Projeto 4: Prevendo o Risco de Calote

Complete cada seção. Quando estiver pronto, salve seu arquivo como um documento PDF e envie-o aqui: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

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?

A decisão que precisa ser tomada neste problema de negócio é se aprovamos ou não um crédito para os novos clientes, ou seja, precisamos classificar os novos clientes em duas categorias, APROVADO ou NÃO APROVADO.

2. Que dados são necessários para informar essas decisões?

Precisamos ter os dados do passado para podermos treiná-los e aplicar o melhor modelo na base nova. Os dados do passado estão na planilha *credit-data-training.xlsx* e a planilhas que usaremos para aplicar o modelo é *customers-to-score*. Em ambas planilhas, precisaremos dos seguintes campos *Account-Balance, Age-years, Credit-Amount, Credit-Application-Result, Duration-of-Credit-Month, Instalment-per-cent,Length-of-current-employment, Most-valuable-available-asset, No-of-Credits-at-this-Bank, Payment-Status-of-Previous-Credit, Purpose, Type-of-Apartment and Value-Savings-Stocks*

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

Para este problema, precisamos usar um Modelo Binário, sendo que a variável resposta é SIM ou NÃO para a aprovação do crédito ao cliente.

Andy: Ótimo trabalho no passo 1!

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)*

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.

Durante o processo de limpeza, os campos abaixo foram removidos do nosso data set.

Concurrent-Credits, Occupation - Ambos campos têm apenas 1 categoria como resultado.

Guarantors, Foreign Workers, No. of Dependents - Campos com baixa variabilidade.

Phone Number - Variável que não é necessária para a criação dos modelos, nenhuma importância.

Duration in Current Address - Alto missing values, 69%.

Em relação a input de dados, achei necessário fazer a inclusão de informações no campo *Age*, visto que percebemos um missing de 2%. A premissa utilizada no input foi a *Mediana*, pois este indicador minimiza o efeito de extremos.

Andy: Excelente trabalho no passo 2!

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)

a. Logistic Regression + Step Wise

- Considerando que a variável *Credit-Application-Result* é nossa variável target, podemos dizer que as variáveis *Account-Balance, Purpose e Credit-Amount* são as variáives preditoras com maior significância, isso pois como vemos na Figura 1 estas variáveis tem p-valor inferior à 0.05.

- O nosso modelo tem uma acurácia boa, de 76.0%, conforme Figura 2. Enquanto temos uma acurácia ainda maior para *Creditworthy* de 87.62%, porém encontramos que o resultado para *Non-Creditworthy* pode estar enviesado, pois o resultado é muito baixo, apenas 48.89%.

Report for Logistic Regression Model StepWise_Risk

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, 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	***
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)

Figura 1: Report Logistic Regression

Model Comparison Report

Fit and error measures

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

Confusion matrix of StepWise_Risk

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

Figura 2: Model Comparison Report for Stepwise Logistic Regression

b. Decision Tree

- Considerando que a variável *Credit-Application-Result* é nossa variável target, podemos dizer que as variáveis *Account-Balance, Value-Saving-Stocks e Duration-of-Credit-Month* são as variáives preditoras com maior significância, isso pois como vemos na Figura 3 estas variáveis estão como mais importantes na *Variável Importância*.

- O nosso modelo tem uma acurácia boa, de 79.1% (melhor que o modelo anterior), conforme Figura 4. Para as variáveis *Creditworthy e Non-Creditworthy* temos o mesmo cenário do modelo anterior, uma alta acurácia para *Creditworthy*, de 86.67% e baixa para *Non-Creditworthy*, de apenas 46.67%. Portanto, também podemos enviesar a variável *Non-Creditworthy*.

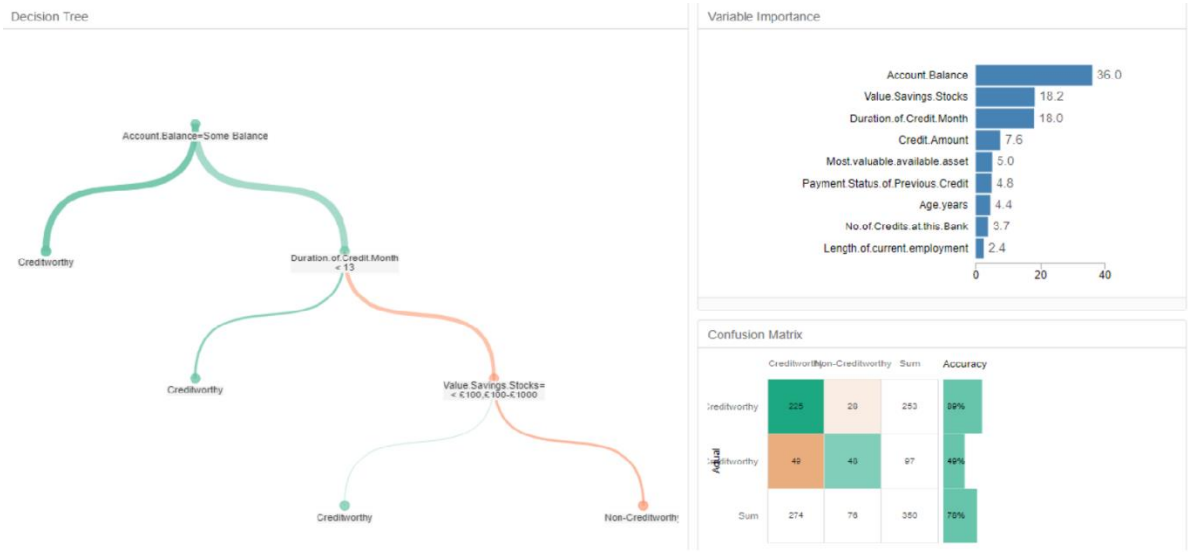

Figura 3: Decision Tree, Variable Importance and Confusion Matrix

Model Comparison Report

Fit and error measures

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

Confusion matrix of DecisionTree_Risk

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

Figura 4: Model Comparison Report for Decision Tree

c. Forest Model

- Considerando que a variável *Credit-Application-Result* é nossa variável target, podemos dizer que as variáveis *Credit-Amount, Age-Years e Duration-of-Credit-Month* são as variáives preditoras com maior significância, isso pois como vemos na Figura 5 estas variáveis estão como mais importantes no gráfico de *Variable Importance Plot*.

- Este modelo mostra uma acurácia geral maior que todos os outros modelos até aqui analisados, de 80.0% conforme Figura 5. Como um alta acurácia para *Creditworthy*, de 96.19% e novamente baixa para *Non-Creditworthy*, de apenas 42.22%. Portanto, também temos enviesamento da variável *Non-Creditworthy*.

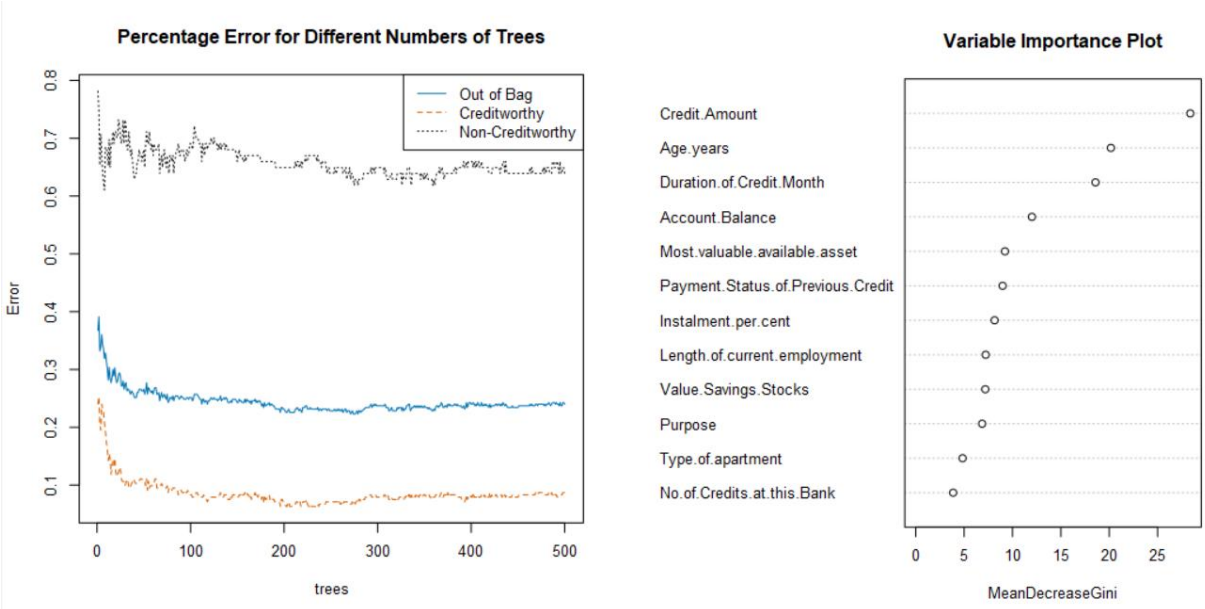

Figura 5: Percentage Error for Different Number of Trees and Variable Importance Plot

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
RandomForest_Risk	0.8000	0.8707	0.7361	0.9619	0.4222

Confusion matrix of RandomForest_Risk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Figura 6: Model Comparison Report for Forest Model

d. Boosted Model

- Considerando que a variável *Credit-Application-Result* é nossa variável target, podemos dizer que as variáveis *Account-Balance, Credit-Amount e Payment-Status-of-Previous-Credit* são as variáives preditoras com maior significância, isso pois como vemos na Figura 7 estas variáveis estão como mais importantes no gráfico de *Variable Importance Plot*.

- Este modelo tem uma acurácia geral de 78.67%, conforme Figura 8. Como um alta acurácia para *Creditworthy*, de 96.16% e novamente baixa para *Non-Creditworthy*, de apenas 37.78%. Também temos enviesamento da variável *Non-Creditworthy*.

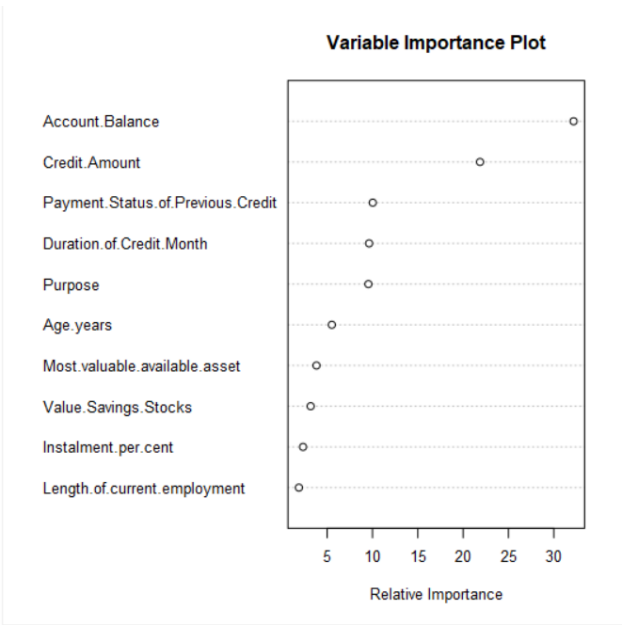

Figura 7: Variable Importance Plot for Boosted Model

Model Comparison Report

Fit and error measures

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

Confusion matrix of Boosted_Risk

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

Figura 8: Model Comparison Report for Boosted Model

Andy: Muito bom trabalho no passo 3!

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:

O modelo escolhido como melhor fit nos dados disponíveis foi o modelo de *Forest Model* pois conta com uma maior acurácia em relação aos outros modelos, conforme podemos ver na Figura 9.

a. Precisão geral contra o seu conjunto de validação

A precisão geral foi a maior em comparação com os outros modelos, de 80.0%.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree_Risk	0.7467	0.8273	0.7054	0.8667	0.4667
RandomForest_Risk	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_Risk	0.7867	0.8632	0.7524	0.9619	0.3778
StepWise_Risk	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boosted_Risk

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

Confusion matrix of DecisionTree_Risk

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

Confusion matrix of RandomForest_Risk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of StepWise_Risk

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

Figura 9: Model Comparison Report for all 4 classification models

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

O segmento *Creditworthy* também conta com uma alta acurácia no modelo *Forest Model* de 96.16%, maior em comparação com os outros modelos. Enquanto o segmento *Non-Creditworthy* não demonstra muita acurácia, apenas de 42.22% porém um dos maiores em relação aos outros modelos.

c. Gráfico ROC

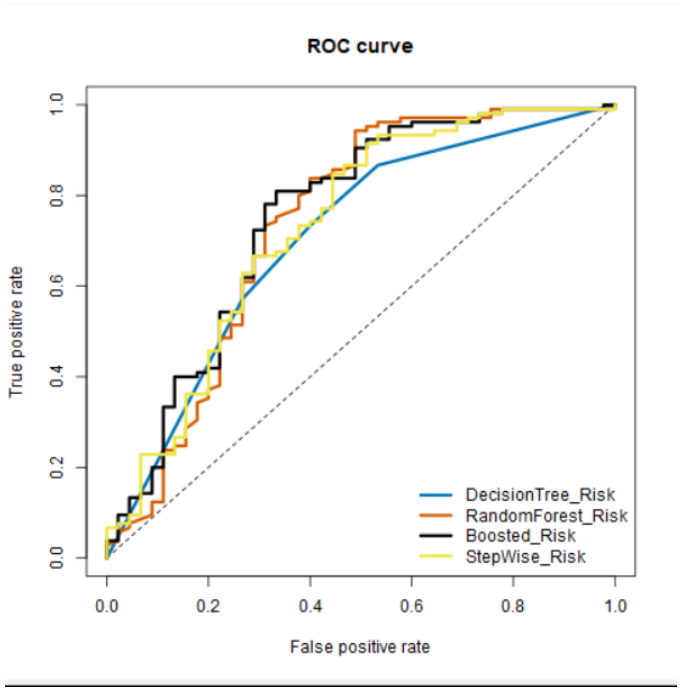

Figura 10: ROC curve for all 4 classification models

d. Bias nas Matrizes de Confusão

Analisando as matrizes de confusão, percebemos que a variável *Creditworthy* ficou bem alinhada com a modelagem, quase em sua maioria o modelo acertou. Enquanto tivemos uma acurácia menor para *Non-Creditworthy* porém podemos mesmo assim considerar que o modelo não tem um bias significativo.

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

2. Quantos indivíduos são bons pagadores?

Por fim, analisando após aplicarmos o nosso modelo na base de novos customers, vimos que **408** pessoas estão aptas a receber o empréstimo do nosso banco, sendo que o critério de seleção foi de pessoas com Score >=50% seriam consideradas como *Creditworthy*.

Antes de Enviar

Por favor, verifique suas respostas contra os requisitos do projeto ditados pela rubrica aqui. Os revisores usarão esta rubrica para classificar seu projeto.

Andy: Sugestão: Adicione uma breve explicação sobre o que a curva ROC representa e como devemos interpretá-la. A razão disso é que algumas pessoas que podem ler o relatório podem não estar familiarizadas com a curva ROC

Andy: Aqui é interessante colocar também um breve explicação de como o gráfico aponta para o modelo floresta como o melhor:

'Ao visualizar o gráfico ROC, pode-se observar que o modelo floresta é a linha mais "alta" para a maior parte do gráfico, o que significa que estamos obtendo uma taxa mais alta de positivo-real vs. falso-positivo. Isso é importante porque não queremos conceder empréstimos a pessoas que não são dignas de crédito.'