DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

BOAS VINDAS

Olá! Que bom que você veio!

Este é o segundo desafio assinado da trilha de desafios que você terá ao longo da jornada. Legal, né?

Em um mundo que muda cada vez mais rápido, é essencial encontrar uma **SOLUÇÃO QUE GARANTA RESKILLING & UPSKILLING CONSTANTES**.

O objetivo dos desafios é cultivar **interface responsiva** entre os avanços tecnológicos, transformações de mercado e mudanças comportamentais dos consumidores, preparando a empresa para emergência de uma nova cultura orientada por dados e preparada para lidar com o novo paradigma trazido pelas tecnologias e metodologias da **Ciência de Dados** e Inteligência Artificial.

ELEMENTOS DOS DESAFIOS

- Contexto de negócio.
- Problema a ser resolvido.
- Macro etapas de uma possível resolução.
- Toolbox (técnicas e algoritmos que podem ser aplicados para resolver o desafio).
- Base de dados.
- Extra (questão extra a ser debatida, por exemplo: estrutura da solução, ética da aplicação, privacidade dos usuários, etc).
- Aula síncrona de resolução e discussão.
- **As soluções elaboradas podem ser utilizadas para enriquecer seus portfólios.**

#FICAADICA

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

Por isso, incentivamos que sempre tente resolver o desafio. Teremos um momento para discussão com a expert sobre os pontos de dificuldades que surgirem, não se preocupe! O importante será participar deste momento!

Caso o desafio aborde um assunto que você não conhece ainda, não se preocupe! Ao longo dos dias você terá aulas com as experts, antes do dia da resolução, que acompanham a complexidade ferramental do case.

Você tem liberdade para fazer o desafio de forma individual, mas também poderá se juntar com o colega caso queira. Pode sair muita coisa legal daí! :)

Vamos começar?

AVANÇAR

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

DESAFIO REGRESSÃO

POR MARCUS OLIVEIRA DA SILVA

VÍDEO CONTEXTUAL: **https://www.youtube.com/watch?v=NC1P4AFTj54**

CONTEXTO ANALÍTICO

Nesse desafio, você deve resolver um case de precificação de imóveis. Esse desafio foi construído em parceria entre a Tera e o QuintoAndar, onde o objetivo é simular um projeto de machine learning com características semelhantes ao que ocorre no dia a dia da empresa.

Imagine-se na seguinte situação: a área de marketing do QuintoAndar quer montar uma calculadora de preço (como esta [aqui: **https://mkt.quintoandar.com.br/quanto-cobrar-de-aluguel/**), e nesse projeto, os analistas negociais e corretores querem, também, entender as principais variáveis e características chaves que influenciam no valor de venda do imóvel (Ex: quantificar o impacto do aumento da área do imóvel no preço, ou quantificar o impacto de ter piscina, ou não no preço). Você é o cientista de dados que atuará na resolução desse case.

Para tanto, existem dois objetivos principais:

OBJETIVO 1 - INTERPRETABILIDADE

construir uma regressão linear simples, com poucas variáveis importantes, visando gerar insights para corretores e proprietários no quesito precificação dos imóveis. Ou seja, o foco será na interpretação dos coeficientes (ex: se aumentar a área do imóvel em uma unidade isso irá aumentar em Y o preço deste imóvel).

OBJETIVO 2 - PREDIÇÃO

construir um modelo com alto poder preditivo, com mais variáveis, visando um bom desempenho e com o intuito de ser usado em uma página web como a calculadora de preço. Note que, em uma situação real, um alto erro de inferência pode gerar grande insatisfação em um proprietário de imóvel,

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

tenha uma interpretação mais difícil.

Vamos lá?

AVANÇAR

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

BASE DE DADOS

O conjunto de dados descreve a venda de propriedades residenciais individuais de uma cidade americana, de 2006 a 2010. O conjunto de dados contém **2.930 observações** e um grande número de features (23 nominais, 23 ordinais, 14 discretas e 20 contínuas) envolvidas na avaliação do valor dos imóveis, ou seja, são **80 variáveis explicativas**.

Geralmente, as **20 features contínuas** estão relacionadas com várias dimensões de área para cada imóvel. Além do típico tamanho do lote e da metragem quadrada total da área habitável, outras variáveis mais específicas são quantificadas no conjunto de dados. Medidas da área do porão, área da sala de estar e até mesmo das varandas estão presentes e divididas em categorias individuais com base na qualidade e no tipo. O grande número de variáveis contínuas neste conjunto de dados deve fornecer muitas oportunidades de discretização e construção de novas features.

As **14 features discretas** normalmente quantificam o número de itens que existem na casa. Como, por exemplo: o número de cozinhas, quartos e banheiro discretizados por sua localização (porão ou acima do térreo). Adicionalmente, a capacidade da garagem e as datas de construção/reforma também estão registadas.

As **features nominais** normalmente identificam várias categorias de moradias, garagens, materiais e condições ambientais, enquanto as variáveis ordinais normalmente classificam vários itens na propriedade.

Segue a documentação do dataset com o glossário de todas as variáveis:

Base de dados:
[https://drive.google.com/file/d/1THulRo680Wqf5MPjMvfNQ1ouhCeMvQaX/view?usp=sharing]
(https://drive.google.com/file/d/1THulRo680Wqf5MPjMvfNQ1ouhCeMvQaX/view?usp=sharing)

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

xBI/view?usp=sharing]
(https://drive.google.com/file/d/1B3AJBLhDSyNghSVrVnCIr0WudSwQqxBI/view?usp=sharing).

AVANÇAR

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

ROTEIRO PARA O DESAFIO

ANÁLISE EXPLORATÓRIA

Você pode tentar o seguinte:

- Verificar a distribuição da variável de interesse (valor de venda)

- Contar o número de valores faltantes

- Verificar a matriz de correlação entre as features continuas

- Scatterplots são úteis para visualizar duas variáveis continuas

- Plotar a distribuição do valor de venda (histogramas ou boxplots) para as diferentes variáveis - categóricas

- Ao final, escreva um pouco sobre o que você conseguiu entender, extrapolar e interpretar a partir da análise exploratória

O número de variáveis é alto, então é importante ser criterioso na montagem dos gráficos exploratórios. Use sua intuição e raciocínio crítico para mostrar as variáveis e encontrar a informação que importa para prosseguir com sua modelagem.

PRÉ-PROCESSAMENTO, LIMPEZA DOS DADOS E CONSTRUÇÃO DE FEATURES

Você pode tentar o seguinte:

- Dropar colunas(features) com muitos valores faltantes

- Buscar algum erro de preenchimento no dataset

- Tente criar features (Exemplo: dividir o número de quartos pela área, elevar a área ao quadrado)

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

OBJETIVO 1 - INTERPRETABILIDADE USANDO UMA REGRESSÃO LINEAR

- Utilize apenas exemplos onde a variável **SaleCondition** é igual à **Normal** (isso ajuda a diminuir o ruído no dado e levar a uma melhor interpretação). Crie um novo dataset aplicando esse filtro (esse dataset será usado apenas nessa etapa)

- Aplique uma **transformação logarítmica** na sua variável de interesse (**SalePrice**)

- Selecione **6 features** para o seu modelo: onde pelo menos uma das features é uma feature construída com os valores de área do imóvel

- As outras features devem estar relacionadas ao estado de conservação da casa, suas amenidades ou instalações internas

- Trate os valores faltantes

- Aplique as transformações nas variáveis categóricas que você julgar necessárias (One hot encoding, ordinal encoding, etc…)

- Utilize a lib statsmodel para fitar a regressão linear; use a função summary para conseguir interpretar os coeficientes

- Importante lembrar que a interpretação do coeficiente muda ao aplicar uma transformação lograritmica (ver mais aqui: **http://www.cazaar.com/ta/econ113/interpreting-beta**, ou aqui: **https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf**)

- Verifique a distribuição dos resíduos da regressão linear, e quais as implicações do resultado obtido

- Reporte o R^2 dessa regressão, e a sua interpretação desse resultado

- As features não podem ter alta correlação (utilizar o EDA feito previamente para encontrar as features que você julgue relevantes)

- Verifique se os pressupostos da regressão linear estão sendo atendidos (dicas: [aqui](**https://lamfo-unb.github.io/2019/04/13/Diagnostico-em-**

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

Regression4.html))

- Em um breve sumário, discorra sobre a interpretação dos coeficientes obtidos pela regressão linear (sobre interpretação ver [aqui] (**https://christophm.github.io/interpretable-ml-book/limo.html#interpretation**))

Constraints do modelo: essa regressão linear deve ter um **R^2 mínimo de 0.85** e conter exatamente 6 features/variáveis. Todas as features devem ter um P-value maior do que 0.95 e não podem ser colineares. Não é necessário separar esse dataset em treino e teste, já que o foco nessa etapa é na interpretação dos coeficientes de uma regressão linear.

OBJETIVO 2 - PODER PREDITIVO, REGRESSÃO VIA RANDOM FOREST

A ideia dessa segunda parte é treinar um modelo mais robusto visando o poder preditivo e a obtenção de um modelo para uso em produção (uso real em uma aplicação web)

- Transforme o sua variável de interesse usando uma transformação logarítmica (ex: y_log=np.log(y))

- Separe seu dataset original em **treino**, **teste**, **validação** (ver instruções de como fazer a separação abaixo)

- Impute os valores faltantes das variáveis **numéricas com a mediana** e os valores faltantes das **variáveis categóricas com a moda**, os imputers devem ser fitados usando o dataset de treino para depois serem aplicadas nos datasets de validação e teste, isso evitará data leakage (Dica: usar os **simpleimputer** do sklearn)

- Aplique as transformações nas variáveis categóricas que você julgar necessárias (One hot encoding, ordinal encoding e etc…), as transformações também devem ser fitadas usando o dataset de treino para depois serem aplicadas nos datasets de validação e teste, isso evitará data leakage (Dica: usar os **transformers** do sklearn)

- **Treine uma árvore inicial (modelo baseline)**, usando todas as features e sem mexer nos hiperparâmetros do modelo (usando n_estimators = 150),

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

- Tente tunar o seu random forest. **Teste diferentes hiperparâmetros**, veja as instruções abaixo, use a documentação do sklearn para entender os hiperparâmetros que você for testar

- Compute a **importância das features** no dataset de validação (usar permutation_importance do sklearn), usando a importância das features remova do seu treinamento as features menos importantes para que o seu modelo tenha no máximo 40 features, verifique novamente a performance com esse número reduzido de features (isso pode melhorar a performance e a velocidade do seu modelo)

- Adicione um breve texto com **sua interpretação** em relação à **importância das features**

- Finalmente, compute as **métricas de avaliação no dataset de teste** para obter o proxy de performance do seu modelo em um ambiente em produção (ambiente real online).

- Adicione uma **conclusão** para fechar o seu case

Dica: sempre que você for avaliar o seu modelo, você deve reverter suas predições da escala log para a escala normal usando uma função exponencial (ex: y_pred = np.exp(y_pred)),

Separação dos dados em treinamento e validação: os dados devem ser separados em treino, validação e teste, na fase de exploração e modelagem você pode avaliar o modelo usando o dataset de validação para evitar overfitting, e depois, com estudo fechado aplicar as métricas de avaliação no dataset de teste (simulando a performance em exemplos nunca vistos). Para esse caso você deve separar os datasets usando a função **train_test_split** do sklearn, usando como **random state o número 42**:

- Primeiro use a função função train_test_split para separar 70% para treino e 30% para validação e teste

- Segundo, aplique novamente essa função para quebrar esses 30% em dois datasets, sendo 50% para teste e 50% para validação. Assim obtendo 70% para treino, 15% para validação e 15% para teste

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

melhorar. Em seguida, usando o número de árvores da primeira etapa, tente alguns valores de max_features e escolha aquele que forneça a melhor métrica. Finalmente, usando o melhor max_features, execute min_samples_leaf de 1 até 15, novamente escolhendo o melhor.

- n_estimators (10, 30, 50, 70, 100, 150, 200)

- max_features ('sqrt', 0.1 até 0.6)

- min_samples_leaf(1 até 15)

Avaliação do modelo de regressão: Para fazer a avaliação do seu modelo você deve aplicar métricas de avaliação no dataset de validação (e no final do estudo no dataset de teste), as seguintes métricas são comuns em modelos de regressão:

- **R^2**: pense nesse score como uma medida do desempenho do nosso modelo em comparação com um modelo trivial que retorna sempre a média para qualquer previsão solicitada. (o valor de 1.0 representa um modelo perfeito, já um valor de 0.0 representa um modelo equivalente a um modelo aleatório)

- **Valor absoluto médio (MAE)**: que é apenas a diferença absoluta média entre os valores previstos e verdadeiros. O valor absoluto evita que desvios negativos e positivos se cancelem.

- Em vez de tomar o valor absoluto, poderíamos elevar ao quadrado as diferenças, dando-nos o **erro quadrático médio (MSE)**. Elevar a diferença também tem o efeito de enfatizar quaisquer previsões que estejam muito longe de seus verdadeiros valores.

- Para ignorar algumas previsões significativamente desviantes (outliers), é melhor usar o MAE no lugar MSE. Tudo depende do que você está buscando

- Como as unidades do MSE são o quadrado das unidades da variável de interesse, é útil usar a **raiz do erro quadrático médio** (RMSE) como métrica de avaliação

---

AVANÇAR

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

TOOLBOX E REFERÊNCIAS

OBJETIVO 1

- **Modelo**: [Statsmodels] (https://www.statsmodels.org/devel/example_formulas.html)

- **Visualização**: [Seaborn](https://seaborn.pydata.org/)

- **Manipulação do dataset**: Pandas e Numpy (use np.log para realizar transformações logarítmicas)

OBJETIVO 2

- **Modelo**: [Sklearn](https://scikit-learn.org/stable/)

- [RandomForest](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

- [Imputar valores faltantes](https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html)

- Enconders de categóricas: [OrdinalEnconder](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html), [OneHotEncoder](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)

- [Importância das features](https://scikit-learn.org/stable/modules/permutation_importance.html#:~:text=The%20permutation%20feature%20importance%20is,model%20depends%20on%20the%20feature.)

- [Métricas de avaliação](https://scikit-learn.org/stable/modules/model_evaluation.html) (R^2, MAE, MSE, RMSE)

DESAFIO DE MERCADO: REGRESSÃO (QUINTOANDAR)

- **Manipulação do dataset**: Pandas e Numpy (use np.log para realizar transformações logarítmicas)

REFERÊNCIAS

[https://mlbook.explained.ai/bulldozer-intro.html]
(https://mlbook.explained.ai/bulldozer-intro.html

https://christophm.github.io/interpretable-ml-book/limo.html

https://christophm.github.io/interpretable-ml-book/limo.html

AULA AVALIADA COM SUCESSO!

Obrigada pelo seu feedback!

REFAZER

VOLTAR PARA O CURSO