Inteligência Computacional

Treinando Classificadores para Previsão de Transações Fraudulentas no Cartão de Crédito

Integrantes

Andrei Magalhães
Gabriel Negreiros
Pedro Elias de Abreu
Vitor Gabriel Pereira Santos

O Problema

- Fraudes bancárias cada vez mais comuns

- No Brasil, cerca de 12,1 milhões de pessoas já foram vítimas de alguma fraude financeira no último ano

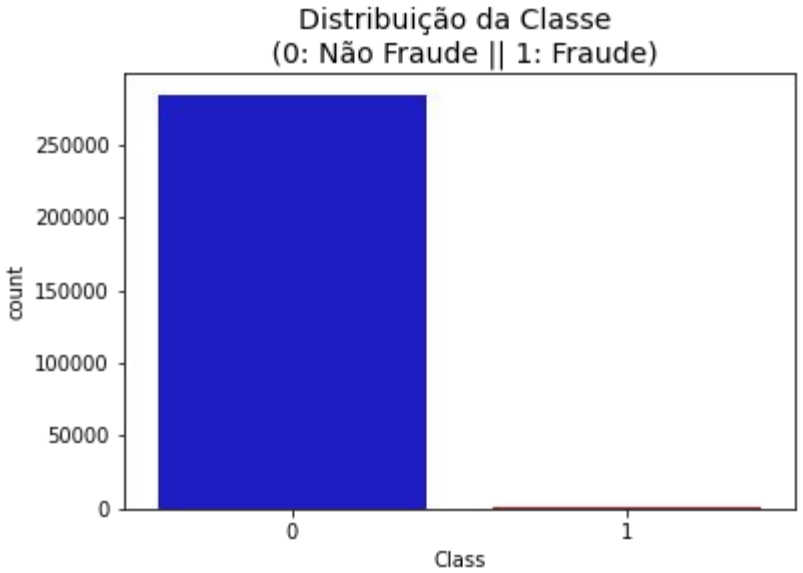
- Com análise das transações pode ser possível prever se uma transação futura é fraudatória ou não

A Base de Dados

O conjunto de dados contém transações feitas por cartões de crédito em dois dias de setembro de 2013 por titulares de cartões europeus. O conjunto de dados é altamente desequilibrado, a classe positiva (fraudes) corresponde a 0,172% de todas as transações.

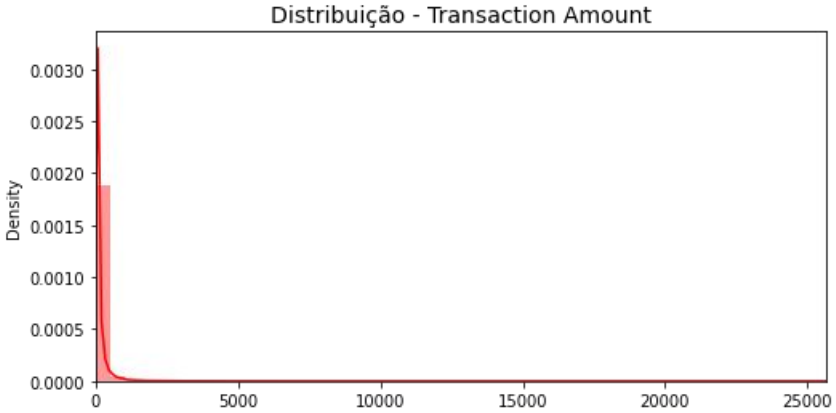
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12		V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	-0.551600	-0.617801	-0.99	-0.470401	0.207971	0.025791	0.403993	0.251412	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.065235	0.48	0.463917	-0.114805	-0.183361	-0.145783	-0.069083	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	0.624501	0.066084	0.71	-2.890083	1.109969	-0.121359	-2.261857	0.524980	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	-0.226487	0.178228	0.50	-1.059647	-0.684093	1.965775	-1.232622	-0.208038	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	-0.822843	0.538196	1.34	-0.451449	-0.237033	-0.038195	0.803487	0.408542	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

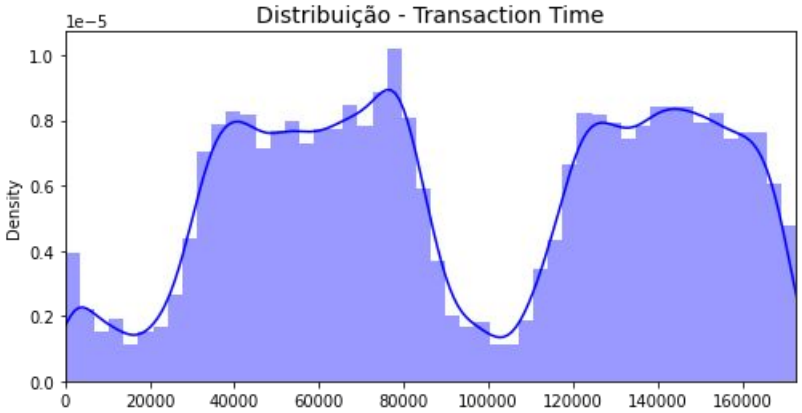
Distribuição das classes



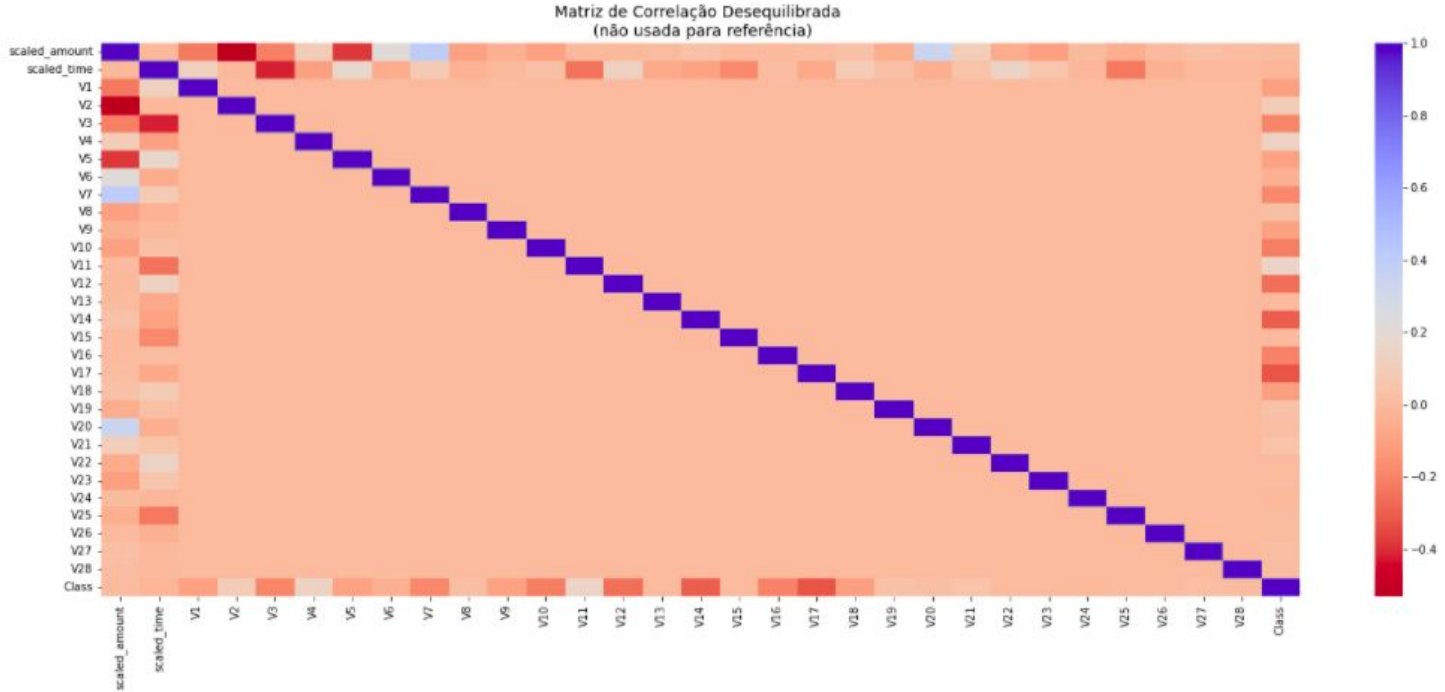
No Frauds: 284.315 Fraud: 492

Distribuição das Features Valor e Tempo





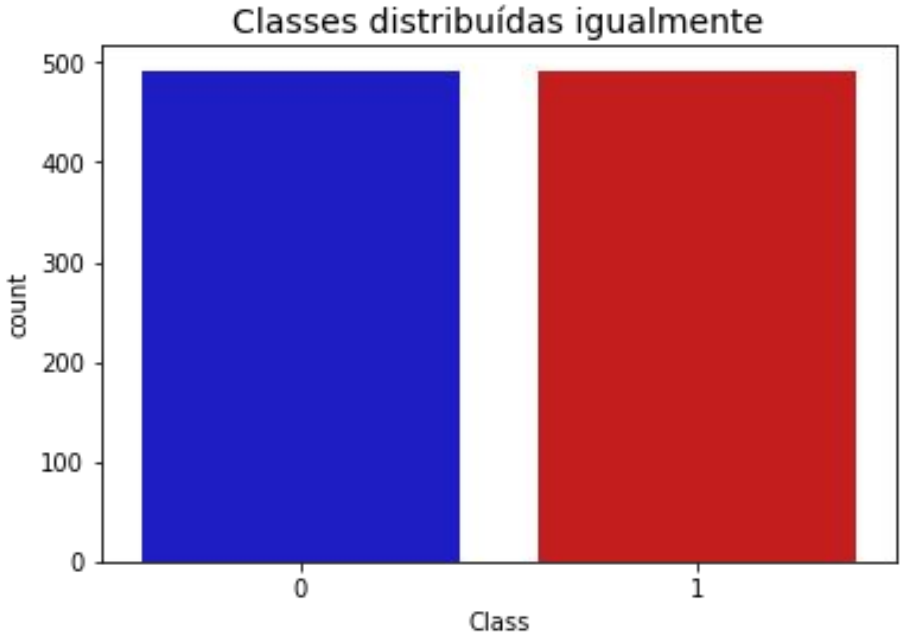
Matriz de correlação desequilibrada



Tratamento do DataSet

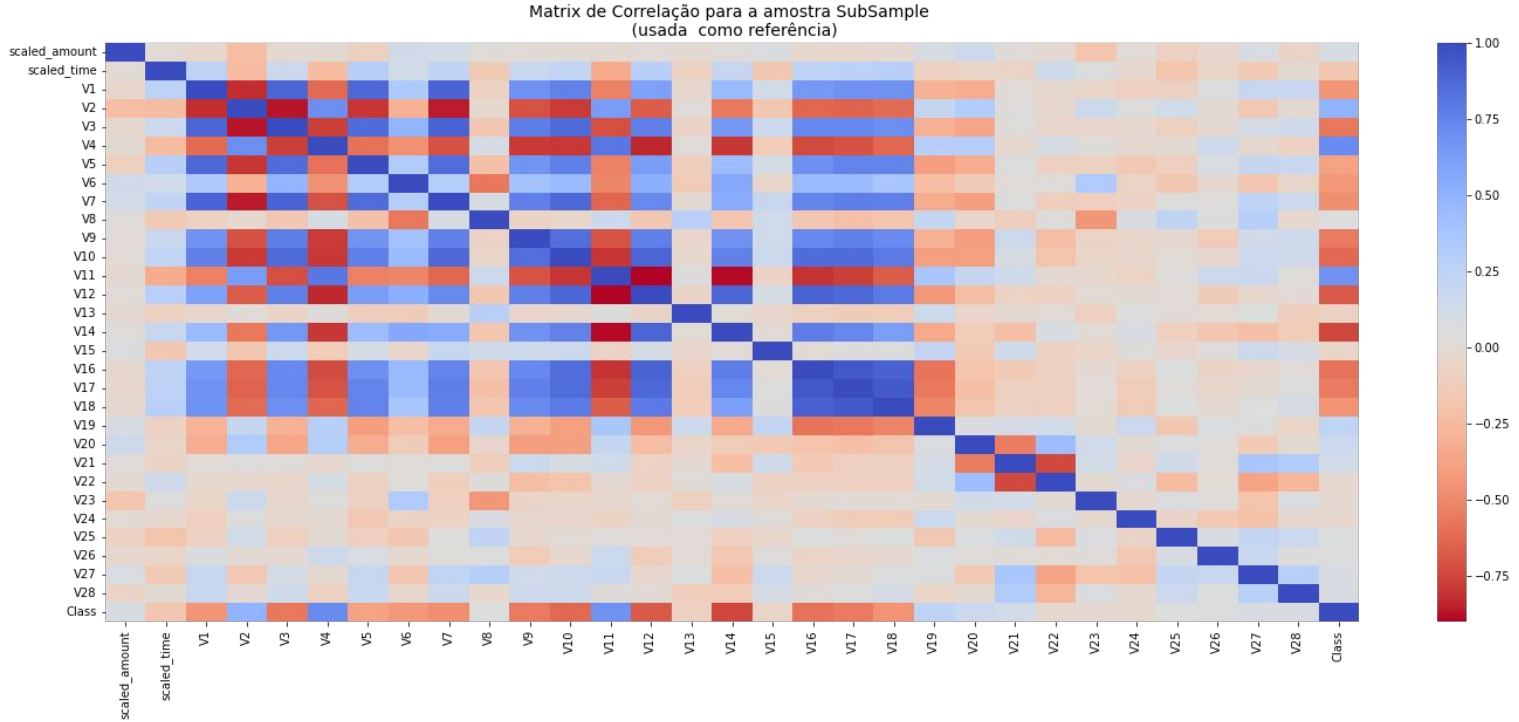
- Feature Scaling
- PCA Transformation
- Remoção de Outliers
- Random Under-Sampling

Distribuição pós Random UnderSampling



No Frauds: 492 Fraud: 492

Matriz de Correlação Equilibrada



Classificadores

- Logistic Regression
- K Neighbors Classifier
- Support Vector Classifier - SVC
- Decision Tree Classifier
- Foi usado o GridSearchCV para testar encontrar os melhores parâmetros para cada classificador.

Métricas de Avaliação

- Accuracy

```
LR - acurácia 0.9052631578947369
KN - acurácia 0.9
SVC - acurácia 0.8947368421052632
Tree - acurácia 0.8789473684210526
```

Métricas de Avaliação

- F1 Score

```
LR - F1 Score 0.90322580645161290
KN - F1 Score 0.89729729729729720
SVC - F1 Score 0.89361702127659580
Tree - F1 Score 0.87700534759358300
```

Métricas de Avaliação

- Confusion Matrix

Logistic Regression

	0	1
0	88	2
1	16	84

K Neighbors Classifier

	0	1
0	88	2
1	17	83

SVC Classifier

	0	1
0	86	4
1	16	84

Decision Tree Classifier

	0	1
0	85	5
1	18	82

Métricas de Avaliação

- Recall

```
LR - Recall de 0.84
KN - Recall de 0.83
SVC - Recall de 0.84
Tree - Recall de 0.9425287356321839
```

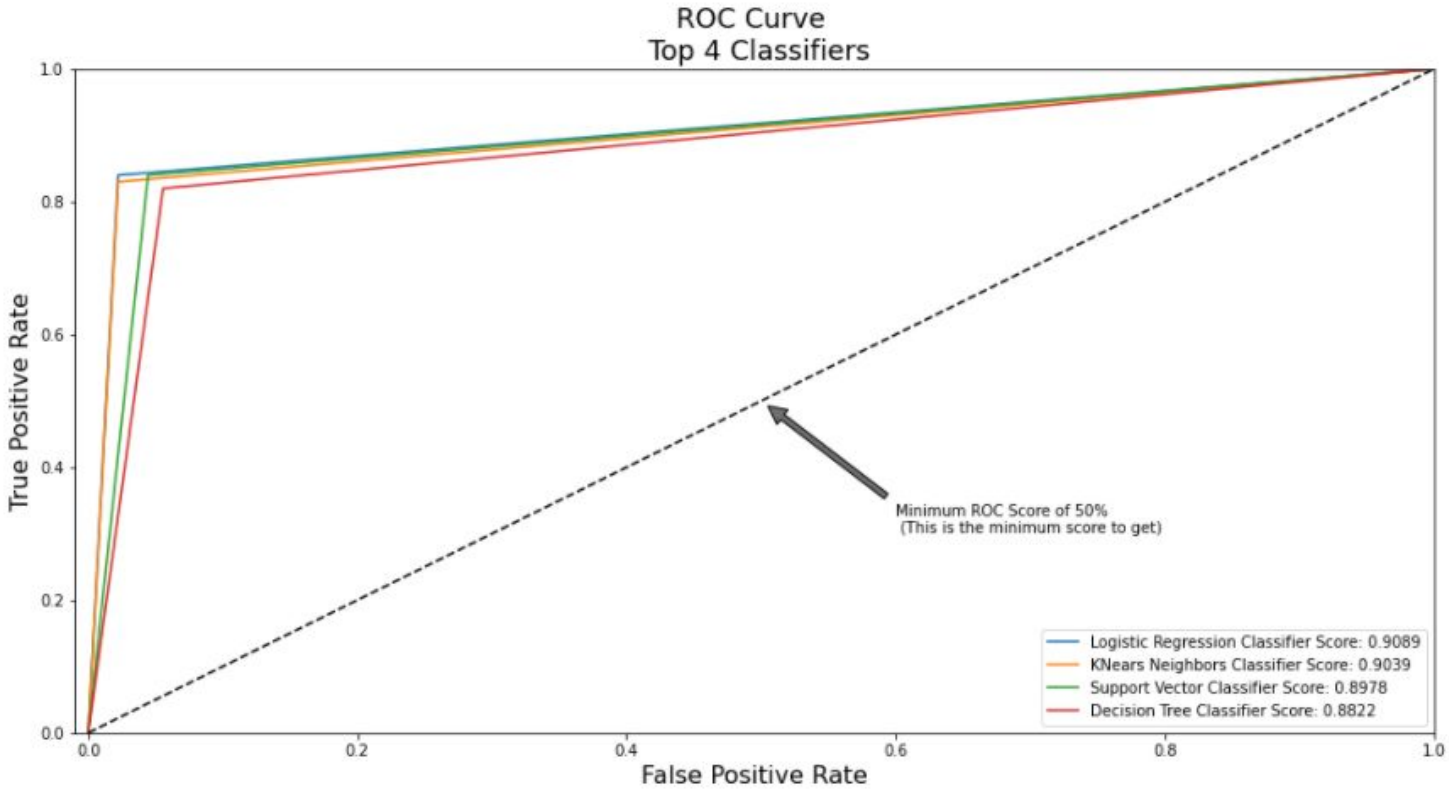
Métricas de Avaliação

- Precision

```
LR  - Precision: 0.9767441860465116
KN  - Precision: 0.9764705882352941
SVC  - Precision: 0.9545454545454546
Tree  - Precision: 0.9425287356321839
```

Métricas de Avaliação

- Roc Curve



Conclusão

Analisando as métricas de avaliação para os classificadores, podemos perceber que os quatro tiveram bom desempenho. Como a base de dados foi balanceada, o TradeOff Precision/Recall apresentou valores próximos, bem equilibrados para todos os classificadores. Contudo o classificador Logistic Regression teve um desempenho levemente superior, mas com métricas bem próximas às do classificador K Neighbors Classifier.

É importante salientar a importância de analisar previamente o DataSet, principalmente em casos onde os dados são altamente desbalanceados, de modo que o tratamento dos dados é essencial para que os classificadores treinem modelos com capacidade real de generalizar as classes.