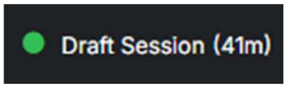
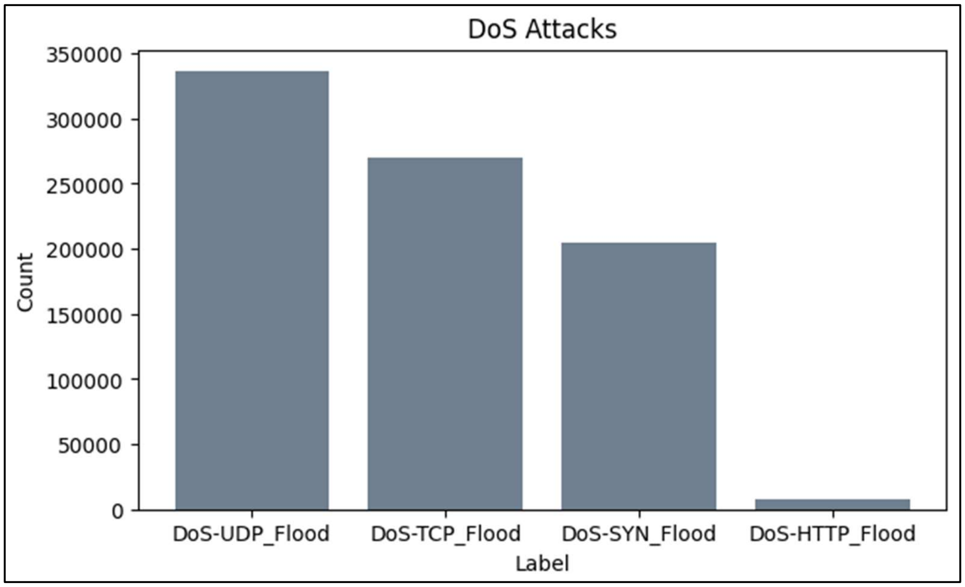
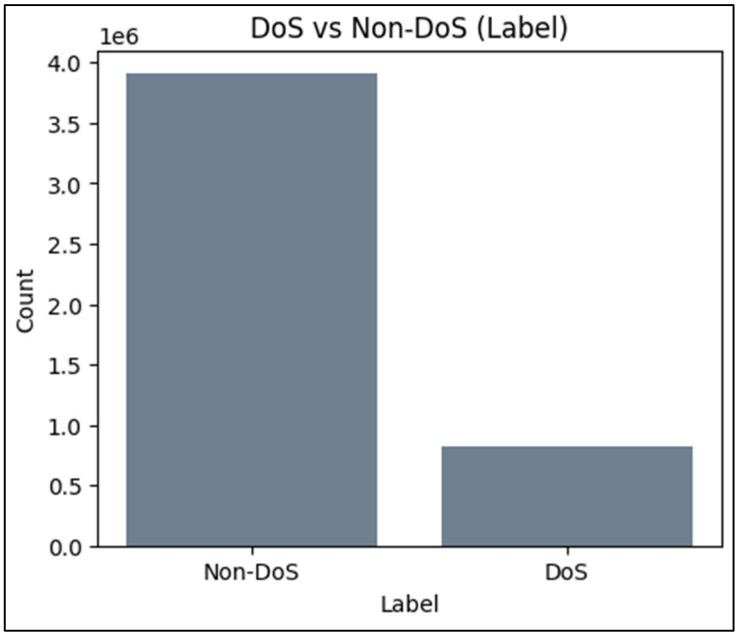
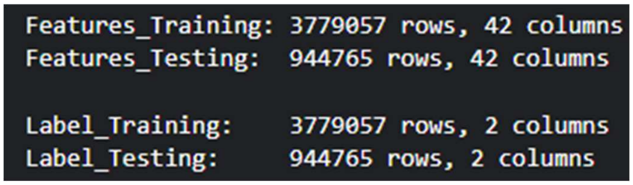
Time Taken to Run All Code in Kaggle Notebook (70 cells): 41 minutes

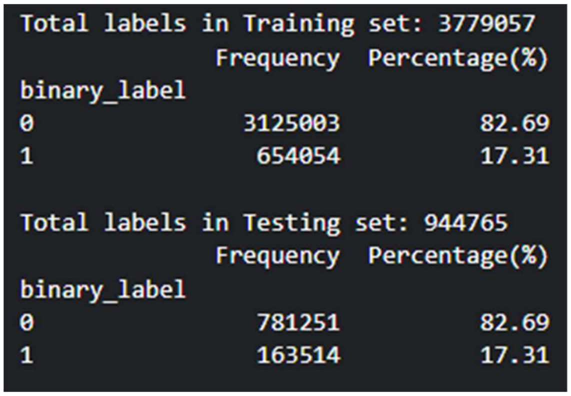




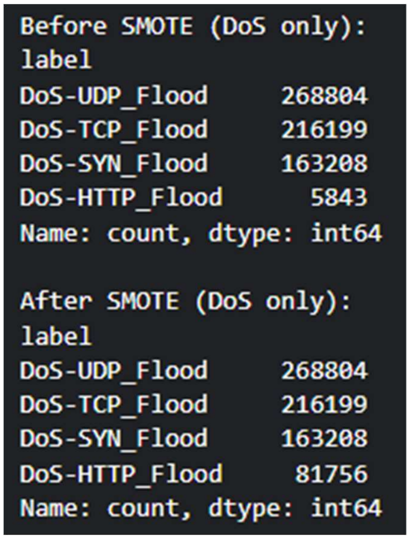




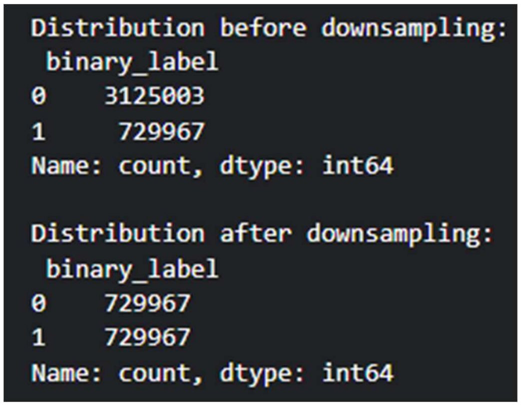
```
Features_Training: 3779057 rows, 42 columns
Features_Testing:  944765 rows, 42 columns

Label_Training:    3779057 rows, 2 columns
Label_Testing:     944765 rows, 2 columns
```



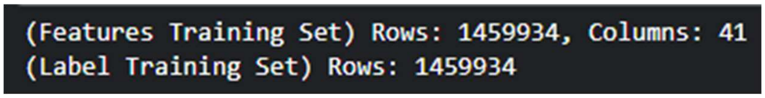
```
Total labels in Training set: 3779057
              Frequency  Percentage(%)
binary_label
0               3125003          82.69
1                654054          17.31

Total labels in Testing set: 944765
              Frequency  Percentage(%)
binary_label
0                781251          82.69
1                163514          17.31
```



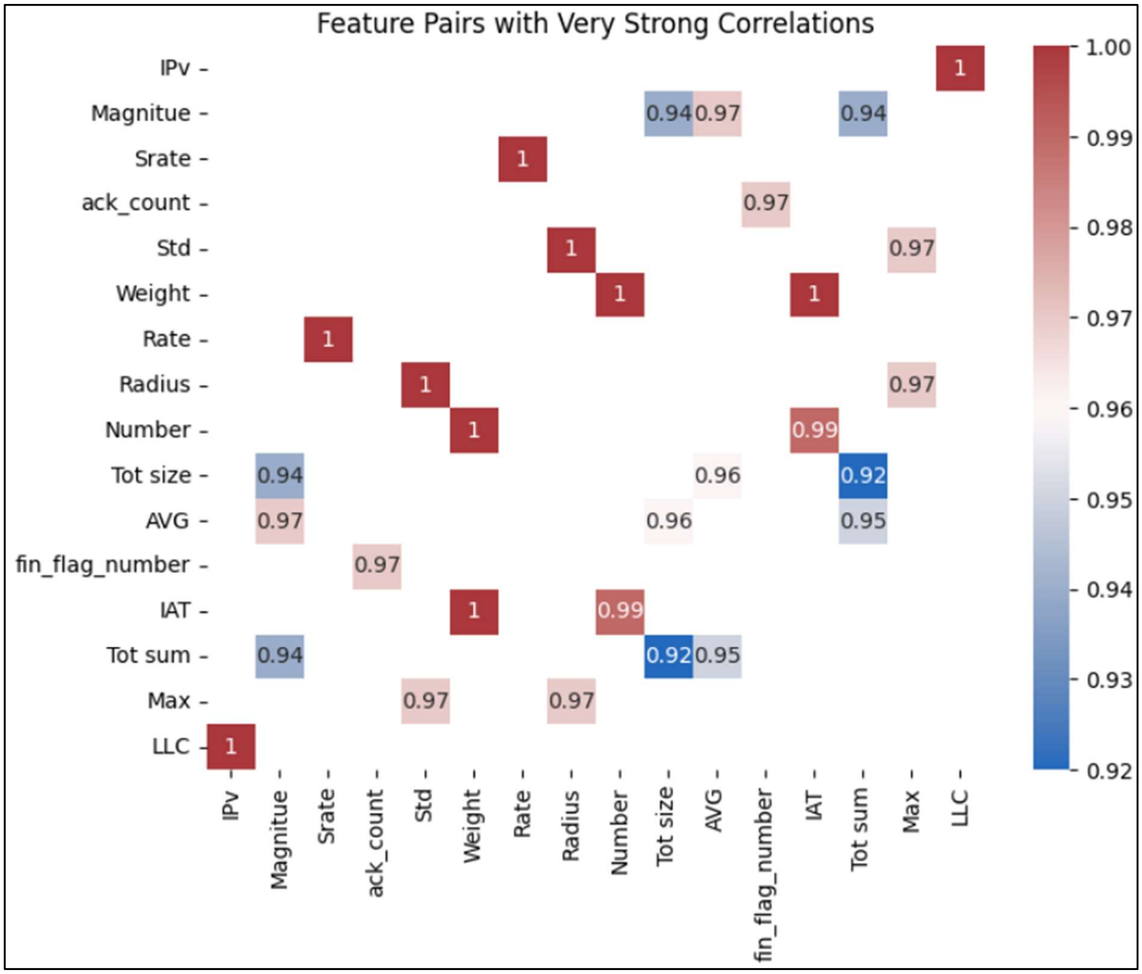
```
Before SMOTE (DoS only):
label
DoS-UDP_Flood      268804
DoS-TCP_Flood      216199
DoS-SYN_Flood      163208
DoS-HTTP_Flood       5843
Name: count, dtype: int64

After SMOTE (DoS only):
label
DoS-UDP_Flood      268804
DoS-TCP_Flood      216199
DoS-SYN_Flood      163208
DoS-HTTP_Flood      81756
Name: count, dtype: int64
```

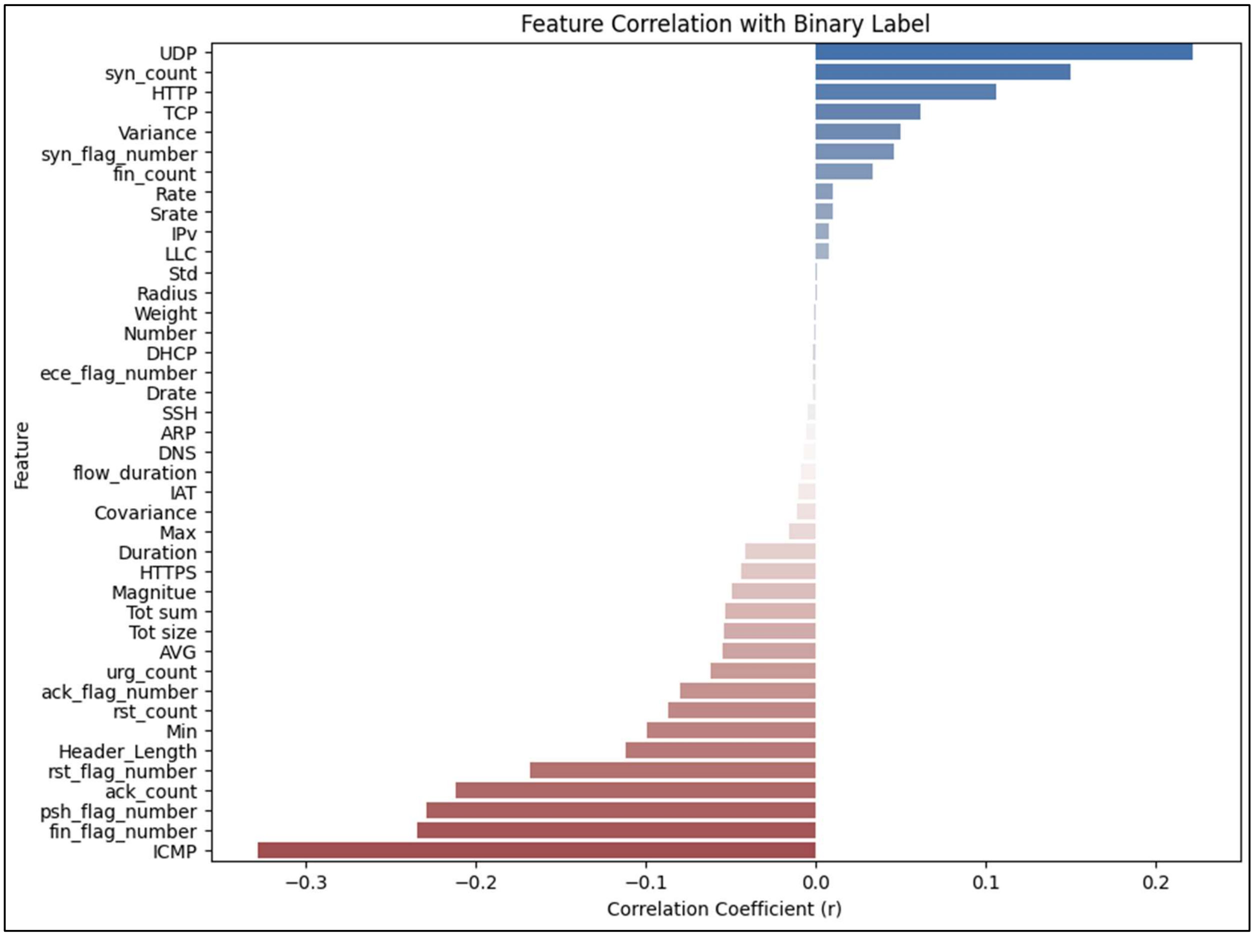


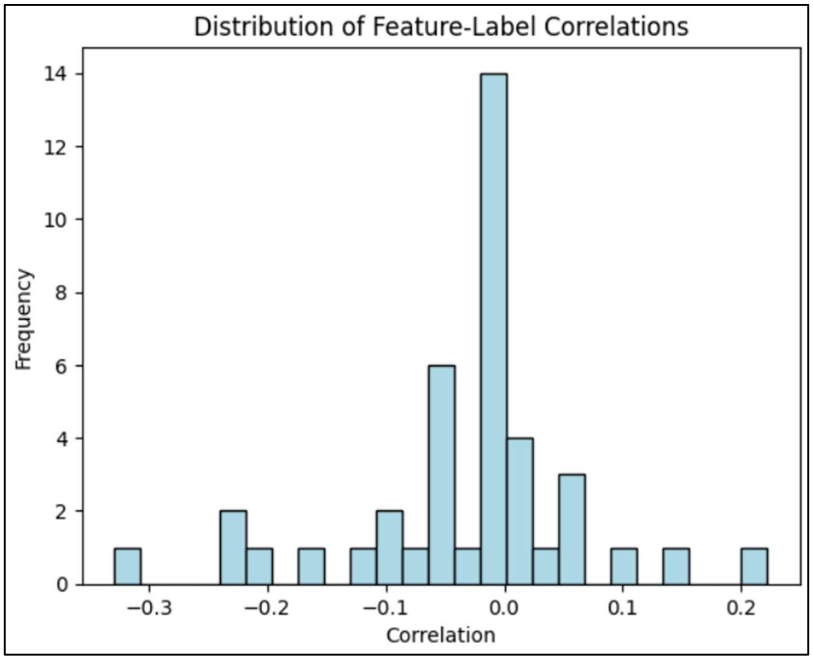
```
Distribution before downsampling:
 binary_label
0    3125003
1     729967
Name: count, dtype: int64

Distribution after downsampling:
 binary_label
0    729967
1    729967
Name: count, dtype: int64
```

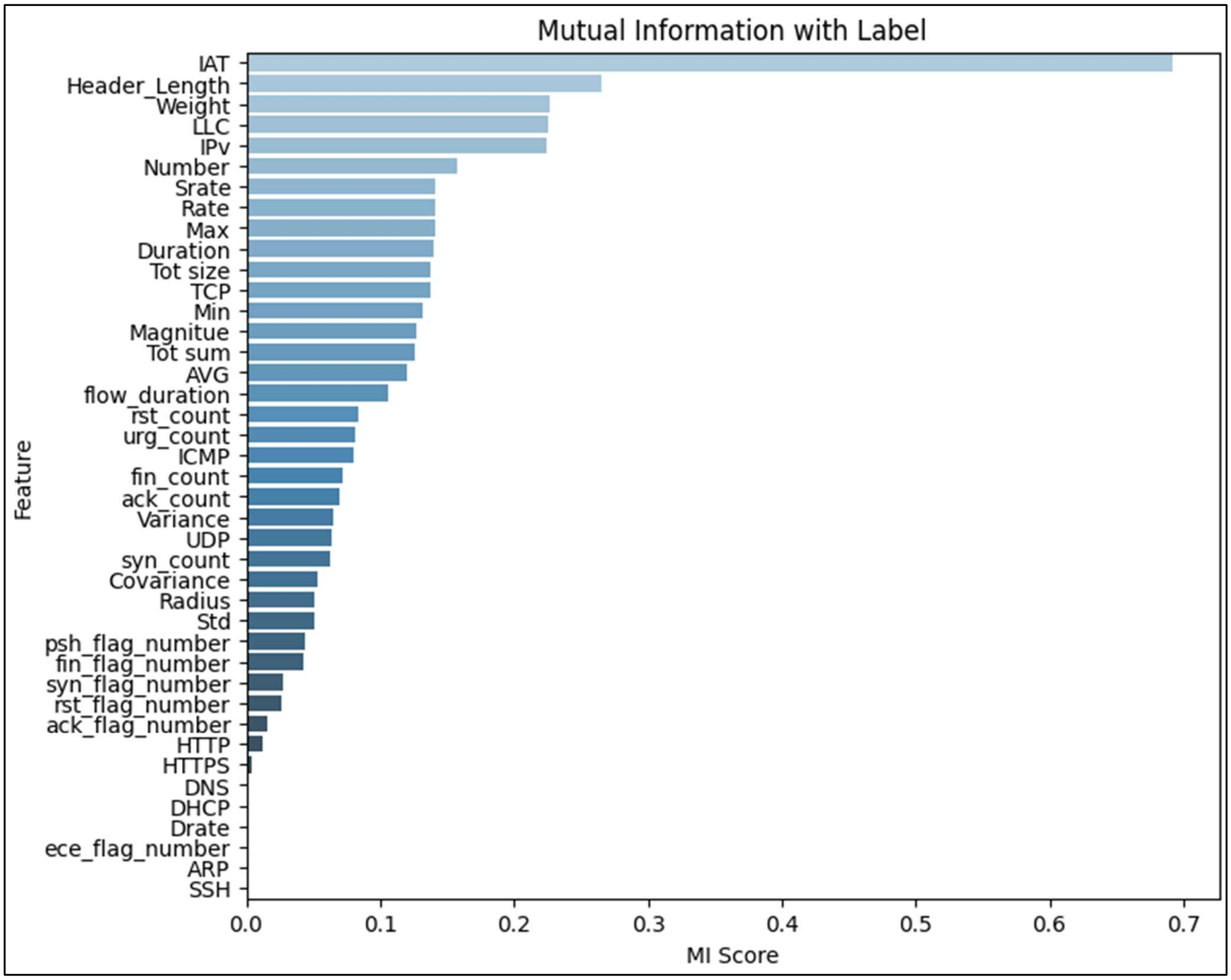


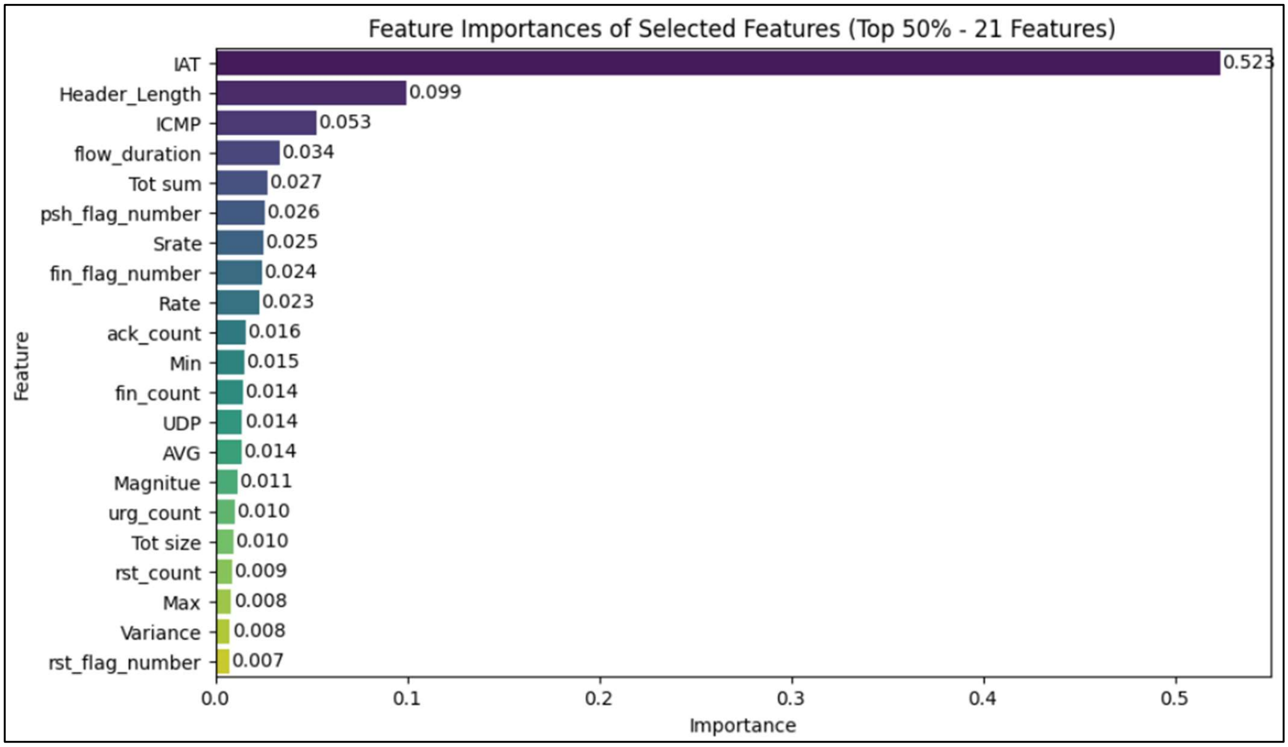
```
(Features Training Set) Rows: 1459934, Columns: 41
(Label Training Set) Rows: 1459934
```

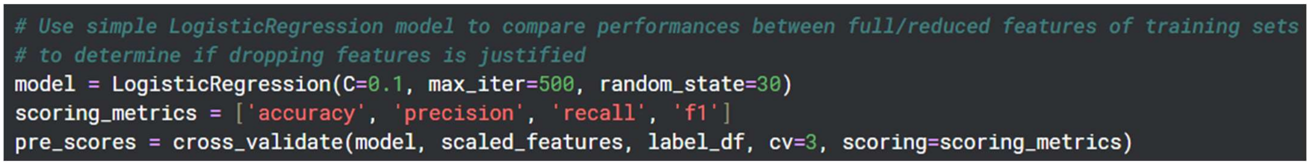



Feature Correlation with Binary Label
UDP
syn_count
HTTP
TCP
Variance
syn_flag_number
fin_count
Rate
Srate
IPv
LLC
Std
Radius
Weight
Number
DHCP
ece_flag_number
Drate
SSH
ARP
DNS
flow_duration
IAT
Covariance
Max
Duration
HTTPS
Magnitue
Tot sum
Tot size
AVG
urg_count
ack_flag_number
rst_count
Min
Header_Length
rst_flag_number
ack_count
psh_flag_number
fin_flag_number
ICMP
Feature
−0.3
−0.2
−0.1
0.0
0.1
0.2
Correlation Coefficient (r)


Distribution of Feature-Label Correlations
14
12
10
8
6
4
2
0
Frequency
−0.3
−0.2
−0.1
0.0
0.1
0.2
Correlation

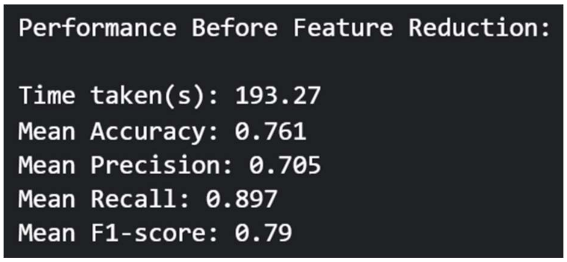

Mutual Information with Label
Feature
IAT
Header_Length
Weight
LLC
IPv
Number
Srate
Rate
Max
Duration
Tot size
TCP
Min
Magnitue
Tot sum
AVG
flow_duration
rst_count
urg_count
ICMP
fin_count
ack_count
Variance
UDP
syn_count
Covariance
Radius
Std
psh_flag_number
fin_flag_number
syn_flag_number
rst_flag_number
ack_flag_number
HTTP
HTTPS
DNS
DHCP
Drate
ece_flag_number
ARP
SSH
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
MI Score


Feature Importances of Selected Features (Top 50% - 21 Features)
Feature
IAT
0.523
Header_Length
0.099
ICMP
0.053
flow_duration
0.034
Tot sum
0.027
psh_flag_number
0.026
Srate
0.025
fin_flag_number
0.024
Rate
0.023
ack_count
0.016
Min
0.015
fin_count
0.014
UDP
0.014
AVG
0.014
Magnitue
0.011
urg_count
0.010
Tot size
0.010
rst_count
0.009
Max
0.008
Variance
0.008
rst_flag_number
0.007
0.0
0.1
0.2
0.3
0.4
0.5
Importance

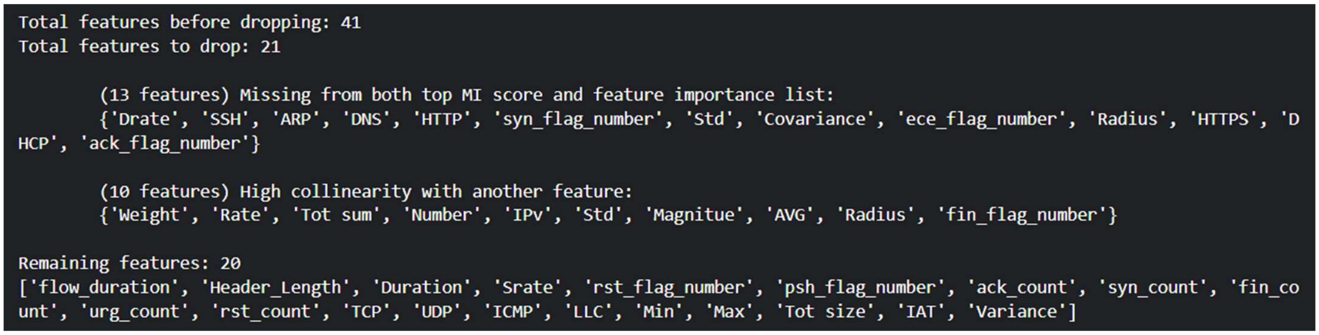


```
# Use simple LogisticRegression model to compare performances between full/reduced features of training sets
# to determine if dropping features is justified
model = LogisticRegression(C=0.1, max_iter=500, random_state=30)
scoring_metrics = ['accuracy', 'precision', 'recall', 'f1']
pre_scores = cross_validate(model, scaled_features, label_df, cv=3, scoring=scoring_metrics)
```

# Full feature set (41 features)

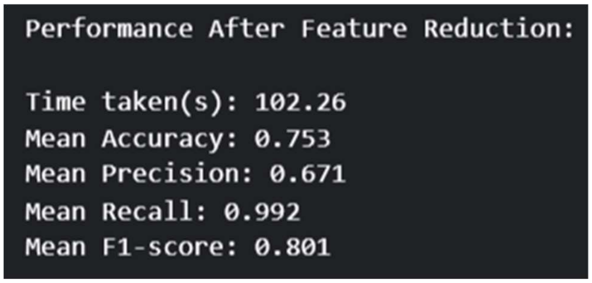


```
Performance Before Feature Reduction:

Time taken(s): 193.27
Mean Accuracy: 0.761
Mean Precision: 0.705
Mean Recall: 0.897
Mean F1-score: 0.79
```

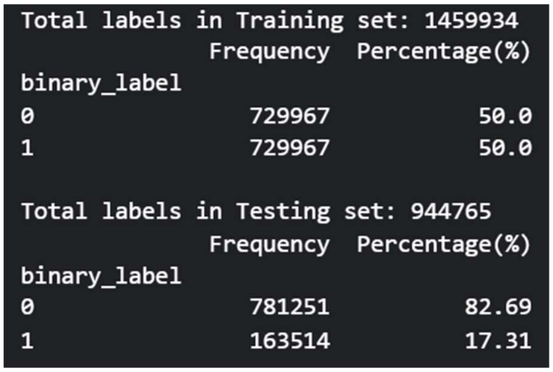


```
Total features before dropping: 41
Total features to drop: 21

        (13 features) Missing from both top MI score and feature importance list:
        {'Drate', 'SSH', 'ARP', 'DNS', 'HTTP', 'syn_flag_number', 'Std', 'Covariance', 'ece_flag_number', 'Radius', 'HTTPS', 'D
HCP', 'ack_flag_number'}

        (10 features) High collinearity with another feature:
        {'Weight', 'Rate', 'Tot sum', 'Number', 'IPv', 'Std', 'Magnitue', 'AVG', 'Radius', 'fin_flag_number'}

Remaining features: 20
['flow_duration', 'Header_Length', 'Duration', 'Srate', 'rst_flag_number', 'psh_flag_number', 'ack_count', 'syn_count', 'fin_co
unt', 'urg_count', 'rst_count', 'TCP', 'UDP', 'ICMP', 'LLC', 'Min', 'Max', 'Tot size', 'IAT', 'Variance']
```

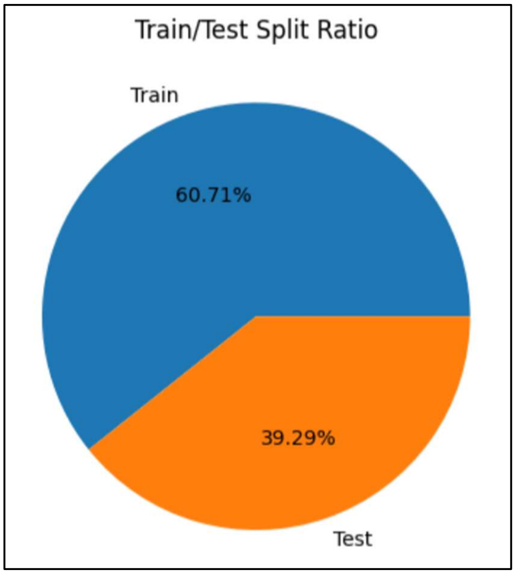
# Reduced feature set (20 features)

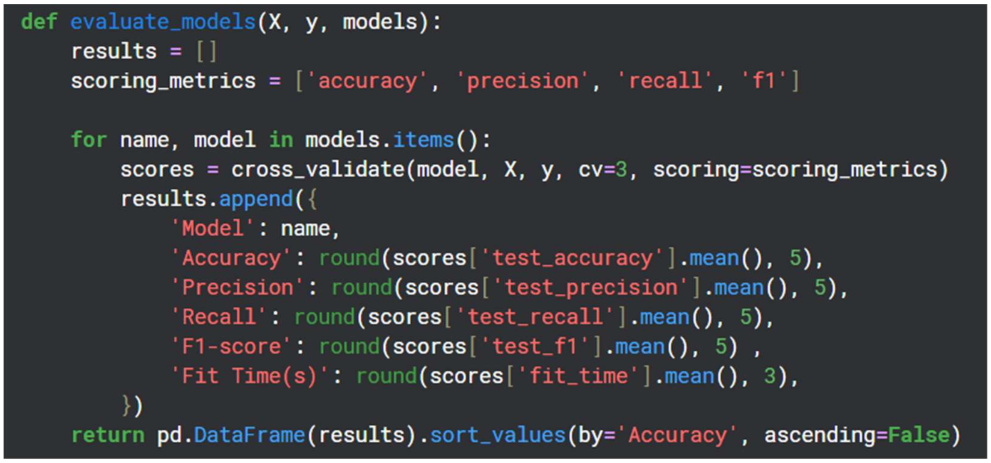


```
Performance After Feature Reduction:

Time taken(s): 102.26
Mean Accuracy: 0.753
Mean Precision: 0.671
Mean Recall: 0.992
Mean F1-score: 0.801
```

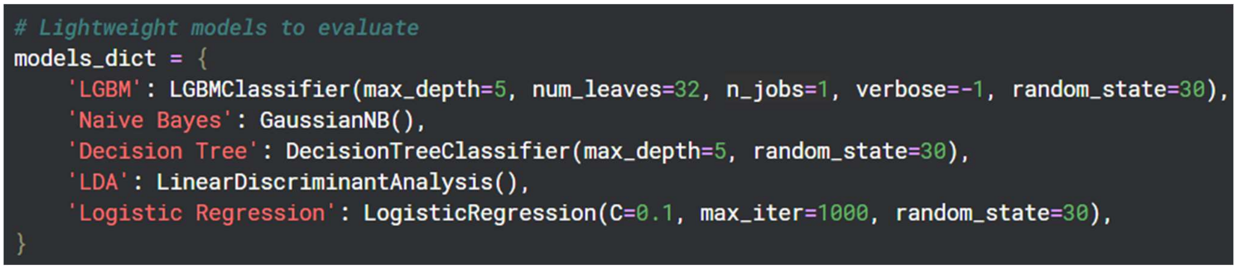


```
Total labels in Training set: 1459934
              Frequency  Percentage(%)
binary_label
0                729967           50.0
1                729967           50.0

Total labels in Testing set: 944765
              Frequency  Percentage(%)
binary_label
0                781251          82.69
1                163514          17.31
```



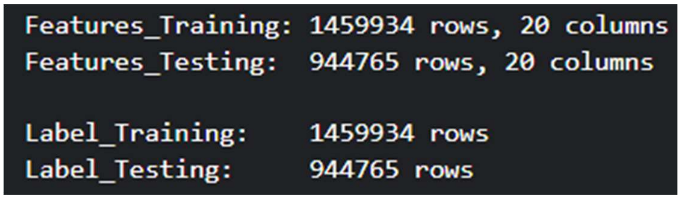


```
def evaluate_models(X, y, models):
    results = []
    scoring_metrics = ['accuracy', 'precision', 'recall', 'f1']

    for name, model in models.items():
        scores = cross_validate(model, X, y, cv=3, scoring=scoring_metrics)
        results.append({
            'Model': name,
            'Accuracy': round(scores['test_accuracy'].mean(), 5),
            'Precision': round(scores['test_precision'].mean(), 5),
            'Recall': round(scores['test_recall'].mean(), 5),
            'F1-score': round(scores['test_f1'].mean(), 5) ,
            'Fit Time(s)': round(scores['fit_time'].mean(), 3),
        })
    return pd.DataFrame(results).sort_values(by='Accuracy', ascending=False)
```

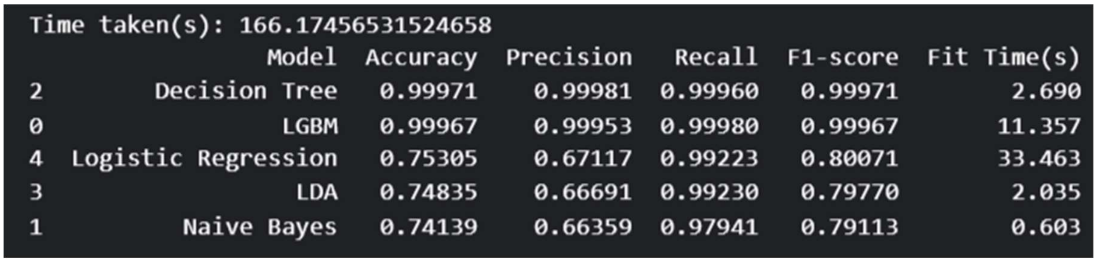


```
# Lightweight models to evaluate
models_dict = {
    'LGBM': LGBMClassifier(max_depth=5, num_leaves=32, n_jobs=1, verbose=-1, random_state=30),
    'Naive Bayes': GaussianNB(),
    'Decision Tree': DecisionTreeClassifier(max_depth=5, random_state=30),
    'LDA': LinearDiscriminantAnalysis(),
    'Logistic Regression': LogisticRegression(C=0.1, max_iter=1000, random_state=30),
}
```

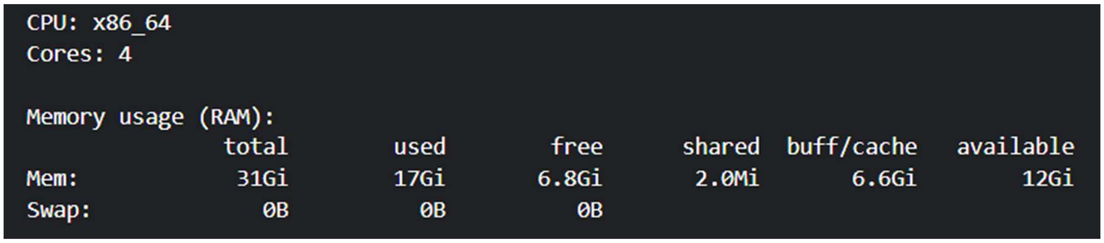
# Performance With IAT (20 Features)



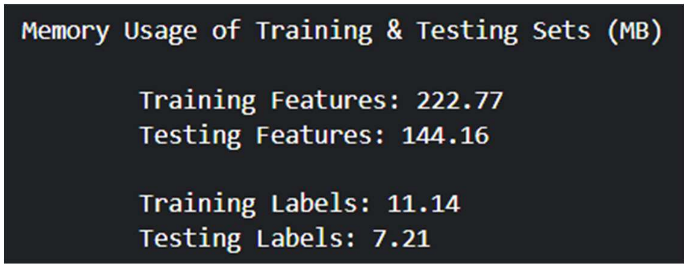
```
Features_Training: 1459934 rows, 20 columns
Features_Testing:  944765 rows, 20 columns

Label_Training:    1459934 rows
Label_Testing:     944765 rows
```



```
Time taken(s): 166.17456531524658
                 Model  Accuracy  Precision   Recall  F1-score  Fit Time(s)
2        Decision Tree   0.99971    0.99981  0.99960   0.99971        2.690
0                 LGBM   0.99967    0.99953  0.99980   0.99967       11.357
4  Logistic Regression   0.75305    0.67117  0.99223   0.80071       33.463
3                  LDA   0.74835    0.66691  0.99230   0.79770        2.035
1          Naive Bayes   0.74139    0.66359  0.97941   0.79113        0.603
```

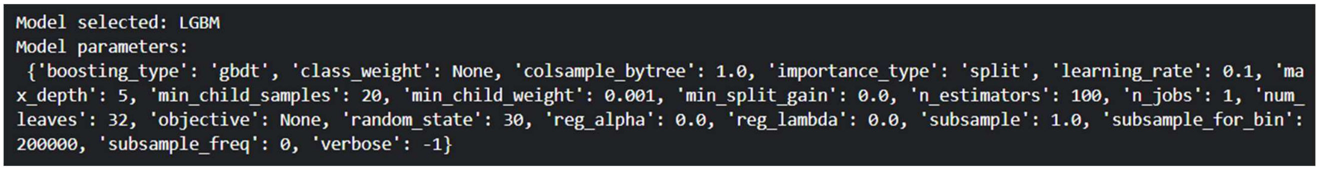


```
CPU: x86_64
Cores: 4

Memory usage (RAM):
               total        used        free      shared  buff/cache   available
Mem:            31Gi        17Gi       6.8Gi       2.0Mi       6.6Gi        12Gi
Swap:             0B          0B          0B
```

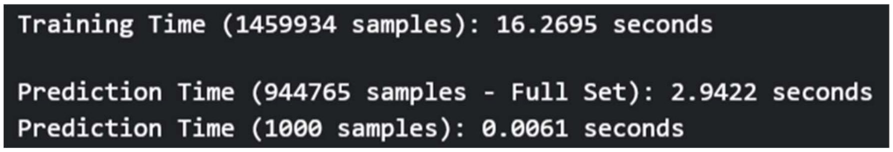


```
Memory Usage of Training & Testing Sets (MB)

        Training Features: 222.77
        Testing Features: 144.16

        Training Labels: 11.14
        Testing Labels: 7.21
```

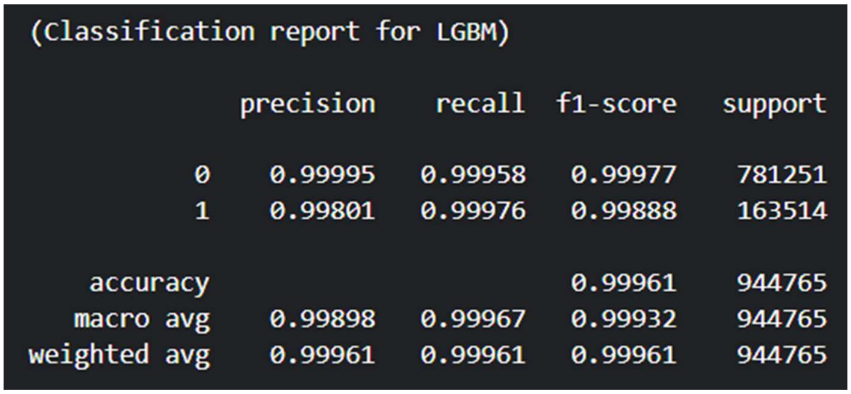
(LGBM) – Selected Model



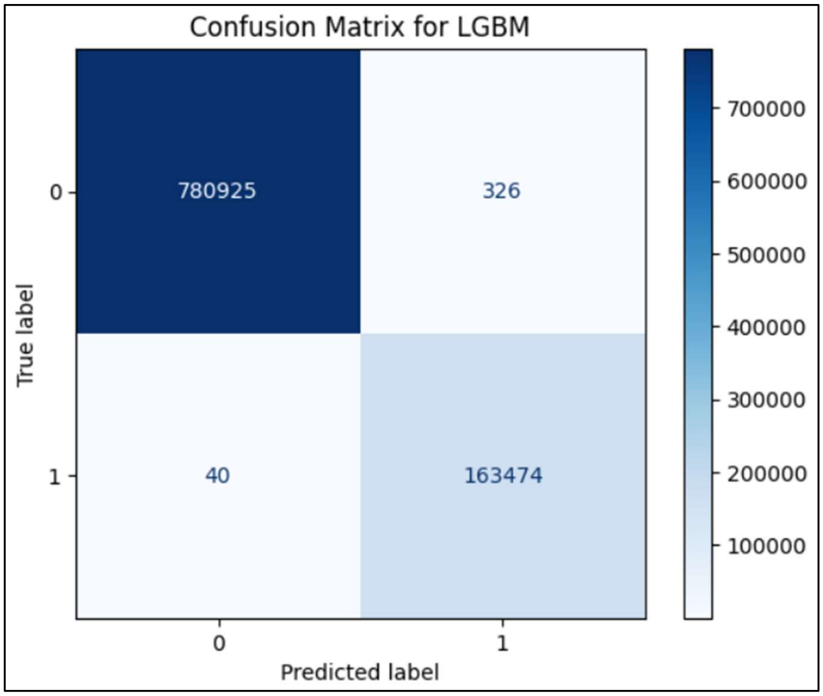
```
Model selected: LGBM
Model parameters:
 {'boosting_type': 'gbdt', 'class_weight': None, 'colsample_bytree': 1.0, 'importance_type': 'split', 'learning_rate': 0.1, 'ma
x_depth': 5, 'min_child_samples': 20, 'min_child_weight': 0.001, 'min_split_gain': 0.0, 'n_estimators': 100, 'n_jobs': 1, 'num_
leaves': 32, 'objective': None, 'random_state': 30, 'reg_alpha': 0.0, 'reg_lambda': 0.0, 'subsample': 1.0, 'subsample_for_bin':
200000, 'subsample_freq': 0, 'verbose': -1}
```

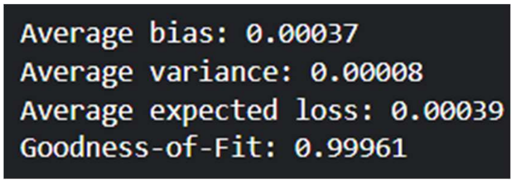


```
Training Time (1459934 samples): 16.2695 seconds

Prediction Time (944765 samples - Full Set): 2.9422 seconds
Prediction Time (1000 samples): 0.0061 seconds
```

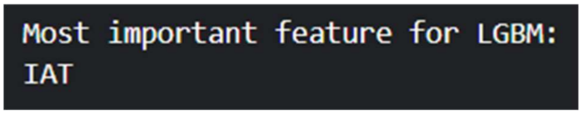


```
 (Classification report for LGBM)

              precision    recall  f1-score   support

           0    0.99995   0.99958   0.99977    781251
           1    0.99801   0.99976   0.99888    163514

    accuracy                        0.99961    944765
   macro avg    0.99898   0.99967   0.99932    944765
weighted avg    0.99961   0.99961   0.99961    944765
```

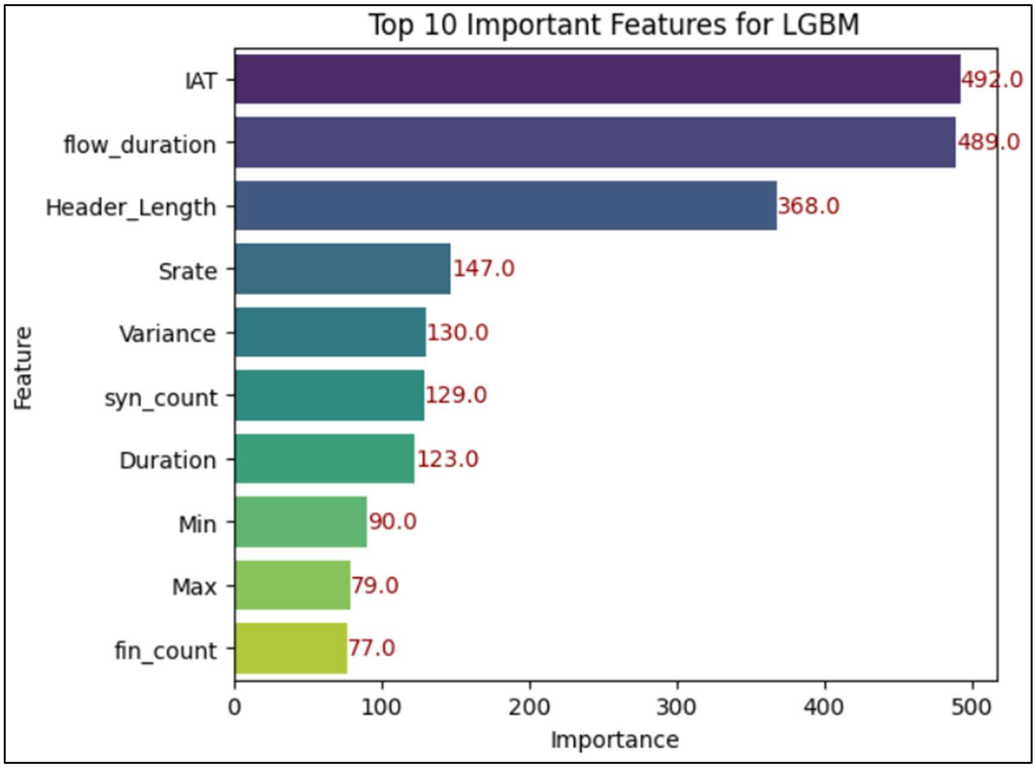


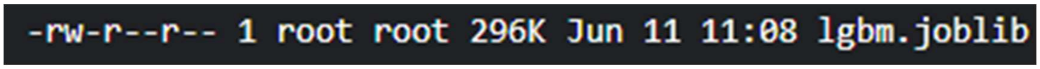


```
Average bias: 0.00037
Average variance: 0.00008
Average expected loss: 0.00039
Goodness-of-Fit: 0.99961
```



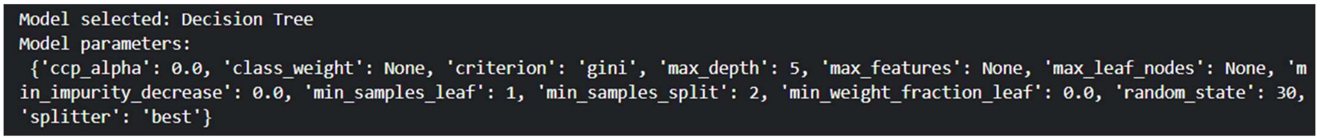
```
Most important feature for LGBM:
IAT
```



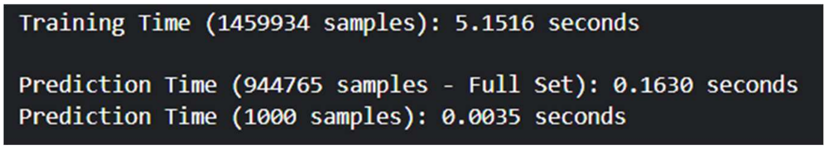


```
-rw-r--r-- 1 root root 296K Jun 11 11:08 lgbm.joblib
```

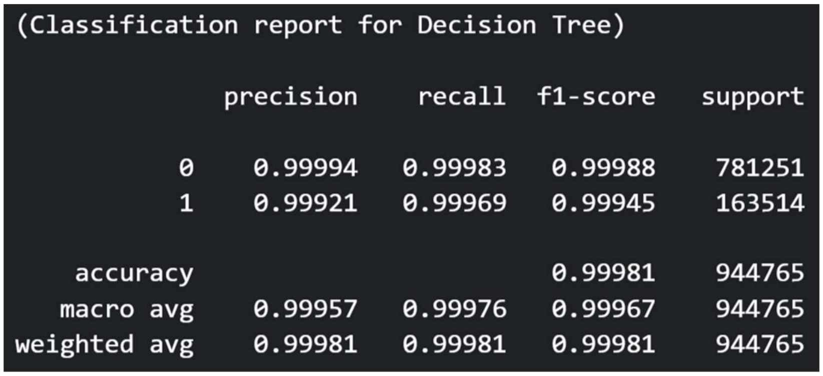
(Decision Tree)



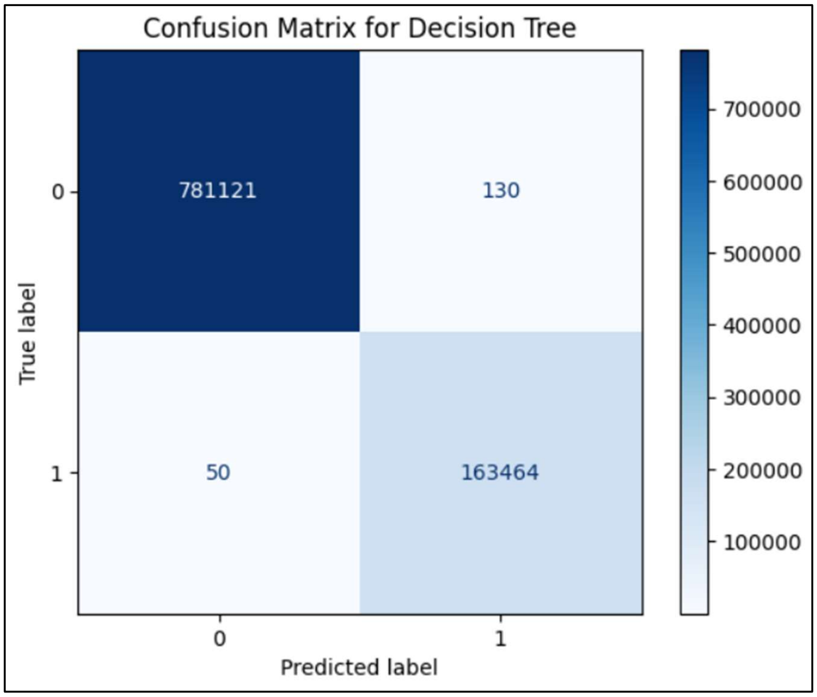
```
Model selected: Decision Tree
Model parameters:
 {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 5, 'max_features': None, 'max_leaf_nodes': None, 'm
in_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'random_state': 30,
 'splitter': 'best'}
```

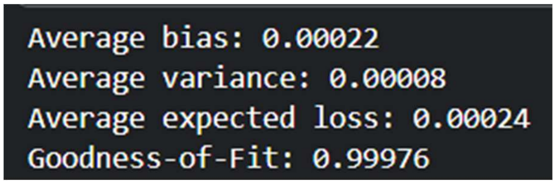


```
Training Time (1459934 samples): 5.1516 seconds

Prediction Time (944765 samples - Full Set): 0.1630 seconds
Prediction Time (1000 samples): 0.0035 seconds
```

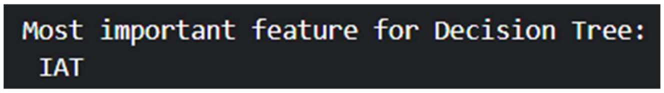


```
(Classification report for Decision Tree)

              precision    recall  f1-score   support

           0    0.99994   0.99983   0.99988    781251
           1    0.99921   0.99969   0.99945    163514

    accuracy                        0.99981    944765
   macro avg    0.99957   0.99976   0.99967    944765
weighted avg    0.99981   0.99981   0.99981    944765
```

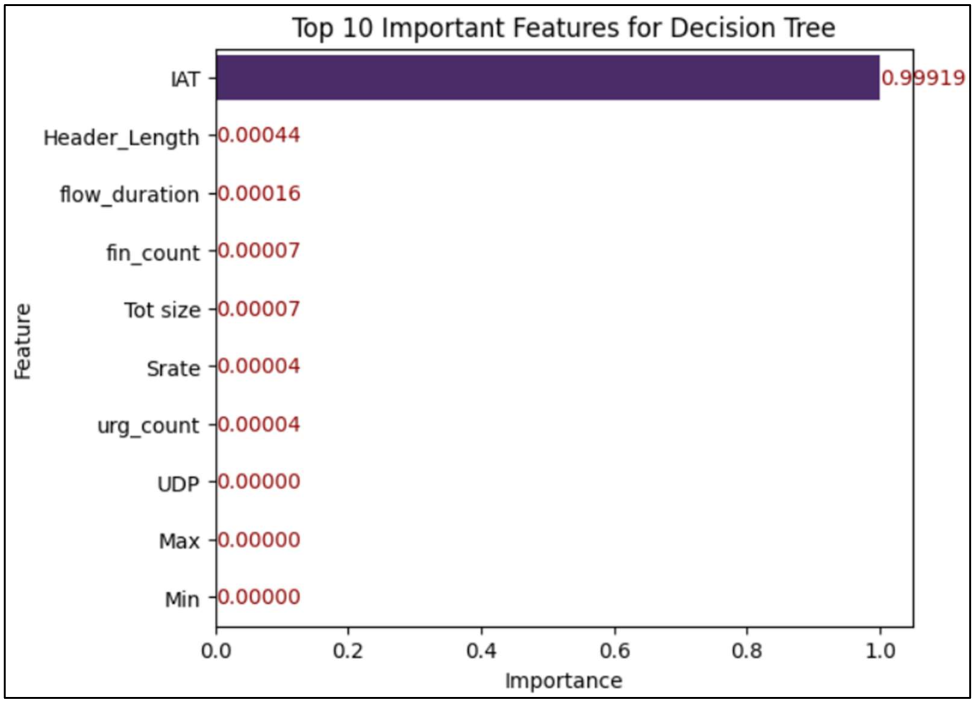


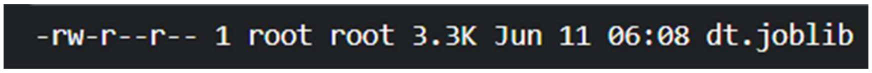


```
Average bias: 0.00022
Average variance: 0.00008
Average expected loss: 0.00024
Goodness-of-Fit: 0.99976
```



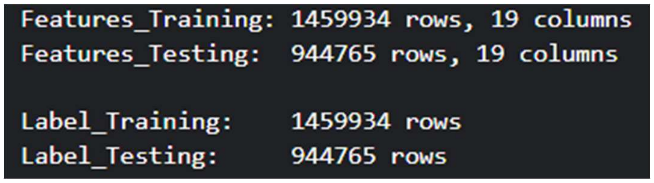
```
Most important feature for Decision Tree:
 IAT
```



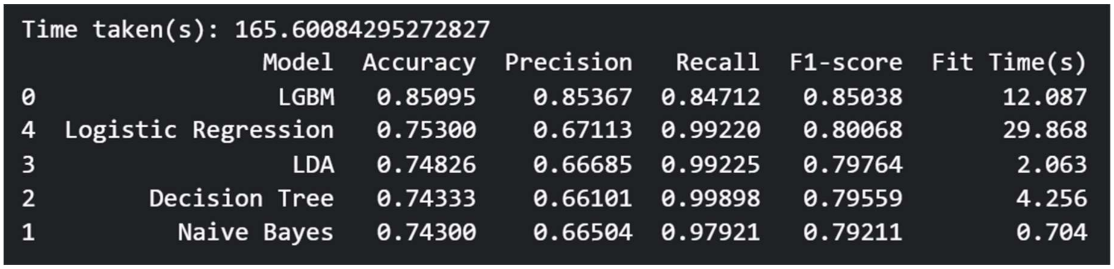


```
 -rw-r--r-- 1 root root 3.3K Jun 11 06:08 dt.joblib
```

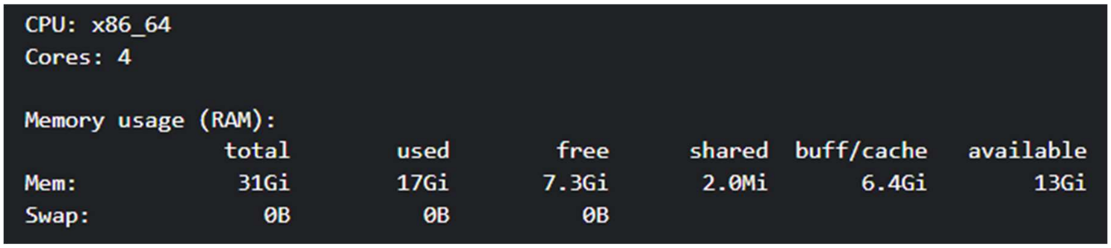
# Performance Without IAT (19 Features)



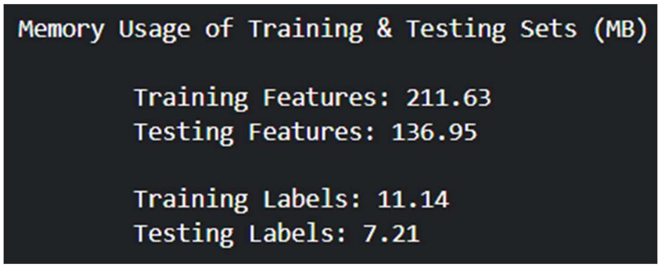
```
Features_Training: 1459934 rows, 19 columns
Features_Testing:  944765 rows, 19 columns

Label_Training:    1459934 rows
Label_Testing:     944765 rows
```



```
Time taken(s): 165.60084295272827
                 Model  Accuracy  Precision   Recall  F1-score  Fit Time(s)
0                 LGBM   0.85095    0.85367  0.84712   0.85038       12.087
4  Logistic Regression   0.75300    0.67113  0.99220   0.80068       29.868
3                  LDA   0.74826    0.66685  0.99225   0.79764        2.063
2        Decision Tree   0.74333    0.66101  0.99898   0.79559        4.256
1          Naive Bayes   0.74300    0.66504  0.97921   0.79211        0.704
```

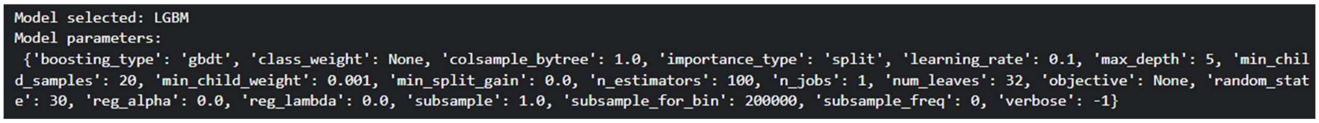


```
CPU: x86_64
Cores: 4

Memory usage (RAM):
               total        used        free      shared  buff/cache   available
Mem:            31Gi        17Gi       7.3Gi       2.0Mi       6.4Gi        13Gi
Swap:             0B          0B          0B
```

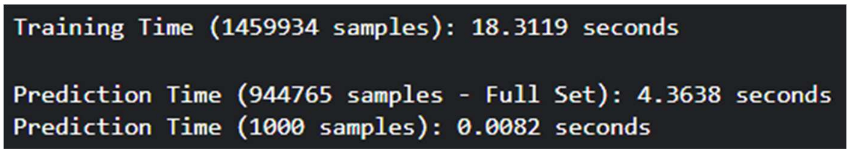


```
Memory Usage of Training & Testing Sets (MB)

        Training Features: 211.63
        Testing Features: 136.95

        Training Labels: 11.14
        Testing Labels: 7.21
```

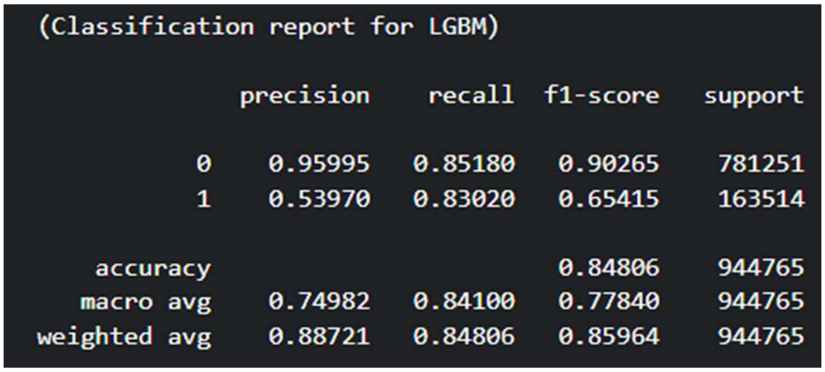
(LGBM)



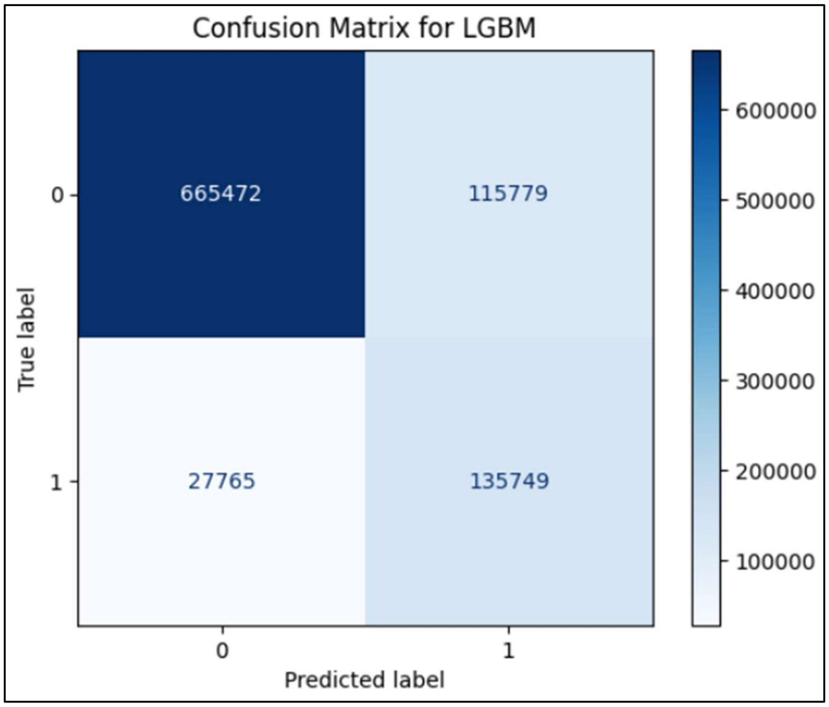
```
Model selected: LGBM
Model parameters:
 {'boosting_type': 'gbdt', 'class_weight': None, 'colsample_bytree': 1.0, 'importance_type': 'split', 'learning_rate': 0.1, 'max_depth': 5, 'min_chil
d_samples': 20, 'min_child_weight': 0.001, 'min_split_gain': 0.0, 'n_estimators': 100, 'n_jobs': 1, 'num_leaves': 32, 'objective': None, 'random_stat
e': 30, 'reg_alpha': 0.0, 'reg_lambda': 0.0, 'subsample': 1.0, 'subsample_for_bin': 200000, 'subsample_freq': 0, 'verbose': -1}
```

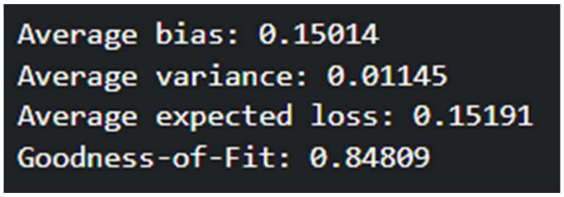


```
Training Time (1459934 samples): 18.3119 seconds

Prediction Time (944765 samples - Full Set): 4.3638 seconds
Prediction Time (1000 samples): 0.0082 seconds
```

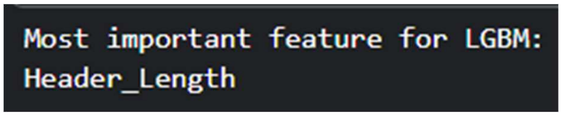


```
 (Classification report for LGBM)

              precision    recall  f1-score   support

           0    0.95995   0.85180   0.90265    781251
           1    0.53970   0.83020   0.65415    163514

    accuracy                        0.84806    944765
   macro avg    0.74982   0.84100   0.77840    944765
weighted avg    0.88721   0.84806   0.85964    944765
```

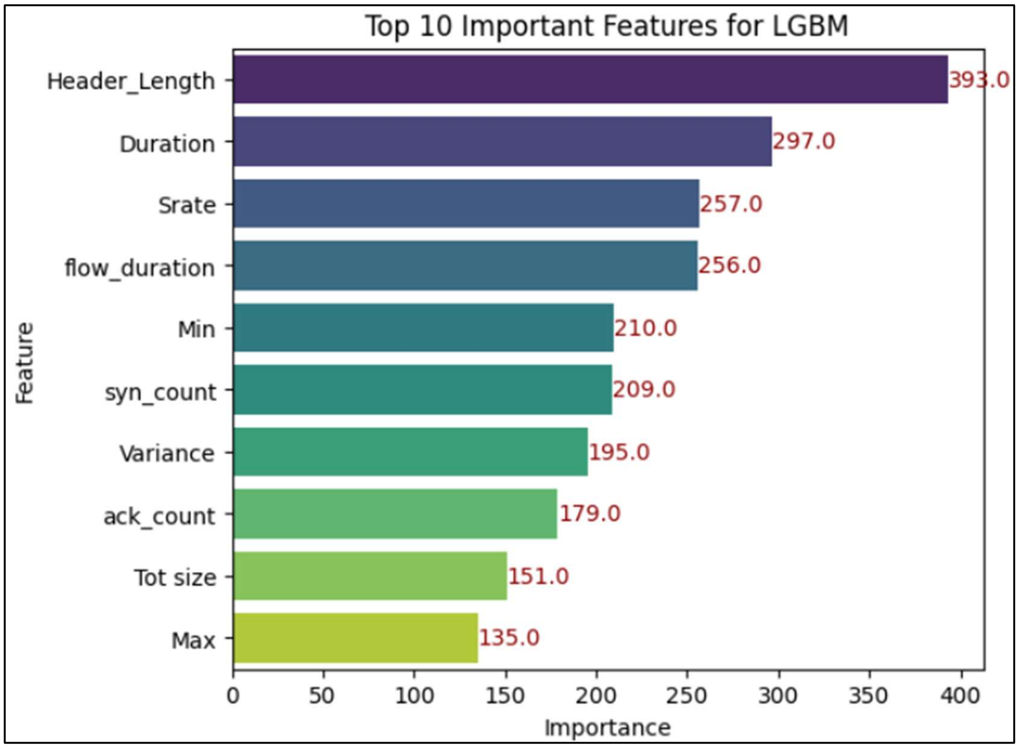


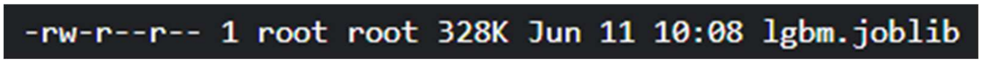


```
Average bias: 0.15014
Average variance: 0.01145
Average expected loss: 0.15191
Goodness-of-Fit: 0.84809
```



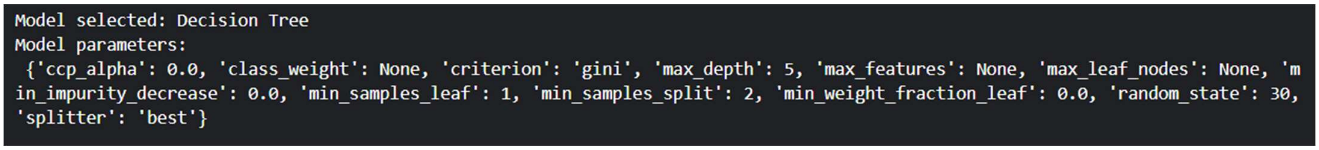
```
Most important feature for LGBM:
Header_Length
```



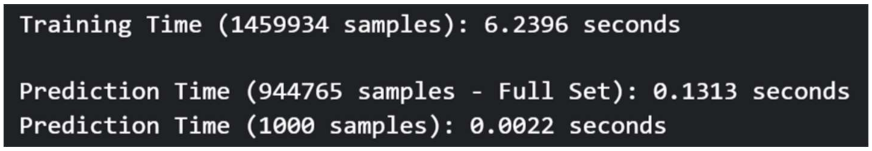


```
-rw-r--r-- 1 root root 328K Jun 11 10:08 lgbm.joblib
```

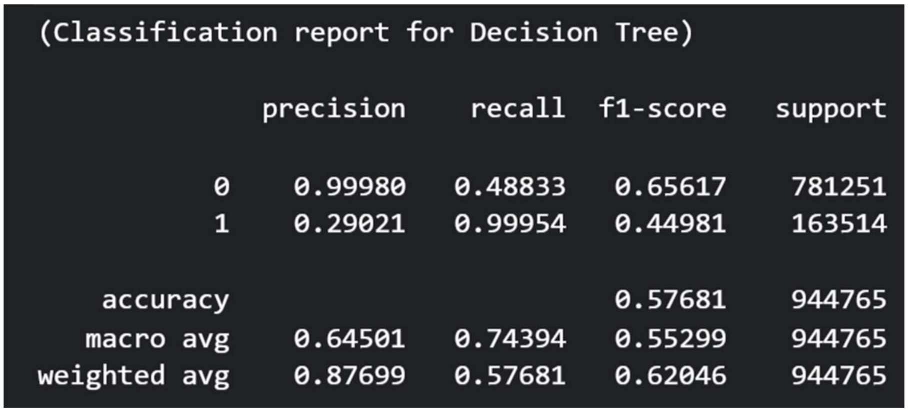
(Decision Tree)



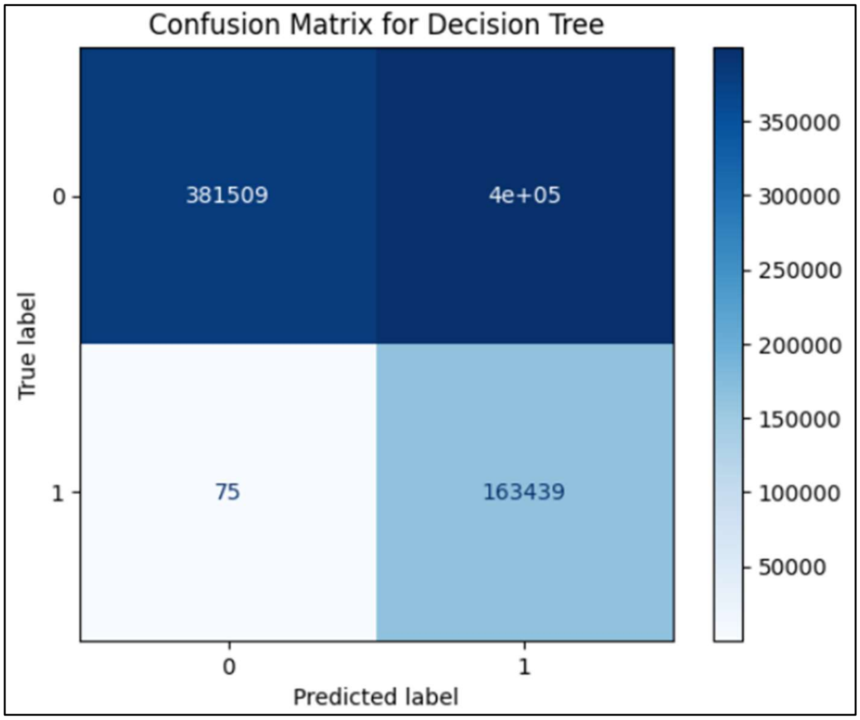
```
Model selected: Decision Tree
Model parameters:
 {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 5, 'max_features': None, 'max_leaf_nodes': None, 'm
in_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'random_state': 30,
'splitter': 'best'}
```

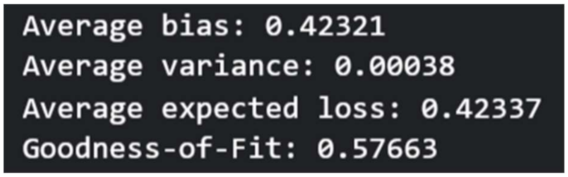


```
Training Time (1459934 samples): 6.2396 seconds

Prediction Time (944765 samples - Full Set): 0.1313 seconds
Prediction Time (1000 samples): 0.0022 seconds
```

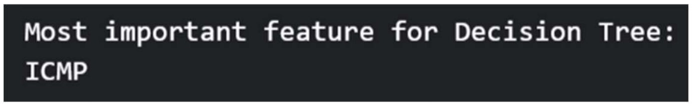


```
 (Classification report for Decision Tree)

              precision    recall  f1-score   support

           0    0.99980   0.48833   0.65617    781251
           1    0.29021   0.99954   0.44981    163514

    accuracy                        0.57681    944765
   macro avg    0.64501   0.74394   0.55299    944765
weighted avg    0.87699   0.57681   0.62046    944765
```

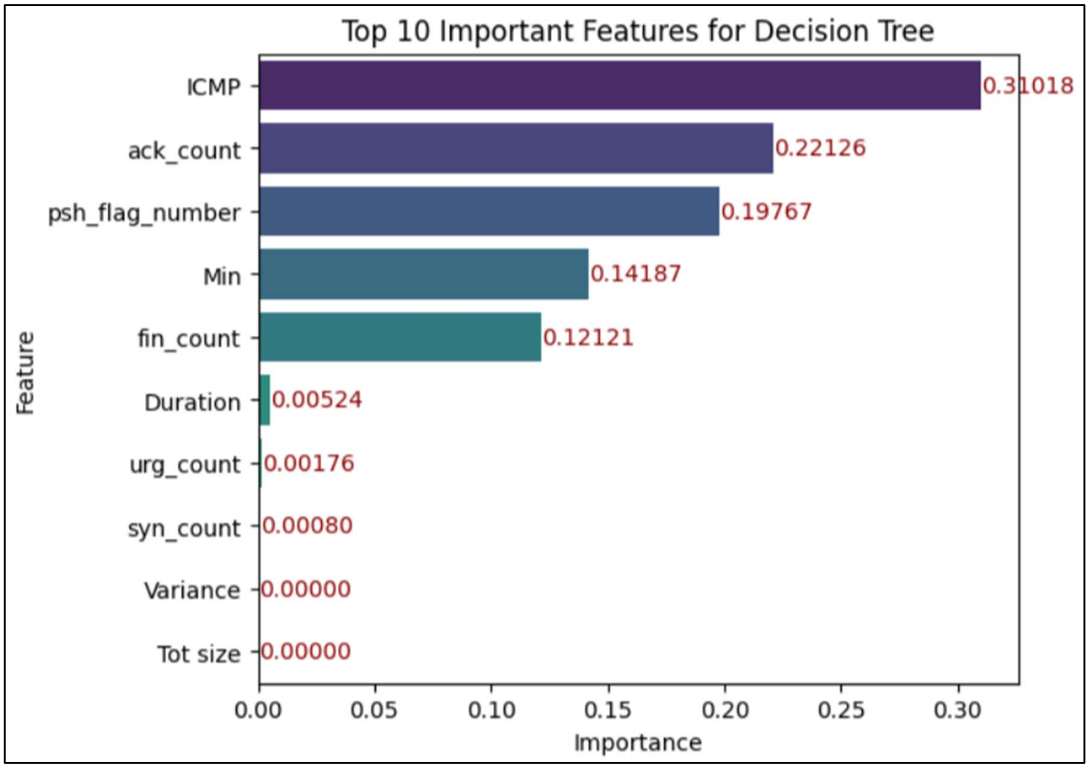


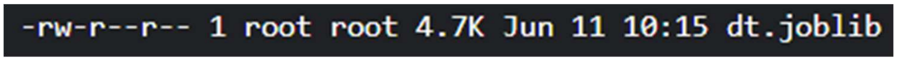


```
Average bias: 0.42321
Average variance: 0.00038
Average expected loss: 0.42337
Goodness-of-Fit: 0.57663
```



```
Most important feature for Decision Tree:
ICMP
```





```
-rw-r--r-- 1 root root 4.7K Jun 11 10:15 dt.joblib
```