Comparisons of Classification Machine Learning Models

tgkh12
Durham University Dept. of Computer Science
Durham, United Kingdom
tgkh12@durham.ac.uk

***Abstract*—Since the end of 2019, the coronavirus (COVID-19) has been spreading all across the world causing one of the worst global pandemics in history. In light of this, a group of researches, Bo Xu et al. have collectively gathered epidemiological data from this outbreak and uploaded it to their GitHub repository. The models trained on this dataset are evaluated using 3 different metrics, performance, accuracy and computational time. The research has shown that there is no overall best performing/most accurate model, but were the best in certain criteria.**

***Index Terms*—logistic regression, GBM, SVM, machine learning, COVID-19, ROC, PRC, AUC, SMOTE**

I. Introduction

The use of classification methodology has been growing in the medical space to assist doctors in giving early-diagnosis of a patient. By way of individuality, every person differs in some physical or socio-economic way such as age, sex, nationality, etc which makes it difficult to make accurate predictions about their condition or outcome quickly. Hence, the aim of this project was to implement and compare 3 different machine learning models on a medical-based classification problem based on the aforementioned open source dataset. This document will first outline the problem of predicting the outcome of a COVID-19 patient, followed by detailing the methodologies applied to train 3 widely usedc predictive models - **Logistic Regression**, **Gradient Boosting Machine**, and **Support Vector Machine**. The document will conclude by comparing these models based on their performance and accuracy on this dataset.

II. Problem Framing

On first glance of the dataset, some potential aspects to further look into were:

- Predicting how soon a patient is confirmed to have COVID-19
- Predicting a patient's recovery period
- Predicting a patient's condition (Dead or Alive)

The third problem stood out as there was a recent news in February about researchers from University of Copenhagen in Denmark that have managed to create a medical AI system that is able to predict a patient's life or death outcome at about 90% certainty [9]. Conversely, there have been quite a lot of research papers done on using ML to predict COVID-19 results and recovery periods in comparison to models used for vital status prediction.

In this case, predicting that a COVID-19 infected person of a certain sex or age while having some chronic disease will die for example will alert doctors to provide earlier intense care and make more informed decision. However, there are so many classification models that are available currently and it is hard to make an optimal choice without initial experimentation. Hence, the work documented in this paper focuses on observing and comparing the predictive behaviour of 3 models on the aforementioned binary classification problem.

III. Data Preprocessing

A. Data Removal and Substitution

The COVID-19 epidemiological dataset used was really large as it contained over a million rows of data entries. However, many of its columns were either not relevant to the problem (such as administration columns) or contained a large number of empty values which would affect the accuracy and the performance of the model. "Fig 1" shows the pattern of missing values occurrences for each column.

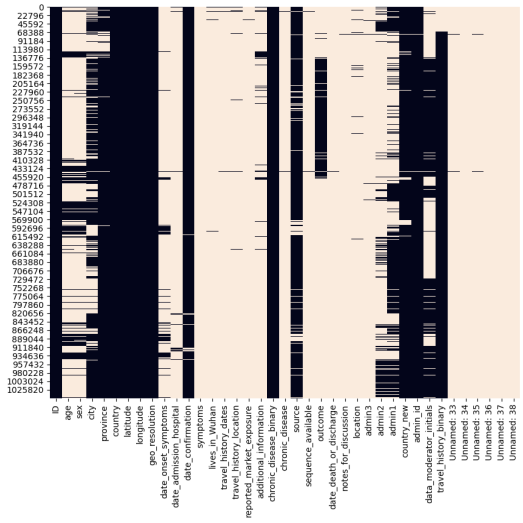


Fig. 1: Heatmap of missing values in each column (Black = Non-empty)

There were **8** features left after the removal of unnecessary columns. Rows which contained missing values in the remaining columns were dropped except for columns *symptoms* and *chronic disease* where their empty values were replaced with a string in the format '*unknown_*{*name of feature*}', e.g. *unknown_symptoms*. While these columns contained a significantly large number of missing values, they were still considered essential in determining a patient's outcome as, for example, existing symptoms or long term medical conditions could contribute towards the effects of COVID-19 leading to a patient's death.

B. Data Standardization

The feature columns age, symptoms and chronic diseases contained unformatted data. The age column had a combination of a single age (20, 30, ...), an age range (50-55) or ages in months (11 months) whereas the other two features had inputs in the form of long strings separated by ":", i.e. "asthma:hypertension:atherosclerosis:coronary artery disease". To resolve this issue, a function was made to ensure all *ages* were non-ranged or in string form, hence, individual ages were converted to floats, mean age was found for age ranges and ages

in months were converted to years. Moreover, one-hot encoding was used on *symptoms* and *chronic diseases* with manually created unique categories (grouped) instead of dummy variables as to hopefully reduce the risk of multicollinearity [3]. Additionally, the countries were also one-hot encoded.

For this project, the outcome column is converted to binary too (0 = Dead, 1 = Alive) as one of the models utilised in this project, logistic regression, is built mainly for binary classification problems.

IV. Model Selection

A. Logistic Regression (LR)

LR is a commonly used statistical classification model to predict an outcome where its probability (log odds) is determined by a series of potential predictors [4]. The equation for the probability is as follows:

$$log[p/(1-p)] = \beta_0 + \beta_1 X_1 + ...\beta_i X_i$$

where p is probability, β_0 is an intercept term and $\beta_i X_i$ is a product of a coefficient with the value of an independent variable [4]. This statistical modeling technique works really well when dealing with binary classification in the medical field where outcomes of interests are mutually exclusive [4], in this case life and death. However, this idea of a linearity relationship for LR is also an issue as it acts on the assumption that there is no multicollinearity between the independent variables [1] - this will potentially be an issue as the dataset has a lot of features. Moreover, the same reason also leads to nonlinear relationships between outcomes and predictors going unnoticed [4]. Despite this, LR is still suitable for this experiment as, due to the limitation of time, it is able to train on huge datasets such as this epidemiological dataset on regular computers fairly quickly [4].

B. Gradient Boosting Machines (GBM)

GBM is a version of the Random Forest model that employs the *boosting* technique instead of *bagging*. This means that it builds each decision tree model sequentially, allowing it to correct mistakes made by the previous models in order to improve discrimination accuracy [13]. Furthermore, the model at each step will add new models which minimizes a loss function which, overall, will minimize training error. Hence, the model is quite robust and is not easily influenced by outliers or unrelated features in the dataset [5]. However, the number of iterations in tree base model generation needs to be controlled as overfitting may still occur due to excessive iterations, resulting minor error or changes in the training data to be exaggerated [13]. This can cause a reduction in prediction accuracy. As GBM is flexible, there are many crucial hyperparameters that have to be throughly tuned which results in a larger search grid [12].

C. Support Vector Machines (SVM)

SVMs, unlike Naive Bayes, are linear binary classifiers that do not rely on probability [8]. Moreover, they are robust models which are able to generalize data well and not fall into multilocal minima easily [7]. However, SVMs do take a long time to train, especially on large datasets [2]. Hence, it is not optimal for proper hyperparameter tuning for this experiment as it will take too much time and the impact is hard to visualize. Despite these limitations, SVM is a good model to use in general when doing comparisons as it works relatively well on any type of data and dimensonality of dataset.

V. Experiments

A. Experimental Protocol

The cleaned and preprocessed dataset was split 80:20 (based on the Pareto Principle for a good initial division) for the training set and testing set respectively. As shown in "Fig 2", it is to note that there is a huge discrepancy (about 25 times difference) in the outcome columns of the dataset.

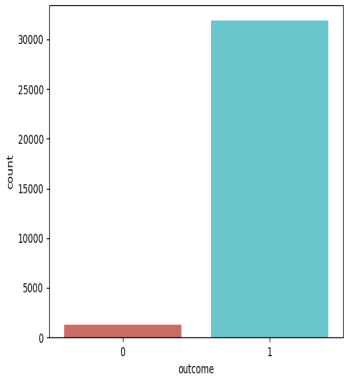


Fig. 2: Number of victims alive or dead

While this is an undersampling of the minority class, it is actually a good sign illustrating that the virus is not too deadly (will have to find a better way to elaborate this). Synthetic Minority Oversampling Technique (*SMOTE*) is used to synthesize new examples to balance the outcomes without adding additional information to the training set for the model. In this project, the Python **imbalanced-learn** module employs k-Nearest Neighbours (*kNN*) to select a random instance of a victim who died due to the virus out of the k nearest entries of the same outcome to create a synthetic example at a randomly selected point in between the two instance in the feature space. "Fig 3" shows the distribution of before and after the use of SMOTE:

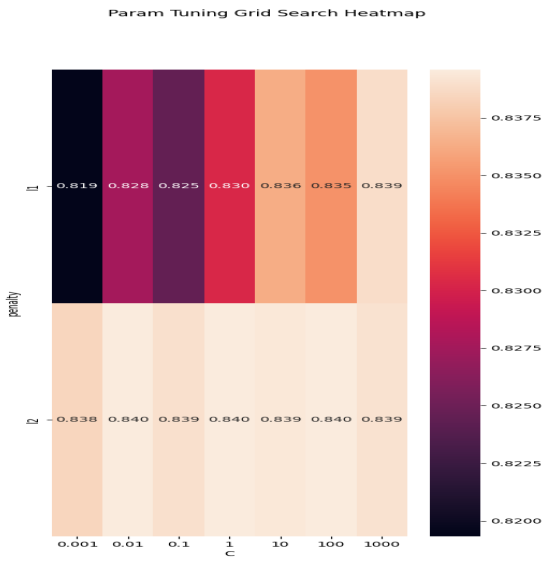
```
X_train.shape (26531, 75), y_train.shape (26531,)
Class distribution in training set:
outcome
1          25508
0           1023
dtype: int64
Percentage of positive class samples: 96.1441332780521

----------------------------------------------------------------
X_train.shape (50970, 75), y_train.shape (50970, 1)
Class distribution in training set:
outcome
0          25485
1          25485
dtype: int64
Percentage of positive class samples: 50.0
```

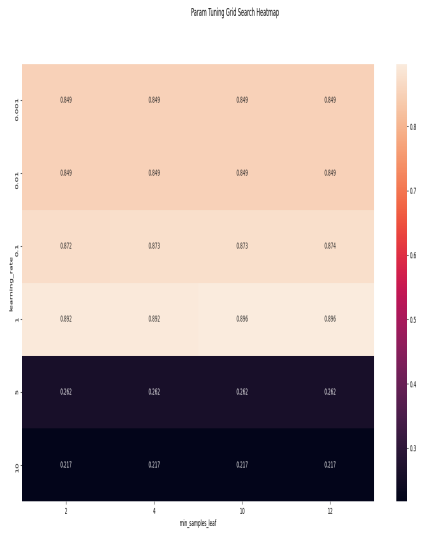
Fig. 3: Distribution of outcome in training set (Top: Before; Bottom: After)

The experimentation will involve the three models being trained on both these instances of training data to analyse the results. Furthermore, all models used are from the Python **Scikit-learn (sklearn) module** to ensure consistency of internal functionality and practices used when training and testing on the dataset. Moreover, optimal hyperparameter tuning was also conducted on the logistic regression and gradient boosting model using sklearn's **Grid Search Cross Validation**.

This is an exhaustive search method which splits the training set into training and validation data to repeatedly try a given dictionary of parameters until an optimal one is found and applied. "Fig 4"



(a) LR



(b) GBM

Fig. 4: Hyperparameter Tuning using GridSearchCV

shows that the LR performs more optimally using L1 penalty and higher C while GBM is better on 1 (preferably) or lower learning rate with at least 10 samples per leaf node. On the other hand, cross validation was not done for the SVM due to the limitation of time and processing capabilities as the dataset was large.

ROCs and *AUCs* [6] were calculated for all the models to evaluate their ability in classifying the outcome of a COVID-19 patient in the balanced training dataset. As the training set with imbalanced classes was still experimented on in this research, a classification report was created to evaluate a model's accuracy based on precision and recall which is a more accurate metric in these circumstances. Using these values, the harmonic mean (F1-Score) and area under precision recall curve can be calculated. These values tend to provide more intuitive and informative visualization for these types of datasets as they express the susceptibility of classifiers with clear visual cues [11].

To improve the estimated performance of our models, a **10-fold stratified cross validation** methodology was also used that runs the models on the dataset 6 times to generate a set of results from the testing data. This cross validation was done on a model's accuracy, precision, recall and ROC AUC. This will aid the understanding of our models' performance range.

B. Results

The initial model accuracy scores are shown in "Table I" - grid search best score relates to the highest accuracy of the model achieved based on a randomly sampled validation set whereas the prediction accuracy is based on its performance on the unseen testing set. We can deduce from this result that GBM and SVM can be statistically considered similar in their prediction abilities while Logistic Regression lacks behind. In all tables I and II, the number outside the bracket corresponds to the mean result after the 10-fold stratified cross-validation whereas the number inside the bracket is the standard deviation of said result.

TABLE I: Model Accuracy Scores

Model Type	Grid Search Best Score	Grid Search Best Score (w/ SMOTE)	Prediction Accuracy on Testing Dataset	Prediction Accuracy on Testing Dataset (w/ SMOTE)
LR	0.849	0.859	0.852 (0.012)	0.8568 (0.014)
GBM	0.966	0.896	0.964 (0.002)	0.951 (0.001)
SVM	-	-	0.963 (0.004)	0.966 (0.002)

Similarly, the box plots shown in "Fig 5" helps to better illustrate the range of accuracy of each model with and without the use of SMOTE on the training data. It can be seen that, out of the 10 folds and 6 repetitions on the unbalanced training set, the GBM has the most precise classification of the outcomes whereas SVM had the best predictions when the data was balanced with SMOTE.

"Fig 6" is an instance of ROC for both types of the training set, however, with cross-validation, the range of area under multiple curves generated can be visualized in "Fig 7" which better summarizes each model's performance.

TABLE II: Classification Report (W/O SMOTE)

Model Type	Precision	Recall	F1-Score	PRC AUC
LR	0.990 (0.004)	0.860 (0.017)	0.920	0.994
GBM	0.970 (0.004)	0.992 (0.005)	0.981	0.995
SVM	0.964 (0.002)	0.998 (0.002)	0.981	0.994

The classification report containing the precision and recall values and their respective combined metrics are shown in "Table II". It can be noted that all 3 models had exceptionally good accuracy based on the F1-Score but LR fell short as it did not, on average, catch as high percentage of positive cases as GBM and SVM.

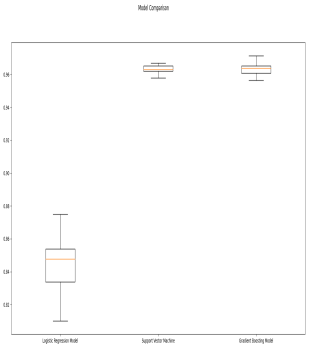
The runtime metrics for each model is shown on the Table III; this includes the training times and cross-validation times on the testing set. These were all performed on the same Linux laptop with a 16-core Intel i7 10th Gen. processor of speed 2.2GHz and 16GB RAM.

TABLE III: Computational Time Comparisons

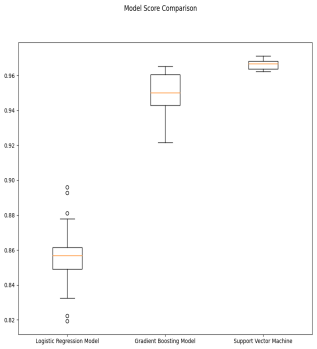
Model Type	Training Time (s)	Repeated Stratified 10-Fold CV (s)
LR	42.95	32.97
GBM	315.80	13.77
SVM	19905.90	2570.61

C. Discussion

It is worth noting that all 3 models performed really well and made quite precise classification as the discrimination accuracy was above 80% while F1-Score were all above 90%. However, it must

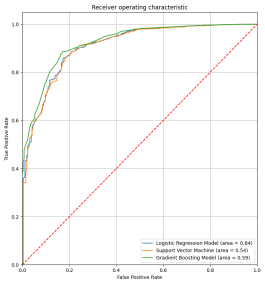


(a) Without SMOTE

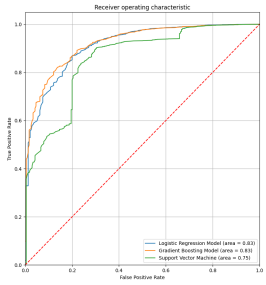


(b) With SMOTE

Fig. 5: Model Accuracy Scores



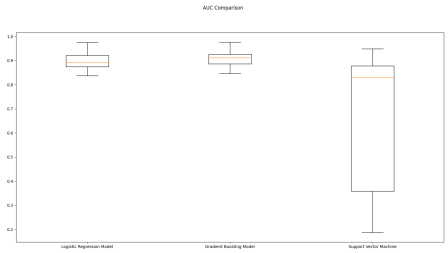
(a) Without SMOTE



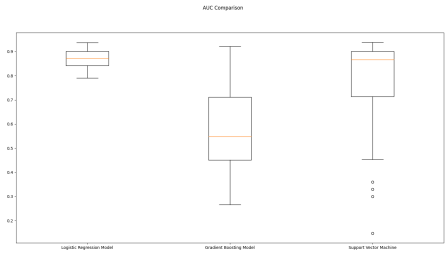
(b) With SMOTE

Fig. 6: ROC Comparison

be taken into account that the testing set itself also has a higher frequency of the majority class outcome and thus the bias developed by the unbalanced training set can influence the accuracy to be good as there is a higher chance for the model to guess that a patient would survive. This bias could be one of the reasons behind the high accuracy scores seen on training set without SMOTE. Furthermore, the use of hyperparameter tuning would have also played a small role in mitigating the effects of an unbalanced training set towards the models, aside from SVM. Based on the boxplots in figures 5 and 7, LR seemed to be the least influenced by the imbalance dataset as its accuracy and AUC ranges are generally small but high in value (on both the undersampled and oversample dataset). This is despite it being the least well-performed model in this experiment based on the mean data in the tables. Conversely, the findings of this experiment suggests that GBM and SVM were statistically really similar in terms of their classification performance and accuracy - looking into it more



(a) Without SMOTE



(b) With SMOTE

Fig. 7: ROC AUC Comparison

specifically, GBM is more precise but has weaker recall than SVM. However, they are more sensitive to the balance of class distributions in the dataset as evident from the ranges shown in boxplots.

Nevertheless, the gradient boosting decision tree model was the best performing model out of the 3 based on precision but logistic regression was the best for its stability. Furthermore, these two models have also outperformed SVM in terms of computation time which makes them more suitable in the medical field due to the large amount of existent medical data and the need for models to learn new data quick to make accurate predictions in a timely manner.

VI. Conclusion

In summary, the work carried out in this paper was to conduct an experimental study to statistically compare the performance and accuracy of 3 machine learning models: logistic regression, gradient boosting machine, and support vector machine to classify the vital status of a COVID-19 patient based on an epidemiological dataset. Regression models and decision trees are widely used in the medical field and the performance, accuracy and computational time that LR and GBM had are certainly evident of their suitability. The current study has focused on predicting a binary outcome. Future works could include either a multi-class classification (i.e. alive, dead, critical condition) or determining the probability of a certain prediction/outcome. Besides that, more techniques in dealing with imbalanced dataset could be experimented on such as setting thresholds, one class learning, cost sensitive learning, etc [10] or different ratio of dataset splits can be tested on (e.g 70:30, 50:50, etc.) for a more thorough model comparison to mitigate overlook the bias present. Overall, this research and experimentation done has exposed me to the practical side of machine learning and the capabilities and wide ranges of models available for various tasks, classification problems in the medical field especially. Furthermore, I also have a better grasp of implementing LR, GBM and SVM models and cleaning/preprocessing datasets and fixing imbalanced classes.

REFERENCES

[1] A. R. Rout, "Advantages and Disadvantages of Logistic Regression." GeeksforGeeks. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ (accessed April. 29, 2021)

[2] C. Chen, "Advantages and Disadvantages of SVM and NRWRH in Drug-gene Interaction Prediction," in International Conference on Industrial Technology and Management Science, 2015, pp. 1030.

[3] I. L. Arevalo et al, "A Memory-Efficient Encoding Method for Processing Mixed-Type Data on Machine Learning," in Entropy, 2020. [Online]. Available at: https://doi.org/10.3390/e22121391

[4] J. V. Tu, "Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes," in Journal of Clinical Epidemiology, Nov. 1996. [Online]. Available at: https://doi.org/10.1016/S0895-4356(96)00002-9

[5] J. Ye, J. H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 2061-2064.

[6] K. H. Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," in Capsian Journal of Internal Medicine, Vol. 4, 2013. pp. 627-635.

[7] M. Awad and R. Khanna, "Support Vector Machines for Classification," in Efficient Vector Mahines, 1st ed. Berkeley, California, USA: Apress, 2015, pp. 39-66. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_3

[8] M. Feldman, "Computational Fairness: Preventing Machine-Learned Discrimination," M.S.thesis, Dept. Comp. Sci., Haverford College, Haverford, Penn. USA, 2015. [Online]. Available: http://hdl.handle.net/10066/17628

[9] P. Awasthi, "Covid-19: New AI tool can predict if infected person will die or survive." The Hindu Business Line. https://www.thehindubusinessline.com/news/science/covid-19-new-ai-tool-can-predict-if-infected-person-will-die-or-survive/article33773251.ece (accessed April. 29, 2021)

[10] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," in GESTS International Transactions on Computer Science and Engineering, Vol. 30, 2006. pp. 4-5.

[11] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," in Public Library of Science, March. 2015. [Online]. Available at: https://doi.org/10.1371/journal.pone.0118432

[12] V. Kurama, "Gradient Boosting In Classification: Not a Black Box Anymore!" PaperspaceBlog. https://blog.paperspace.com/gradient-boosting-for-classification/ (accessed April. 29, 2021)

[13] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," in Transportation Research Part C: Emerging Technologies, Sept. 2015. [Online]. Available at: https://doi.org/10.1016/j.trc.2015.02.019