**Assignment 2**

-Vaishnavi Ayyadurai, wb9452

Firstly, data is pre-processed to remove redundancy and to improve the

performance of the model:

1. Eliminated attributes which has Correlation >0.9.
2. Checked for NullVariance and corresponding attributes are removed.
3. Factors of length 1 are removed.
4. There are 18 attributes that are finally left out of 21 after pre-processing.
5. 40% are data are stratified sampled and are used as training dataset and

rest 60% data is used as test dataset. This proportion of dataset is chosen

after testing various proportions of dataset(like 75%:25%),(60%:40%), out

of which (40-training): (60-testing) executed in efficient timeframe during

training.

**Qn1. With 10-fold cross validation:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Parameters** | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** |
| **Neural Networks** | 0.8873 | 0.9999 | 0.0003 | 0.8873 |
| **SVM** | 0.9045 | 0.9797 | 0.3125 | 0.9182 |
| **Decision Tree** | 0.9098 | 0.9746 | 0.3997 | **0.9275** |
| **Naïve Bayes** | 0.8988 | 0.9711 | 0.254 | 0.8774 |

**Other Detailed parameters:**

**Neural Network:**

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9999543962 0.0003591954 0.8873735330 0.5000000000

Precision Recall F1 Prevalence

0.8873735330 0.9999543962 0.9403061881 0.8873421819

Detection Rate Detection Prevalence Balanced Accuracy

0.8873017158 0.9999190677 0.5001567958

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

0.8873421819 0.0005562166 0.8833342832 0.8912587374 0.8873421819

AccuracyPValue McnemarPValue

0.5050493951 0.0000000000

With PCA

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

9.041356e-01 5.109298e-01 9.003971e-01 9.077788e-01 8.873422e-01

AccuracyPValue McnemarPValue

7.293774e-18 1.084523e-02

**SVM:**

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9797063 0.3125000 0.9181946 0.6615970

Precision Recall F1 Prevalence

0.9181946 0.9797063 0.9479537 0.8873422

Detection Rate Detection Prevalence Balanced Accuracy

0.8693347 0.9467870 0.6461032

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

9.045403e-01 3.796529e-01 9.008087e-01 9.081765e-01 8.873422e-01

AccuracyPValue McnemarPValue

1.151494e-18 1.121929e-200

**Decision tree:**

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9746443 0.3997845 0.9274834 0.6668664

Precision Recall F1 Prevalence

0.9274834 0.9746443 0.9504792 0.8873422

Detection Rate Detection Prevalence Balanced Accuracy

0.8648430 0.9324620 0.6872144

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

9.098818e-01 4.537581e-01 9.062430e-01 9.134239e-01 8.873422e-01

AccuracyPValue McnemarPValue

3.625990e-31 3.332842e-123

**Qn2:**

**“Precision”** is considered to be of higher importance in this case, since precision is average probability of relevant retrieval which makes sure that there no false positive results, while recall is the average probability of the **complete** retrieval that makes sure there are no false Negative result. For example, I cannot have a lot of applications falsely tested for opening a term deposit, if I know that all of the ones who are willing to open a term deposit are tested. In this case, all the relevant hits are identified so that there are no false positive results. Precision can test and make sure that opening a term deposit for an unsanctioned or otherwise undesirable party can incur large fine or even endanger banking license and thus will prevent

them. Also, Staffs required to review high volumes of false positive results may

inadvertently **overlook** a **true crucial positive hit.**

Generally,

F1=(2\*recall\*precision)/(recall+precision)

Which gives equal importance to Recall & Precision.

Since, Precision is preferred for this case, Beta =0.5(0<beta<1)

So Weighted-F becomes:

**F0.5=(1.25\*recall \*precision)/(0.25\*precision)+recall**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Parameters** | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** | **F1** | **F2** |
| **Neural Networks** | 0.8873 | 0.9999 | 0.0003 | 0.8873 | 0.9403 | 0.907744 |
| **SVM** | 0.9045 | 0.9797 | 0.3125 | 0.9182 | 0.9479 | 0.929874 |
| **Decision Tree** | 0.9098 | 0.9746 | 0.3997 | **0.9275** | 0.9504 | 0.936552 |
| **Naïve Bayes** | 0.8988 | 0.9711 | 0.254 | 0.8774 | 0.9401 | 0.894665 |

* Based on F1 value (which considers recall and precision to be of equal importance), the leader by marginal difference beating SVM is “**Decision Tree”.**
* Also, Weighted F-value (F2 value) is more for Decision tree compared to

any other models.

Hence, recommended model with 10-fold cross validation would be

**“Decision tree”.**

**Qn3. Dimensionality Reduction With PCA:**

Preprocessed data is tested with PCA & Decision tree is a winner.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Parameters** | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** | **F1** | **F2** |
| **Neural Networks** | 0.8873 | 0.9999 | 0.0003 | 0.8873 | 0.9474 | 0.907744 |
| **SVM** | 0.899 | 0.9839 | 0.2298 | 0.9096 | 0.9453 | 0.923549 |
| **Decision Tree** | 0.9027 | 0.9693 | 0.3785 | 0.9247 | 0.9464 | 0.933289 |
| **Naïve Bayes** | 0.888 | 0.9611 | 0.194 | 0.8564 | 0.9371 | 0.875474 |

**Other parameters:**

**Neural Networks:**

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9522528 0.5434626 0.9426237 0.5910156

Precision Recall F1 Prevalence

0.9426237 0.9522528 0.9474138 0.8873422

Detection Rate Detection Prevalence Balanced Accuracy

0.8449741 0.8964066 0.7478577

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

9.061994e-01 5.137599e-01 9.024962e-01 9.098068e-01 8.873422e-01

AccuracyPValue McnemarPValue

3.655749e-22 3.625311e-06

**SVM**:

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9839931 0.2298851 0.9096160 0.6458123

Precision Recall F1 Prevalence

0.9096160 0.9839931 0.9453438 0.8873422

Detection Rate Detection Prevalence Balanced Accuracy

0.8731386 0.9598980 0.6069391

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

8.990369e-01 2.975219e-01 8.952133e-01 9.027664e-01 8.873422e-01

AccuracyPValue McnemarPValue

1.889619e-09 7.260757e-282

**Decision Tree:**

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9693086 0.3785920 0.9247335 0.6103069

Precision Recall F1 Prevalence

0.9247335 0.9693086 0.9464966 0.8873422

Detection Rate Detection Prevalence Balanced Accuracy

0.8601084 0.9301149 0.6739503

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

9.027598e-01 4.170136e-01 8.989980e-01 9.064266e-01 8.873422e-01

AccuracyPValue McnemarPValue

2.749991e-15 6.286223e-103

**Comparing Without and With PCA**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Parameters** | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** | **F1** | **F2** |
| **Neural Networks** | 0.8873 | 0.9999 | 0.0003 | 0.8873 | 0.9403 | 0.9077 |
| **Neural Networks(With PCA)** | 0.8873 | 0.9999 | 0.0003 | 0.8873 | 0.9474 | 0.9077 |
| **SVM** | 0.9045 | 0.9797 | 0.3125 | 0.9182 | 0.9479 | 0.9299 |
| **SVM(With PCA)** | 0.899 | 0.9839 | 0.2298 | 0.9096 | 0.9453 | 0.9235 |
| **Decision Tree** | 0.9098 | 0.9746 | 0.3997 | **0.9275** | 0.9504 | 0.9366 |
| **Decision Tree(With PCA)** | 0.9027 | 0.9693 | 0.3785 | 0.9247 | 0.9464 | 0.9333 |
| **Naïve Bayes** | 0.8988 | 0.9711 | 0.254 | 0.8774 | 0.9401 | 0.8947 |
| **Naïve Bayes(With PCA)** | 0.888 | 0.9611 | 0.194 | 0.8564 | 0.9371 | 0.8755 |

**Inference:**

Neural networks have no impact with and without PCA, while the other three model’s *accuracy, precision, specificity, F1 and Measured F values* have been marginally reduced with PCA and takes more time for execution.

Even with PCA, Decision Tree has higher F1 and F2 values with more accuracy and precision compared to other models.

To culminate, Decision tree is the suggested model for this case study.