Springboard Data Science Course

Data Science Capstone Project 1

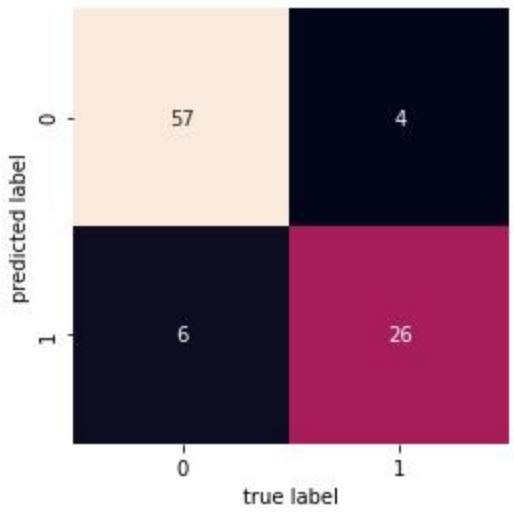
Orthopedic Biomechanical Features

Michelle Ide - 12/29/2020

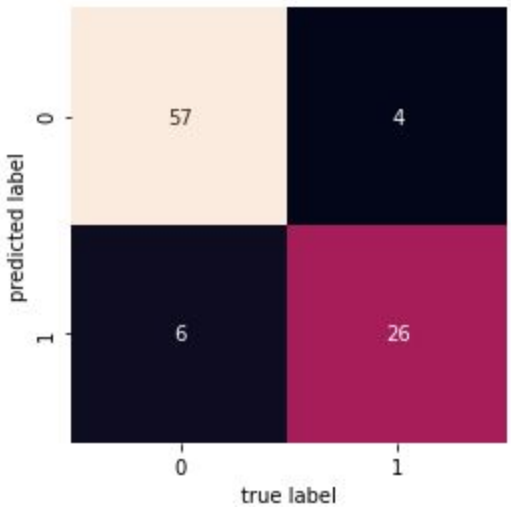
~~~ MACHINE LEARNING ~~~

Machine learning models were tested for classification of orthopedic features into Normal and Abnormal. Below is a description of the data specifics, models tested, and results. This data was collected from Kaggle, cleaned and analyzed. Hypotheses testing validated statistical significance existed between target values and correlation independence between features.

Data Descriptions

- 6 quantitative Features
- 1 binomial Target of Normal/Abnormal
- alpha of 0.5 for hypothesis test
- Supervised learning
- Unbalanced dataset addressed with resampling using ADASYN
- Cross-validation used to prevent over-fitting
- Accuracy will be measured with F1 score, ROC AUC plots, and confusion matrix.

Models Tested

5 Discriminative classification models include:

- ➢ Logistic Regression
- ➢ KNearest Neighbors
- ➢ Random Forest
- ➢ SVM
- ➢ Gradient Boosting

1 Generative model:

- ➢ Naive Bayes
~~~

Results

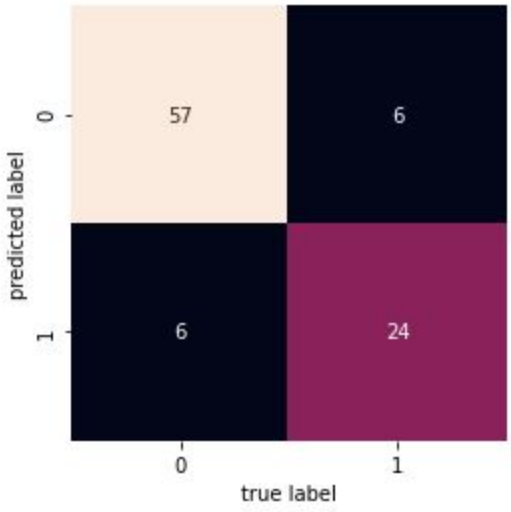
Logistic Regression



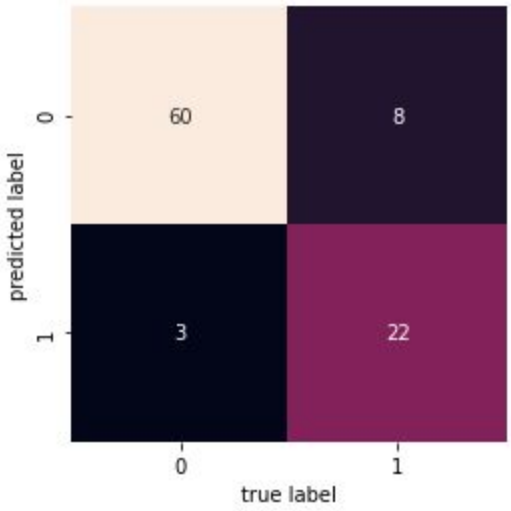
Gradient Boosting Classifier



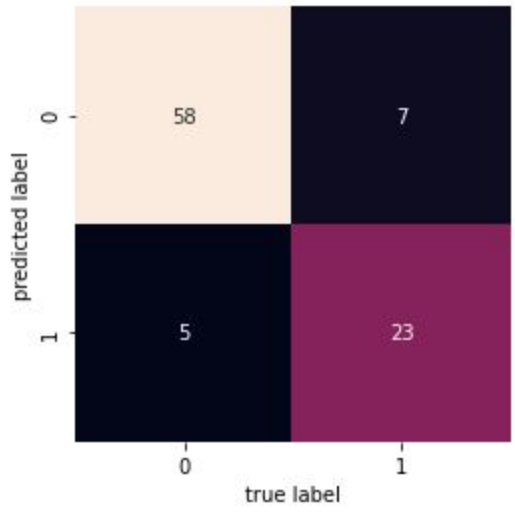
Support Vector Machine



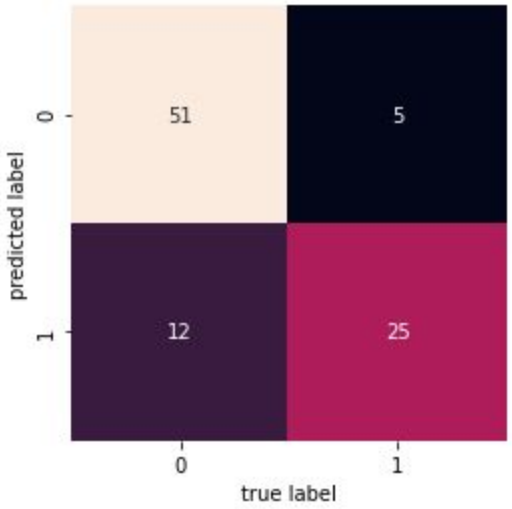
Random Forest Classifier



KNneighbors



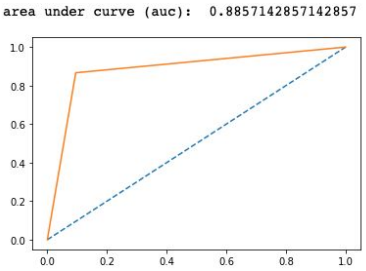
Gaussian Naive Bayes



LogisticRegression(random_state=42)

	precision	recall	f1-score	support
0	0.93	0.90	0.92	63
1	0.81	0.87	0.84	30
accuracy			0.89	93
macro avg	0.87	0.89	0.88	93
weighted avg	0.90	0.89	0.89	93

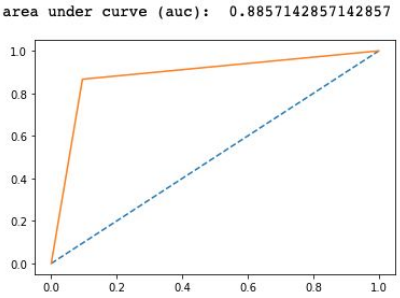
area under curve (auc): 0.8857142857142857



GradientBoostingClassifier(random_state=42)

	precision	recall	f1-score	support
0	0.93	0.90	0.92	63
1	0.81	0.87	0.84	30
accuracy			0.89	93
macro avg	0.87	0.89	0.88	93
weighted avg	0.90	0.89	0.89	93

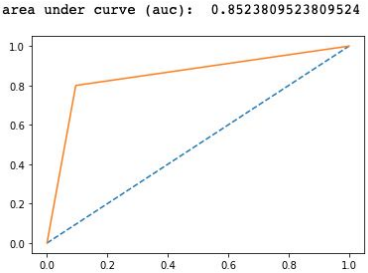
area under curve (auc): 0.8857142857142857



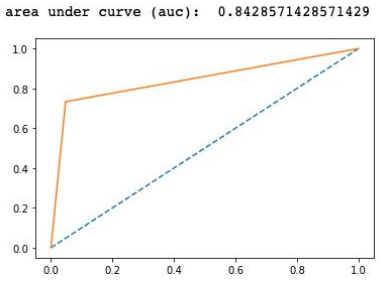
SVC(random_state=42)

	precision	recall	f1-score	support
0	0.90	0.90	0.90	63
1	0.80	0.80	0.80	30
accuracy			0.87	93
macro avg	0.85	0.85	0.85	93
weighted avg	0.87	0.87	0.87	93

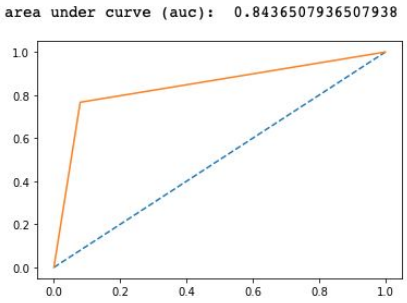
area under curve (auc): 0.8523809523809524



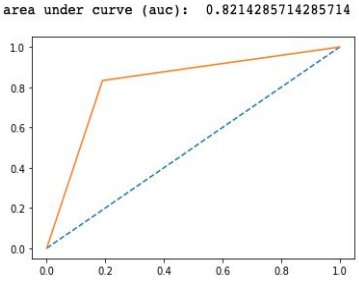
```
RandomForestClassifier(random_state=42)
              precision    recall  f1-score   support

           0       0.88      0.95      0.92        63
           1       0.88      0.73      0.80        30

    accuracy                           0.88        93
   macro avg       0.88      0.84      0.86        93
weighted avg       0.88      0.88      0.88        93

area under curve (auc):  0.8428571428571429
```




```
KNeighborsClassifier()
              precision    recall  f1-score   support

           0       0.89      0.92      0.91        63
           1       0.82      0.77      0.79        30

    accuracy                           0.87        93
   macro avg       0.86      0.84      0.85        93
weighted avg       0.87      0.87      0.87        93

area under curve (auc):  0.8436507936507938
```




```
GaussianNB()
              precision    recall  f1-score   support

           0       0.91      0.81      0.86        63
           1       0.68      0.83      0.75        30

    accuracy                           0.82        93
   macro avg       0.79      0.82      0.80        93
weighted avg       0.83      0.82      0.82        93

area under curve (auc):  0.8214285714285714
```



With parameter tuning, both Logistic Regression and Gradient Boosting algorithms performed best, and identically: ROC AUC of 88.56%, an f1 score for normal: 84% and abnormal: 92% with only 11% of test data mislabeled.

SVM, RandomForest also performed well with f1 scores at or above 80% for all and ROC AUC averaging 84%.

KNeighbors and GaussianNB broke down below 80% for normal results.