Breast Cancer Predictions

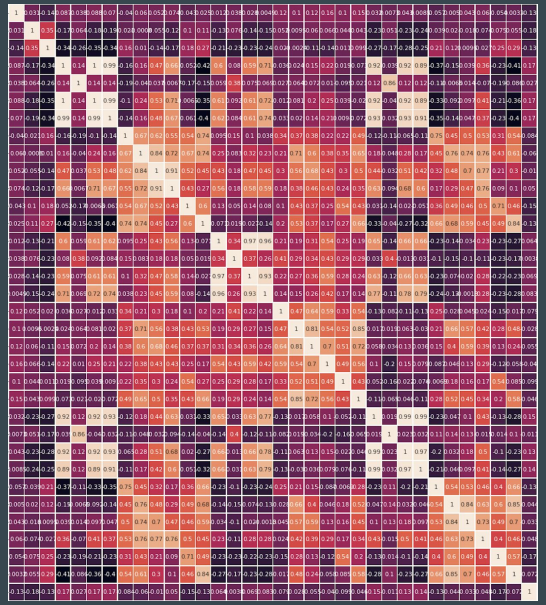
19th April, 2020

-Eklavaya Singh

Overview of the dataset & Pre-Processing

The given data set, has 35 features and 198 values with a class imbalance of 151 negatives and 47 positives. One feature has missing data(Lymph_Node_Status) and there is high correlation among some sets. Hence, following actions taken:

1) Missing values handled using imputer class and replacing with most frequent
2) Features with corr>0.9 filtered and only one kept
3) Values scaled using standard scaler for improved accuracy
4) Oversampling using SMOTE for handling class imbalance
5) Feature reduction using OLS regressor for better accuracy, increased speed and less overfitting



Correlation heat map of features

Classification(Three classifiers used)

Logistic Regression

-> The highest mean accuracy(80%) after stratified K-fold cross validation, but high standard deviation nearing 4%

XGboost

- Initially overfitting, but avoided by hyperparameter tuning
- Most stable result with 76% mean accuracy and 2% stddev after stratified k-fold cross validation

K-NN

Hyperparameters tuned using Grid Search CV, for n-neighbours = 2.

Good accuracy

THE FINAL ALGORITHM

To get a merged accuracy(the high unstable accuracy of log reg merged with K-NN and XGBOOST's stability) an ensemble method was used to get an accuracy of 78% with mere 1.5% std dev(checked by k-fold cross validation)

Regression for predicting Time in cases of Recurrence

Regression Used:- Polynomial Regression

High accuracy in prediction

The polynomial regression fitted very at degree 5.

High error below degree 5 and constant slight increase in error per degree increased

RMSE around 1.2755741081188927e-08