Projeto 4: Prevendo o Risco de Calote

Passo 1: Entendimento de negócios e dados

1. Que decisões precisam ser tomadas?

Consideramos um banco de pequeno porte com dois anos no mercado. A empresa deve implementar uma solução efetiva para automatizar as decisões de aceitar ou rejeitar um pedido de empréstimo, diferenciando clientes com probabilidades de inadimplência. Para chegar ao nosso objetivo vamos testar os diferentes modelos de classificação e escolher o mais eficiente para o problema.

2. Que dados são necessários para informar essas decisões?

- Características socioeconômicas, como idade, gênero, estado civil, numero de dependentes, profissão, salários, entre outras.

- Atividade de credito, como histórico de pagamentos, tiempo de pagamento, preferencias de consumo e hábitos de pagamento.

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

Precisamos usar um modelo binário, ja que o nosso objetivo e prever se o cliente vai sim pagar o empréstimo, ou não vai pagar o empréstimo, estamos lidando com o processo de “score de credito”, onde queremos saber se o cliente e bom ou mal pagador.

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.

Para o sucesso do nosso modelo preditivo, primeiramente devemos realizar uma limpeza dos dados.

“Duration-in-Current-address” e “Age-years” tem dados faltantes, 69% e 2% respectivamente, nesse caso a primeira por ter um alto porcentagem de dados faltantes, vamos a retirar de nosso modelo, no caso de “Age-years”, vamos imputar a mediana (33) aos 2% dos dados, decidi escolher a mediana para evitar os outliers de idade. As variáveis “Concurrent-Credits” , “Occupation”, “Guarantors”, ”No-of-dependents” e “Foreign-Worker” serão removidas do conjunto de dados ja que apresentam baixa variabilidade, “Telephone” sera removida devido a falta de importância como variável preditiva.

Passo 3: Treinar seus Modelos de Classificação

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.

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?

Desenvolvi os modelos de Regressão Logística, Arvores de Decisão, Modelo de Floresta e Boosted Model, tendo com a nossa variável alvo “Credit-Application-Result".

- Regressão Logística - Stepwise:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

	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

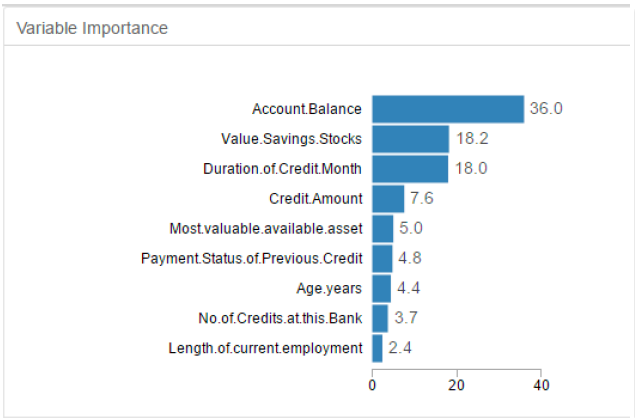
(Dispersion parameter for binomial taken to be 1)

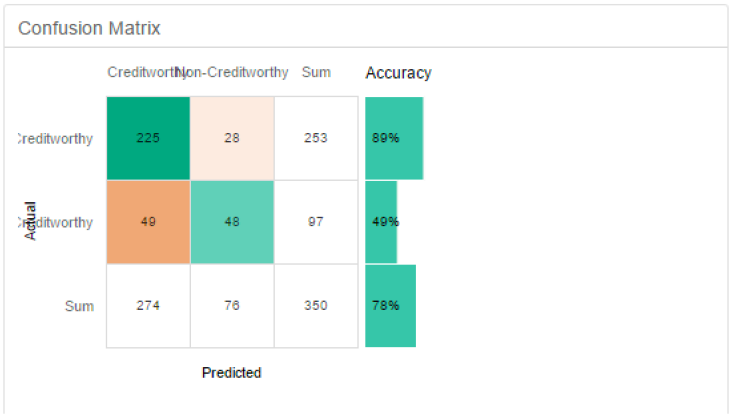
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

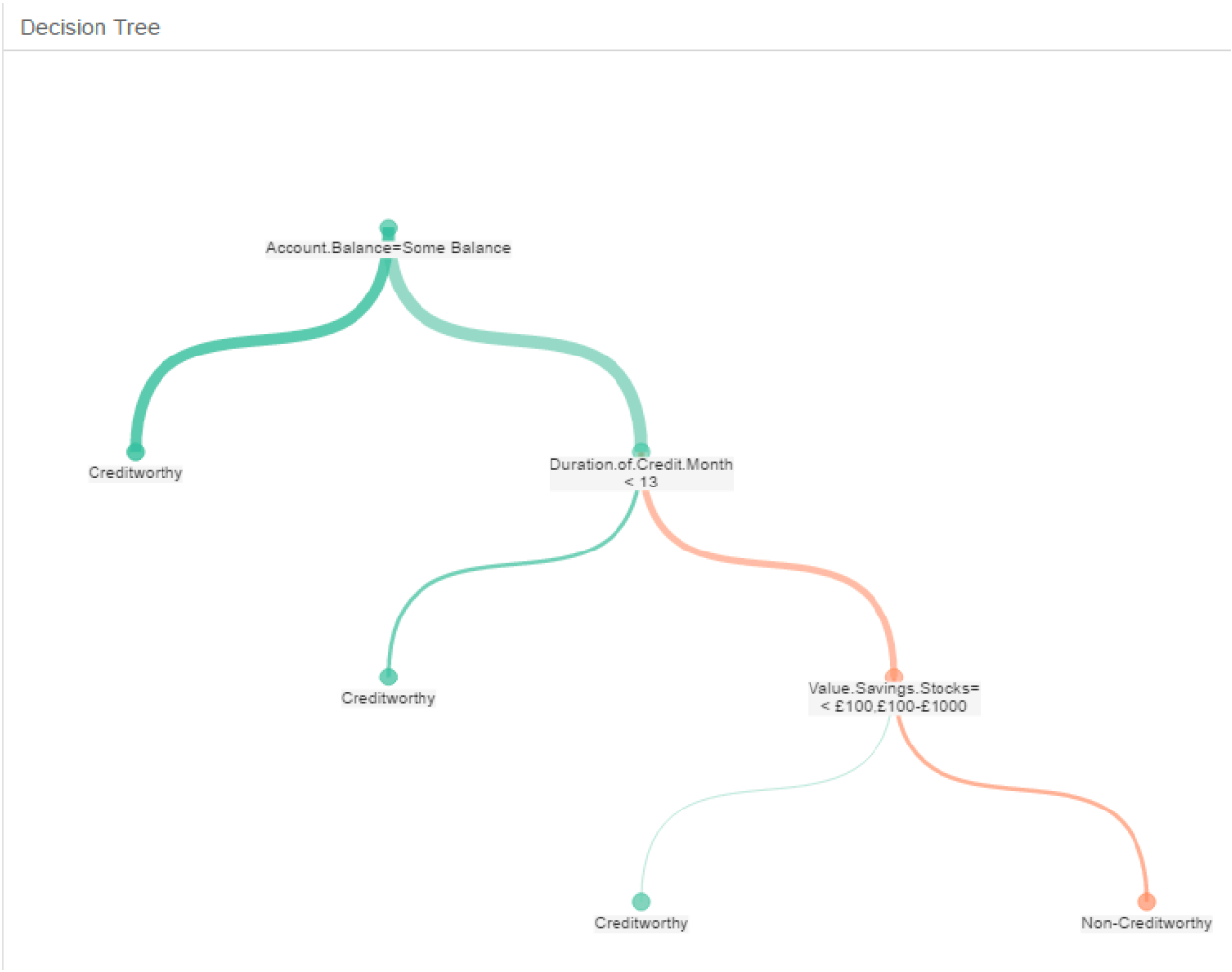
Number of Fisher Scoring iterations: 5

Com os correspondentes coeficientes e p-values, podemos chegar a conclusão que as variáveis mais importantes são "Account-Balance", “Purpose" e “Credit-Amount" tendo p-values menores que 0.05

- Arvore de Decisão:







Podemos apreciar que as variareis mais importantes são "Account-Balance", "Value-Savings-Stocks" e “Duration-of-Credit-Month”. A media geral da matriz de confusão e de 78%, com 89% de Bom pagador e 49% de mal pagador.

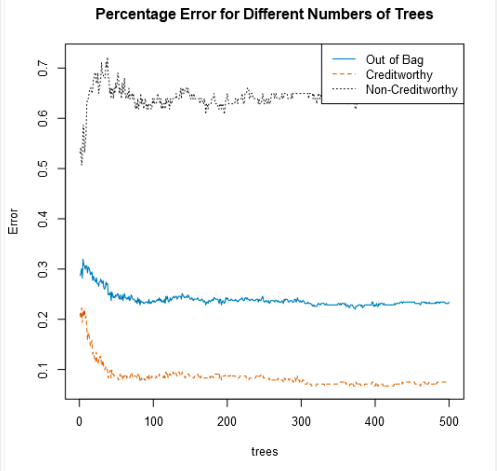
- Modelo de Floresta:

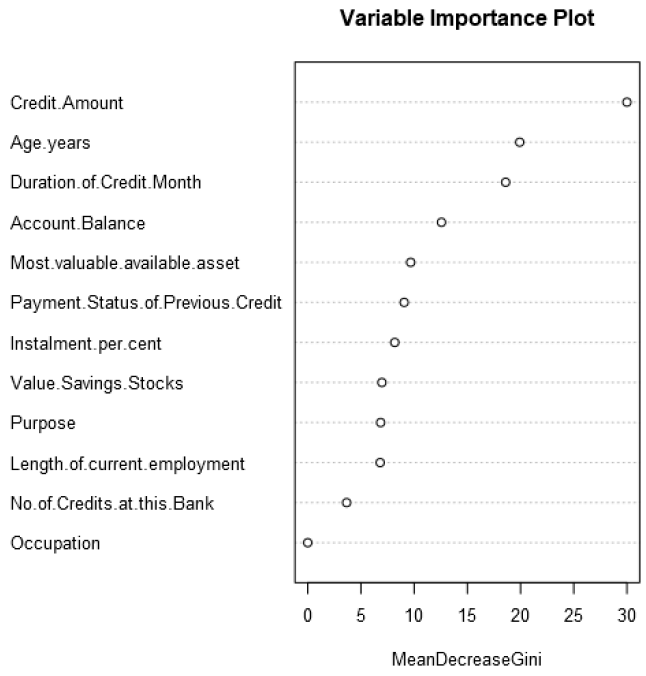
OOB estimate of the error rate: 36.2%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.075	234	19
Non-Creditworthy	0.649	63	34

Plots



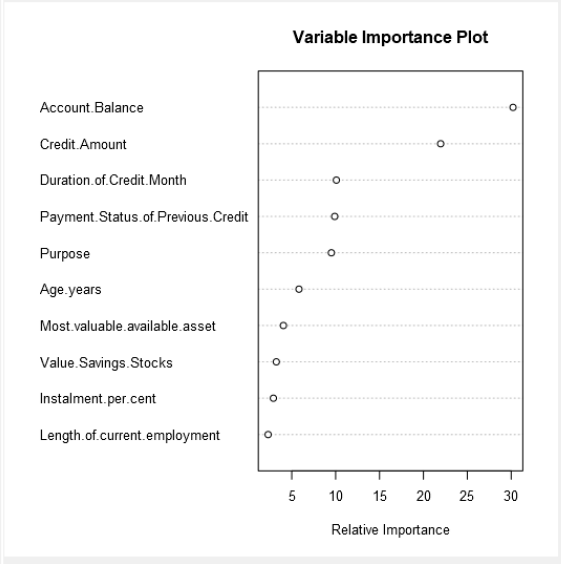


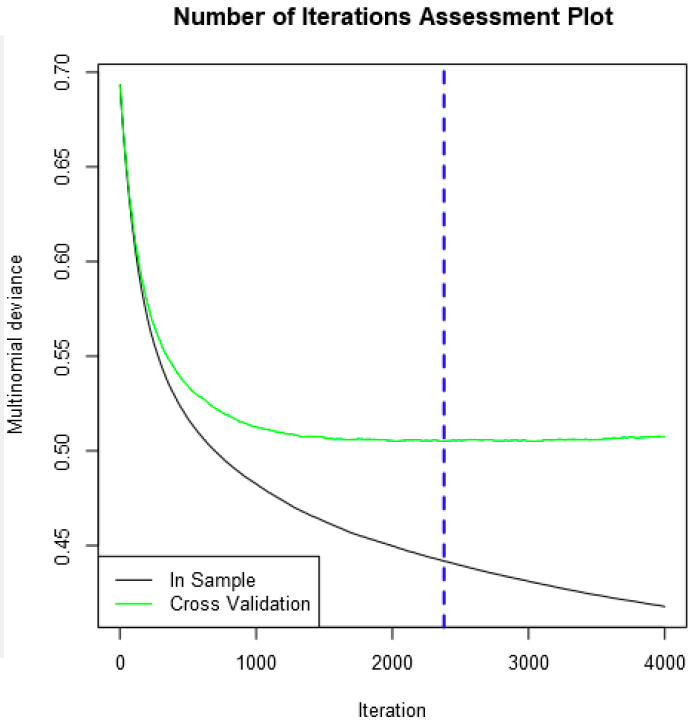
No Gráfico de importância das variáveis podemos apreciar que "Credit-Amount", "Age- year" e “Duration-of-Credit-Month”, são as mais importantes.

- Boosted Model

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2379

Plots:





No Gráfico de importância das variáveis podemos apreciar que "Account-Balance", "Credit-Amount" e "Duration-of-Credit-Month", são as mais importantes.

- Validando os Modelos:

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CREDITWORTH	0.7467	0.8273	0.7054	0.7913	0.6000
FM_CREDITWORTH	0.8067	0.8745	0.7490	0.8016	0.8333
BM_CREDITWORTH	0.7800	0.8584	0.7524	0.7813	0.7727
LR_CREDITWORTH	0.7600	0.8364	0.7306	0.8000	0.6286

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM_CREDITWORTH

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Confusion matrix of DT_CREDITWORTH

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

Confusion matrix of FM_CREDITWORTH

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	25
Predicted_Non-Creditworthy	4	20

Confusion matrix of LR_CREDITWORTH

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

DT = Arvore de Decisão, FM = Modelo de Floresta, BM = Boosted Model, LR = Regressão Logística.

LR = Regressão Logística:
Para a regressão logística a precisão geral do modelo é de 76%. A precisão para os Pagadores e de 80%, mas a dos Não Pagadores e de 60%, tornando o modelo inviável, devido a alta taxa de erro de predição de falsos positivos.

BM = Boosted Model:
O Boosted Model tem um comportamento das classes mais estável, 78% para pagadores e 77% para não pagadores, mas a acurácia em media não e muito boa com 78%, não seria a melhor escolha de modelo.

DT = Arvore de Decisão:
Em este modelo a precisão geral e de aproximadamente 75%, com 79% para os pagadores, e 60% para os não pagadores, por tanto não e um modelo viável devido a grande quantidade de falsos positivos que pode gerar.

FM = Modelo de Floresta:
Este modelo apresenta o melhor performance comparado aos outros, com uma acurácia geral

de 81% aproximadamente, com 80% de acurácia para os pagadores, e 83% para os não pagadores, o modelo e estável e não tendencioso, e também o que menor taxa de erro para falsos positivos apresenta, sendo o modelo escolhido para esta tarefa.

Step 4: Escrita

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

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CREDITWORTH	0.7467	0.8273	0.7054	0.7913	0.6000
FM_CREDITWORTH	0.8067	0.8745	0.7490	0.8016	0.8333
BM_CREDITWORTH	0.7800	0.8584	0.7524	0.7813	0.7727
LR_CREDITWORTH	0.7600	0.8364	0.7306	0.8000	0.6286

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM_CREDITWORTH

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Confusion matrix of DT_CREDITWORTH

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

Confusion matrix of FM_CREDITWORTH

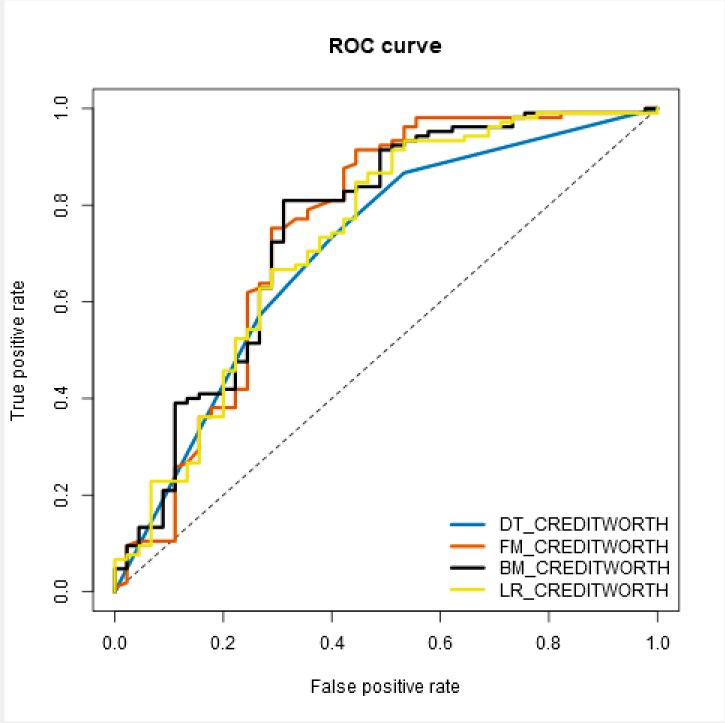
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	25
Predicted_Non-Creditworthy	4	20

Confusion matrix of LR_CREDITWORTH

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

DT = Arvore de Decisão, FM = Modelo de Floresta, BM = Boosted Model, LR = Regressão Logística.

Com base na comparação entre os modelos, podemos destacar, para este problema, que o Modelo de Floresta e quem melhor prediz os resultados com base na acurácia de 0.8067, sendo maior que os outros modelos. A acurácia para os creditworthy e de 0.8016 e para os NonCreditworthy vai para 0.8333. Para esta tarefa, apresentando maior risco para o negocio, o modelo Floresta apresenta uma maior acurácia que os outros modelos, levando assim a menos falsos positivos, ou seja, evitar aprovar credito para um não pagador, evitando assim o risco de calote.



Com o gráfico ROC podemos apreciar o falso positivo e o verdadeiro positivo dos diferentes modelos.

	Area Under the ROC curve
Arvore de Decisão:	0.7054
Modelo Floresta	0.7490
Boosted Model	0.7524
Regressão Logistica	0.7306

2. Quantos indivíduos são bons pagadores?

Os clientes classificados como “Creditworthy" somam 409.