Investigating the relationship between mammographic density and risk of breast cancer, via logistic regression and random forest.

Ben McGuirk

January 2024

1 Variables

Relevant measures of my variable summaries include mean, median, mode, range, standard deviation, skewness and kurtosis.

1.1 Age

- Mean: 50.17
- Median: 49
- Mode: 48
- Range: 35
- Standard Deviation: 6.21
- Skewness: 0.43
- Kurtosis: 0.04

1.2 Body Mass Index

- Mean: 26.72
- Median: 25.7
- Mode: 30.1
- Range: 32.8
- Standard Deviation: 4.83
- Skewness: 1.09
- Kurtosis: 1.46

1.3 Mammographic Density

- Mean: 44.45
- Median: 40
- Mode: 0
- Range: 100
- Standard Deviation: 30.18
- Skewness: 0.002
- Kurtosis: -1.34

2 Data Preprocessing

2.1 Train Test Split

Splitting the data into train, cv and test sets helps to prevent overfitting. Overfitting leads to high accuracy and performance when measured against the training data, however when measured against new unseen data it performs poorly. The data was split 60% training, 20% cv and 20% test.

2.2 Isolation Forest

After the split outliers were removed using the isolation forest algorithm. In total, 54 outliers were found; 32 in the training set, 11 in the cv set and 11 in the test set.

2.3 Z-score Normalisation

Finally features were scaled with z-score normalisation. Logistic regression can be sensitive to large scale features. Z-score normalisation is defined by the following equation:

$$z = \frac{x - \mu}{\sigma}$$

where:

- z is the z-score
- x is the original value
- μ is the mean of the variable
- σ is the standard deviation of the variable

3 Logistic Regression

3.1 Method

Sigmoid Function

The first step of building a logistic regression model is to define a hypothesis function. The sigmoid function was chosen with parameters initialised to zero. The sigmoid function is defined by the following equation:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

where:

- $h_\theta(x)$ is the predicted probability.
- θ is the parameter vector.
- x is the input feature vector.
- e is the base of the natural logarithm.
- T indicates the dot product of θ and x.

Logistic Regression Cost Function

The cost function determines the average loss across the data set (cross-entropy loss):

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

where:

- $J(\theta)$ is the cost function.
- m is the number of training examples.
- $x^{(i)}$ is the feature vector for the i-th example.
- $y^{(i)}$ is the true label for the i-th example (0 or 1).

Gradient Descent

To minimise the cost function gradient descent updates the parameters over a number of iterations. Gradient descent is defined by the following equation:

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

3.2 Results

Gradient descent was carried out with an initial value for α of 0.01. An automatic convergence test was applied and the model converged after 80 iterations, with a cost of 0.59. As you increase α the model takes a steeper step towards a global minimum, however if α becomes too large, the model can 'overshoot' and increase cost. Here a value for α of 1 was determined as the most effective, converging after 14 iterations with a cost of 0.35.

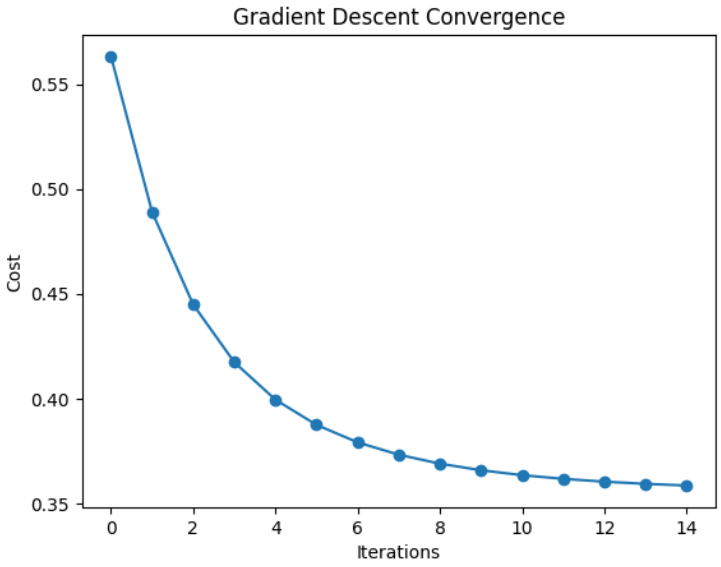

Figure 1: convergence of logistic regression model ($\alpha = 1$).

Accuracy when tested on the train, cv and test sets were 88.47%, 89.11% and 88.61% respectively. These scores demonstrate that the model has a high accuracy and has not been overfit to the train dataset.

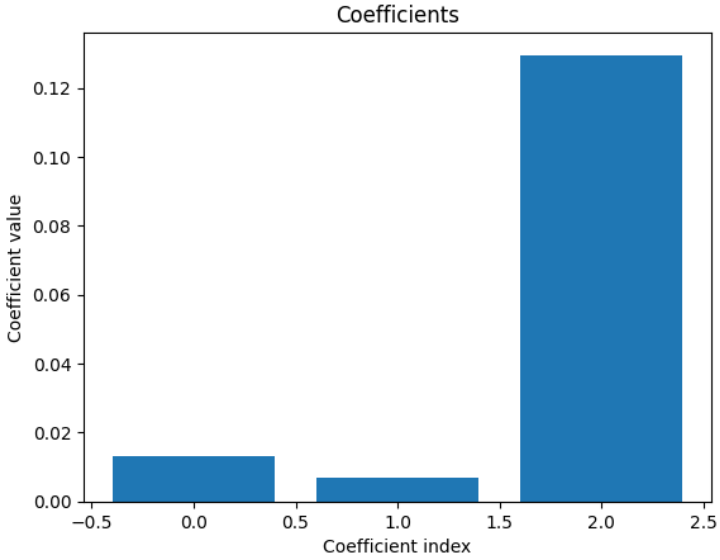

Figure 2: coefficient plot of logistic regression model.

4 Random Forest

4.1 Method

For the second machine learning model a random forest was chosen. The random forest model was applied using the Scikit-learn implementation.

4.2 Results

Accuracy scores on the train, cv and test sets were 99.84%, 87.62% and 87.13% respectively. The difference between the accuracy on the train set vs the cv and test sets suggest that the model might be slightly overfit, however accuracy is still high across each set. Ways of reducing high variance include increasing the size of the dataset, feature selection and applying regularisation to the cost function.

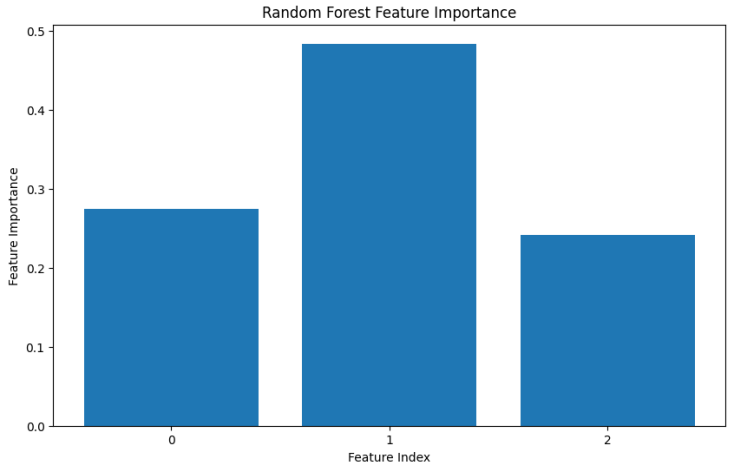

Figure 3: feature importance scores for each variable.

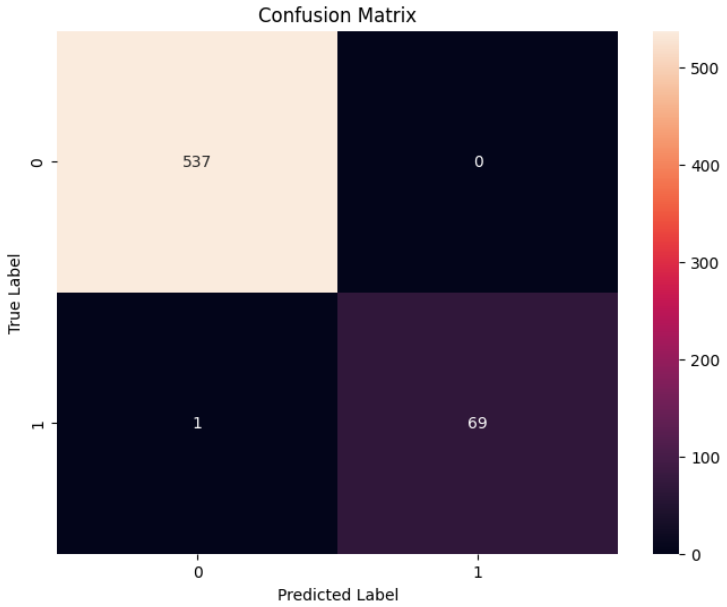

Figure 4: The confusion matrix shows that the model performed extremely well, predicting 0 false negatives and just one false positive, out of 607 training examples.

5 Discussion

5.1 Logistic Regression vs Random Forest

The coefficient plot showed that mammographic density has the strongest influence on the model. High coefficient value in logistic regression does not necessarily indicate that the same variable will have a high feature importance score in a random forest model. The lower feature importance score attributed to mammographic density just indicates that it has a low effect on the reduction of impurity across all trees, and so may still have a large influence over a patient's risk of breast cancer.

The random forest model was extremely accurate when tested against the train dataset, however achieved similar accuracy scores to logistic regression when tested on the cv and test datasets.

5.2 Future thoughts and directions

A convolutional neural network could be used for determining mammographic density, rather than the manual approach used by a radiologist. A hybrid approach, where the radiologist can assess scans where the confidence of the neural network is low, may help to maintain efficiency and accuracy.

Future investigation could include testing the respective models with different proportions of train, cv and test sets, as well as the performance of different polynomials of x. Implementation of regularisation may prevent random forest overitting.

Overall, mammographic density has a clear influence on breast cancer risk, demonstrated by its high coefficient for x. Investigation of additional variables may uncover further risk factors, helping to build well rounded diagnostic tools for breast cancer.

6 Appendix

GitHub repository: `https://github.com/BenMcGuirk/QMUL_PhD_task`