Lighthouse 2023 - Indicium

Desafio Lighthouse para vaga de Cientista de Dados (jul/2023).

Candidato: Renato Massamitsu Zama Inomata.

Linkedin: https://www.linkedin.com/in/renato-inomata/

Github: https://github.com/renatoinomata/

Link para este repositório: https://github.com/renatoinomata/indicium-data-science

Estruturação do trabalho

Neste README está um resumo das análises e resultados obtidos neste trabalho.

O conteúdo foi dividido em dois notebooks, um para a Análise Exploratória dos Dados (EDA)e outro para a parte de Machine Learning (ML).

Para a EDA, o conteúdo foi dividido da seguinte maneira:

- I) Introdução, onde importamos os dados e pacotes necessários, bem como visualizaremos uma pequena porção dos dados;
- II) Análise das features, onde fizemos uma análise mais minunciosa de cada variável, e também iremos comparar suas relações com a variável *target*;
- III) Perguntas propostas, onde respondemos às perguntas de negócio propostas pela Incidium para este trabalho;
- IV) Outras hipóteses, onde criamos nossas próprias hipóteses e tentaremos respondê-las utilizando os dados disponíveis;
- V) Conclusões, onde reunimos as informações relevantes levantadas acerca do conjunto de dados.

Para o ML, a estrutura foi:

- I) Introdução, onde importamos as bibliotecas e dados necessários;
- II) Modelos, onde conduzimos os testes para as predições, tanto no conjunto de dados quanto nos algoritmos de ML;
- III) Seleção do modelo e tuning de hiperparâmetros, onde definimos o modelo final a ser utilizado e tunamos seus hiperparâmetros;
- IV) Exportação do modelo e dos resultados, onde obtivemos os resultados para o conjunto de teste e exportamos para um .csv e o modelo para um .joblib;
- V) Conclusões, onde discorremos sobre os resultados encontrados e algumas sugestões de trabalhos futuros.

O problema

Você foi alocado(a) em um time da Indicium que está trabalhando atualmente junto a um cliente que o core business é compra e venda de veículos usados. Essa empresa está com dificuldades na área de revenda dos automóveis usados em seu catálogo.

Para resolver esse problema, a empresa comprou uma base de dados de um marketplace de compra e venda para entender melhor o mercado nacional, de forma a conseguir precificar o seu catálogo de forma mais competitiva e assim recuperar o mau desempenho neste setor.

Seu objetivo é analisar os dados para responder às perguntas de negócios feitas pelo cliente e criar um modelo preditivo que precifique os carros do cliente de forma que eles fiquem o mais próximos dos valores de mercado.

Repositório

Os arquivos deste repositório estão organizados da seguinte maneira:

datasets

O repositório possui uma pasta `dataset` com os conjuntos de dados utilizados. Dentre eles:

- `cars_train.csv`: dataset para treinamento composto por 29584 linhas, 28 colunas de informação (features) e a variável a ser prevista (“preco”);
- `cars_test.csv`: dataset para teste composto por 9862 linhas e 28 colunas, sendo que este dataset não possui a coluna “preco”;
- `aliquotas_ipva.xslx`: arquivo com os dados das aliquotas de IPVA dos estados presentes no conjunto de treino. Fonte: idinheiro;
- `populacao_pib.xlsx`: população e PIB per capita dos estados presentes no conjunto de treino. Fontes: Wikipédia - População, Wikipédia - PIB per capita;
- `populacao_municipios.xlsx`: população por município. Fonte: IBGE.

enunciado

Pasta com o enunciado em pdf do presente trabalho.

model

Pasta com o modelo gerado ao final do trabalho em .joblib e os resultados obtidos `predicted.csv`.

analises

Os notebooks para a Análise Exploratória dos Dados (EDA) e para os modelos de Machine Learning estão na própria pasta principal do repositório.

Também estão disponibilizadas versões em HTML dos notebooks.

- `01_cars_eda.ipynb`: arquivo com a análise exploratória de dados
- `02_cars_ml.ipynb`: arquivo com os testes realizados para obter o modelo de Machine Learning.

Conclusões

Nesta seção apresentaremos um breve resumo das conclusões de cada etapa do desafio.

Análise Exploratória dos Dados

O conjunto de dados parece apresentar inconsistências, principalmente nas colunas `versao`, `cambio` e `tipo`. Essas inconsistências são percebidas porque a descrição das versões dos veículos não condiz com as demais features. Além disso, esperava-se que as features `ano_de_fabricacao` e `ano_modelo` estivessem com no máximo um ano de diferença entre elas, o que não foi observado para os dados apresentados.

Não existe correlação forte entre as features do conjunto de dados e o *target* preço. O maior valor encontrado foi uma correlação negativa com o hodômetro, de 0.36.

Das hipóteses elaboradas, notou-se que:

- Os vendedores PJ apresentaram comportamento diferenciado com relação aos vendedores PF. Apenas vendedores PJ fazem delivery, e também são os únicos que possuem veículos trocados anteriormente. Todos os vendedores PJ aceitam trocas. Em geral, vendedores PJ anunciam com mais fotos do veículo, não pagam IPVA, não realizam revisões nas concessionárias, não estão com as revisões em dia e não apresentam garantia de fábrica;
- O pagamento do IPVA não está ligado com a alíquota do estado;
- O número de vendas está fortemente correlacionado à população de cada estado. O ticket médio apresenta correlações menores, tanto para o PIB per capita quanto para a população do estado. Para as cidades, o número de vendas está fortemente correlacionado à sua população, porém o valor de cada veículo não apresenta correlação significativa.

Machine Learning

Os modelos baseados em árvores de decisão apresentaram uma performance superior aos modelos baseados em regressão linear.

Uma dificuldade que estes modelos poderão encontrar são com o tratamento de outliers, visto que o RandomForest não é recomendado para extrapolação de valores.

Foram utilizadas algumas técnicas de NLP (Natural Language Processing) sendo gerados vários tokens para as descrições dos carros, e espera-se que, caso novos carros sejam adicionados, o modelo consiga absorver os textos das descrições dos veículos e possa interpretar e atribuir um valor a ele com base em suas características.

Algumas outras ideias para realizarmos predições mais assertivas seriam:

- Consultar valores da tabela FIPE: é bastante comum a utilização dessa tabela para negociação de veículos, e é inclusive usada como critério para valor venal no cálculo do IPVA.

- Confrontar as vendas com outras características demográficas das cidades/estados: utilizou-se o PIB e a população para indicar correlação nas vendas, porém percebeu-se apenas correlação na quantidade e não nos preços. Número de veículos da cidade, por exemplo, poderia ser uma informação relevante para a análise dos dados.

- Fazer uma limpeza no dataset mais profunda, e principalmente entender como os dados foram obtidos: durante a exploração de dados percebeu-se que algumas colunas não condiziam com as demais, porém como não sabíamos a origem dos dados é difícil de se afirmar quais dados realmente estavam incorretos. Exemplo, um veículo poderia apresentar em sua descrição 2P e no seu número de portas apresentar 4, mas qual dessas informações estaria realmente correta?

- Aplicar outros tipos de modelos: neste trabalho concentramos em apenas métodos de regressão linear e o RandomForest. Existem vários outros métodos como SVMs, Gradient Boosting, Redes Neurais, etc. que poderiam apresentar resultados diferentes dos obtidos aqui.