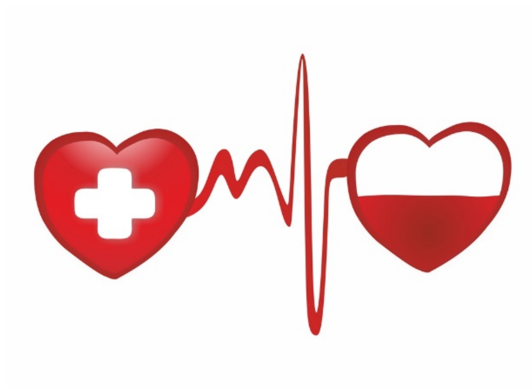
Midterm Project



Warm Up: Predict Blood Donations

Name: Young, Brent

DrivenData Name: bdy3176

DrivenData Public Score: 0.4371

MSDS 454 Section #: 55

Quarter: Summer 2018

Introduction

Problem

The purpose of the midterm project is to analyze blood donation data *(includes information about each donor's history*) from a mobile blood donation vehicle in Taiwan using machine learning methods to predict the probability that a donor made a donation in March 2007. The Blood Transfusion Service Center drives to various universities in the area and collects blood as part of a blood drive. As a result, we ultimately want to predict whether or not a donor will give blood the next time the vehicle comes to campus. However, predicting the probability a person will donate blood can be difficult given that the data provided only contains 4 predictor variables (e.g., Months since Last Donation, Number of Donations, Total Volume Donated, and Months since First Donation) and has only 576 records in the training dataset and 200 in the test dataset.

Significance

The problem is significant/interesting because good data-driven systems for tracking and predicting blood donations and supply needs can help improve the supply chain, ensuring that more patients receive the blood transfusions they need. This is particularly important given that blood remains a critical resource during emergencies, for sickle cell patients, cancer patients, and car accident victims. In fact, according to the American Red Cross, every two seconds someone in the U.S. needs blood, more than 41,000 blood donations are needed every day, and 30 million blood components are transfused each year in the United States.

Applicability to Data Scientists

This real-world data problem is applicable to data scientists because it serves as a great resource for practicing our data science skills and testing out various machine learning algorithms. For instance, since this is a classification problem, it'll provide data scientists the opportunity to practice classification modeling techniques such as Logistic Regression, LDA, QDA, Decision Trees, Bagging, Random Forest, Gradient Boosting Machines, and Neural Network.

Data Exploration

Structure and Description of Training, Validation, and Test Datasets

To help streamline the analysis, both the train and test datasets have been consolidated into one dataset called total. The combined dataset prior to feature engineering consists of 6 variables (includes identification and response variable) and 776 observations. The data was then split into a 48/26/26: train/validation/test so that the validation and test sets have the same amount of observations. For instance, there are 376 training observations (48%), 200 validation observations (26%), and 200 test observations (26%). The training set will be used to fit the models, the validation set will be used to estimate prediction error for model selection, and the test set will be used for assessment of the prediction error of the final chosen model.

Data Mining/ Cleaning

Cleaning

```
       ID      MSLD        NUM     VOLUME       MSFD TARGET_FLAG         DPM      TENRAT
        0         0          0          0          0         200           0           0
       DF  MSLD_bin     REPEAT  SQRT_MSLD   SQRT_NUM SQRT_VOLUME   SQRT_MSFD SQRT_TENRAT
        0         0          0          0          0           0           0           0
  SQRT_DF   LOG_NUM LOG_VOLUME   LOG_MSFD
        0         0          0          0
```

In order to make the data more R friendly, we then added ID to help identify distinct records. We also renamed Months since Last Donation, Number of Donations, Total Volume Donated, Months since First Donation, and Made Donation in March 2007 (response variable) to MSLD, NUM, VOLUME, MSFD, and TARGET_FLAG, respectively. After conducting summary statistics on the dataset. The results showed that there were no missing values, except the 200 TARGET_FLAG NA's in the test set (see top right). As a result, missing value imputation was not conducted.

Summary of Variables, Feature Creation, Measurement Levels, and Standardization

ID, which is used for identification purposes will be ignored. TARGET_FLAG represents our classification response variable of whether a person made a donation in March 2007 (1 = Yes, 0 = No). There is a total of 90 people who donated in March 2007 in the training dataset and 48 who donated in March 2007 in the validation dataset. The other 4 variables are numeric and consist of Months since Last Donation, Number of Donations, Total Volume Donated, and Months since First Donation. Through feature creation, we also created three new numeric variables: Donations per Month (DPM = MSFD/NUM), Ratio of Months since Last Donation to First Donation (TENRAT = MSLD/MSFD), and Donation Frequency (DF = MSFD-MSLD/NUM)). We also created two qualitative variables: MSLD_bin and REPEAT. MSLD_bin creates groups for Months since Last Donation of 0 to 4, 5 to 8, 9 to 12, and 13+, while REPEAT identifies repeat donors (e.g., if a donor donated more than once = 1, else 0). Furthermore, SQRT transformations were conducted on MSLD, NUM, VOLUME, MSFD, TENRAT, and DF, while LOG transformations were conducted on NUM, VOLUME, and MSFD due to the skewness of the variables. Lastly, all variables, except, ID, TARGET_FLAG, MSLD_bin, AND REPEAT have been standardized to have a mean of 0 and standard deviation of 1 in the training, validation, and test datasets. Note: Standardization was conducted after EDA.

Data Visualization

Descriptive Statistics & Univariate Plots

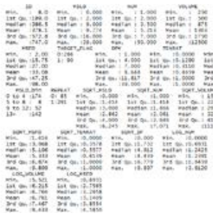


Figure 1

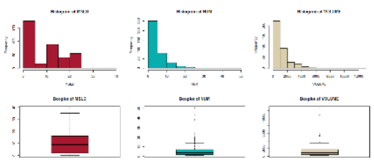


Figure 2

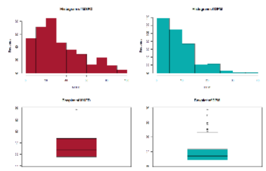


Figure 3

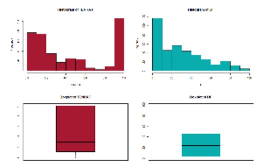
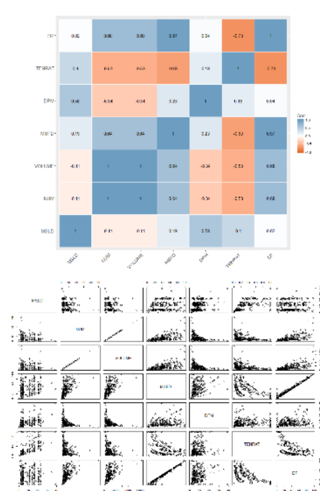


Figure 4

Observations: Figure 1 shows summary statistics of the variables in the training dataset so that we can check for missing values, outliers, distributions, etc. The data shows that 90 people donated in March 2007, while 286 did not donate. Additionally, the average MSLD is 10, NUM is 5, VOLUME is 1303, and MSFD is 33. We also wanted to point out some interesting insights based on the "minimum" of NUM and VOLUME. For instance, the fact that the minimum number of donations is 1 and minimum volume donated is 250cc indicates that they are essentially duplicate variables. This makes sense given that 250cc of blood is the maximum amount of blood that is extracted per visit. The summary statistics also shows that there are no missing values in the training dataset. Furthermore, the histograms and boxplots above illustrates that there are some outliers in all the numeric variables (figure 2 & 3), except TENRAT and DF (figure 4). Furthermore, all the numeric variables have a right skew, except for TENRAT (left side of figure 4), which has a bimodal distribution. These distributions and outliers were validated using quantiles as well. However, I decided not to handle/truncate the outliers because it did not improve my models when I tested them on the validation dataset. In regards to the qualitative variables, most of the donors/non-donors are from 0 to 4 and 13+ range for MSLD_bin and are repeat donors.

Multivariate Plots for Numeric Variables

Figure 5: Correlation and Scatterplot Matrix



Observations: Figure 5 shows a correlation and scatterplot matrix of the numeric variables DF, TENRAT, DPM, MSFD, VOLUME, NUM, and MSLD. This allows us to see which variables may be correlated with each other so that we can gleam interesting insights. The plots show that VOLUME and NUM are perfectly positive correlated, which validates what we saw in our summary statistics. As a result, given the duplicative nature of these two variables, VOLUME was removed from the dataset. Furthermore, the plots revealed strong positive correlations between MSFD vs. DF, which makes sense given that the longer someone has been a donor for, the more frequently he/she will donate.

Figures 6a to 6c: Boxplot of MSLD, NUM, DF vs. TARGET_FLAG

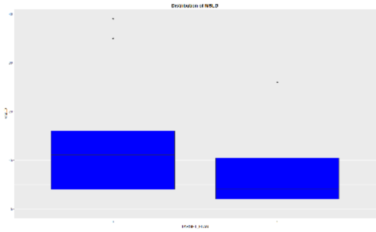


Figure 6a

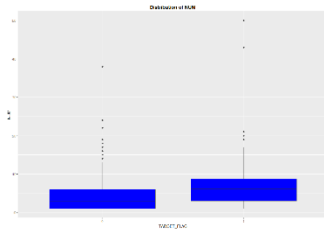


Figure 6b

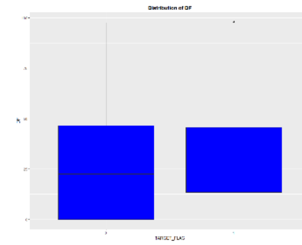


Figure 6c

Observations: Figure 6a shows a boxplot of MSLD vs. TARGET_FLAG, so that we can compare the median differences and variability between the numeric variable and TARGET_FLAG. The results show that the lower the number of months since the donor's most recent donation, the more likely he/she made a donation in March 2007 (vice versa). Figure 6b shows a boxplot of NUM vs. TARGET_FLAG. The results show that as the number of donations that the donor has made increases, the more likely he/she made a donation in March 2007 (vice versa). Figure 6c shows a boxplot of DF vs. TARGET_FLAG. The results show that as donation frequency increases, the more likely he/she made a donation in March 2007 (vice versa). As a result, this provides evidence that MSLD, NUM, and DF are strong predictors to include in our models since the median difference between whether a person made a donation in March 2007 (1 = Yes, 0 = No) is wide. I also conducted additional boxplots of TARGET_FLAG (x-axis) vs. the other numeric variables (y-axis) and found that DPM and TENRAT have median differences between TARGET_FLAG, while MSFD did not. However, these variables do not seem as strong as the predictors mentioned above since they either had a lot of variability or there were subtle differences between whether a person made a donation in March 2007 (1 = Yes, 0 = No). This was also validated using the variable importance feature in the Caret package.

Multivariate Plots for Qualitative Variables

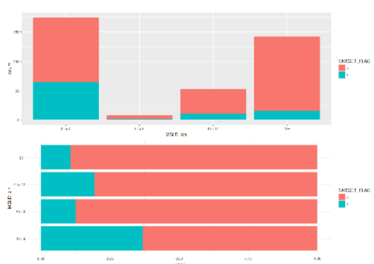


Figure 7a

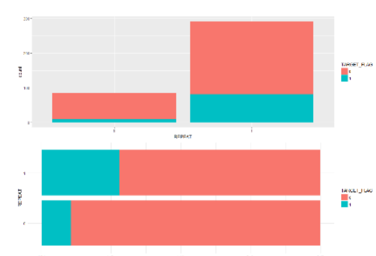


Figure 7b

Observations: Figures 7a and 7b shows bar plots of MSLD_bin and REPEAT vs. donors who made a donation in March 2007 (blue) and those who did not (red). The data shows that donors who fell between 0 to 4 months in regards to the number of months since their most recent donation, donated in March 2007 more than the other categories ranges (36.8%) (e.g., 5 to 8, 9 to 12, and 13+ months). For instance, those who were in the 13+ month range donated only 10.6% of the time. Furthermore, the data shows that those who were repeat customers (27.8%) donated in March 2007 more than those who were not (10.6%). For example, 81 out of the 90 people who donated in March 2007 were repeat customers.

Reviews of Literature & Formulation of Models

Reviews of Literature *(see last page for references)*

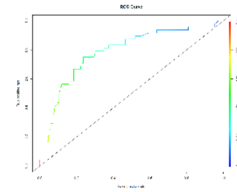
There were many peer reviewed journals in the NU library database that used logistic regression, LDA, QDA, decision trees etc. to predict the probability that a donor made a donation within a specified time period or predicted the probability on a closely related subject area. For instance, in the *Transfusion Medicine (2000)*, Flegel, Besenfelder, and Wagner use logistic regression to calculate the probability that someone will donate blood within a preselected time frame (e.g., 6-9 months after an index donation). Interestingly, first-time donors had a donation probability of 33% and were more likely to return than repeat donors. Second, in the *Environmental Pollution* (2018), Wang, Li, Ma, Li, Wang, Huang, Xu, Chenzi; and An, Yi use QDA, logistic regression, and decision trees to predict the probability of cadmium pollution (Cd) in rice in China. The results showed that the accuracy rate of 74% with QDA was significantly higher than the decision tree (67%) and logistic regression (68%) models. Lastly, Bhoi, Sherpa, and Khandelwal use LDA and decision trees to predict the probability of cardiovascular diseases such as myocardial ischemia and cardiac arrhythmias.

Modeling Strategy

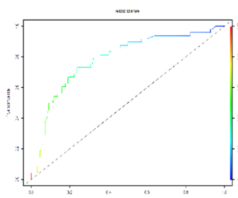
Given that the goal of the DrivenData competition Warm Up: Predict Blood Donations is to predict the probability that that a donor made a donation in March 2007, I began my analysis using linear classification techniques such as logistic regression (with stepAIC) and LDA to serve as initial baselines prior to conducting more sophisticated modeling techniques. Additionally, given that this data set is the smallest and least complex dataset on DrivenData (e.g., only 4 variables such as Months since Last Donation, Number of Donations, Total Volume Donated, and Months since First Donation are included), I felt like linear classification modeling techniques would perform admirably. After building logistic regression and LDA models, I then moved to QDA, Decision Trees, Bagging, Random Forest, Gradient Boosting Machines, and Neural Network. The next section provides a summary of my results for each modeling technique.

Application of Tools

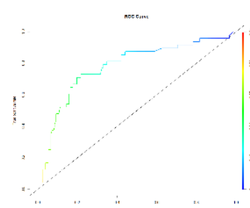
Logistic Regression: Logistic regression models the probability that the response variable belongs to a specific category and assumes a linear decision boundary (James, Witten, Hastie, & Tibshirani, 2013). For instance, it models the probabilities of the K classes using linear functions in x, while also ensuring that they sum to 1 and remain in-between 0 and 1 (Hastie, Tibshirani, & Friedman, 2009). This is accomplished using the logistic function and maximum likelihood, which is used to fit the model (James, et al., 2013). As a result, using the glm function, we produced a logistic regression model of TARGET_FLAG ~ MSLD + LOG_NUM + DF. These variables were chosen using stepwise regression (stepAIC). The Analysis of Deviance table showed that all the variables that were included in this model were statistically significant, which illustrates that these variables improved the model. Additionally, the Analysis of Deviance table and varimp showed that LOG_NUM, followed by MSLD, and DF impacted the model the most. In regards to the coefficients of the model, variables such as LOG_NUM, which has a positive coefficient and MSLD, which has a negative coefficient make intuitive sense and were statistically significant. For instance, as the number of donations that the donor has made increases, the more likely he/she will donate (vice versa). Additionally, the lower the number of months since the donor's most recent donation, the more likely he/she will donate (vice versa). The model produced the following performance metrics on the training dataset: AIC: 361.6423, BIC: 377.3606 and the following accuracy metrics and evaluation criteria on the validation dataset: accuracy: 0.77, AUC: 0.785636 (see ROC curve on the right), and LogLoss: 0.4608874. Note: We also tried logistic regression GAM, but the performance was exactly the same as a standard logistic regression model (e.g., model.gam1 <- glm(TARGET_FLAG ~ MSLD + s(NUM,1) + s(MSFD,50)+s(TENRAT,1)).



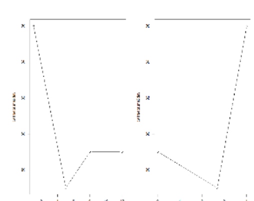
Linear Discriminant Analysis: LDA is very similar in form to logistic regression (distributions are assumed to be normal), except it models the distribution of the predictors separately in each of the response classes and then applies Bayes theorem (James, et al., 2013). This model also uses Gaussian densities, which arises when we assume that the classes have a common covariance matrix, and assumes a linear decision boundary (Hastie, et al., 2009). Using the lda function, we then produced a linear discriminant analysis model of MSLD + LOG_NUM + DF. These variables were chosen using the same variables from the logistic regression model (with stepAIC). The model also produced the following accuracy metrics and evaluation criteria: accuracy: 0.77, AUC: 0.785636 (see ROC curve on the right) and LogLoss: 0.4601341. LDA performed similarly to logistic regression.



Quadratic Discriminant Analysis: QDA is similar to LDA, in which the QDA classifier results from assuming the observations in each class are drawn from a Gaussian distribution and then Bayes theorem is applied to perform prediction (James, et al., 2013). However, unlike LDA, QDA assumes that each class has its own covariance matrix and assumes a quadratic decision boundary (James, et al., 2013). Using the qda function, we then produced a quadratic discriminant analysis model of MSLD + LOG_NUM + DF. These variables were chosen using the same variables from the logistic regression model (with stepAIC). The model also produced the following accuracy metrics and evaluation criteria: accuracy: 0.795, AUC: 0.7783717 (see ROC curve on the right), and LogLoss: N/A. QDA had higher accuracy than logistic and LDA, but slightly lower AUC.



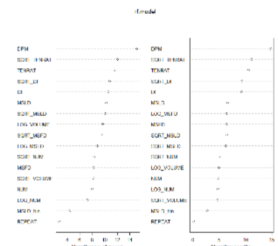
Decision Tree: A decision tree is a tree-based method that involves stratifying or segmenting the predictor space into a number of simple regions (James, et al., 2013). For instance, predictions are made by assigning an observation in a given region to the most common occurring class of training observations in that region (James, et al., 2013). Using the tree function, we produced a decision tree model with 5 predictor variables, after cross-validation helped eliminate 12 predictor variables (see plot on the right). The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.76,



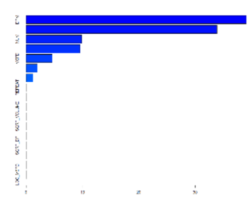
AUC: 0.6761239, and LogLoss: 0.5129595 on the validation dataset. The performance was worse than logistic regression, LDA, and QDA.

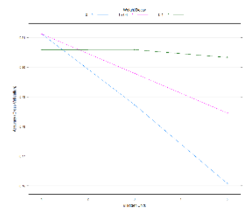
Bagging: Bagging is a technique for reducing the variance of an estimation prediction function. For classification, a committee of trees each cast a vote for the predicted class (aka: majority vote) (Hastie, et al., 2009). As a result, using the randomForest function (mtry=17, ntree=100), a bagged decision tree model was produced using all 17 predictor variables. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.755, AUC: 0.6057429, and LogLoss: N/A on the validation dataset. Interestingly, the performance was worse than logistic regression, LDA, QDA, and a basic decision tree.

Random Forest: Random forest provides an improvement over bagged trees by incorporating a small tweak that decorrelates the trees and then averages them (e.g., forces each split to only consider a subset of predictors and will not consider strong predictors so that other predictors will have more of a chance (James, et al., 2013)). As a result, using the randomForest function (mtry=4, ntree=400), a random forest model was produced using all 17 predictor variables. Using the variable importance function (which measures prediction strength), the plot on the right shows that DPM, SQRT_TENRAT, and TENRAT are the most important variables. This is different than what we saw earlier using stepAIC. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.78, AUC: 0.683114, and LogLoss: N/A on the validation dataset. This was an improvement over a basic decision tree and bagging. The accuracy score was also the second highest compared to all the models we've fitted thus far. However, AUC was dramatically lower than logistic regression, LDA, and QDA.



Gradient Boosting Machines: Boosting provides another approach for improving the predictions resulting from a decision tree by fitting each tree on an altered version of the original dataset (James, et al., 2013). In other words, trees are grown sequentially (e.g., each tree is grown using information from previously grown trees). As a result, using the gbm function, a boosted decision tree model was produced using all 17 predictor variables. I also incorporated n.trees =50, shrinkage=0.1, and depth=1, which was determined using a grid search. Relative importance showed than DPM and TENRAT are the most important variables, similar to what was seen in randomForest (see plot on the right). The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.765, AUC: 0.6387061, and LogLoss: 0.4992688 on the validation dataset. However, according to these metrics and evaluation criteria, boosting performed inadequately compared to the other models.





Neural Network: Neural network (aka: single hidden layer back-propagation network) is a nonlinear statistical model that is basically a nonlinear generalization of a linear model (Hastie, et al., 2009). It contains inputs, a hidden layer, and outputs that are typically represented by a network diagram (Hastie, et al., 2009). Additionally, neural network has unknown parameters called weights that introduce nonlinearities where needed (Hastie, et al., 2009). Using the nnet function, we then produced a neural network model of MSLD + LOG_NUM + DF. These variables were chosen using the same variables from the logistic regression model (with stepAIC). Additionally, using the variables from the logistic regression model makes sense given that neural network is essentially a bunch of logistic regressions, fed into a multinomial logit model. I also incorporated 1 hidden layer into the model with a decay= 1e-04 and maxit=1000, which was determined using a grid search (see plot on the right). The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.77, AUC: 0.7871436, and LogLoss: 0.4579446 on the validation dataset. As a result, neural network performed better than all the tree-based methods and had similar performance to logistic regression, LDA, and QDA.

Performance/Accuracy of Classification Models on Validation Set & DrivenData

Model Name	Accuracy	AUC	LogLoss
Logistic Regression	0.77	0.785636	0.4609
Linear Discriminant Analysis	0.77	0.785636	0.4601
Quadratic Discriminant Analysis	**0.795**	**0.778372**	**N/A**
Decision Tree	0.76	0.676124	0.5130
Bagging	**0.755**	**0.605743**	**N/A**
Random Forest	0.78	0.683114	N/A
Gradient Boosting Machines	0.765	0.638706	0.4993
Neural Network	0.77	0.787144	0.4579

Figure 8

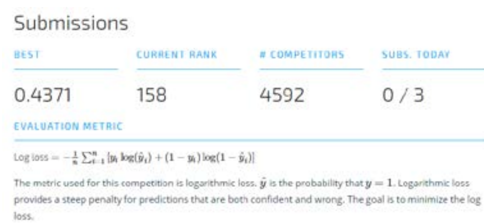


Figure 9

Observations: Figure 8 shows the performance/accuracy metrics and evaluation criteria for the following models in the validation dataset: logistic regression, LDA, QDA, decision trees, bagging, random forest, boosting, and neural networks. The results show that QDA had the highest accuracy, while bagging had the lowest accuracy out of all the models. Additionally, neural network, logistic regression, LDA, and QDA had similar AUC around 0.77 to 0.78. Given that QDA had the highest accuracy and a high AUC, I applied the model to the test dataset, and then submitted it to DrivenData. The LogLoss score was 0.4371, which currently places me in the top 3% out of 4,592 competitors (as of 7/18/18) (figure 9).

Conclusion

Future Work

In regards to future work, there are three areas that could help improve my models. First, it would be beneficial to obtain additional data and variables such gender, age, ethnicity, education level, job status, job type, college/university, donation region/city, climate/temperature variables, income (individual and family/parents), average time spent waiting prior to donating at the university, and average dollars spent on advertising at the university. Second, it could be helpful exploring different mixing of models using an ensemble approach or model averaging since model diversity can help increase accuracy and performance. Lastly, it could be helpful to explore different transformations of the variables, interactions, and other R packages.

Learnings

In the end, I learned three primary things from building these models. First, I learned how to build different classification and prediction models using various methods. Second, I learned that it's really important to conduct a thorough EDA and that a lot can be learned from it. Lastly, I learned that trying different modeling approaches can result in better performance and to never settle on a model due to gut instinct. For instance, I found it interesting that vanilla approaches such as logistic regression, LDA, and QDA performed better than more sophisticated modeling techniques such as Bagging, Random Forest and GBM.

References

1. Flegel, W.A., Besenfelder, W., & Wagner, F.F. (2000). Predicting a donor's likelihood of donating within a preselected time interval. *Transfusion medicine*, 10(3), 181-92. [Peer Reviewed Journal].

2. Wang, X., Li, X., Ma, R., Li, Y., Wang, W., Huang, H., Xu, C., and An, Y. (2018). Quadratic discriminant analysis model for assessing the risk of cadmium pollution for paddy fields in a county in China. *Environmental Pollution*, 236, 366-372. [Peer Reviewed Journal].

3. Bhoi, A.K., Sherpa, K.S., and Khandelwal, B. (2015). Classification Probability Analysis for Arrhythmia and Ischemia Using Frequency Domain Features of QRS Complex. *International Journal Bioautomation*, 19(4), 531-542. [Peer Reviewed Journal].

4. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer Science + Business Media.

5. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer Science + Business Media.