

## Projeto 4: Prevendo o Risco de Calote

### Passo 1: Entendimento de negócios e dados

*Fornecer uma explicação das principais decisões que precisam ser feitas. (Limite de 250 palavras)*

R: Devido a um escândalo financeiro em um banco concorrente, houve um grande aumento do número de pessoas pedindo empréstimo para o banco onde trabalho. De uma hora para outra, é preciso analisar 500 pedidos de empréstimo em uma semana. Antes do escândalo financeiro no banco concorrente, analisávamos em média 200 pedidos de empréstimo por mês, e cada um deles foram aprovados de forma manual. Precisamos descobrir uma forma de analisar todos esses novos pedidos de empréstimo em uma semana.

*Decisões chave:*

*Responda estas perguntas*

1. Que decisões precisam ser tomadas?

R: Precisamos avaliar sistematicamente se esses novos pedidos de empréstimo merecem crédito ou não.

2. Que dados são necessários para informar essas decisões?

R: Dados de todos os antigos pedidos de empréstimo, onde foram analisados variáveis como *Account-Balance* (Saldo em conta), *Duration-of-Credit-Month* (Quantidade de parcelas), *Payment-Status-of-Previous-Credit* (Status de pagamento de crédito anterior), *Purpose* (Propósito do empréstimo), *Credit-Amount* (Valor do empréstimo) entre outras. Temos também uma lista de clientes cujos pedidos devem ser analisados nos próximos dias.

3. Que tipo de modelo (Contínuo, Binário, Não-Binário, Time-Series) precisamos usar para ajudar a tomar essas decisões?

R: Binário. Devido se tratar apenas de duas condicionantes, crédito aprovado ou não.

### Passo 2: Construindo o Conjunto de Treinamento

*Construa seu conjunto de treinamento dado os dados fornecidos a você. Os dados foram limpos para você já assim você **não deve precisar converter quaisquer campos de dados para os tipos de dados apropriados.***

*Aqui estão algumas diretrizes para ajudar a orientar sua limpeza de dados:*

- Para campos de dados numéricos, existem campos que se correlacionam entre si? A

correlação deve ser de pelo menos 0,70 para ser considerada "alta".

- Existem dados em falta para cada um dos campos de dados? Campos com muitos dados em falta devem ser removidos
- Existem apenas alguns valores em um subconjunto de seu campo de dados? O campo de dados parece muito uniforme (há apenas um valor para todo o campo?). Isso é chamado de "baixa variabilidade" e você deve remover os campos que têm baixa variabilidade. Consulte a seção "Dicas" para encontrar exemplos de campos de dados com baixa variabilidade.
- Seu conjunto de dados limpos deve ter 13 colunas onde a média de **Age Years** deve ser 36 (arredondado para cima)

**Nota:** Por uma questão de consistência no processo de limpeza de dados, impute dados usando a média de todo o campo de dados em vez de remover alguns pontos de dados. (Limite de 100 palavras)

**Nota:** Para alunos que usam software diferente do Alteryx, por favor, formate cada variável como:

| Variable                          | Data Type |
|-----------------------------------|-----------|
| Credit-Application-Result         | String    |
| Account-Balance                   | String    |
| Duration-of-Credit-Month          | Double    |
| Payment-Status-of-Previous-Credit | String    |
| Purpose                           | String    |
| Credit-Amount                     | Double    |
| Value-Savings-Stocks              | String    |
| Length-of-current-employment      | String    |
| Instalment-per-cent               | Double    |
| Guarantors                        | String    |
| Duration-in-Current-address       | Double    |
| Most-valuable-available-asset     | Double    |
| Age-years                         | Double    |
| Concurrent-Credits                | String    |
| Type-of-apartment                 | Double    |
| No-of-Credits-at-this-Bank        | String    |
| Occupation                        | Double    |
| No-of-dependents                  | Double    |
| Telephone                         | Double    |
| Foreign-Worker                    | Double    |

*Para alcançar resultados consistentes os revisores esperam.*

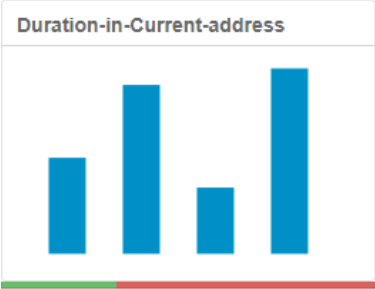
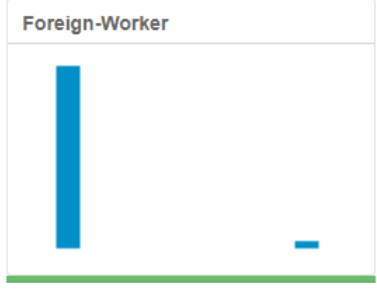
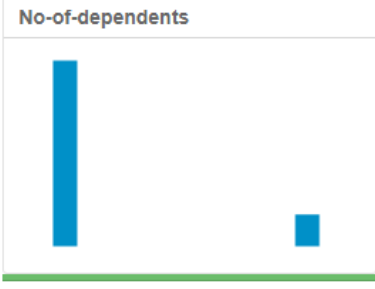
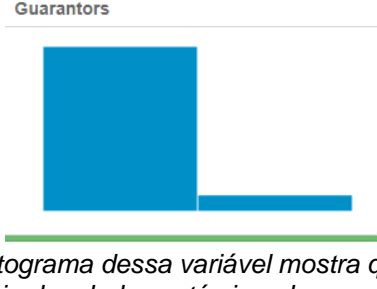


*Responda esta pergunta:*

1. Em seu processo de limpeza, quais campos você removeu ou imputou? Por favor, justifique por que você removeu ou imputou esses campos. As visualizações são

incentivadas.

R: Imputei dados para o campo "Age-years", porque para uma análise de crédito é interessante saber qual é a idade do cliente. Para isso fiz a imputação com o valor mediano de "Age-years".

Removi os campos listados na tabela abaixo.

| <i>Campos removidos</i>  | <i>Campos removidos</i>  |
|--|--|
| <div><p><b>Duration-in-Current-address</b></p><p>Para esse campo existir muitos valores nulos. Optei por removê-lo.</p></div>                 | <div><p><b>Foreign-Worker</b></p><p>O histograma dessa variável mostra que a maioria dos dados está viesada para um único valor.</p></div> |
| <div><p><b>No-of-dependents</b></p><p>O histograma dessa variável mostra que a maioria dos dados está viesada para um único valor.</p></div> | <div><p><b>Guarantors</b></p><p>O histograma dessa variável mostra que a maioria dos dados está viesada para um único valor.</p></div>    |
| <div><p><b>Telephone</b></p><p>Não há nenhuma razão lógica para incluir a variável</p></div>  | <p><b>Occupation</b></p> <p>Está variável é uma variável de "baixa variabilidade", por isso removi ela. Esta variável contém apenas um valor "1".</p>  |
| <div><p><b>Concurrent-Credits</b></p><p>Está variável é uma variável de "baixa variabilidade", por isso removi ela.</p></div>               |  |

## Passo 3: Treinar seus Modelos de Classificação

Primeiro, crie suas amostras de Estimação e Validação, onde 70% de seu conjunto de dados deve ir para Estimativa e 30% de seu conjunto de dados inteiro deve ser reservado para Validação. Defina a Semente Aleatória como 1.

Crie todos os modelos a seguir: regressão logística, árvore de decisão (decision trees), modelo de floresta (forest model), e boosted model.

Responda a estas perguntas para **cada modelo** criado:

1. Quais variáveis preditoras são significativas ou as mais importantes? Por favor, mostre os p-values ou gráficos de importância para todas as suas variáveis de previsão.

a) Regressão logística

As variáveis mais importantes para o modelo de regressão logística são:

Account.Balance; Credit.Amount; Purpose; Instalment.per.cent;

Length.of.current.employment; Payment.Status.of.Previous.Credit

Coefficientes:

|  | Estimate   | Std. Error | z value | Pr(> z ) |     |
|--|------------|------------|---------|----------|-----|
| (Intercept)                                    | -2.9621914 | 6.837e-01  | -4.3326 | 1e-05    | *** |
| Account.BalanceSome Balance                    | -1.6053228 | 3.067e-01  | -5.2344 | 1.65e-07 | *** |
| Credit.Amount                                  | 0.0001704  | 5.733e-05  | 2.9716  | 0.00296  | **  |
| Instalment.per.cent                            | 0.3016731  | 1.350e-01  | 2.2340  | 0.02549  | *   |
| Length.of.current.employment4-7 yrs            | 0.3127022  | 4.587e-01  | 0.6817  | 0.49545  |     |
| Length.of.current.employment< 1yr              | 0.8125785  | 3.874e-01  | 2.0973  | 0.03596  | *   |
| Most.valuable.available.asset                  | 0.2650267  | 1.425e-01  | 1.8599  | 0.06289  | .   |
| Payment.Status.of.Previous.CreditPaid Up       | 0.2360857  | 2.977e-01  | 0.7930  | 0.42775  |     |
| Payment.Status.of.Previous.CreditSome Problems | 1.2154514  | 5.151e-01  | 2.3595  | 0.0183   | *   |
| PurposeNew car                                 | -1.6993164 | 6.142e-01  | -2.7668 | 0.00566  | **  |
| PurposeOther                                   | -0.3257637 | 8.179e-01  | -0.3983 | 0.69042  |     |
| PurposeUsed car                                | -0.7645820 | 4.004e-01  | -1.9096 | 0.05618  | .   |

Significance codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

b) *Árvore de decisão*

As variáveis mais importantes para o modelo de árvore de decisão são:  
*Account.Balance; Value.Savings.Stocks; Duration.of.Credit.Month.*

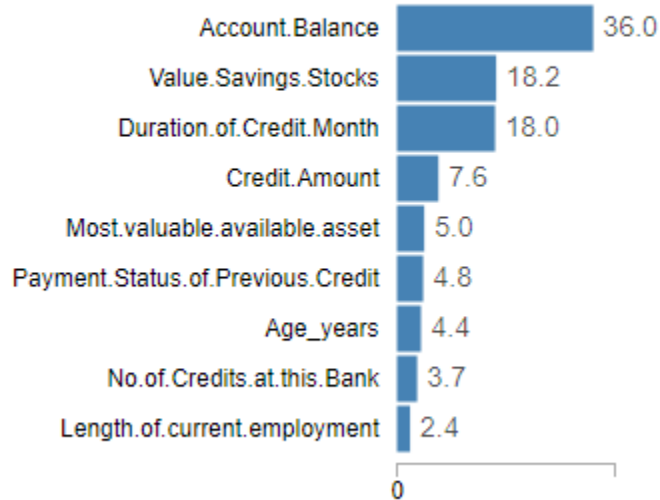
#### Leaf Summary

node), split, n, loss, yval, (yprob)

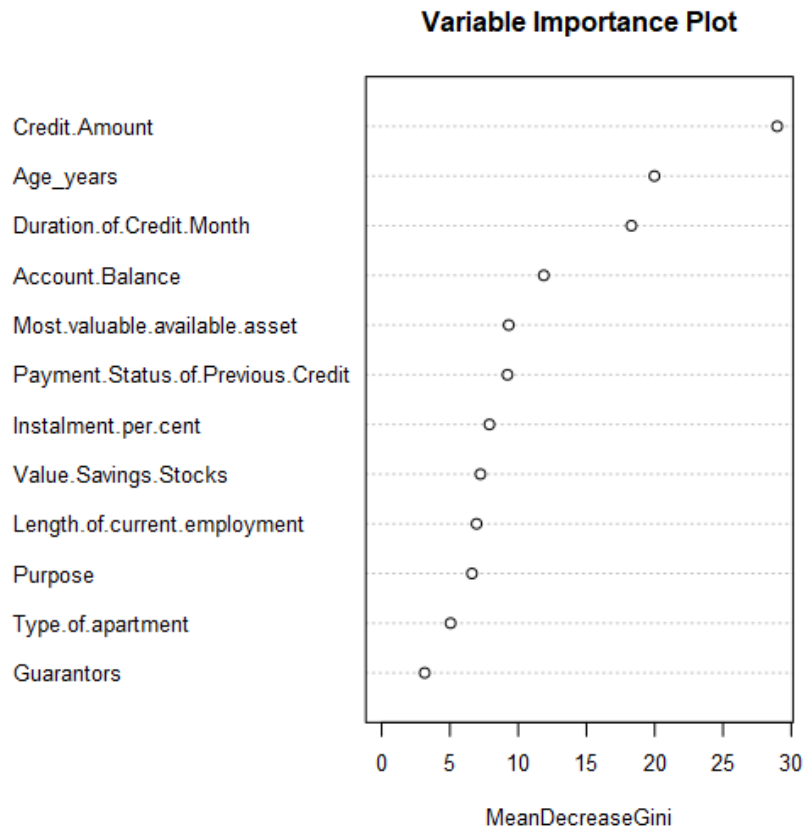
\* denotes terminal node

- 1) root 350 97 Creditworthy (0.7228571 0.2771429)
- 2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) \*
- 3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
- 6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) \*
- 7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
- 14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) \*
- 15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) \*

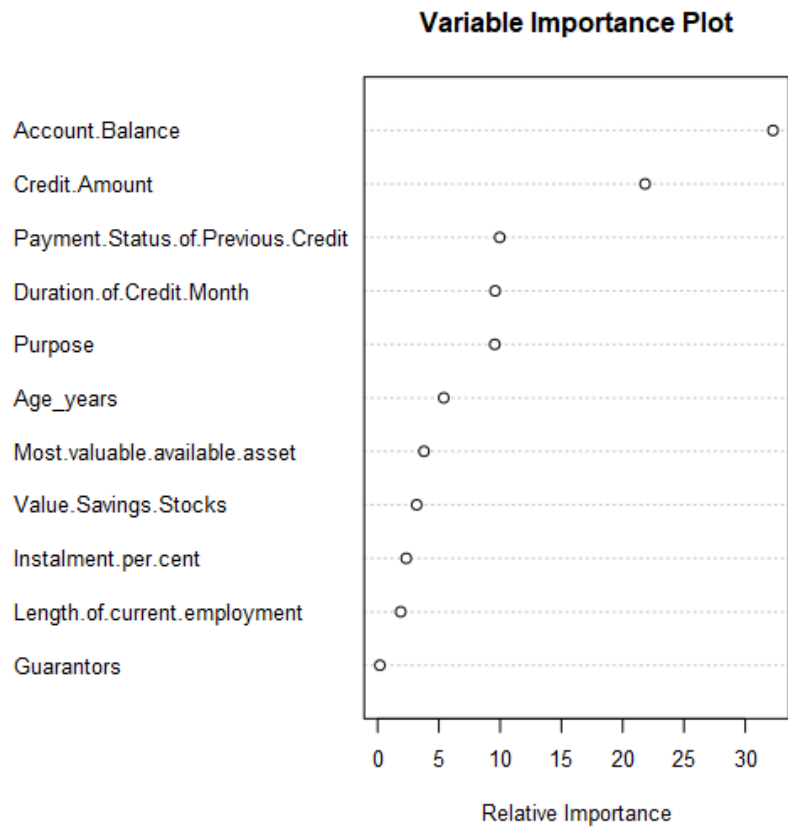
#### Variable Importance



c) *Modelo Floresta*



d) *Boosted Model*



2. Valide seu modelo em relação ao conjunto de Validação. Qual foi a porcentagem geral de precisão? Mostre a matriz de confusão. Existe algum viés (bias) nas previsões do modelo?

R: Ao analisar o relatório de precisão onde fizemos a comparação dos modelos, podemos ver que o modelo Floresta tem 82% de precisão. Este modelo apresenta ser melhor do que os outros, sendo que o modelo “Árvore\_decisão” teve uma precisão de 74%, o modelo “Passo\_a\_Passo” 76% e o modelo “Boosted” 78%.

Na matriz de confusão podemos ver quantos “Creditworthy” e “Non-Creditworthy” foram previstos corretamente. Novamente o modelo Floresta se mostrou melhor aqui, prevendo corretamente 102 registros de 105 como “Creditworthy”. Os outros modelos também classificaram melhor os “Creditworthy” em comparação com os “Non-Creditworthy”. Percebemos que os “Non-Creditworthy” foram um pouco mais difíceis de prever, isso aconteceu porque temos muito mais valores “Creditworthy” que valores “Non-Creditworthy”.

| Model Comparison Report |          |        |        |                       |                           |
|-------------------------|----------|--------|--------|-----------------------|---------------------------|
| Fit and error measures  |          |        |        |                       |                           |
| Model                   | Accuracy | F1     | AUC    | Accuracy_Creditworthy | Accuracy_Non-Creditworthy |
| Árvore_decisão          | 0.7467   | 0.8273 | 0.7054 | 0.8667                | 0.4667                    |
| Floresta                | 0.8200   | 0.8831 | 0.7363 | 0.9714                | 0.4667                    |
| Boosted                 | 0.7867   | 0.8632 | 0.7524 | 0.9619                | 0.3778                    |
| Passo_a_Passo           | 0.7600   | 0.8364 | 0.7306 | 0.8762                | 0.4889                    |

| Confusion matrix of Boosted |                     |                         |
|-----------------------------|---------------------|-------------------------|
|                             | Actual_Creditworthy | Actual_Non-Creditworthy |
| Predicted_Creditworthy      | 101                 | 28                      |
| Predicted_Non-Creditworthy  | 4                   | 17                      |

| Confusion matrix of Floresta |                     |                         |
|------------------------------|---------------------|-------------------------|
|                              | Actual_Creditworthy | Actual_Non-Creditworthy |
| Predicted_Creditworthy       | 102                 | 24                      |
| Predicted_Non-Creditworthy   | 3                   | 21                      |

| Confusion matrix of Passo_a_Passo |                     |                         |
|-----------------------------------|---------------------|-------------------------|
|                                   | Actual_Creditworthy | Actual_Non-Creditworthy |
| Predicted_Creditworthy            | 92                  | 23                      |
| Predicted_Non-Creditworthy        | 13                  | 22                      |

| Confusion matrix of Árvore_decisão |                     |                         |
|------------------------------------|---------------------|-------------------------|
|                                    | Actual_Creditworthy | Actual_Non-Creditworthy |
| Predicted_Creditworthy             | 91                  | 24                      |
| Predicted_Non-Creditworthy         | 14                  | 21                      |

## Step 4: Escrita

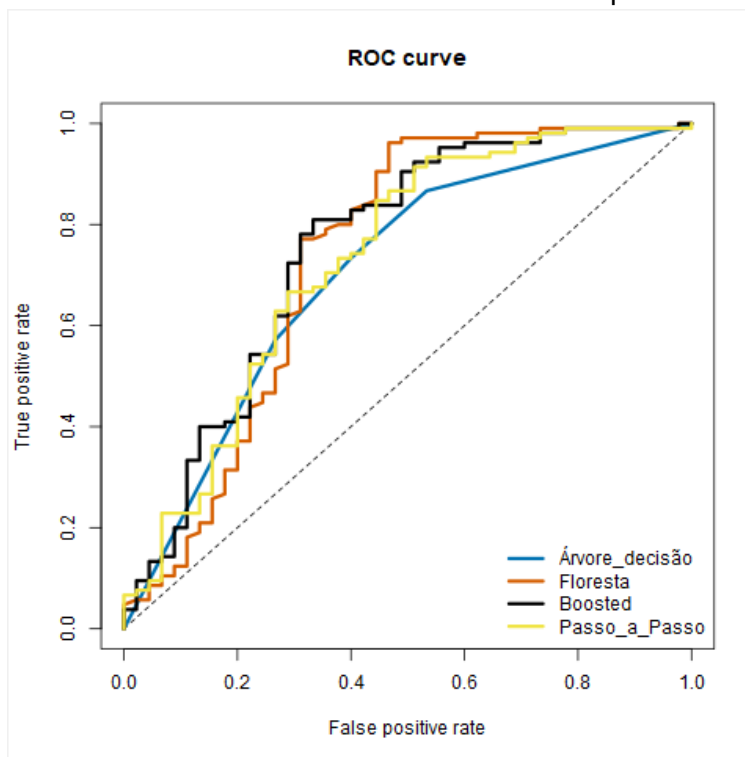
*Decidir sobre o melhor modelo e pontuação de seus novos clientes. Para revisar a consistência, se Score\_Creditworthy for maior que Score\_NonCreditworthy, a pessoa deve ser rotulada como "Creditworthy"*

*Escreva um breve relatório sobre como você criou o seu modelo de classificação e anote quantos dos novos clientes se qualificariam para um empréstimo. (Limite de 250 palavras)*

*Responda estas perguntas:*

1. Qual modelo você escolheu usar? Por favor, justifique sua decisão usando apenas as seguintes técnicas:
  - a. Precisão geral contra o seu conjunto de validação
  - b. Exatidão dentro dos segmentos "Creditworthy" e "Non-Creditworthy"
  - c. Gráfico ROC
  - d. Bias nas Matrizes de Confusão

R: Usei o Modelo Floresta. Ao analisar o indicador "Accuracy" este foi o modelo mais forte com 82% de precisão. Este modelo apresentou ser melhor do que os outros, sendo que o modelo "Árvore\_decisão" teve uma precisão de 74%, o modelo "Passo\_a\_Passo" 76% e o modelo "Boosted" 78%. O modelo Floresta também apresentou uma melhor exatidão dentro dos segmentos "Creditworthy" e "Non-Creditworthy". Analisando o gráfico ROC abaixo, percebemos que modelo Floresta juntamente com o modelo Boosted tiveram entre os outros modelos a maior taxa de verdadeiros positivos. O modelo Floresta foi escolhido porque em determinado momento a curva permanece mais à esquerda e acima dos outros modelos, dessa forma esse modelo tem uma maior taxa de verdadeiros positivos e uma menor taxa de falsos positivos.





**Nota:** Lembre-se de que seu chefe só se preocupa com a precisão das previsões para os segmentos Creditworth e Non-Creditworthy.

2. Quantos indivíduos são bons pagadores?

R: 407