Nome: Leonardo Santos Miranda Mestrado

Relatório Final - Trabalho Prático 1

Análise das características de um carro e como elas se comportam

Base de dados e contextualização

A base de dados escolhida foi encontrada no Kaggle, denominada como Car Features and MSRP (https://www.kaggle.com/CooperUnion/cardataset). A base de dados está no formato CSV e é composta por 16 atributos e 11914 registros diferentes, fornecidos pelos websites Edmunds e Twitter, que descrevem um carro. Dentre os atributos, temos a montadora, o modelo do carro, o ano de fabricação, o tipo do motor de gasolina, o tipo de transmissão, o número de portas, o tamanho do carro e entre outros. O dataset foi escolhido por possuir uma documentação clara e com um bom número de registros e atributos de fácil entendimento e de uso para a mineração de padrões frequentes.

Objetivo

O objetivo inicial deste primeiro trabalho prático é fazer análises nas características de um carro, levando hipóteses como:

- Qual o tipo de transmissão mais comumente fabricado? Carros manuais, automáticos ou ambos?
- Quais as montadoras que mais fabricam carros?
- Quais os estilos de carros mais comumente fabricados?
- Quais os conjuntos de categoria de mercado mais frequentes nos carros fabricados? (Ex.: Luxúria, performance, alta-performance, nenhuma, etc.)

Motivação

Através das análises das características pode-se ter uma base do que, por exemplo, é mais popular e preferido pelos clientes. Desta forma, a montadora em questão pode levantar suposições através dos resultados das análises e melhorar, por exemplo, a quantidade de carros a serem produzidos com uma determinada característica, aumentando suas vendas.

Trabalhos Relacionados

Dentre os trabalhos relacionados com a base de dados, alguns notebooks na página do Kaggle (https://www.kaggle.com/CooperUnion/cardataset/notebooks) foram importantes para um maior entendimento dos dados, dentre eles o trabalho que mais se destacou, sendo relevante para este trabalho prático foi o:

- Exploratory Data Analysis, do usúario Tharun

Neste trabalho foi realizada uma exploração completa dos dados utilizando as bibliotecas do Python, obtendo informações relevantes como:

```
df = pd.read_csv("../input/cardataset/data.csv")
df.head(5)
```

- Uma lista completa dos atributos e seus tipos

```
df.dtypes
```

```
Make                object
Model               object
Year                 int64
Engine Fuel Type    object
Engine HP          float64
Engine Cylinders   float64
Transmission Type   object
Driven_Wheels       object
Number of Doors    float64
Market Category     object
Vehicle Size        object
Vehicle Style       object
highway MPG          int64
city mpg             int64
Popularity           int64
MSRP                 int64
dtype: object
```

- Quantidade de linhas repetidas no dataset

```
duplicate_rows_df = df[df.duplicated()]
print("number of duplicate rows: ", duplicate_rows_df.shape)
```

```
number of duplicate rows:  (989, 10)
```

- Se há algum atributo com dados nulos (quantidade)

```
print(df.isnull().sum())
```

```
Make            0
Model           0
Year            0
HP             69
Cylinders      30
Transmission    0
Drive Mode      0
MPG-H           0
MPG-C           0
Price           0
```

Metodologia

A metodologia seguida foi inspirada no CRISP-DM e ilustrada por fluxos de trabalhos realizados na plataforma Lemonade.

Resultados experimentais e análise

Nesta seção haverá imagens contendo ilustrações dos resultados obtidos através dos fluxos de trabalhos realizados na plataforma Lemonade, com as hipóteses descritas na seção de objetivo.

- **Hipótese 1:** Qual o tipo de transmissão mais comumente fabricado? Carros manuais, automáticos ou ambos?

Através da mineração de sequências, com um suporte mínimo de 20%, os seguintes resultados foram gerados:

Figura 1 (Fluxo de Trabalho para hipótese 1)

Trabalho Prático 1 - Mineração de Padrões Frequentes. Imagem gerada em Mon Jan 25 2021 14:44:47 GMT-0300 (Hora padrão de Brasília)

Figura 2 (Resultado obtido para hipótese 1) // Execução #3100

Results

Sequence	Freq
[["AUTOMATIC"]]	8266
[["MANUAL"]]	2935

2 registros

Com essa execução, foi possível observar que de fato, carros automáticos são mais produzidos pelas montadoras. Um detalhe curioso foi que os carros de ambos os tipos (tanto automático e manual) nem foram apresentados por possuir um suporte mínimo menor do que 20%. Para que eles sejam apresentados, o suporte mínimo teve que ser alterado para 1%, gerando uma nova tabela de resultados:

Figura 3 (Novo resultado obtido para hipótese 1) // Execução #3102

Results

Sequence	Freq
[["AUTOMATIC"]]	8266
[["AUTOMATED_MANUAL"]]	626
[["MANUAL"]]	2935

3 registros

- **<u>Hipótese 2</u>:** Quais as montadoras que mais fabricam carros?

Ainda com a mineração de sequências, utilizando desta vez um suporte mínimo de 5%, os seguintes resultados foram gerados:

Figura 1 (Fluxo de Trabalho para hipótese 2)

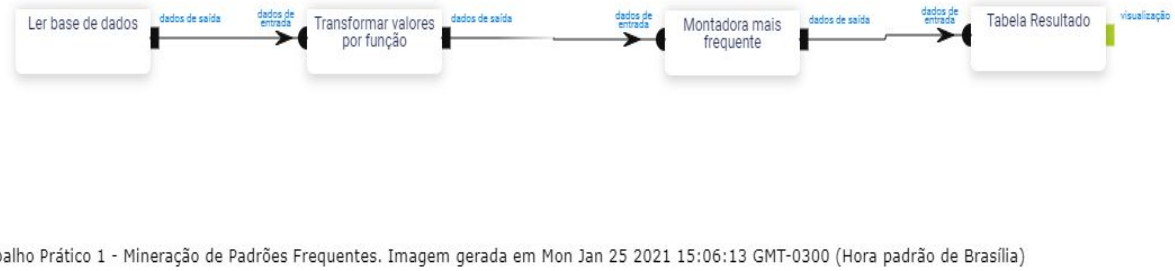


Figura 2 (Resultado obtido para hipótese 2) // Execução #3108

Results

Sequence	Freq
[["Chevrolet"]]	1123
[["Dodge"]]	626
[["Volkswagen"]]	809
[["Toyota"]]	746
[["Ford"]]	881

5 registros

Confirmando assim que a montadora Chevrolet é a mais frequente na base de dados, seguida por Ford e Volkswagen. As outras montadoras não atingiram o suporte mínimo de 5% e não foram apresentadas.

- **Hipótese 3:** Quais os estilos de carros mais comumente fabricados?

Finalizando a utilização da mineração de sequências, para esta hipótese um suporte de 10% foi utilizado, gerando os seguintes resultados:

Figura 1 (Fluxo de trabalho para hipótese 3)

Trabalho Prático 1 - Mineração de Padrões Frequentes. Imagem gerada em Mon Jan 25 2021 15:36:25 GMT-0300 (Hora padrão de Brasília)

Figura 2 (Resultado obtido para hipótese 3) // Execução #3119

Results

Sequence	Freq
[["Coupe"]]	1211
[["4dr"]]	3190
[["Hatchback"]]	1208
[["Pickup"]]	1696
[["SUV"]]	2655
[["Cab"]]	1696
[["Sedan"]]	3048
[["Cab"], ["Pickup"]]	1696
[["4dr"], ["SUV"]]	2488

9 registros

Confirmando assim que os carros de estilo “4dr” são mais frequentes nessa base de dados.

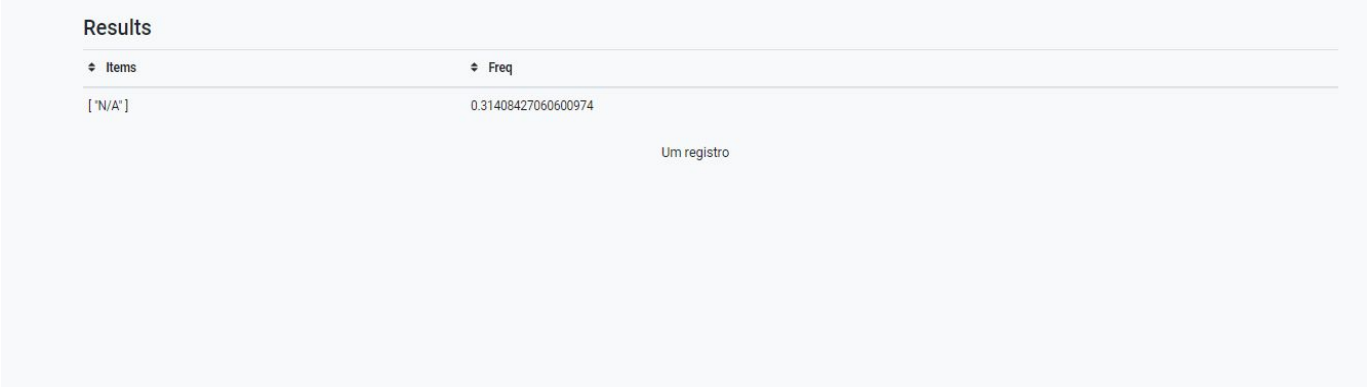
- **<u>Hipótese 4:</u>** Quais os conjuntos de categoria de mercado mais frequentes nos carros fabricados? (Ex.: Luxúria, performance, alta-performance, nenhum, etc.)

Por último, para gerar os resultados desta hipótese, foi utilizado no lemonade a operação de mineração de itemsets frequentes com um suporte mínimo de 20% e confiança mínima de 60% para geração das regras, obtendo os seguintes resultados:

Figura 1 (Fluxo de trabalho para hipótese 4)

Trabalho Prático 1 - Mineração de Padrões Frequentes. Imagem gerada em Mon Jan 25 2021 15:48:30 GMT-0300 (Hora padrão de Brasília)

Figura 2 (Resultado obtidos para hipótese 4)


Results

Items	Freq
["N/A"]	0.31408427060600974

Um registro

O resultado desta hipótese me deixou bastante surpreso: na verdade, os carros que não possuem uma categoria de mercado específica são os mais frequentes.

Conclusões e perspectivas

No desenvolvimento deste trabalho comecei a entender de maneira todo o material das videoaulas que assisti. Também percebi que na proposta de trabalho prático que eu entreguei havia muitas coisas sem sentido para o tema de mineração de padrões frequentes. Espero que os resultados que obtive através das diversas execuções realizadas no Lemonade sejam satisfatórios e tenham relevância para este trabalho prático. De forma resumida, foi um trabalho que me fez correr atrás e me trouxe experiência em mineração de dados. Vou anexar juntamente com este pdf uma pasta com todas as imagens dos resultados obtidos, caso sejam difíceis de visualizar.