PROJETO MTCD E MP

AVALIAÇÃO DE CRÉDITO

Jéssica Generoso - nº125765
Joan Concha - nº126550
Manuel Martins - nº131408
Meda Račaitytė - nº127575

Business Understanding

Problema

Uma empresa financeira procura melhorar a eficiência na gestão de crédito

Solução

Desenvolvimento de um sistema inteligente para segmentar clientes em escalões de pontuação de crédito

Data Understanding

Segmentação dos Clientes

3 níveis:

- Poor;
- Standard;
- Good;

Segmentos dos Dados

- Dados identificadores;
- Informações financeiras;
- Comportamento de crédito;
- Comportamento de pagamento;
- Outras variáveis;

Data Preparation

Funções

- Substituição de valores pela mediana, moda ou valor escolhido
- Formulação de gráficos (gráfico de pontos, histograma e *boxplot*)
- Cálculos e determinação de valores (*outliers*, moda, *upper bound* e *lower bound*)

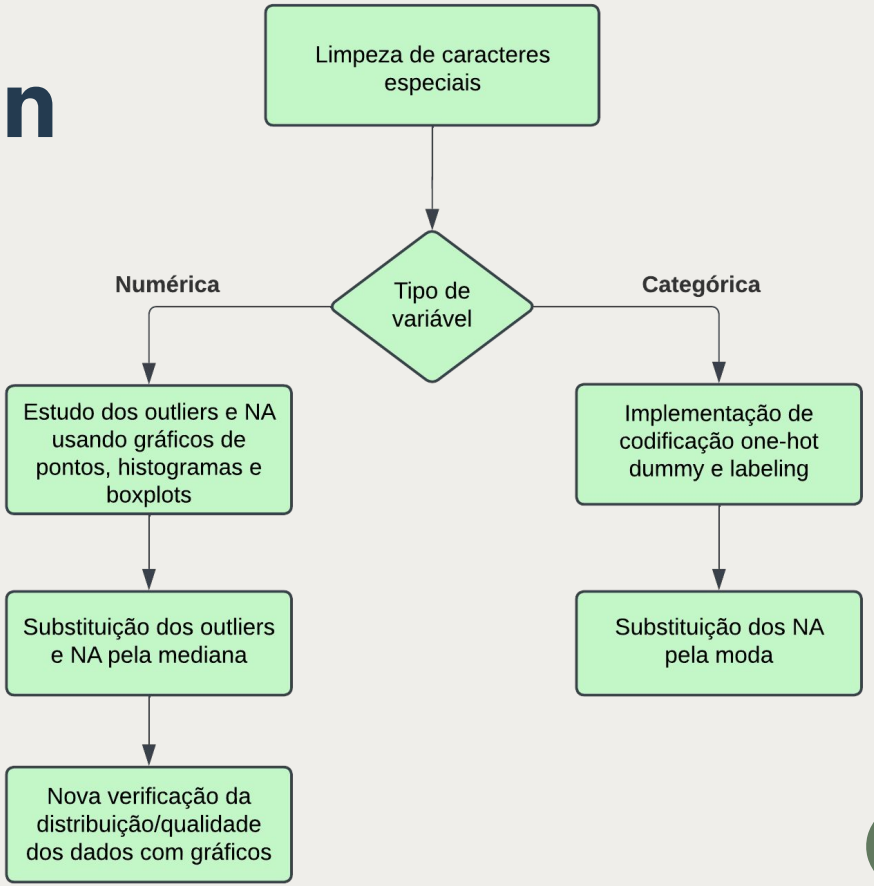
Resultam em...

- Maior eficiência
- Menor complexidade
- Normalização dos métodos

Data Preparation

Outras decisões

- Eliminação das variáveis de identificação (**ID de empréstimo**, **ID de cliente**, **Nome** e **NSS**)



Data Preparation

Variáveis numéricas

- Identificação de valores estranhos ou omissos;
- Substituição de valores estranhos ou omissos pela mediana;
- Determinação dos *lower bound* e *upper bound* (IQR);
- Vizualização dos *outliers*;
- Substituição dos *outliers* pela mediana;
- Visualização dos dados (histograma e *boxplot*);

Exceções

- **Idade;**
- **Mês;**

Data Preparation

Variáveis categóricas

- Identificação de valores estranhos ou omissos;
- Substituição de valores estranhos ou omissos pela moda;
- Implementação da codificação *one-hot-encoding dummy*, *labeling* e *lumping*;

Modeling

Divisão de Dados

- Conjunto de treino ⅔ e conjunto de teste ⅓;
- Classe maioritária “Standard” e minoritária “Good”;
- Criação da amostra estratificada;

Normalização

- Aplicação da padronização;

Criação de samples

- *Under-sample*: 17.796 observações;
- *Over-sample*: 53.346 observações;

Modeling

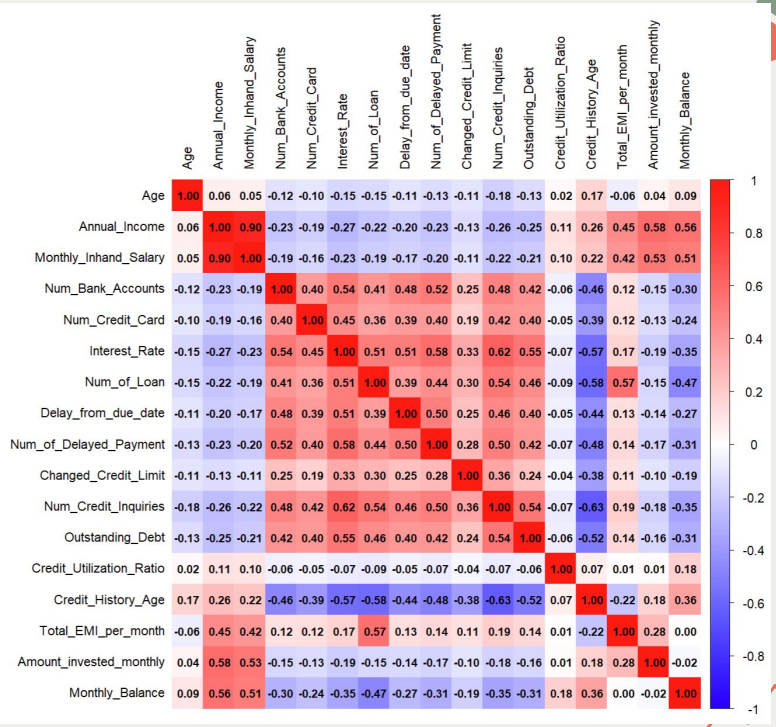
Estudo das correlações - Variáveis numéricas

- Correlação de *Spearman* escolhida para captar relações não-lineares.
- Forte correlação identificada (*cutoff* = 0.75):
  - **Rendimento Anual** vs **Salário Mensal Líquido** (*Spearman* = 0.9).

Cinco maiores diferenças entre os coeficientes de *Pearson* e *Spearman

1ª Variável	2ª Variável	Coeficiente de *Pearson*	Coeficente de *Spearman*	Diferença relativa (%)
Num_of_Loan	Total_EMI_per_month	0,45	0,57	12,13
Annual_Income	Amount_invested_monthly	0,52	0.58	6,59
Num_of_Loan	Outstanding_Debt	0,52	0,46	6,21
Monthly_Inhand_Salar	Amount_invested_monthly	0,48	0,53	5,38
Num_of_Loan	Monthly_Balance	-0,43	-0,47	4,32

Heatmap* de correlações - *Spearman



* *Heatmap* de correlações para *Pearson* nos Anexos.

Modeling

Estudo das correlações - Variáveis categóricas

Nominais (V de *Cramér*)

* Rácio de correlação η (**Tipo de Empréstimo** vs **Número de Empréstimos**) = 0,963737

Ordinais (τ de *Kendall*)

Valores de τ na ordem de 0,03 para todas as variáveis (**Grupo de Idades**, **Mix de Crédito**, e **Avaliação de Crédito**).

Conclusões:

- Possível exclusão das variáveis **Ocupação**, **Grupo de Idades**, **Mix de Crédito**.
- Possível exclusão do **Tipo de Empréstimo**: fortemente correlacionado com o **Número de Empréstimos** (ambas indicadores de empréstimos).

Modeling

Seleção das variáveis importantes - Recursive Feature Elimination (RFE)

Número ótimo de variáveis para a modelação

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD  KappaSD Selected
         5   0.7114 0.5670   0.017406 0.026112
         6   0.7348 0.6022   0.010430 0.015645
         7   0.7421 0.6132   0.010302 0.015451
         8   0.7520 0.6280   0.011645 0.017469
         9   0.7546 0.6318   0.010109 0.015163
        10   0.7532 0.6297   0.011538 0.017307
        11   0.7536 0.6304   0.008324 0.012485
        12   0.7547 0.6321   0.008141 0.012210
        13   0.7565 0.6347   0.009318 0.013976
        14   0.7593 0.6390   0.010789 0.016183        *
        15   0.7562 0.6344   0.008637 0.012955
        38   0.7547 0.6320   0.006356 0.009527
```

Importância das variáveis selecionadas

	Good	Poor	Standard	Overall	var	Variables	Resample
1	86.605379	30.298184	24.8359803	47.246514	Delay_from_due_date	38	Fold02
2	80.005719	28.330615	22.0517793	43.462705	Num_Credit_Card	38	Fold06
3	36.897872	39.870251	27.7885167	34.852213	Num_of_Delayed_Payment	38	Fold09
4	41.117889	4.133939	49.6361248	31.629318	Changed_Credit_Limit	38	Fold06
5	32.645842	32.228333	29.0510903	31.308422	Total_EMI_per_month	38	Fold10
6	32.182460	42.799597	17.6231052	30.868387	Outstanding_Debt	38	Fold07
7	30.480505	24.049243	38.0331900	30.854313	Credit_MixStandard	38	Fold01
8	33.038072	49.830585	9.2175089	30.695389	Interest_Rate	38	Fold07
9	39.715794	28.974291	17.9609927	28.883692	Annual_Income	38	Fold05
10	35.553355	28.515188	19.1158484	27.728131	Monthly_Inhand_Salary	38	Fold05
11	36.826681	14.689230	30.5460026	27.353971	Credit_MixGood	38	Fold02
12	37.530968	28.510482	12.5295157	26.190322	Num_Credit_Inquiries	38	Fold03
13	26.535529	29.232666	17.9979110	24.588702	Num_Bank_Accounts	38	Fold03
14	32.628105	22.863381	8.2603045	21.250597	Age	38	Fold01

- Número ótimo de variáveis preditoras = 14 (12 numéricas e 2 codificadas em *one-hot*).
- Variáveis selecionadas ordenadas pela sua importância (*Overall*).

Modeling

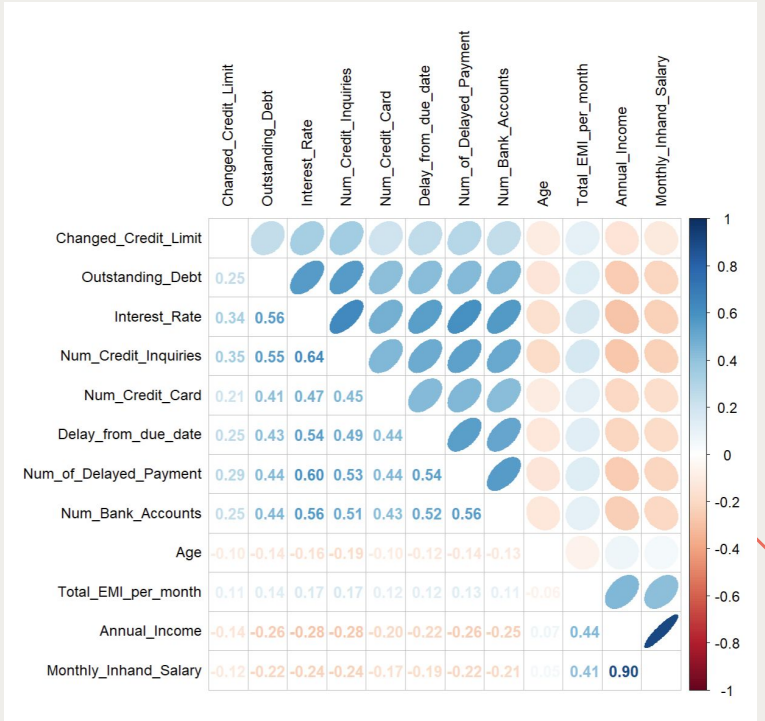
Estudo das correlações (pós-RFE) - Variáveis numéricas

- Exclusão da variável que representa o **Salário Mensal Líquido** (menor importância do que a **Remuneração Anual** para o modelo).

Importância das variáveis selecionadas

	Good	Poor	Standard	Overall	var	Variables	Resample
1	86.605379	30.298184	24.8359803	47.246514	Delay_from_due_date	38	Fold02
2	80.005719	28.330615	22.0517793	43.462705	Num_Credit_Card	38	Fold06
3	36.897872	39.870251	27.7885167	34.852213	Num_of_Delayed_Payment	38	Fold09
4	41.117889	4.133939	49.6361248	31.629318	Changed_Credit_Limit	38	Fold06
5	32.645842	32.228333	29.0510903	31.308422	Total_EMI_per_month	38	Fold10
6	32.182460	42.799597	17.6231052	30.868387	Outstanding_Debt	38	Fold07
7	30.480505	24.049243	38.0331900	30.854313	Credit_MixStandard	38	Fold01
8	33.038072	49.830585	9.2175089	30.695389	Interest_Rate	38	Fold07
9	39.715794	28.974291	17.9609927	28.883692	Annual_Income	38	Fold05
10	35.553355	28.515188	19.1158484	27.728131	Monthly_Inhand_Salary	38	Fold05
11	36.826681	14.689230	30.5460026	27.353971	Credit_MixGood	38	Fold02
12	37.530968	28.510482	12.5295157	26.190322	Num_Credit_Inquiries	38	Fold03
13	26.535529	29.232666	17.9979110	24.588702	Num_Bank_Accounts	38	Fold03
14	32.628105	22.863381	8.2603045	21.250597	Age	38	Fold01

Matriz de correlações - *Spearman*



Modeling - Regressão logística

Pós-RFE com GridSearch

Conjunto de treino

- Conjunto de dados: *Down-sample*
- Categoria de referência: “Good”
- Métricas:
  Accuracy- 69,45%
  Sensitivity- 69,45%
  Specificity- 84,73%
  AUC- 82,37%

Conjunto de teste

- Dados normalizados
- Métricas:
  Accuracy- 64,54%
  Sensitivity- 64,54%
  Specificity- 82,27%
  AUC- 81,91%

Evaluation - Regressão logística

- **Taxa de juro**, **número dos cartões de crédito**, **dívida pendente** e **mix de crédito** foram as variáveis mais relevantes;
- Clientes com limites de crédito estáveis e dívidas reduzidas tendem ser classificados como "Good";
- O modelo identifica bem clientes de maior risco - “Standard” e “Poor”;

Modeling- Árvore de Decisão

Pré-RFE

Conjunto de treino

- Conjunto de dados: *Down-sample*
- Métricas:
  Accuracy - 69,33%
  Sensitivity - 69,32%
  Specificity - 84,66%
  AUC- 78,52%

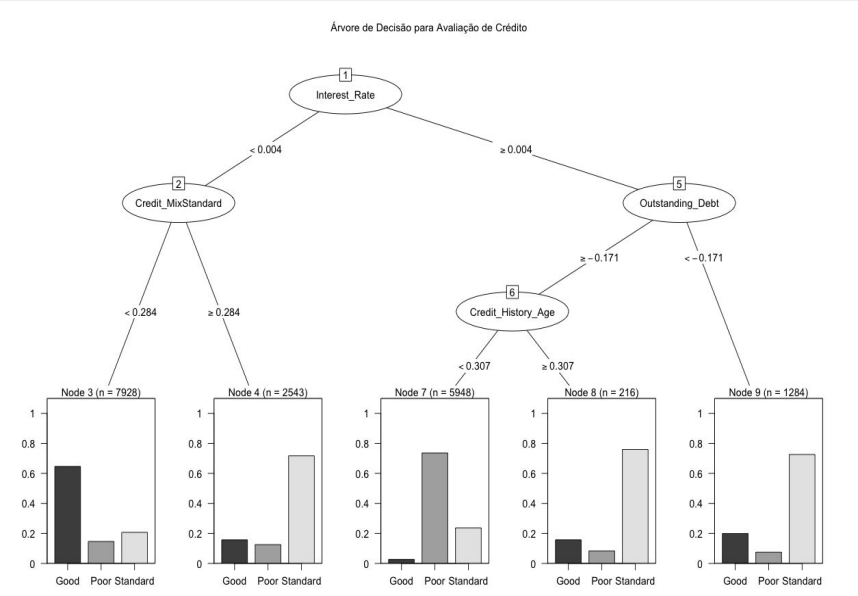
Conjunto de teste

- Dados normalizados
- Métricas:
  Accuracy - 66,45%
  Sensitivity - 66,45%
  Specificity - 83,23%
  AUC- 77,63%

Evaluation - Árvores de Decisão

Conjunto de Treino

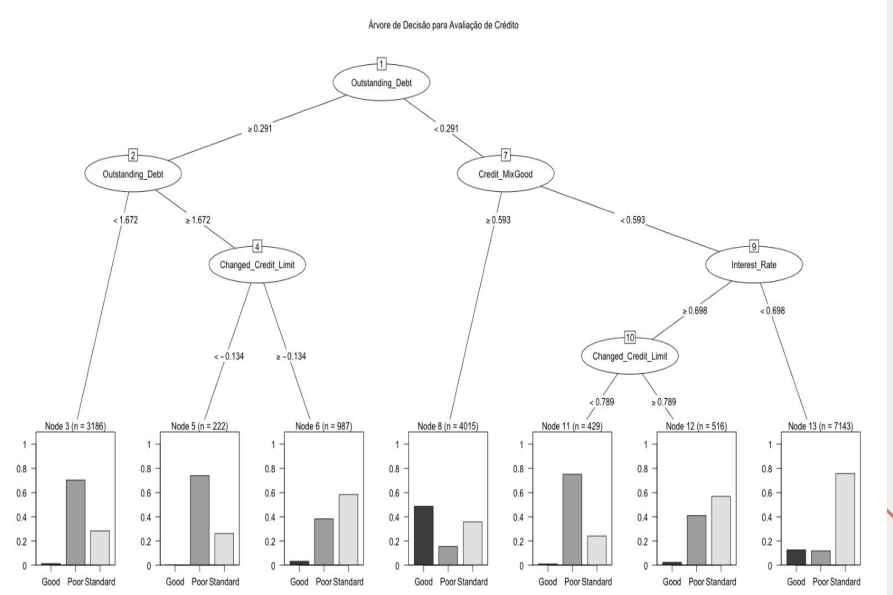
- **Taxa de juros** é o fator mais relevante;
- Taxas mais baixas levam a classificação “Standard” e “Good”;
- **Dívidas Pendentes** elevadas e **Históricos de empréstimo** curtos levam a classificação “Poor” ou “Standard”;



Modeling- Árvores de Decisão

Conjunto de Teste

- **Dívida pendente** é o fator mais relevante;
- **Dívidas pendentes** mais baixas e **Limites de Crédito** estáveis tendem a ser classificados como “Standard”;
- **Dívidas** elevadas e **taxas de juro** altas aumentam probabilidade de classificação “Poor”;
- Valores baixos de **Mix do Crédito** levam a classificação “Good”;



Modeling - Florestas Aleatórias

Pós-RFE com GridSearch

- Validação cruzada com 10 divisões
- Objetivo: reduzir o risco de overfitting e avaliar a robustez

Conjunto de treino

- Conjunto de dados: *Down-sample*
- Métricas:
  Accuracy - 74,80%
  Sensitivity - 74,80%
  Specificity - 84,70%
  AUC- 89,52%

Conjunto de teste

- Dados normalizados
- Métricas:
  Accuracy - 67,23%
  Sensitivity - 67,22%
  Specificity - 83,61%
  AUC- 85,75%

Evaluation - Florestas Aleatórias

- O modelo apresenta *overfitting*;
- Perfis financeiros estáveis com baixo **endividamento** e **histórico** consistente são classificados como "Good";
- Características financeiras instáveis levam a classificações como "Standard" ou "Poor";

Modeling - Redes Neuronais

Pré-RFE

Conjunto de treino

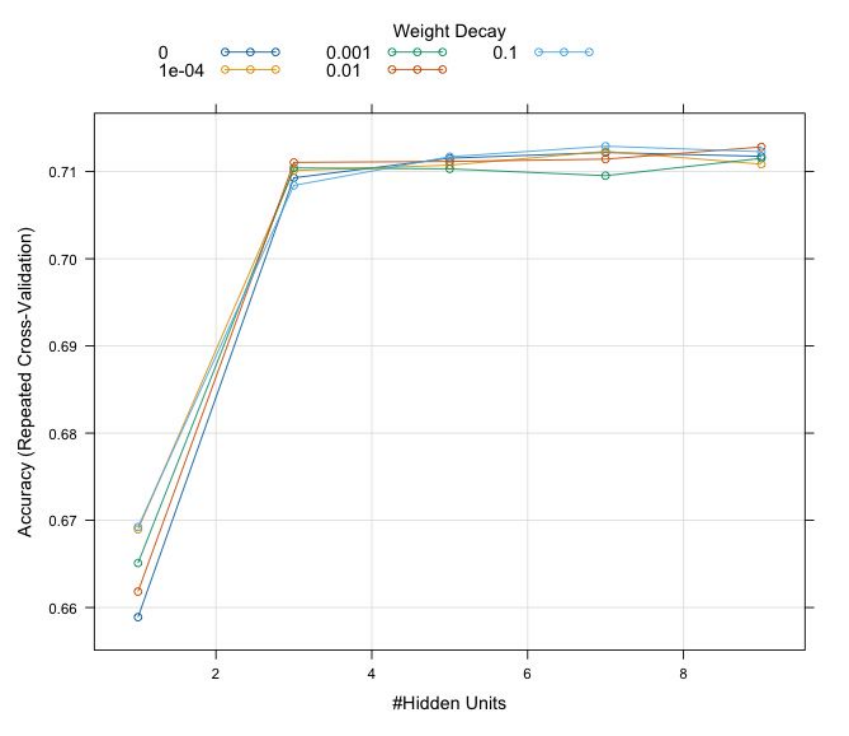
- Conjunto de dados: *Down-sample*
- Métricas:
  Accuracy - 72,39%
  Sensitivity - 72,39%
  Specificity - 86,20%
  AUC- 86,14%

Conjunto de teste

- Dados normalizados
- Métricas:
  Accuracy - 66,34%
  Sensitivity - 66,34%
  Specificity - 83,17%
  AUC- 84,39%

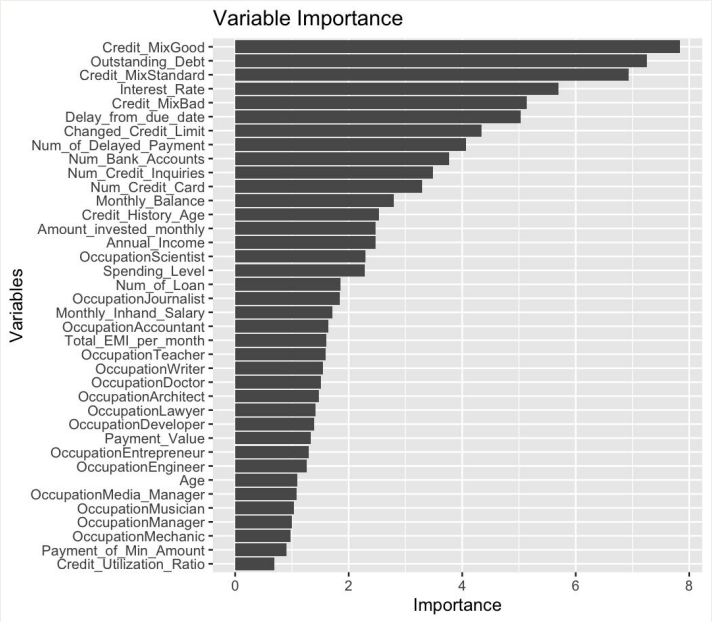
Modeling- Redes Neuronais

- **Unidades ocultas:**
  Para menos de 4, desempenho menor
  Para maior de 4, não existem melhorias significativas
- **Impacto do weight decay:** Baixo
- **Melhor configuração:**
  Entre 4 a 6 unidades ocultas com valores baixos ou moderados de *weight decay*, pois maximizam a *accuracy*



Evaluation - Redes Neuronais

- **Mix de crédito** e **dívidas pendentes** são os fatores de maior importância no modelo;
- **Taxas de juros** têm significância na classificação;

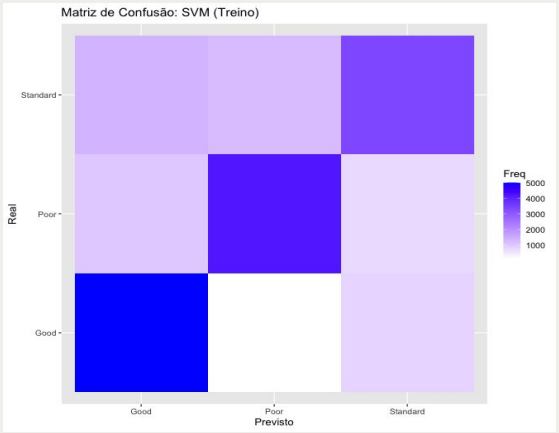


Modeling- SVM

Pós-RFE com GridSearch

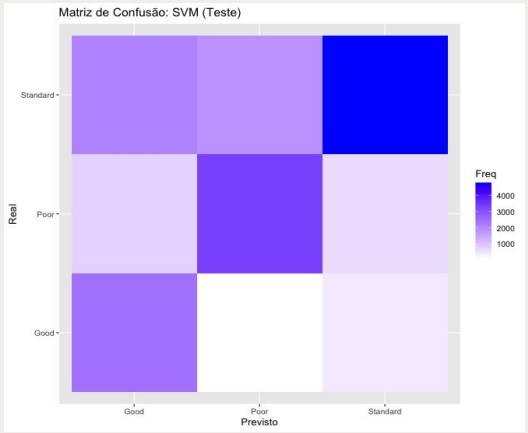
Conjunto de treino

- Conjunto de dados: *Down-sample*
- Métricas:
  Accuracy - 70,34%
  Sensitivity - 70,34%
  Specificity - 85,17%
  AUC- 81,66%



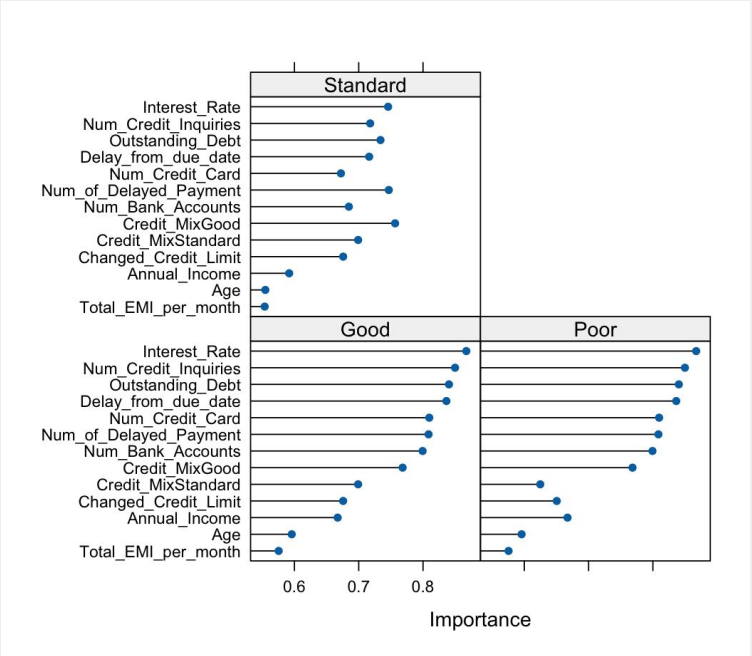
Conjunto de teste

- Dados normalizados
- Métricas:
  Accuracy - 64,04%
  Sensitivity - 64,04%
  Specificity - 82,02%
  AUC- 80,71%



Evaluation - SVM

- A **taxa de juro** é a variável que se destaca como importante em todas as classes de avaliação de crédito;
- **Número de créditos solicitados** e **dívida pendente** têm um forte impacto para classificação “Good” e “Poor”;
- Classificação “Standard” é influenciada pelo **mix de créditos** e **número de pagamentos em atraso**;



Deployment

- **Melhores versões:**
  Accuracy entre 69% e 74% no conjunto de treino;
  Accuracy entre 64% e 67% no conjunto de teste;

- **Modelo com melhores resultados:**
  Florestas Aleatórias

- **Implementação:**
  Integração no sistema interno para automatizar análise de risco;
  Suporte a decisões rápidas e precisas;

- **Monitorização:**
  Plano contínuo com revisões periódicas;
  Adaptação a mudanças nos dados ou no comportamento dos clientes;

Sugestões/Melhorias

- **Data Preparation**:
  Aplicar transformações para corrigir enviesamentos e melhorar a qualidade dos dados usados para modelação;

- **Data Preprocessing**:
  Testar técnicas de *oversampling*, como *SMOTE*, para aumentar/equilibrar o número de classes da variável alvo;
  Explorar escalas alternativas, como Min-Max, para normalizar os dados;

- **Modeling:**
  Avaliar o desempenho de algoritmos como *XGBoost* e *Naive Bayes Classifier*, para potencial melhoria da precisão e robustez do modelo;
  Aplicar métodos avançados de procura de hiperparâmetros, como a Otimização Bayesiana.

Titanic

Tarefa

Conseguimos criar um modelo capaz de prever a sobrevivência de um passageiro na tragédia do Titanic?

Data Understanding

- Identificação das variáveis;
- Identificação dos valores omissos;
- Seleção das variáveis a serem incluídas no modelo de previsão;

Valores Omissos

- Age → 177/891;
- Cabin→ 687/891;
- Embarked → 2/891;

Titanic

Limpeza dos dados

- Eliminação das variáveis de identificação e da variável **Cabine**;
- Substituição de valores omissos pela moda
- Implementação da codificação *one-hot-encoding dummy* (**Embarque**)

Modelação

Modelos formulados:

- Regressão Logística;
- Árvore de Decisão;
- Floresta Aleatória;
- SVM;
- Redes Neuronais;

Titanic

Métricas de avaliação

Metric	Value
Accuracy	0.83959
Sensitivity	0.939227
Specificity	0.678571
AUC	0.884299

Importância das variáveis

Variable	Overall
Sex	60.69027
Fare	43.8036
Pclass	20.11386
Age_Group	16.95763
SibSp	12.24627
Parch	10.44823
Embarked_S	4.939207
Embarked_Q	1.620579

Perguntas?

Anexos

Heatmap de correlações - *Pearson*