topminds
challenge
stone.co

Data Challenge

uma iniciativa Stone.

Case Data Science

Previsão de probabilidade de default (PD)

Responsável: Bruno Pinheiro de Oliveira

Cientista de Dados (BU Produtos Financeiros)

Descrição do case

"*The problem is not a loss of money or credit, it's a loss of trust*" (David Perry)

Nos últimos dois anos a Stone tem trabalhado arduamente para facilitar o acesso ao crédito para empreendedores, empreendedoras e pequenos e médios negócios em todo o Brasil. Mesmo no momento mais intenso da pandemia provocada pela Covid liberamos linha de crédito e, nessa jornada, já emprestamos mais de R$ 1,2 bi para apoiar mais de 100 mil clientes a crescer, manter e reestruturar seus negócios. Ajudamos a alavancar investimentos, a salvar empreendimentos... e até mesmo uma vaca do abate.

Para que isso fosse possível, criamos um produto diferente dos existentes no mercado. Os clientes pagam os empréstimos com um percentual de retenção fixo sobre as transações em cartão de suas vendas físicas e digitais. Então, conseguimos mensurar o risco de uma maneira diferente, usando a relação histórica que temos com nossos clientes nos meios de pagamento.

Como em qualquer business de crédito precisamos gerenciar o risco das operações de forma consistente. Essa gestão de risco é essencial para garantir os resultados da empresa, mas essencialmente para garantir que nossas decisões são as melhores para nossos clientes. Isso demanda um uso robusto de dados. É com esse foco que o nosso time de dados trabalha para produzir previsões acuradas capazes de orientar as tomadas de decisão.

E é disso que esse case se trata. O desafio é **estimar a probabilidade de default (PD) de uma base de 3.000 clientes** da Stone. Essa é uma informação crucial para determinar se podemos ou não conceder crédito para um cliente e para tomarmos decisões sobre limite, taxa de juros etc.

Atualmente usamos uma variedade de métodos estatísticos e de machine learning para realizar essas previsões. Desafiamos você a desenvolver um modelo de PD capaz de contribuir com a nossa missão de ampliar o acesso ao crédito no Brasil e apoiar cada vez mais o crescimento e o fortalecimento dos negócios.

Dados

Para a resolução do case você trabalhará com dados reais de transação nos meios de pagamento da Stone:

- **train.parquet** – contém os dados para treinamento, com features raw a serem trabalhadas
- **test.parquet** – são os dados de teste

Requisitos

Você deverá usar os dados em train.parquet para treinar o seu modelo e os dados em test.parquet para validar gerar os resultados. Use os dados em train.parquet para validar o seu modelo. O modelo precisa ser treinado com os ids presentes em train.parquet e as

predições devem ser feitas para os ids em test.parquet, gerando como outputs tanto a probabilidade de default como a classificação (o threshold é por sua conta).

O seu case deverá conter duas entregas:

Entrega 1

Um data set chamado submission.parquet com as classificações (ypred) e probabilidades de default (yprob) geradas pelo modelo que você desenvolveu.

O arquivo submission.parquet deverá conter as seguintes colunas: id, ypred, yprob.

Entrega 2

O projeto apresentado em um ou mais notebooks (Jupyter ou RMarkdown), contendo o código, a documentação e o racional de cada etapa do projeto:

1. Manipulação dos dados e feature engineering
2. Análise exploratória dos dados
3. Modelo final
4. Explicação dos resultados

Avaliação dos cases

A seleção dos cases que chegarão à etapa final do desafio, de elaboração e apresentação de estratégias, será feita da seguinte forma.

1. Scores mínimos

Em primeiro lugar, seu modelo deve ter uma assertividade mínima e resultar em métricas AUC e Recall superiores aos pisos, que são:

- **AUC**: 75%
- **Recall**: 70%

Todos os modelos que superarem estes valores passarão para a segunda etapa de avaliação, que envolve a análise do projeto pela banca avaliadora

2. Projeto

Aqueles que atingirem os valores mínimos nas métricas AUC e Recall terão seus projetos avaliados. Na etapa de avaliação do projeto os pontos indicados no requisito 2 dos cases serão avaliados gerando 4 notas de 0 a 10 para os seguintes elementos avaliados:

1. Criatividade na abordagem do problema

2. Uso apropriado de técnicas de análise e modelagem

3. Interpretabilidade do modelo (encorajamos o uso de técnicas avançadas de machine learning, mas é preciso garantir que entendemos as decisões que tomamos)

4. Qualidade da documentação

5. Qualidade do código

A nota final será uma média aritmética calculada da seguinte maneira:

$$\textbf{nota final} = (a + b + c + d) / 4$$

As melhores notas serão selecionadas para a última etapa do desafio. Em caso de empate, será selecionado o case com a maior nota ***e*** (qualidade do código).

Os(as) responsáveis pelos cases selecionados para esta etapa ganharão um curso de Story Telling. Eles(as) vão elaborar e apresentar uma estratégia de uso dos resultados dos seus modelos para a aprovação de operações de crédito.

Como os projetos serão avaliados

Para cada item dos requisitos do projetos serão considerados elementos de avaliação diferentes:

1. Manipulação dos dados e feature engineering

   I. Criatividade na abordagem do problema

   IV. Qualidade da documentação

   V. Qualidade do código

2. Análise exploratória dos dados

   I. Criatividade na abordagem do problema

   II. Uso apropriado de técnicas de análise e modelagem

3. Modelo final

   I. Criatividade na abordagem do problema

   II. Uso apropriado de técnicas de análise e modelagem

   IV. Qualidade da documentação

V. Qualidade do código

4. Explicação dos resultados (avaliação do modelo)

II. Uso apropriado de técnicas de análise e modelagem

III. Interpretabilidade do modelo

V. Qualidade do código

Banca avaliadora

Composta por engenheiros e cientistas de dados da Stone.