Fraud Detection

Case

•

Marcelo Bianchi Barata Ribeiro

Presentation summary

- Introduction
- EDA & Data Cleansing
- Feature Engineering
- Modeling
- Strategies

Good practices

- Agile
  - MVP (Minimum Viable Product)
- 6 sigma
  - Pareto Analysis (80/20)
- SoC (Separation of Concerns)

EDA and Cleansing

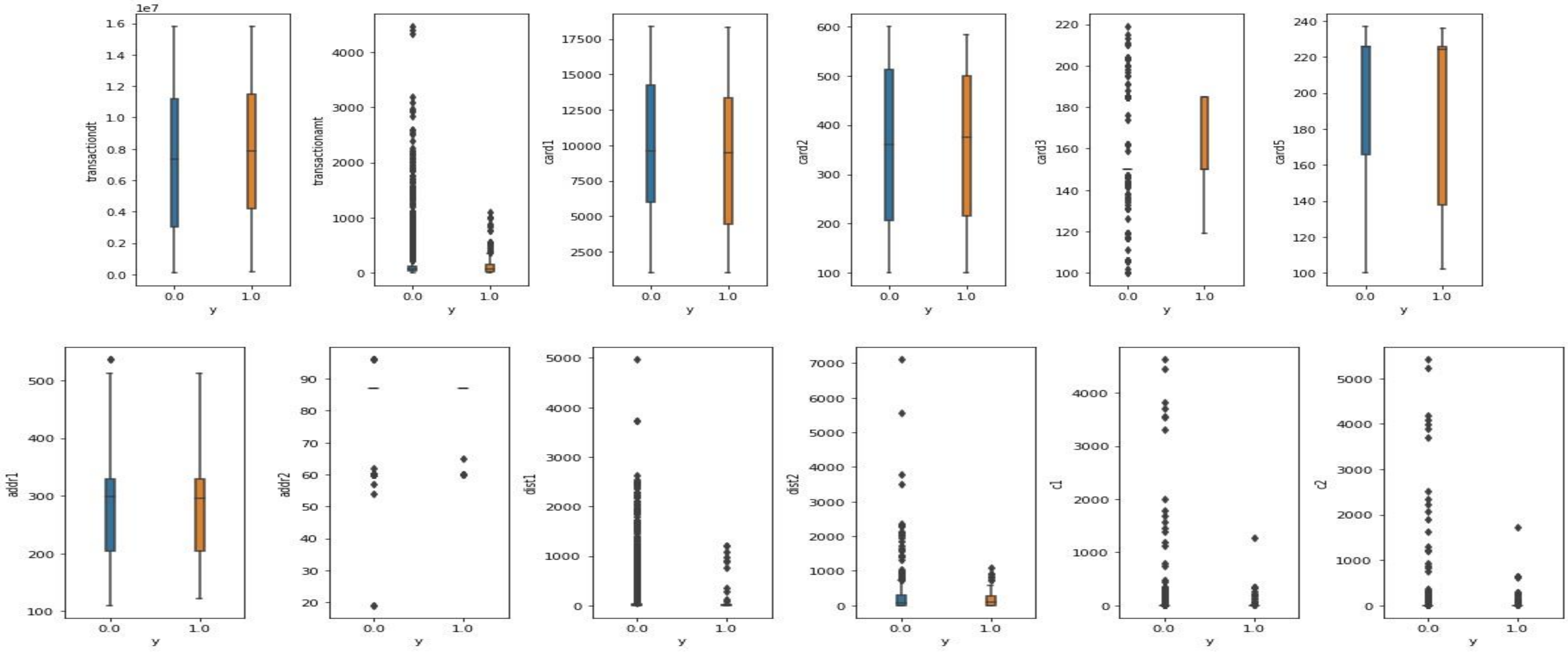
Data Cleansing

- Missing values:
  - We can't assume beforehand if those are Missing at Random (MAR) or Missing not at Random (MNAR). I am assuming they are Missing at Random until further investigation.
  - Threshold for column removal: 30%
  - Threshold for row removal: 50%
  - Median applied for the remaining missing data
    - I saved a functions for KNN (K nearest neighbor), but didn’t use it because of processing speed.

Outliers

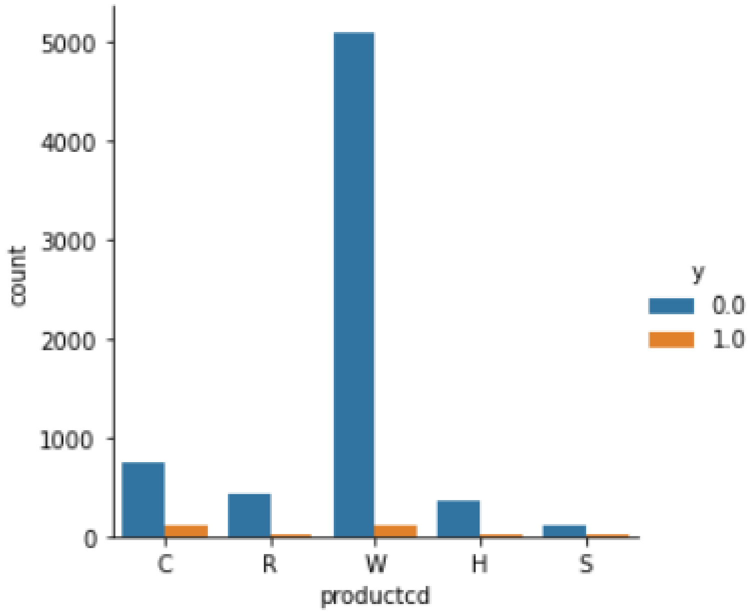
- Isolation Forest
  - Contamination threshold: 1%
  - 1603 annotated outliers
  - There wasn’t a rational to justify outlier removal
  - Alternative: usa a band of values instead of removing.

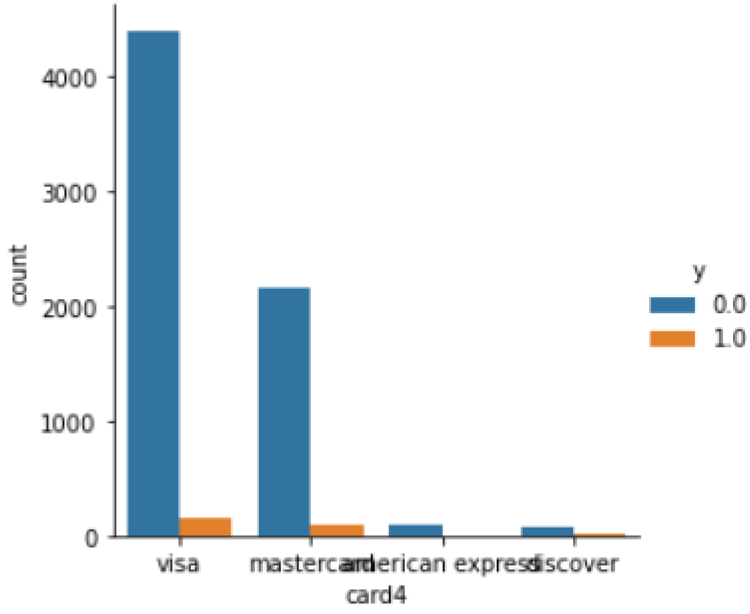
Boxplots

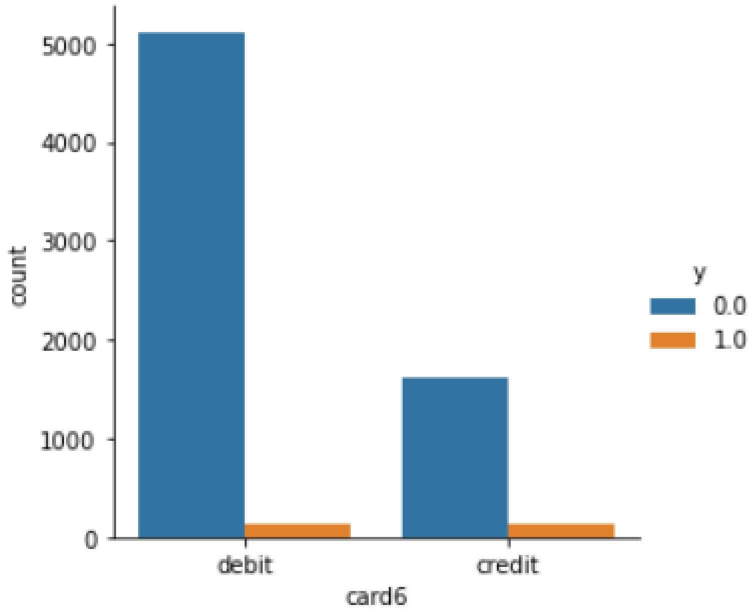


1e7
transactiondt
transactionamt
card1
card2
card3
card5
addr1
addr2
dist1
dist2
c1
c2
y
0.0
1.0

Distributions







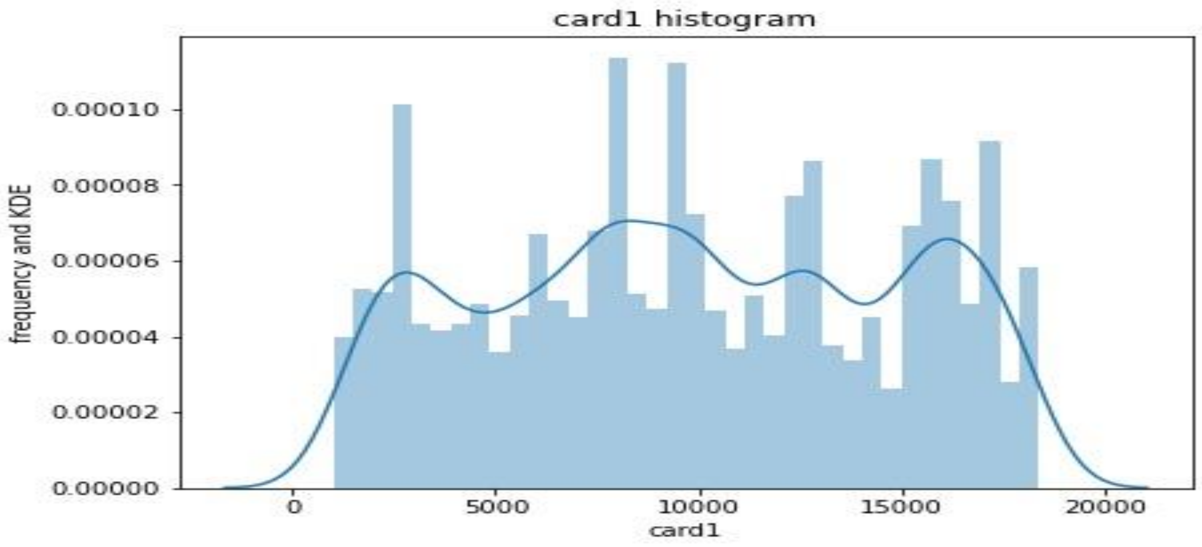
2 = solteiro, 4 = comunhão parcial

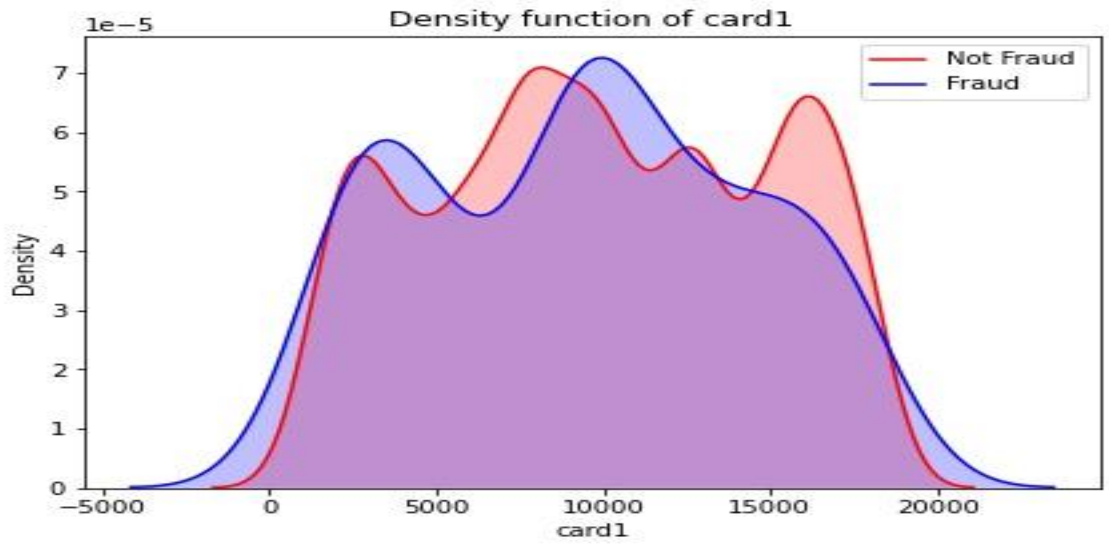
C	0.131920
S	0.093750
R	0.038031
H	0.029491
W	0.021142

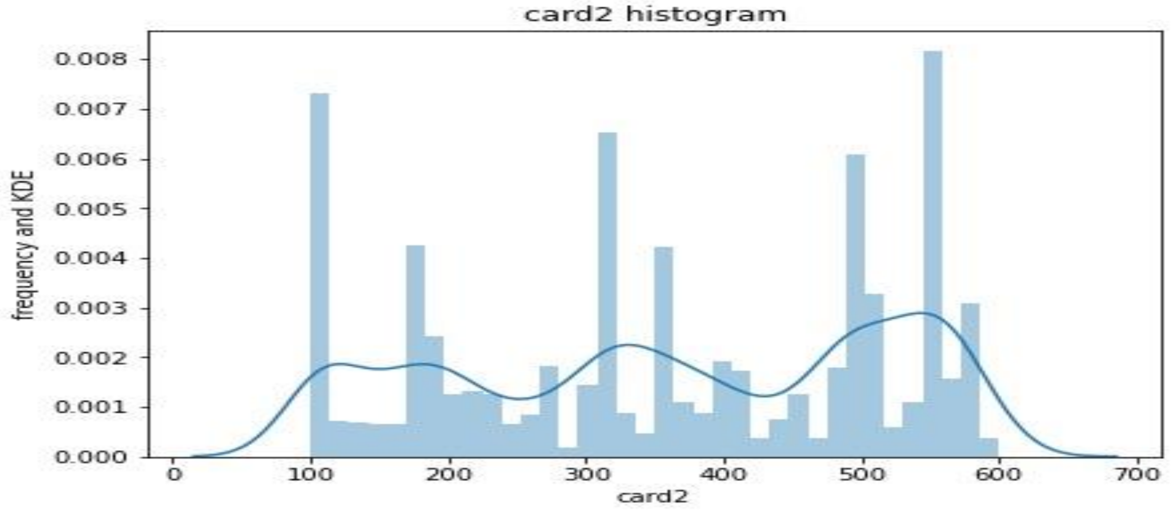
discover	0.105882
mastercard	0.042241
visa	0.034385
american express	0.010000

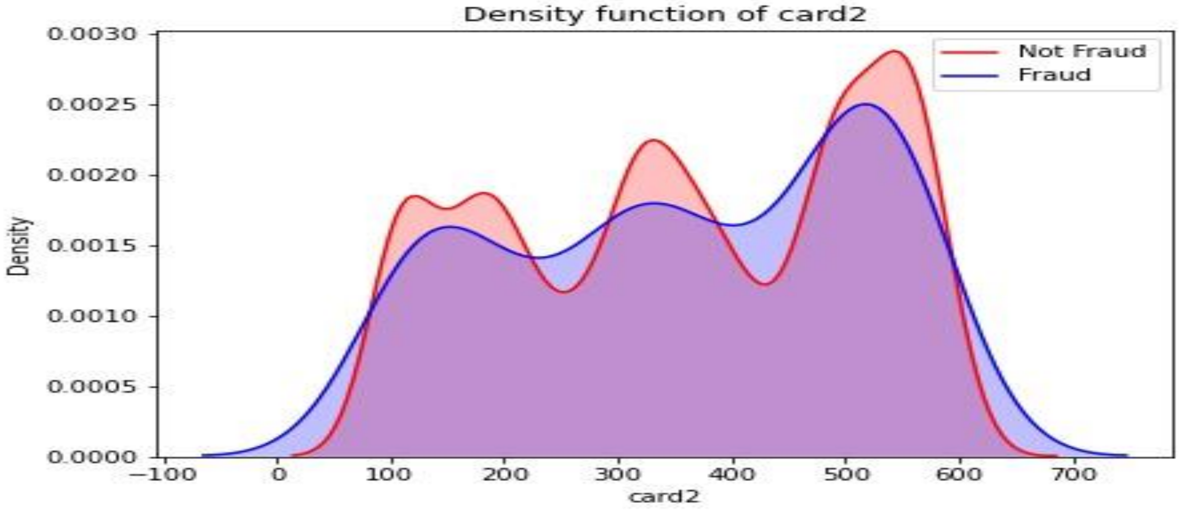
credit	0.069500
debit	0.026811

Distributions

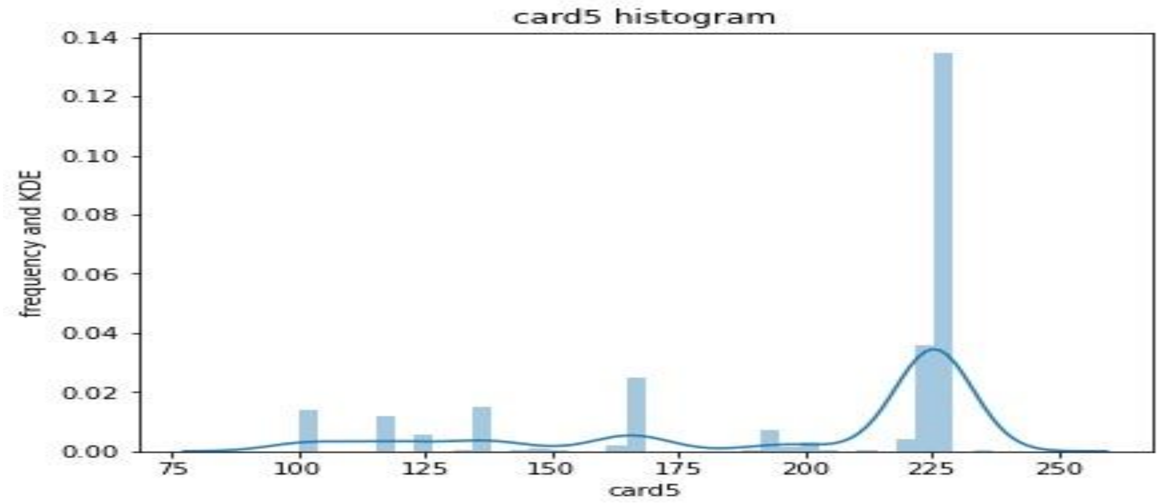


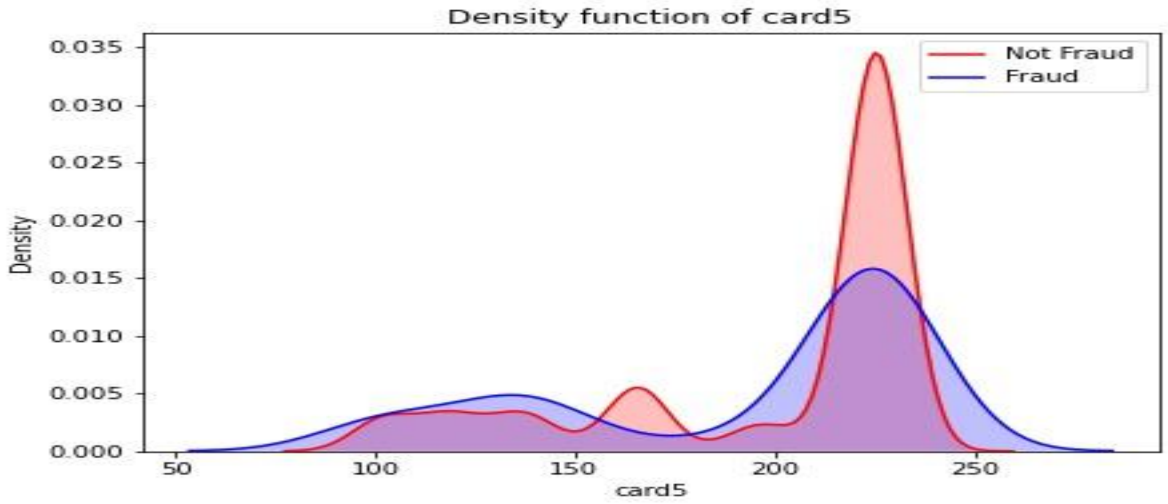


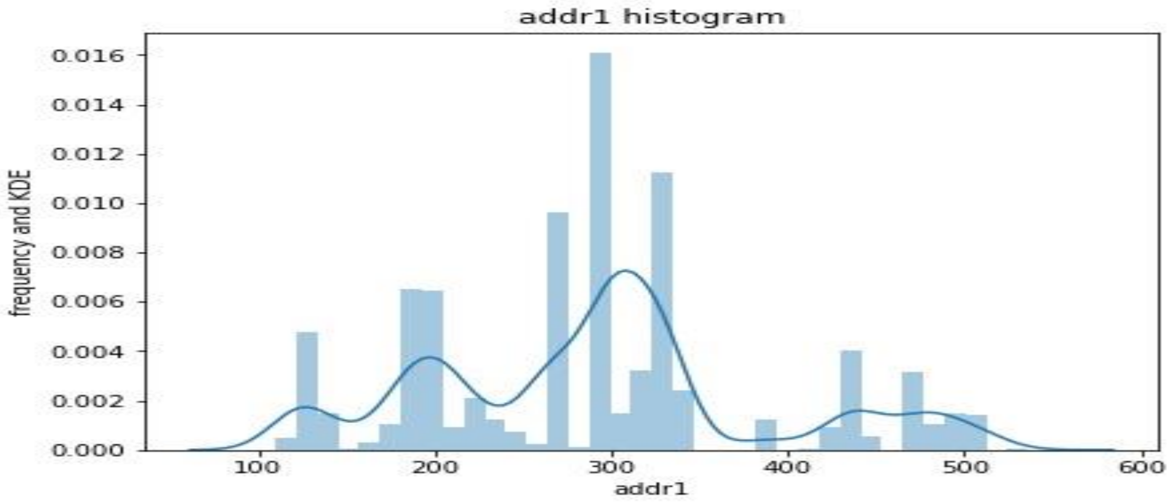


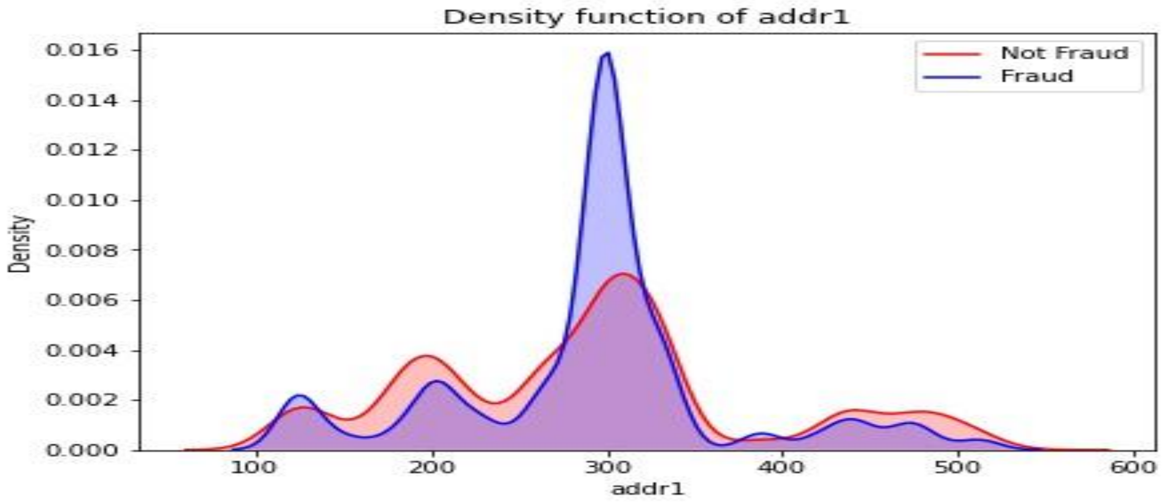


Distributions

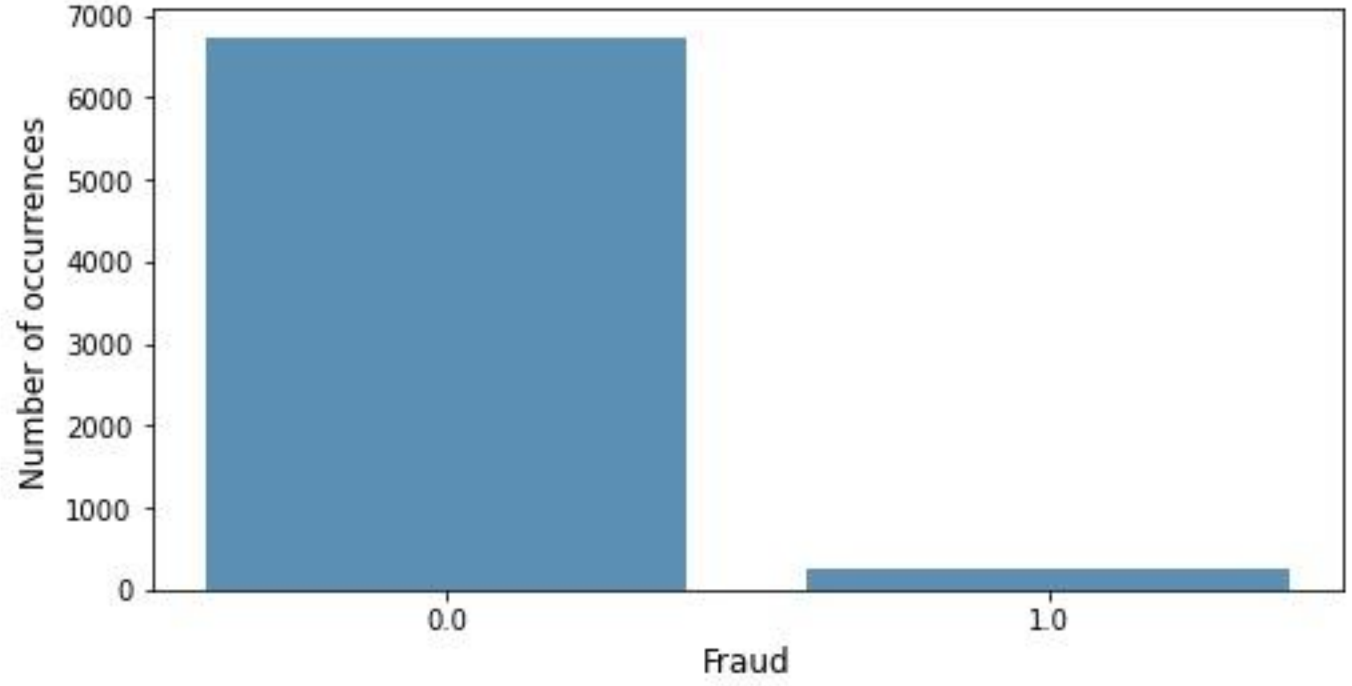








Imbalance



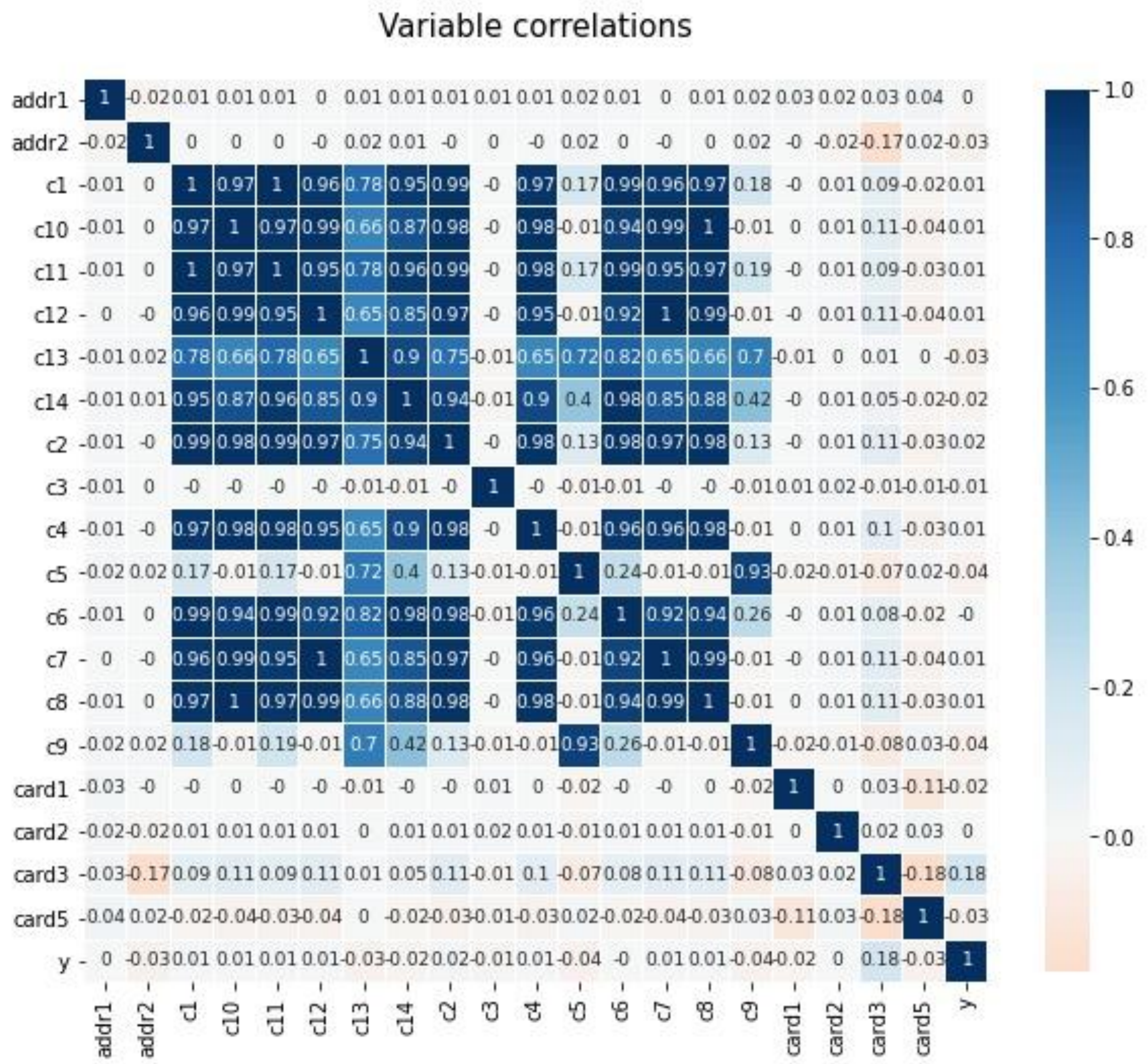
Ratio between classes: 26.6
Proportions:
0 0.96
1 0.04

Feature Engineering

Feature engineering (continuation)

- Encoding
  - Ordinal Encoding: tree models
  - One-Hot Encoding: linear regression
- New features
  - …
- Feature selection
  - Visualizations and intuition
  - Multicollinearity: Variance Inflation Factor (VIF)
    - Attention to non-linear relationships!
  - Feature Importances: tree models

Correlação com variáveis numéricas



Multicollinearity

- Variance Inflation Factor (VIF)
  - Threshold: VIF = 5
- Important step for linear regression models
- Potentially removable variables:
  - ...

Feature Selection

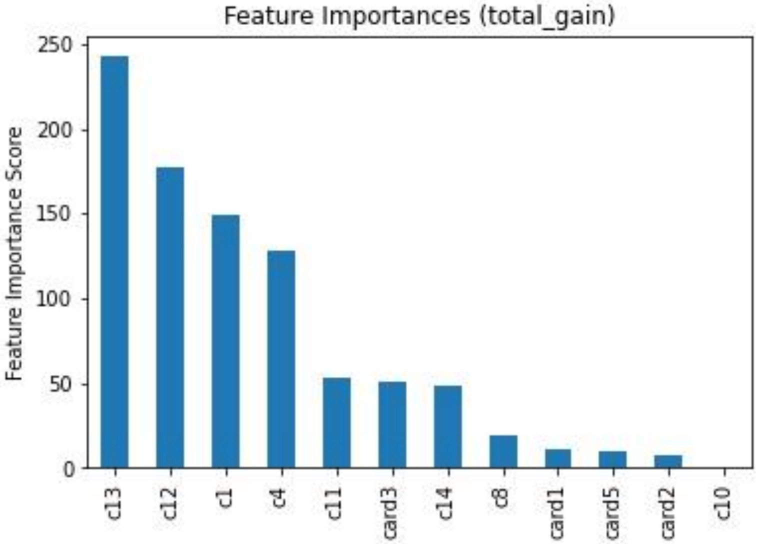
All variables

Only ‘c’ and

card variables

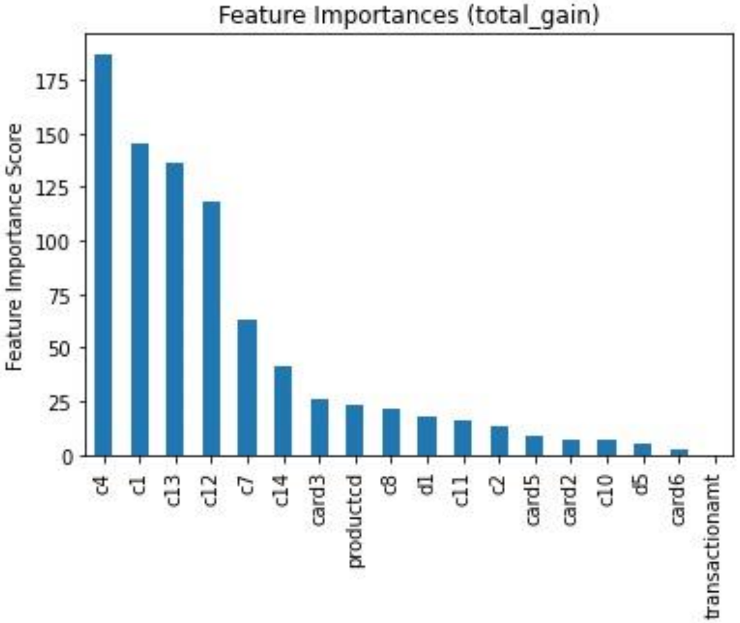
...

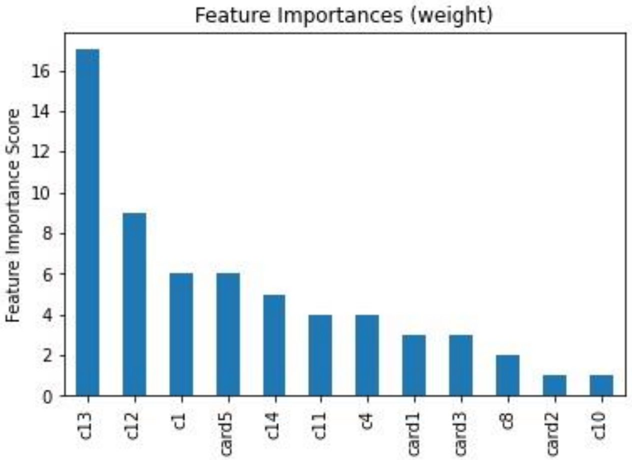
Feature Importances

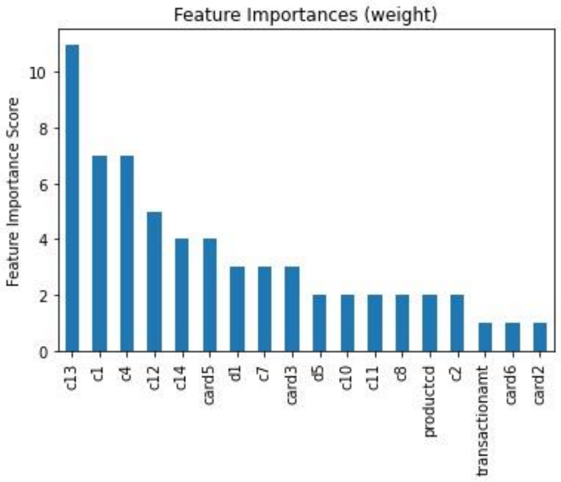


F1: 0.466

F1: 0.49







Modeling

Modeling

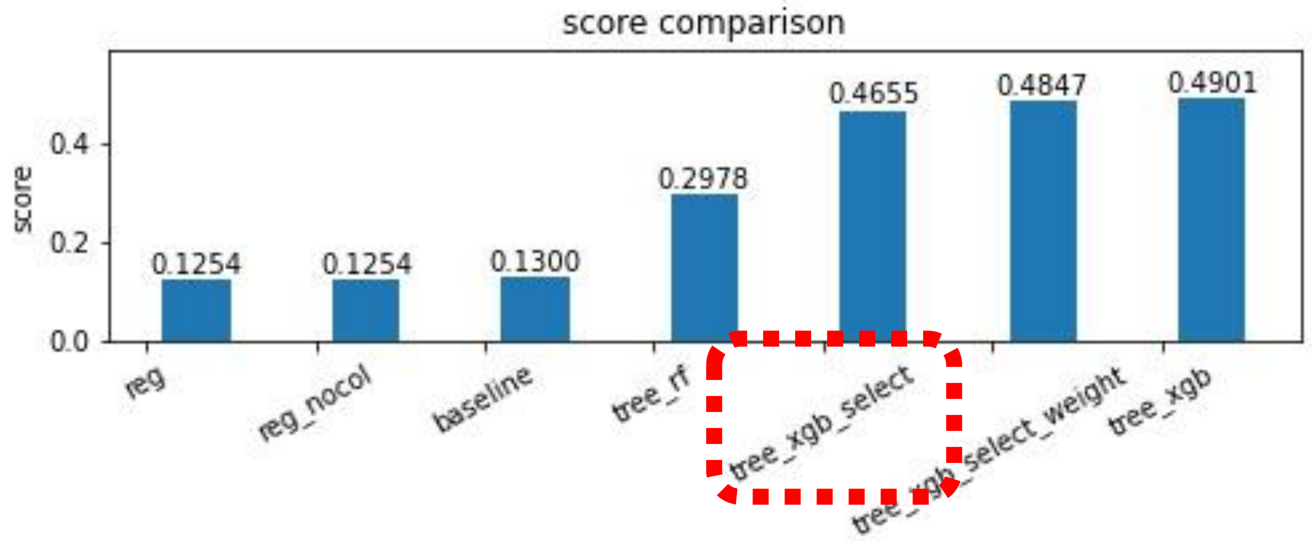
- Occam's razor (parcimony): the simplest solution is usually the best one.
- Algorithm choice depends on optimizations depend on several criteria:
  - Prediction score
  - Explainability
  - Time of development (train and prediction)
  - Computational cost: Big-O notation, memory use
  - Reprodutibility
- The choice depends on the business problem and stakeholder's knowledge.

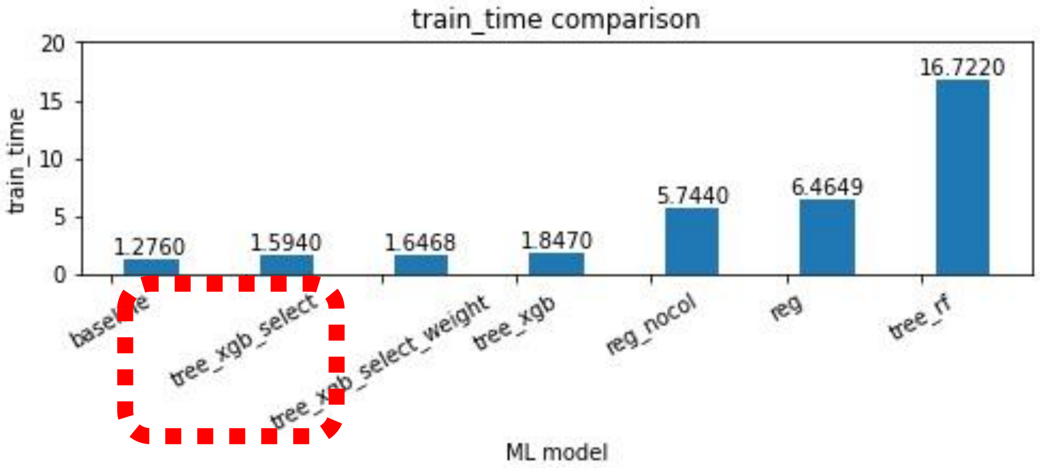
Modeling

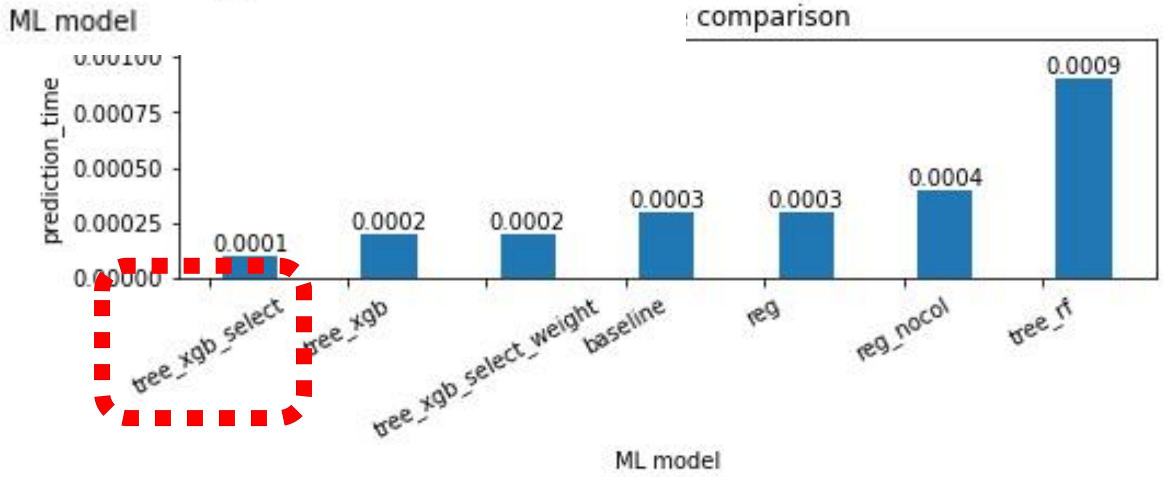
- Score metric: F1 (harmonic mean between precision and recall)
- Validation
  - Train-validation-test split
  - k-fold cross-validation
  - Hyperparameter tuning: gridsearchcv
  - Grids: data/04_models
  - Final metrics: data/05_model_outputs
- Models (fixed seed for reproducibility)
  - Logistic Regression with no tuning (baseline)
  - Logistic Regression
  - Random Forest
  - Xgboost

Model selection (validation phase)

Métrica de score: F1







Final model

Model: Xgboost com 5 variáveis

Validation

- **F1:** 0.466

Features:
C...
Card...

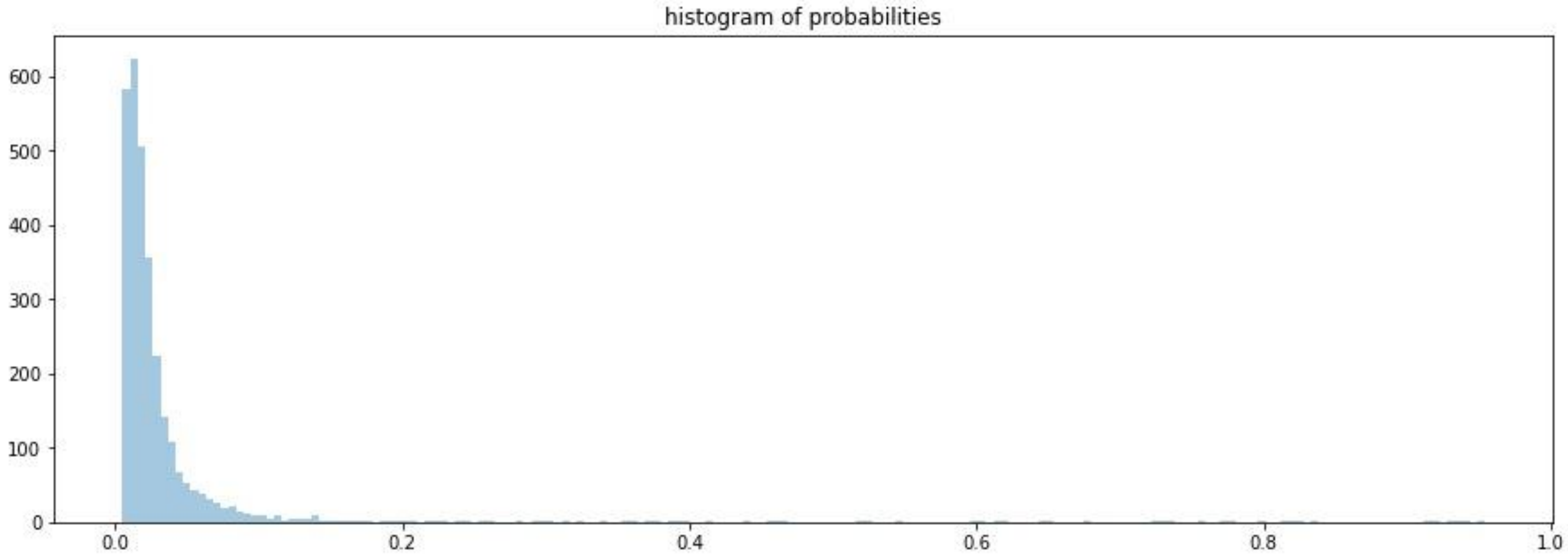
Test

- **F1**: 0.364
- Accuracy: 0.971
- Recall: 0.25
- Precision: 0.71
- Log Loss: 0.368

Sample comparison of estimation and true value

	estimated	true_value
id		
3300259	0.954047	0.0
3390348	0.939576	1.0
3241013	0.935662	1.0
3494143	0.930268	1.0
3378721	0.919951	1.0
...	...	...
3485599	0.005046	0.0
3182041	0.004997	0.0
3233681	0.004978	0.0
3294267	0.004927	0.0
3179811	0.004874	0.0

Histogram of probabilities

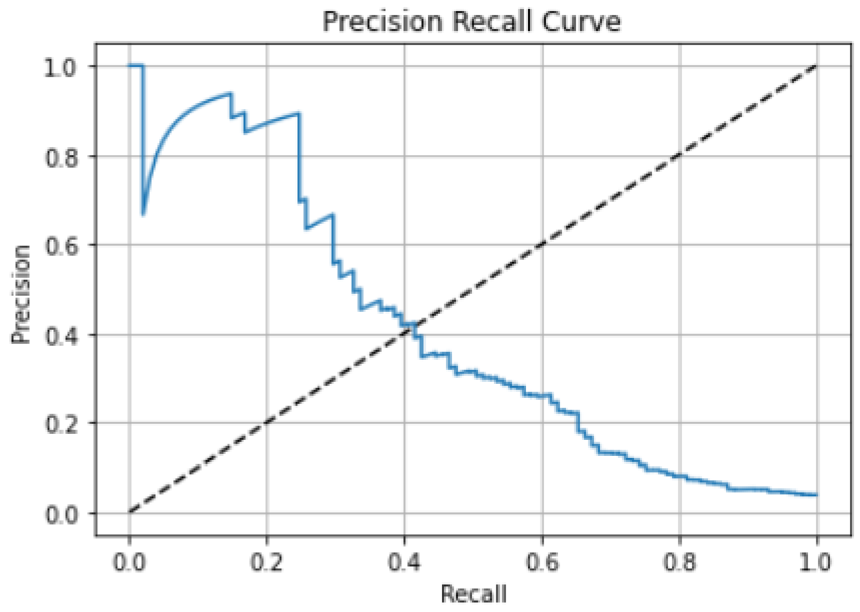


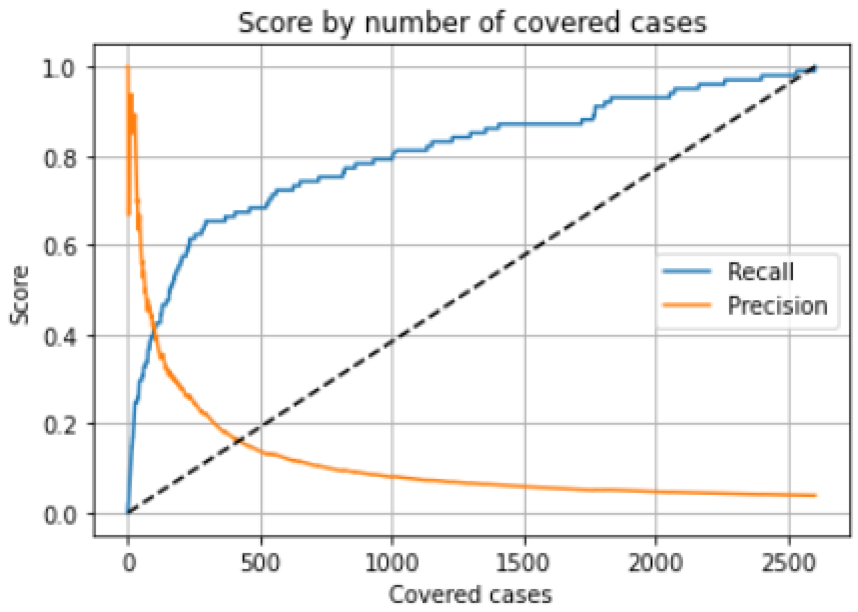
Strategy

Strategy

- In a real business problem, we have key metrics to use so that we can establish final decisions.
- By contrast, on kaggle problems, generally business data are missing. So we need to work on assumptions. So let's consider:
  - \$5,000 as the median cost of fraud: how much the company loses from a fraudulent transaction.
  - \$1,000 as the median remaining lifetime value of users.

Precision vs. Recall





Strategy

4 comparações

theoretical baseline (80% TN, 40% TP)			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	20	0	0
normal transactions correctly classified (TN)	776	1000	776000
normal transactions misclassified (FP)	194	0	0
fraudulent transactions misclassified (FN)	14	-5000	-68000
Total			$\boxed{708000}$

$\boxed{\text{threshold .5}}$ Recall .24, Precision .77			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	8	0	0
normal transactions correctly classified (TN)	964	1000	964000
normal transactions misclassified (FP)	2	0	0
fraudulent transactions misclassified (FN)	26	-5000	-130000
Total			834000

$\boxed{\text{threshold .3}}$ Recall .32, Precision .62			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	11	0	0
normal transactions correctly classified (TN)	960	1000	960000
normal transactions misclassified (FP)	7	0	0
fraudulent transactions misclassified (FN)	23	-5000	-115000
Total			845000

$\boxed{\text{threshold .15,}}$ Recall .37, Precision .44			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	12	0	0
normal transactions correctly classified (TN)	951	1000	951000
normal transactions misclassified (FP)	16	0	0
fraudulent transactions misclassified (FN)	21	-5000	-105000
Total			846000

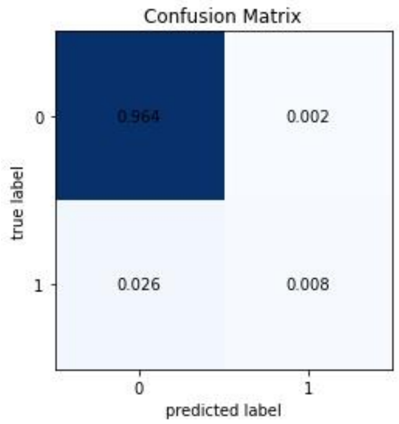
Hipothetical median remaining lifetime value of users = 1000

Hipothetical median cost of fraud = 5000

Strategy

theoretical baseline (80% TN, 40% TP)			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	20	0	0
normal transactions correctly classified (TN)	776	1000	776000
normal transactions misclassified (FP)	194	0	0
fraudulent transactions misclassified (FN)	14	-5000	-68000
Total			708000

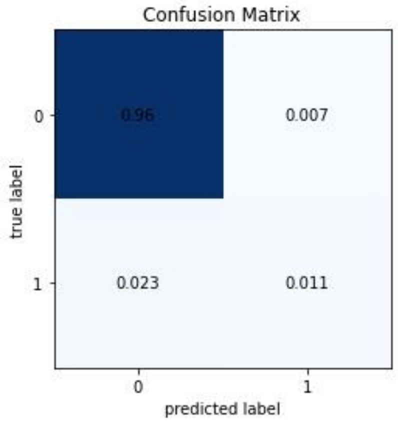
threshold .5 Recall .24, Precision .77			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	8	0	0
normal transactions correctly classified (TN)	964	1000	964000
normal transactions misclassified (FP)	2	0	0
fraudulent transactions misclassified (FN)	26	-5000	-130000
Total			834000



18% improvement over theoretical baseline

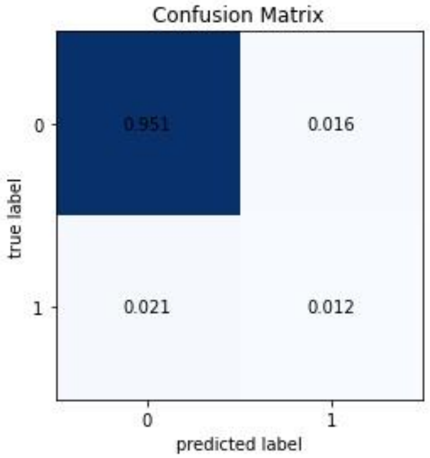
Strategy

threshold .3 Recall .32, Precision .62			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	11	0	0
normal transactions correctly classified (TN)	960	1000	960000
normal transactions misclassified (FP)	7	0	0
fraudulent transactions misclassified (FN)	23	-5000	-115000
Total			845000



19% improvement over theoretical baseline.

threshold .15, Recall .37, Precision .44			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	12	0	0
normal transactions correctly classified (TN)	951	1000	951000
normal transactions misclassified (FP)	16	0	0
fraudulent transactions misclassified (FN)	21	-5000	-105000
Total			846000



Final Threshold

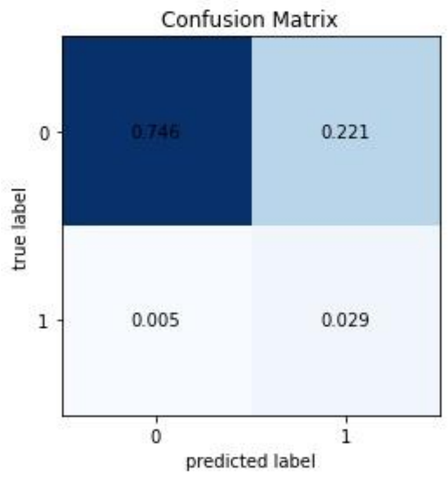
19% improvement over theoretical baseline.

1.5% improvement over original threshold.

Stress test

What if we set an aggressive threshold?

threshold .01, Recall .85, Precision .11			
Description	obs	value	total
fraudulent transactions correctly classified (TP)	33	0	0
normal transactions correctly classified (TN)	188	1000	188000
normal transactions misclassified (FP)	779	0	0
fraudulent transactions misclassified (FN)	1	-5000	-5000
Total			183000



74% loss over theoretical baseline.

Thanks

Marcelo B. Barata Ribeiro

•

marcelobbribeiro@gmail.com

•

linkedin.com/in/marcelobarataribeiro

•

https://github.com/Marcelobbr