Homework 1

2b)

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Deceased Patients | Alive Patients | Function to complete |
| Event Count  1. Average Event Count  2. Max Event Count  3. Min Event Count | 982.014  8635  1 | 498.118  12627  1 | event\_count\_metrics |
| Encounter Count  1. Average Enc. Count  2. Max Enc. Count  3. Min Enc. Count | 23.038  203  1 | 15.452  391  1 | encounter\_count\_metrics |
| Record Length  1. Average Rec. Length  2. Max Rec. Length  3. Min Rec. Length | 127.532  1972  0 | 159.2  2914  0 | record\_length\_metrics |

4.1b)

Results on training set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | AUC | Precision | Recall | F1 |
| Logistic Regression | 0.95454 | 0.94540 | 0.98692 | 0.89880 | 0.94080 |
| SVM | 0.994019 | 0.994511 | 0.98820 | 0.99702 | 0.99259 |
| Decision Tree | 0.77631 | 0.74759 | 0.79215 | 0.60119 | 0.68358 |

4.1c)

Results on test set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | AUC | Precision | Recall | F1 |
| Logistic Regression | 0.73809 | 0.7375 | 0.68041 | 0.73333 | 0.70588 |
| SVM | 0.73809 | 0.73888 | 0.67676 | 0.74444 | 0.70899 |
| Decision Tree | 0.67142 | 0.65694 | 0.63291 | 0.55555 | 0.59171 |

4.1d)

From the results, we can see that our train accuracy is significantly greater than the test accuracy. This means that our model has high variance. Strategies to combat this include: simplify the model by selecting one with fewer parameters e.g. dropping features in the training data, constraining the model complexity via regularization, etc. Another option is to gather more training data is this will help the model generalize better and thus improve test accuracy. And finally, another option is to try to reduce the noise in the training data by fixing data errors, removing outliers, etc.

4.2b)

|  |  |  |
| --- | --- | --- |
| CV Strategy | Accuracy | AUC |
| K-Fold | 0.72132 | 0.70757 |
| Randomized | 0.73571 | 0.71882 |

4.3b) The model’s tried in this stage of the homework include: Logistic Regression (chosen), Random Forests, SVM(with various kernels), KNN, AdaBoost, and Naïve Bayes. Logistic Regression was found to be the best performer after performing randomized search on all the models. Furthermore, since the dimensionality of the feature space is quite high but also sparse, I decided to employ a dimensionality technique. SVD (Singular Value Decomposition) made the most sense as it is known that SVD is the decomposition technique to use if you have a sparse dataset (and as a bonus it works well/efficiently with scipy sparse matrices). The overall model performs a fair bit better than the model from the previous part.