Fundação Getulio Vargas

Aprendizado Supervisionado

Prof. Sergio Polimante

FGV

"Essentially, all models are wrong, but some are useful."

— George E. P. Box

Projeto Final: Aprimoramento de Modelo de Previsão de Preços de Imóveis na Califórnia

Objetivo

Aplicar os conceitos aprendidos em aula para melhorar as métricas de qualidade do modelo de regressão linear que prevê o valor mediano das casas na Califórnia, utilizando o conjunto de dados apresentado no notebook de referência.

Atividade

O aluno utilizará os notebooks fornecidos em aula como referência para desenvolver seu modelo de machine learning de regressão linear.

É esperado que o aluno aplique as técnicas vistas em aula para desenvolvimento do modelo, com o objetivo de melhorar as métricas de avaliação. Utilizaremos as métricas RMSE e MAPE para fazer a avaliação do modelo.

No desenvolvimento do projeto, o aluno pode utilizar as seguintes técnicas vistas em aulas (não é obrigatório usar todas):

1. Exploração e Análise de Dados:
    - Exploração dos dados, utilizando gráficos e análises estatísticas básicas para entender como as variáveis estão distribuídas e se relacionam entre si. Isso inclui o uso do pandas para examinar os dados (média, mediana, contagens), criar visualizações (histogramas, gráficos de dispersão, boxplots) e identificar possíveis problemas como valores ausentes ou extremos que possam afetar o modelo.
    - Visualizações relevantes para entender as relações entre variáveis
    - Identificação de padrões, tendências e anomalias nos dados
2. Pré-processamento dos Dados:
    - Tratamento de valores ausentes (caso existam)

- Identificação e tratamento adequado de outliers
- Transformações de variáveis para melhorar a linearidade

3. Engenharia de Features:
    - Criação de novas variáveis relevantes, através da combinação de características existentes (por exemplo, calcular a proporção de quartos por residência dividindo AveRooms por AveBedrms)
    - Seleção de features importantes
    - Análise de multicolinearidade e remoção de variáveis redundantes
4. Modelagem com Regularização:
    - Implementação de regressão linear com regularização:
        - Ridge Regression (L2)
        - Lasso Regression (L1)
        - ElasticNet (combinação de L1 e L2)
    - Validação cruzada para ajuste de hiperparâmetros
5. Avaliação do Modelo:
    - Análise de métricas (R^2, RMSE, MAE, MAPE, etc.)
    - Análise de resíduos para validar suposições do modelo
    - Comparação com o modelo base apresentado no notebook 01

Formato de Entrega

- Fazer a submissão do link para o seu notebook no Google Colab
    - Caso tenha feito em seu computador local, fazer o upload para o Colab.
- O notebook deve ser executado sem erros (use "Runtime->Run All" antes do envio)

- A primeira célula deve conter os nomes de todos os integrantes do grupo
- Apenas um dos integrantes deve enviar o trabalho no ECLASS

Recomendações

- Recomenda-se fortemente trabalhar em grupo, preferencialmente com pessoas de diferentes habilidades (por exemplo, alguém com perfil de negócios junto com alguém com perfil técnico)
- O trabalho pode ser feito individualmente, mas o trabalho em equipe é encorajado

Requisitos de Código e Documentação

- O código deve estar bem comentado, explicando os passos principais
- Os comentários também serão considerados na avaliação
- Cada etapa deve ter uma breve explicação da abordagem e justificativa
- Faça seus comentários como se estivesse me explicando o que fez, use tanto comentários no código, como também nas células de texto.

Conclusão

Na última célula do notebook, o aluno deve incluir uma conclusão sobre o projeto com suas considerações finais. Essa conclusão deve incluir respostas para seguintes questões como:

- O modelo desenvolvido é satisfatório para prever o preço das casas?
- Quais foram os principais problemas identificados durante o desenvolvimento?
- Quais ações poderiam ser tomadas para melhorar ainda mais as métricas de qualidade?
- Justifique suas respostas com base nos resultados obtidos e na teoria aprendida no curso.

Avaliação

A avaliação considerará:

- Qualidade do pré-processamento e tratamento dos dados
- Implementação correta dos métodos de treino do modelo
- Melhoria nas métricas de desempenho em relação ao modelo base (modelo com uma única feature de maior correlação).
- Qualidade da análise e das conclusões
- Organização e documentação do código

Boa sorte com o projeto!