Predicting Blood Donations Using Machine Learning Techniques

Deepti Bahel, Prerana Ghosh, Arundhyoti Sarkar, Matthew A. Lanham

Purdue University Krannert School of Management

dbahel@purdue.edu; ghoshp@purdue.edu; sarkar13@purdue.edu; lanhamm@purdue.edu

Abstract

We study the performance of machine learning algorithms that have not been previously investigated to support the 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 gauge the future supply of blood and plan accordingly to entice voluntary blood donors to meet demand. We found that in a non-cluster C.50 tree realized the best accuracy, a clustered (k=4) ANN model yielded the best, while a clustered (k=4) SVM model yielded the best specificity, which might be the best for targeted targeted advertisement. Our current solution is within the top 8% of all current participants in the DataDriven.org blood prediction competition.

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 expiration. 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 calendar year.

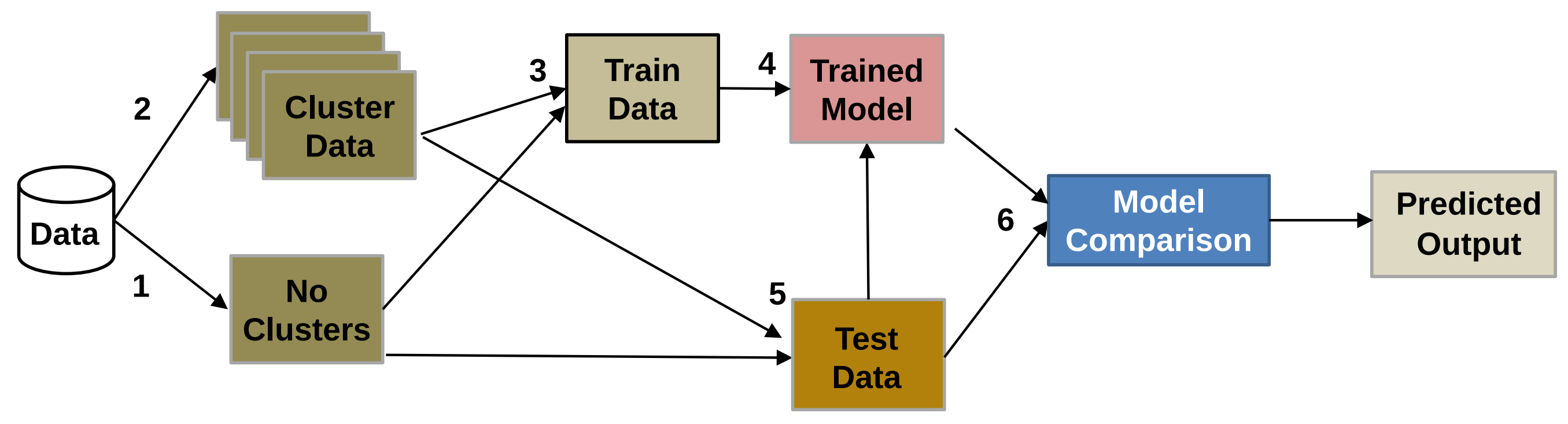
Amazingly, only around 5% of the eligible donor population actually donate (Linden, Gregorio et al. 1988, Katsaliaki 2008). This low percentage highlights the risk humans are faced with today as blood and blood products are forecasted to increase year-on-year. This is likely why so many researchers continue 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.

Data

The dataset used in our study is one used by others researchers studying the problem posted on the UCI Machine Learning Repository. 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).

Methodology

We followed two approaches: a non-clustered and clustered approach as shown:



The dataset was randomly partitioned into training set and testing set using a 70/30 train/test partition. Models are trained [3] 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 referred to as a validation-set approach where there is one training set and one test set.

Once models are trained, the [5] test (i.e. holdout) data is fed into each trained model to measure [6] 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 with respect to the training data.

The statistical performance measures we obtained were overall accuracy, sensitivity, specificity, and area under the curve (AUC). 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)). AUC is generated from a receiver operating characteristic (ROC) curve.

Models

Some of the models tested in our study were investigated by others: **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 (i.e. **logit**), and **ensemble-type models**, boosted and bagged versions of the logit, and **random forests**.

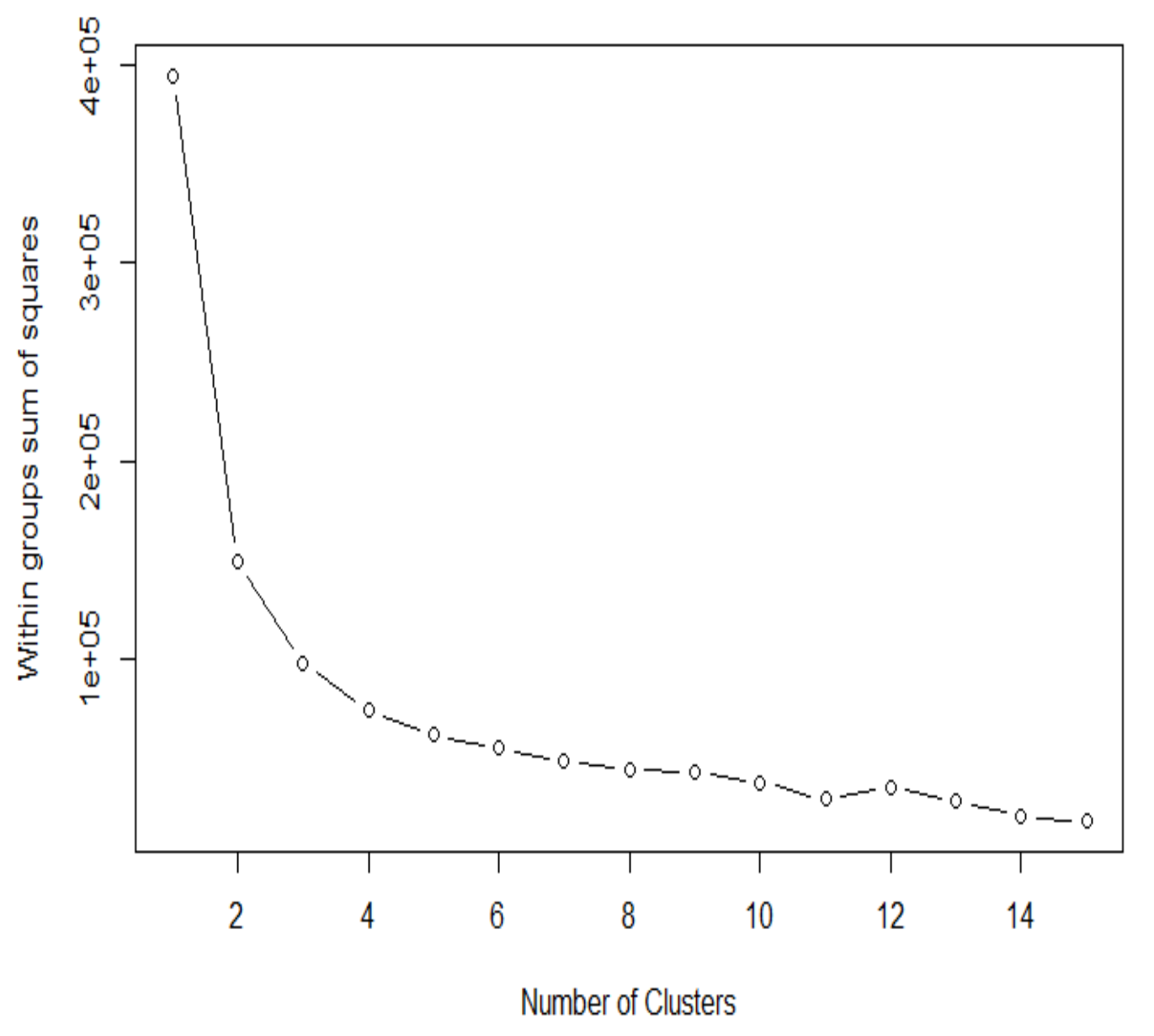
Results

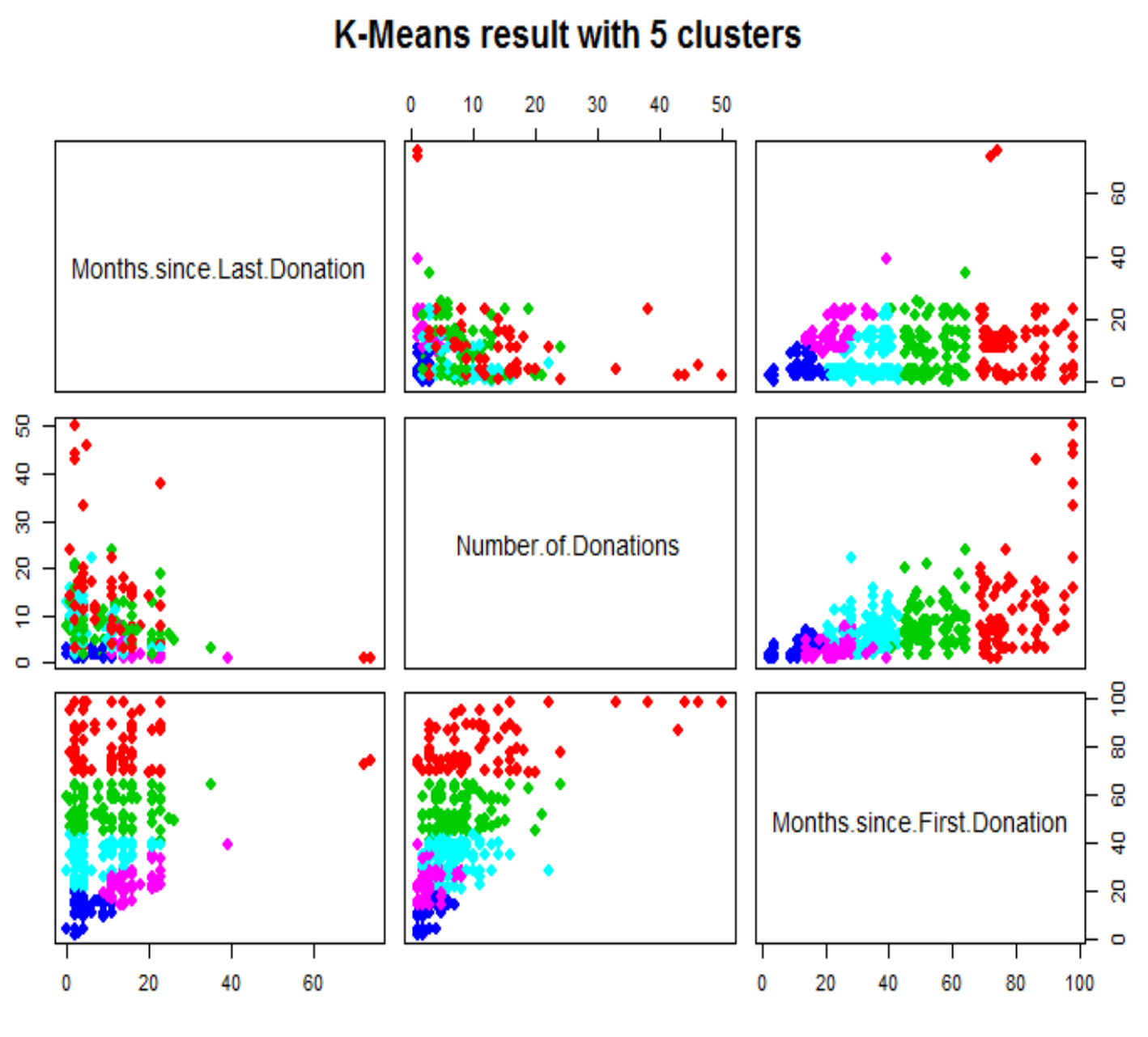
We realized the following results as shown in the table below when we did not cluster.

		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

We observed that the C5.0 decision tree yielded the best accuracy (88.3%), while logistic regression had the best AUC (0.726)

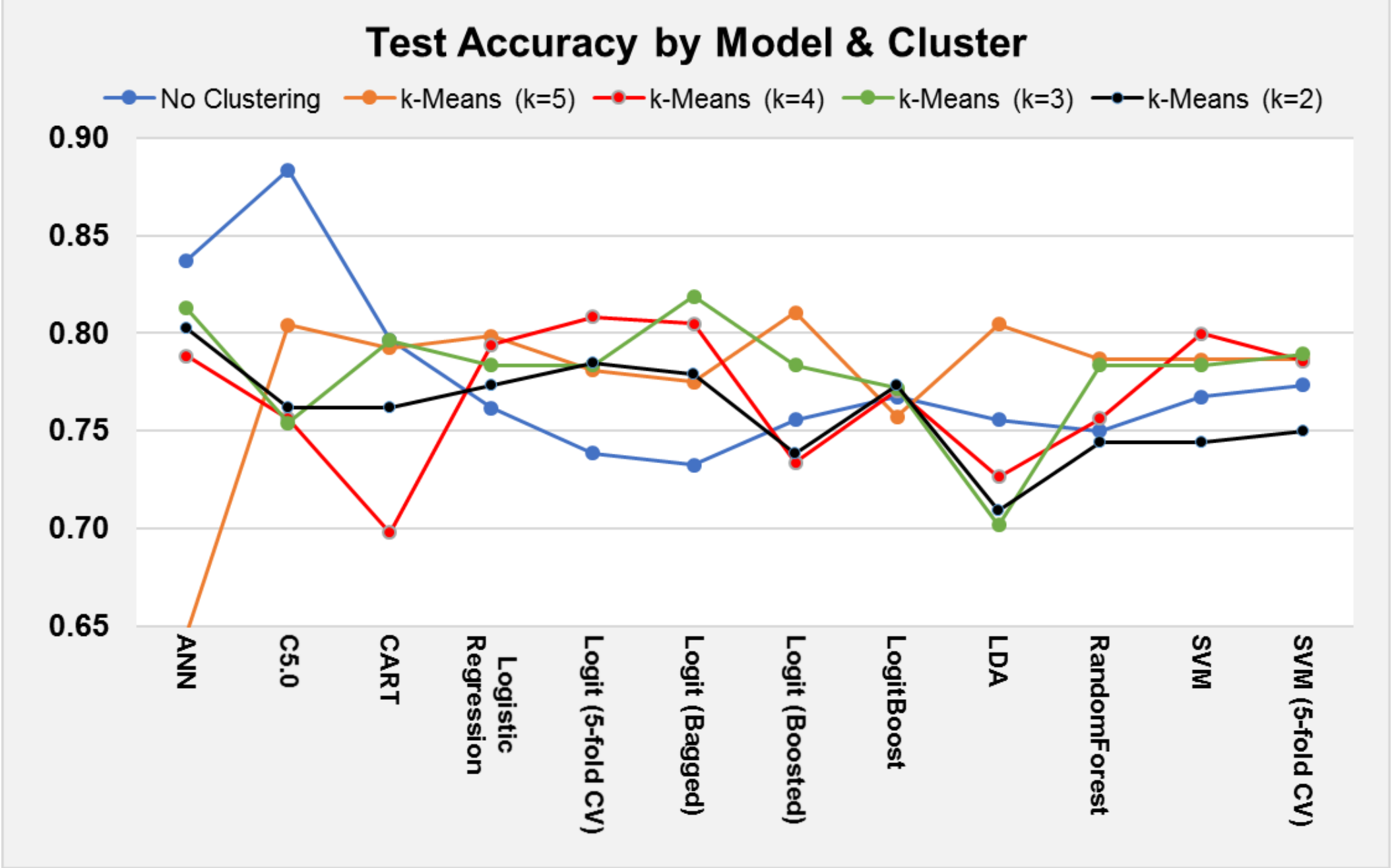
Clusters generated from the k-Means algorithm and evaluated using an elbow plot of mean squared error (MSE) versus the number of clusters as shown below. Next, we trained models clusters k=2,3,4, and 5, which were points around the elbow of the elbow plot.



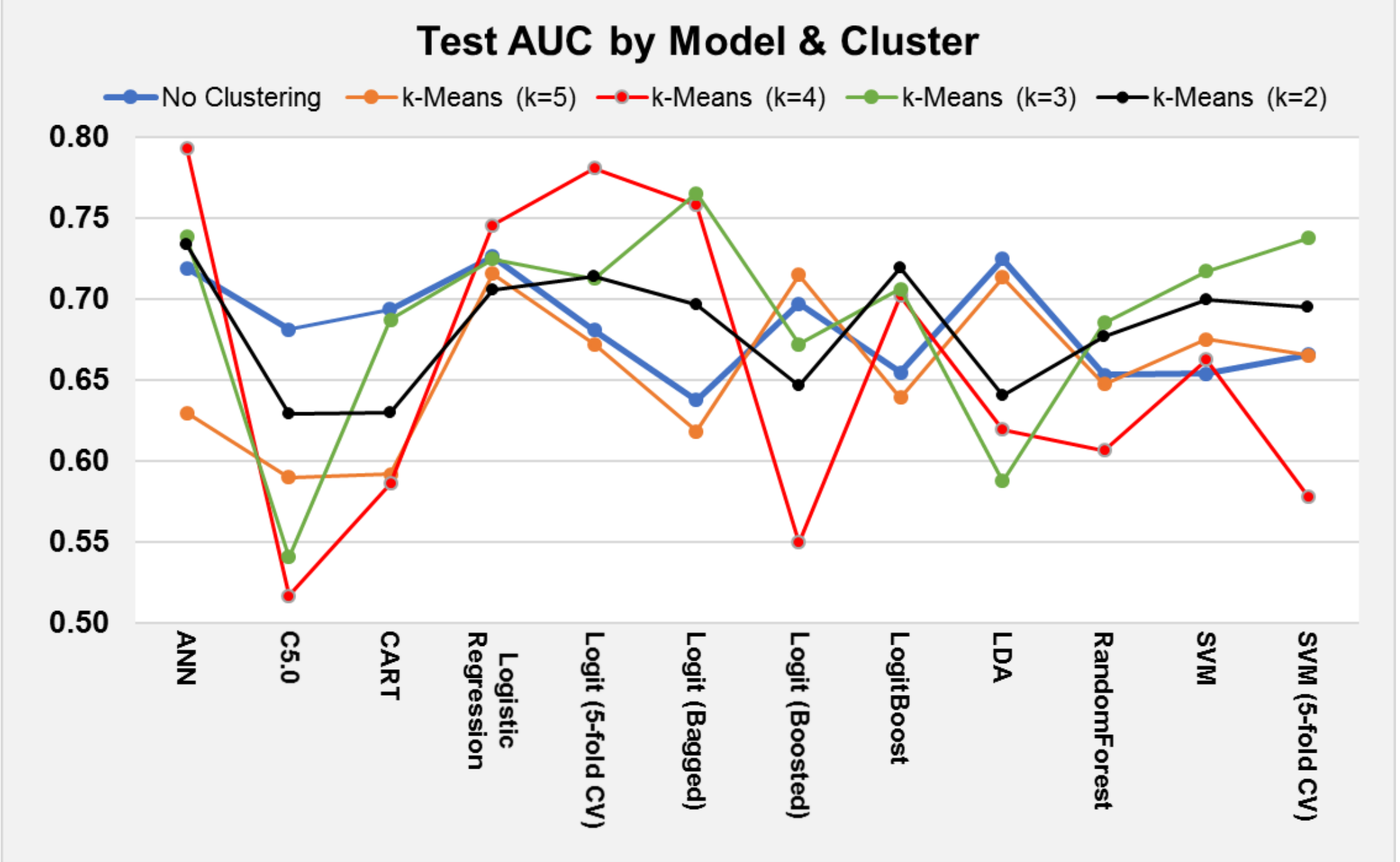


Interestingly, there is are clear grouping by the number of months since first donation. We depict these clustering using k-5.

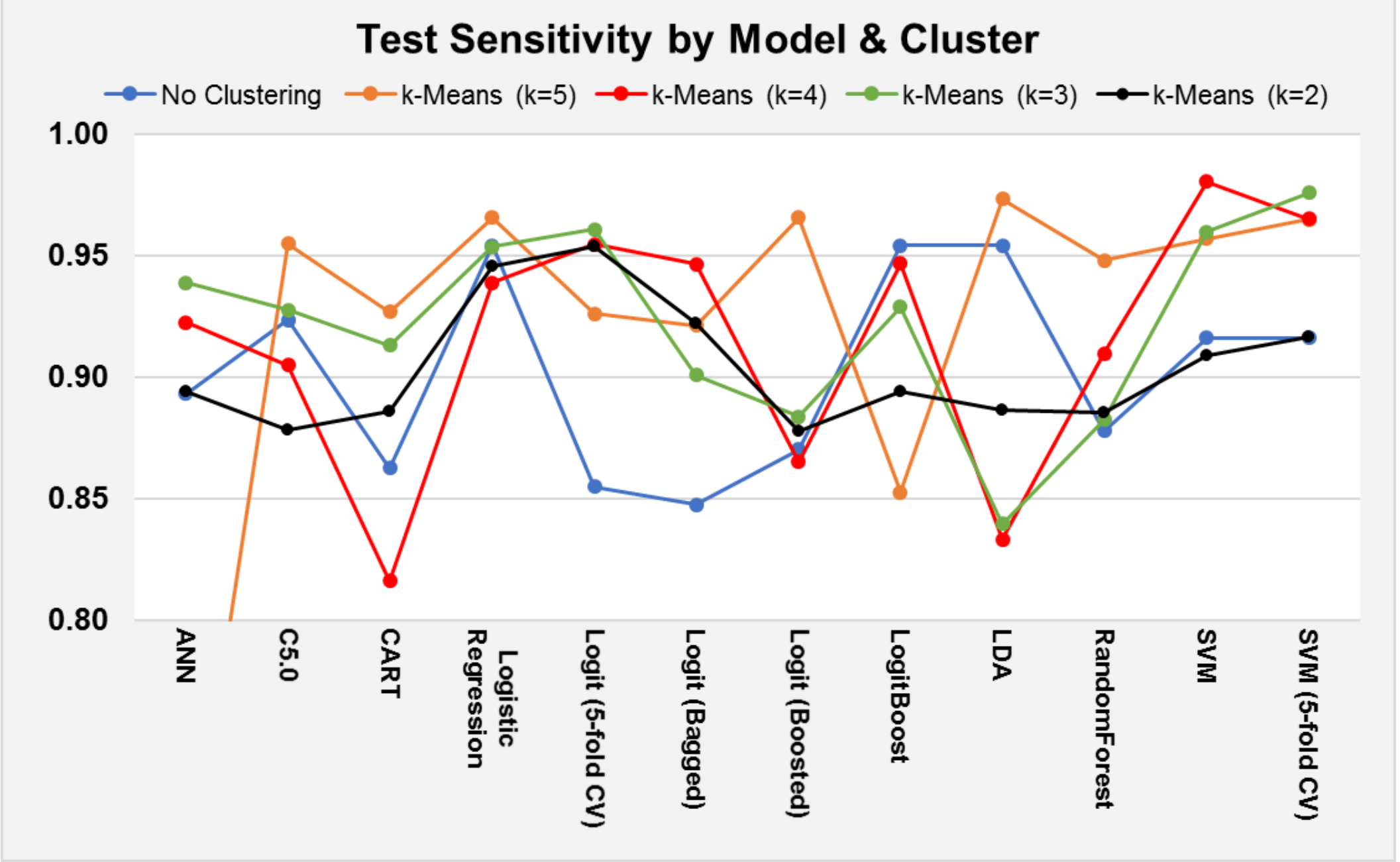
The best overall accuracy achieved on the test set was 88.37% from the C5.0 tree without clustering.



The ROC curves for each model combination was cluttered, so we examined the AUC as shown below. Typically the model with the best AUC is chosen in practice. The ANN model with k=4 yielded the greatest AUC of 0.793.



From the blood bank perspective, some studies we read made the argument that knowing whom will be a repeat donor is more important than knowing whom will not donate. Focusing on sensitivity in such cases we could achieve the best results using a clustered (k=4) SVM model.



Conclusions

We have compared the performance of various binary classification algorithms not investigated 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 cluster (k=4) ANN model performed the best based on the test set AUC, and C.50 based on accuracy. However, C.50 alone may not be the best measure with respect to likelihood to predict blood. The AUC considers the area determined by Sensitivity (TPR) and 1- Specificity (FPR). Our model could be used for targeted advertisement. Here we are more interested in the TPR which would be to target the actual donors who would be interested in donating blood regularly. Thus, focusing on sensitivity leads to using a clustered (k=4) SVM model.

Acknowledgements

We would like to thank Purdue's Business Information and Analytics Center (BAIC) for partially funding this work.