# Trabalho de Formatura MAP2040

Gustavo Soares Gomes

IME-USP

2024

# Contexto

- O mercado de crédito possui papel fundamental na economia brasileira: consumos de familias e investimentos empresariais;
- No cenário atual de déficit da divída pública e demais instabilidades, há uma elevação das taxas de juros e o aumento de inflação levando a uma elevação no custo de crédito;
- Esse cenário implica em altos graus de endividamento com aumento da probabilidade de inadimplência.

# Contexto

- Consequente piora do perfil médio de risco dos tomadores de empréstimos no Brasil;
- Como formular políticas de crédito que favoreçam a inclusão financeira de maneira sustentável para as instituições ?
- Será necessário que as instituições financeiras adotem modelos mais conservadores na concessão de crédito?
- Necessidade de modelos preditivos eficazes para decisão de crédito.

# Objetivos

- Apresentar um breve contexto do cenário atual do crédito no Brasil;
- Desenvolver, implementar e comparar um modelo preditivo para análise e concessão de crédito em uma população que simule o cenário de crédito atual no Brasil;
- Comparar a eficácia frente a modelos tradicionais de Regressão Logística e Random Forest a partir de métricas adequadas para tal.

# Fonte dos Dados

- Dados obtidos da plataforma Kaggle com origem do Lending Club, plataforma de empréstimos pessoais peer to peer online nos EUA;
- Empréstimos pessoais sem garantia cujo propósito principal era a consolidação de débitos e dívidas de cartão de crédito realizados dentro do período de 2007 a 2018;
- 2 milhões de registros com 147 variáveis observadas na base original, somado à capacidade de processamento limitada.

# Refinamento dos Dados

- Foram obtidas duas amostras aleatórias da base original: amostra de treino com 350 mil registros e amostra de teste com 150 mil, com iguais quantidades de "bons" e "maus" pagadores em cada amostra;
- Bom pagador é aquele que pagou completamente seu empréstimo, mesmo com eventuais parcelas pagas depois do vencimento;
- Mau pagador é aquele em que não há expectativa de pagamento total do empréstimo devido a atrasos recorrentes.

# Preparação dos Dados

- Variáveis do tipo texto transformadas em numéricas, observações nulas substituídas por 0 e remoção de variáveis de identificação;
- Clusterização das variáveis conforme correlação: variáveis com mais de 90% de correlação entre si agrupadas no mesmo cluster (algoritmo Linkage);
- Para cada cluster com mais de uma variável (32), foi escolhida uma variável representante;
- Exclusão de variáveis observadas pós fechamento do contrato (14): pagamento até o momento, valor recuperado em caso de perda, etc.

# Compactação dos Dados

- Mesmo após aplicação dessas etapas, a base ainda apresentava 68 variáveis;
- Utilização do algoritmo de *Random Forest* para medir a importância de cada variável na previsão da variável alvo e redução da base às 10 variáveis com maior importância.

| Variável | Descrição | Importância |
|---|---|---|
| `int rate` | A taxa de juros do empréstimo | 8.53% |
| `dti` | A relação dívida/renda do mutuário | 4.34% |
| `bc open to buy` | Crédito disponível em linhas rotativas | 3.59% |
| `revol bal` | O saldo total de crédito rotativo do mutuário | 3.56% |
| `avg cur bal` | Saldo médio atual de todas as contas | 3.55% |
| `loan amnt` | O valor do empréstimo solicitado | 3.55% |
| `annual inc` | A renda anual do mutuário | 3.55% |
| `tot cur bal` | Saldo total atual de todas as contas | 3.39% |
| `bc util` | Utilização da linha de crédito atual | 3.36% |
| `total bc limit` | Total de limites de crédito do Banco | 3.35% |

# Padronização de Dados

- Teste de Shapiro Wilk: Rejeita-se a hipótese nula de que as variáveis possuem distribuição normal a 5% de siginificância;
- Remoção de outliers baseada em intervalo interquartil e normalização pelo método min-max.

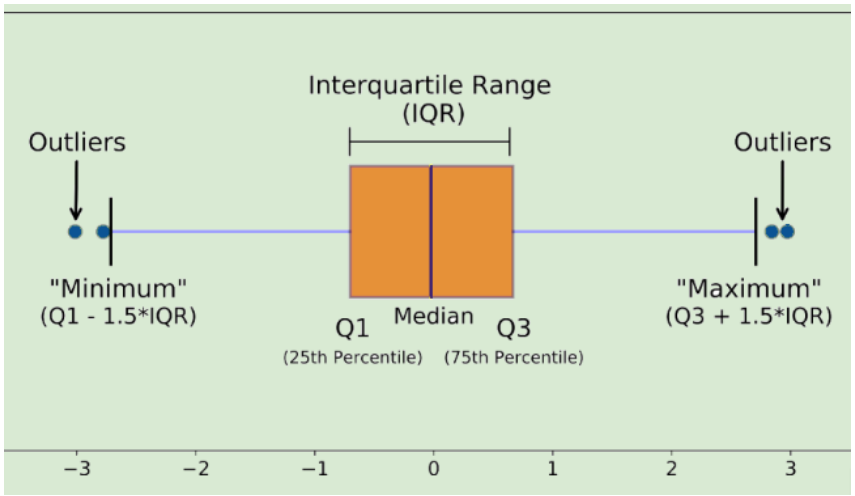


Figure: Reprodução Varsha Saini

$$x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

# Desenvolvimento Modelo de Risco

- Modelos tradicionais como a Regressão Logística e Árvores de decisão não englobam o sentido do risco do crédito:
- A penalização pelo modelo em aceitar um mutuário "ruim" é a mesma do que quando rejeitamos um mutuário "bom", o primeiro possui maior perda financeira envolvida;
- Definiu-se uma variável proxy $I$ (index) que estima a capacidade de pagamento do empréstimo pelo mutuário:

$$I = \frac{V\dot{(}1+i)^{60}}{60R}$$

- Demais variáveis escolhidas foram: relação dívida/renda do indivíduo (dti) e saldo médio atual das contas do mutuário (avg cur bal).

# Desenvolvimento Modelo de Risco

- Mais variáveis poderiam ter sido escolhidas para aumentar o desempenho do modelo, não o fizemos por uma questão de complexidade e tempo para implementação;
- Por questão de interpretabilidade, iremos utilizar um modelo de árvore de decisão;
- Erros que podem ser cometidos pelo modelo:
  - Erro Tipo I : Consiste em aceitar a concessão de empréstimo a um mutuário que futuramente deixará de cumprir as suas obrigações com a instituição;
  - Erro Tipo II: Consiste em não aceitar conceder o empréstimo a um mutuário que futuramente não deixaria de cumprir as suas obrigações com a instituição.

# Desenvolvimento Modelo de Risco

- Parâmetro de Flexibilidade:

$$\alpha = \frac{\text{Contagem do Erro Tipo 1}}{\text{Contagem dos Erros Totais}}$$

- Desejamos encontrar modelos tais que $\alpha < 0.5$. Quanto mais próximo de 0.5, mais tolerável é o modelo às perdas financeiras e quanto mais próximo de 0 mais rígida é sua decisão;
- Parâmetro de flexibilidade será utilizado para a definição dos valores dos pontes de corte de cada variável conforme o algoritmo a seguir:

# Desenvolvimento Modelo de Risco

- Etapa I : Escolher um ponto de corte inicial para a variável *index* a partir de um valor entre entre as medianas dos maus e bons pagadores para essa variável. Em nosso caso, utilizamos $index = 15.0$ como estimativa inicial;
- Etapa II: Escolher da base de treino uma amostra aleatória de 1000 bons pagadores e 1000 maus pagadores;
- Etapa III: Para o ponto de corte da iteração verificar se o erro tipo 1 representa menos que $\alpha$ do total de erros. Em caso positivo, este será o ponto de corte e em caso negativo reduz-se o ponto de corte em -0.5 até encontrar o ponto de corte que satisfaça essa condição;
- Etapa IV: Repete-se as etapas I, II, III 1000 vezes e tira-se a média dos pontos de cortes obtidos.

# Desenvolvimento Modelo de Risco

- Ponto de corte index $= 2.5$
- Para representar o conservadorismo em relação à essa população, optamos por seguir trabalhando apenas com os mutuários que possuissem index $< 2.5$;
- Etapa V consiste em rodar novamente as etapas I a IV com as demais variáveis (dti e avg cur bal);

# Desenvolvimento Modelo de Risco

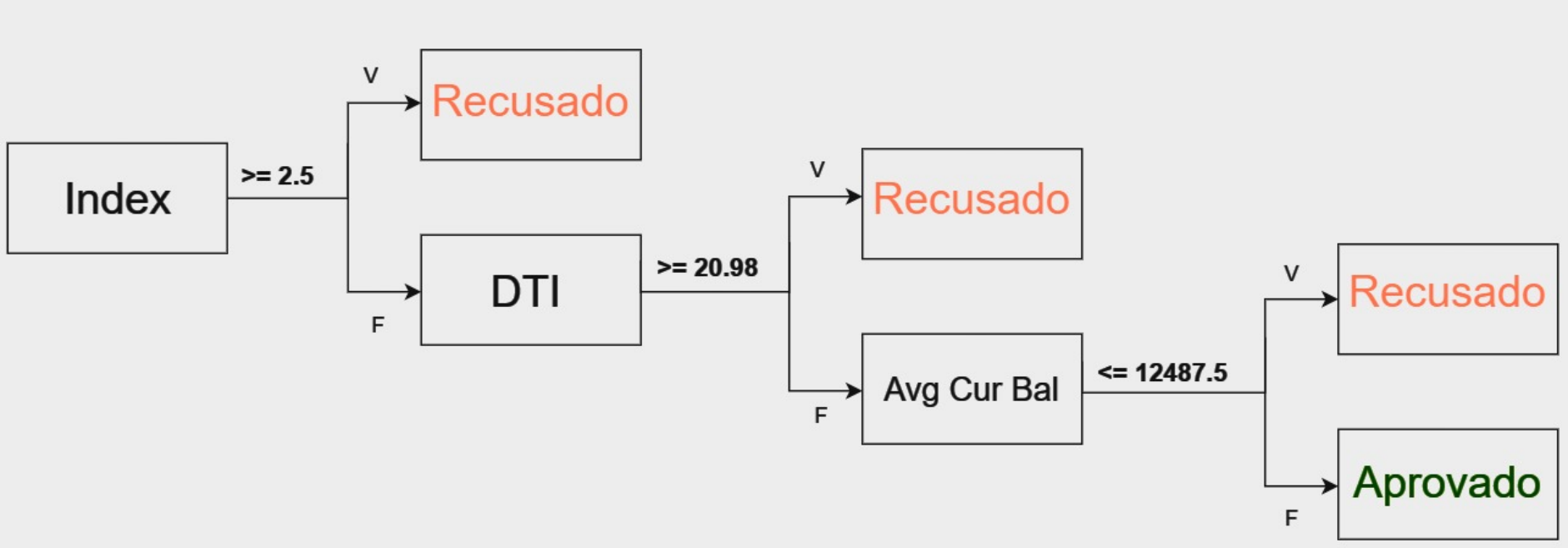


Figure: Modelo final para $\alpha = 0.33$

# Métricas de Avaliação

- A acurácia mede a proporção de previsões corretas em relação ao total de previsões realizadas:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

- A precisão mede a proporção de previsões positivas corretas entre todas as previsões positivas realizadas:

$$\text{Precisão} = \frac{TP}{TP + FP}$$

# Métricas de Avaliação

- A sensibilidade, ou *recall*, mede a proporção de positivos verdadeiros corretamente identificados:

$$\text{Sensibilidade} = \frac{TP}{TP + FN}$$

- O F1-Score é a média harmônica entre precisão e sensibilidade:

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}$$

# Métricas de Avaliação

- O **F1-Score ponderado** é uma média ponderada dos F1-Scores de cada classe;

$$F1_{\text{ponderado}} = \frac{\sum_{i=1}^{n} w_i \cdot F1_i}{\sum_{i=1}^{n} w_i}$$

- Em nosso caso $n = 2$, $w_0 = 1 - \alpha$ e $w_1 = \alpha$.
- Olhamos principalmente para os indicadores de precisão, sensibilidade e F1- score;
- *trade-off* entre precisão e sensibilidade.

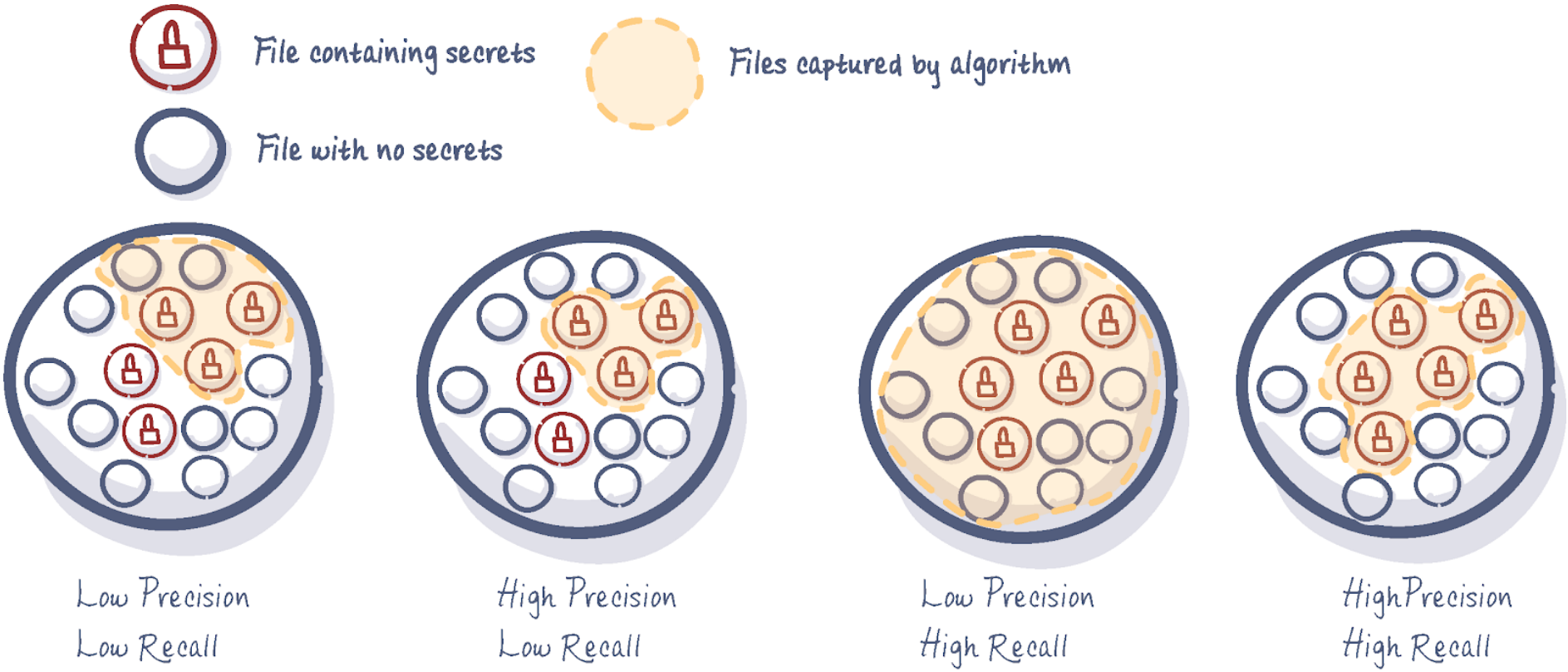


Figure: Reprodução GitGardian

# Aplicação dos modelos tradicionais

- Floresta Aleatória
  - Predição da variável *loan status* a partir das 10 variáveis obtidas no slide 8;
  - Pelo funcionamento do algoritmo, não há influência ativa de uma variável sobre a outra;
- Regressão Logística
  - Existem $2^{10} - 1$ possíveis modelos;
  - Critérios de informação: equilibrio entre verossimilhança e quantidade de variáveis utilizadas, mais comum são BIC e AIC;
  - Iremos utilizar o critério BIC por penalizar de forma mais rigorosa modelos complexos;
  - Modelo escolhido possui todas os coeficientes $\beta$ das variáveis significativos (Wald).

# Simulações

| Simulação 1 com $\alpha = 0.25$ | | | | | |
|---|---|---|---|---|---|
| **Modelo** | **Acurácia** | **Precisão** | **Sensibilidade** | **F1-Score** | **Aprovação** |
| Regressão Logística | 63,76% | 62,58% | 62,32% | 64,35% | 48,15% |
| Floresta Aleatória | 67,34% | 73,18% | 51,24% | 69,28% | 33,86% |
| Modelo de Risco | 55,14% | 77,09% | 10,29% | 66,56% | 6,45% |

Table: Resultados da simulação com $\alpha = 0.25$

| Simulação 2 com $\alpha = 0.33$ | | | | | |
|---|---|---|---|---|---|
| **Modelo** | **Acurácia** | **Precisão** | **Sensibilidade** | **F1-Score** | **Aprovação** |
| Regressão Logística | 63,76% | 62,57% | 62,32% | 64,15% | 48,15% |
| Floresta Aleatória | 67,34% | 73,18% | 51,24% | 68,32% | 33,86% |
| Modelo de Risco | 55,84% | 74,82% | 13,07% | 66,82% | 8,45% |

Table: Resultados da simulação com $\alpha = 0.33$

# As simulações indicam que

- O Modelo de Risco possui menor acurácia do que os demais modelos, conforme esperado;
- Modelo de Risco possui maior precisão do que demais modelos e essa métrica de "conservadorismo" aumenta conforme diminuímos o parâmetro de flexibilidade;
- Modelos possuem F1-scores muito semelhantes, o que indica que o aumento de precisão do Modelo de Risco é compensado por uma diminuição consideravel na sensibilidade.
- Em outras palavras, o Modelo Risco leva mais em conta não trazer maus pagadores para a instituição do que perder a captação de vários potenciais bons pagadores.

# Aplicações

- A instituição poderá utilizar o Modelo de Risco então para populações em que não é compensatório captar o máximo de clientes possíveis devido a alta probabilidade de inadimplência;
- A calibragem do parâmetro $\alpha$ depende essencialmente das perdas registradas no portfólio e dos juros gerados pelos potenciais clientes recusados, então recomenda-se a instituição em um primeiro momento adotar um $\alpha$ pequeno (cenário mais conservador), para ir ajustando seu valor conforme crescimento do seu portfólio.