Quero Empréstimo

Data Science Case

Cientistas Responsáveis: Alisson Felipe Lima Santos
Natan Nascimento Oliveira Matos
Natália Braga da Fonseca

Para solucionarmos o problema da empresa Quero Empréstimo, podemos realizar a utilização de modelos de score de crédito, onde calculam a probabilidade de inadimplência e são uma das principais ferramentas utilizadas para aprovar ou negar um crédito.

Com isso, em um primeiro momento precisamos analisar o dataset que nos foi concedido e dessa forma realizar uma análise exploratória dos dados, entendo todas as suas características. Após entender inicialmente como o dataset está estruturado, podemos realizar uma análise univariada para compreendermos a relação entre as variáveis e a variável *target* que é a inadimplência, com isso podemos responder algumas perguntas como “Como está a distribuição de clientes com classe boa e ruim?”, “Como está a distribuição do histórico de crédito?”, “Quantos clientes estão no cheque especial?”, entre outros.

A partir do momento que identificamos pontos em cada uma das variáveis, vamos entender um pouco melhor como o dataset está distribuído e com isso poderemos tomar melhores decisões. Para melhorarmos ainda mais o entendimento sobre o dataset, podemos realizar uma análise bivariada e analisar as relações entre variáveis, em um primeiro momento podemos gerar uma matriz de correlação para termos como base e a partir disso, vamos identificando as relações e criando algumas perguntas chaves para respondermos.

Após finalizarmos toda a análise exploratória entendendo as variáveis e realizando um *deep dive* em cima do nosso dataset, podemos partir para o pré-processamento, nesta etapa vamos realizar o tratamento de valores faltantes e remover os *outliers* do nosso dataset. Finalizando essa primeira etapa, vamos normalizar os dados, podemos utilizar o método “*RobustScaler*” presente na biblioteca do sklearn, esse método leva em conta os percentis da distribuição para fazer a normalização, com isso podemos ter um melhor uso dos dados presentes no nosso dataset mesmo que ainda não tenhamos removido por completo os *outliers*. E por fim, podemos realizar o *Oversampling* e *Undersampling* por meio do método *Smoteenn*, ou seja, se avaliarmos que o nosso conjunto de dados está desbalanceado utilizamos o “*Smoteenn*”, que utiliza os dois principais métodos para lidar com dados desbalanceados: oversampling e undersampling. A etapa de *oversampling* ele realiza com o SMOTE, já o *undersampling* ele realiza com o ENN.

Por fim, podemos realizar uma modelagem preditiva, selecionando alguns modelos como Regressão Logística, Árvore de Decisão e Random Forest, e avaliando qual modelo se encaixa melhor para o nosso caso de uso. Realizamos as previsões e identificamos clientes inadimplentes e tomamos a decisão de qual cliente devemos ou não ceder crédito, dessa forma poderemos reduzir alguns porcentos a taxa de inadimplência da empresa Quero Empréstimo.