Desafio Cientista de Dados - Lighthouse

Nome: Angela Pastorello

Introdução

Iniciamos o processo importando as bibliotecas que serão utilizadas para as a análise exploratória e a construção do modelo. Então, carregamos os datasets no formato de arquivo csv com codificação UTF-16 e separador do tipo tab. Checamos com que tipo de dados estamos trabalhando. Após isso, conferimos a presença de valores NaN e os substituímos por 0 pois se tratavam de variáveis numéricas ou fortes candidatas a serem transformadas em tipo booleano. Prosseguimos fazendo as mesmas transformações nos dois datasets para manter a uniformidade.

Retiramos variáveis redundantes ('veiculo_alienado', 'elegivel_revisao') e com baixo impacto ('versao', 'cidade_vendedor') na precificação para melhorar o desempenho do modelo.

Transformamos variáveis do tipo float em inteiras para facilitar sua uniformização, arredondamos a variável preço para ser correspondente a um valor monetário. Convertemos variáveis categóricas para binárias quando conveniente para evitar o uso de colunas com correlação duplicadas no modelo após o one-hot-encoding.

Em seguida, limpamos os dados categóricos do df_test e no df_train que não estão presentes em ambos dataframes para não criar colunas inutilizáveis no one-hot Manter um conjunto de categorias único. Isso acontece pois esses dados são ocorrências únicas que não possuem correspondente no outro dataframe e podem prejudicar o funcionamento do modelo se não forem tratados.

Normalização, padronização e codificação

A partir disso, as variáveis numéricas com intervalos pequenos foram normalizadas (‘num_fotos’, ‘num_portas’, ‘ano_de_fabricacao’, ‘ano_modelo’) e as variáveis numéricas de grande escala foram padronizadas (‘hodometro’). Essas técnicas são importantes para transformar as variáveis de forma que elas estejam em uma escala similar, e são necessárias para beneficiar o desempenho do modelo.

Depois foram feitas as análises exploratórias explicadas no final do documento e partimos para a codificação das variáveis categóricas ('marca', 'modelo', 'cambio', 'tipo', 'blindado', 'cor', 'vendedor_PF', 'estado_vendedor', 'anunciante', 'entrega_delivery', 'troca', 'dono_aceita_troca', 'veiculo_único_dono', 'revisoes_concessionaria', 'ipva_pago', 'veiculo_licenciado', 'garantia_de_fábrica', 'revisoes_dentro_agenda'). O one-hot-encoding é uma técnica de transformação de variáveis categóricas em representação numérica binária, onde cada categoria se torna uma nova coluna com valores 0 ou 1. Para aplicar o one-hot-encoding corretamente, é importante garantir que a codificação seja consistente entre os conjuntos de treinamento e teste como fizemos nos passos anteriores. Assim foram criadas as versões codificadas dos dataframes de treinamento e teste.

Verificação e modelo

Foi feita uma matriz de correlação para avaliar o impacto das variáveis no preço e partimos para o treinamento do modelo. Inicialmente foi feito um modelo teste separando dados do dataframe de treinamento codificado para que usá-los teste e analisarmos a eficácia do modelo. Foi obtido um R^2 de 0.6845 com esse modelo. O resultado poderia ser melhor, mas ainda é um resultado aceitável para a quantidade de variáveis processados e ao grande desempenho do modelo requer.

Agora utilizando nosso modelo final, nomeamos a variável a ser prevista ‘preco’ como target ao separá-la do dataset de treinamento codificado e as variáveis de treino como df_train para a preparação do modelo em que então foram aplicadas. Com o

modelo preparado, o dataframe de teste codificado foi aplicado ao modelo para prever a variável ‘preco’.

Estamos resolvendo uma regressão e o modelo escolhido foi um Random Forest. é um tipo de modelo muito utilizado para análises preditivas de preço e que permite aproveitar os dados disponíveis

Regressão Random Forest

O modelo Random Forest pode lidar com conjuntos de dados grandes, alta dimensionalidade e uma variedade de tipos de dados incluindo dados numéricos e categóricos, tornando-o adequado para conjuntos de dados complexos. Além disso, é menos sensível a valores atípicos e ruído nos dados em comparação com alguns outros algoritmos.

A estrutura de árvore do modelo e a agregação de várias árvores (floresta) reduzem o risco de overfitting e torna o modelo capaz de capturar relações não lineares entre as variáveis de entrada e a variável de saída.

O modelo Random Forest pode ser mais difícil de interpretar do que modelos mais simples. Devido à combinação de várias árvores, pode ser complicado entender como cada variável contribui individualmente para a predição. Também possui hiperparâmetros que precisam ser ajustados para o melhoramento do modelo, como o número de árvores na floresta, a profundidade máxima das árvores, entre outros. Em conjuntos de dados muito grandes ou com muitas variáveis, o treinamento do modelo Random Forest pode exigir mais tempo e recursos computacionais em comparação com algoritmos mais simples

Análise Exploratória

Através da análise exploratória tentamos descobrir o melhor estado para se vender ou comprar um carro, de acordo com as características preteridas. Para

delimitar as marcas populares, foram selecionadas as marcas mais procuradas no ano passado e as marcas dos modelos mais comprados neste ano.

a) O melhor estado para se vender um carro de marca popular dependerá do que o vendedor está procurando. Considerando que para o vendedor seria vantajoso considerar como objetivo conseguir vender pelo preço mais alto ou ter maior facilidade em venda, podemos analisar as seguintes possibilidades. Podemos considerar valor total arrecadado com vendas de carro no estado, maior quantidade de carros vendidos e maior preço médio por carro vendido. Sendo assim, o estado com valor total arrecadado com vendas mais alto e maior quantidade de carros vendidos é São Paulo, enquanto o estado com maior preço médio por carro vendido é o Piauí seguido pelo Sergipe.

b) Considerando que o melhor estado para se comprar uma picape com transmissão automática seja o que o comprador encontre o menor preço médio por carro vendido, a melhor oferta estaria na Paraíba.

c) Ao procurar o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica, o comprador pode levar em consideração o menor preço médio por carro vendido. Nesse caso, é a Paraíba seguida do Pará e do Amazonas.