

#### Online Resource 4:

Confusion matrix in 5 machine learning and logistic regression models for training dataset of original data

|               | RF                     |                        | LR                     |                        | SVM                    |                        |
|---------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
|               | Prediction<br>Negative | Prediction<br>Positive | Prediction<br>Negative | Prediction<br>Positive | Prediction<br>Negative | Prediction<br>Positive |
| True Negative | 57,683                 | 10                     | 57,693                 | 0                      | 57,693                 | 0                      |
| True Positive | 84                     | 10                     | 94                     | 0                      | 94                     | 0                      |

  

|               | GB                     |                        | KNN                    |                        | BRF                    |                        |
|---------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
|               | Prediction<br>Negative | Prediction<br>Positive | Prediction<br>Negative | Prediction<br>Positive | Prediction<br>Negative | Prediction<br>Positive |
| True Negative | 57,679                 | 14                     | 57,691                 | 2                      | 50,953                 | 6,740                  |
| True Positive | 77                     | 17                     | 91                     | 3                      | 5                      | 89                     |

RF: Random Forest

LR: Logistic Regression

SVM: Support Vector Machine

GB: Gradient Boosting

KNN: k-Nearest Neighbor

BRF: Balanced Random Forest