

**A Comparison of Logistic Regression(LR) and Random Forest(RF) Applied for Breast Cancer Prediction**

**Asha Guruvayurappan**

Research Glossary

***accuracy.*** The match between a sample and the target population is referred to as accuracy. It also indicates how close a value obtained from prediction is equal to the actual value.

***bagging (Bootstrap Aggregation).*** Bagging in ensemble machine learning creates several weak models, aggregating the predictions to select the best prediction.

***benign****.* A growth that is not cancerous, which does not invade nearby tissue or spread to other parts of the body.

***correlated predictors.*** Correlation is a statistical measure that expresses the extent to which two variables are linearly related, meaning correlated predictors change together at a constant rate.

***ensemble learning.*** A process in which multiple ML models are created and combined to solve a particular problem.

***features.*** Each feature is a column that represent measurable data which can be used for analysis.

***malignant.*** Malignant tumors have cells that grow uncontrollably. These tumors are cancerous , which spread locally or to distant sites via the bloodstream.

***negative Correlation.*** Relationship between two variables in which an increase in one variable is associated with a decrease in the other.

***normal Distribution.*** A normal distribution of data is one in which the majority of data points are relatively similar, meaning they occur within a small range of values with fewer outliers on the high and low ends of the data range.

***Overfitting.*** A overfitting model performs exceptionally on training data but fail to perform on test or unseen data.

***positive correlation.*** Relationship between two variables in which both move in same direction, i.e. when one increases the other increases.

***scatter Plot.*** A display of the relationship between two quantitative or numeric variables. A scatter plot shows the value of one variable plotted against the value of another variable.

***sigmoid curve.*** Sigmoid is used to predict the probabilities of a binary outcome. Sigmoid function will transform values between the range of 0 and 1

***standardization.*** Data standardization is the process of bringing data into a uniform format.

Data is received from various sources in various formats, necessitating standardisation for analysis.

***target*.** The target is a feature or column in the dataset that is classified or predicted based on all other features. These values can be both categorical or numerical.

**Abb. SMOTE**

Synthetic minority over sampling is a technique to over sample imbalanced class of data.

Formulas

Precision =

Recall =

F1 Score =

Attributes

1) Unique patient ID number   
2) Diagnosis (M = malignant, B = benign)   
3-32) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from centre to points on the perimeter)   
b) texture (standard deviation of grey-scale values)   
c) perimeter   
d) area   
e) smoothness (local variation in radius lengths)   
f) compactness (perimeter^2 / area - 1.0)   
g) concavity (severity of concave portions of the contour)   
h) concave points (number of concave portions of the contour)   
i) symmetry   
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is standard error Radius and field 23 is Worst Radius.

Data Analysis

In order to understand the depth of correlation, mean grouped variable correlation was visualized using heatmap. It was clear that some of the variables were highly correlated like area, radius, perimeter etc.

Even after normalization data was seen to follow the same tread as basic data. Figure 1 shows the histplots and its distribution up on normalization of mean data

Chart, histogram

Figure 1

Figure 1

Implementation Details

Logistic Regression

The model was trained with fitglm as the problem was to classify between Malignant and Benign. The model performed very well when trained and cross-validated on basic data, but adding noise drastically reduced the performance of the model. The Lasso trained hyperparameter[Figure 2] model did not perform as expected.

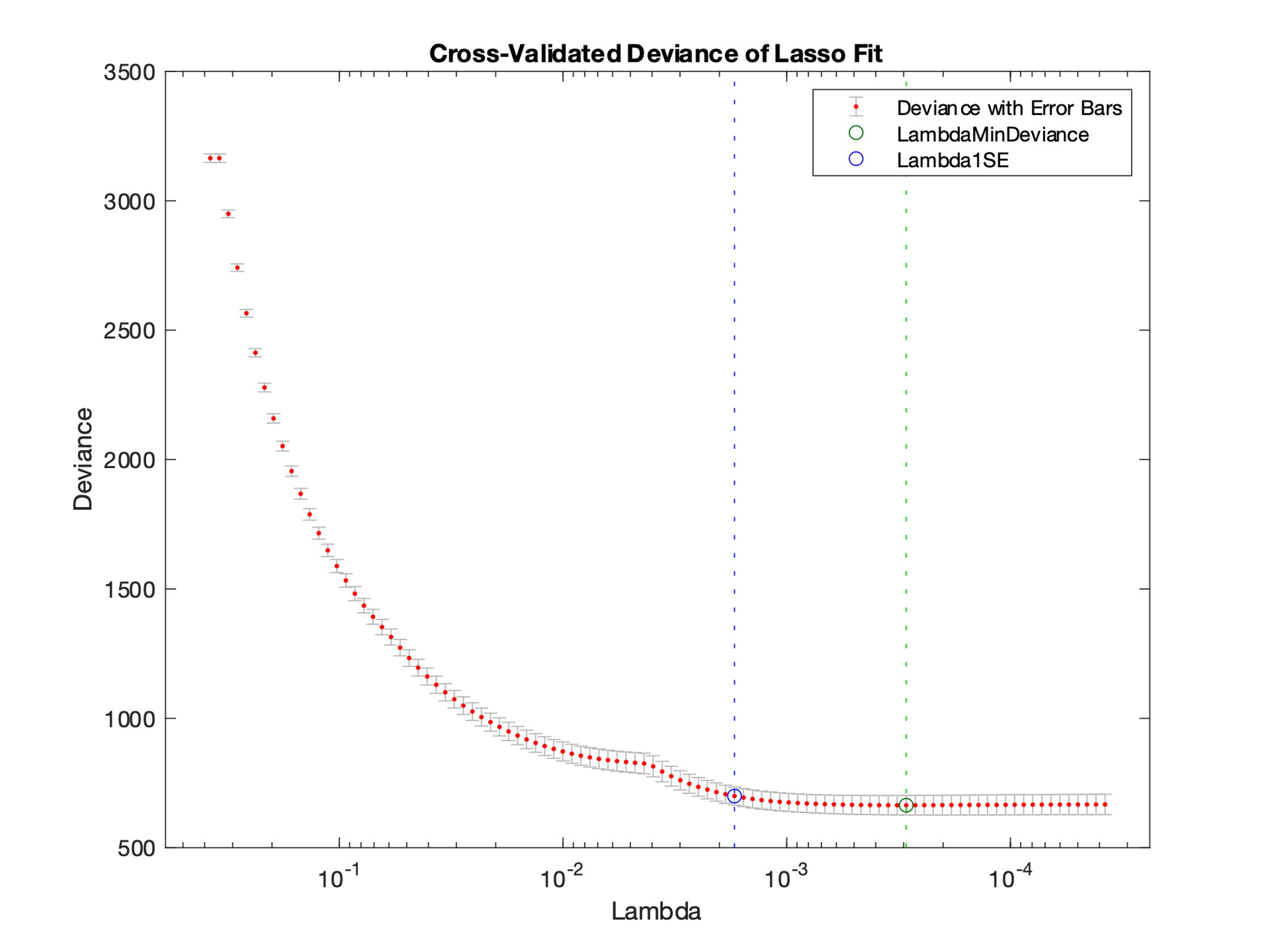
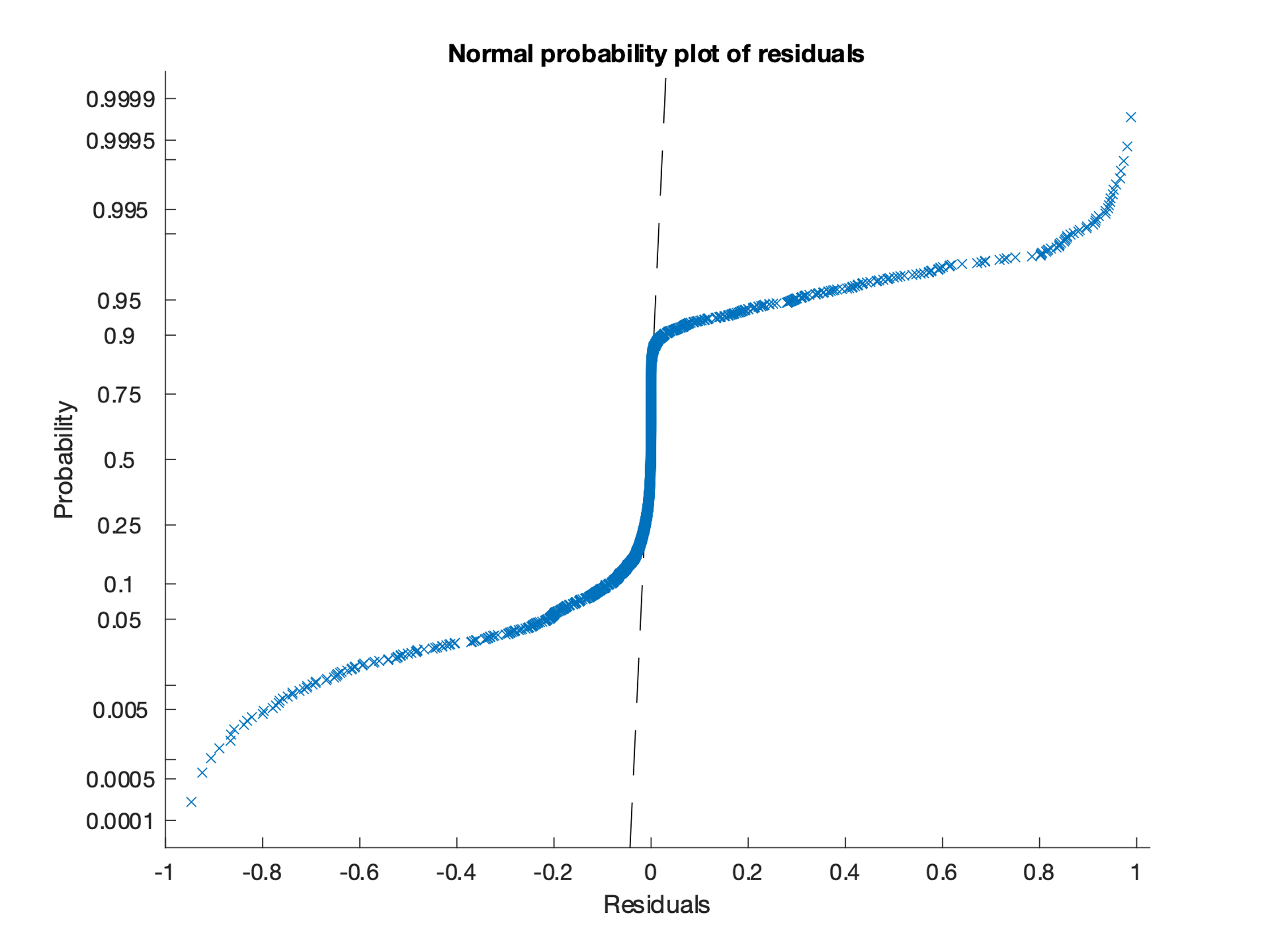
 

Figure 2 Figure 3

The residual plot [Figure 3] reveals that the probability curve between -1 to 0 is less steep than between 0 and 1

Random Forest

TreeBagger was trained with multiple combinations of tree size and leaf size before finalizing the best fit for model. A dynamic array stored the tree size, leaf size and accuracy of each combination from which the best tree - leaf size combination was obtained. A larger tree with minimum leaf was identified as the best working model, this was deduced by calculating out-of-bag accuracy and plotting out-of-bag errors for each tree - leaf combination.

Hyperparameter tuning was done by setting OOBPredictiors and calculation out-of-bag predictor delta error [Figure 4] of all features, the threshold was set at 0.5 after multiple model analysis.

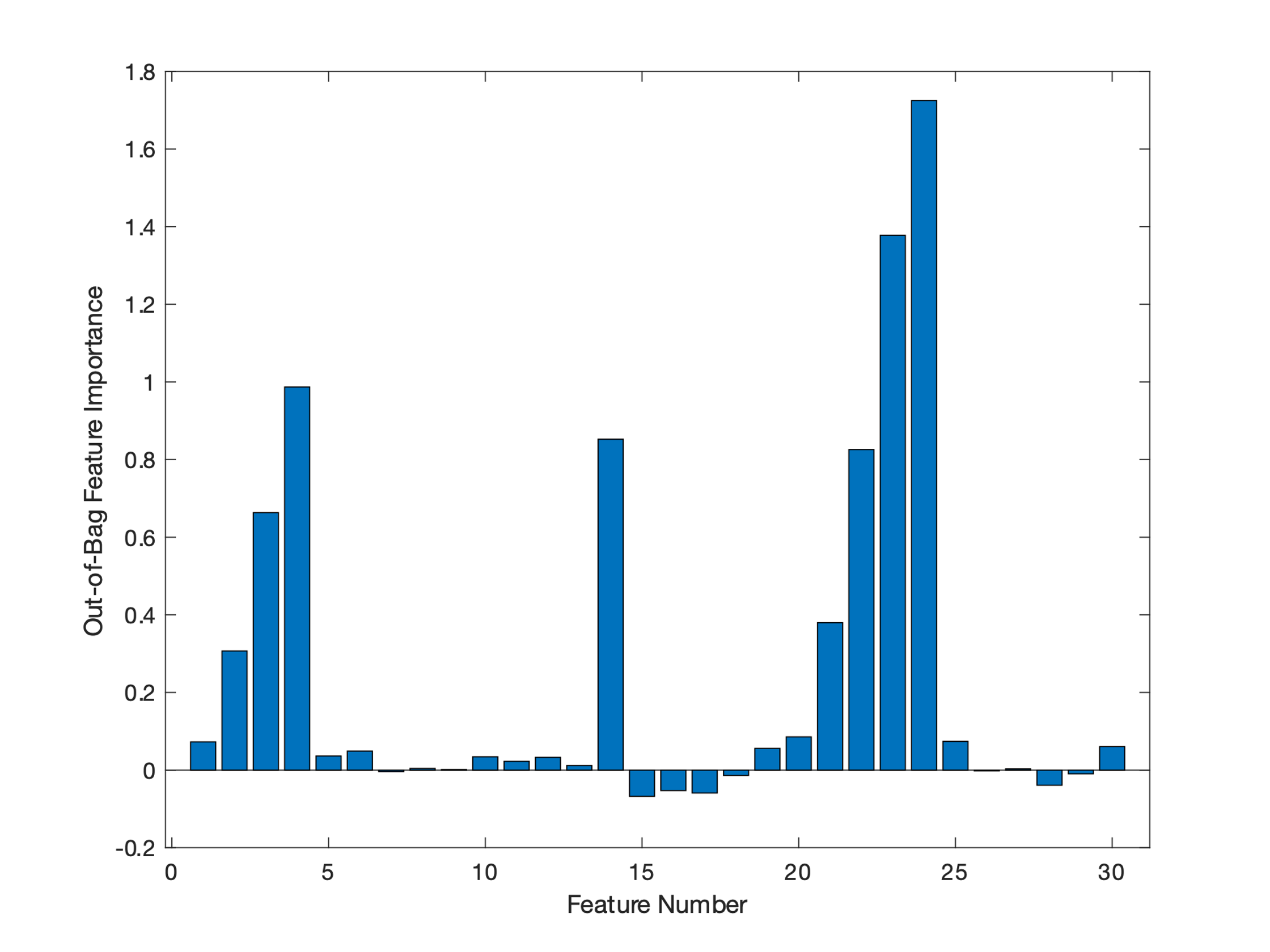


Figure 4

Conclusions

Both the models performed equally well when training and cross-validated on basic data, but on addition of noise, Random Forest performed much better when compared to logistic regression. Hyper parameter tuning improved only random forest models where logistic regression’s remind almost same.

Recall for Benign cases were almost equal in both Logistic Regression and Random Forest whereas recall for malignant cases was best predicted by random forest. if the goal is to create a model that predicts malignant cases accurately then Random Forest works much better. The overall performance of Random Forest is much better when compared to Logistic Regression for this dataset.