Mortality Post Hospital Discharge Results

1. Descriptive Statistics

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 Encounter Count 2. Max Encounter Count 3. Min Encounter Count	 23.038 203 1	 15.452 391 1	encounter_count _metric
Record Length 1. Average Record Length 2. Max Record Length 3. Min Record Length	 127.532 1972 0	 159.2 2914 0	record_length_metrics

Table 2: Descriptive statistics for alive and dead patients

2. Model Creation

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.954545	0.961389	0.898810	0.986928	0.940810
SVM	0.994019	0.993094	0.997024	0.988201	0.992593
Decision Tree	0.776316	0.780760	0.601190	0.792157	0.683587

Table 3: Model performance on training data

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.738095	0.734011	0.733333	0.680412	0.705882
SVM	0.738095	0.734780	0.744444	0.676768	0.708995
Decision Tree	0.680952	0.677450	0.677778	0.616161	0.645503

Table 4: Model performance on test data

Based on performance metrics on training and test data, something we can do to improve the test performance: 1. Provides more data samples to the training algorithm. The more data training algorithms learn from, the more cases it will be able to identify correctly. 2. Reduces noise or uses meaningful data and features that most heavily impact decision made by algorithm. 3.

Cross-validation. Cross-validation can be robust to the model by holding out data from the training process, which helps to improve performance on unseen observations. 4. Hyperparameter tuning. It help to guide the training process to find optimal model

3. Model Validation

CV Strategy	Accuracy	AUC
K-Fold	0.746386	0.739560
Randomized	0.735714	0.730846

Table 5: Cross Validation on Logistic Regression

I experimented the data with Random Forest and Gradient Boosting classifiers, and used cross validation(K-fold and RandomizedCV) to compute average accuracy and average AUC as in *table 6* and *table 7* below. I chose Random Forest because Random Forest combines multiple decision trees to make predictions, thus it's great for capturing complex interaction between features. I chose Gradient Boosting for its capability to build trees one at a time, where each new tree helps to correct errors made by previously trained trees, thus it's great for anomaly detection. Based on the result, teh Gradient Boosting appears to outperform Random Forest and pretty comparable to Logistic Regression, which we performed in the previous section.

CV Strategy	Accuracy	AUC
K-Fold	0.709310	0.704239
Randomized	0.701190	0.695461

Table 6: Cross Validation on Random Forest

CV Strategy	Accuracy	AUC
K-Fold	0.739165	0.734412
Randomized	0.742857	0.737387

Table 7: Cross Validation on Gradient Boosting