação de células contendo apenas travessões, a nossa equipe executou um comando específico. Os resultados revelam que nas colunas 'fuel_type', 'engine', 'transmission', 'ext_col' e 'int_col', observamos a presença exclusiva de travessões.

```
def contar_travessoes_na_coluna(coluna):
    return coluna.apply(lambda x: isinstance(x, str) and x.strip() == '-').sum()


contagem_por_coluna = {}


for coluna in df.columns:
    contagem_por_coluna[coluna] = contar_travessoes_na_coluna(df[coluna])

# Exiba os resultados
print("Contagem de células com apenas travessões em cada coluna:")
print(contagem_por_coluna)
```

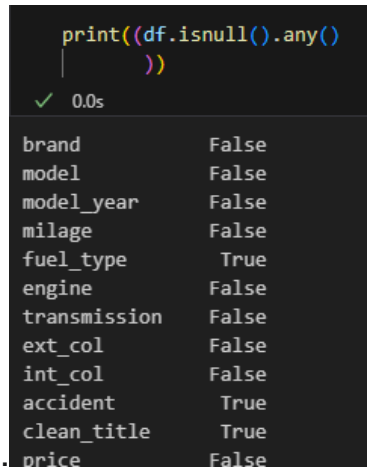
✓ 0.0s

```
Contagem de células com apenas travessões em cada coluna:
{'brand': 0, 'model': 0, 'model_year': 0, 'milage': 0, 'fuel_type': 38, 'engine': 38, 'transmission': 4, 'ext_col': 11, 'int_col': 98, 'accident': 0, 'clean_title': 0, 'price': 0}
```

Carateres Especiais
Relativamente a caracteres especiais, a nossa equipa não detetou nenhum no DataSet.

Dados Incoerentes
Após uma análise dos dados verificamos que não existe incoerência ao nível de pontos e vírgulas.

Linhas em Branco
Após uma análise detalhada, identificámos a presença de valores nulos em colunas específicas. As colunas 'fuel_type', 'accident' e 'clean_title' contêm valores nulos.

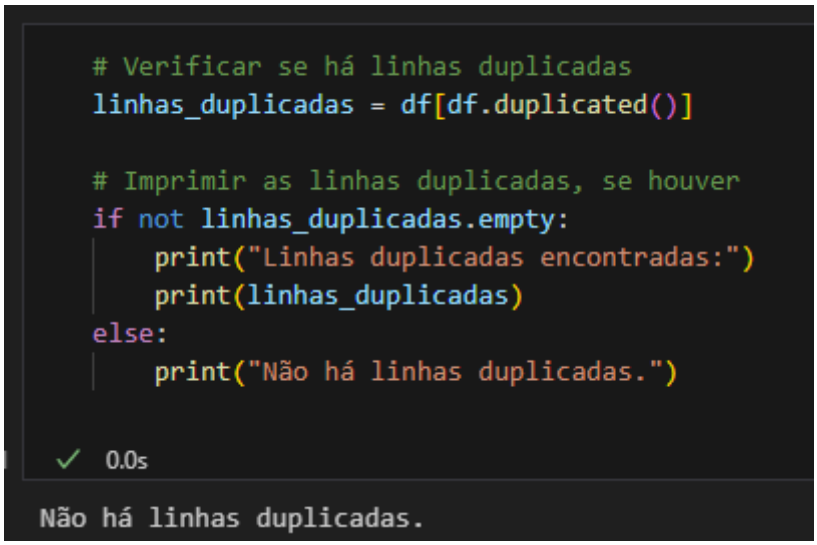


```
print((df.isnull().any()
       ))
```

✓ 0.0s

```
brand           False
model           False
model_year      False
milage          False
fuel_type        True
engine          False
transmission    False
ext_col         False
int_col         False
accident         True
clean_title      True
price           False
```

Valores Duplicados
É essencial verificar a presença de dados duplicados no DataSet, pois várias linhas idênticas comprometem a precisão do resultando e produz um impacto negativo durante a análise e exploração dos dados.
Dito isto concluímos que o nosso DataSet não contem nenhuma linha duplicada.



```
# Verificar se há linhas duplicadas
linhas_duplicadas = df[df.duplicated()]

# Imprimir as linhas duplicadas, se houver
if not linhas_duplicadas.empty:
    print("Linhas duplicadas encontradas:")
    print(linhas_duplicadas)
else:
    print("Não há linhas duplicadas.")
```

✓ 0.0s

```
Não há linhas duplicadas.
```

Valores Únicos

A nossa equipa, para a exploração de dados, decidiu também fazer a contagem de valores distintos presentes em cada coluna do 'train.csv'. Isto é útil para saber qual diversidade dos dados e mais especificamente de cada coluna.

"brand":

```
Coluna: brand
Distinct Count: 56
Distinct Percentage: 1.75%
```

"model":

```
Coluna: model
Distinct Count: 1670
Distinct Percentage: 52.07%
```

"model_year":

```
Coluna: model_year
Distinct Count: 34
Distinct Percentage: 1.06%
```

"milage":

```
Coluna: milage
Distinct Count: 2325
Distinct Percentage: 72.50%
```

"fuel_type":

```
Coluna: fuel_type
Distinct Count: 7
Distinct Percentage: 0.22%
```

"engine":

```
Coluna: engine
Distinct Count: 1015
Distinct Percentage: 31.65%
```

"transmission":

```
Coluna: transmission
Distinct Count: 54
Distinct Percentage: 1.68%
```

"ext_col":

```
Coluna: ext_col
Distinct Count: 265
Distinct Percentage: 8.26%
```

"int_col":

```
Coluna: int_col
Distinct Count: 140
Distinct Percentage: 4.37%
```

“accident”:

```
Coluna: accident
Distinct Count: 2
Distinct Percentage: 0.06%
```

“clean_title”:

```
Coluna: clean_title
Distinct Count: 1
Distinct Percentage: 0.03%
```

“price”:

```
Coluna: price
Distinct Count: 1332
Distinct Percentage: 41.53%
```

Máximos, Mínimos, Média e Quadris

Para uma melhor compreensão dos dados decidimos também calcular média, mínimos, quadris e máximos das colunas com valores categóricos que são as colunas ‘model_year’, ‘price’ e ‘milage’. Estes estão representados na seguinte coluna.

```
df[['model_year','price','milage']].describe()
```

	model_year	price	milage
count	3207.000000	3.207000e+03	3207.000000
mean	2015.517930	4.487024e+04	64594.983785
std	6.131963	8.382969e+04	52387.260707
min	1974.000000	2.000000e+03	100.000000
25%	2012.000000	1.705000e+04	23151.500000
50%	2017.000000	3.169800e+04	52253.000000
75%	2020.000000	4.999650e+04	93450.000000
max	2024.000000	2.954083e+06	405000.000000

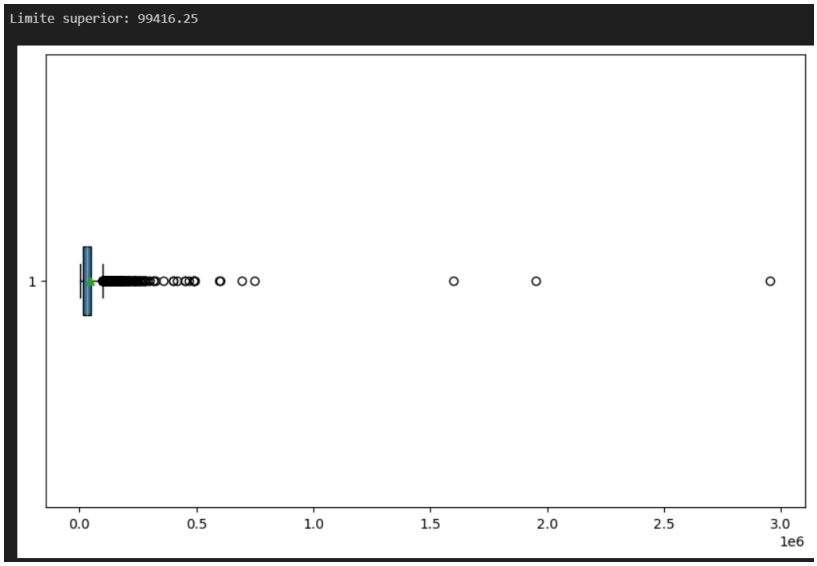
Outliers

Os outliers são pontos de dados que se destacam pela sua diferença em relação à maioria das outras observações num conjunto de dados. Eles podem indicar variações extremas, erros experimentais ou até mesmo revelar características excecionais nos dados. Os outliers podem influenciar análises estatísticas e distorcer a interpretação dos resultados, tornando crucial a identificação e compreensão desses valores atípicos para uma análise mais precisa e confiável.

Para visualizar a distribuição dos dados e identificar esses valores atípicos, o nosso grupo optou por usar box plots. Através disto, pudemos observar a magnitude e a posição desses pontos discrepantes em relação aos restantes dados, o que nos ajudou na compreensão mais profunda da distribuição estatística do conjunto de dados analisado.

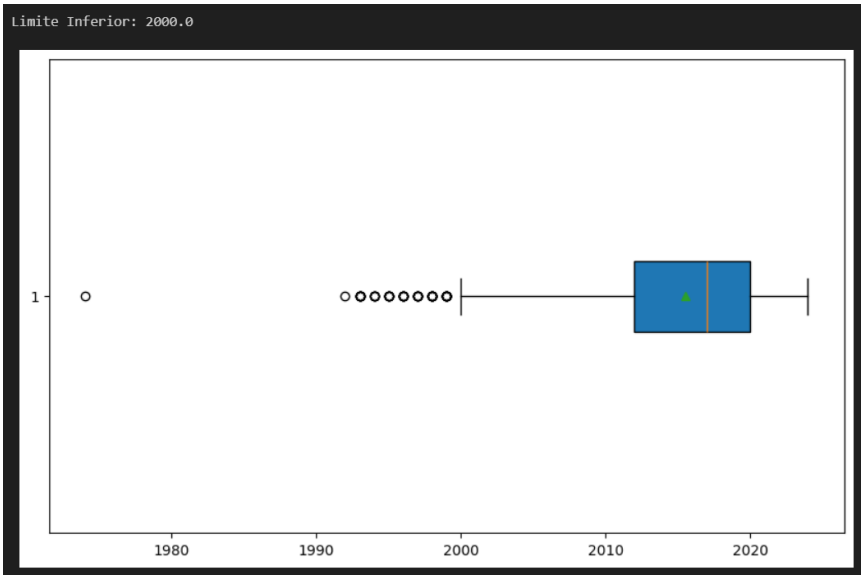
“price”:

Para calcular os outliers da coluna ‘price’ a nossa equipa utilizou o gráfico BoxPlot, e dentro deste gráfico calculou os limites dos quadris e apercebemo-nos que o limite superior da coluna price é 99416.25 dólares por isso vamos eliminar todas as linhas que possuem valores maiores que ‘99416’ dólares na coluna price.



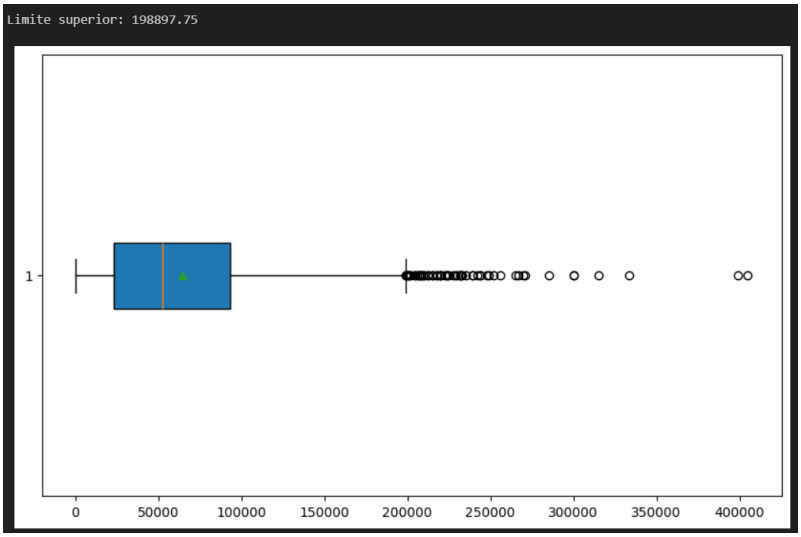
“model_year”:

Já para calcular os outliers da coluna ‘model_year’ a nossa equipa utilizou também o gráfico BoxPlot. Calculamos também os limites dos quadris e apercebemo-nos que o limite inferior da coluna ‘model_year’ é o ano 2000 por isso vamos eliminar todas as linhas que possuem um valor menor que ‘2000’ na coluna ‘model_year’. A razão de eliminarmos estes valores é por serem muito dispersos e afetarem a qualidade do algoritmo de previsão que estamos a desenvolver.



“milage”:

Por fim para a coluna ‘milage’ também fizemos uma BoxPlot e calculamos o limite superior do gráfico. Isto permitiu-nos descobrir a partir de que valores devemos considerar os nossos outliers. No caso da coluna ‘milage’ o limite superior é 198897.75 milhas por isso vamos eliminar todos os valores acima de ‘198898’ milhas.

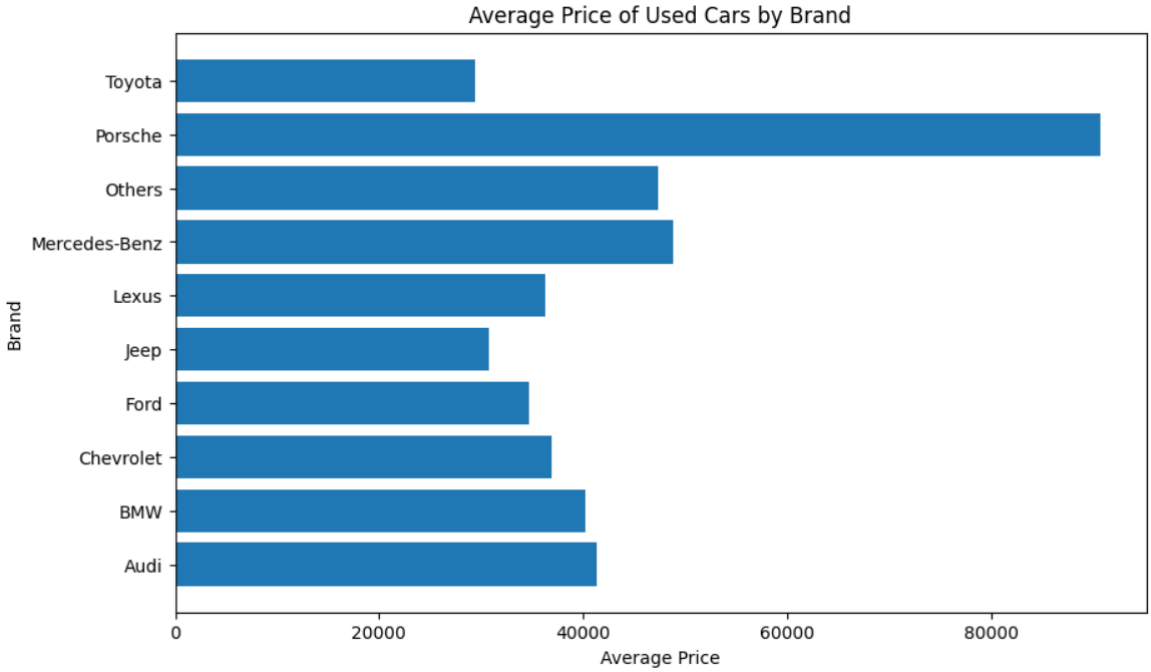


Relações entre Atributos

Neste instante, a equipa teve de examinar os dados para determinar se eles tinham impacto na escolha do cliente, visando compreender quais deveriam ou não ser levados em conta na etapa subsequente. Assim, encontram-se apresentados a seguir diversos gráficos relativos à análise dos dados realizada.

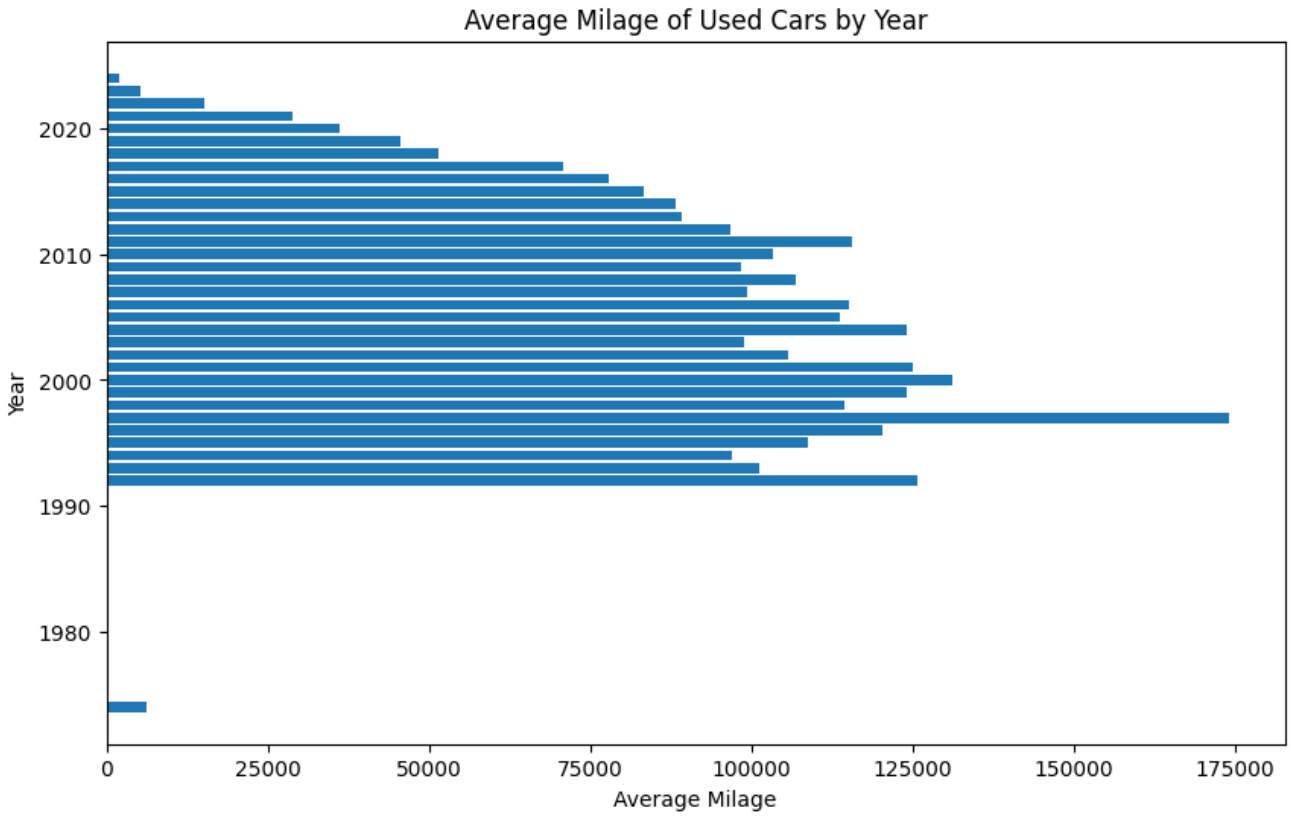
Relação do atributo *Brand* com o atributo *Price*

O seguinte gráfico representa a relação da marca dos carros usados com o preço dos mesmos. Pelo gráfico conseguimos verificar que a marca com o preço médio mais caro é a “Porsche” e a marca com o preço médio mais barato é a “Toyota”.



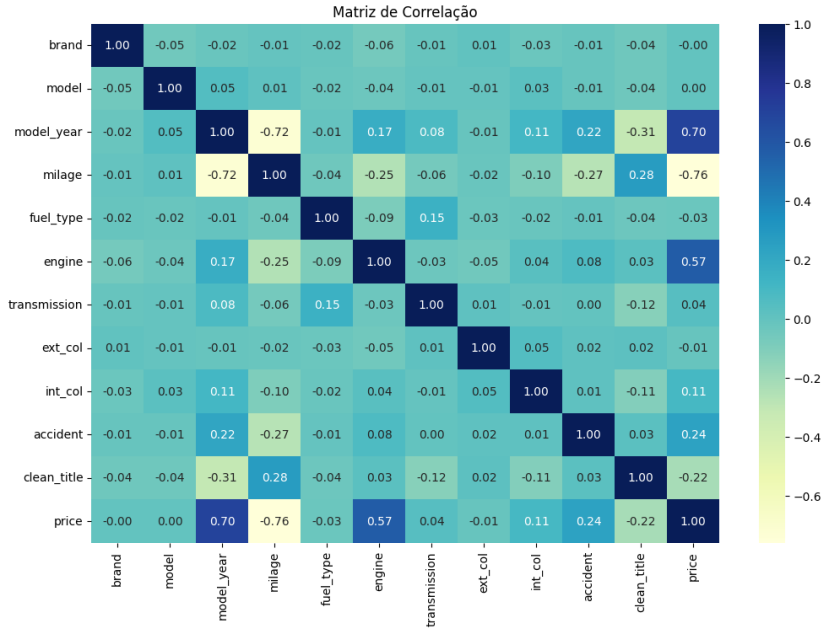
Relação do atributo *Year* com o atributo *Milage*

O seguinte gráfico representa a relação do ano com a média de milhas percorridas dos mesmos. Pelo gráfico conseguimos verificar que o ano com a maior média de milhas percorridas é “2003” e o ano com menos milhas percorridas é “2023”.



Relação entre atributos

Para representar a relação entre pares de variáveis utilizamos uma matriz Spearman. Esta matriz é calculada através da linguagem Python e viabiliza a avaliação de como uma variável se comporta em relação à outra, mesmo quando a relação entre elas não é linear. A matriz indica que várias características do carro estão correlacionadas com o seu preço. As correlações mais fortes são observadas entre o preço e a marca, o tipo do motor e o ano de fabricação do veículo. Também se verifica que outras características, como o “clean_title” e o “milage” estão correlacionadas com o preço, embora essas correlações sejam o contrário das outras, pois se uma sobe a outra desce.



Dividir coluna em cc, cylinder, horse_power

```
# Remove ".0HP" suffix from the "horse_power" column
df["cylinder"] = df["cylinder"].str.replace("V", "")

# Convert "horse_power" column to integer, handle non-numeric values with errors='coerce'
df["cylinder"] = pd.to_numeric(df["cylinder"], errors='coerce').astype("Int64")

# Display the resulting DataFrame
print(df.to_string())
```

```
# Remove ".0HP" suffix from the "horse_power" column
df["horse_power"] = df["horse_power"].str.replace(".0HP", "")

# Convert "horse_power" column to integer, handle non-numeric values with errors='coerce'
df["horse_power"] = pd.to_numeric(df["horse_power"], errors='coerce').astype("Int64")

# Display the resulting DataFrame
print(df.to_string())
```

```python
# Define a function to extract the desired values
def extract_engine_details(engine_string):
    parts = engine_string.split(" ")

    horse_power = None
    cc = None
    cylinder = None

    for part in parts:
        if 'HP' in part:
            horse_power = part
        elif 'L' in part:
            cc = part
        elif part.isdigit() or (part.startswith('V') and part[1:].isdigit()):
            cylinder = part

        if 'Liter' in part:
            cc = None

    return pd.Series({
        "horse_power": horse_power,
        "cc": cc,
        "cylinder": cylinder
    })

# Convert the engine column to string type
df["engine"] = df["engine"].astype(str)

# Extract engine details for each row
df[["horse_power", "cc", "cylinder"]] = df["engine"].apply(extract_engine_details)

df.drop('engine', axis=1, inplace=True)

# Display the resulting DataFrame
print(df.to_string())
```

Através deste código dividimos a coluna engine em cc, cylinder e horse_power em que cc significa cilindrada, cylinder significa o número de cilindros, e horse_power significa o número de cavalos do motor. Também transforma as colunas horse_power e cylinder em string.

3.4. Verificar a Qualidade dos Dados

Relatório de Qualidade de Dados

Na nossa exploração sobre a qualidade dos dados presentes no dataset, realizamos uma análise que revelou a existência de problemas de qualidade significativos. Concluímos que vários aspetos dos dados apresentam questões de qualidade que exigem modificação.

Procedemos assim à resolução e remoção de certos outliers que não vão de acordo com o projeto. Relativamente ao atributo “Milage” iremos remover valores acima de “198898” milhas. No que toca ao atributo “Model_year” vamos efetuar a remoção de valores abaixo de “2000” e no atributo Price vamos retirar todos os valores acima de “99416” dólares.

Atributo	Problemas	Soluções
Brand	Grande diversidade de marcas	Agrupar marcas para 9 marcas com maior frequência e agrupar o resto das marcas em ‘Others’
Model	n/a	-------
Model_year	Outliers	Remover valores abaixo de ‘2000’
Milage	Valores com ‘,’ e ‘mi.’ e tabela em String Outliers	Tirar ‘,’ e ‘mi.’ e transformar tabela em Integer Remover valores acima de ‘198898’
Fuel_type	Linhas com valores nulos e travessões Grande diversidade de tipos de combustível	Agrupar em ‘E85 Flex Fuel, ‘Gasoline’, ‘Hybrid’ e ‘Others’
Engine	Valores com ‘-‘	Agrupar em ‘Others’
Transmission	Grande diversidade de dados	Agrupar em ‘Automatic’, ‘Manual’ e ‘Others’
Ext_col	Grande diversidade de cores	Agrupar cores em 'Red', 'Black', 'Blue', 'White', 'Beige', 'Gray', 'Brown', 'Green', 'Gold', 'Silver', 'Yellow', e em ‘Others’
Int_col	Grande diversidade de cores	Agrupar cores em 'Red', 'Black', 'Blue', 'White', 'Beige', 'Gray', 'Brown', 'Green', 'Gold', 'Silver', 'Yellow', e em ‘Others’
Accident	Linhas com valores nulos	Trocar valores nulos por ‘None reported’
Clean_title	Linhas com valores nulos	Trocar valores nulos por ‘No’
Price	Outliers	Remover valores acima de ‘99416’

A tabela apresentada representa os atributos, os erros encontrados em cada um e as suas respetivas soluções.

4. Preparação de Dados

Descrição do Dataset

O foco principal da próxima etapa é alterar os dados usados em todo o projeto, visando torná-los mais pertinentes e compreensíveis. Visto isto, decidimos alterar a composição das colunas *Brand, Model_year, Milage, Fuel_type, Engine, Transmission, Ext_col, Int_col, Accident, Clean_title* e *Price.* Estas alterações dos dados foram efetuadas utilizando a linguagem Python.

4.1. Seleção de Dados

A seleção dos dados representa a lista dos atributos excluídos. A nossa equipa chegou à conclusão de que todos os atributos presentes no dataset são relevantes e que dispensam remoção.

4.2. Limpeza dos dados

Após a limpeza dos dados e eliminação dos outliers das colunas 'model_year', 'price' e 'milage' a coluna relativamente aos mínimos máximos média e quadris vai ser atualizada como está explicito na imagem seguinte. A nossa equipa eliminou carros do ano abaixo de 2000, distância percorrida acima de 198898 milhas e preço acima de 99416, pois estes valores se encontravam muito dispersos dentro da amostra que nos foi fornecida e poderá interferir de forma negativa na precisão do algoritmo de previsão de preço de carros usados que estamos a desenvolver.

```
df[['model_year','price','milage']].describe()
✓ 0.0s
```

	model_year	price	milage
count	2904.000000	2904.000000	2904.000000
mean	2015.758609	34323.280647	64005.247245
std	5.509263	21179.586527	46337.138808
min	2000.000000	2000.000000	100.000000
25%	2013.000000	17500.000000	26000.000000
50%	2017.000000	30351.000000	54141.000000
75%	2020.000000	46139.750000	92895.750000
max	2024.000000	99000.000000	198868.000000

Em relação aos seguintes atributos, foram corrigidos pequenos erros devidamente explicados, uma vez que não contribuíam para o projeto.

Atributo	Solução	Motivos
Brand	Agrupar marcas para 9 marcas com maior frequência e agrupar o resto das marcas em ‘Others’	Melhorar a visualização e analise dos dados.
Model	n/a	----------
Model_year	Remover valores abaixo de ‘2000’	Dados muito dispersos o que pode afetar o algoritmo de previsão de dados
Milage	Tirar ‘,’ e ‘mi.’ e transformar tabela em Integer Remover valores acima de ‘198898’	Padronização, Facilitar Operações Matemáticas, Evitar Erros em Análises e Dados muito dispersos o que pode afetar o algoritmo de previsão de dados
Fuel_type	Agrupar em ‘Diesel’, ‘Gasoline’, ‘Hybrid’ e ‘Others’	Melhorar a visualização e analise dos dados.
Engine	Agrupar em ‘Others’	Melhorar a visualização e analise dos dados.
Transmission	Agrupar em ‘Automatic’, ‘Manual’ e ‘Others’	Melhorar a visualização e analise dos dados.
Ext_col	Agrupar cores em 'Red', 'Black', 'Blue', 'White', 'Beige', 'Gray', 'Brown', 'Green', 'Gold', 'Silver', 'Yellow', e em ‘Others’	Melhorar a visualização e analise dos dados.
Int_col	Agrupar cores em 'Red', 'Black', 'Blue', 'White', 'Beige', 'Gray', 'Brown', 'Green', 'Gold', 'Silver', 'Yellow', e em ‘Others’	Melhorar a visualização e analise dos dados.
Accident	Trocar valores nulos por ‘None reported’	Melhorar a visualização e analise dos dados.
Clean_title	Trocar valores nulos por ‘No’	Melhorar a visualização e analise dos dados.
Price	Remover valores acima de ‘99416’	Dados muito dispersos o que pode afetar o algoritmo de previsão de dados

4.3. Construção dos dados

Para a construção de dados a nossa equipa criou 4 cenários distintos que irão ser desenvolvidos e usados na terceira e última entrega. Estes cenários vão ser estudados de modo a identificar qual o que apresenta melhores resultados em relação ao preço dos carros usados.
Os cenários envolvem combinações entre remoção de outliers e remoção de colunas. A nossa equipa removeu outliers nas tabelas, “model_year”, “milage” e “price”. Após o estudo do heatmap do ponto Relação entre Atributos concluímos que as colunas relevantes para a target são a ‘model_year’, ‘engine’ e ‘milage’, pois apresentem uma relação em valor absoluto de >0.50, o que o nosso grupo achou adequado. Também vai ser necessário substituir possíveis valores nulos por average dessa coluna.

Os cenários são:

Cenário 1 - Com outliers e com todas as colunas do dataset.

Cenário 2 - Sem outliers e com todas as colunas do dataset.

Cenário 3 – Com outliers e com apenas as colunas selecionadas.

Cenário 4 – Sem outliers e com apenas as colunas selecionadas.

Integração dos dados

A nossa equipa não achou pertinente adicionar dados pois após criar os cenários concluímos que estavam reunidos todos os dados necessários para prosseguir para a fase final de modelação do nosso projeto.

4.4. Formatação dos dados

A formatação de dados é uma fase crucial no processo de preparação para análise e modelagem. Consiste em efetuar ajustes na estrutura e na apresentação dos dados, sem alterar o seu significado essencial. Estas modificações são frequentemente necessárias para cumprir os requisitos específicos das ferramentas de análise ou modelagem utilizadas. Quanto ao atributo "Transmission", foi essencial realizar uma conversão de categórico para numérico, uma vez que inicialmente estava no formato de string. Essa

adaptação possibilitou a obtenção de resultados com uma margem de erro significativamente reduzida.

Transmission	Valor Numérico
Automatic	0
Manual	1
Others	2

No atributo "Int_col", foi necessário agrupar as cores em 12 opções devido à grande quantidade de escolhas disponíveis no dataset.

Int_col	Valor Numérico
Red	0
Black	1
Blue	2
White	3
Beige	4
Gray	5
Brown	6
Green	7
Gold	8
Silver	9
Yellow	10
Others	11

No atributo "Ext_col", foi necessário agrupar as cores em 12 opções devido à grande quantidade de escolhas disponíveis no dataset.

Ext_col	Valor Numérico
Red	0
Black	1
Blue	2
White	3
Beige	4
Gray	5
Brown	6
Green	7
Gold	8
Silver	9
Yellow	10
Others	11

No atributo "Fuel_type", foi necessário agrupar os tipos de combustíveis em 4 opções devido à grande quantidade de escolhas disponíveis no dataset.

Fuel_type	Valor Numérico
Diesel	0
Gasoline	1
Hybrid	2
Others	3

No atributo "Brand", foi necessário agrupar as marcas dos veículos em 8 opções devido à grande quantidade de escolhas disponíveis no dataset.

Brand	Valor Numérico
BMW	0
Audi	1
Ford	2
Mercedes-Benz	3
Jeep	4
Porsche	5
Chevrolet	6
Toyota	7
Lexus	8
Others	9

O atributo “accident” é do tipo booleano e por isso só assumia dois valores que transformamos em 0 e 1.

Accident	Valor Numérico
None reported	0
At least 1 accident or damage reported	1

Igualmente ao anterior o atributo “clean_title” é booleano e por isso só assume 2 valores que também transformamos em 0 e 1.

Clean_title	Valor Numérico
Yes	0
No	1

5. Modelação

5.1. Seleção de técnicas de modelação

Nesta fase do projeto, foram selecionadas várias técnicas de modelação, escolhidas cuidadosamente pela sua adequação aos diversos cenários criados e testados. Assim, temos a oportunidade para aprofundar significativamente a análise, fazendo uma comparação entre as diversas técnicas de modelação escolhidas para cada um dos cenários. Para avaliar o desempenho das técnicas, vamos utilizar os critérios de sucesso de data mining referidos anteriormente no relatório.

As técnicas que utilizamos foram as seguintes:

Random Forest

Uma abordagem de aprendizagem supervisionada que reúne várias árvores de decisão, com o objetivo de aprimorar a precisão do modelo e mitigar o sobreajuste.

Gradient Boosted Trees

Uma técnica de aprendizagem supervisionada que utiliza um algoritmo de impulsionamento para construir uma sequência de árvores de decisão, permitindo que estas se complementem de forma sinérgica.

Neural Net-Deep Learning

Uma metodologia de aprendizagem não supervisionada que recorre a uma rede neural artificial para extrair padrões complexos e aprender representações significativas a partir dos dados.

Decision Tree

Um método de aprendizagem supervisionada que cria uma estrutura de árvore de decisão, proporcionando uma representação visual do fenómeno do mundo real e facilitando a interpretação das relações entre variáveis.

Com a utilização destas técnicas, vai ser possível determinar qual dos cenários é o melhor.

5.2. Gerar modelos de teste

Relativamente a esta fase, a equipa optou por definir vários cenários e mecanismos para testar a qualidade do modelo em questão. Neste sentido, podemos afirmar que cada um dos cenários criados é composto por uma combinação de variáveis fornecidas pelo conjunto de dados, técnicas utilizadas e métodos de teste e validação. Assim, apresentam-se de seguida os diversos cenários elaborados pela equipa.

Cenários	Atributos Selecionados	Técnicas a usar	Método de teste e Validação
1	Todas as colunas com outliers e sem nulos(substituídos por average): brand model_year milage fuel_type engine transmission ext_col int_col accident clean_title price horse_power cylinder cc	Todas	Cross-Validation
2	Todas as colunas sem outliers e sem nulos(substituídos por average): brand model_year milage fuel_type engine transmission ext_col int_col accident clean_title price horse_power cylinder cc	Todas	Cross-Validation

3	Apenas as colunas selecionadas com outliers e sem nulos(substituídos por average): cylinder horse_power milage model_year price	Todas	Cross-Validation
4	Apenas as colunas selecionadas sem outliers e sem nulos(substituídos por average): cylinder horse_power milage model_year price	Todas	Cross-Validation

5.3. Construir Modelos

Configuração dos Parâmetros

Para construir os modelos o nosso grupo decidiu utilizar a ferramenta Rapid Miner que é uma plataforma de ciência de dados e análise preditiva que nos permite extrair insights valiosos a partir de dados complexos. Esta ferramenta oferece uma interface gráfica intuitiva para a construção, treino e avaliação de modelos de Aprendizagem Automática, sem a necessidade de programação extensiva.

No componente “Set Role” o parâmetro utilizado foi o ‘price’ pois é o que estamos a variável que estamos a tentar prever.

No componente “Cross Validation” para o parâmetro ‘number of folds’ colocamos 10 e o ‘sampling type’ é automatic.

O componente “Apply Model” permaneceu inalterado enquanto no componente “Performance (Regression)” selecionamos ‘root mean squared error’, ‘absolute error’, ‘relative error strict’ e ‘squared correlation’.

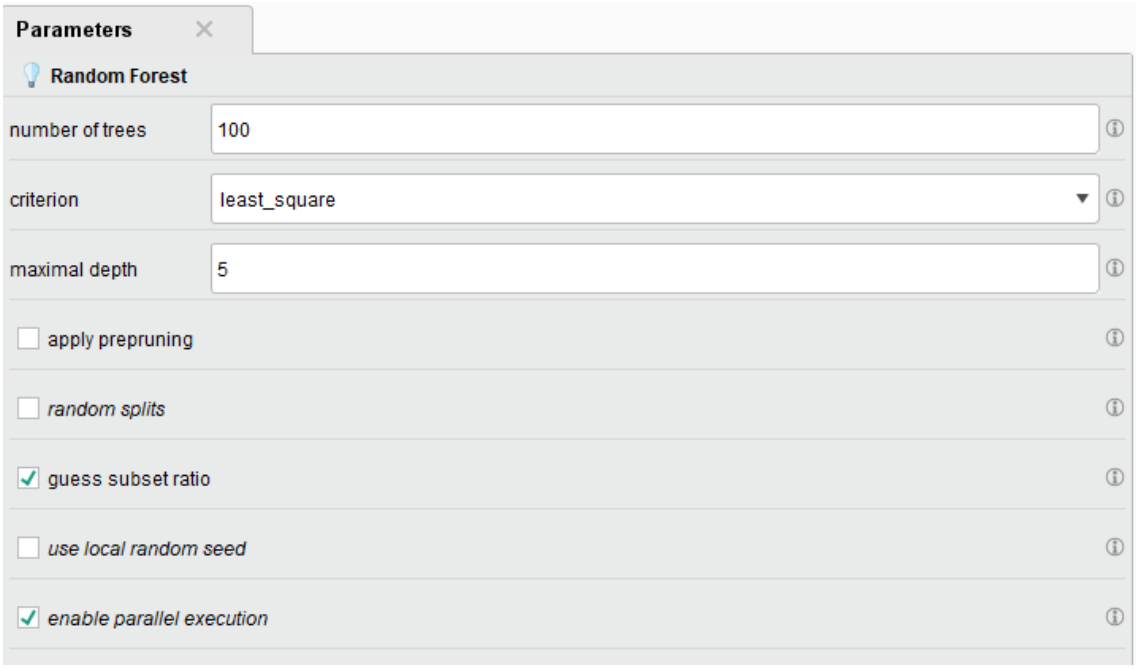
Para além destes utilizamos os componentes Sellect Attibutes e Replace Missing Values de modo a podermos construir os cenários dentro do ambiente do Rapid Miner. No componente Sellect Attibutes, o parâmetro ‘type’ escolhemos include attributes, e para o parâmetro ‘attribute filter type’ escolhemos a subset. No componente Replace Missing Values no ‘attribute filter type’ escolhemos subset e no parâmetro ‘default’ escolhemos average, deste modo substitui valores nulos por average dessa coluna, apenas nas colunas selecionadas por nós que são as colunas ‘cc’ ‘cylinder’ e ‘horse_power’.

5.4. Modelo Gerado

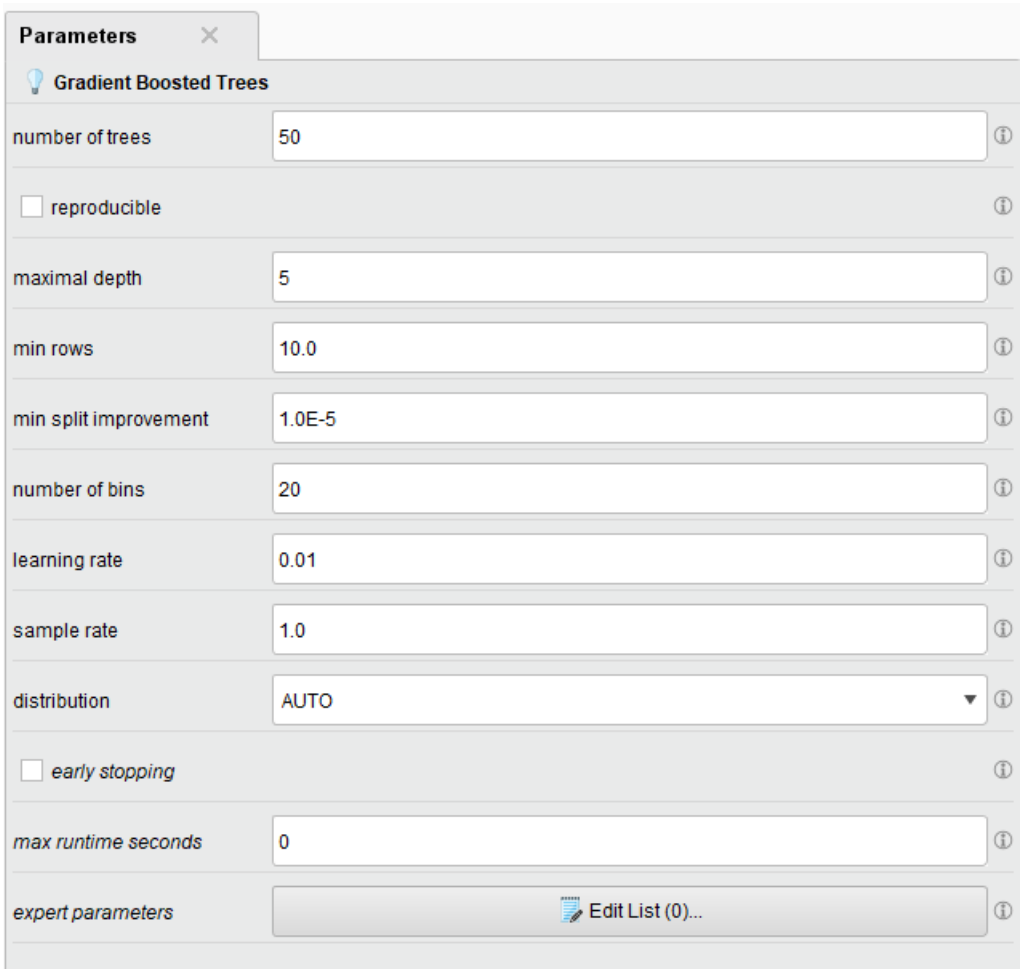
A nossa equipa estudou quatro modelos de aprendizagem automática que são e cada um deles foi configurado com os parâmetros seguintes:

Decision Tree

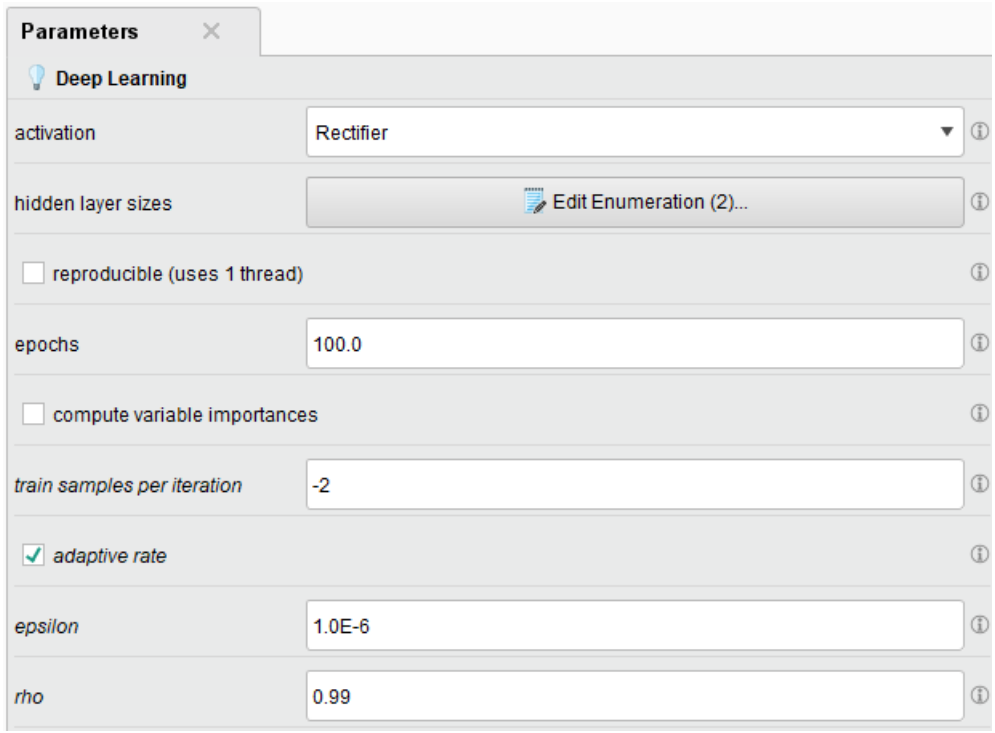
Random Forest



Gradient Boosted Trees



Neural Net- DeepLearning



5.5. Avaliação dos modelos

Os diversos cenários apresentados a seguir foram analisados utilizando o método Cross-Validation, tendo em atenção os fatores explicados anteriormente. Em todos os cenários, os valores nulos foram substituídos pelo valor average dessa coluna.

Cenário 1

O seguinte cenário contém todos os atributos presentes com outliers no dataset. Os atributos inseridos são brand, model_year, milage, fuel_type, engine, transmission, ext_col, int_col, accident, clean_title, price, horse_power, cylinder e cc.

Na seguinte tabela estão representadas quatro interações para cada uma das métricas.

Métrica / Técnica	Erro Médio Quadrático (MSE)	Erro Médio Absoluto (MAE)	Erro Percentual Médio (MAPE)	Coeficiente de Determinação (R^2)
Random Forest	57254.388	15271.395	59.75%	0.522
Gradient Boosted Trees	64015.819	22723.799	107.07%	0.449
Neural Net-Deep Learning	54793.658	14001.346	50.25%	0.531
Decision tree	63333.179	20050.750	62.49%	0.382

Cenário 2

Relativamente ao cenário 2, este contém todos os atributos presentes sem outliers no dataset. Os atributos inseridos são: brand, model_year, milage, fuel_type, engine, transmission, ext_col, int_col, accident, clean_title, price, horse_power, cylinder e cc.

Na seguinte tabela estão representadas quatro interações para cada uma das métricas.

Métrica Técnica	Erro Médio Quadrático (MSE)	Erro Médio Absoluto (MAE)	Erro Percentual Médio (MAPE)	Coeficiente de Determinação (R^2)
Random Forest	10366.085	7549.702	35.47%	0.793
Gradient Boosted Trees	15475.121	12105.684	68.08%	0.772
Neural Net-Deep Learning	8054.302	5608.722	26.09%	0.855
Decision tree	12478.072	9199.921	41.15%	0.654

Cenário 3

O cenário 3, contém os seguintes atributos com outliers, que são: cylinder, horse_power, milage, model_year e price.

Na seguinte tabela estão representadas quatro interações para cada uma das métricas.

Métrica Técnica	Erro Médio Quadrático (MSE)	Erro Médio Absoluto (MAE)	Erro Percentual Médio (MAPE)	Coeficiente de Determinação (R^2)
Random Forest	60529.158	18386.987	57.79%	0.416
Gradient Boosted Trees	65289.153	23296.292	106.67%	0.399
Neural Net-Deep Learning	58583.372	20857.560	132.95%	0.431
Decision tree	66558.369	19987.893	60.83%	0.362

Cenário 4

Relativamente ao cenário 4, este contém os seguintes atributos sem outliers: cylinder, horse_power, milage, model_year e price.

Na seguinte tabela estão representadas quatro interações para cada uma das métricas.

Métrica / Técnica	Erro Médio Quadrático (MSE)	Erro Médio Absoluto (MAE)	Erro Percentual Médio (MAPE)	Coeficiente de Determinação (R^2)
Random Forest	11605.648	8390.332	36.74%	0.705
Gradient Boosted Trees	15753.554	12262.035	68.69%	0.714
Neural Net-Deep Learning	11306.023	8107.935	51.53%	0.723
Decision tree	12432.716	9185.959	40.31%	0.656

6. Avaliação

6.1. Avaliação dos resultados

Validação dos objetivos de Data Mining

O nosso grupo decidiu utilizar 4 critérios de sucesso de data mining já anteriormente mencionados, pois estes vão permitir avaliar o sucesso do nosso modelo.
Após examinar o ponto Avaliação de Resultados concluímos que o melhor modelo é o Deep Learning no Cenário 2.
Para o critério de sucesso **Raiz do Erro Médio Quadrático (RMSE)** a nossa esquipa estipulou um valor abaixo de **10000** e o valor obtido depois do treino do modelo Deep Learning é **8054.302**, o que **satisfaz** o critério.
Para o critério de sucesso **Erro Médio Absoluto (MAE)** a nossa esquipa estipulou um valor abaixo de **5000** e o valor obtido depois do treino do modelo Deep Learning é **5608.722**, o que **não satisfaz** o critério.
Para o critério de sucesso **Erro Percentual Médio (MAPE)** a nossa esquipa estipulou um valor abaixo de **25%** e o valor obtido depois do treino do modelo Deep Learning é **26.09%**, o que **não satisfaz** o critério.
Para o critério de sucesso **Coeficiente de Determinação (R^2)** a nossa esquipa estipulou um valor acima de **0.8** e o valor obtido depois do treino do modelo Deep Learning é 0.**855**, o que **satisfaz** o critério.

Aprovação do modelo

A aprovação do modelo de Deep Learning no Cenário 2 representa um marco significativo para a nossa equipa, evidenciando não apenas sua excelência, mas também sua capacidade de atender a dois dos quatro critérios de sucesso de Data Mining. Isso reforça não apenas a robustez do modelo em si, mas também sua relevância para atender às metas e expectativas específicas que delineamos no início do processo.

6.2. Revisão do processo

Ao longo do projeto, deparamo-nos com desafios relacionados à qualidade do dataset fornecido pelo docente. A análise revelou falta de uniformidade nos dados, exigindo intervenções manuais para padronizar terminologias e garantir uma avaliação consistente. Adicionalmente, a existência de lacunas em alguns dados levou-nos a adotar estratégias distintas: procurámos informações externas para o preenchimento manual em colunas específicas.
Estas medidas corretivas foram cruciais para manter a integridade do dataset e garantir a fiabilidade nas análises. A uniformização das terminologias e o preenchimento adequado das lacunas contribuíram significativamente para resultados mais robustos, superando os desafios iniciais relacionados à qualidade dos dados fornecidos.
Quanto ao modelo desenvolvido, consideramos que ele alcança os nossos objetivos de negócio e é capaz de responder eficazmente às necessidades da empresa.

6.3. Determinação dos próximos passos

Ao rever os vários modelos desenvolvidos, a equipa identificou uma possibilidade de aprimoramento em alguns cenários que ainda não atingiram o seu potencial máximo. Para concretizar essa otimização, tornou-se claro que seria necessário realizar uma análise mais aprofundada desses cenários, adotando ferramentas mais adequadas que assegurassem uma qualidade de dados superior. Esse refinamento na qualidade dos dados, consequentemente, abriria portas para uma ampliação nas soluções ideais disponíveis.
É importante ainda salientar que foi proposta à equipa a realização de um novo projeto, abrangendo o desenvolvimento e a implementação da solução escolhida. Esta etapa representa a continuação da metodologia CRISP-DM, incluindo a elaboração do relatório correspondente.

7. Conclusão

Na disciplina de Aprendizagem Automática em Sistemas Empresariais, a nossa equipa estudou um dataset sobre o preço de carros usados e com ferramentas de data mining iremos prever se é viável ao nosso cliente adquirir o veículo para revenda. Para alcançar esse objetivo, adotamos a metodologia CRISP-DM, que ofereceu o suporte essencial ao longo do projeto.

A linguagem principal utilizada ao longo do projeto foi o Python. A equipa considera que concluiu com êxito todas as metas estabelecidas, alcançando resultados positivos.

Acreditamos que esta experiência nos permitiu aprofundar os nossos conhecimentos, os quais podem ser aplicados no futuro por todos os membros do grupo.

8. Bibliografia

Pandas. (2018). *Python Data Analysis Library — pandas: Python Data Analysis Library*. Pydata.org. https://pandas.pydata.org/

Altair RapidMiner Community. (2021, November 24). RapidMiner Community. https://community.rapidminer.com/

Hotz, N. (2023, January 19). *CRISP-DM*. Data Science Project Management. https://www.datascience-pm.com/crisp-dm-2/

Machine Learning and RapidMiner Tutorials | Altair Engineering Inc. Academy. (n.d.). Academy.rapidminer.com. Retrieved January 14, 2024, from https://academy.rapidminer.com/catalog

IBM SPSS Modeler CRISP-DM Guide. (n.d.). https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf

Cleaning, Feature Eng. and EDA - Car Prices. (n.d.). Kaggle.com. Retrieved January 14, 2024, from https://www.kaggle.com/code/francofaundez/cleaning-feature-eng-and-eda-car-prices#EDA