APPLYING MACHINE LEARNING REGRESSION AND CLASSIFICATION TO PREDICT RELAPSE FREE SURVIVAL AND PATHOLOGICAL COMPLETE RESPONSE FOR BREAST CANCER PATIENTS

Ayush Ranjan (20601369), Amit Kumar (20594315), Rohan Sood (20600751), Dominic Mwita (20442271)

School of Computer Science University of Nottingham
Computer Science
Nottingham

ABSTRACT

Different regression and classification models were trained to predict relapse free survival (RFS) and pathological complete response (PCR), respectively to find the best model for each task. The paper used breast cancer patients dataset from The American College of Radiology Imaging Network (I-SPY 2 TRIAL). To predict RFS of patients XGBoost model delivered the best results compared with random forest, support vector machines. Classification problem is the case of imbalanced classes where the class of interest, patients that achieved PCR is the minority class. Thus, the interest was to achieve high balanced classification accuracy. En ensemble of voting classifiers consisting logistic regression, random forest and XGboost outperformed individual models by achieving 61.4 percent balanced accuracy, albeit at the cost of slightly higher variance. More data or relevant features may be necessary for the trained models to achieve better prediction outcomes.

***Index Terms*—** Imbalanced classes, logistic regression, random forest, support vector machines, XGBoost, balanced accuracy score

1. INTRODUCTION

This paper applies supervised machine learning algorithms on breast cancer patients public dataset extracted from The American College of Radiology Imaging Network to predict pathological complete response - PCR (classification) and relapse free survival - RFS of the patients for stratification before provision chemotherapy treatment. The aim is to build different machine learning models and select the best one for the task given nature of dataset.

Similar studies have been conducted, applying different methods with varying performance in prediction based on selected metrics. Regression models such as General Linear Regression, Gradient Boosted Machines (GBM) and Random Forests were applied in [1] to predict survival time for lung cancer patients in months using data from Surveillance, Epidemiology, and End Results (SEER). The methods achieved varying performances in different conditions, for instance length of survival time–implying that no best model for every task–the choice and performance of an algorithm depends on many factors and specifics of the problem itself.

The classification task aimed at predicting PCR status is a binary in nature because the target has two outcomes–achieved PCR (positive outcome) or not. The dataset contains imbalanced classes with respect to target variable PCR where it contains around 22 percent of positive samples, the rest are negative samples. This makes prediction accuracy misleading because the model can do a good job in predicting the majority classes but poorly on minority class because the model tends to over-fit majority class [2, 3]. Imbalanced or skewed data is among biggest challenges in machine learning classification in real-world scenario as most of data is imbalanced and the cases of interest is from minority class. Various workarounds have been proposed in literature to enhance model prediction. As in [2, 4], these include sampling based approaches of oversampling minority classes, undersampling majority class, cost sensitive-learning and generating synthetic samples for minority class known as Synthetic Minority-Oversampling Technique (SMOTE). This study adopts cost-sensitive learning which works by assigning higher weight on misclassification errors for instances belonging to the minority class.

2. METHODS

2.1. Data Preprocessing

The dataset used contains 400 instances in with 5 missing cases in the target variable–PCR for classification, these were then dropped to minimize the risk of misleading results by imputing them. Before further processing, 20 percent of instances in dataset was reserved for testing to the so called avoid data snooping bias which provide a semblance of good performance because data from the test set leaked [5]. Train-test split was stratified to ensure that the training set contains

positive instances in the same proportion as in the original dataset. For regression task–predicting Relapse Free Survival (RFS), all 400 cases were preserved as no missing values were present in target variable.The missing cases for features were imputed using k-nearest neighbour method. Different approaches were used to transform features depending on the model being trained.

Both feature selection and dimensionality reduction methods were applied to avoid the curse of dimensionality where by data points are sparse in the feature space leading to over-fitting [5]. ANOVA test was used to select categorical features for regression task while for continuous features (extracted from image), Principal Components Analysis (PCA) was applied. Nevertheless, correlation between continuous features and target RFS was very weak suggesting non-linear relationship. To handle outliers–which were numerous in image-based data–two methods were applied: robust scaler and replacement of feature values with median for values falling above or below 3 times the interquartile range. The robust scaler from sklearn library works by substracting the median from data and dividing by interquartile range. These approaches were used depending the model being trained.

2.2. Models

To predict RFS three non-linear methods were applied, namely Random Forest (RF), Support Vector Machines (SVM) and XGBoost regressor. Their performance was evaluated using mean absolute error (MSE). For PCR classification RF, SVM, Logistic Regression and XGBoost were trained with cost-sensitive learning methods to handle imbalanced classes. SMOTE was attempted but resulted in models with high variance, hence dropped. Cost-sensitive learning was chosen because it handles misclassification better than minority oversampling which is prone to overfitting and majority undersampling which leads to information loss [6].

Random Forest

Random forests algorithm is an ensemble of decision trees which grows numerous decision trees using different subsets of features and training examples, and the final prediction is obtained by aggregating predictions from individual decision trees. This makes random forests are amongst the most powerful machine learning algorithms can decipher complex non linear patterns which makes them prone to overfitting. Random forests have inbuilt feature selection and can handle large datasets [5].

For predicting RFS two random forests models were trained, one with scaled data and the other without scaled data. Model hyperparameters that were optimized by grid search cross-validation. The best estimator for tree built with scaled data was with 200 trees, maximum number of features of 6, and bootstrapped samples. For the one with unscaled data, optimal results were 100 trees, 10 maximum features with bootstrap sampling applied.

In classifying whether patients achieved PCR, a cost-sensitive balanced random forest with bootstrapping applied was built where by errors on misclassification of positive samples (PCR=1) are more penalized to improve model classification on minority class. Feature selection was applied to reduce dimensionality of data despite the fact that random forests have inbuilt feature selection. Three Categorical features were selected by chi-square test and 5 percent significance level. Continuous features were chosen by ANOVA test (equivalent to t-test on binary data) where 10 out of 108 features were chosen. Parameters of the model were optimized using cross validation and grid search. The best model was with 100 trees, maximum tree depth of 2, and maximum number of features of 8.

Support Vector Machines

Support Vector Machines for regression works by ensuring most of the instances fall within the hyperplane while minimizing errors for those lying outside the plane. Given a continuous target variable y and feature matrix x, SVM regression finds weights w and biases b that optimizes:

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\zeta_i \tag{1}$$

Subjet to

$$|y_i - (w^T x + b)| < \epsilon + \zeta_i$$

Where C is a regularization hyperparameter that controls the width of the hyperplane and ζ is a slack term that represents instances outside the hyperplane. The increasing C means tolerating more errors and reducing the hyperplane width. SVM is a very powerful algorithm that can perform complex non-linear regression and classification tasks but it is computationally expensive, thus suited for small and medium sized datasets.

Non-linear Support Vector Machines (SVM) was trained because exploratory data analysis did not support linear models as evidenced by weak correlation between features and Relapse Free Survival (RFS)–the target variable–the highest correlation coefficient was less than 0.25 in absolute value. LASSO with small penalty turned all feature coefficients to zero. Thus, dimensionality reduction using PCA was performed on 107 continuous features extracted from image data and retain 2 components which captures 99 percent of their variance. Based on ANOVA test categorical features were also not found to be good predictors for RFS with insignificant p-values at 5 percent level. Hyperparameter tuning by grid search suggested the best model with 'rbf' kernel, regularization parameter C=1, and margin of error tolerance (epsilon) of 0.2.

SVM trained for PCR classification with cost sensitive-learning and SMOTE had high variance hence dropped.

XGBoost
XGboost is one of most powerful algorithm in machine learning that is computationally efficient and provide best prediction performance. It is an ensemble tree-based algorithm that works by sequentially adding trees to minimize errors made by previous tree. The trees individually are weak predictors but combined together are very powerful predictors. The subsequent trees instead of predicting target are fitted to errors made by predecessor and the final prediction is obtained by summing prediction of all trees in an ensemble [7]. The algorithm works by fitting ensemble of k trees using n instances and x_i features by minimizing the loss function in equation 2.

$$L^t = \sum_{i=1}^{n} l(\hat{y_i}^{t-1}, y_i) + f_t(x_i)) + \Omega(f_k) \quad (2)$$

Where $\hat{y_i}^t$ is prediction of example i at iteration t, f_t is a tree, k is number of trees the term Ω penalizes complexity of the model to prevent overfitting. The algorithm greedily add tree f_t that improves the model. Nonetheless, the algorithm is complex and has so many hyperparameters to tune to obtain the best results.

For predicting RFS, an XGBoost ensemble of 100 trees optimized by cross validation with grid search, other hyperparameters were proportion of features to build trees (0.2), learning rate (0.05), maximum tree depth (4) and regularization term for individual tree complexity - gamma (0).

XGBoost with 100 trees for classification was trained considering class imbalances by assigning higher weight to the positive class equal to the ratio of negative to positive instances in data which was 0.78/0.22. Other model hyperparameters were optimized by cross-validation and the best model was the one with maximum fraction of features to build trees (0.15), learning rate (0.05), and maximum tree depth (3).

Logistic Regression
Logistic regression is among widely used and simple classification algorithm that is primarily used for binary classification but can be extended to handle multi-class classification. The logistic model is expressed as:

$$p_i(x) = \frac{1}{1 + e^{-(wx+b)}} \quad (3)$$

The function $p_i(x)$ computes the probability that a sample i belongs to a positive class, RFS=1 in case of this study. The threshold used to assign samples to a positive class is usually $p_i >= 0.5$ but it can be changed based on specific context of the problem of interest [8]. The intrinsic parameters w and b are optimized by minimizing the binary cross entropy loss in Equation 4. C is regularization parameter to penalize weak predictors.

$$L = \sum_{i=1}^{n} -[y_i ln(p_i) + (1 - y_i) ln(1 - p_i)] + C||w|| \quad (4)$$

The binary logistic regression to predict PCR was fitted using cost-complexity learning to improve classification on positive samples. The model was fitted using 4 categorical features selected by chi-square test and 9 continuous features selected by ANOVA test.

Voting classifier ensemble
Voting classifiers works by aggregating predictions from a group of classifiers (ensemble), individual classifiers may be weak learners when considered separately but together they become strong learners giving better predictive performance. This is, however, conditioned upon, independence of individual classifiers, because it allows them to make different mistakes. This paper adopts soft voting ensemble because it provides better results– compared with hard voting ensemble–as more weight is assigned to votes with high probability [5]. An ensemble of Logistic Regression, Random Forest, and XGBoost was used to form a voting classifier.

3. RESULTS

3.1. Regression

As shown in Table 3.1 there are small differences in mean absolute errors across trained models for both cross validation and independent test set. Nevertheless, XGBoost model produced more stable results with lowest cross validation standard deviation compared to other models. Thus XGBoost was selected among the four models presented. Noteworty, Random Forest 1 was trained with scaled data while Random Forest 2 was trained with unscaled data. XGBoost model produced the best results among selected models by having lowest standard deviation of cross-validation MAE.

Model	Cross Validation	SD	Test
Random Forest 1	20.124	2.400	25.199
Random Forest 2	20.093	2.178	25.094
SVM	20.118	2.661	25.559
XG boost	20.207	1.967	25.361

Table 1. Mean Absolute Error (MAE) for trained models

Results obtained suggest that possibly the features used to predict RFS may not be the best for the purpose. Thus, using different features may lead to better results.

3.2. Classification

To predict whether breast cancer patients achieved PCR results from estimated logistic regression, random forest and XGboost vary slightly in terms of balanced accuracy score achieved. The models were also found to be relatively stable having small standard deviation of cross validation scores. The three models when combined together in an ensemble

achieved the highest balanced classification accuracy for both cross-validation and test sets, however, at the cost of slightly higher variance as shown in Table 3.2. Therefore, a voting ensemble was chosen as the best classifier.

Model	Cross Validation	SD	Test
Logistic	0.615	0.028	0.593
Random Forest	0.609	0.070	0.587
XGBoost	0.610	0.086	0.591
Voting Ensemble	0.653	0.067	0.614

Table 2. Balanced accuracy for trained models

Because the models were optimized on balanced classification accuracy metric, other metrics especially precision deteriorated because balanced classification accuracy is an average of recall of both classes $PCR = 1$ and $PCR = 0$. Optimizing balanced classification accuracy, however reduces overall accuracy because models were trained to penalize more misclassification on minority class, Table 3.2.

	precision	recall	f1-score	support
0	0.84	0.76	0.80	62.00
1	0.35	0.47	0.40	17.00
accuracy	0.70	0.70	0.70	0.70
macro avg	0.59	0.61	0.60	79.00
weighted avg	0.73	0.70	0.71	79.00

Table 3. Classification report for ensemble voting classifier

4. CONCLUSION

A dataset containing 400 patients of which 20 percent was retained as an independent test-set was used to build machine learning models to enhance better stratification of patients before prescribing chemotherapy treatment. Supervised machine learning models of Random Forests, Support Vector Machines and XGBoost were trained and compared for predicting RFS for breast cancer patients. The absolute mean error for all three methods was around 20 for validation sets and higher, around 25 for a test set. XGBoost outperformed other methods considered by producing more stable results by having lowest standard deviation of cross validation MAE. The results obtained suggests that the given clinical and image-based features may not be the best for predicting RFS. With more informative features and/ or data, prediction can be improved. For predicting patients response to initial treatment (PCR), class imbalances were considered because the dataset contained about 22 percent of patients who achieved PCR and the remaining 78 percent did not. As such, the models were trained with cost sensitive learning which penalizes more on errors. Model performance was evaluated based on balanced accuracy score which makes more sense when the distribution of classes is skewed. An ensemble of voting classifiers outperformed individual models that were aggregated by achieving highest balanced classification accuracy, but with slightly high variance. The built model may not be the best for the task, more data may be required to train better models and if more informative could be available then performance can be increased considerably.

5. REFERENCES

[1] James A Bartholomai and Hermann B Frieboes, "Lung cancer survival prediction via machine learning regression, classification, and statistical techniques," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 632–637.

[2] Vimalraj S Spelmen and R Porkodi, "A review on handling imbalanced data," in *2018 international conference on current trends towards converging technologies (ICCTCT)*. IEEE, 2018, pp. 1–11.

[3] Yoga Pristyanto, Irfan Pratama, and Anggit Ferdita Nugraha, "Data level approach for imbalanced class handling on educational data mining multiclass classification," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 310–314.

[4] S. Sukhanov, A. Merentitis, C. Debes, J. Hahn, and A. M. Zoubir, "Bootstrap-based svm aggregation for class imbalance problems," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 165–169.

[5] Aurélien Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*, " O'Reilly Media, Inc.", 2022.

[6] Somiya Abokadr, Azreen Azman, Hazlina Hamdan, and Nurul Amelina, "Handling imbalanced data for improved classification performance: Methods and challenges," in *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, 2023, pp. 1–8.

[7] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 2016, KDD '16, ACM.

[8] Andriy Burkov, "The hundred-page machine learning book," 2020.

Fig. 1. Contribution table

Name	Data pre-processing (10%)	Feature selection (25%)	ML Method development (25%)	Method evaluation (10%)	Report writing (30%)
Ayush Ranjan	20	20	20	20	20
Amit Kumar	20	20	20	20	20
Rohan Sood	20	20	20	20	20
Dominic Mwita	20	20	20	20	20
Sanyog Chavhan	20	20	20	20	20