Projeto 4: Prevendo o Risco de Calote

Passo 1: Entendimento de negócios e dados

Decisões chave:

1. Que decisões precisam ser tomadas?

   Se um empréstimo deve ser concedido ou não.

2. Que dados são necessários para informar essas decisões?

   Ocupação atual, histórico de pagamento, propósito do empréstimo, reserva atual, valor do patrimônio atual, prazo para pagamento e quaisquer outros dados que possam ilustrar a saúde financeira atual do cliente.

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

   Modelo binário.

Passo 2: Construindo o Conjunto de Treinamento

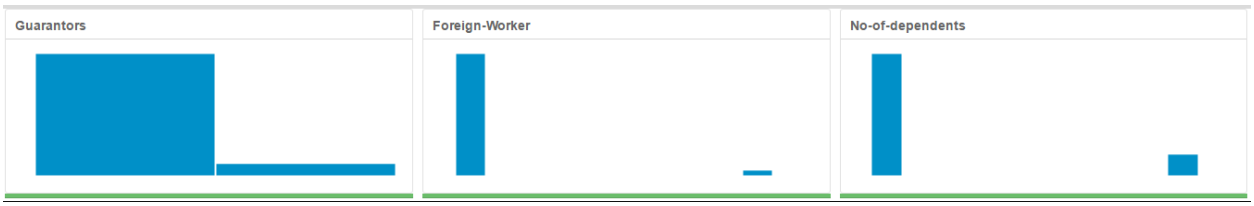
1. Em seu processo de limpeza, quais campos você removeu ou imputou? Por favor, justifique por que você removeu ou imputou esses campos. As visualizações são incentivadas.

   No processo de limpeza foi rodado um *field summary* e uma *association analysis*, levando a exclusão dos campos abaixo:

   1. *Telephone*
   2. *No-of-dependents*
   3. *Guarantors*
   4. *Foreign-worker*
   5. *Occupation*
   6. *Concurrent-Credits*
   7. *Duration-in-Current-address*

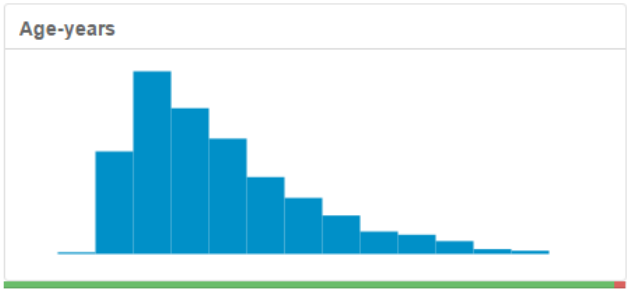
   A variável ***1*** foi excluída por não haver razão lógica de a mesma influenciar no modelo.

   Já as variáveis ***2*** a ***6***, não serão mantidas por possuírem baixa variabilidade, como pode ser visto nos histogramas abaixo.

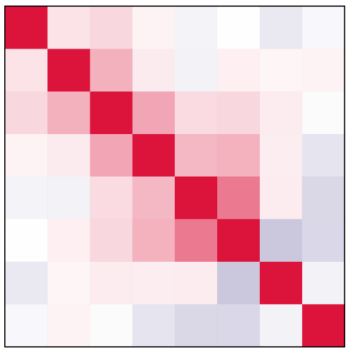


Por fim, foi excluída a variável ***7*** (*Duration-in-Current-address)*, por conta da alta taxa de dados faltantes (69%).

Além da exclusão dos campos citados, a variável *Age-years* apresentou 2% dos registros com dados faltantes, que foram imputados com a média das outras observações, conforme instrução prévia.



Não foi encontrado nenhum caso de alta correlação interna entre as possíveis variáveis preditoras.



Passo 3: Treinar seus Modelos de Classificação

Regressão Logística:

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.*

   As variáveis mais importantes foram *Account.balance, Purpose, Credit.Amout e Instalment.Per.Cent*

6 Coefficients:

7

	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	***
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	.
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
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	*
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.

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

(Dispersion parameter for binomial taken to be 1)

8 Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

Record Report

11 Response: Credit.Application.Result

	LR Chi-Sq	DF	Pr(>Chi-Sq)	
Account.Balance	31.129	1	2.41e-08	***
Payment.Status.of.Previous.Credit	5.687	2	0.05823	.
Purpose	12.225	3	0.00665	**
Credit.Amount	9.882	1	0.00167	**
Length.of.current.employment	5.522	2	0.06324	.
Instalment.per.cent	5.198	1	0.02261	*
Most.valuable.available.asset	3.509	1	0.06104	.

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

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 (bais) nas previsões do modelo?*

   O modelo tem precisão geral de 76%, com leve viés para taxações *Creditworthy.*

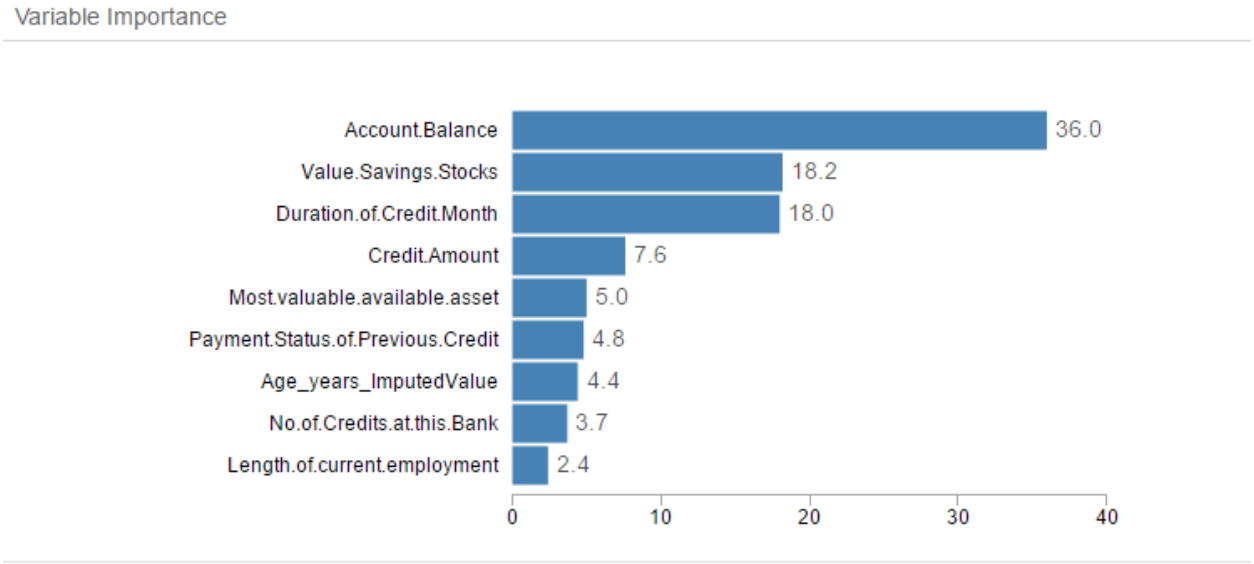
Confusion matrix of SW_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Árvore de Decisão:

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.*

   As variáveis usadas no modelo foram: *Account.Balance Duration.of.Credit.Month Value.Savings.Stocks.*



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 (bais) nas previsões do modelo?*

   O modelo tem precisão geral de 75%, com leve viés para taxações *Creditworthy.*

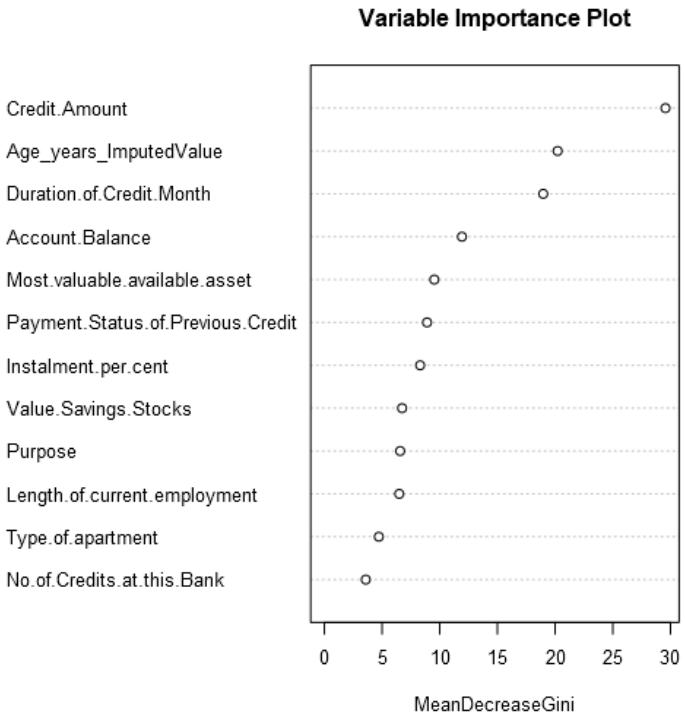
Confusion matrix of DT_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Modelo de Floresta:

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.*

   As variáveis de maior importância foram: *Credit.Amount, Age_Years* e *Duration.of.Credit.Month*.



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 (bais) nas previsões do modelo?*

   O modelo tem precisão geral de 80% e não apresentou viés contra o conjunto de validação.

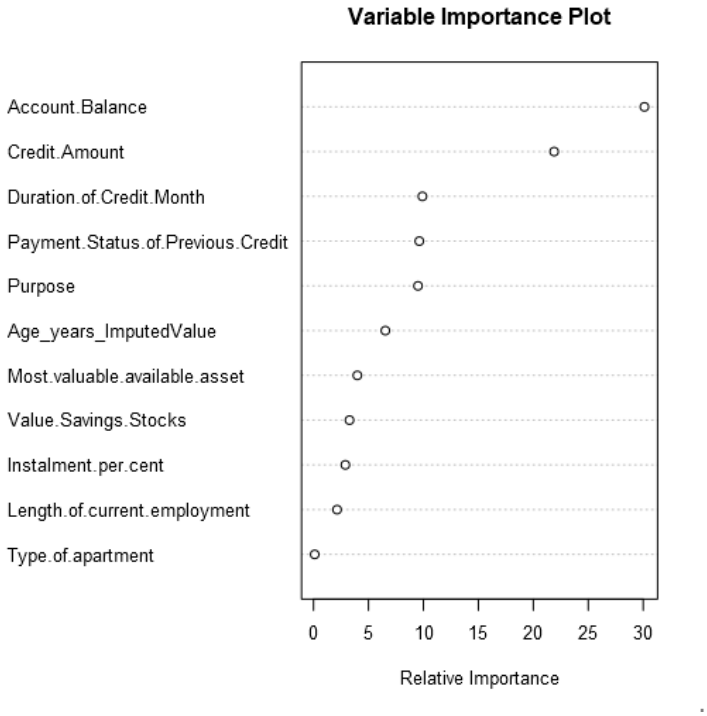
Confusion matrix of FM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Boosted Model:

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*

   As variáveis de maior importância foram: *Account.Balance* e *Credit.Amount.*



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 (bais) nas previsões do modelo?*

   O modelo tem precisão geral de 79% e não apresentou viés contra o conjunto de validação.

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Step 4: Escrita

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

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

Com base na acurácia geral apresentada pelos modelos, a disputa ficou entre o Modelo de Floresta e o Modelo Boosted.

Com isto em mente, ao analisar a acurácia dentro dos segmentos para os dois modelos, observamos que o Modelo de Floresta apresenta melhor acurácia na previsão de clientes *Non-Creditworthy*, reduzindo a chance de concessão de crédito a maus pagadores (falsos positivos).

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0,7467	0,8273	0,7054	0,7913	0,6000
FM_Credit	0,8000	0,8718	0,7379	0,7907	0,8571
BM_Credit	0,7867	0,8621	0,7526	0,7874	0,7826
SW_Credit	0,7600	0,8364	0,7306	0,8000	0,6286

Ao observarmos a matriz de confusão, percebemos que os modelos apresentam resultados bastante similares, sem exibir nenhum viés claro.

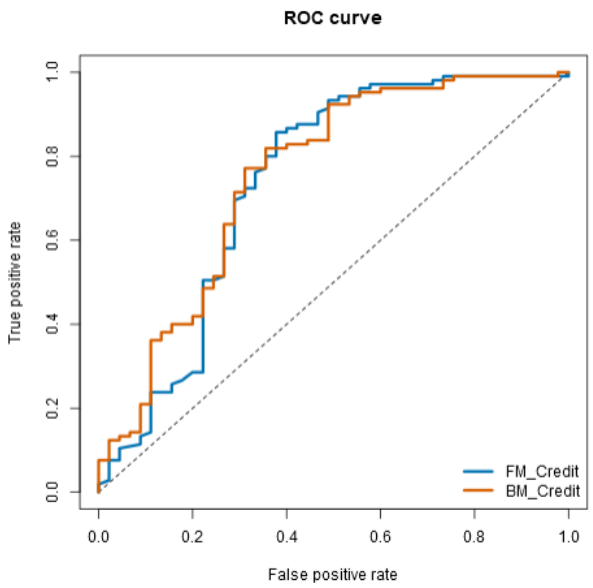
Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of FM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Outro ponto a ser observado é que, apesar de o Modelo Boosted alcançar um patamar alto mais rapidamente na curva ROC, podemos ver que os resultados finais dos AUCs foram bem similares, com leve vantagem para o Modelo Boosted (+0,015)



Assim sendo, acredito que o **Modelo de Floresta seja a melhor opção** por uma leve margem, já que apresenta uma acurácia ligeiramente maior.

2. Quantos indivíduos são bons pagadores?

   Usando o modelo de floresta, 412.