Introduction

A mammogram is an X-ray picture of the breast, used to identify masses which may indicate signs of breast cancer. Physicians can use the mammogram to score the likelihood of malignancy from 1 (definitely benign) to 5 (highly suggestive of malignancy). However, the 'BI-RADS' (Breast Imaging Reporting and Database System) score tends to be overly conservative meaning between 55-80% of subsequent biopsies turn out to be benign (Chhatwal et al, 2010). The goal of this report is to develop a model which will identify malignancy with more precision than the BI-RADS score, whilst maintaining a sufficiently high true positive rate. That is, with high probability, the model should correctly predict malignancy in patients with malignant masses. The model will be a supervised machine learning model with inputs based on the mammographic features captured in the mammographic mass data set (MMDS) located in the UCI machine learning repository (Elter and Schulz-Wendtland, 2007). This dataset contains data on 961 patients each labelled with a true malignancy value of either 0 (benign) or 1 (malignant) based on full field digital mammograms collected by the Institute of Radiology at the University of Erlangen-Nuremberg between 2003 and 2006. The input features are:

- Patient's age in years
- Mass shape coded as round=1, oval=2, lobular=3 or irregular=4
- Mass margin coded as circumscribed=1, microlobulated=2, obscured=3, ill-defined=4 or spiculated=5
- Mass density coded as high=1, iso=2, low=3 or fat-containing=4

130 patients (13.5%) have missing data. Missingness by feature is Age: 5, Shape: 31, Margin: 48 and Density: 76.

Each patient also has an associated BI-RADS score, from which a prediction of malignancy can be derived based on different threshold values (e.g., BI-RADS $\geq 4 \implies$ malignant). This will provide the baseline for which the model must outperform. Note however, the BI-RADS score is *not* a feature of the model. The figures below show class label frequency by feature class.

Class Label Frequency vs Patient Age

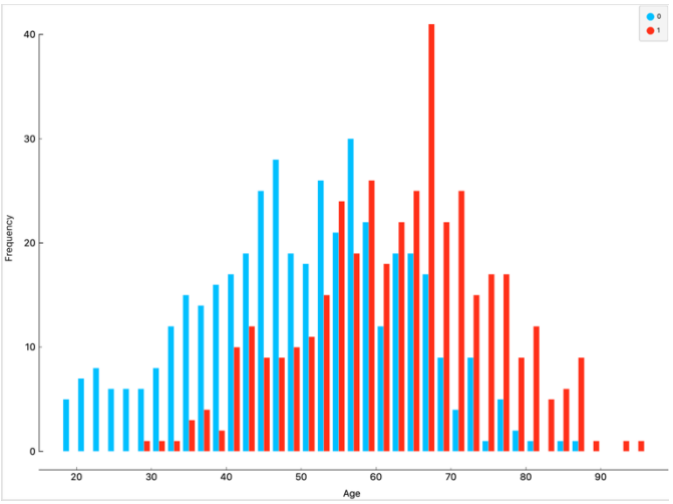

Figure 1

Class Label Frequency vs Mass Shape

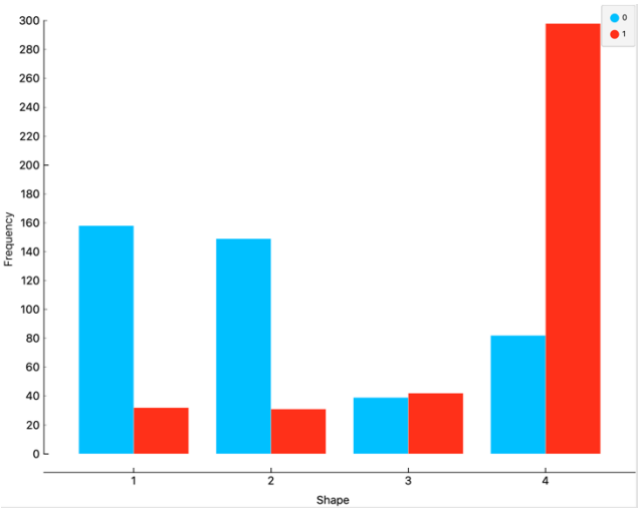

Figure 2

Class Label Frequency vs Mass Margin

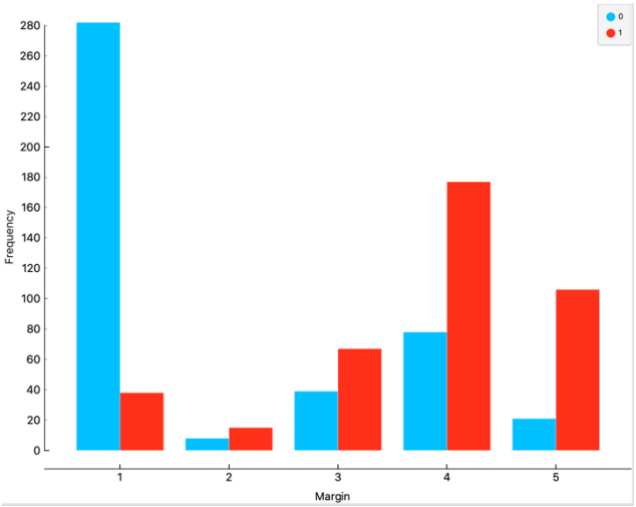

Figure 3

Class Label Frequency vs Mass Density

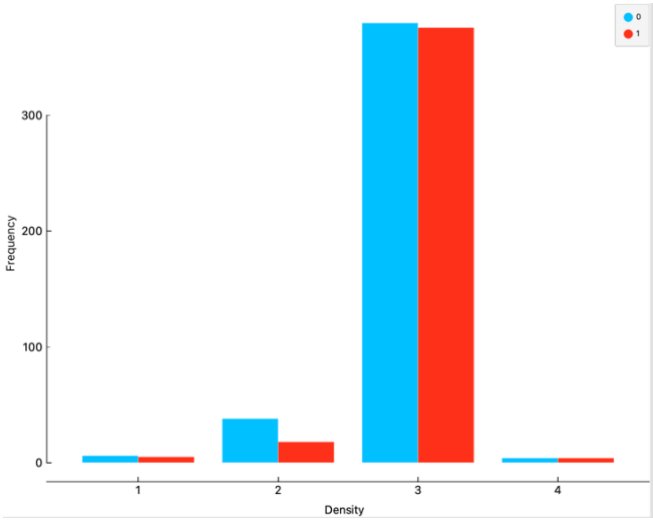

Figure 4

These distributions suggest being younger or having a mass which is round or oval in shape, circumscribed in margin and possibly iso in density will be protective characteristics. Whereas being older or having a mass which is irregular in shape, obscured, ill-defined, spiculated or possibly microlobulated in margin will increase risk of malignancy. Most of these characteristics are expected to be formally identified during feature selection. Age has no obviously erroneous data (negative or very large values), which could have otherwise skewed parameter estimates or affected pre-processing steps such as normalization. Finally, the classes are relatively well balanced (51.5% benign vs 48.5% malignant), which should prevent the model from simply predicting the dominant class.

Methodology

Models

The two models will be outlined first as this will inform some pre-processing decisions. Below $\boldsymbol{X}$ is the vector of features, Y the class label, $\boldsymbol{W}$ the vector of model parameters, and $\boldsymbol{I}$ the identity matrix

Random Forest: For a dataset of size N create lots of bootstrapped datasets (random sampling with replacement) of size N and overfit decision trees on each (i.e., each tree will be deeper than optimal). Random Forest is the average of these classifiers and as such can be considered pseudo-probabilistic. The idea is that because the bootstrapped data sets are different, each tree overfits in different places so when the trees are averaged the general pattern is retained, but areas of overfitting get averaged away. This means random forest is typically good out of the bag as less hyperparameter tuning is needed to control for overfitting, although some gain in performance may still be achieved by setting a lower limit on the node size splits can occur on (N_{limit}), and restricting the number of features considered at any given split (N_{split}). Algorithm pseudo-code (Ranganathan et al, 2018):

```
Precondition: A training set S: = (x1, y1),...,(xn, yn),
features F, and number of trees in forest B.
function RandomForest (K, L)
        H←∅
        for i ∈ 1,...,B do
                K(i) ← A bootstrap sample from S
                hi ← RandomizedTreeLearn(K(i), L)
                H ← H ∪ {hi}
        end for
        return H
end function
function RandomizedTreeLearn(K, L)
        At each node:
                f ← very small subset of L
                Split on best feature in f
        return the learned tree
end function
```

Logistic Regression: A logistic regression model is a probabilistic model of the form $P(Y = 1 \mid \boldsymbol{X}) = \sigma(w_0 + w_1 x_1 + \cdots + w_n x_n)$, where $\sigma(z) = (1 + e^{-z})^{-1}$. It is highly interpretable, since each unit in increase in x_i corresponds to an increase in the odds of $Y = 1$ by a factor of $\sigma(w_i)$. To control for overfitting, parameters are constrained such that $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, C\boldsymbol{I})$, for some hyperparameter C. So instead of simply choosing $\boldsymbol{W}$ which

maximizes the likelihood of the observed data D via $\max_{W} P\,(D\,;\boldsymbol{W})$, $\boldsymbol{W}$ is now chosen according to $\max_{W}\ \{P\,(D \mid \boldsymbol{W}) * P(\boldsymbol{W}\,;\ C)\}$. That is, $\boldsymbol{W}$ must balance how well it fits the data, accounting for the fact that smaller $w's$ are more likely. A smaller C defines a tighter prior distribution encouraging smaller parameters. Algorithm pseudo-code (Chapelle and Li, 2011):

Require: Regularization parameter $\lambda > 0$.
$m_i = 0,\ q_i = \lambda$. {Each weight w_i has an independent prior $\mathcal{N}(m_i, q_i^{-1})$}
for $t = 1, \ldots, T$ **do**
Get a new batch of training data $(\mathbf{x}_j, y_j),\ \ j = 1, \ldots, n$.
Find $\mathbf{w}$ as the minimizer of: $\frac{1}{2}\sum_{i=1}^{d} q_i(w_i - m_i)^2 + \sum_{j=1}^{n} \log(1 + \exp(-y_j \mathbf{w}^\top \mathbf{x}_j))$.
$m_i = w_i$
$q_i = q_i + \sum_{j=1}^{n} x_{ij}^2 p_j(1 - p_j),\ \ p_j = (1 + \exp(-\mathbf{w}^\top \mathbf{x}_j))^{-1}$ {Laplace approximation}
end for

<u>Pre-processing</u>

<u>Step 1 - Imputation</u>: If missingness is not completely at random, removing these patients could bias the model. For example, perhaps the missing margin values were generally ill-defined (margin = 4) which figure 2 suggests likely increases risk of malignancy. Imputation can mitigate against this bias. Imputation was done using Orange's default model-based imputer, which builds a classification or regression tree on the other features in order to impute missing values in the target feature. This method should give imputed values closer to the true (unknown) values vs constant-value imputers based on the features mode/average.

<u>Step 2 - Feature Construction</u>: All features except age are categorical. The default encoding assumes some natural ordering. For example, an irregular mass (shape = 4) takes a higher value than a round one (shape = 1). For the features shape and margin, this ordering is completely arbitrary. Although density is ordinal, this encoding may still constrain the model unnecessarily, since a logistic model $\sigma(w_{density} x_{density} + \cdots)$ must assume the density term $w_{density} x_{density}$ grows linearly with density. To remove these arbitrary and restrictive encodings, each class is encoded as its own feature ('one-hot' encoding). Often, one of the newly 'one-hot' encoded features needs dropping to avoid perfectly correlated features (i.e., to account for the fact $x_{shape=1} = 1 - (x_{shape=2} + x_{shape=3} + x_{shape=4})$). However, regularisation prevents the singularity that this dependency induces during parameter estimation, so in this instance dropping features is unnecessary.

<u>Step 3 - Feature Pre-processing</u>: The regularised logistic model assumes $w_i \sim \mathcal{N}(0, C)$, meaning the degree to which each w_i is encouraged to be small is the same (they share the same C). But if the features vary in scale, then it is not reasonable to expect the parameters to be of comparable magnitude, so this shared prior wouldn't make sense. Therefore, age is normalized to lie in the interval [0,1].

Step 4 - Feature Selection: Features are ranked by importance. The $i'th$ features importance is derived by fitting a random forest on the data and computing, on the average tree, the total reduction in Gini impurity across the nodes which split on feature i. Splits on nodes with more datapoints are weighted to give a higher impurity reduction. Features with more classes may have inflated importance as more classes creates more opportunity for an impurity decreasing split to occur by chance. However, the one-hot encoding should remove this bias. For robustness other metrics of importance based on the Gini ratio and information gain are also computed.

Models, Training and Evaluation

Models: Random Forest for N_{limit} in $\{3,5,7\}$ and for each of these values N_{split} in $\{2,3,4\}$. Logistic regression for C in $\{0.001, 0.3, 0.5, 0.7, 1, 3, 5, 10, 50\}$.

Training: Each model will be trained and evaluated via 10-fold cross validation. That is, the data will be split into 10 pieces on which 9 will be used for model training and 1 for evaluation. Model performance is the average performance over the 10 possible choices of evaluation fold.

Performance Metrics:

- AUC: Measures the trade-off between the True Positive Rate (TPR) = $P(Classified\ as\ Malignant \mid Mass\ Maligant)$ and the False Positive Rate (FPR) = $P(Classified\ as\ Malignant \mid Mass\ Benign)$ at different probability thresholds for classifying a mass as malignant. An AUC close to 1 indicates a high TPR can be achieved whilst maintaining a low FPR.
- Precision = $P(Mass\ Malignant \mid Classified\ as\ Maligant)$: Measures how likely a mass predicted as malignant is truly malignant.
- The TPR as a stand-alone metric. Note TPR is also referred to as recall.

Model Pipeline

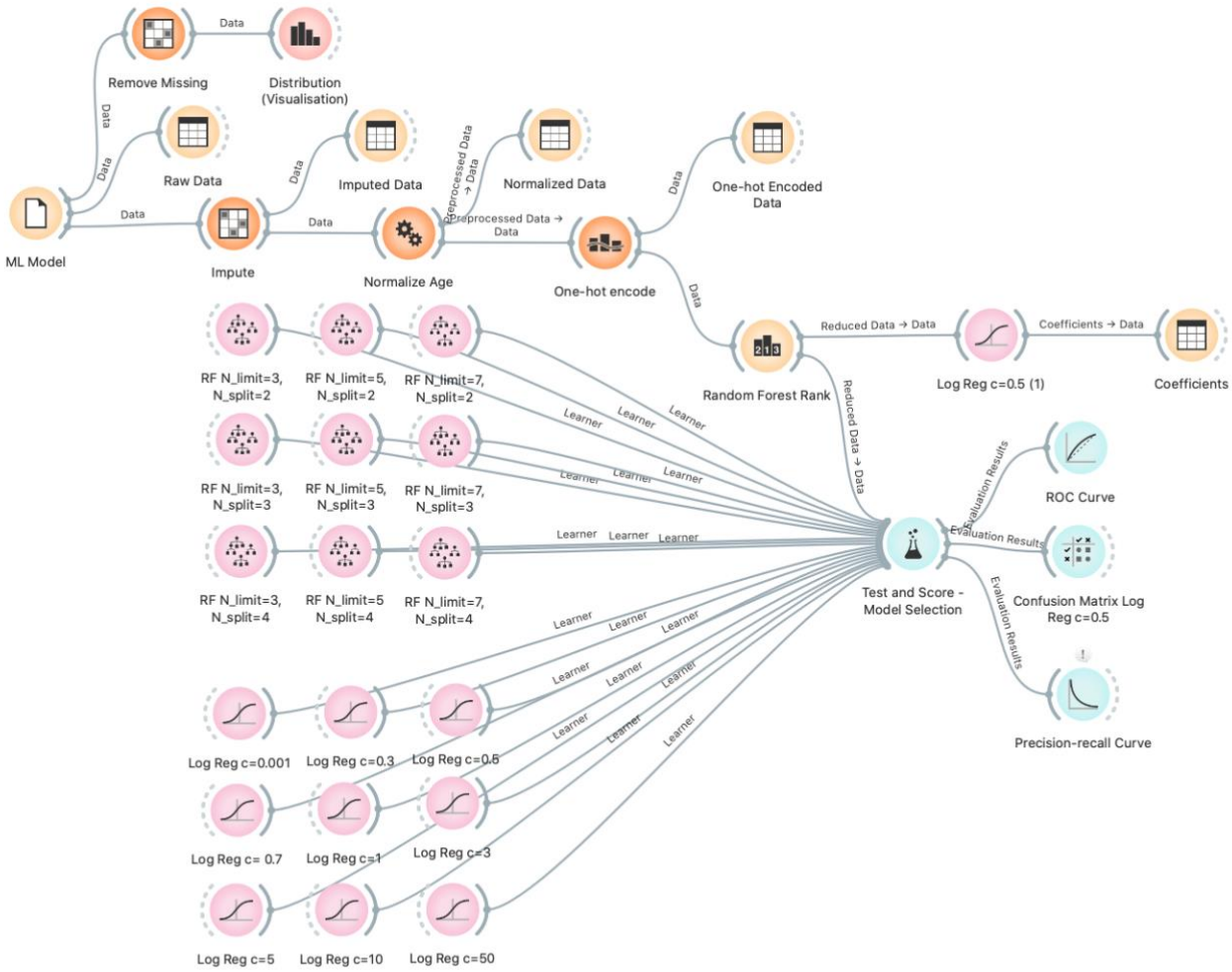

Figure 5

Note, before computing the confusion matrices for the BI-RADS models, 19 patients with either missing or out-of-bound BI-RADS scores were removed.

Results

The top 9 Random Forest ranked features were each noted above as potentially predictive characteristics. Since importance is minimal after feature 7 (across all three measures), just the top 7 are kept.

Random Forest Feature Ranking

		#	Info. gain	Gain ratio	Gini
1	N Margin=1		0.253	0.260	0.159
2	N Shape=4		0.231	0.234	0.151
3	N Age		0.129	0.064	0.084
4	N Shape=1		0.087	0.108	0.056
5	N Shape=2		0.079	0.104	0.050
6	N Margin=5		0.072	0.122	0.046
7	N Margin=4		0.053	0.060	0.036
8	N Margin=3		0.010	0.020	0.007
9	N Density=2		0.005	0.015	0.003
10	N Density=3		0.004	0.009	0.003
11	N Margin=2		0.001	0.006	0.001
12	N Density=4		0.000	0.001	0.000
13	N Shape=3		0.000	0.000	0.000
14	N Density=1		0.000	0.000	0.000

Figure 6

Performance Metrics

Model Performance

Model	AUC	CA	F1	Precision	Recall
Log Reg c=3	0.866	0.799	0.790	0.765	0.818
Log Reg c=5	0.866	0.798	0.789	0.764	0.816
Log Reg c=50	0.866	0.799	0.790	0.767	0.813
Log Reg c=10	0.866	0.799	0.790	0.767	0.813
Log Reg c=1	0.865	0.797	0.791	0.756	0.829
Log Reg c= 0.7	0.865	0.800	0.794	0.760	0.831
Log Reg c=0.5	0.863	0.802	0.796	0.763	0.831
Log Reg c=0.3	0.863	0.801	0.794	0.763	0.827
Log Reg c=0.001	0.854	0.685	0.562	0.789	0.436
RF N_limit=7, N_split=2	0.830	0.768	0.755	0.739	0.771
RF N_limit=7, N_split=3	0.830	0.767	0.754	0.737	0.773
RF N_limit=5, N_split=3	0.828	0.768	0.755	0.738	0.773
RF N_limit=7, N_split=4	0.827	0.763	0.749	0.734	0.764
RF N_limit=5, N_split=2	0.826	0.766	0.756	0.731	0.782
RF N_limit=5, N_split=4	0.826	0.761	0.747	0.732	0.762
RF T_depth=3, N_split=2	0.824	0.763	0.750	0.732	0.769
RF N_limit=3, N_split=3	0.822	0.766	0.752	0.737	0.769
RF N_limit=3, N_split=4	0.820	0.763	0.750	0.732	0.769

Figure 7

The logistic models clearly outperform the Random Forest models, giving better AUC, precision and recall rates. From a patient safety perspective, it's better to have a model with a stronger TPR even if AUC and precision are slightly worse. This is because misclassifying a benign mass primarily incurs a resource/time cost (from an unnecessary biopsy), whereas misclassifying a malignant mass could put the patients' health at significant risk if the mass develops. Therefore $C = 0.5$ is selected as the final model as its TPR (83.1%) is better than the less regularised models by ~1.3-1.8% but its precision is only marginally worse by 0.1-0.4% and AUC by 0.003.

Therefore, a logistic model with $C = 0.5$ is trained over the data. Table 1 summarises the parameter estimates and their corresponding interpretations.

Term	Coefficient Estimate	Increases odds of malignancy by a factor of…
Intercept	-1.38	NA
Circumscribed (Margin = 1)	-1.17	0.31
Irregular (Shape = 4)	0.92	2.5
Age	2.63	13.88
Round (Shape = 1)	-0.47	0.62
Oval (Shape = 2)	-0.68	0.5
Spiculated (Margin = 5)	0.79	2.2
Ill-defined (Margin = 4)	0.19	1.21

Table 1

Note in particular, that increasing age is the greatest risk factor of malignancy whereas a circumscribed (well-defined) margin is the most protective feature.

The ROC curve demonstrates that there is a diminishing trade-off between the TPR and FPR rate as the classification threshold is lowered below 50% (I.e., a more conservative classifier of malignancy). A rapidly increasing FPR indicates the model may be losing precision quickly as this threshold is lowered.

Log Reg (C=0.5) ROC Curve (TPR vs FPR)

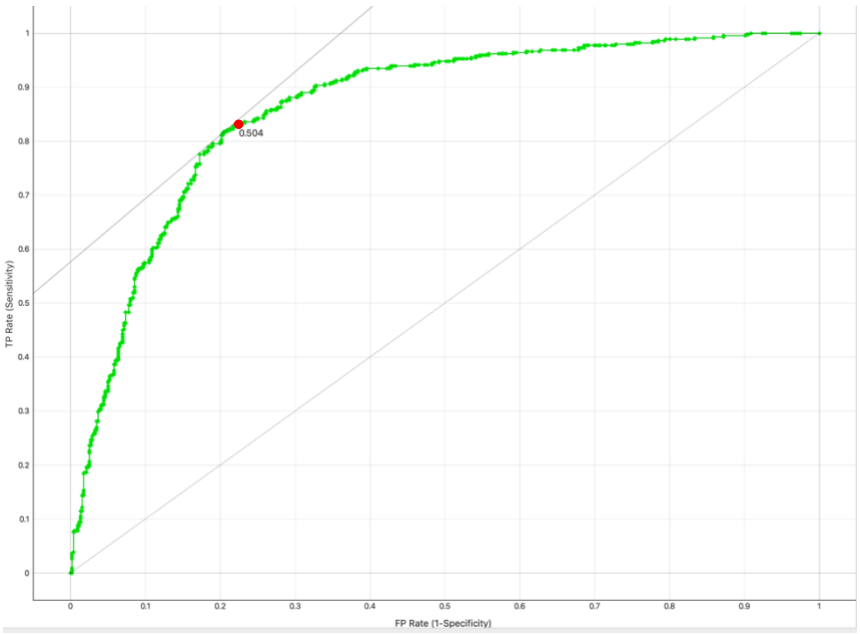

Figure 8

A precision-recall curve makes the exact nature of this relationship clearer. This is a better curve to consider for this problem as it directly measures how resource efficiency (precision) trades-off with patient safety (TPR). Figure 8 suggests the optimal *mathematical* trade-off occurs around a classification threshold for malignancy of 50% (red dot on curve), since at lower thresholds, each unit increase in the TPR incurs a more rapid drop in precision. However, for the purposes of patient safety, it may be necessary to accept a less 'optimal' trade-off, if the current TPR of 83.1% (at threshold 50%) is deemed too low. More on this in discussion.

Log Reg (C=0.5) Precision-Recall Curve

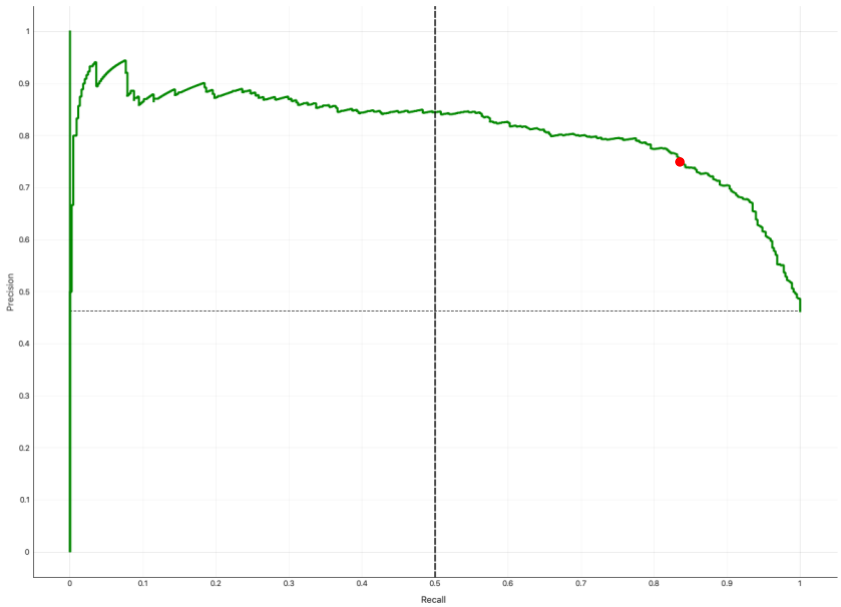

Figure 9

The confusion matrix and the key performance metrics are summarised in the tables below.

Log Reg (C=0.5) Confusion Matrix, Threshold = 50%		
	Predicted = 0	Predicted = 1
Actual = 0	401	115
Actual = 1	75	370

Table 2

Log Reg (C=0.5) Key Performance Metrics, Threshold = 50%		
TPR	FPR	Precision
$\frac{370}{75+370} = 0.831$	$\frac{115}{401+115} = 0.223$	$\frac{370}{115+370} = 0.763$

Table 3

For comparison, the same metrics for two BI-RADS scoring-based models, at classification thresholds of 4 and 5 respectively, are summarised below.

BI-RADS Confusion Matrix, Threshold = 4		
	Predicted = 0	Predicted = 1
Actual = 0	43	467
Actual = 1	7	425

Table 4

BI-RADS Confusion Matrix, Threshold = 5		
	Predicted = 0	Predicted = 1
Actual = 0	470	40
Actual = 1	127	305

Table 5

BI-RADS Models Key Performance Metrics			
Model	TPR	FPR	Precision
Threshold = 4	$\frac{425}{7+425} = 0.984$	$\frac{467}{43+467} = 0.916$	$\frac{425}{467+425} = 0.476$
Threshold = 5	$\frac{305}{127+305} = 0.706$	$\frac{40}{470+40} = 0.078$	$\frac{305}{40+305} = 0.884$

Table 6

A BI-RADS threshold of 4 (and below) is too conservative and captures too many false positives. As a result, the model is imprecise and will lead to around 52% of biopsies being unnecessary. This model is therefore poor from a resource perspective. On the other hand, a threshold of 5 returns a precise model, but one which will miss around 29% of malignant masses, meaning this model is poor from a patient safety perspective.

Discussion and Conclusion

Clinical Use

For the BI-RADS models either the TPR or precision is too poor for clinical use. This is not true of the logistic model. With an unnecessary biopsy rate of 24%, and misclassification of malignant masses around 17%, it is much more applicable for clinical use. The markedly better precision vs a typical BI-RADS scoring model, should mean less unnecessary biopsies are carried out. Within the UK, this equates to a cost saving of between £1000-£2000 for each unnecessary biopsy avoided (Leeds Teaching Hospital, 2020). Note however, model performance needs to be reassessed over a test set of British patients prior to roll out, to ensure it generalises sufficiently well from the (presumably) German cohort it was developed on.

Missing 17% of malignant masses may still be deemed unacceptably high so a more conservative threshold may be required. A TPR of 90% can be achieved whilst maintaining a precision of 70%. Although this would mean 30% of biopsies are unnecessary, this is still better than the threshold 4 BI-RADS model, which sits at 52%.

Methodology Limitations

Ideally pre-processing should not involve the dataset the model gets evaluated over. This can exaggerate performance as the model has been fit using the very data it is being evaluated over. Orange has no easy way of separating the data before pre-processing, so this issue is present in the pipeline. Furthermore, standard practice is to reserve a separate test set for assessing final model performance. This is because the choice of hyperparameter will be slightly biased towards the training set, so is not completely 'optimal' in the sense of giving the best generalisability. This means *validation* set performance (i.e., performance over the dataset used to tune hyperparameters) is a bit better than actual 'real-world' performance. In this case only validation set performance was measured, so again performance might be inflated. For the logistic model only $'L_2'$ regularisation was considered, which corresponds to the normal prior introduced. Orange also provides the option of using L_1 regularisation, which corresponds to a Laplacian prior. This prior pushes some parameters to 0 and *generally* this parsimony comes at the expense of slightly worse performance than L_2 regularisation. However, this could have been confirmed by rerunning the model with L_1 regularisation.

Model Performance

Performance of other models across the same dataset is summarised below.

Model	Notes	AUC	TPR	Precision	Our model's precision at this TPR	Methodology Notes	Source
SVM	Polynomial Kernel	0.83	0.85	0.78	0.74	70:30 train: test split. Imputation via classification/regression tree for categorical/continuous features respectively. Feature selection method unclear. Classification threshold = 50%	(Mokhtar and Elsayad, 2013)
ANN	4 layers with 12:30:18:1 nodes.	0.81	0.85	0.76	0.74		
Decision Tree	Splits based on chi-squared tests	0.81	0.86	0.73	0.73		
Multivariate Adaptive Regression Spline (MARS) Algorithm	For an overview of MARS models see Friedman, 1991.	0.88	0.9	0.74	0.7	80:20 train: test split. Imputation via KNN (K=5). No features removed. Classification threshold unspecified. Assumed default = 50% based on models R documentation: https://cran.r-project.org/web/packages/earth/earth.pdf)	(Pfob et al, 2022)

Table 7

At a classification threshold of 50%, both the SVM and MARS models have a notably better precision-recall trade-off than our model. Given the inflated nature of our performance metrics, it is likely that *all* these models will outperform ours on unseen data. The best model comes from Pfob et al, who achieved a TPR of 90% whilst maintaining a precision of 74%. They ran five other algorithms on the data, including a regularised logistic regression model and note 'the five algorithms showed equally high performance on the [test] set' (only the MARS model had the confusion matrix provided to compute the summary metrics). A review of their methodology suggests ways to improve our performance. In particular, they:

- Apply a Box-Cox transform to age to induce a more normally distributed feature. Currently age looks to have some left skew. Although normality is not an assumption of logistic regression it can improve performance (Osborne, 2010).
- Don't remove features. In our case, features such as margin = 3 (obscured) may naturally have a lower importance simply due to rarity. Since most patients don't have obscured masses, splits on this feature will typically only occur deeper in the tree on nodes with fewer patients. It may therefore have been discarded due to a low 'importance' value, despite being a predictive feature. Instead, they pool less frequent features (present in $\leq$ 10% of patients) into an "other" category. This limits the ranking problem above, by pooling 'rarer' features into a more prevalent feature.
- Allow for more hyperparameter tuning. For logistic regression, L_1 and L_2 regularisation correspond to powers in the penalty term of the loss function, of 1 and 2 respectively. Taking this power term as an additional hyperparameter α potentially allows for more optimal w's to be selected.

References

Chhatwal, J., Alagoz, O. and Burnside, E.S., 2010. Optimal breast biopsy decision-making based on mammographic features and demographic factors. *Operations research*, *58*(6), pp.1577-1591.

Elter, Matthias. Schulz-Wendtland, Rüdiger, 2007. *Mammographic Mass Data Set.* [Online]. Available from: http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Ranganathan, J., Hedge, N., Irudayaraj, A.S. and Tzacheva, A.A., 2018, July. Automatic detection of emotions in Twitter data: a scalable decision tree classification method. In *Proceedings of the Workshop on Opinion Mining, Summarization and Diversification* (pp. 1-10).

Chapelle, O. and Li, L., 2011. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, *24*.

Leeds Teaching Hospital, 2020. *Interactive Costing Tool (iCT) Investigation and Intervention Tariff 2020/21.* [Online]. Available from: https://www.leedsth.nhs.uk/assets/71432c14fa/NIHR-2020-Investigation-and-Intervention-Tariff-1-v2.2-1.pdf

Friedman, J.H., 1991. Multivariate adaptive regression splines. *The annals of statistics*, *19*(1), pp.1-67.

Mokhtar, S.A. and Elsayad, A., 2013. Predicting the severity of breast masses with data mining methods. *arXiv preprint arXiv:1305.7057*.

Pfob, A., Lu, S.C. and Sidey-Gibbons, C., 2022. Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison. *BMC medical research methodology*, *22*(1), pp.1-15.

Osborne, J., 2010. Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, *15*(1), p.12.