Predicting Cancer Diagnoses from Breast Mass Imaging Data

Jack Rogers (903848157)

Predictive Modeling II

Problem Definition

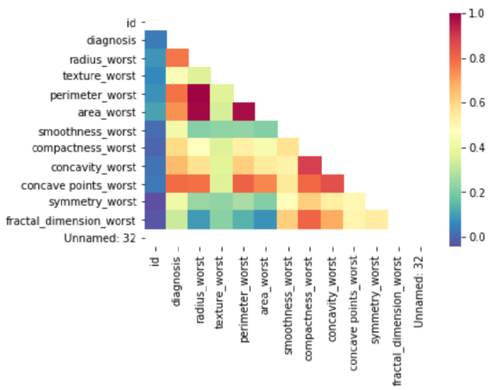
The goal of this project is to predict cancer diagnoses; specifically, whether a breast mass is malignant or benign. This is a very important issue, and high accuracy can improve both the speed and confidence of diagnoses as well as the treatment of the patient. Having even a small head start on an aggressive condition like breast cancer can be life-saving for those with malignant masses.

This project utilizes a dataset from Kaggle (originally from the UCI ML Department) that contains digitized imaging data of breast masses from patients in Wisconsin. From the image data, 30 features were extracted measuring the size and shape of the masses. The dataset fortunately had no missing data and required very little cleaning.

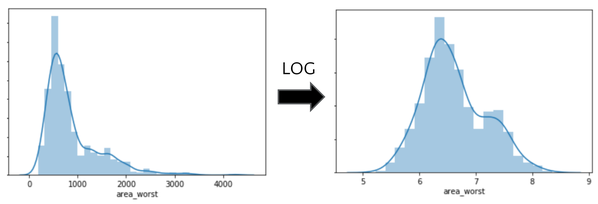
Feature Selection

The thirty features that were initially present in the dataset would have made the model quite complex, so the "mean" and "se" (standard error) attributes of each feature were dropped, leaving just the "worst" (largest). This resulted in 10 predictor variables.

A correlation heatmap was plotted for these predictors, showing that a few of them were correlated. The highly-correlated predictors were dropped to reduce model complexity while keeping predictive power.



A pair-plot of the variable distributions revealed that a couple of the predictors were heavily right-skewed and would benefit from a log-transformation. An example of a log-transformed variable, "area_worst", is shown before and after the logarithmic transformation.



Modeling & Evaluation

After feature selection, there were 7 predictor variables:

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      569 non-null    int64
 1   diagnosis               569 non-null    int32
 2   texture_worst           569 non-null    float64
 3   log_area_worst          569 non-null    float64
 4   smoothness_worst        569 non-null    float64
 5   log_compactness_worst   569 non-null    float64
 6   concave points_worst    569 non-null    float64
 7   symmetry_worst          569 non-null    float64
 8   fractal_dimension_worst 569 non-null    float64
```

Legend
- Identifier
- Target Variable
- Predictors

The data was then split into training and test sets, with 75% allocated for training:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state=10)
```

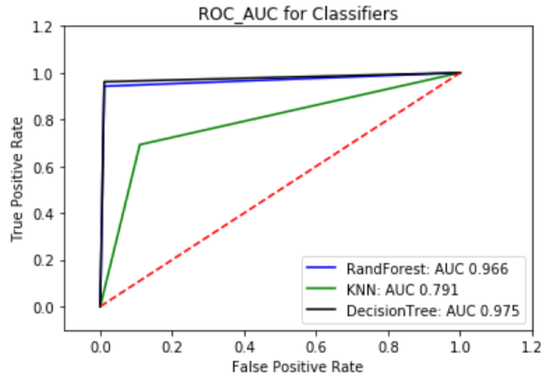
3 models total were chosen to classify the breast cancer diagnoses: two regular classifiers (Decision Tree, K-Nearest Neighbors) and one ensemble classifier (Random Forest). One tuning parameter was optimized for each model through an iterative fit-and-predict process to improve performance. The best-performing version of each model was fit on the training data. They were then evaluated on accuracy, specificity, and sensitivity from test data.

Classification Model	Hyperparameter Tested	Accuracy	Sensitivity	Specificity
Random Forest	*Max Depth* (Optimal = 3)	97.20%	94.23%	**98.90%**
Decision Tree	*Max Depth* (Optimal = 3)	**97.90%**	**96.15%**	**98.90%**
K-Nearest Neighbors	*N Neighbors* (Optimal = 1)	81.82%	69.23%	89.01%

After each model made predictions on the test data, they were evaluated on the metrics listed in the chart above. As shown, there was a clear best performer. The Decision Tree model outperformed the other two models in accuracy and sensitivity. It also had the greatest specificity, recording the same value as the Random Forest model.

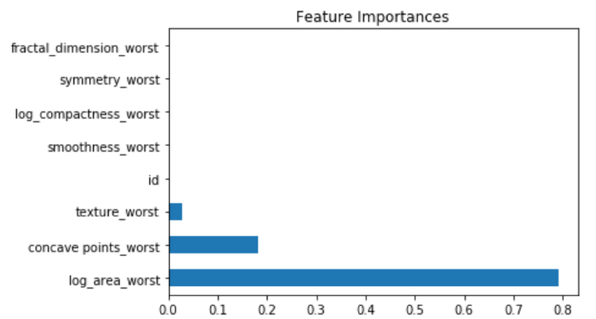
Notably, the Decision Tree attained over 96% sensitivity, which outperformed the second-best model by almost 2 percentage points. Since the goal is to predict the presence of a malignant, or dangerous, form of breast cancer, sensitivity is a very important metric.

To the right, "area under the curve" scores were calculated and plotted for the classifiers. Once again, the Decision Tree model has the best score, recording a very impressive 97.5 ROC AUC. This confirms that the Decision Tree model is the best predictor of the group.

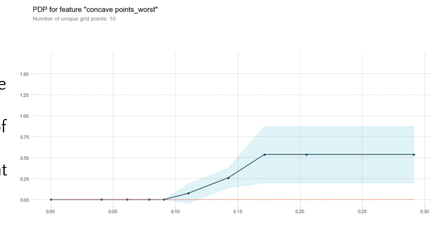


Results

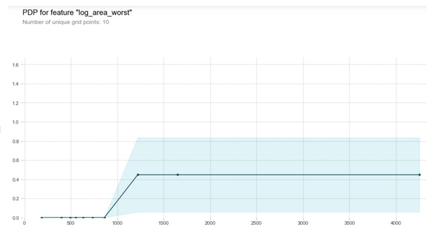
The Decision Tree has been determined to be the best model, so now the most important features can be extracted. One feature, the log of the worst area measurement, is by far the most important.



The second-most important feature determining diagnoses is concave points on the contour of the cell nuclei. As the number of points increase, the likelihood of a malignant mass increases, then flattens out.



The most important feature determining diagnoses is magnitude of the largest area measurement of the cell nuclei. As the log of area increases, the likelihood of a malignant mass increases sharply, then flattens out.



Conclusion

Using a Decision Tree Classifier, breast cancer diagnoses can be predicted with 96% sensitivity and nearly 98% accuracy using only 3 predictors. The texture, number of concave points, and area (logarithmically-transformed) of cell nuclei in breast masses are shown to be very effective in classifying instances as benign or malignant.

In addition, a more high-level view of predictors can be established by grouping the important features. The significantly high importance of the area variable strongly suggests that the **size** of the breast mass is the most relevant predictor, followed by the **shape**. The variables measuring concavity (concave points) and texture can be considered shape-related attributes.

This model can aid in recognition of dangerous masses and result in better care for patients that develop potentially malignant tumors.