Desafio Cientista de Dados

Introdução

Olá candidato(a), o objetivo deste desafio é testar os seus conhecimentos sobre a resolução de problemas de negócios, análise de dados e aplicação de modelos preditivos. Queremos testar seus conhecimentos dos conceitos estatísticos de modelos preditivos, criatividade na resolução de problemas e aplicação de modelos básicos de machine learning. É importante deixar claro que não existe resposta certa e que o que nos interessa é sua capacidade de descrever e justificar os passos utilizados na resolução do problema.

Desafio

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.

Para isso são fornecidos dois *datasets*:

1. Um *dataset* para treinamento chamado *cars_training* composto por 29584 linhas, 28 colunas de informação (*features)* e a variável a ser prevista (“preco”).
2. Um segundo *dataset* para teste chamado de *cars_test* composto por 9862 linhas e 28 colunas, sendo que este *dataset* não possui a coluna “preco”.

Seu objetivo é prever a coluna "preco" coluna a partir dos dados enviados e nos enviar para avaliação dos resultados.

Você poderá encontrar em anexo um dicionário dos dados.

Entregas

1. Utilizando as variáveis (*features*), faça um relatório com uma análise das principais estatísticas da base de dados. Descreva graficamente essas variáveis (*features*), apresentando as suas principais estatísticas descritivas. Comente o porquê da escolha destas estatísticas e o que elas nos informam..
2. Faça uma EDA. Nesta EDA, crie e responda 3 hipóteses de negócio. Além disso, responda também às seguintes perguntas de negócio:
    a. Qual o melhor estado cadastrado na base de dados para se vender um carro de marca popular e por quê?
    b. Qual o melhor estado para se comprar uma picape com transmissão automática e por quê?
    c. Qual o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica e por quê?
3. Explique como você faria a previsão do **preço** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

4. Envie o resultado final do modelo em uma planilha com apenas duas colunas (id, preco).
5. A entrega deve ser feita através de um repositório de código público que contenha:
    a. README explicando como instalar e executar o projeto
    b. Arquivo de requisitos com todos os pacotes utilizados e suas versões
    c. Relatórios das análises estatísticas e EDA em PDF, Jupyter Notebook ou semelhante conforme passo 1 e 2.
    d. Códigos de modelagem utilizados no passo 3.
    e. Arquivo final com o nome *predicted.csv* conforme passo 4 acima.

    Todos os códigos produzidos devem seguir as boas práticas de codificação.

Prazo

Você tem até **7 dias corridos** para a entrega, contados a partir do recebimento deste desafio.

Envie o seu relatório dentro da sua data limite para o email:
selecao.lighthouse@indicium.tech

O arquivo de entrega deve ser nomeado como: **LH_CD_SEUNOME**

Bom trabalho!

Dicionário dos dados

A base de dados de treinamento contém 29 colunas, sendo que 28 delas são colunas de *features* e uma coluna *target*. Seus nomes são auto-explicativos, mas, caso haja alguma dúvida, a descrição das colunas é:

- **id:** Contém o identificador único dos veículos cadastrados na base de dados
- **num_fotos:** contém a quantidade de fotos que o anuncio do veículo contém

- **marca:** Contém a marca do veículo anunciado
- **modelo:** Contém o modelo do veículo anunciado
- **versao:** Contém as descrições da versão do veículo anunciando. Sua cilindrada, quantidade de válvulas, se é flex ou não, etc.
- **ano_de_fabricacao:** Contém o ano de fabricação do veículo anunciado
- **ano_modelo:** Contém o modelo do ano de fabricação do veículo anunciado
- **hodometro:** Contém o valor registrado no hodômetro do veículo anunciado
- **cambio:** Contém o tipo de câmbio do veículo anunciado
- **num_portas:** Contém a quantidade de portas do veículo anunciado
- **tipo:** Contém o tipo do veículo anunciado. Se ele é sedã, hatch, esportivo, etc.
- **blindado:** Contém informação se o veículo anunciado é blindado ou não
- **cor:** Contém a cor do veículo anunciado
- **tipo_vendedor:** Contém informações sobre o tipo do vendedor do veículo anunciado. Se é pessoa física (PF) ou se é pessoa jurídica (PJ)
- **cidade_vendedor:** Contém a cidade em que vendedor do veículo anunciado reside
- **estado_vendedor:** Contém o estado em que vendedor do veículo anunciado reside
- **anunciante:** Contém o tipo de anunciante do vendedor do veículo anunciado. Se ele é pessoa física, loja, concessionário, etc
- **entrega_delivery:** Contém informações se o vendedor faz ou não delivery do veículo anunciado
- **troca:** Contém informações o veículo anunciado já foi trocado anteriormente
- **elegivel_revisao:** Contém informações se o veículo anunciado precisa ou não de revisão
- **dono_aceita_troca:** Contém informações se o vendedor aceita ou não realizar uma troca com o veículo anunciado
- **veiculo_único_dono:** Contém informações o veículo anunciado é de um único dono
- **revisoes_concessionaria:** Contém informações se o veículo anunciado teve suas revisões feitas em concessionárias

- **ipva_pago:** Contém informações se o veículo anunciado está com o IPVA pago ou não
- **veiculo_licenciado:** Contém informações se o veículo anunciado está com o licenciamento pago ou não
- **garantia_de_fábrica:** Contém informações o veículo anunciado possui garantia de fábrica ou não
- **revisoes_dentro_agenda:** Contém informações se as revisões feitas do veículo anunciado foram realizadas dentro da agenda prevista
- **veiculo_alienado:** Contém informações se o veículo anunciado está alienado ou não
- **preco (*target*):** Contém as informações do preço do veículo anunciado