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)

Decisões chave:

Responda estas perguntas

1. Que decisões precisam ser tomadas?

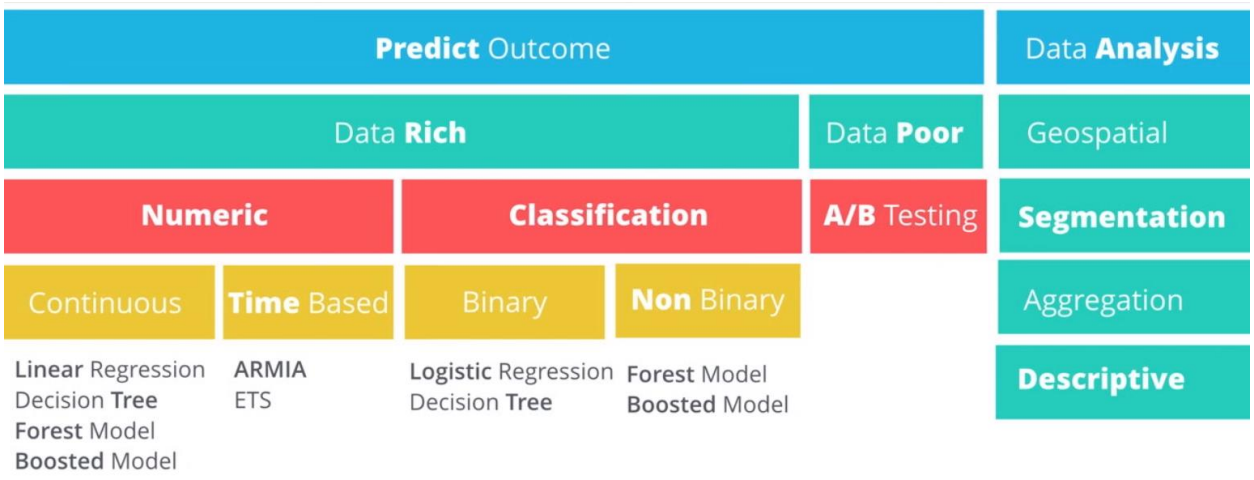
R: Objetivo é identificar se os clientes que solicitaram os empréstimos merecem crédito ou não.

2. Que dados são necessários para informar essas decisões?

R: Dados históricos sobre os empréstimos feitos, como por exemplo saldo da conta e valor de crédito que estão no arquivo “credit-data-training.xlsx” para treinar nosso modelo e a lista de clientes a serem processados (arquivo “customers-to-score.xlsx”).

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

R: Usando o diagrama abaixo, podemos chegar a conclusão que utilizaremos o modelo de classificação binário pois estaremos prevendo algo com dados ricos em informações e classificar se o cliente merece ou não crédito.



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: Quando resumi todos os campos dos dados, variável “Duration-in-Current-address” tinha 69% dos dados faltantes. Sendo assim foi removido esse campo.

Embora a variável “Age-years” tenha 2% dos dados ausentes, decidi colocar o valor da mediana nesses campos pois está mais inclinado para a esquerda o gráfico nesse campo e de se tratar de um dos campos importantes para o treinamento do modelo.

Os campos “Concurrent-Credits” e “Occupation” foram removidos pois tinha somente 1 valor igual para todos os registros.

Os campos “Guarantors”, “Foreign-Worker” e “No-of-dependents” foram removidos pois mostraram baixa variabilidade, onde mais de 80% dos dados se inclinam para um valor.

O campo “Telephone” foi removido devido a sua irrelevância.

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.*

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?*

Você deve ter quatro conjuntos de perguntas respondidas. (Limite de 500 palavras)

1) Logistic Regression (Stepwise)

Utilizando o campo “Credit-Application-Result” como variável target, os campos “Account-Balance”, “Purpose” e “Credit-Amount” foram os 3 campos mais significativos com o p-value menor que 0.05.

Record Report

Report for Logistic Regression Model Stepwise_Results

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Credit.Amount + Instalment.per.cent + Length.of.current.employment + Most.valuable.available.asset + Payment.Status.of.Previous.Credit + Purpose, family = binomial(logit), data = the.data)

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	***
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

(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

Type II Analysis of Deviance Tests

A precisão geral é de 76%, enquanto a precisão de prever se o cliente merece é 80% e 62.9% para aqueles que não merecem. Modelo tende a prever melhor os que merecem o crédito.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Results	0.7600	0.8364	0.7306	0.8000	0.6286

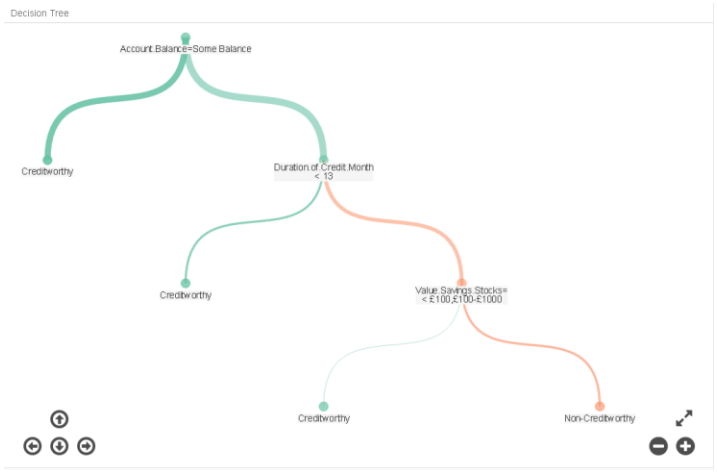
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of 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)

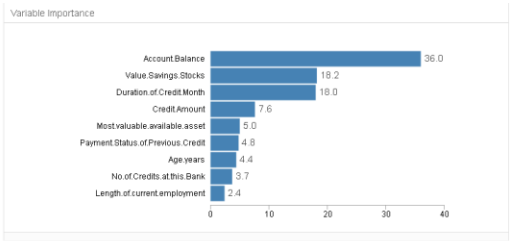
Confusion matrix of Stepwise_Results

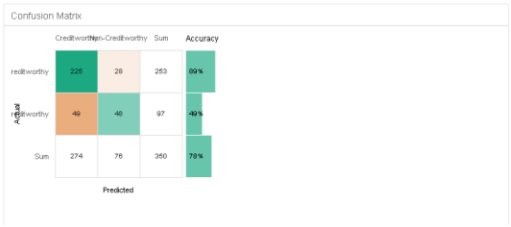
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

2) Decision Tree

Utilizando o campo “Credit-Application-Result” como variável target, os campos “Account-Balance”, “Value.Savings.Stocks” e “Duration.of.Credit.Month” foram os 3 campos mais significativos.







A precisão geral é de 74.67%, enquanto a precisão de prever se o cliente merece é 79.13% e 60% para aqueles que não merecem. Modelo tende a prever melhor os que merecem o crédito.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Results	0.7467	0.8273	0.7054	0.7913	0.6000

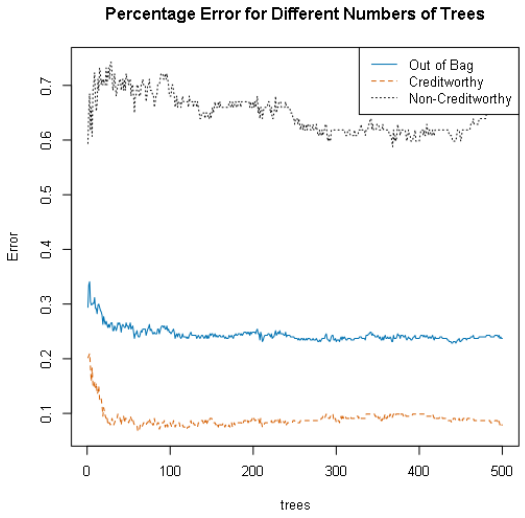
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of 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)

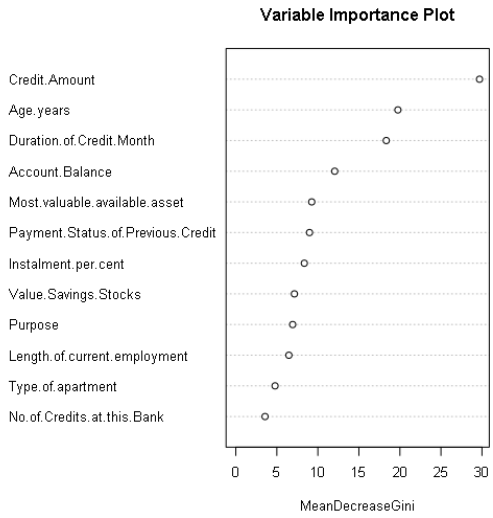
Confusion matrix of DT_Results

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

3) Forest Model

Utilizando o campo “Credit-Application-Result” como variável target, os campos “Credit-Amount”, “Age-years” e “Duration-of-Credit-Month” foram os 3 campos mais significativos.





A precisão geral é de 80.67%, enquanto a precisão de prever se o cliente merece é 79.69% e 86.36% para aqueles que não merecem. Modelo tende a prever melhor os que não merecem o crédito.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Results	0.8067	0.8755	0.7423	0.7969	0.8636

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 FM_Results

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

4) Boosted Model

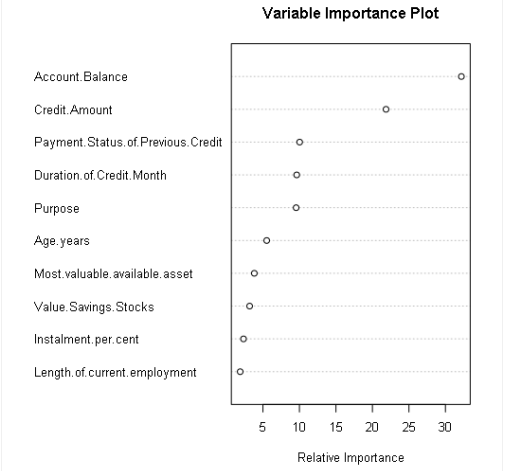
Utilizando o campo “Credit-Application-Result” como variável target, os campos “Account-Balance” e “Credit-Amount” foram os campos mais significativos.

Report for Boosted Model Bosted_Results

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:



A precisão geral é de 78.67%, enquanto a precisão de prever se o cliente merece é 78.29% e 80.95% para aqueles que não merecem. Modelo tende a prever melhor os que não merecem o crédito.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Bosted_Results	0.7867	0.8632	0.7524	0.7829	0.8095

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 Bosted_Results

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

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:

R: Foi escolhido o modelo “Forest Model”, pois ofereceu uma precisão geral maior que 80% no conjunto de dados de validação.
O modelo “Forest Model” atinge uma taxa positiva, pois a diferença de precisão entre aqueles que merecem e não merecem são mais ou menos iguais. O que torna menos tendencioso em comparação com os demais modelos. Evitando assim a não emprestar o dinheiro a clientes com alta probabilidade de inadimplência, garantido oportunidade para aqueles clientes que mereçam os empréstimos. O mais importante é evitar falso positivo por ter maior risco a empresa, pois estaríamos aprovando crédito a um não pagador.
Entre os modelos o “Decision Tree” é o que apresenta maior risco para o negócio, pois tem a menor precisão de prever os que não merecem aprovação no crédito e sem contar também na taxa de precisão para prever um bom pagador.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Results	0.7467	0.8273	0.7054	0.7913	0.6000
FM_Results	0.8067	0.8755	0.7423	0.7969	0.8636
Bosted_Results	0.7867	0.8632	0.7524	0.7829	0.8095
Stepwise_Results	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 Bosted_Results

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of DT_Results

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

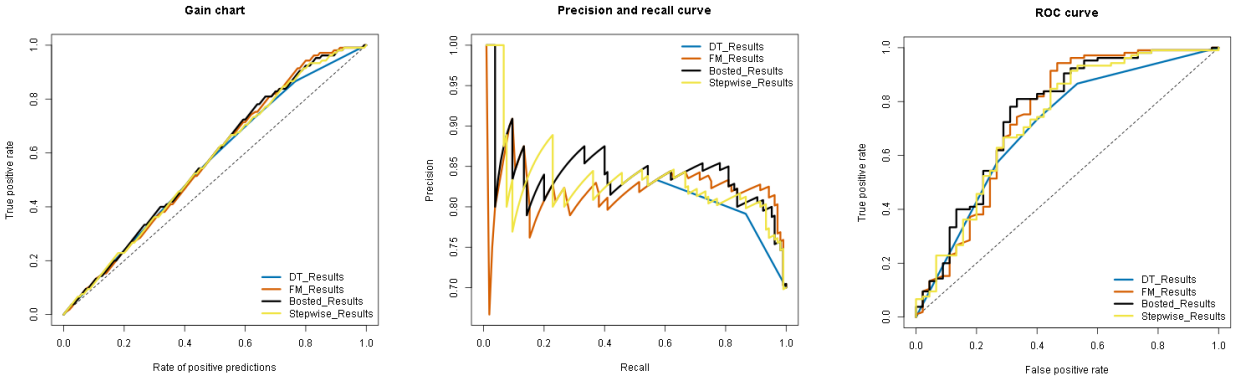
Confusion matrix of FM_Results

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of Stepwise_Results

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

Verificando o gráfico de curva ROC, podemos ver trade-off entre a taxa “verdadeiro positivo” e “falso positivo” entre os modelos e na qual mostra o modelo “Forest Model” como segundo melhor com AUC de 74.23%.



2. Quantos indivíduos são bons pagadores?

R: Há 410 clientes utilizando o modelo “Forest Model”