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 |  |
| Undersampling | 0.86 | 0.54 | 0.79 | 0.046875 |  |
| Oversampling | 0.91 | 0.53 | 0.86 | 0.09375 |  |
| SMOTE | 0.92 | 0.55 | 0.86 | 0.109375 |  |

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 |