Machine Learning - FGV RI

Thiago Curado

April 14, 2022

Instruções de Realização e Entrega

- Cada aluno deve realizar sua própria resolução da lista de exercícios;
- A entrega deve ser feita até o final do dia **11 de maio**;
- **A resolução deve ser feita, obrigatoriamente, em formato de *notebook***, contendo tanto o arquivo-base como seu output final (arquivo html ou pdf). Aceitarei envios de arquivos zipados ou links para pastas armazenadas na nuvem (Dropbox, Google Cloud, etc), mas dê preferencia pela entrega via compartilhamento de repositório GIT (aproveite para se familiarizar mais com essa ferramenta!);
- A resolução deve ser feita, necessariamente,e utilizando-se a linguagem *Python*

Divirta-se!

Exercício 1

Utilizando o notebook "linear-regularized.ipynb", disponível no repositório do curso conseguimos em aula uma acurácia em torno de 8-9% para previsão de preços de imóveis.

Usando apenas os modelos vistos até aqui - isto é, Regressão Linear, Ridge e Elastic Net - o seu desafio será aumentar a acurácia dos modelos, conforme medido pelo medida "MAPE" *(Mean Absolute Percentage Error)*.

Para isso, você deverá trabalhar no pré-tratamento de dados, construir novas features e eventualmente eliminar algumas variáveis explicativas de seu dataset original

No notebook desenvolvido, pede-se que você:

- Justifique todas as escolhas para tratamento de outliers e missing values. Você tentou mais de uma abordagem? Houve impactos relevantes nos resultados?
- Você selecionou um subconjunto de features para treinar os modelos? Qual foi o impacto? Qual foi o critério utilizado para essa seleção de variáveis?
- Você construiu novas features a partir do dataset original? Em caso positivo, justifique suas construções, e avalie se as novas variáveis se mostraram relevantes.

Exercício 2

Acesse o seguinte dataset do kaggle https://www.kaggle.com/datasets/uciml/student-alcohol-consumption?datasetId=251sortBy=voteCount

Os dados contêm informações sobre o desempenho de estudantes do ensino médio nos cursos de português e matemática. Aqui, o nosso foco não será na previsão em si, mas na utilização de modelos lineares (e interpretáveis) para determinação dos principais condicionantes do desempenho acadêmico. Em especial, estamos interessados em determinar os impactos do consumo de álcool sobre as notas.

Mais uma vez, foque na utilização de modelos lineares, como regressões regularizadas, vistas até aqui no curso.

Nota: *ignore as variáveis G1 e G2. Elas são muito correlacionadas com G3, nossa variável resposta, e sua utilização atrapalharia a identificação dos impactos de interesse.*

Mais especificamente, construa um notebook que percorra as seguintes etapas:

- Percorra as etapas usuais de análise exploratória e preparação dos dados. Nesse sentido, avalie e, se necessário, trate a presença de outliers, missing values, obser-

vações repetidas e outras questões que possam afetar o aprendizado dos modelos. Justifique todas suas escolhas

- Quais são os principais determinantes para a nota observada? Avalie e interprete os impactos quantitativos das principais variáveis.
- Foque agora na avaliação dos impactos do consumo do álcool sobre o desempenho dos alunos. Quais modificações no seu conjunto de dados você propõe para melhorar essa identificação? E, utilizando tais estratégias, qual a sua avaliação acerca desses impactos?
  - Nota: *tente ir além da estimativa do impacto médio da variável avaliando, por exemplo, a heterogeneidade dos impactos ao longo dos níveis de consumo.*

Exercício 3 - Bônus

Este exercício é opcional, e será tratado como bônus.

Amplie agora o leque de modelos utilizados até aqui, utilizando,por exemplo, implementações de Random Forest ou Boosting.

Pergunta-se:

- A utilização de tais algoritmos consegue melhorar a acurácia dos modelos preditivos do exercício 1? Discuta
- A utilização de tais algoritmos traz novos insights, ou complementa as identificações discutidas no exercício 2?