Classificação de Clientes Inadimplentes Utilizando Algoritmos de Aprendizagem de Máquina Supervisionados

Trabalho final da Disciplina de Aprendizagem de Máquina (INF07004)
Prof. Dr. Luiz Oliveira

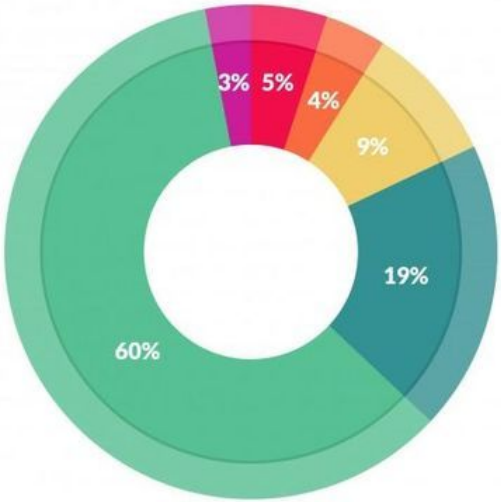
Fernando Nakayama
Agnaldo de Souza Batista

Curitiba, 20 de novembro de 2019

Exploração dos Dados

Exploração e limpeza da base de dados

- Remoção de NA
- 672 para 197 características

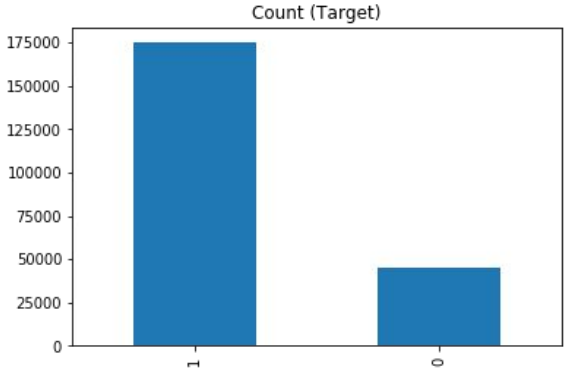


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

CrowdFlower https://visit.figure-eight.com/data-science-report.html

Base Desbalanceada



Seleção de características (Base desbalanceada)

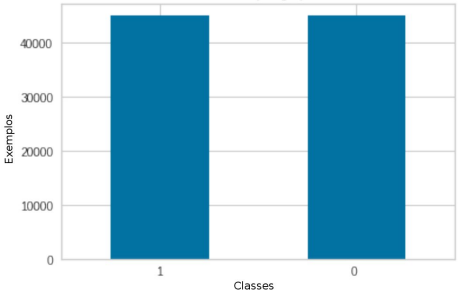
- RFECV (LR e RF) e AG

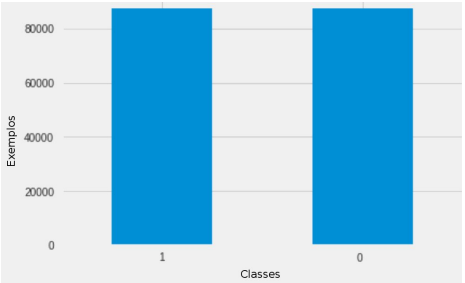
- Características
  - AG: 102
  - RFECV+LR: 24
  - RFECV+RF: 67

Testes (Base Desbalanceada)

1. Decision Tree
2. Random Forest
3. Adaptive Boosting
4. Extreme Gradient Boosting
5. Voting Classifier

Base Balanceada





Balanceamento da Base de Dados

- Oversampling (SMOTE)
- Undersampling (Random)

Seleção de características (Base Balanceada)

- RFECV (LR e RF)
- AG

- Características
  - AG: 91
  - RFECV+LR: 86
  - RFECV+RF: 31
  - RFECV+RF+F1: 37

Testes (Base Balanceada)

- Grid search
- Voting classifier (RF, ADA e XGBC)

Metodologia

1	2	3	4	5	6
Exploração e limpeza da base de dados	Seleção de características (Base desbalanceada)	Testes na Base Desbalanceada	Balanceamento da Base de Dados	Seleção de características (Base Balanceada)	Testes na Base Balanceada
- Remoção de NA - 672 para 197 características	- RFECV (LR e RF) - AG - Características - AG: 102 - RFECV+LR: 24 - RFECV+RF: 67	-DT -RF -ADA -XGB -Voting	- Random Undersampling -SMOTE	- RFECV (LR e RF) - AG - Características - AG: 91 - RFECV+LR: 86 - RFECV+RF: 31 - RFECV+RF+F1: 37	- Grid search - Voting classifier (RF, ADA e XGBC)

"Data Science is 99% preparation, 1% misinterpretation " - Big Data Borat

Resultados

Base
- Desbalanceada

Seleção de Características
- RFECV com Random Forest

Voting Classifier
- Random Forest
- Decision Tree
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Resultados

Base
- Desbalanceada

Seleção de Características
- RFECV com Random Forest

Voting Classifier
- Random Forest
- Decision Tree
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.10	0.17	13442
1	0.81	0.98	0.89	52553
accuracy			0.80	65995
macro avg	0.71	0.54	0.53	65995
weighted avg	0.77	0.80	0.74	65995

Resultados

Base
- Desbalanceada

Seleção de Características
- RFECV com Random Forest

Voting Classifier
- Random Forest
- Decision Tree
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.10	0.17	13442
1	0.81	0.98	0.89	52553
accuracy			0.80	65995
macro avg	0.71	0.54	0.53	65995
weighted avg	0.77	0.80	0.74	65995

Resultados

Base
- Desbalanceada

Seleção de Características
- RFECV com Random Forest

Voting Classifier
- Random Forest
- Decision Tree
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.10	0.17	13442
1	0.81	0.98	0.89	52553
accuracy			0.80	65995
macro avg	0.71	0.54	0.53	65995
weighted avg	0.77	0.80	0.74	65995

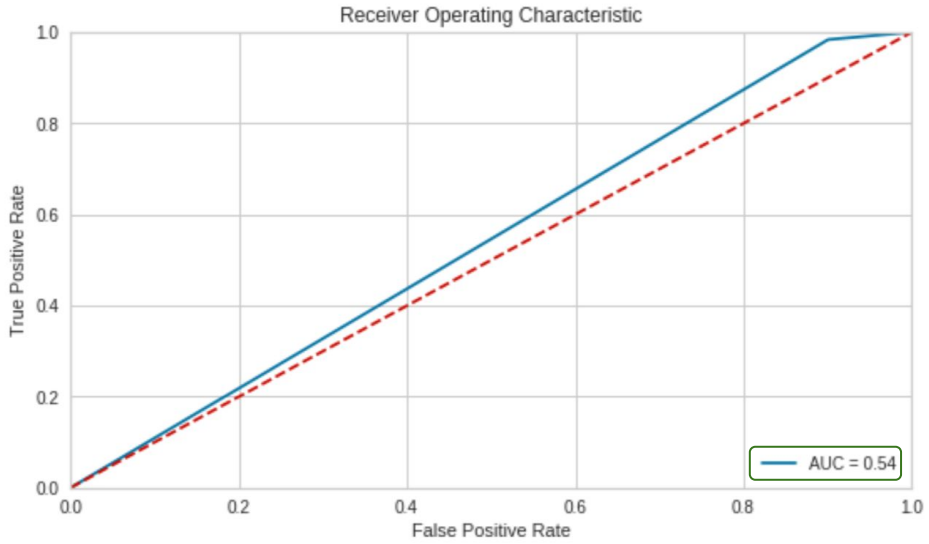
Resultados

Base
- Desbalanceada

Seleção de Características
- RFECV com Random Forest

Voting Classifier
- Random Forest
- Decision Tree
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC



Resultados

Base
- Desbalanceada

Seleção de Características
- RFECV com Random Forest

Voting Classifier
- Random Forest
- Decision Tree
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

- Atende ao enunciado do trabalho
- Identifica 81% dos maus pagadores
- Modelo conservador

Resultados

Base
- Desbalanceada

Seleção de Características
- RFECV com Random Forest

Voting Classifier
- Random Forest
- Decision Tree
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

- Atende ao enunciado do trabalho
- Identifica 81% dos maus pagadores com precisão
- **Modelo conservador**

- Prejudica a classe de bons pagadores

Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.56	0.60	13530
1	0.66	0.71	0.66	13508
accuracy			0.63	27038
macro avg	0.64	0.63	0.63	27038
weighted avg	0.64	0.63	0.63	27038

Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.56	0.60	13530
1	0.66	0.71	0.66	13508
accuracy			0.63	27038
macro avg	0.64	0.63	0.63	27038
weighted avg	0.64	0.63	0.63	27038

Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.56	0.60	13530
1	0.66	0.71	0.66	13508
accuracy			0.63	27038
macro avg	0.64	0.63	0.63	27038
weighted avg	0.64	0.63	0.63	27038

Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.56	0.60	13530
1	0.66	0.71	0.66	13508
accuracy			0.63	27038
macro avg	0.64	0.63	0.63	27038
weighted avg	0.64	0.63	0.63	27038

Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.56	0.60	13530
1	0.66	0.71	0.66	13508
accuracy			0.63	27038
macro avg	0.64	0.63	0.63	27038
weighted avg	0.64	0.63	0.63	27038

Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

Classification Report

	precision	recall	f1-score	support
0	0.61	0.56	0.60	13530
1	0.66	0.71	0.66	13508
accuracy			0.63	27038
macro avg	0.64	0.63	0.63	27038
weighted avg	0.64	0.63	0.63	27038

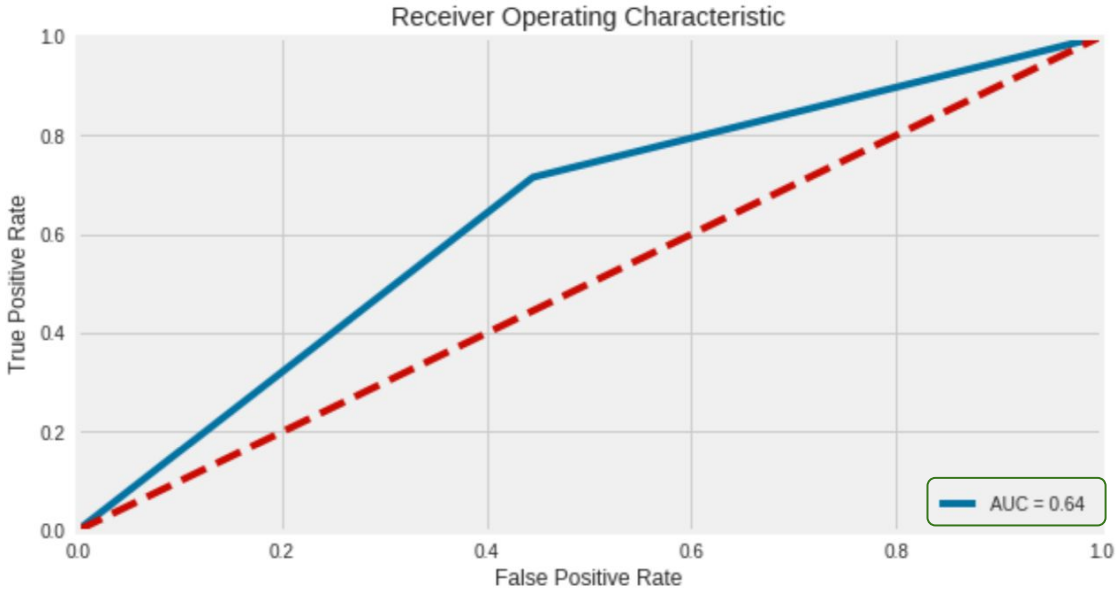
Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC



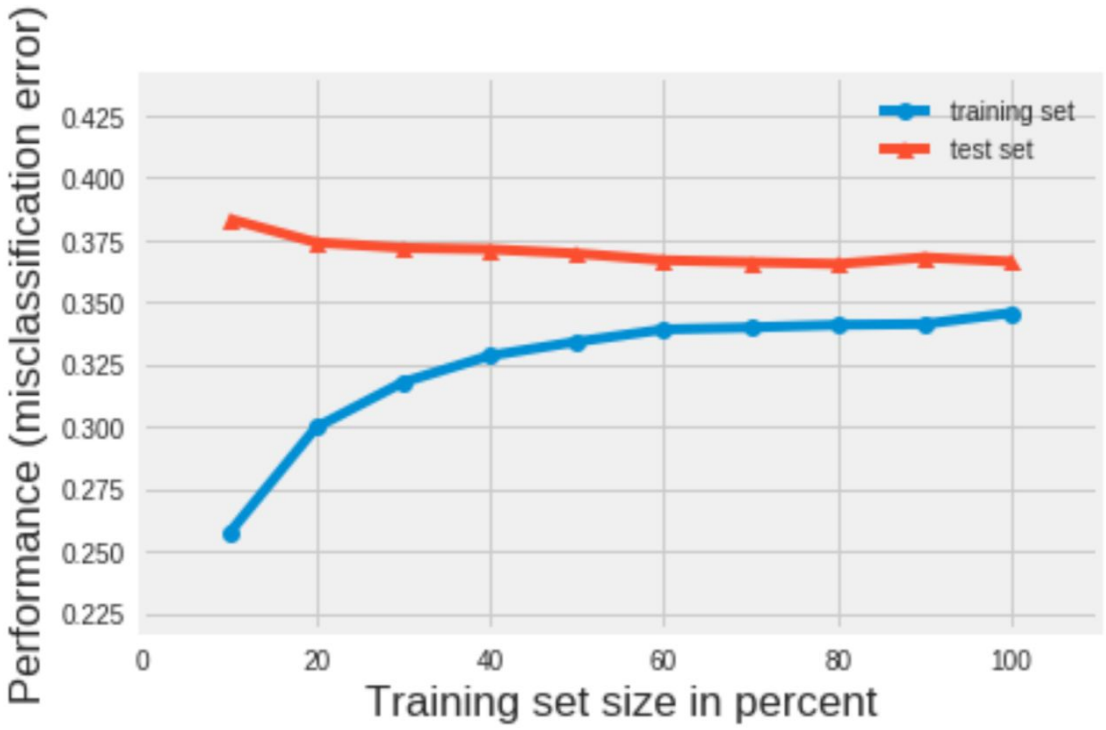
Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC



Resultados

Base
- Balanceada

Seleção de Características
- RFECV com Random Forest e F1-Score

Voting Classifier
- Random Forest
- AdaBoost
- Extreme Gradient Boosting

Métricas
- Acurácia
- Precisão
- Recall
- F1-Score
- Curva ROC e AUC

- **Modelo de Risco**
- Precisão razoável para ambas as classes
- Boa sensibilidade para identificar maus pagadores, sem descartar os bons completamente
- Bom valor de F1-*Score* médio e para cada classe individualmente

Conclusão

- Classificação de clientes inadimplentes usando algoritmos de aprendizado de máquina supervisionados
- Modelo conservador classifica os maus pagadores com boa precisão
- Modelo de risco atende ambas as classes de maneira mais justa