Predição de qualidade do vinho

Diego de Almeida Miranda - 133603

# Introdução

A inadimplência no setor bancário é uma das variáveis mais observadas na área. Para realizar o cálculo da inadimplência, já foram propostos diversos métodos diferentes. Há uma dificuldade em definir precisamente o que é a inadimplência, o que leva os pesquisadores a deixarem de medir diversos desses casos (de inadimplência). Objetivamente, podemos definir a inadimplência como a falta do cumprimento de uma obrigação. Entretanto, apesar de possuir uma definição clara, a tarefa de definir um inadimplente para analisar o risco de crédito não é trivial. Por conta disso, diversos autores apresentam e usam diferentes definições para inadimplência, mas todas tendo em foco um devedor que não sanou suas dívidas.

Ainda assim, devido ao mercado e nível de competitividade empresarial crescente no decorrer dos anos, se faz necessária maior eficiência no gerenciamento dos recursos capitais de todas as organizações. O acesso facilitado ao crédito também abriu portas para a concorrência de maus pagadores. Nitidamente, a existência de mau pagadores é um problema para as empresas de todo porte, pois na concessão de crédito é feito um planejamento de quando obteria novamente aquele valor.Ainda que as empresas e consumidores tenham aproximado relações nos últimos anos, através de diversas obrigações contratuais, é corriqueiro que haja o descumprimento destes acordos. Isso tanto de quebra de acordo do cliente para a empresa, quanto da quebra da empresa para com o cliente.

Com um eventual não cumprimento do cliente com as cláusulas estabelecidas no contrato, como o não pagamento do valor acordado, a empresa precisa tomar medidas, podendo cancelar o serviço que está sendo prestado, até que o cliente acerte suas dívidas com a empresa em questão. Mesmo assim, é necessário que a empresa tenha maneiras de se proteger contra clientes com alto potencial de se tornarem inadimplentes.

O Serasa realizou um levantamento em 2021 indicando que em Abril haviam 62,98 milhões de brasileiros que estão inadimplentes. Em setembro, esta dívida atingiu o valor de R$\$$245,3 Bilhões. No mesmo mês, havia R$\$$3.944,65 de dívida por pessoa inadimplente. Além disso, bancos e cartões de crédito são os credores de 28,70% das dívidas, sendo seguidos pelas empresas relacionadas ao abastecimento de água, luz e contas básicas, correspondendo a 23,5% das dívidas. Empresas do comércio varejistas aparecem ocupando o terceiro lugar, com 13% das dívidas totais no Brasil.

Para fins de estudo, utilizaremos este conjunto de dados sintéticos obtidos através da plataforma Kaggle, intitulado como \_Loan Default Prediction\_. Segundo o autor, são dados obtidos através de instituições financeiras, mas que por segurança das instituições envolvida, foram removidos quaisquer dados que possam ser utilizados para rastrear e identificar algum desses clientes.

Aqui, temos 3 principais atributos para identificar uma amostra de indivíduo/instituição inadimplente, sendo elas a condição de emprego, se está empregado ou não, qual o valor de dinheiro no banco no momento atual, e por fim qual o seu salário anual. Através destas variáveis teremos de realizar as análises que poderão dizer ou não quando um indivíduo tem chances consideráveis de ser inadimplente. Normalmente, para este tipo de análise, também são considerados os dados de percepção do mercado, setor de atividade da empresa/individuo, tempo de atividade, nível de informatização da empresa, nível de escolaridade do sócio, entre outros. Dependendo da população estudada, i.e., microempresa, empresas médias, de grande porte ou até mesmo indivíduos, esses atributos podem variar.

É notável que o volume de dados é algo a ser fortemente considerado, uma vez que não podemos fazer as análises puramente através de análises gráficas e exploratória dos dados. A grande quantidade de amostras e parâmetros para serem analisados torna a tarefa de compreender os padrões de comportamento dos clientes uma atividade extremamente complicada, mesmo que para grandes equipes de pessoas.

Em advento do desenvolvimento tecnológico e matemático, podemos aplicar métodos de aprendizado de máquina para encontrar relações entre as variáveis descritivas e objetivas do problema. Quando pensamos em dados de inadimplência, podemos buscar compreender quanto a relação entre o salário anual e o extrato do banco de um indivíduo interfere no problema de classificação de um indivíduo inadimplente.

Para este conjunto de dados em específico, foram escolhidos métodos de classificação a partir da regressão logística, modelo este mais indicado pela literatura sobre o tema. O fato de termos uma variável objetivo binária, ou é inadimplente ou não é, torna o uso da regressão linear convencional algo impróprio, uma vez que não estamos prevendo dados contínuos. A adequação de um modelo para com os dados pode ser feita através das análises de significância estatística das variáveis descritivas. Além da Regressão logística convencional, será utilizado a regressão logística com penalização e, por fim, o método de classificação Random Forest.

Como estamos trabalhando com variáveis categóricas, não existem razões para utilizar um método de regressão que nos retorne valores contínuos. Da mesma forma, vale para a métrica de avaliação que será aqui utilizada. Os modelos aqui apresentados serão avaliados utilizando as métricas de acurácia, área sob a ROC - conhecido como AUC ROC - e pontuação F1. A acurácia é dada pela divisão entre a soma dos resultados verdadeiros (positivo e negativo) pela quantidade de amostras que temos. Sendo \*VP\* amostras classificadas corretamente como positivas, e \*VN\* amostras classificadas corretamente como negativas, então a acurácia \*acc\* pode ser dada por $acc = \frac{VP + VN}{|amostras|}$. A métrica de pontuação F1 é a média harmônica entre precisão e revocação, e é dada pela seguinte equação $f1 = 2\frac{Precisao\times Revocaçao}{Precisao + Revocaçao}$. Por fim, a curva ROC é realizada através das taxas de verdadeiro positivo e verdadeiro negativo. No eixo X da curva ROC temos o \*False positive Rate\*, onde $FPR = \frac{FP}{VP + FP}$. De forma análoga, temos no eixo Y temos o \*True Positive Rate\*, dado por $FPR = \frac{VP}{VP + FN}$. A métrica de AUC nada mais é do que o valor da área sob a curva ROC.

Nenhuma métrica de validação cruzada será implementada neste trabalho, o que pode acabar por não assegurar a capacidade de generalização dos modelos. Por fim, podemos dizer que todas as análises aqui apresentadas foram feitas utilizando a linguagem de programação R e suas bibliotecas.

# Análises

As análises podem ser verificadas no pdf associado a este trabalho.

# Conclusão

Agora que já realizamos todas as análises e extraímos as informações necessárias, podemos então discorrer sobre o que foi aqui apresentado. Neste conjunto de dados, temos apenas 3.33% das variáveis classificadas como inadimplentes. Isso somam 333 amostras das 10 mil coletadas. De maneira contra intuitiva o salário anual de um indivíduo não se mostrou como atributo pertinente para a classificação de uma pessoa possivelmente inadimplente. A variável salarial apresentou uma média de aproximadamente 400K de unidades monetárias para pessoas adimplentes e inadimplentes. A baixa correlação, com valor atribuído de -0.02, também é um indicador para podermos pressupor isto antes mesmo de aplicarmos os modelos. A variável que mais se correlaciona com o salário anual é o indicador de emprego , i.e. \_Employed\_, com indice de 0.75 pontos. O atributo \_Employed\_, por sua vez, também não apresentou correlação com a inadimplência. Entretanto, a variável de saldo bancário, \_Bank Balance\_, se mostrou muito eficiente enquanto variável descritiva para a variável alvo de inadimplência. A média do saldo bancário em indivíduo inadimplentes passa de 20.000, enquanto a média do saldo bancário de pessoas adimplentes é ligeiramente abaixo de 10.000. Do conjunto de pessoas adimplentes, existem alguns outliers que passam de 25 mil em saldo bancário, valor que estariam no "bigode"(whisker) do boxplot de inadimplentes.

Ajustando o primeiro modelo de regressão logística, com todas as variáveis e também com todas as amostras, pudemos notar - mais uma vez - que a variável \_Annual Salary\_ não é significante no nosso modelo, isso por ela possuir um P valor de 0.62 > 0.05, que é o nível significância. Neste modelo, também foi considerado pouco significante a variável \_Employed\_. No geral, pudemos ver que este primeiro modelo teve acurácia de 0.97, um valor alto o suficiente para aplicações práticas. Para um segundo modelo, removemos não só a variável salarial, como também as amostras influentes no modelo, i.e., aquelas com valor absoluto residual padronizado maior que 3 Neste modelo, todas as variáveis são significativas, entretanto não houve otimização no desempenho, onde obtivemos novamente 0.97 de acurácia. Por fim, o modelo de classificação Random Forest obteve 0.97 de acurácia, além disso, também atribui-se que a variável de salário anual também não é relevante no processo de classificação.

Como todos os modelos obtiveram resultados muito parecidos em acurácia, observamos que em comparação aos outros dois modelos, o modelo de regressão logistica (com todos atributos e amostras) teve destaque, pois apresentou melhor métrica de auc (0.87).

Dessa forma, podemos dizer que o saldo bancário atual de um indivíduo pode vir a ser um bom indicador de potenciais clientes inadimplentes. Entretanto, as variáveis de condição de emprego e de salário anual não parecem ser interessantes para este tipo de predição. Outras possíveis variáveis positivas para o desenvolvimento de modelos preditivos para este modelo, são aquelas omitidas aqui neste conjunto de dados, ou então variáveis temporais, de forma a que possamos acompanhar o comportamento econômico dos clientes. Com estas diferentes variáveis, poderíamos entender o contexto em que cada cliente está inserido sócio-economicamente, e podemos a partir disto buscar por modelos mais sofisticados na área.

# Referências bibliográficas

[1] Camargos, Marcos Antônio de, Mirela Castro Santos Camargos, Flávio Wagner Silva, Fabiana Soares dos Santos, and Paulo Junio Rodrigues. "Fatores condicionantes de inadimplência em processos de concessão de crédito a micro e pequenas empresas do Estado de Minas Gerais." Revista de Administração Contemporânea 14, no. 2 (2010): 333-352.

[2] Daros, Mariane, and Nelson Guilherme Machado Pinto. "Inadimplência no brasil: uma análise das evidências empíricas." Revista de Administração IMED 7, no. 1 (2017): 208-229.

[3] Annibal, Clodoaldo Aparecido. Inadimplência do Setor Bancário Brasileiro: uma avaliação de suas medidas. No. 192. 2009.