Prevendo o Risco de Calote

Entendimento de negócios e dados

Que decisões precisam ser tomadas?

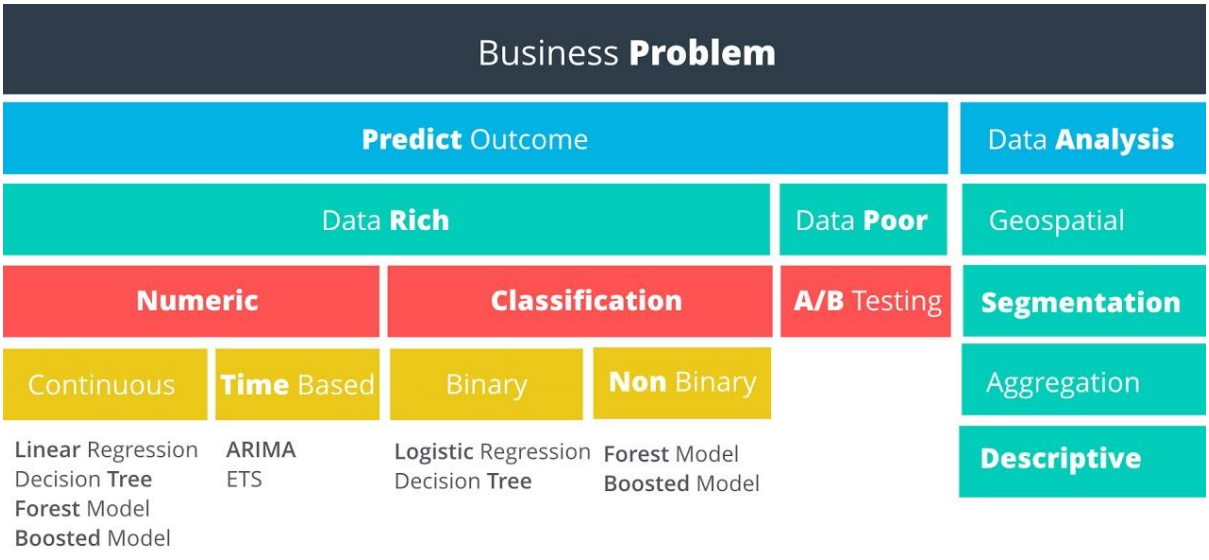
Determinar quais dos novos clientes receberá ou não o empréstimo solicitado.

Que dados são necessários para informar essas decisões?

Para tomar essa decisão, precisamos saber se um cliente é digno de receber o crédito ou não; para determinar isso, há alguns fatores que podem ser considerados. Algumas das condições que podem influenciar na decisão são por exemplo; o tempo no emprego atual, renda, pontuação de crédito, idade, status de pagamento de crédito anterior e suas economias atuais.

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

Usando o guia abaixo, precisamos usar um modelo binário para tomar essa decisão, pois estamos prevendo se o cliente será digno de crédito ou não, ou seja, estamos “rotulando” o cliente.



Construindo o Conjunto de Treinamento

Em seu processo de limpeza, quais campos você removeu ou imputou?.

Os campos removidos estão apresentados abaixo.

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

As variaveis removidas foram:

1. Concurrent-Credits, foi removido por conter baixa variabilidade “Other Banks/Depts”.
2. Guarantors, foi removido por descrepancia nos dados, onde contem 457 “none”.
3. Os campos Foreing-Worker, No-of-dependents, Telephone, foram removidos por não apresentarem uma ligação lógica com o resultado esperado.
4. Occupation, foi removido por conter baixa variabilidade.
5. Duration.in.current.address, foi removido por faltar dados, ou seja muitos dados (null) no segmento.

A imagem abaixo mostra os graficos gerados pela ferramenta resumo de campo, onde podemos visualizar o porque de alguns dados serem removidos.

O campo age_years tem apenas 2.4% dos dados faltando, e precisa ser imputado, e para isso eu usei o valor da mediana de 33 anos; porque a média é uma medida de centralidade muito sensível na presença outliers e se usássemos ela poderia estar enviesando os resultados, já a mediana é um pouco menos sensível.
Outro detalhe a ser considerado é, se imputar os dados faltantes com o valor da média, e depois calcular a média novamente o resultado será o mesmo nas duas médias.

Treinar seus Modelos de Classificação

Para treinar os modelos, os dados foram divididos em 70% para treino e 30% para validação.

Regressão logística + Stepwise

Quais variáveis preditoras são significativas ou as mais importantes?

As variáveis significativas segundo este modelo estão com o * mostrado na imagem abaixo, junto com o p-values.

Coeficientes:

	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	0.3127022	4.587e-01	0.6817	0.49545	
Length.of.current.employment< 1	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

(Parâmetro de dispersão para binomial obtido para ser 1)

Desvio nulo: 413.16 em 349 graus de liberdade
Desvio residual: 328.55 em 338 graus de liberdade
R quadrada McFadden: 0.2048, critério de informações Akaike 352.5

Validação

A precisão do modelo foi de 0.7600

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stp_credit	0.7600	0.8364	0.7306	0.8762	0.4889

Matrix de confusão.

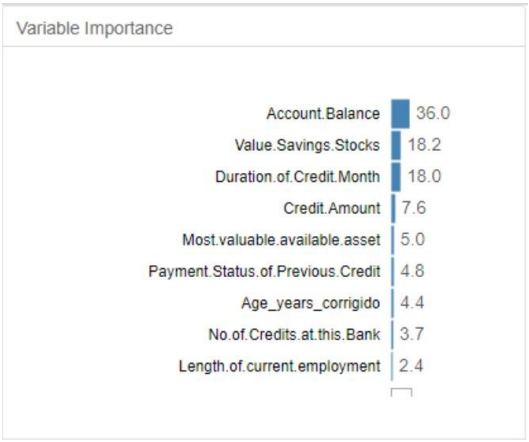
Confusion matrix of stp_credit

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

A precisão de um modelo "*Regressão logística + Stepwise*" em predizer corretamente os indivíduos "creditworthy" é de 92% e a precisão na previsão de indivíduos "noncreditworthy" foi de 22%. Isso significa que esse modelo em particular tem um viés com relação a prever corretamente os indivíduos "creditworthy" porque sua precisão nesse segmento é relativamente mais alta do que na "noncreditworthy".

Decision Trees

Quais variáveis preditoras são significativas ou as mais importantes?



As variáveis de maior importância são: Account balance, value.saving.stokcs e duration.of.credit.month.

Validação

A precisão do modelo foi de 0.7467

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_credit	0.7467	0.8273	0.7054	0.8667	0.4667

Matriz de confusão

Confusion matrix of DT_credit

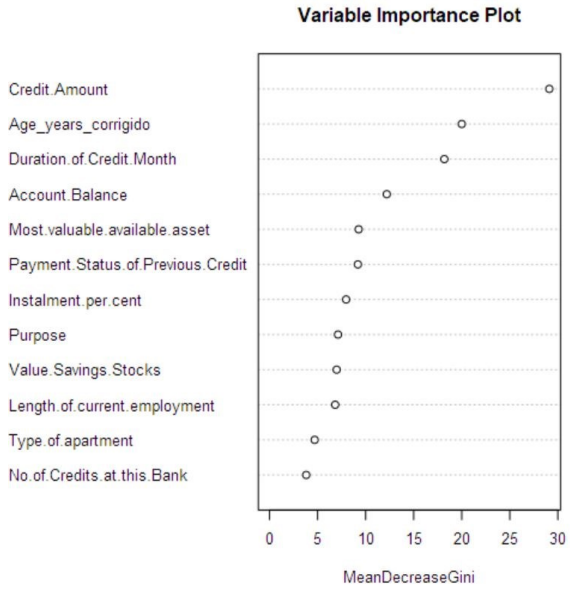
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

A precisão do modelo “*Decision trees”* em predizer corretamente os indivíduos "creditworthy" é de 91% e a precisão na previsão de indivíduos "noncreditworthy” foi de 21%. Isso significa que esse modelo em particular tem um viés com relação a prever corretamente os indivíduos "creditworthy" porque sua precisão nesse segmento é mais alta do que na “noncreditworthy”.

Forest Model

Quais variáveis preditoras são significativas ou as mais importantes?

As variáveis mais importantes são: Credit.amount, age_years e duration.of.credit.month.



Validação

Precisão foi de 0.8067 = 81%

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_credit	0.8067	0.8755	0.7440	0.9714	0.4222

Matriz de confusão, mostrada abaixo.

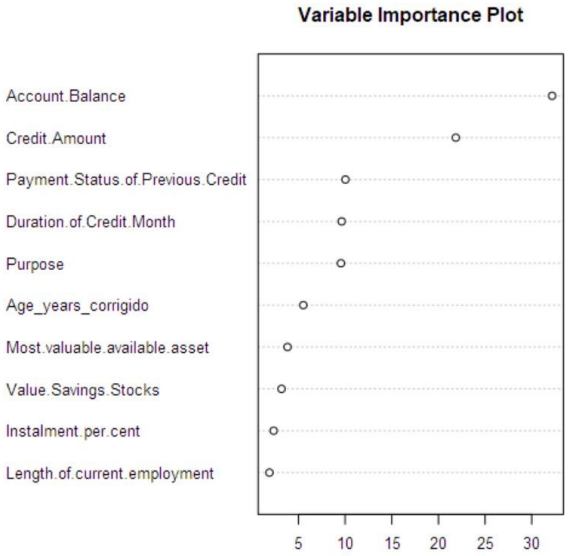
Confusion matrix of FM_credit

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

A precisão do modelo “*Forest model”* em predizer corretamente os indivíduos "creditworthy" foi de 102% e a precisão na previsão de indivíduos "noncreditworthy” é de 19%. Isso significa que esse modelo em particular tem um viés com relação a prever corretamente os indivíduos "creditworthy" porque sua precisão nesse segmento é muito mais alta do que na “noncreditworthy”.

Boosted model

Quais variáveis preditoras são significativas ou as mais importantes?
As variaveis de maior importância são: Account.balance, credit.amount e payment.status.of.previous.credit.



Validação

A precisão foi de 0.7867 = 79%

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Btd_credit	0.7867	0.8632	0.7524	0.9619	0.3778

Matriz de confusão

Confusion matrix of Btd_credit

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

A precisão do modelo “*Boosted model*” em predizer corretamente os indivíduos "creditworthy" foi de 101% e a precisão na previsão de indivíduos "noncreditworthy” é de 17%. Isso significa que esse modelo em particular tem um viés com relação a prever corretamente os indivíduos "creditworthy" porque sua precisão nesse segmento é mais alta do que na “noncreditworthy”.

Escrita

Qual modelo você escolheu usar?

O modelo escolhido foi o Forest Model, por apresentar uma boa Accuracy, F1 alto, e porcentagem de 97% em predizer corretamente.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Árvore_de_decisão	0.7467	0.8273	0.7054	0.8667	0.4667
Floresta_de_Decisão	0.8067	0.8755	0.7431	0.9714	0.4222
Modelo_impulsionador	0.7867	0.8632	0.7524	0.9619	0.3778

Precisão geral contra o seu conjunto de validação

A precisão geral do modelo foi de 0.083 para creditworthy e 0.639 para non-creditworthy como mostra a matriz de confusão do modelo, abaixo.

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.083	232	21
Non-Creditworthy	0.639	62	35

Contra a 0.8067 do modelo de validação, onde 0.9714 para creditworthy e 0.4222 para non-creditworthy.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_credit	0.8067	0.8755	0.7440	0.9714	0.4222

Podendo afirmar que houve um ganho de 0.8884 para o segmento creditworthy e 0.3583 para o non-creditworthy.

Exatidão dentro dos segmentos "Creditworthy" e "Non-Creditworthy"

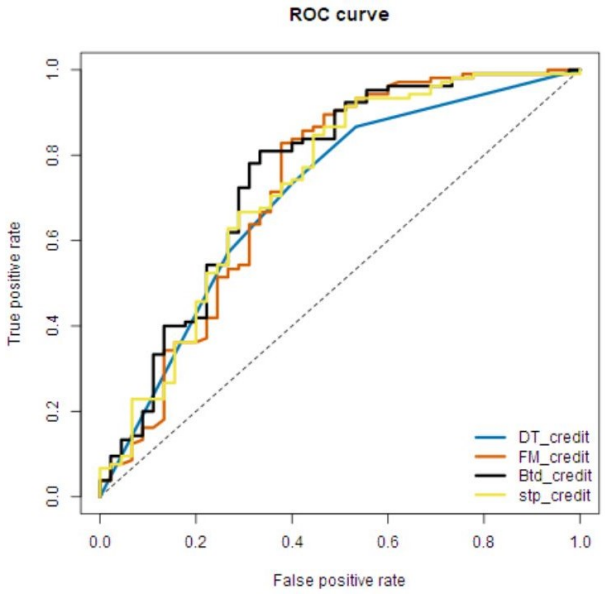
A exatidão é explicada pela matriz de confusão, onde a precisão de creditworthy foi de 102 e de non-creditworthy de 19

Confusion matrix of FM_credit

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

Gráfico ROC

Com o gráfico ROC, podemos observar que a modelo floresta é a linha mais alta para a maior parte do gráfico e também com maior permanência na parte mais alta, o que significa que estamos obtendo uma taxa mais alta de positivo-real vs. falso-positivo. Isso é importante para uma tomada de decisão/predição correta, porque não queremos conceder empréstimos a pessoas que não são dignas do crédito



Bias nas Matrizes de Confusão

A precisão do modelo *Forest Model* em predizer corretamente os indivíduos "creditworthy" é de 102% e a precisão na previsão de indivíduos "noncreditworthy" de 19%. Isso significa que esse modelo em particular tem um viés com relação a prever corretamente os indivíduos "creditworthy" porque sua precisão nesse segmento é muito mais alta do que na "noncreditworthy". Em relação as previsões errôneas temos 3% "creditworthy" classificadas como "noncreditworthy" e 26% de "noncreditworthy" classificadas como "creditworthy".

Confusion matrix of FM_credit

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

Quantos indivíduos são bons pagadores?

O número de pessoas dignas de crédito do conjunto de dados foi de 411 dos 500 novos clientes.