PREDICTING BLOOD DONATIONS USING MACHINE LEARNING TECHNIQUES

Deepti Bahel[1], Prerana Ghosh, Arundhyoti Sarkar, Matthew A. Lanham
Purdue University Krannert School of Management
403 W. State St., Krannert Bldg. 466, West Lafayette, IN 47907
dbahel@purdue.edu, ghoshp@purdue.edu, asarkar13@gmail.com, lanhamm@purdue.edu

ABSTRACT

We study the performance of machine learning algorithms that have not been previously investigated to support this problem of blood donation prediction. We build models on clustered data sets using k-means clustering and not using clustering to see if performance is significantly improved using clustering or not. The motivation for this research is that blood demand is gradually increasing by the day due to needed transfusions due to accidents, surgeries, diseases etc. Accurate prediction of the number of blood donors can help medical professionals know the future supply of blood and plan accordingly to entice voluntary blood donors to meet demand. We found that in a SVM clustered model with = 4 led to the best test set sensitivity (98.4%), which beat other studies. Our current solution is within the top 8% of all current participants in the DataDriven.org blood prediction competition.

[1] Corresponding author (Phone: 765-400-8848)

INTRODUCTION

The donation of blood is important because most often people requiring blood do not receive it on time causing loss of life. Examples include severe accidents, patients suffering from dengue or malaria, or organ transplants. Extreme health conditions such as Leukemia and bone marrow cancer, where affected individuals experience sudden high blood loss and need an urgent supply of blood and do not have it can also lead to loss of life. Sound data-driven systems for tracking and predicting donations and supply needs can improve the entire supply chain, making sure that more patients get the blood transfusions they need, which can reduce mortality risk.

One of the interesting aspects about blood is that it is not a typical commodity. First, there is the perishable nature of blood. Grocery stores face the dilemma of perishable products such as milk, which can be challenging to predict accurately so as to not lose sales due to an expired product. Blood has a shelf life of approximately 42 days according to the American Red Cross (Darwiche, Feuilloy et al. 2010). However, what makes this problem more challenging than milk is the stochastic behavior of blood supply to the system as compared to the more deterministic nature of milk supply. Whole blood is often split into platelets, red blood cells, and plasma, each having their own storage requirements and shelf life. For example, platelets must be stored around 22 degrees Celsius, while red blood cells 4 degree Celsius, and plasma at -25 degrees Celsius. Moreover, platelets can often be stored for at most 5 days, red blood cells up to 42 days, and plasma up to a one calendar year.

Amazingly, only around 5% of the eligible donor population actually donate (Linden, Gregorio et al. 1988, Katsaliaki 2008). This low percentage highlights the risk humans are faced today as blood and blood products are forecasted to increase year-on-year. This is likely why so many researchers continue to try to understand the social and behavioral drivers for why people donate to begin with. The primary way to satisfy demand is to have regularly occurring donations from healthy volunteers.

In our study, we focus on building a data-driven system for tracking and predicting potential blood donors. We investigate the use of various binary classification techniques to estimate the probability that a person will donate blood in March 2007 or not based on his past donation behavior. There is a time lag between the demand of blood required by patients suffering extreme blood loss and the supply of blood from blood banks. We try to improve this supply-demand lag by building a predictive model that helps identify the potential donors.

Based on our understanding of the problem, we follow a structured analytical process widely known in the data mining community, called the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman, Clinton et al. 2000). The idea behind this analysis framework is to develop and validate a model (or solution) that satisfies the requirements of problem and needs of stakeholders. We used guidance in the academic literature to get ideas of how others have modeled this problem and followed a similar process. Some authors clustered data before building their predictive models and some did not. We tried both and used some algorithms that others have not yet investigated to see if our solution was as good or better than what others have found.

We structured this paper as follows. We performed a review on the literature to see what methodologies have found to be successful at understanding this problem. We discuss the data set used in our study. Next, we discuss the methodology/design we implemented and discuss the models we investigated. Lastly, we present our results, discuss our conclusions, and how we plan to extend this research.

LITERATURE REVIEW

We examined the academic literature and grouped what we found into a couple different categories. First, blood banks often will survey donor volunteers to try and understand the factors that led them to donate. For example Godin, Conner et al. (2007) found that the important factors that lead to repeated blood donation among experienced donors were intention, perceived control, anticipated regret, moral norm, age, and past donation frequency. Moreover, the factors leading to repeated blood donation among new donors were only intention and age.

Others have designed studies to understand one's motives for donating blood. Sojka and Sojka (2008) surveyed over five hundred donators and found that the most commonly reported motivator among their participants was friend influence (47.2%), followed by media requests (23.5%). Lastly, they found that altruism (40.3%), social responsibility (19.7%), and friend influence (17.9%) were the primary drivers for blood donors to continue to be blood donors in the future.

As stated previously, only around 5% of eligible donor population actually donate (Katsaliaki 2008). The reasons for this are regularly reviewed by social and behavior scientists to help improve population participation (Ferguson, France et al. 2007).

The studies just discussed are outlined below in Table 1 are just a fraction of the many studies being performed to better understand the social and behavioral aspects of why people donate blood. We have found that many studies are trying to extend what is known as a theory of planned behavior (TPB), which continues to be developed in this area. This theory predicts the occurrence of certain behavior given that it is intentional and under volitional control (Veldhuizen, Ferguson et al. 2011). A systematic review and meta-analysis was performed in this area by Bednall, Bove et al. (2013).

Authors	Methods	Data	Drivers
(Godin, Conner et al. 2007)	Logistic Regression	Survey (2070 experience donors, 161 new donors)	*Experienced donors*: intention, perceived control, anticipated regret, moral norm, age, and past donation frequency. *New donors*: intention and age
(Sojka and Sojka 2008)	Descriptive statistics	Survey (531 participants)	*General motivators*: friend influence (47.2%), media requests (23.5%). *Continued donations*: altruism (40.3%), social responsibility (19.7%), friend influence (17.9%)
(Masser, White et al. 2009)	Structural equation modeling	Survey 1 (263 participants); Follow-up survey (182 donors)	Moral norm, donation anxiety, and donor identity indirectly predicted intention through attitude.
(Masser, Bednall et al. 2012)	Path analysis	Survey1 (256 participants)	Their extended TPB model showed intention was predicted by attitudes, perceived control, and self-identify

Table 1: Social and psychological studies investigating drivers of blood donations

The focus of our study is to understand the performance that using traditional machine learning techniques can have at predicting future blood donation. Table 2 outlines what we believe is an exhaustive list of all published studies in this domain, the data set used, methods employed, and results achieved. The "-" symbol indicates that nothing is reported in their paper in this table field.

Authors	Methods	Data	Results
(Mostafa 2009)	ANN (MLP), ANN (PNN), LDA	Survey (430 records, 8 features)	ANN (MLP): Test accuracy (98%) ANN (PNN): Test accuracy (100%) LDA: Test accuracy (83.3%)
(Santhanam and Sundaram 2010) (Sundaram 2011)	CART CART vs. DB2K7	UCI ML blood transfusion data[2] (748 donors, 5 features)	Precision/PPV (99%), Recall/Sensitivity (94%)
(Darwiche, Feuilloy et al. 2010)	PCA for feature reduction ANN (MLP) vs SVM (RBF)	UCI ML blood transfusion data (748 donors, 5 features)	SVM (RBF) using PCA: Test Sensitivity (65.8%); Test Specificity (78.2%); AUC (77.5%) MLP with features recency & monetary: Test Sensitivity (68.4%); Test Specificity (70.0%); AUC (72.5%)
(Ramachandran, Girija et al. 2011)	*J48* algorithm in Weka (aka C4.5)	Indian Red Cross Society (IRCS) Blood Bank Hospital (2387 records, 5 features)	Recall/Sensitivity (95.2%), Precision/PPV (58.9%), Specificity (4.3%)
(Lee and Cheng 2011)	k-Means clustering, J48, Naïve Bayes, Naïve Bayes Tree, Bagged ensembles of (CART, NB, NBT)	Blood transfusion service center data set (748 records/donors, 5 features)	Bagged (50 times) Naïve Bayes: Accuracy (77.1%), Sensitivity (59.5%), Specificity (78.1%), AUC (72.2%) * model had best AUC among competing models
(Zabihi, Ramezan et al. 2011)	Fuzzy sequential pattern mining	Blood transfusion service center data set (748 records/donors, 5 features)	Precision/PPV (Frequency feature 88%, Recency feature 72%, Time feature 94%)
(Sharma and Gupta 2012)	J48 algorithm in Weka (aka C4.5)	Blood bank of Kota, Rajasthan, India (3010 records, 7 features)	Accuracy (89.9%)
(Boonyanusith and Jittamai 2012)	Artificial Neural Network (ANN), J48 algorithm (aka C4.5)	Survey (400 records, 5 features)	ANN: Accuracy (76.3%); Recall/Sensitivity (81.7%); Precision/PPV (87.9%); Specificity (53.8%) J48: Recall/Sensitivity (81.2%); Precision/PPV (87.3%); Specificity (52.5%)
(Testik, Ozkaya et al. 2012)	Two-Step Clustering with CART This is fed into a serial queuing network model	Blood donation center (1095 donors, 3 clusters)	-
(Bhardwaj, Sharma et al. 2012)	-	-	-
(Khalid, Syuhada et al. 2013)	-	-	-
(Ashoori, Alizade et al. 2015)	C5.0, CART, CHAID, QUEST	Blood transfusion center in Birjand City in North East Iran (9231 donors, 6 features)	Model accuracy (train/test): C5.0 (57.49/56.4%), CART (55.9/56.4%), CHAID (55.56/55.61%), QUEST (55.34/56.11%)
(Ashoori, Mohammadi et al. 2017)	Two-step clustering, C5.0, CART, CHAID, QUEST	Census survey from a blood transfusion centers from Birjand, Khordad, & Shahrivar (1392 participants)	Important features: Blood pressure level, blood donation status, temperature Model accuracy: C5.0 (99.98%), CART (99.60%), CHAID (99.30%), QUEST (89.13%)

Table 2: Predicting blood donation with a focus on data mining/machine learning techniques

[2] https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

The first published study we found investigating machine learning classification techniques to identify donors versus non-donors was by Mostafa (2009). They show that it is possible to identify factors of blood donation behavior using machine learning techniques. They train and test two artificial neural network (ANN) variants; one using a multi-layer perceptron (MLP); the other a probabilistic neural network (PNN). They then compare these non-linear models to a linear discriminant analysis (LDA) model. They conclude that the ANN models both perform very well compared to LDA due the nonlinearities that exist in their data.

Santhanam and Sundaram (2010) used the Classification and Regression Tree (CART) from the University of California – Irvine Machine Learning repository. They showed on this data set that this algorithm has the ability to classify future blood donors accurately that had donated previously (i.e. recall/sensitivity of 94%). We found a very similar study published by one of the original authors the following year with a comparison of what they call a Regular Voluntary Donor (RVD) versus a DB2K7 (Donated Blood in 2007), which led to slightly better recall and precision (Sundaram 2011). Their key contribution was that the RVD model realized better accuracy than DB2K7. Darwiche, Feuilloy et al. (2010) extend this investigation of this data set by testing ANN with a radial basis function (RBF) as well as investigate performance using Support Vector Machines (SVMs). Even though the feature space is limited they also build and evaluate these models using principal components analysis (PCA) as feature inputs instead of the raw feature inputs. The SVM (RBF) model performed best using PCA as inputs because this model achieved the highest area under the curve (AUC) on the test set (i.e. 77.5%). The ANN model achieved the best AUC of 72.5% using only the features recency and monetary value. Lastly, we found the study design of (Darwiche, Feuilloy et al. 2010) better than (Santhanam and Sundaram 2010) and (Sundaram 2011) because their models are assessed on a test (i.e. holdout) set, which provides more realistic performance on future observations. Furthermore, this design allows one to identify if a model has overfit to the data by comparing the testing set statistics to the training set statistics.

Zabihi, Ramezan et al. (2011) investigate the use of fuzzy sequential pattern mining to try and predict future blood donating behavior. The features investigated in this study were (1) months since last donation, (2) total number of donations, (3) time (in months) since first donation, and (4) a binary feature indicating whether blood was donated in March 2007 or not. These features are similar in nature to those we investigated in our study.

Ramachandran, Girija et al. (2011) investigated the performance of the J48 algorithm provided in Weka[3]. The J48 algorithm is an implementation of the C4.5 decision tree written in Java (Wikipedia , Quinlan 1993). They found this methodology to also perform well at predicting blood donors whom had donated before having a sensitivity of 95.2%, but performed poorly at future non-donors. Sharma and Gupta (2012) also used the J48 algorithm in Weka on a different blood donation data set obtained from a blood bank in Kota, Rajasthan, India. While they were attempting to predict the "number" of donors through their age and blood group, they actually performed a classification of donors versus non-donors which raised concerns over the validity of this study.

Boonyanusith and Jittamai (2012) performed a blood donation survey in Thailand. Like previous studies they used the J48 decision tree, but also tried an artificial neural network. Both models yielded similar performance with sensitivity (81.7% vs 81.2%) and specificity (53.8% vs. 52.5%).

[3] Weka is a data mining software for Java users. http://www.cs.waikato.ac.nz/ml/weka/

Bhardwaj, Sharma et al. (2012) provided a very limited review of data mining in blood donation and do not actually train and test any models. They propose to do this in the future research. Likewise Khalid, Syuhada et al. (2013) provide a slightly more extensive review of the literature, but also do not perform any modeling or analysis.

Testik, Ozkaya et al. (2012) use the idea of trying to group similar donors based on arrival patterns using Two-Step clustering (SPSS 2001). Then once clusters are formed, CART was implemented on the individual clusters to try to improve predictive accuracy. This approach has been tested in other domains and is an approach we investigate in our study. However, instead of Two-Step clustering we implement models based on more widely known k-Means clustering algorithm. The authors do not report the predictive accuracy of their approach, nor provide a comparison of using Two-Step clustering-CART versus using CART alone. Their primary contribution is the formulation of a serial queuing network model that could be used in the case of blood center operations where arrival patterns could be estimated and used to support workforce size utilization.

Ashoori, Alizade et al. (2015) collected census data collected from a blood transfusion center located in Birjand City, North East Iran. This data set consisted of 9,231 donors and measured six features. They tried to predict future blood donors using four types of decision trees (C5.0, CART, CHAID, and QUEST). Their cross-validated models all yielded poor performance ranging from 55 to 57 percent accuracy. One interesting aspect of their results was that the best performing model, the C5.0 tree, had 41 rules compared to only 13 (CHAID), 8 (CART), and 5 (QUEST). With trees the more rules (or splits) used often will lead to overfitting to the data, but can also lead to more distinct probability values in the prediction. Ashoori, Mohammadi et al. (2017) extend research into the performance of these techniques by first using two-step clustering before employing the same decision tree algorithms used in their previous study. They conclude that this approach helped them predict faster and more precisely compared to their previous study.

DATA

The dataset used in our study is one used by others researchers studying the problem posted on the UCI Machine Learning Repository [4]. The source data has been taken from blood donor database of the Blood Transfusion Service Center in Hsin-Chu City in Taiwan. 748 donors were randomly selected from the donor database for the study. The features measured include R (Recency - months since last donation), F (Frequency - total number of donation), M (Monetary - total blood donated in c.c.), T (Time - months since first donation), and a binary variable representing whether the donor donated blood in March 2007 (1 stands for donating blood; 0 stands for not donating blood) as shown in Table 3.

Variable	Type	Description
X	Integer	Donor ID
Months since Last	Integer	This is the number of months since this donor's most
Number of	Integer	This is the total number of donations that the donor has
Total Volume	Integer	This is the total amount of blood that the donor has

[4] https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

Months since First	Integer	This is the number of months since the donor's first
Donated blood in	Binary	This gives whether person donated blood in March 2007

Table 3: Data dictionary

METHODOLOGY

Figure 1 outlines the flow of our study. First, we used k-Means clustering to cluster the data into similar groups. The idea is to group like items with like items before building predictive models, as this can lead to better predictive model performance per cluster, and thus lead to improved performance over all clusters.

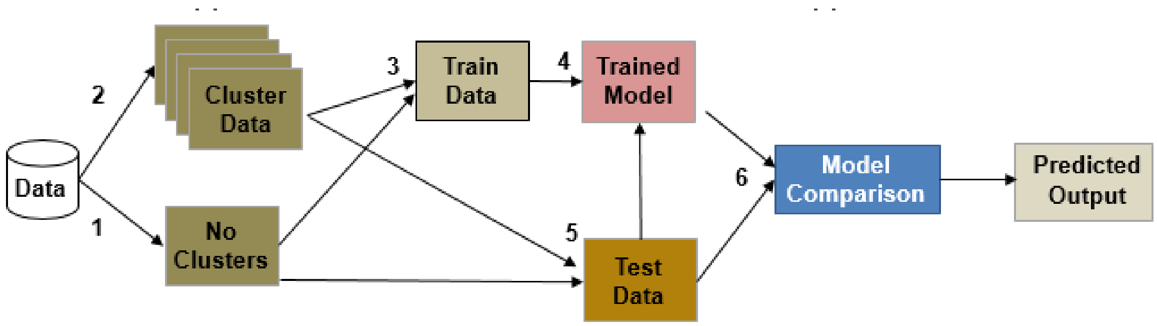
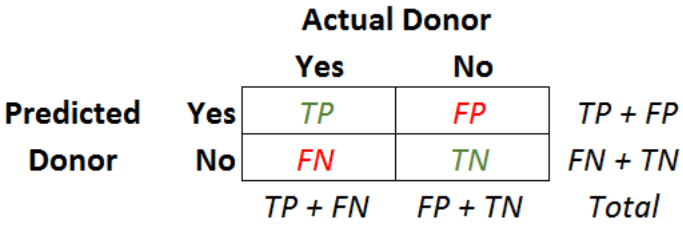


Figure 1: Methodology flow diagram

The dataset was randomly partitioned into training set and testing set using a 70/30 train/test partition. Models are trained using various algorithms using the entire training set, as well as trained on each cluster generated within the training set. Each model was trained once using what is sometimes referred to as a validation-set approach. We did try 5-fold 10-fold cross-validation on two of the models we researched. The idea here is to estimate a model over multiple folds (i.e. random partitions) instead of just one random training set. Cross-validation averages model fit performance measures such as prediction error to correct for the optimistic nature observed from the training error and thus provide an estimate of prediction risk that is more transparent (Seni and Elder 2010). The problem with k-fold and is the reason we only tried this on a couple models is we have a very small data sets when the data sets are clustered. Folds will make these training sets even smaller which might not provide any algorithm enough examples to learn.

Once models are trained, the test (i.e. holdout) data is fed into each trained model to measure model performance. These measures allow us to gauge the generalizability of the remaining subset of data not used in the study, and provides us a feel to the degree of how overfit any models are to the training data.

The statistical performance measures we obtained were overall accuracy, sensitivity, specificity, and area under the curve (AUC). The first three measures are easily calculated using a confusion matrix as shown in Figure 2. The overall accuracy measures how well you classify donors versus non-donors (TP+TN/Total). Sensitivity measures how well we are able to correctly predict donors whom have actually donated (TP/(TP+FN)). Specificity allows us to gauge how well we are able to predict non-donors among those whom did not donate (FP/(FP+TN)).



		Actual Donor		
		Yes	No	
Predicted	Yes	*TP*	*FP*	*TP + FP*
Donor	No	*FN*	*TN*	*FN + TN*
		TP + FN	*FP + TN*	*Total*

Figure 2: Confusion matrix used to generate statistical performance measures of donating

AUC is generated from a receiver operating characteristic (ROC) curve. This curve plots sensitivity versus 1-specificity for varying probability cutoff values. The typically used cutoff in binary classification model performance evaluation is 0.50, but it need not be. It is often standard practice to use the AUC statistic as the preferred statistic to use to compare which model performs better than another. Using varying cutoff thresholds (e.g. 0.01, 0.02, … , 0.99), one can construct a confusion matrix for each threshold and plot the sensitivity vs. 1-specificity performance. The estimated area under this curve provides the modeler a better feel of performance across a plethora of cutoffs, where the AUC closer to 1 indicates a perfect classifier, while a value close to 0.50 indicates a model that is not able to learn.

MODELS

Some of the models tested in our study have been used by other authors in their studies, but we also try other techniques to see how the perform. We used k-means to cluster our data as did (Lee and Cheng 2011). The predictive models similar to other studies include **CART** (Santhanam and Sundaram 2010, Lee and Cheng 2011, Sundaram 2011, Testik, Ozkaya et al. 2012, Ashoori, Alizade et al. 2015, Ashoori, Mohammadi et al. 2017), **J48/C4.5/C5.0** (Ramachandran, Girija et al. 2011, Boonyanusith and Jittamai 2012, Sharma and Gupta 2012, Ashoori, Alizade et al. 2015, Ashoori, Mohammadi et al. 2017), **artificial neural network (ANN)** (Mostafa 2009, Darwiche, Feuilloy et al. 2010, Boonyanusith and Jittamai 2012), **support vector machines (SVM)** (Darwiche, Feuilloy et al. 2010). The additional models we investigate that are not investigated in the literature is logistic regression (often referred to as just a logit model), boosted and bagged versions of the logit, and random forests.

K-means clustering is an unsupervised learning method which is used when labelled data is not available. In this method, the number of clusters or groups that the data needs to be divided is determined beforehand and it assigned to the variable K. Randomly k points are chosen as the centroids of the cluster. Each data point is assigned to one of the K clusters based on the vicinity of the point to the centroid of the Kth cluster. The assignment is determined by calculating the least Euclidean distance of the data point to all the K centroids. After formation of the clusters, the process is repeated several times until the centroids converge, that is they stop moving. This is determined by calculating the Euclidean distance between the new centroid and the old centroid of the same cluster.

CART

Classification and Regression Techniques (CART) coined by Leo Breiman is used to refer to decision trees used for classification and regression techniques in predicting modelling. Trees have nodes and leaves. Every node is a question. In a binary tree, it is a yes or no question. The parent node is split into two child nodes. The splitting continues until a decision is reached for the target variable. The last node where the data cannot be split further are called leaves or terminal nodes. Every node is split on certain variable which gives the maximum information gain. Various methods like minimum entropy and Gini index is used for this. Sometimes probability is also used. Due to the nature of trees, it is quite easy for the model to overfit. Hence methods such as pruning are used where the nodes are not split after reaching a predetermined maximum depth. The nodes are then replaced by a leaf.

C5.0 Decision Tree

J48/C4.5 and C5.0 are the successive versions of CART which accept both continuous and discrete features along with missing data points. Pruning is also implemented along with normal tree construction. C5.0 incorporates variable misclassification cost. It considers the fact that certain kind of misclassification penalize the model more. Hence, C5.0 is designed to minimize expected misclassification costs rather than error rates. It also gives different weightage to different cases minimizing case weight attribute. It is also found to be faster than C4.5.

Artificial Neural Network

Artificial neural networks (ANNs) are learning algorithms inspired by human brains. The main architecture of ANN is the input layer, the hidden layer and the output layer. Except for the input layer, all other layers are connected to their previous layer by weights in the form of a directed graph. The nodes represent a neuron which has a linear or non-linear activation function. The learning happens in two parts, feed-forward and back-propagation. In feed forward, weights are assigned and in back-propagation, actual learning happens. The error is calculated at each node and the weights are updated. This process is repeated until the algorithm converges. We investigated the multi-layered perceptron (MLP) neural network where the activation layer is a linear function which maps the weighted inputs to the output of each node.

Support Vector Machines

Support vector machines (SVMs) are supervised classification or regression techniques widely used for non-linear datasets. The kernel trick allows the user to deal with non-linear data without having to worry about its linear separability. In SVM-RBF, a radial basis function is used as the kernel. The algorithm transforms the data into higher dimension into a linearly separable space and implements quadratic programming to increase the speed. The crux of the algorithm is that the data is transformed into a linear space. The data is then separated using a hyperplane which is supported by data points. The best separating hyperplane is one with maximum support vectors and the maximum margin. However, the drawback is that a very complex or wiggly hyperplane is likely to overfit the data.

Logistic regression

One of the benefits of the traditional logit is it is a parametric model that allows one to interpret the effect each variable has on the response. Often public health researchers use this model to estimate odds ratios which provide a meaningful statistic for interpretation. Logistic regression is a binary class classification algorithm which is used to predict a binary outcome. Given a set of independent variables, it gives the probability of a data point by fitting it to a logit function. In generalized linear model (GLM) terms, the logit is the link function that transforms the response into a logistic s-curve.

The initial exploratory analysis revealed that there was a high correlation between total volume donated and number of donations made. The first logit model estimated included all variables, and we found that the effect of total volume donated was not able to be estimated in the software we used. We then created two: one models removed total volume, the other model removed total number of donations and kept the other variables in model. We found both the variables are statistically significant when not estimated together, but are when estimated together, indicating an issue of multicollinearity. We concluded to use total number of donations, months since first donation, and months since last donation as predictors.

> Model: Made Donation in March 2007 ~ f (Total Number of Donations, Months Since First Donation + Months Since Last Donation)

LDA

Sir Ronald Fisher published developed Linear Discriminant Analysis (LDA) in 1936, which makes it one of the first classification techniques every developed. LDA takes a different approach to estimating probabilities by modeling the distribution of the features separately in each of the response classes (blood donor vs non-blood donor). Then uses Bayes' theorem to flip these around into conditional probabilities. It turns out that the model is very similar in form to logistic regression when the distributions of the features are assumed to be normal. According to James, Witten et al. (2013), when the response classes are well-separated, the logit will tend to be unstable, while LDA tends to not have this issue. Also, if the number of observations is small and the distribution of features are normally distributed in each of the classes, the LDA model tends to be more stable than a logit.

Ensemble approaches

Ensembling approaches are a family of machine learning algorithms which tend to convert weak learners to stronger ones. Random Forest is the most popular ensembeling technique used today. This algorithm ensembles decision trees which can be used for both classification and regression problems. In this method, multiple decision tree models are built on smaller samples. The final output of a classifier is determined by the mode of output of all trees and mean of the outputs if it is a regression problem. Many have found random forests to be one of the more competitive approaches in machine learning.

Bootstrap aggregation or bagging is an ensembling technique primarily used to reduce bias and variance in supervised learning. Bootstrapping is random selection with replacement. From the entire dataset, a fixed number of data points are randomly selected with replacement. A full blown logistic regression model is constructed on the bootstrapped sample, which involves applying validation and regularization techniques. This process of generating bootstrapped samples and training a model is done several times. The final output is either the mode or the mean of the combination of all model outputs depending upon classification or regression problem respectively.

RESULTS

In the initial exploratory analysis phase of our study, we tried to find a visible line of distinction between donors and non-donors. Figure 3 shows a 3D graph of months since first donation, number of donations and months since last donation. We found that the groups of donors and non-donors were not visibly distinct.

Since the number of features available for modeling was so few, exploratory data analysis (EDA) was very limited. We investigated interactions among features as well as tried two-way and three-way interactions as model inputs but either led to the same performance or poorer performance. In such cases, we decided to use the main effects and no interactions in any of the methods we investigated. We followed the commonly accepted philosophy in predictive modeling that a simpler, less complex model is preferred when the statistical performance measures are no different.

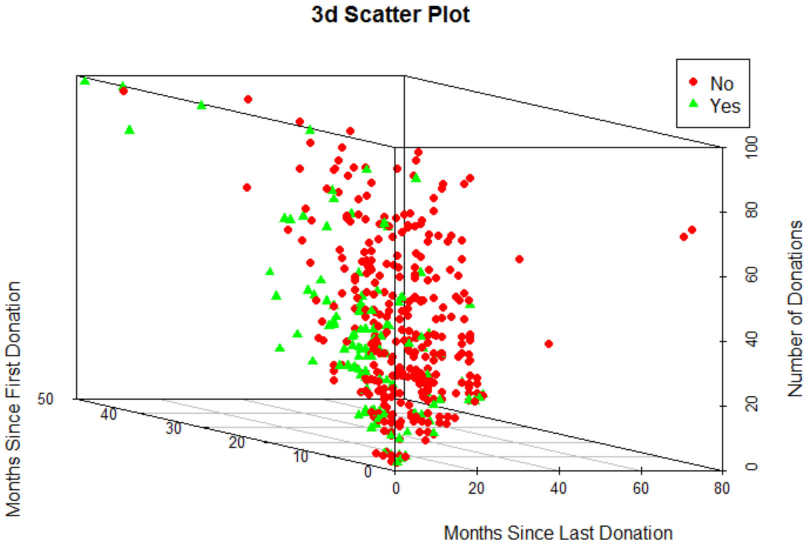


Figure 3: A 3D scatter plot depicting the distribution of patients donating blood or not

We evaluated the clusters generated from the k-Means algorithm using an elbow plot of mean squared error (MSE) versus the number of clusters as shown in Figure 4. We ran models by varying cluster size from 2 to 5. We found that five clusters was the ideal number of clusters for this data set as the MSE had marginal improvement with more groups. We also needed to consider the number of observations available in each cluster to train intelligent predictive models on each cluster, so having more than five would questionable.

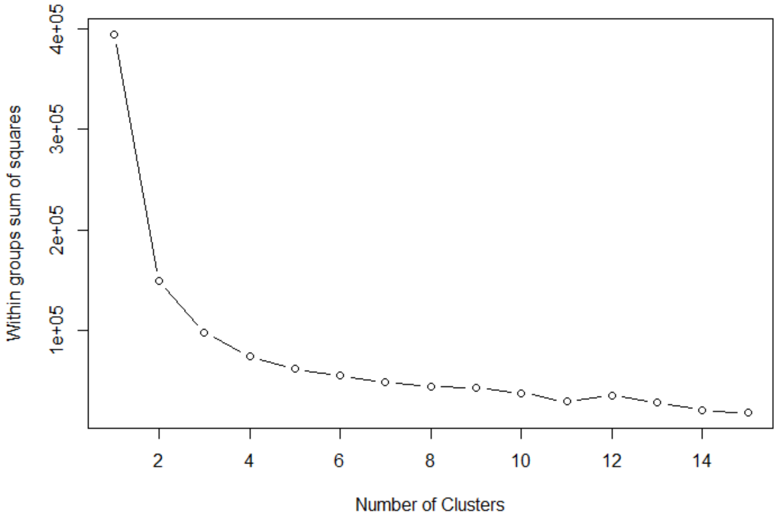


Figure 4: k-Means elbow plot

Figure 5 shows how the clusters are formed with regard to the features available. There is a clear grouping by the number of months since first donation.

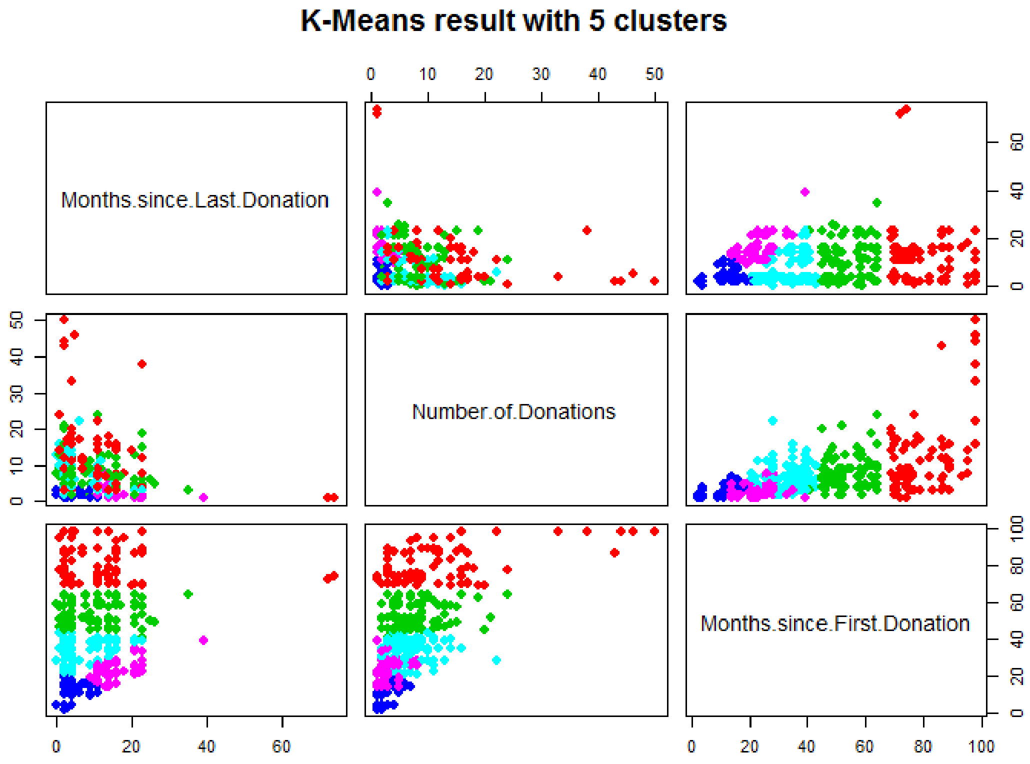
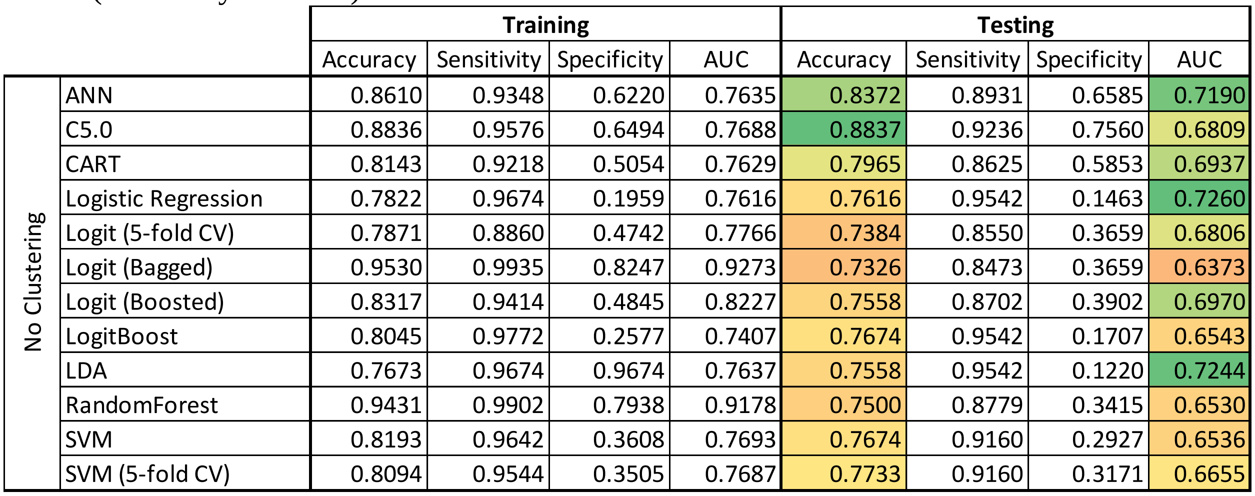


Figure 5: Colored clusters by features

Figure 6, Figure 7, Figure 8 summarizes the statistical performance for every model we investigated. The results are rather conclusive that using k-Means clustering with k=4 does improve performance of our models. High sensitivity on the testing set consistently is higher compared to not clustering. Compared to (Lee and Cheng 2011), our 5-fold cross-validated logit model performed the best among all models generated and slightly beat their best AUC of 72.2% using a boosted Naïve Bayes model.

They also tend to get a better balance of sensitivity versus specificity (59.5% vs 78.1%) compared to our model (85.5% vs. 36.59%). From the blood bank perspective, this means our best model predicts true blood donors better, but does not predict non-donors as well. Depending on how the blood bank markets to donors each of these models could lead to significantly different business performance and marketing costs. The argument made in some studies is that knowing whom will be a repeat donor is more important than knowing whom will not donate. Blood banks have invested significant effort to get recurring customers because those individuals help make the supply of blood more consistent over time, which helps supply chain management less risky.

Focusing on sensitivity in such cases we could achieve the best results using a clustered SVM model (sensitivity = 98.4%).



		Training				Testing			
		Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
No Clustering	ANN	0.8610	0.9348	0.6220	0.7635	0.8372	0.8931	0.6585	0.7190
	C5.0	0.8836	0.9576	0.6494	0.7688	0.8837	0.9236	0.7560	0.6809
	CART	0.8143	0.9218	0.5054	0.7629	0.7965	0.8625	0.5853	0.6937
	Logistic Regression	0.7822	0.9674	0.1959	0.7616	0.7616	0.9542	0.1463	0.7260
	Logit (5-fold CV)	0.7871	0.8860	0.4742	0.7766	0.7384	0.8550	0.3659	0.6806
	Logit (Bagged)	0.9530	0.9935	0.8247	0.9273	0.7326	0.8473	0.3659	0.6373
	Logit (Boosted)	0.8317	0.9414	0.4845	0.8227	0.7558	0.8702	0.3902	0.6970
	LogitBoost	0.8045	0.9772	0.2577	0.7407	0.7674	0.9542	0.1707	0.6543
	LDA	0.7673	0.9674	0.9674	0.7637	0.7558	0.9542	0.1220	0.7244
	RandomForest	0.9431	0.9902	0.7938	0.9178	0.7500	0.8779	0.3415	0.6530
	SVM	0.8193	0.9642	0.3608	0.7693	0.7674	0.9160	0.2927	0.6536
	SVM (5-fold CV)	0.8094	0.9544	0.3505	0.7687	0.7733	0.9160	0.3171	0.6655

Table 4: Table comparing the statistical performance metrics of non clustered model investigated

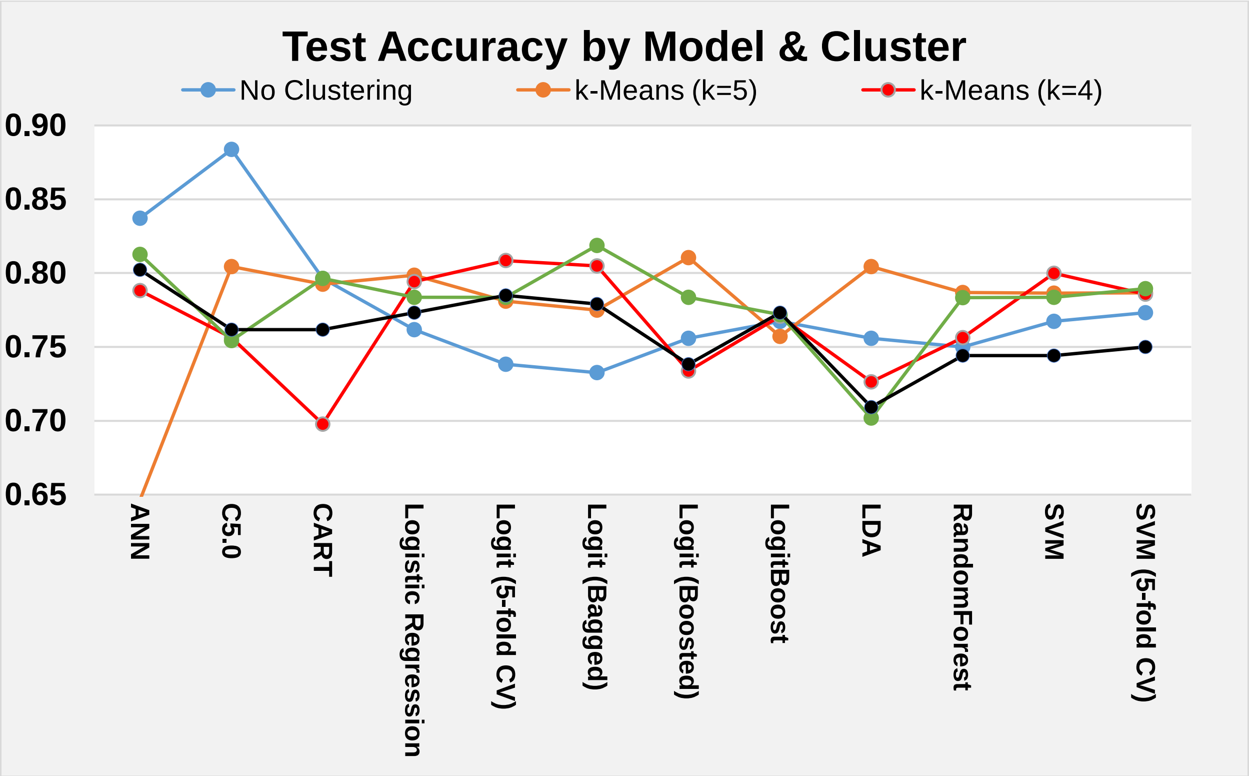


Figure 6 Graph comparing accuracy of different model on clustered and non clustered dataset

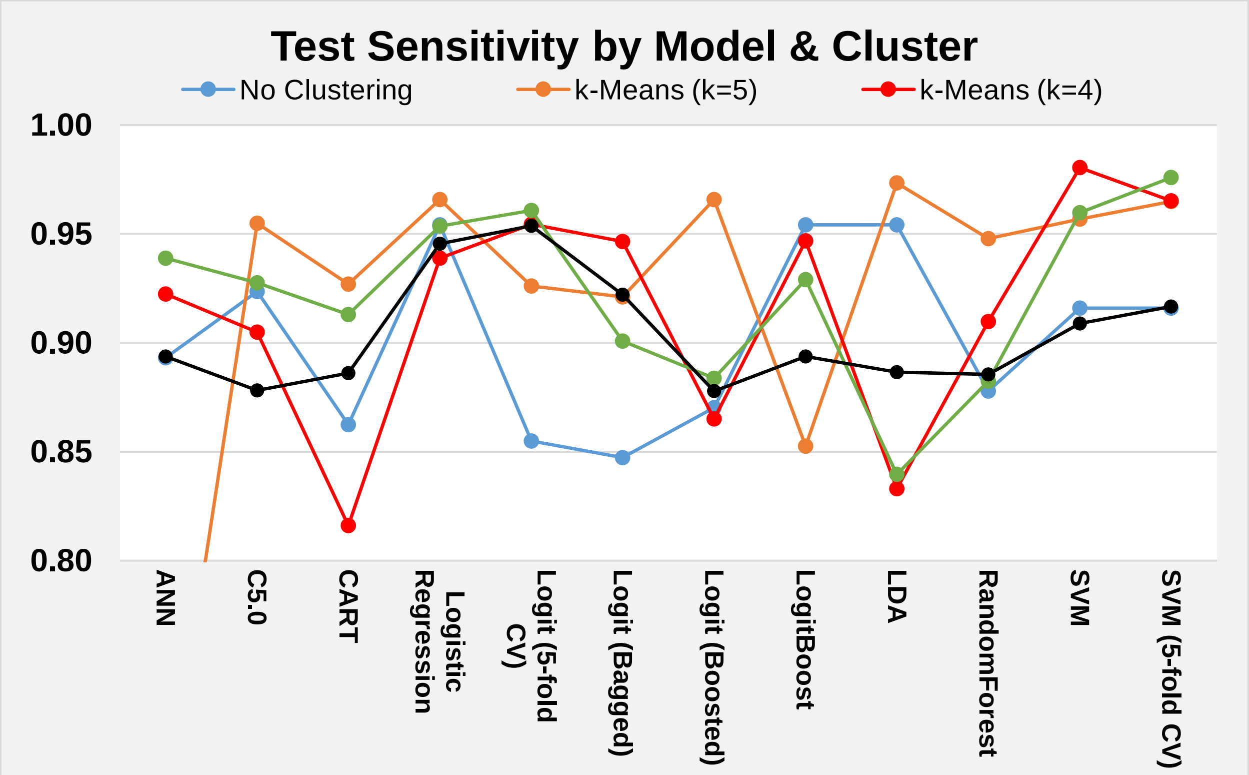


Figure 7 Graph comparing sensitivty of different model on clustered and non clustered dataset

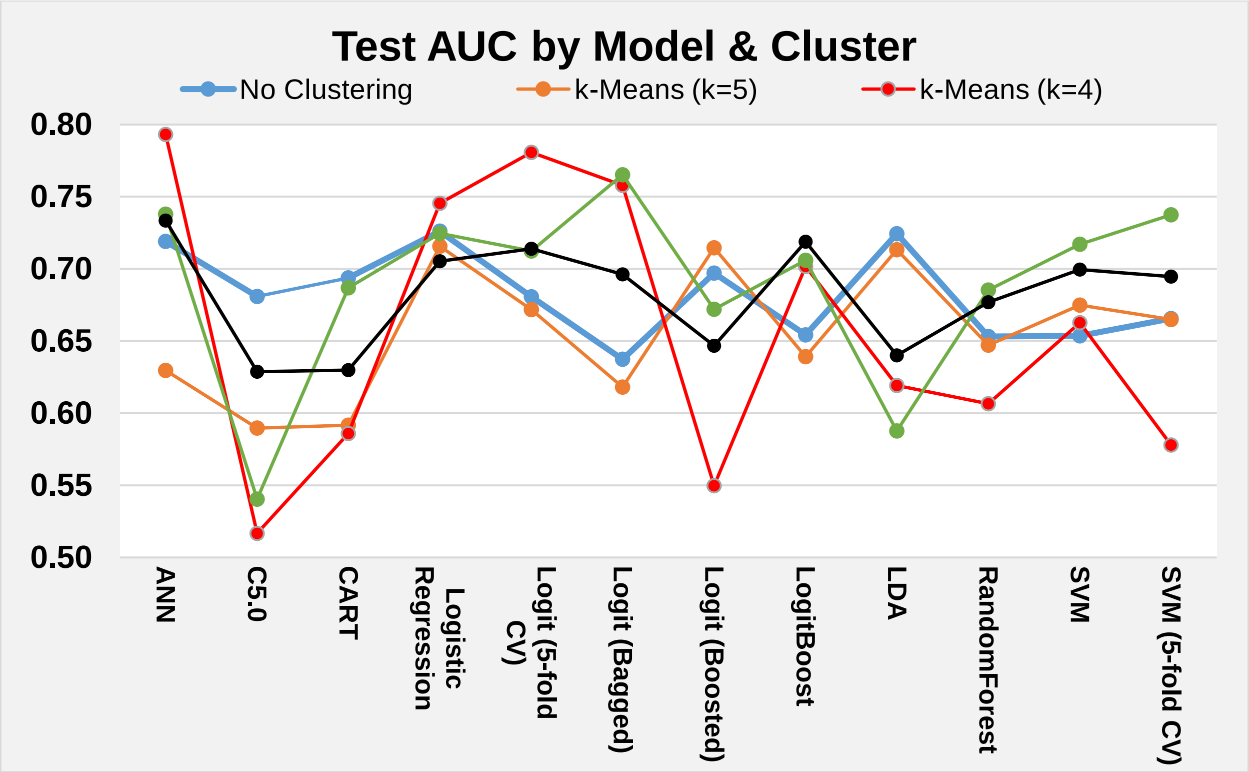


Figure 8 Graph comparing AUC of different model on clustered and non clustered dataset

Figure 96 shows that various models perform significantly better than others on the test/holdout set. AUC is the statistic used frequently by practitioners to compare model performance among competing models. The bagged logit (blue line) performed very poorly as its closest to the 45-degree line (i.e. 50% AUC). The basic logit (red line) appears to be one of the strongest performers, while the random forest (neon green line) is surprisingly not.

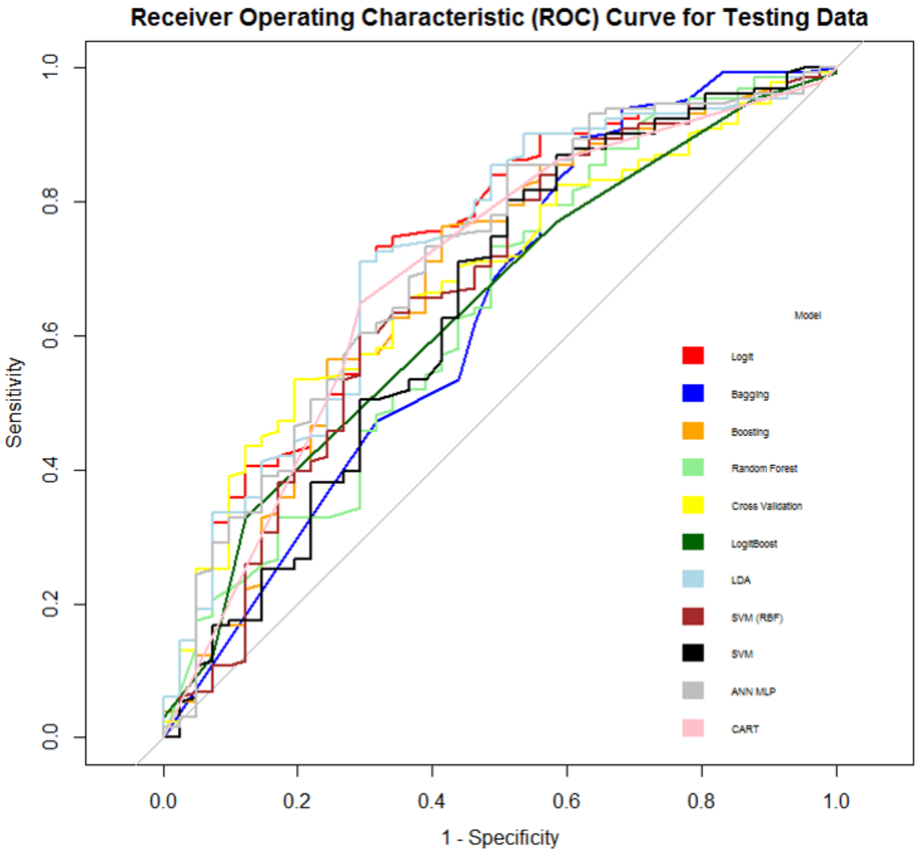


Figure 96: ROC curves for each model

CONCLUSIONS

In this study, we have compared the performance of various binary classification algorithms not investated previously on clustered data and non-clustered data to see if we can better predict if a person is going to donate blood or not.

Among the algorithms examined, the un-clustered 5-fold cross-validated logistic regression model performed the best based on the test set AUC. However, AUC alone may not be best performance measure with respect to likelihood to predict blood. That is because AUC considers the area determined by True Positive Rate (TPR)/sensitivity and False Positive Rate (FPR)/(1- Specificity). Our model could be used for targeted advertisement. In such a case, we are more interested in the TPR which would be to target the actual donors who would be interested in donating blood regularly. Hence, our performance would focus more on sensitivity leading us to recommend a clustered SVM model.

We believe this study could be a valuable extension to the academic literature in blood donation modeling if we add the following. First, some authors used algorithms that are specific to a certain software package. For example, CHAID and QUEST are decision tree algorithms specific to SPSS Modeler. Two-Step clustering is also specific to SPSS. We believe we should test and compare these algorithms to have a more complete comparison to what other authors have done. Secondly, there are many other machine learning algorithms that could be tested where one (or a combination of several) might yield to significantly better results. Lastly, we believe the statistical performance found to date would likely be deemed as “good” among blood banks practitioners in the field. We believe assessing the models from a cost-benefit perspective where the financial cost of misclassification and reward of correct classification is incorporated into the confusion matrix and assessment statistics likely provide additional insights to the blood bank than just statistical performance measures alone.

REFERENCES

Ashoori, M., et al. (2015). "A model to predict the sequential behavior of healthy blood donors using data mining."

Ashoori, M., et al. (2017). "Exploring Blood Donors’ Status Through Clustering: A Method to Improve the Quality of Services in Blood Transfusion Centers." Journal of Knowledge & Health **11**(4): page: 73-82.

Bednall, T. C., et al. (2013). "A systematic review and meta-analysis of antecedents of blood donation behavior and intentions." Social science & medicine **96**: 86-94.

Bhardwaj, A., et al. (2012). "Data mining techniques and their implementation in blood bank sector–a review." International Journal of Engineering Research and Applications (IJERA) **2**(4): 1303-1309.

Boonyanusith, W. and P. Jittamai (2012). Blood donor classification using neural network and decision tree techniques. Proceedings of the World Congress on Engineering and Computer Science.

Chapman, P., et al. (2000). "CRISP-DM 1.0 Step-by-step data mining guide."

Darwiche, M., et al. (2010). Prediction of blood transfusion donation. Research Challenges in Information Science (RCIS), 2010 Fourth International Conference on, IEEE.

Ferguson, E., et al. (2007). "Improving blood donor recruitment and retention: integrating theoretical advances from social and behavioral science research agendas." Transfusion **47**(11): 1999-2010.

Godin, G., et al. (2007). "Determinants of repeated blood donation among new and experienced blood donors." Transfusion **47**(9): 1607-1615.

James, G., et al. (2013). "An Introduction to Statistical Learning."

Katsaliaki, K. (2008). "Cost-effective practices in the blood service sector." Health policy **86**(2): 276-287.

Khalid, C., et al. (2013). Classification Techniques in Blood Donors Sector–A Survey. E-Proceeding of Software Engineering Postgraduates Workshop (SEPoW) 2013, Universiti Teknikal Malaysia Melaka.

Lee, W.-C. and B.-W. Cheng (2011). "An intelligent system for improving performance of blood donation." 品質學報 **18**(2): 173-185.

Linden, J. V., et al. (1988). "An estimate of blood donor eligibility in the general population." Vox sanguinis **54**(2): 96-100.

Masser, B. M., et al. (2012). "Predicting the retention of first-time donors using an extended Theory of Planned Behavior." Transfusion **52**(6): 1303-1310.

Masser, B. M., et al. (2009). "Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model." Transfusion **49**(2): 320-329.

Mostafa, M. M. (2009). "Profiling blood donors in Egypt: a neural network analysis." Expert Systems with Applications **36**(3): 5031-5038.

Quinlan, J. R. (1993). C4. 5: programs for machine learning, Morgan kaufmann.

Ramachandran, P., et al. (2011). "Classifying blood donors using data mining techniques." IJCST **1**.

Santhanam, T. and S. Sundaram (2010). "Application of CART algorithm in blood donors classification." Journal of Computer Science **6**(5): 548.

Seni, G. and J. F. Elder (2010). "Ensemble methods in data mining: improving accuracy through combining predictions." Synthesis Lectures on Data Mining and Knowledge Discovery **2**(1): 1-126.

Sharma, A. and P. Gupta (2012). "Predicting the number of blood donors through their age and blood group by using data mining tool." International Journal of communication and computer Technologies **1**(6): 6-10.

Sojka, B. N. and P. Sojka (2008). "The blood donation experience: self-reported motives and obstacles for donating blood." Vox sanguinis **94**(1): 56-63.

SPSS (2001). The SPSS TwoStep Cluster Component: A scalable component enabling more efficient customer segmentatio. SPSS**:** 9.

Sundaram, S. (2011). "A Comparison of Blood Donor Classification Data Mining Models." Journal of Theoretical & Applied Information Technology **30**(2).

Testik, M. C., et al. (2012). "Discovering blood donor arrival patterns using data mining: A method to investigate service quality at blood centers." Journal of medical systems **36**(2): 579-594.

Veldhuizen, I., et al. (2011). "Exploring the dynamics of the theory of planned behavior in the context of blood donation: does donation experience make a difference?" Transfusion **51**(11): 2425-2437.

Wikipedia C4.5 algorithm.

Zabihi, F., et al. (2011). "Rule Extraction for Blood donators with fuzzy sequential pattern mining." The Journal of mathematics and Computer Science **2**.