

FIAP

Turma 4DTSR

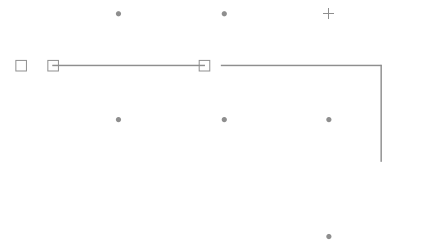
NABBA

Data Science & Artificial Intelligence
FEATURE ENGINEERING



PROJETO INTEGRADO

FEATURE ENGINEERING



CONTEXTO E DESAFIOS

A QuantumFinance possui um modelo de risco de crédito desatualizado, que tem concedido cartões para um número muito alto de mal pagadores e gerando problemas financeiros para a companhia.

Neste contexto, o HEAD de Ciência de Dados chamou os melhores talentos de sua equipe para ajudá-lo na construção de um novo modelo que será utilizado antes da concessão de cartão de crédito aos requerentes. Uma primeira análise feita no dataset mostrou que será necessário despende grande esforço de Feature Engineering antes do desenvolvimento de fato. De qualquer maneira ele está extremamente confiante, pois sabe que possui uma excelente equipe para resolver o problema.

O desafio de sua equipe nesta etapa será realizar melhorias no dataset baseadas nas melhores práticas de Feature Engineering para que, num momento posterior, seja feito o desenvolvimento do modelo.

Boa sorte!

DATASET E METODOLOGIA

O dataset disponibilizado para a sua equipe possui informações apenas de solicitações de cartão de crédito que foram aprovadas, ou seja, todos os registros são de clientes da empresa. Os cliente são marcados como maus pagadores (“bad”) se houver atraso de maior de 60 dias no pagamento da fatura durante o primeiro ano após a contratação.

A melhoria do dataset deverá ser feita com base nas técnicas fundamentais de *Feature Engineering* como imputação de dados, normalização, codificação de variáveis categóricas, redução de dimensionalidade e seleção de características.

A avaliação das melhorias do dataset deverá ser feita através da análise de performance obtida com o treinamento de um KNN Classifier para predição de bons e maus pagadores. Para garantir que não haja interferência de outros fatores na análise, alguns cuidados deverão ser tomados:

- Depois da escolha inicial do valor de k, este parâmetro deverá ser mantido constante durante todas as análises posteriores;
- A divisão do dataset entre treino e teste deverá ser feito sempre com a mesma *random_seed* para garantir que flutuações geradas por mudanças na composição do dataset não afetem na avaliação;

ENTREGÁVEIS

Ao término do projeto a equipe deverá fazer as seguintes entregas:

Arquivo powerpoint com análises e resultados:

A apresentação deverá conter não apenas o resultado final obtido e uma explicação detalhada das decisões tomadas para atingir o resultado alcançado. Em particular, alguns pontos são extremamente importantes:

- Análise para a determinação de k ;
- Detalhamento de todas as Técnicas de Feature Engineering aplicadas no dataset com justificativas das decisões tomadas;
- Resultados intermediário que a equipe julgar importante apresentar para demonstrar a evolução de qualidade do dataset;
- Relatório de acurácia e matriz de confusão dos resultados apresentados;

Jupyter Notebook:

O notebook deverá estar bem organizado com seções claras e textos que facilitem a compreensão da análise e decisões tomadas e que permita a obtenção do resultado final a partir do dataset disponibilizado.

Importante: parte significativa da avaliação do resultado será feita com a execução do notebook. Desta forma, é importante que todas as células executem corretamente e que os resultados sejam integralmente reproduzíveis;



FIAP

OBRIGADO!