Treinamento DDoS Dataset

Felipe Ribeiro Quiles

Ciência de Dados para Segurança 2021/2

ÍNDICE

Apresentação do dataset

- O dataset é composto por 12794627 datapoints
- Cada datapoint corresponde a um fluxo (ida e volta).
- Os datapoints são compostos de 82 características
- Os datapoints são classificados entre:
  - DDoS
  - Benigno
- O dataset balanceado tem 51% DDoS e 49% Benigno

Processamento dos Dados para o Treinamento

- Diminuição do Dataset para 500 mil datapoints (20% oculto + 80% treino e teste)
- 50% ddos e 50% benigno
- Exclusão de 6 características
  - FlowID
  - IP Destino
  - IP Origem
  - Porta Destino
  - Porta Origem
  - Timestamp
- Troca das classes de “ddos” e “benign” para 1 e 0, respectivamente

Amostra

[6.0, 1724944.0, 3.0, 1.0, 360.0, 120.0, 120.0, 120.0, 120.0, 0.0, 120.0, 120.0, 120.0, 0.0, 278.2699032548303, 2.318915860456919, 574981.3333333334, 2771.371922592371, 578021.0, 572595.0, 1150616.0, 575308.0, 3836.7613947182067, 578021.0, 572595.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 60.0, 20.0, 1.7391868953426894, 0.5797289651142298, 120.0, 120.0, 120.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 150.0, 120.0, 120.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 3.0, 360.0, 1.0, 120.0, -1.0, 64.0, 3.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1]

Ambiente de Treinamento

- Processador
  - Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz
- Memória
  - 16 GB
- Sistema Operacional
  - Linux Mint 20.04

Método de Treinamento

- Os modelos a serem treinados são baseados em 3 tipos:
  - KNN
  - RandomForest
  - MLP
- Para realizar o treinamento utilizou-se o GridSearchCV
- GridSearchCV é uma ferramenta utilizada para automatizar o processo de ajuste dos parâmetros de um algoritmo
- Para início, decidiu-se que a classe positiva era a “ddos”, marcada como 1 (Problema Binário)
- Treinamentos utilizando taxas de 80/20 e 50/50
- Utilização do K-Folding para comparação de resultados
  - Foi utilizado K = 5
- Escolha do “precision_score” para a métrica do modelo

KNN - Treinamento

- Definição dos parâmetros para o GridSearch
  - 1 Vizinho
  - 3 Vizinhos
  - 5 Vizinhos
- Estatísticas do Treinamento 80/20

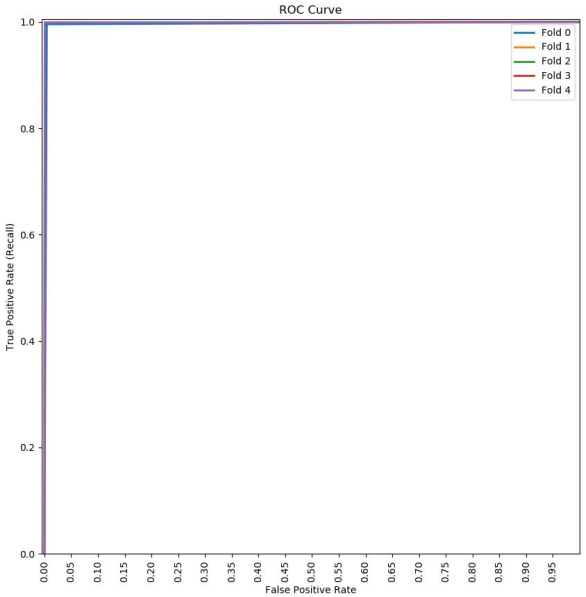
```
Melhores parametros do KNeighbors para o precision_score
{'n_neighbors': 1}

Tempo de treinamento para o KNeighbors: 159.31919622421265

Auc Fold 0: 0.9955827660583146
Auc Fold 1: 0.9990088675790347
Auc Fold 2: 0.9992974465756712
Auc Fold 3: 0.999046279035318
Auc Fold 4: 0.9991473299027351
```

- Estatísticas Treinamento 50/50

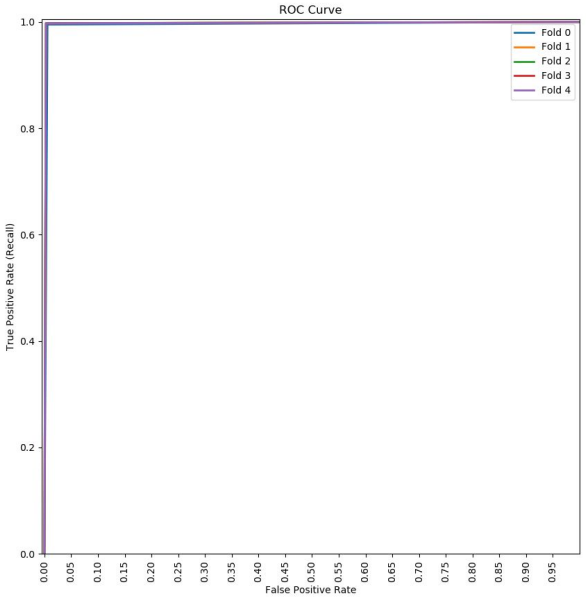
```
Melhores parametros do KNeighbors para o precision_score
{'n_neighbors': 1}

Tempo de treinamento para o KNeighbors: 75.34206819534302

Auc Fold 0: 0.9946326840597697
Auc Fold 1: 0.9973693028338546
Auc Fold 2: 0.9974340483774409
Auc Fold 3: 0.9972134869567371
Auc Fold 4: 0.9972990658683852
```

KNN- Resultados

- Curva ROC - Treinamento 80/20

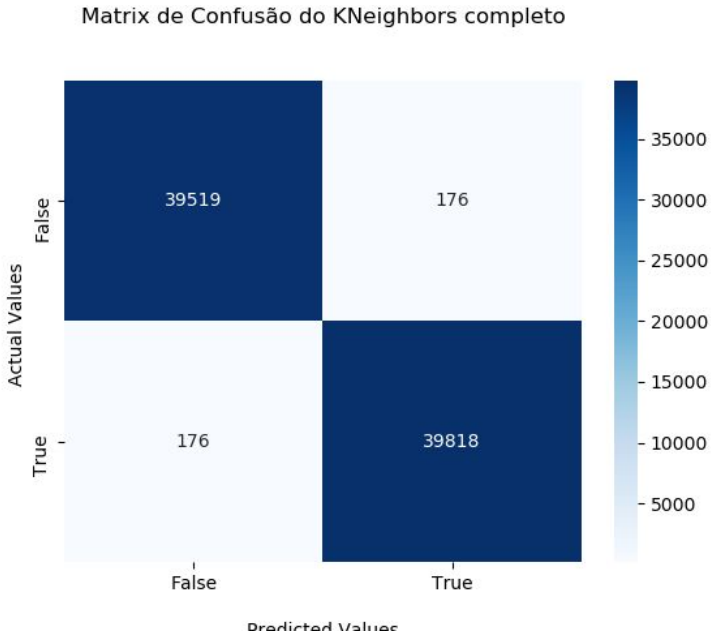


- Curva ROC - Treinamento 50/50

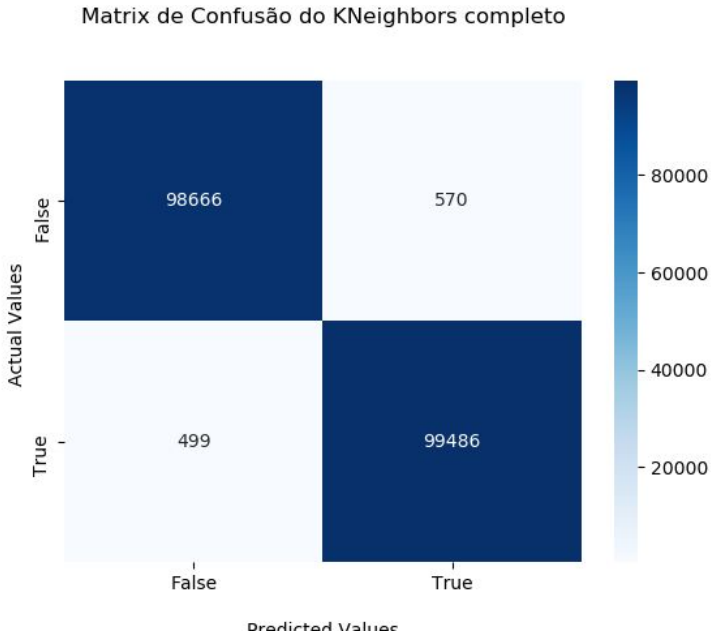


KNN-Resultados

- Matriz de Confusão 80/20



- Matriz de Confusão 50/50



KNN – Comparação

- Autores
  - KNN com K = 6
  - Precisão de 99.95%
  - Misclassification = 2295 de 4222227

- Replicação
  - KNN com K = 1
  - Precisão de 99.12% e 98.93%, 80/20 e 50/50 respectivamente

RF - Treinamento

- Definição dos parâmetros para o GridSearch
  - 50 Árvores
  - 100 Árvores
- Estatísticas do Treinamento 80/20

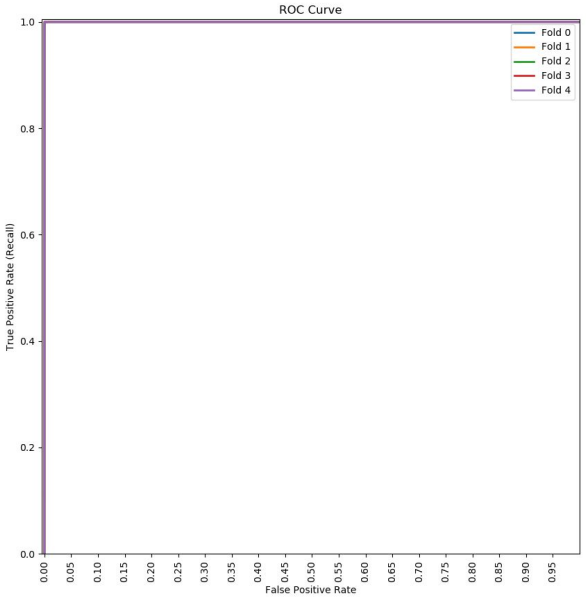
```
Melhores parametros do RandomForest para o precision_score
{'n_estimators': 100}

Tempo de treinamento para o RandomForest: 221.26284837722778

Auc Fold 0: 0.9999029121278382
Auc Fold 1: 0.9999728870402483
Auc Fold 2: 0.999986604616212
Auc Fold 3: 0.999973619295193
Auc Fold 4: 0.9999600695867069
```

- Estatísticas Treinamento 50/50

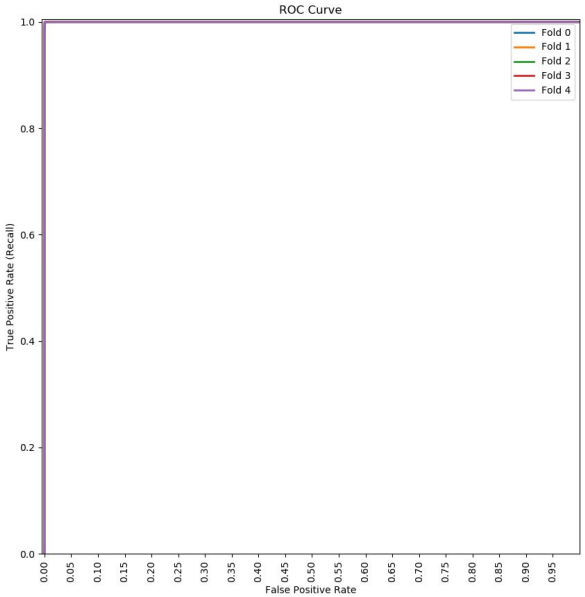
```
Melhores parametros do RandomForest para o precision_score
{'n_estimators': 100}

Tempo de treinamento para o RandomForest: 80.37305212020874

Auc Fold 0: 0.9999229980429992
Auc Fold 1: 0.9999608262816269
Auc Fold 2: 0.9999556606976475
Auc Fold 3: 0.9999709526544666
Auc Fold 4: 0.9999701640118444
```

RF- Resultados

- Curva ROC - Treinamento 80/20

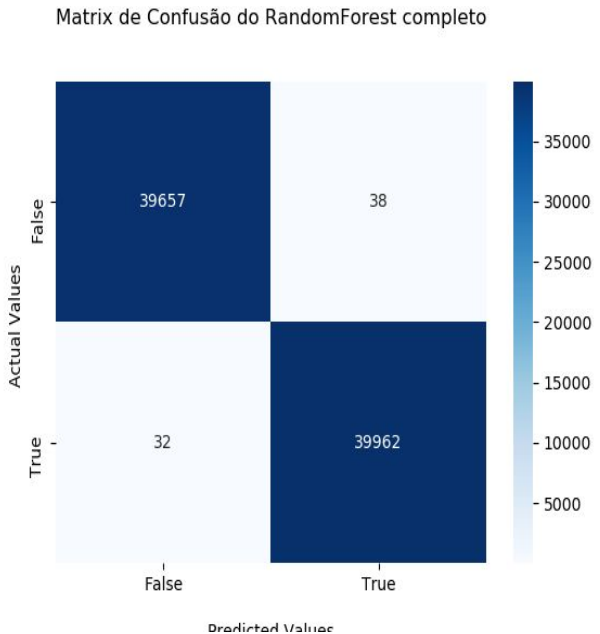


- Curva ROC - Treinamento 50/50

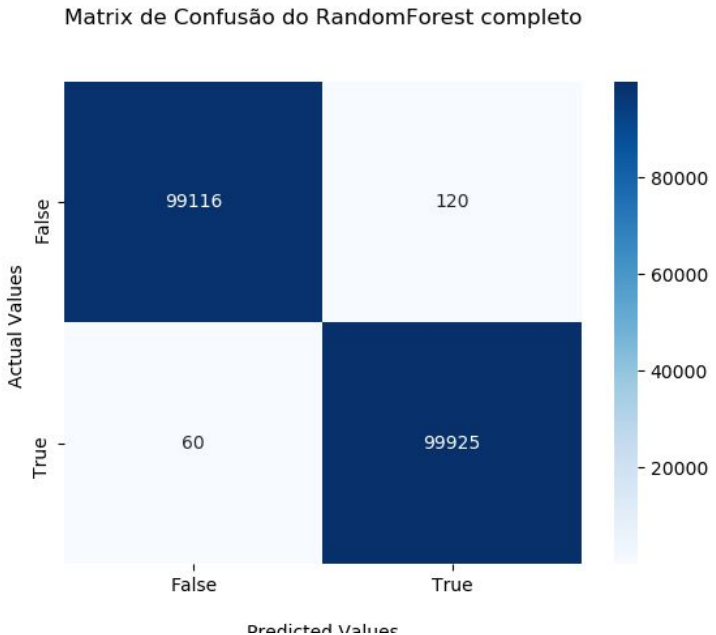


RF-Resultados

- Matriz de Confusão 80/20



- Matriz de Confusão 50/50



RF - Comparação

- Autores
  - RF com n_estimators = 200
  - RF com max_depth = 5
  - Precisão de 99.95%
  - Misclassification = 2315 de 4222227

- Replicação
  - RF com n_estimators = 100
  - RF com max_depth = None
  - Precisão de 99.82% para ambos treinamentos

MLP - Treinamento

- Definição dos parâmetros para o GridSearch
  - 1000 Épocas, erro 0.01
  - 5000 Épocas, erro 0.01
- Estatísticas do Treinamento 80/20

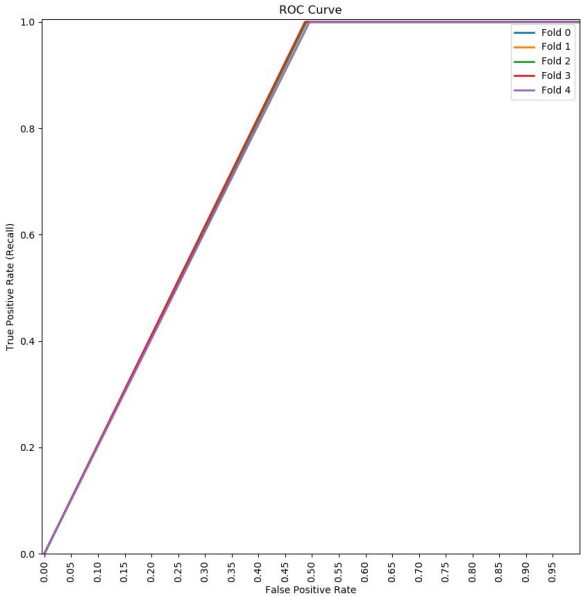
```
Melhores parametros do MLP para o precision_score
{'hidden_layer_sizes': 7, 'max_iter': 1000, 'tol': 0.01}

Tempo de treinamento para o MLP: 133.3489019870758

Auc Fold 0: 0.756177156262317
Auc Fold 1: 0.7553772478896146
Auc Fold 2: 0.7549926279091757
Auc Fold 3: 0.7566618195903589
Auc Fold 4: 0.7523986999611851
```

- Estatísticas Treinamento 50/50

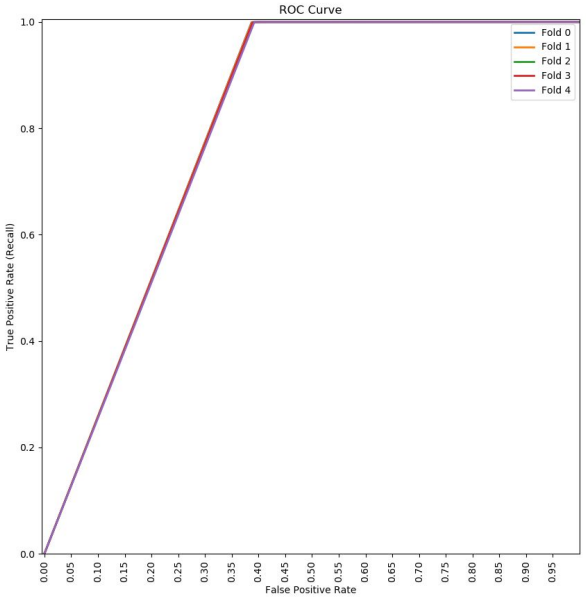
```
Melhores parametros do MLP para o precision_score
{'hidden_layer_sizes': 7, 'max_iter': 1000, 'tol': 0.01}

Tempo de treinamento para o MLP: 90.26089715957642

Auc Fold 0: 0.8064113343471754
Auc Fold 1: 0.8061260448690828
Auc Fold 2: 0.8049639165210506
Auc Fold 3: 0.805409465537288
Auc Fold 4: 0.8038654728033059
```

MLP- Resultados

- Curva ROC - Treinamento 80/20

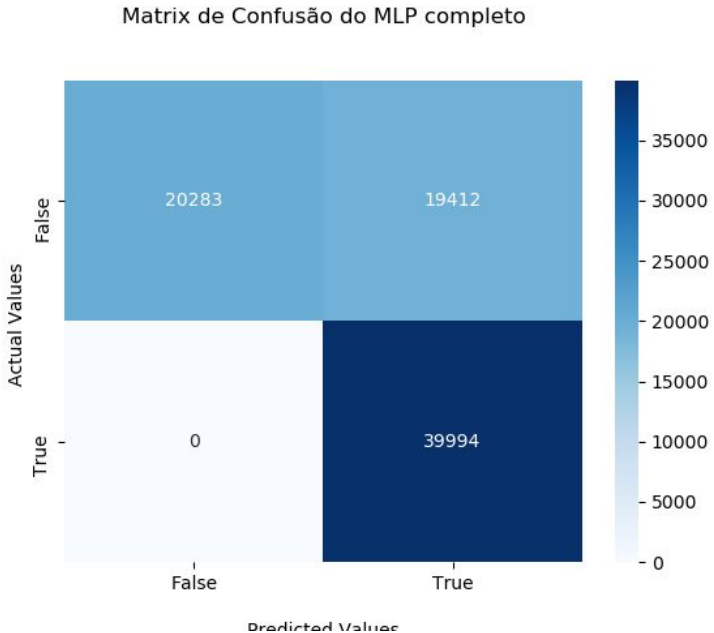


- Curva ROC - Treinamento 50/50

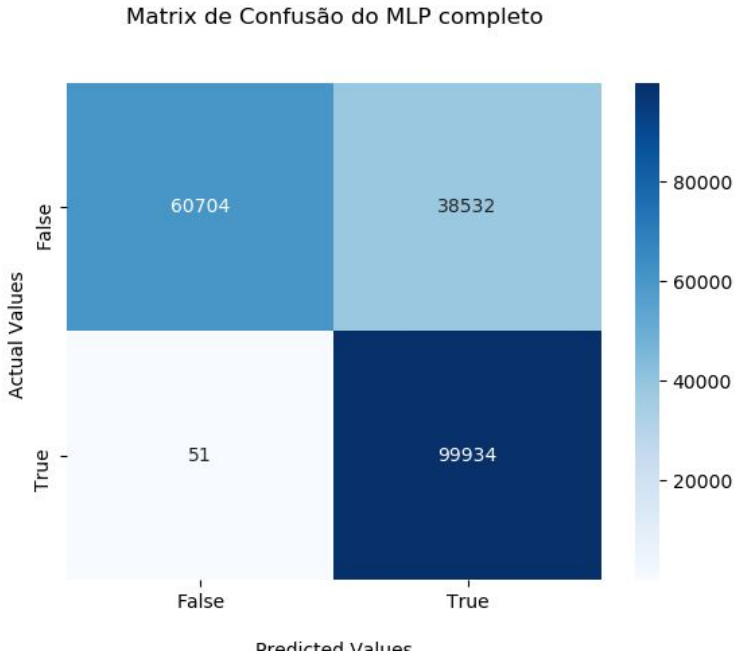


MLP-Resultados

- Matriz de Confusão 80/20



- Matriz de Confusão 50/50



Experimentação dos Modelos

- Experimentação e obtenção de resultados para os 6 seguintes modelos:
  - KNN com 80/20
  - KNN com 50/50
  - RF com 80/20
  - RF com 50/50
  - MLP com 80/20
  - MLP com 50/50
- Será passado os 20% restante do dataset original + arquivo pcap de DDOS

Conclusão

- MLP ficou muito específico em acertar o DDoS, mas resulta em muitos Falso Positivos
- Bons resultados obtidos para o dataset em questão
- Dificuldade de colocar em prática
- KNN e RF não atingem resultados muito expressivos quando testados com outros datasets

Obrigado!