Springboard Data Science Course

Data Science Capstone Project 1

Orthopedic Biomechanical Features

Michelle Ide - 12/29/2020

~~~ SUMMARY REPORT ~~~

This project proposes to improve speed in identifying spondylolisthesis of the lumbar back using quantitative x-ray measurements. As a healthcare issue, the balance of bias prioritizes elimination of false negatives (false Normal) results. The following summarizes the data, analysis, method, modeling and results. Detailed explanations of data and statistical analysis can be found in the "Data and Analysis" report, additional testing and detailed results can be found in the Addendum.

DATA

This project uses a single dataset from Kaggle which include 6 quantitative features and a single string target. There were 0 empty values and one record was removed as an outlier using Tukey's algorithm to identify and visual inspection to validate outlier status for points >2x it's nearest neighbors. The target values were encoded as 0 for Abnormal and 1 for Normal. The Normal targets consisted of 100 data points, Abnormal 209. All 6 features displayed normal-ish distribution with the 'degree' feature skewed to the right.

Data Descriptions

- 6 quantitative Features
- 1 binomial Target of 209 Normal (0) and 100 Abnormal (1) labels
- alpha of 0.5 for hypothesis test
- Supervised learning
- Unbalanced dataset addressed with resampling using SMOTE
- K-Fold Cross-validation to prevent over-fitting
- Accuracy measured with F1 score, ROC AUC plots, and confusion matrix.

METHOD

With only 309 records, any records with empty feature values should be filled by estimation. Non-unique values are expected. Outliers identified using Tukey's
~~~

algorithm, inspected, and selected for removal if data point is > 2 times it's neighbor. Encoding of the string target after outlier removal is performed. An alpha of 0.05 was selected for this project. Hypothesis to be performed states: Features do not have a statistically significant impact describing the difference between Normal and Abnormal results. A limit on feature colinearity of 90% correlation between features was set for feature selection.

DATA ANALYSIS

Statistical analysis proved a significant statistical difference between Normal and Abnormal targets, nullifying the hypothesis test and proving viability of the project. No feature correlations over 90% were found. It should be noted 2 features were very close and could be considered for feature selection if additional tuning were performed, 'incidence' and 'tilt'. The 'degree' feature was most highly correlated with the target results which also contained the greatest degree of variance within the "abnormal" set, more the 2 times the variance of the "normal" portion. Cleaning and EDA concluded with the cleaned data features 'X' and encoded target 'Y' values stored in csv files for easy upload into new notebooks by importing "X.csv", "Y.csv", and "df.csv" for the completed encoded set.

MODELING

This project is a supervised classification problem with 2 target variables and unbalanced quantitative data. Therefore, six classification models including 5 generative and 1 discriminative were selected for testing.

Logistic Regression	Support Vector Machine	KNearest Neighbors
Gradient Boosting	Random Forest	Guassian Naive Bayes

To address the unbalanced data, a stratified train-test split was used with resampling methods, SMOTE and ADASYN tested. To prevent overfitting KFold cross-validation was included within the GridSearch during parameter tuning.

Models Tested

5 Discriminative

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

1 Generative

- Naive Bayes

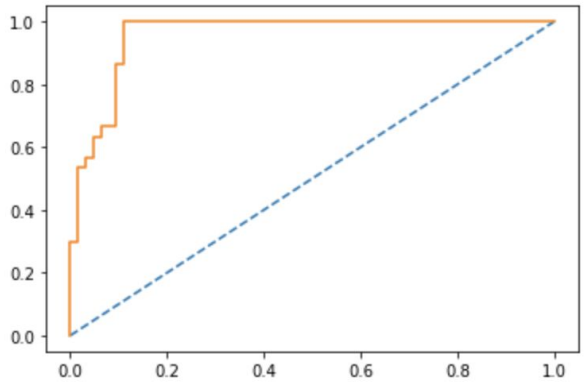
TESTING

Model accuracy was determined measuring ROC AUC, F1, and Recall scores with a confusion matrix plot. Most models performed similarly with the exception of Guassian Naive Bayes, possibly due to some feature correlations that existed within the data. If a NB model were preferred, additional feature selection could improve the accuracy, more testing is needed. Below is a summary of the data and test results. Note each model includes recall scores to determine best models based on bias selection.

RESULTS

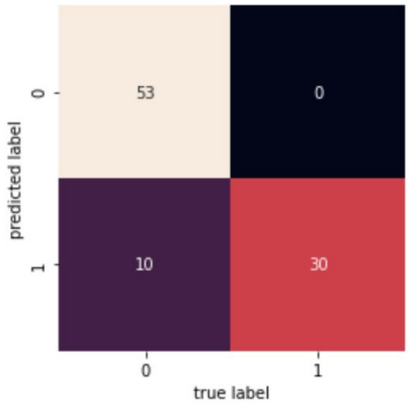
The Logistic Regression resulted in highest recall, precision, and f1-score for both resampling methods, SMOTE and ADASYN. To eliminate or reduce false negatives, it is recommended to use ADASYN resampling. With ADASYN recall was 1.00 for Normal values, meaning, all results labeled Normal were actually Normal. This results in a lower recall for Abnormal and is the trade-off for this safety measure with a recall of 0.83 for Abnormal results meaning 17 out of 100 Abnormal results were in fact Normal.

```
Training score:  0.8425925925925926
area under curve (auc):  0.9560846560846561
```



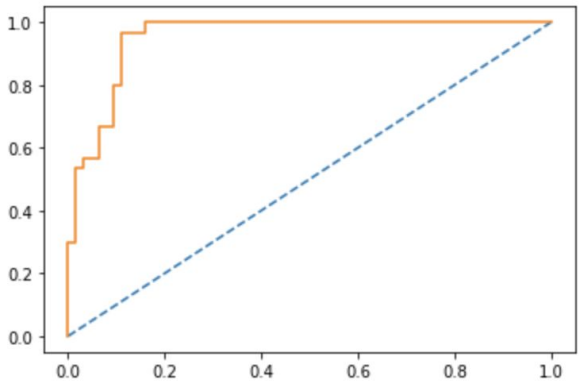

```
              precision    recall  f1-score   support

           0       1.00      0.84      0.91        63
           1       0.75      1.00      0.86        30

    accuracy                           0.89        93
   macro avg       0.88      0.92      0.89        93
weighted avg       0.92      0.89      0.90        93
```

```
Best Parameters:  {'class__C': 10, 'class__penalty': 'l2', 'class__solver': 'newton-cg'}
```



A more balanced result is possible using SMOTE, with .97 recall for Normal values, 3 out of 100 were in fact Abnormal, and .89 for Abnormal, 11 out of 100 were in fact Normal.

```
Training score:  0.8425925925925926
area under curve (auc):  0.9523809523809523
```



	precision	recall	f1-score	support
0	0.98	0.89	0.93	63
1	0.81	0.97	0.88	30
accuracy			0.91	93
macro avg	0.89	0.93	0.91	93
weighted avg	0.93	0.91	0.92	93

```
Best Parameters:  {'class__C': 10, 'class__penalty': 'l2', 'class__solver': 'newton-cg'}
```