Nome: Leonardo Santos Miranda Mestrado

Relatório Final - Trabalho Prático 4

Análise das características de um carro para predizer o seu preço de varejo estimado.

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 11915 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, facilitando o treinamento da rede neural para a aplicação das técnicas de regressão.

Objetivo

O objetivo deste trabalho é treinar um modelo utilizando os algoritmos de regressão, como a Regressão Linear,Random Forest e Gradient Boosted Tre. Para isso, a base de dados será separada em um conjunto de treino e outro de teste e será treinado para que consiga compreender as 16 características e o rótulo msrp, que é o objetivo da predição.

Motivação

_____Ao utilizar os diversos algoritmos de regressão vou entender melhor observando os resultados que foram obtidos em cima da base de dados. Analisar o comportamento dos algoritmos utilizados vai me ajudar a compreender melhor o funcionamento de cada um.

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:

- Car Retail Price Prediction, do usúario Gabriel Atkin

Neste trabalho, foi realizada uma predição do preço de varejo (rótulo msrp), mas antes disso uma breve análise dos dados foi feita:

Figura 1- Exibição dos atributos da base de dados e o seus respectivos 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
```

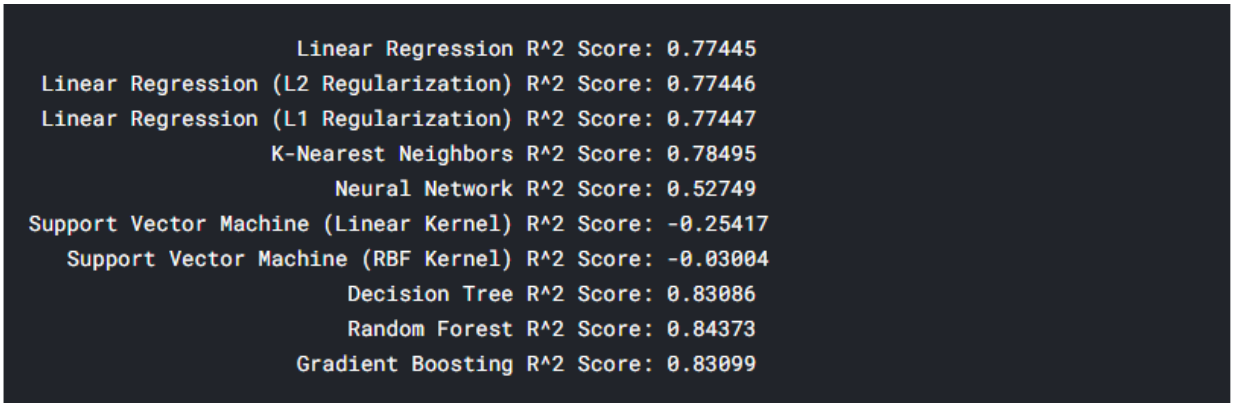
A base de dados é composta por 15 atributos que descrevem um carro mas apenas 7 deles são numéricos que serviram de entrada para os algoritmos de regressão. Seguindo o trabalho relacionado, os seguintes algoritmos de regressão foram selecionados para o treinamento:

Figura 2 - Modelos de regressão utilizados para o treinamento, do autor Gabriel Atkin

```
models = {
    "                    Linear Regression": LinearRegression(),
    " Linear Regression (L2 Regularization)": Ridge(),
    " Linear Regression (L1 Regularization)": Lasso(),
    "                  K-Nearest Neighbors": KNeighborsRegressor(),
    "                       Neural Network": MLPRegressor(),
    "Support Vector Machine (Linear Kernel)": LinearSVR(),
    "   Support Vector Machine (RBF Kernel)": SVR(),
    "                         Decision Tree": DecisionTreeRegressor(),
    "                         Random Forest": RandomForestRegressor(),
    "                     Gradient Boosting": GradientBoostingRegressor()
}
```

Antes do treinamento, o autor do notebook realizou um pré-processamento dos dados, preenchendo possíveis dados nulos. Após a limpeza dos dados, foi dado início ao treinamento dos modelos, obtendo os seguintes resultados para cada modelo:

Figura 3 - Resultados obtidos para cada algoritmo de regressão, do autor Gabriel Atkin



Analisando os resultados obtidos desse autor, o algoritmo de regressão que maior se sobressaiu dentre os outros foi o Random Forest, seguido do Gradient Boosting, uma indicativa para o nosso trabalho.

Metodologia

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

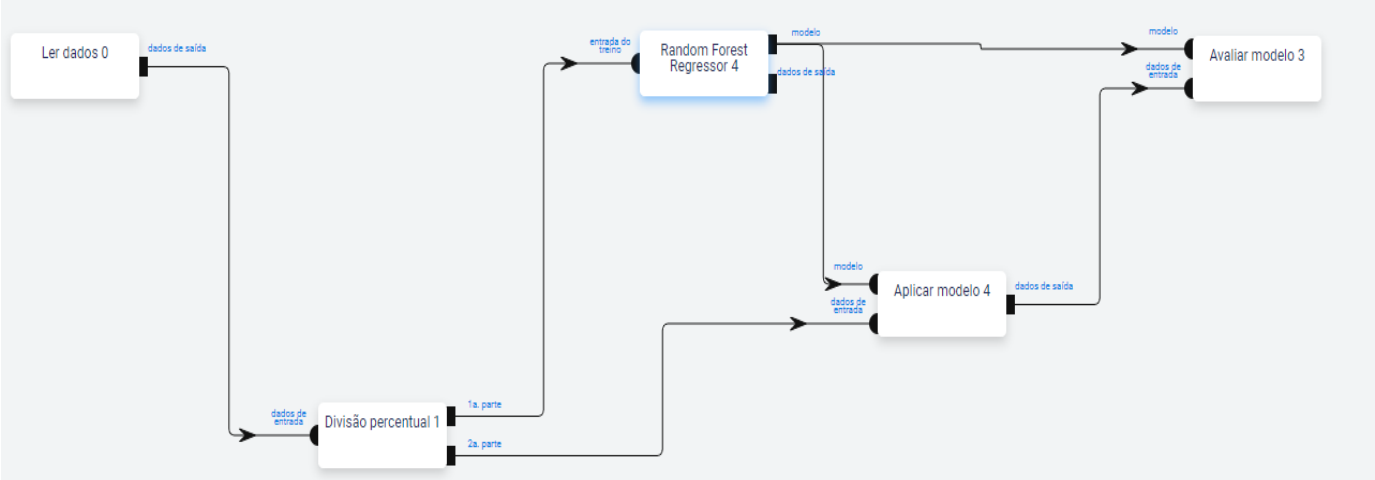
Resultados experimentais e análise

Nesta seção haverá imagens contendo ilustrações dos resultados obtidos através dos fluxos de trabalhos realizados na plataforma Lemonade, utilizando os algoritmos de regressão anteriormente descritos na seção de objetivo.

- **Algoritmo de Regressão 1 - Random Forest**

Como foi observado no trabalho relacionado, o primeiro algoritmo de regressão a ser utilizado e analisado para este projeto será o Random Forest. Dito isso, o seguinte fluxo de trabalho na plataforma Lemonade foi configurado:

Figura 4 - Fluxo de Trabalho definido para o algoritmo Random Forest



No primeiro bloco do fluxo foi feita a leitura da base de dados que foi dividida em duas partes: 70% para o treinamento e 30% para o conjunto de teste, no bloco “Divisão Percentual”. Seguindo, a parte de treinamento foi aplicada no bloco que realiza o algoritmo do Random Forest que por sua vez foi configurado com os parâmetros descritos na figura 5. Como os modelos de regressão aceitam somente características numéricas como entrada, e por isso, apenas 7 das *features* foram selecionadas como atributos previsores: *popularity, year, number_of_doors, engine_hp, engine_cylinders, highway_mpg e city_mpg.* Apesar disso, o modelo foi aplicado e avaliado pela Métrica de Coeficiente de Determinação (R2) – a mesma utilizada no trabalho anteriormente citado. Os resultados obtidos podem ser vistos na figura 6.

Figura 5 - Parâmetros utilizados pelo algoritmo Random Forest

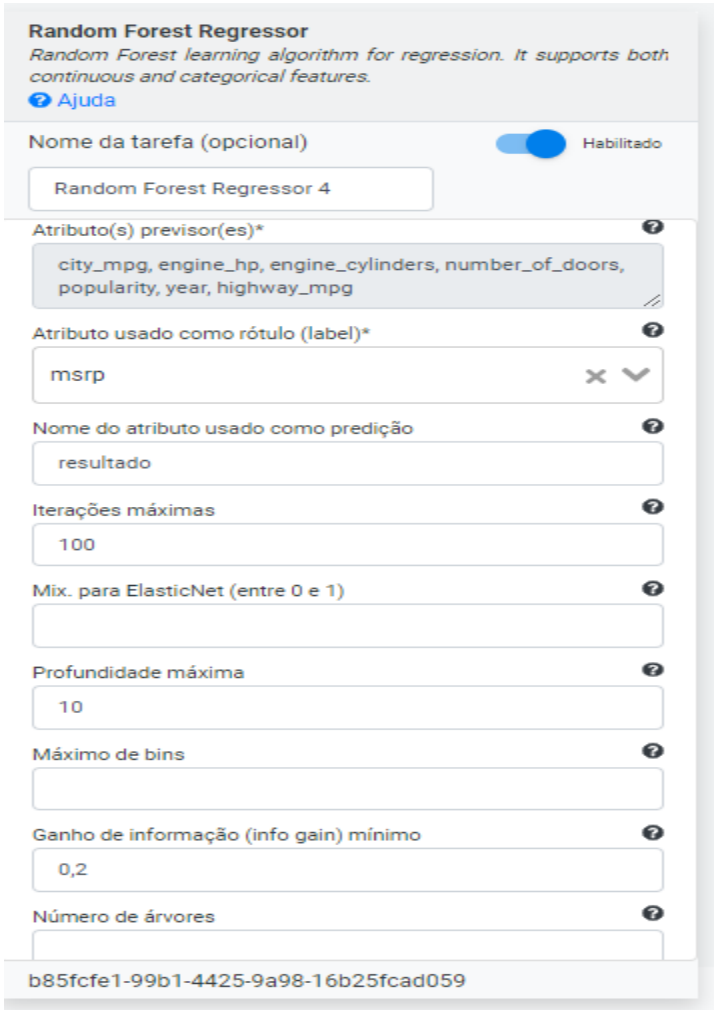


Figura 6 - Resultado R2 e importância dos atributos obtidos pelo algoritmo Random Forest

Importância dos atributos

0	city_mpg	0.10697623712487554
1	engine_hp	0.27887853075835456
2	engine_cylinders	0.39932890405433763
3	number_of_doors	0.014981061667057677
4	popularity	0.07658825418393723
5	year	0.08119219426623019
6	highway_mpg	0.0420548179452071

Com um resultado parecido com o trabalho do Gabriel, o nosso modelo obteve bons resultados. Para o Random Forest, o atributo que teve maior relevância foi o engine_cylinders, seguida do engine_hp, ou seja, para este algoritmo, os critérios mais importantes foi o motor do carro. O próximo algoritmo de regressão a ser treinado será o Gradient Boosting Tree.

- **<u>Algoritmo de Regressão 2</u> - Gradient Boosted Tree**

De forma semelhante, o fluxo de trabalho foi definido e apenas o bloco do algoritmo sofreu mudanças – agora ele aplica o Gradient Boosted Tree. Com os parâmetros (ganho de informação 0,2 e profundidade máxima 10) e treinando apenas os atributos numéricos, os resultados obtidos e a importância dos atributos podem ser vistos na figura 8:

Figura 7 - Fluxo de Trabalho definido para o algoritmo Gradient Boosted Tree

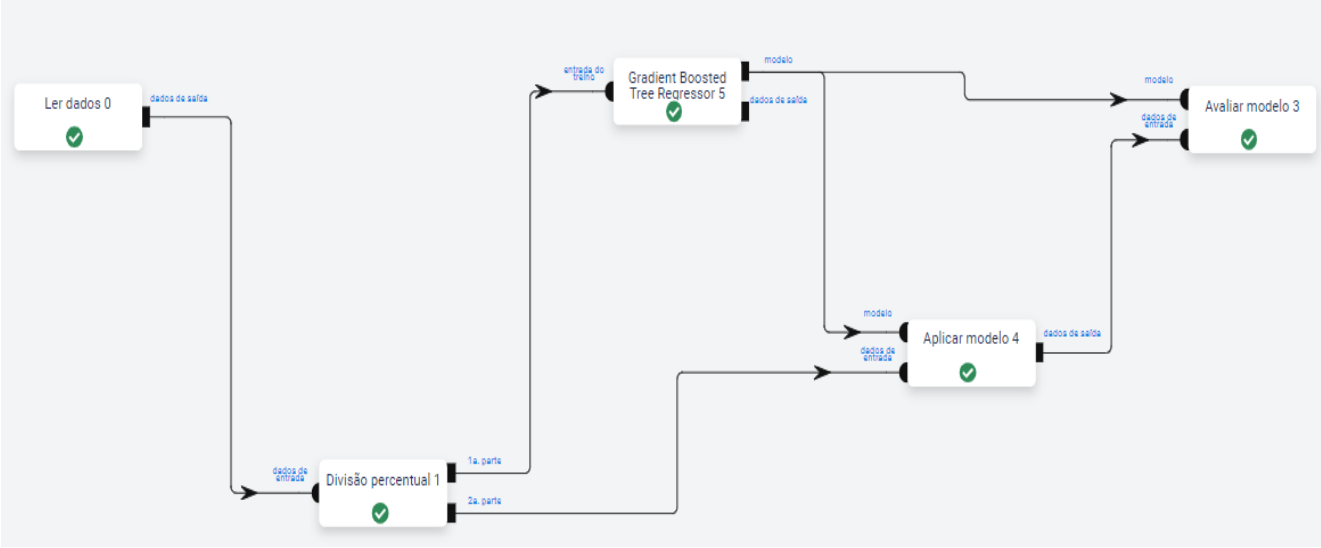


Figura 8 - Resultado R2 e importância dos atributos obtidos pelo algoritmo Gradient Boosted Tree

Importância dos atributos

0	city_mpg	0.12379898127215223
1	engine_cylinders	0.12833792607603736
2	engine_hp	0.12259608845051495
3	highway_mpg	0.11712666935751229
4	number_of_doors	0.04392704534487475
5	popularity	0.24899507212884822
6	year	0.21521821737006017

r2: 0.8296761828530156

Com uma performance um pouco pior que o algoritmo Random Forest, o algoritmo de Gradient Boosted Tree obteve um R2 de aproximadamente 0.82. Diferente do algoritmo de regressão anterior, dessa vez as características que foram importantes relevantes na predição foram a popularidade e o ano do carro. Por último, analisaremos os resultados obtidos pelo algoritmo de Regressão Linear.

- **<u>Algoritmo de Regressão 3</u> - Regressão Linear**

De maneira similar, apenas o bloco do algoritmo de regressão foi alterado para a Regressão Linear e portanto não teve muitas mudanças para o fluxo de trabalho. Em relação aos parâmetros, apenas um limite de 100 iterações foi aplicado. Executando o fluxo de trabalho foi obtido os seguintes resultados:

Figura 9 - Fluxo de Trabalho definido para o algoritmo de Regressão Linear

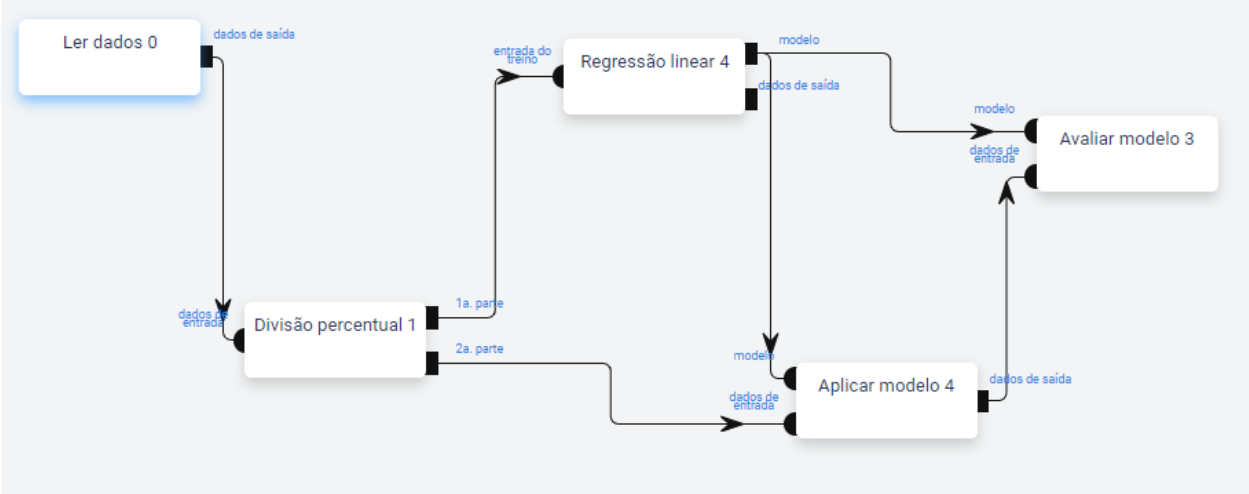


Figura 10 - Coeficiente de Determinação R2 obtido no algoritmo de Regressão Linear

O resultado obtido pelo algoritmo de Regressão Linear foi bem inferior se comparado aos demais analisados. Apesar de ser esperado um resultado pior, visto que aconteceu a mesma coisa com o trabalho relacionado estudado, uma diferença tão grande no Coeficiente de Determinação R2 não foi planejado. Acredito que, por não ter realizado nenhum tipo de pré-processamento na base de dados, o algoritmo de regressão linear não conseguiu estimar o preço de varejo de forma confiável.

Conclusões e perspectivas

Aplicar os algoritmos de Regressão de forma prática no Lemonade e avaliar os resultados foi bem útil para o meu processo de aprendizagem. Analisar os resultados e compará-los com os outros algoritmos de Regressão foi de suma importância para o entendimento e comportamento de cada algoritmo em individual. Espero que o estudo realizado neste documento seja de relevância para o trabalho prático da disciplina. Vou anexar juntamente com este pdf uma pasta com todas as imagens dos resultados obtidos, caso sejam difíceis de visualizar.