Predição do valor de carro usados

T1 - Regressão

Nesse trabalho inicial da disciplina, a ideia é se familiarizar com o processo de criação e utilização de modelos de machine learning, assim como seus pontos fortes e dificuldades. Nesse contexto, iremos trabalhar com uma boa parte dessa pipeline de trabalho. Primeiramente, conseguir os dados a serem utilizados neste estudo, nesse caso os dados já foram coletados. Na sequência, fazer análises iniciais sobre eles, separá-los em treinamento, validação e teste, estudar ainda melhor as variáveis e sua correlação inicial com a saída. Por fim, testar e treinar diversos modelos e observar seus resultados para diferentes casos desejados de aplicação.

Esse dataset contém informações referentes aos anúncios de automóveis retirados de diversos sites de anúncios da Bielorússia, através de web scraping. Ele possui atributos numéricos: ano de produção, odômetro, entre outros; atributos booleanos (binários): se alguma funcionalidade está disponível ou não, se o carro tem garantia ou não, etc; e também atributos categóricos (não-binários): marca do carro, modelo, entre outros; totalizando 30 atributos. Os dados e explicações mais completas estão disponíveis neste link do Kaggle.

Dado esse contexto, a ideia desta atividade é treinar modelos para a predição do valor de um anúncio de automóvel com base nos diversos atributos disponíveis. A variável a ser predita é o **price_usd** que é o valor em dólar do automóvel no anúncio. Idealmente, seria interessante prever o valor real do carro, mas essa informação não está disponível neste dataset, pois o valor negociado do carro pode ser menor que o anunciado ou, até mesmo, um anúncio pode nunca acarretar em uma venda. A métrica a ser utilizada será o ***mean absolute percentage error* (MAPE)**. A escolha por essa métrica deve-se ao fato dela representar o erro de uma forma mais igualitária entre preços diversos de carros, sem privilegiar carros de alto ou baixo valor.

Informações gerais a respeito do trabalho

O trabalho deverá ser entregue em um Jupyter Notebook com todos os resultados salvos (certifique-se que os resultados das células estão de fato disponíveis no notebook antes de enviar). Deve-se seguir a estrutura proposta nos próximos itens para a sequência do notebook e resultados esperados. Além disso, é necessário buscar manter o notebook bem organizado e comentado.

Estamos à disposição para tirar dúvidas a respeito dessa atividade, então não hesite em entrar em contato se ficar travado em algum ponto, é melhor solicitar ajuda antes do que gastar horas sem conseguir avançar.

OBS: Na última seção estão descritos os erros comuns e pontos que devem ser observados, sugere-se fortemente observar antes de começar e depois de finalizar o trabalho para evitar erros.

Contatar Estagiário de docência - victor.bauler@posgrad.ufsc.br

1 - Carregamento dos dados e EDA inicial

Nessa etapa, será carregado o dataset, observado possíveis problemas dele e informações gerais a respeito dos dados.

Atividades

1. Baixe os dados a partir do Kaggle e carregue-os no notebook
2. Mostre se existem dados faltantes no dataset
3. Mostre se existem dados duplicados no dataset
4. Mostre informações gerais sobre o dataset utilizando o df.info
5. Plote histogramas das colunas numéricas
6. Mostre a distribuição das variáveis binárias

DICA: Pode mostrar esses dados em gráficos mostrando a distribuição, em valores absolutos, percentuais, tabelas, como achar melhor. Uma função que pode auxilar é o df.value_counts

7. Comente a respeito do que foi encontrado nessa etapa do trabalho. Existe algum dado que aparente ter inconsistências ou problemas? Discorra a respeito.

2 - Limpeza e separação dos dados

Nessa etapa, será realizada a limpeza dos dados conforme pontos observados na seção 1 e também o conjunto de dados será dividido entre treinamento, validação e teste.

Atividades

(Limpeza)

1. Retire as linhas referentes aos dados duplicados
2. Selecione somente a parte do dataset onde o estado do carro (state) é usado (owned)
3. Selecione somente a parte do dataset onde o odômetro (odometer_value) é menor que 999.999

4. Selecione somente a parte do dataset onde o preço (price_usd) é maior ou igual a 100
5. Crie duas novas colunas booleanas a partir das colunas transmission e engine_type:
    a. Crie uma coluna que será verdadeira se a transmissão for automática e falsa se não for (is_transmission_automatic)
    b. Crie uma coluna que será verdadeira se a engine_type for diesel e falsa se não for (is_engine_type_diesel)
6. Remova as instâncias referentes a modelos de carros raros, quando o modelo de carro aparece menos de 15 vezes no dataset
7. Remova as colunas state, transmission, engine_type, duration_listed e number_of_photos

Nesse momento você deve estar com 34.247 linhas e 27 colunas, confira para observar se não cometeu nenhum erro

(Split)

8. **Divisão estratificada entre train_val e teste**
    a. Crie uma variável auxiliar (price_cat) utilizando a estratificação quantizada com 10 quantis a partir do preço (price_usd)
    b. Utilize essa função para dividir o conjunto de dados de forma estratificada, separe 20% para teste e 80% para treinamento e validação (train_val) e com o random_state igual a 42
9. **Divisão estratificada entre treinamento e validação**
    a. Crie uma nova variável auxiliar (train_val_price_cat) para fazer uma nova estratificação quantizada, mas desta vez no train_val ao invés do dataset completo
    b. Utilize essa nova variável para dividir o train_val entre treinamento (75%) e validação (25%)

As variáveis auxiliares price_cat e train_val_price_cat são somente para a divisão dos datasets e não devem ser adicionadas aos conjuntos de dados utilizados pelos modelos.

10. Separe esses conjuntos entre X_train, y_train, X_val, y_val e X_test e y_test
11. Confira o shape apresentado para cada um deles está coerente com o esperado (60% train, 20% val, 20% test)

3 - EDA de treinamento

Nessa etapa serão observadas correlações entre os atributos e a variável alvo.

Atividades

Somente utilize o conjunto de treinamento nesta etapa, não o conjunto completo.

1. **Variáveis numéricas (int e float):** Observe a correlação de Pearson
    a. Calcule a correlação a partir do conjunto de treinamento
    b. Plote esses resultados em um gráfico de barras
    c. Repita as letras **a.** e **b.,** calculando a correlação com o logaritmo do preço, ao invés do preço em si e comente sobre os resultados encontrados
2. **Variáveis booleanas (binárias):** Para cada variável booleana, trace um gráfico da distribuição das classes e outro cruzando seus valores com os da variável alvo. Uma sugestão seria um gráfico de barras horizontais para a distribuição e o sns.catplot para o cruzamento com o preço.
3. **Variáveis categóricas (não-binárias):** Para cada variável categórica, repita os mesmos passos do item anterior. Fique à vontade para limitar o número de categorias quando for muito elevado, selecionando as que são mais frequentes.

4 - Seleção de hiperparâmetros

Nessa etapa serão encontrados os melhores hiperparâmetros para os modelos testados.

Atividades

Mostre as métricas RMSE e MAPE para todos esses testes (tanto para o conjunto de treinamento quanto validação)

Será utilizada a regressão linear LinearRegression, do sklearn, sem regularização, em todos os exercícios com exceção do **5.b.** e **7.**, nos quais será utilizada a regressão com regularização L2 (Ridge).

1. **Somente com atributos numéricos**
    a. Construa uma pipeline que contenha o pré-processamento (nesse caso, somente seleciona os dados numéricos) e também contenha o modelo de Regressão linear. Realize o treinamento e mostre os resultados no conjunto de treinamento e também no conjunto de validação.
    b. Repita a letra **a.**, mas escalonando os dados com o StandardScaler ao invés de não escaloná-los. Teve alguma diferença nos resultados? Explique o porquê.

A partir desse momento, não é necessário utilizar o StandardScaler, a não ser nos dados numéricos, após o uso de Polinomial Features.

2. **Somente com atributos booleanos:** construa uma nova pipeline de pré-processamento que seleciona somente os dados booleanos e repassa esses dados para um modelo de regressão linear. Treine esse novo modelo e observe os resultados.

3. **Somente com atributos categóricos:** Faça o mesmo que na letra anterior, mas utilizando os dados categóricos ao invés dos binários. Para tal, utilize o OneHotEncoder na pipeline de processamento.
4. **Todos os atributos:** crie uma pipeline que utilizará todos os dados (numéricos, binários e categóricos). Realize o treinamento e observe os resultados.
5. **Polynomial features:** Crie uma pipeline que irá aplicar PolynomialFeatures **somente nos dados numéricos**. Após essa aplicação, faça o escalonamento desses dados utilizando o StandardScaler e junte com o resto dos dados (binários e categóricos).
    a. Primeiramente, realize um treinamento utilizando somente o PolynomialFeatures com a regressão linear (teste ao menos para os valores d=2, 5 e 10). Observe os resultados e comente o que foi obtido.
    b. Repita o treinamento e avaliação, mas dessa vez utilizando o modelo Ridge (fazendo uma busca pelo melhor valor do hiperparâmetro alpha, testando ao menos 3 valores, sendo um deles alpha=1, que é o padrão). O que esses novos resultados demonstram? (Obs: outros valores de hiperparâmetros podem ser avaliados, a tabela abaixo é apenas um exemplo.)

	degree (Poly)		
alpha (Ridge)	**2**	**5**	**10**
0.01	Teste 1	Teste 2	Teste 3
1	Teste 4	Teste 5	Teste 6
100	Teste 7	Teste 8	Teste 9

6. **TTR:** Repita o item **4.,** utilizando o TransformTargetRegressor para transformar a variável alvo (transformando através de Log e retornando através da exponenciação).
7. **TTR+Poly(+Ridge):** Faça o mesmo que no item anterior, mas utilizando o melhor modelo do item **5.** como base.
8. Comente sobre o que foi descoberto nessa etapa da atividade.

5 - Retreinamento e resultados no conjunto de teste

Retreinar modelos com os dados de treinamento e validação e os melhores hiperparâmetros encontrados na seção 4 e observar os resultados no conjunto de teste.

Atividades

Para essa etapa, concatene os dados do conjunto de treinamento e validação em X_train_val e y_train_val **(não realize uma nova divisão do conjunto de dados!)**. Esse será o conjunto de dados utilizados para treinar os modelos nessa etapa. Observe os resultados tanto para esse conjunto quanto para o conjunto de teste. Utilize os melhores hiperparâmetros encontrados na seção 5 para cada um dos casos a seguir.

1. **Regressão linear com todos os atributos (item 4.4)**
2. **Polynomial features (item 4.5)**
3. **TTR (item 4.6)**
4. **TTR+Poly(+Ridge) (item 4.7)**

6 - Análise de resultados

Observar erros e predições do melhor modelo.

Atividades

Fazer essas análises no conjunto de teste

1. Mostre um scatter plot entre os valores reais e preditos para o melhor modelo encontrado
2. Mostre a distribuição do erro das predições
3. As suas conclusões mudariam caso a métrica de avaliação considerada fosse o RMSE ao invés do MAPE? Explique.

7 - Conclusão

Comentários sobre o trabalho.

Atividades

1. Comentários gerais sobre o trabalho e sobre o que você aprendeu ao realizá-lo
2. Existe algum viés nos dados do dataset, por exemplo, por serem dados de anúncios de vendas? O que isso poderia influenciar nos modelos que estão sendo desenvolvidos? E quais limitações isso pode trazer?
3. Outros comentários relevantes (opcional)

Dicas e Erros comuns

Dicas e pontos a serem conferidos no trabalho como um todo.

- **Gerais**
  - Manter o notebook organizado e bem comentado
  - Usar markdown para comentários a respeito da atividade (facilita a organização e leitura)
  - Separar o notebook com as seções e subseções do markdown (#, ##, ###...)
  - Conferir que os resultados das células estão salvos no notebook
  - Confira se as métricas, otimizações e modelos utilizados estão corretas (adaptados para problemas de regressão nesse caso)
- **2 - Limpeza e separação dos dados**
  - Confira que os dados estão sendo separados com stratify e com auxílio do pd.qcut
- **3 - EDA de treinamento**
  - Confira que as análises estão sendo feitas no conjunto de treinamento e não no conjunto completo
- **4 - Seleção de hiperparâmetros**
  - Não observe nem utilize o conjunto de testes nesta etapa