Data Balancing - Decision Tree Algorithm

| Data | Negative Class F1-Score | Positive Class F1-Score | Accuracy | Time  (s) | Memory  (MiB) |
| --- | --- | --- | --- | --- | --- |
| Imbalanced, raw | 0.92 | 0.55 | 0.86 | 0.046875 | 209.79 |
| Undersampling | 0.86 | 0.54 | 0.79 | 0.015625 | 209.77 |
| Oversampling | 0.91 | 0.53 | 0.86 | 0.078125 | 209.77 |
| SMOTE | 0.92 | 0.55 | 0.86 | 0.109375 | 210.83 |

Feature Selection (w/ SMOTE)

Decision Tree Model

| **Data** | **Negative Class F1-Score** | **Positive Class F1-Score** | **Accuracy** | **Time**  **(s)** | **Memory**  **(MiB)** |
| --- | --- | --- | --- | --- | --- |
| Control | 0.91 | 0.57 | 0.85 | 0.9375 | 245.04 |
| Pearson Correlation | 0.91 | 0.59 | 0.86 | 0.078125 | 209.42 |
| Random Forest | 0.92 | 0.55 | 0.86 | 0.9375 | 239.09 |
| Recursive Feature Elimination | 0.81 | 0.34 | 0.71 | 0.46875 | 244.23 |

Random Forest Model

| **Data** | **Negative Class F1-Score** | **Positive Class F1-Score** | **Accuracy** | **Time**  **(s)** | **Memory**  **(MiB)** |
| --- | --- | --- | --- | --- | --- |
| Control | 0.93 | 0.64 | 0.89 | 1.26525 | 265.24 |
| Pearson Correlation | 0.93 | 0.66 | 0.88 | 1.53125 | 227.65 |
| Random Forest | 0.94 | 0.65 | 0.89 | 1.5 | 254.89 |
| Recursive Feature Elimination | 0.83 | 0.34 | 0.73 | 1.0625 | 285.75 |

Logistic Regression Model

| **Data** | **Negative Class F1-Score** | **Positive Class F1-Score** | **Accuracy** | **Time**  **(s)** | **Memory**  **(MiB)** |
| --- | --- | --- | --- | --- | --- |
| Control | 0.92 | 0.52 | 0.87 | 0.921875 | 253.05 |
| Pearson Correlation | 0.92 | 0.69 | 0.87 | 0.140625 | 210.72 |
| Random Forest | 0.92 | 0.53 | 0.86 | 0.578125 | 239.93 |
| Recursive Feature Elimination | 0.74 | 0.38 | 0.63 | 0.046875 | 244.79 |

Features Kept

| Rank | Pearson Correlation | Random Forest | Recursive Feature Elimination |
| --- | --- | --- | --- |
| 1 | PageValues | PageValues | Weekend\_True |
| 2 | ExitRates | ExitRates | Weekend\_False |
| 3 | ProductRelated | ProductRelated\_Duration | Month\_Sep |
| 4 | Month\_Nov | Administrative\_Duration | Month\_Oct |
| 5 | ProductRelated\_Duration | ProductRelated | Month\_Nov |
| 6 | BounceRates | BounceRates | Browser\_12 |
| 7 | Administrative | Month\_May | Browser\_3 |
| 8 | TrafficType\_2 | Administrative | TrafficType\_13 |
| 9 | VisitorType\_New\_Visitor | OperatingSystems\_3 | Month\_Jul |
| 10 | VisitorType\_Returning\_Visitor | VisitorType\_Returning\_Visitor | Month\_Feb |
| 11 | Informational | Month\_Dec | TrafficType\_15 |
| 12 | Administrative\_Duration | TrafficType\_1 | SpecialDay |
| 13 | TrafficType\_3 | Informational\_Duration | TrafficType\_3 |
| 14 | SpecialDay | Month\_Mar | ExitRates |
| 15 | Month\_May | Browser\_2 | BounceRates |
| 16 | OperatingSystems\_3 | OperatingSystems\_1 | TrafficType\_7 |
| 17 | Informational\_Duration | TrafficType\_3 | TrafficType\_1 |
| 18 | TrafficType\_13 | Weekend\_True | TrafficType\_18 |
| 19 | TrafficType\_1 | Operating\_systems\_2 | Month\_Aug |
| 20 | Month\_Mar | Weekend\_False | OperatingSystems\_3 |
| 21 | OperatingSystems\_2 | Browser\_1 | TrafficType\_19 |
| 22 | TrafficType\_8 | Region\_1 | TrafficType\_20 |
| 23 | Month\_Feb | Region\_3 | TrafficType\_6 |
| 24 | TrafficType\_20 | TrafficType\_2 | TrafficType\_16 |

Data Scaling Results (w/ SMOTE)

Control Data Accuracy Score (not scaled)

| **Model** | **Control**  **(not filtered)** | **Pearson Correlation** | **Random Forest** | **Recursive Feature Elimination** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.86 | 0.85 | 0.86 | 0.69 |
| Random Forest | 0.89 | 0.88 | 0.89 | 0.73 |
| Logistic Regression | 0.87 | 0.88 | 0.88 | 0.64 |

Normalized Data Accuracy Scores

| **Model** | **Control**  **(not filtered)** | **Pearson Correlation** | **Random Forest** | **Recursive Feature Elimination** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.86 | 0.84 | 0.86 | 0.69 |
| Random Forest | 0.89 | 0.88 | 0.89 | 0.72 |
| Logistic Regression | 0.87 | 0.86 | 0.86 | 0.63 |

Standardized Data Accuracy Scores

| **Model** | **Control**  **(not filtered)** | **Pearson Correlation** | **Random Forest** | **Recursive Feature Elimination** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.86 | 0.84 | 0.86 | 0.69 |
| Random Forest | 0.90 | 0.88 | 0.89 | 0.72 |
| Logistic Regression | 0.88 | 0.87 | 0.87 | 0.63 |

Data Cross validation (w/ SMOTE and Standardization)

No cross-validation

| **Model** | **Control**  **(not filtered)** | **Pearson Correlation** | **Random Forest** | **Recursive Feature Elimination** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.85 | 0.86 | 0.86 | 0.69 |
| Random Forest | 0.90 | 0.89 | 0.89 | 0.74 |
| Logistic Regression | 0.89 | 0.89 | 0.87 | 0.64 |

K-fold cross-validation

| **Model** | **Control**  **(not filtered)** | **Pearson Correlation** | **Random Forest** | **Recursive Feature Elimination** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.86 | 0.90 | 0.90 | 0.82 |
| Random Forest | 0.90 | 0.90 | 0.90 | 0.82 |
| Logistic Regression | 0.88 | 0.88 | 0.88 | 0.84 |

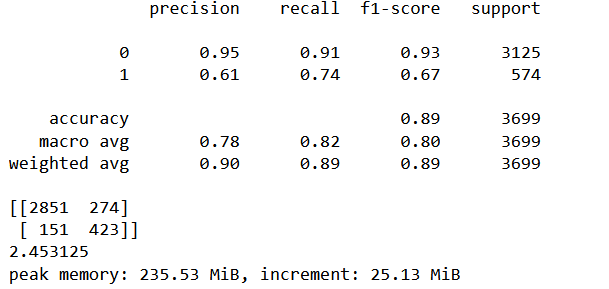
Time-series cross-validation

| **Model** | **Control**  **(not filtered)** | **Pearson Correlation** | **Random Forest** | **Recursive Feature Elimination** |
| --- | --- | --- | --- | --- |
| Decision Tree | 0.86 | 0.87 | 0.86 | 0.76 |
| Random Forest | 0.90 | 0.90 | 0.90 | 0.82 |
| Logistic Regression | 0.88 | 0.88 | 0.88 | 0.84 |

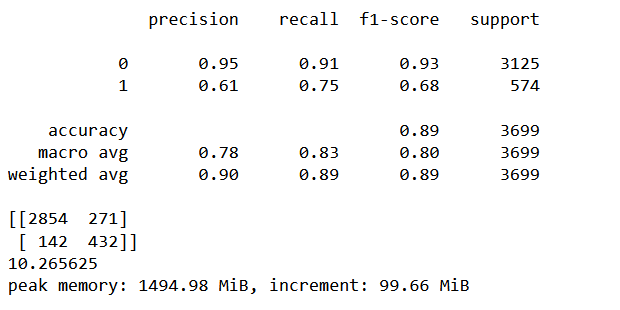
Model Tuning

Changing number of estimators

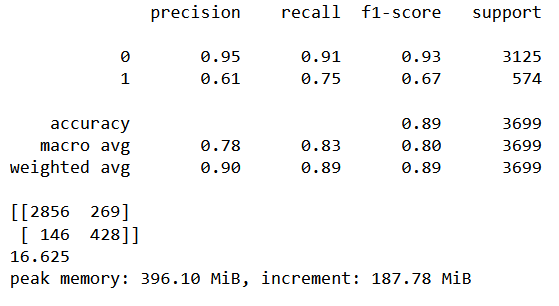
100 (default)



500

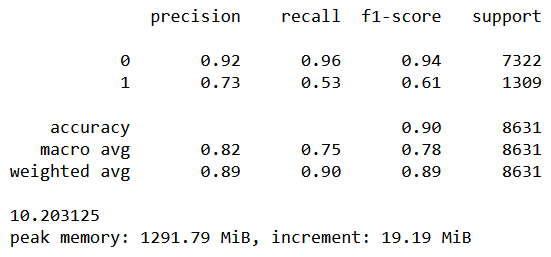


1000

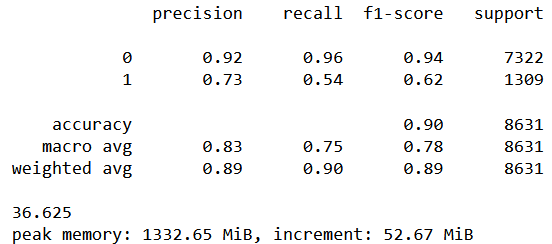


KFold on Random Forest

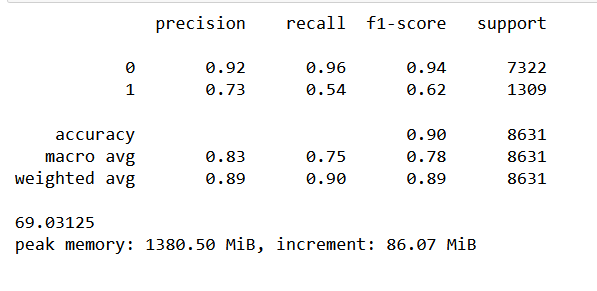
100



500

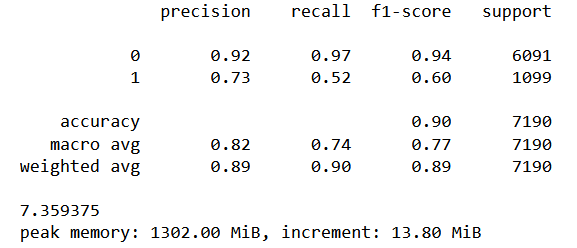


1000

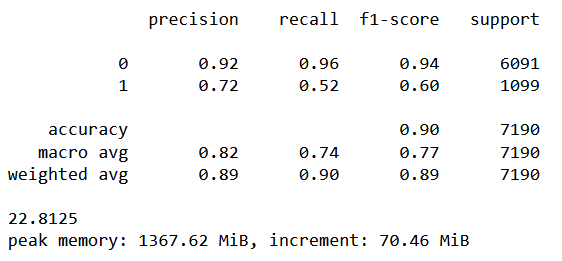


Time-Series on RF

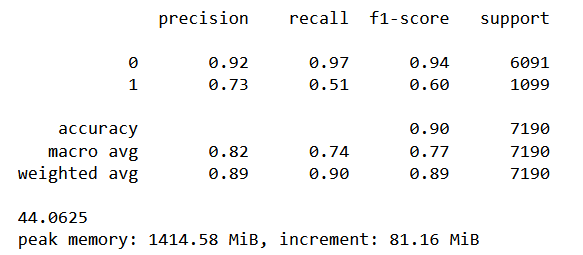
100



500



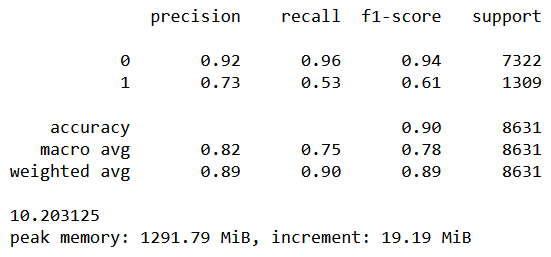
1000



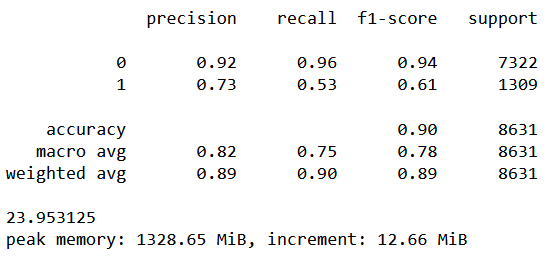
Changing number of folds

K-fold on RF

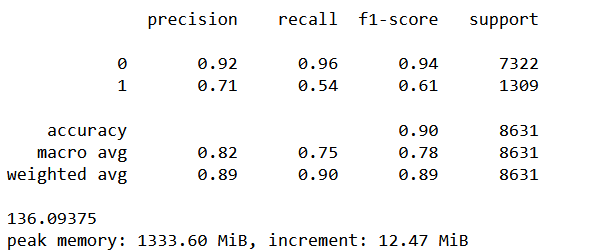
5 folds



10 folds

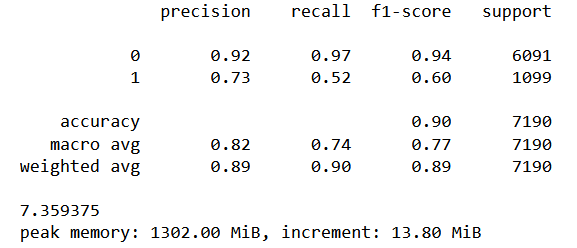


50 folds

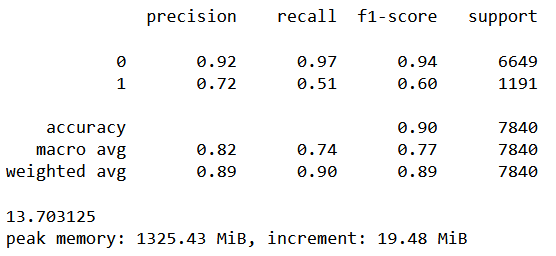


Time-series on RF

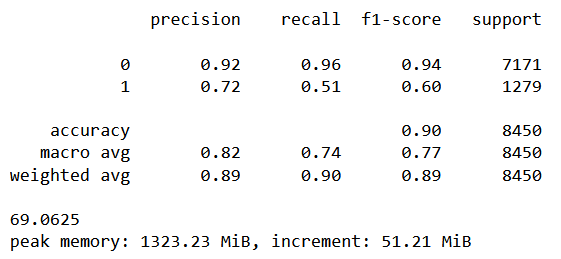
5 folds



10 folds

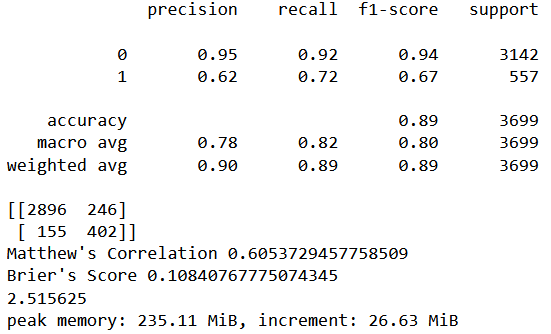


50 folds

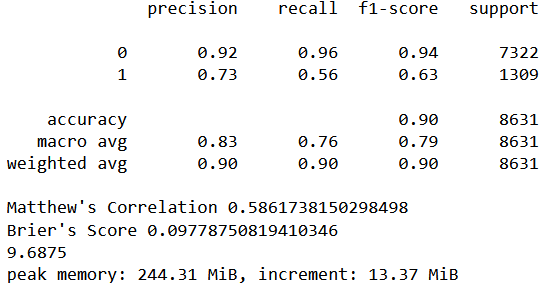


Additional Metrics

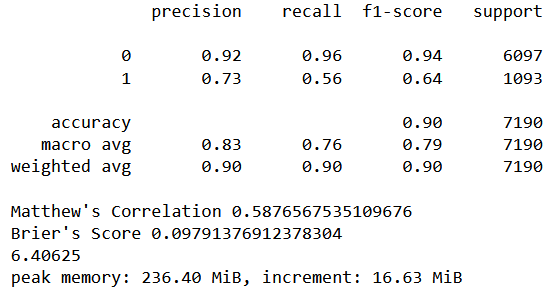
P corr



Kfold on RF



Time-series on RF



**Variance testing**

**Pearson corr**

negative\_precision = [0.95, 0.95, 0.95, 0.95, 0.95]

negative\_recall = [0.92, 0.92, 0.92, 0.92, 0.92]

negative\_f1\_score = [0.94, 0.94, 0.94, 0.94, 0.93]

positive\_precision = [0.62, 0.62, 0.62, 0.63, 0.62]

positive\_recall = [0.72, 0.72, 0.72, 0.71, 0.71]

positive\_f1\_score = [0.67, 0.67, 0.67, 0.67, 0.66]

Negative Precision Variance: 0.0

Negative Recall Variance: 1.232595164407831e-32

Negative F1-score Variance: 1.5999999999999674e-05

Positive Precision Variance: 1.600000000000003e-05

Positive Recall Variance: 2.400000000000004e-05

Positive F1-score Variance: 1.600000000000003e-05

Accuracy Variance: 0.0

Matthew's Correlation Variance: 4.4388481056110005e-06

Brier's Score Variance: 5.379094223386346e-07

Time Variance: 0.00845703125

accuracy = [0.89, 0.89, 0.89, 0.89, 0.89]

matthews\_correlation = [0.6033965770959435, 0.6047371915042924, 0.605388806151251, 0.6041407710751439, 0.5994121189853286]

briers\_score = [0.10867802108678021, 0.10813733441470667, 0.10813733441470667, 0.10759664774263314, 0.10975939443092728]

time = [2.4375, 2.5625, 2.640625, 2.40625, 2.421875]

**Kfold on RF**

negative\_precision = [0.92, 0.92, 0.92, 0.92, 0.92]

negative\_recall = [0.97, 0.97, 0.97, 0.96, 0.96]

negative\_f1\_score = [0.94, 0.94, 0.94, 0.94, 0.94]

positive\_precision = [0.74, 0.74, 0.74, 0.74, 0.74]

positive\_recall = [0.56, 0.55, 0.55, 0.55, 0.56]

positive\_f1\_score = [0.64, 0.63, 0.63, 0.63, 0.64]

accuracy = [0.90, 0.90, 0.90, 0.90, 0.90]

matthews\_correlation = [0.5897665120780788, 0.5833569514943414, 0.584062157761428, 0.5860958315439827, 0.589897709880703]

briers\_score = [0.09674443320958588, 0.09801878823216033, 0.09778744109806689, 0.09743961524448788, 0.09674476868976872]

time = [9.453125, 9.578125, 9.546875, 9.96875, 9.9375]

Negative Precision Variance: 1.232595164407831e-32

Negative Recall Variance: 2.400000000000004e-05

Negative F1-score Variance: 1.232595164407831e-32

Positive Precision Variance: 0.0

Positive Recall Variance: 2.400000000000004e-05

Positive F1-score Variance: 2.400000000000004e-05

Accuracy Variance: 0.0

Matthew's Correlation Variance: 7.621492401814735e-06

Brier's Score Variance: 2.7592693214958056e-07

Time Variance: 0.04556640625

**Time-series on RF**

negative\_precision = [0.92, 0.92, 0.93, 0.92, 0.92]

negative\_recall = [0.97, 0.97, 0.96, 0.96, 0.96]

negative\_f1\_score = [0.94, 0.94, 0.94, 0.94, 0.94]

positive\_precision = [0.74, 0.74, 0.73, 0.74, 0.73]

positive\_recall = [0.56, 0.56, 0.55, 0.55, 0.55]

positive\_f1\_score = [0.64, 0.64, 0.63, 0.63, 0.63]

accuracy = [0.90, 0.90, 0.90, 0.90, 0.90]

matthews\_correlation = [0.5897683347185898, 0.5912042015863317, 0.5893192697464371, 0.5847338869126286, 0.5794291285215211]

briers\_score = [0.0968011126564673, 0.09638386648122392, 0.09735744089012517, 0.09791376912378304, 0.09916550764951321]

time = [6.71875, 6.546875, 6.390625, 6.53125, 6.515625]

Negative Precision Variance: 1.600000000000003e-05

Negative Recall Variance: 2.4000000000000045e-05

Negative F1-score Variance: 1.232595164407831e-32

Positive Precision Variance: 2.400000000000004e-05

Positive Recall Variance: 2.4000000000000048e-05

Positive F1-score Variance: 2.4000000000000048e-05

Accuracy Variance: 0.0

Matthew's Correlation Variance: 1.8622383959054213e-05

Brier's Score Variance: 9.393358493193886e-07

Time Variance: 0.010996093749999996