**Supplementary 6: 100 Iteration average results of individual machine learning and regression models under varying model specifications**

1. Testing a variety of linear and non-linear machine learning and regression models

|  |  |  |
| --- | --- | --- |
| **EC** | Continuous |  |
| **Variables** | All |  |
| **Train-Test** | 70-30 |  |
| **Row Labels** | **Average of R2\_Train** | **Average of R2\_Test** |
| Bagging | 0.883958695 | 0.390130017 |
| RandomForest | 0.886629656 | 0.38153062 |
| Quadratic\_correlated | 0.493063218 | 0.075613038 |
| Stepwise | 0.667250327 | -0.4812077 |
| OLS | 0.684766378 | -0.881249685 |
| ANN | 0.899084686 | -1.589536008 |
| Quadratic\_all | 0.883570262 | -66.78533693 |
| Log-Linear | 0.480977049 | -9.34509E+19 |

|  |  |  |
| --- | --- | --- |
| **EC** | Continuous |  |
| **Variables** | All |  |
| **Train-Test** | 70-30 |  |
| Random Forest maxnodes = 6 (reduced depth to avoid overfitting in small dataset) | | |
| **Row Labels** | **Average of R2\_Train** | **Average of R2\_Test** |
| Bagging | 0.740572046 | 0.40654547 |
| RandomForest | 0.761469885 | 0.402998112 |
| Quadratic\_correlated | 0.485572096 | 0.075818915 |
| Stepwise | 0.65584359 | -0.519069914 |
| OLS | 0.674089184 | -0.821922951 |
| ANN | 0.889559236 | -4.73293199 |
| Quadratic\_all | 0.875804963 | -69.9171603 |
| Log-Linear | 0.465514807 | -2.5877E+14 |

|  |  |  |
| --- | --- | --- |
| **EC** | Continuous |  |
| **Variables** | Pure bands and highly correlated | |
| **Train-Test** | 80-20 |  |
| **Parameter adjustments for small data (maxnodes = 6)** | |  |
| **Row Labels** | **Average of R2\_Train** | **Average of R2\_Test** |
| RandomForest | 0.729574484 | 0.444587276 |
| Bagging | 0.745901662 | 0.432524444 |
| Lasso | 0.385876678 | 0.301727259 |
| Ridge | 0.336485852 | 0.273195735 |
| OLS | 0.519148023 | 0.270856197 |
| Quadratic\_correlated | 0.546587506 | 0.247986491 |
| Stepwise | 0.504797542 | 0.235174696 |
| Quadratic\_all | 0.606914335 | -0.234295277 |
| Log-Linear | -0.113285846 | -6.731768748 |
| ANN | -6.586192881 | -179.2320997 |

|  |  |  |
| --- | --- | --- |
| **EC** | Continuous |  |
| **Variables** | PCA (3) |  |
| **Train-Test** | 80-20 (random) |  |
| **Row Labels** | **Average of R2\_Train** | **Average of R2\_Test** |
| Bagging | 0.454671094 | 0.191897586 |
| Ridge | 0.271008882 | 0.167371934 |
| Quantile Regression | 0.294190471 | 0.156754325 |
| OLS | 0.339315496 | 0.151301889 |
| Lasso | 0.336095118 | 0.145802083 |
| Log-Linear | 0.226656671 | 0.145261048 |
| RandomForest | 0.362814088 | 0.140630829 |
| Stepwise | 0.332848696 | 0.127901676 |
| ANN | -5.770963851 | -30.57203946 |
|  |  |  |

1. Testing variable selection with random sampling

|  |  |  |
| --- | --- | --- |
| **EC** | Binary |  |
| **Variables** | All |  |
| **Train-Test** | 80-20 (random) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.992857143 | 0.847272727 |
| ANN | 0.954166667 | 0.787272727 |
| Logistic | 0.956785714 | 0.762727273 |

|  |  |  |
| --- | --- | --- |
| **EC** | Binary |  |
| **Variables** | Lasso |  |
| **Train-Test** | 80-20 (random) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Logistic | 0.889166667 | 0.848181818 |
| Random Forest | 0.992857143 | 0.841363636 |
| ANN | 0.960595238 | 0.789090909 |

|  |  |  |  |
| --- | --- | --- | --- |
| EC | Binary |  | |
| **Variables** | Pure bands and highly correlated | |
| **Train-Test** | 80-20 (random) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.992857143 | 0.838181818 |
| Logistic | 0.901071429 | 0.833181818 |
| ANN | 0.96 | 0.775909091 |

|  |  |  |
| --- | --- | --- |
| **EC** | Categorical |  |
| **Variables** | All |  |
| **Train-Test** | 80-20 (random) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.995364489 | 0.76482743 |
| ANN | 0.899229683 | 0.671691396 |

|  |  |  |
| --- | --- | --- |
| **EC** | Categorical |  |
| **Variables** | Pure bands and highly correlated | |
| **Train-Test** | 80-20 (random) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.995541963 | 0.74184358 |
| ANN | 0.886974731 | 0.667993751 |

|  |  |  |
| --- | --- | --- |
| **EC** | Categorical |  |
| **Variables** | Lasso variables |  |
| **Train-Test** | 80-20 (random) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.947499565 | 0.730831604 |
| ANN | 0.780972881 | 0.709749143 |

1. Testing Variable selection with stratified sampling

|  |  |  |
| --- | --- | --- |
| **EC** | Binary |  |
| **Variables** | All |  |
| **Train-Test** | 80-20 (stratified) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.992352941 | 0.844285714 |
| ANN | 0.954705882 | 0.798571429 |
| Logistic | 0.947294118 | 0.754761905 |

|  |  |  |
| --- | --- | --- |
| **EC** | Binary |  |
| **Variables** | Lasso |  |
| **Train-Test** | 80-20 (stratified) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Logistic | 0.890470588 | 0.843809524 |
| Random Forest | 0.992352941 | 0.834761905 |
| ANN | 0.956352941 | 0.810952381 |

|  |  |  |  |
| --- | --- | --- | --- |
| **EC** | Binary |  | |
| **Variables** | Pure bands and highly correlated | |
| **Train-Test** | 80-20 (stratified) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Logistic | 0.900941176 | 0.830952381 |
| Random Forest | 0.992352941 | 0.826190476 |
| ANN | 0.960117647 | 0.798095238 |

|  |  |  |
| --- | --- | --- |
| **EC** | Categorical |  |
| **Variables** | All |  |
| **Train-Test** | 80-20 (stratified) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.99173913 | 0.78800641 |
| ANN | 0.900927668 | 0.686840798 |

|  |  |  |
| --- | --- | --- |
| **EC** | Categorical |  |
| **Variables** | Pure bands and highly correlated | |
| **Train-Test** | 80-20 (stratified) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.992463768 | 0.76102381 |
| ANN | 0.881891211 | 0.710176005 |

|  |  |  |
| --- | --- | --- |
| **EC** | Categorical |  |
| **Variables** | Lasso variables |  |
| **Train-Test** | 80-20 (stratified) |  |
| **Row Labels** | **Average of Accuracy\_Train** | **Average of Accuracy\_Test** |
| Random Forest | 0.931594203 | 0.752554487 |
| ANN | 0.781774522 | 0.707827381 |