Predicting Blood Donations Using Machine Learning Techniques

Deepti Bahel
M.S. in Business Analytics & Information Management Candidate

KRANNERT SCHOOL OF MANAGEMENT
PURDUE UNIVERSITY

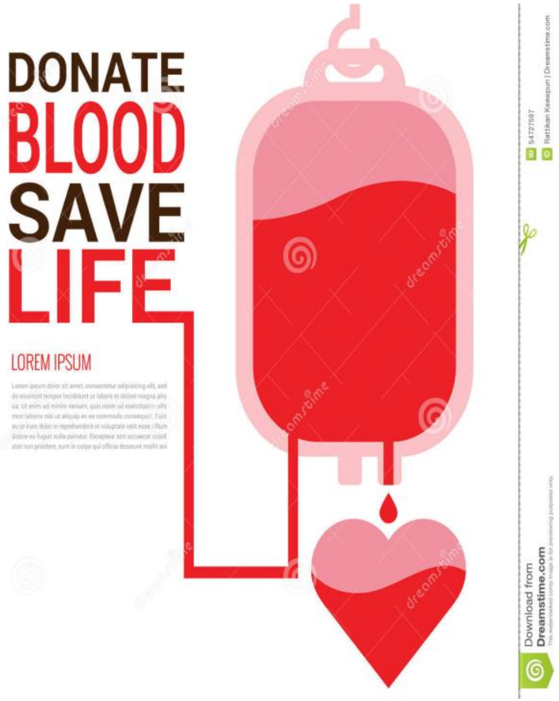
dbahel@purdue.edu

PURDUE'S KRANNERT SCHOOL
PREPARING ANALYTICAL GLOBAL BUSINESS LEADERS

Overview

- Motivation
- Literature Review
- Data
- Methodology
- Models
- Results
- Conclusions

- Shortage of blood in case of fatal accidents and diseases such as dengue, malaria can be life threatening for the patient.
- Every year Red Cross organize blood donation drives to give back to society.
- Only 5% of eligible blood donors donate on regular basis.
- Different components of blood have different shelf life.
- A good data-driven system will help blood donation drives target potential donors effectively.



Deepti Bahel
(dbahel@purdue.edu)

Social and psychological studies investigating drivers of blood donations

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

Predicting blood donation with a focus on data mining/machine learning

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 (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

Deepti Bahel
(dbahel@purdue.edu)

Predicting blood donation with a focus on data mining/machine learning

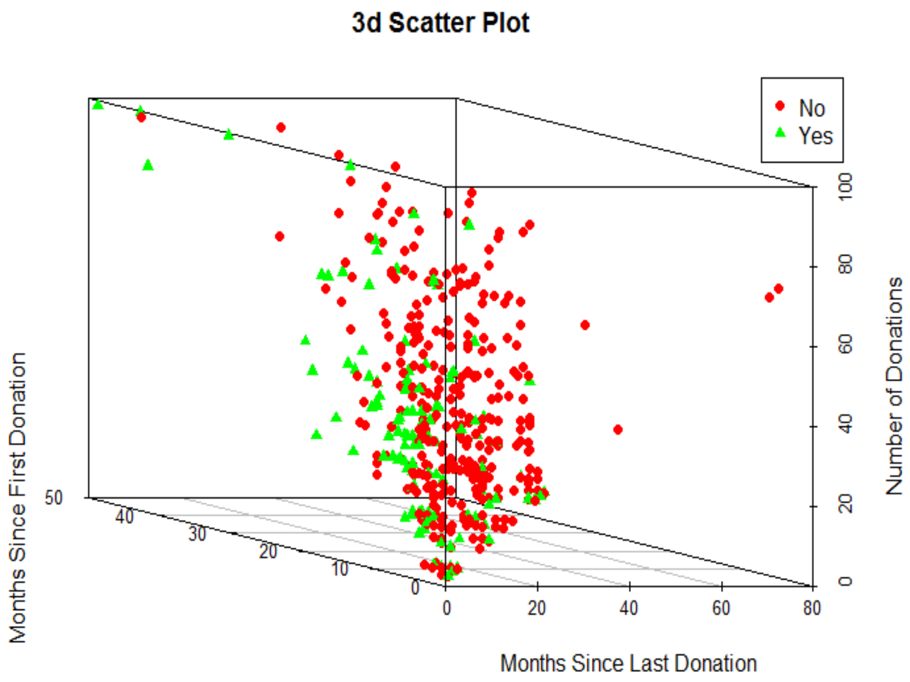
Authors	Methods	Data	Results
(Zabihi, Ramezan et al. 2011)	Fuzzy sequential pattern mining	Blood transfusion service center data set (748 records/donors, 5 features)	Precision/PPV (Frequency feature 88%, Recency feature 72%, Time feature 94%)
(Sharma and Gupta 2012)	J48 algorithm in Weka (aka C4.5)	Blood 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)	-
(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%)

- Source: UCI Machine Learning Repository
- Number of observations : 748
- Number of features : 4

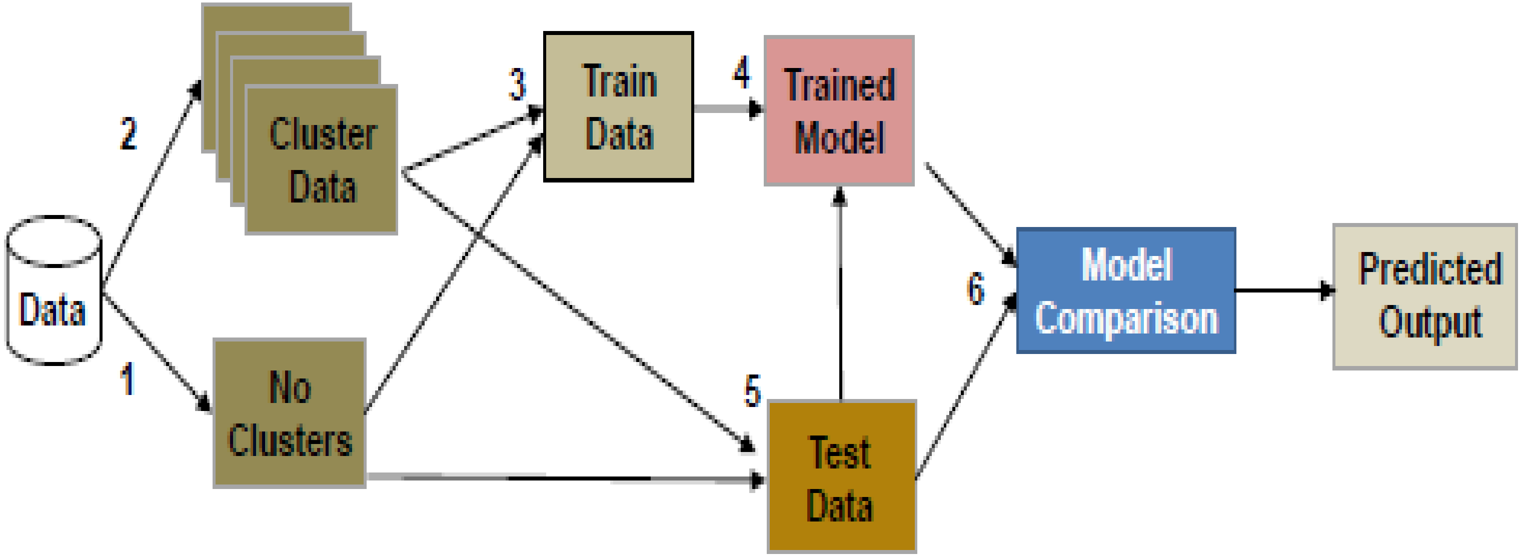
Data Dictionary

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
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

There are no clear boundaries of separation between donors and non donors



We used combination of both supervised and unsupervised learning methods to find the best model

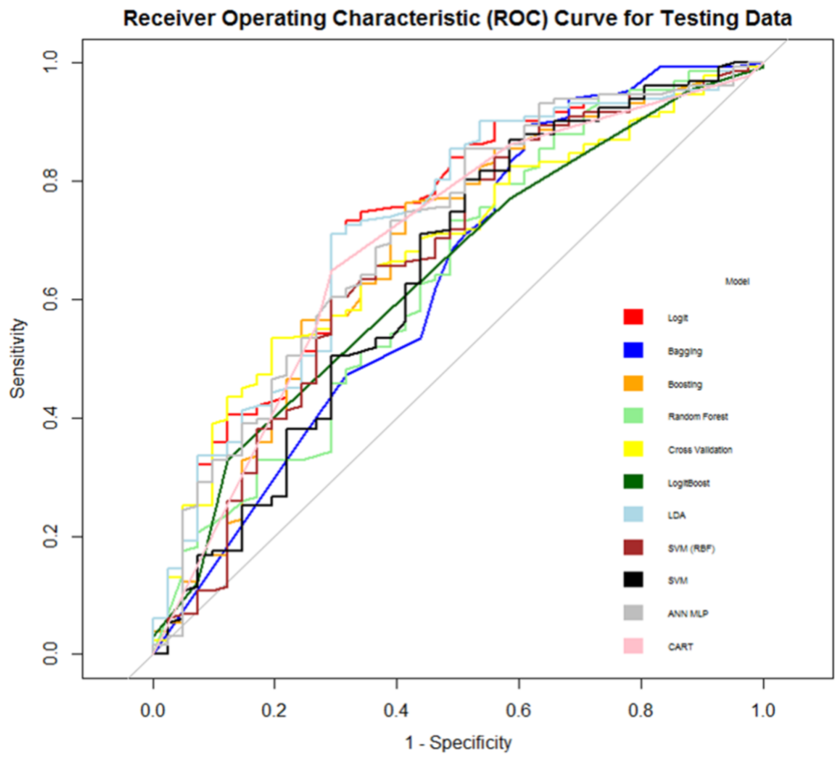


Models from previous studies	New Models
Support Vector Machines	Logistic Regression(logit)
Artificial Neural Network (MLP)	Boosted version of Logit
CART	Bagged version of logit
C5.0	Ensemble Methods
LDA	Random Forest

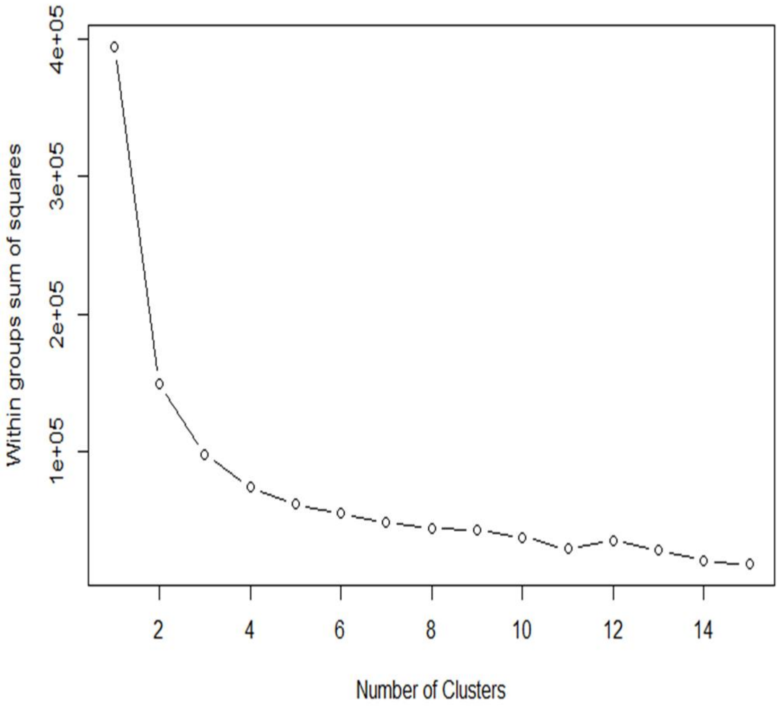
Using supervised learning method, C5.0 method had the highest accuracy

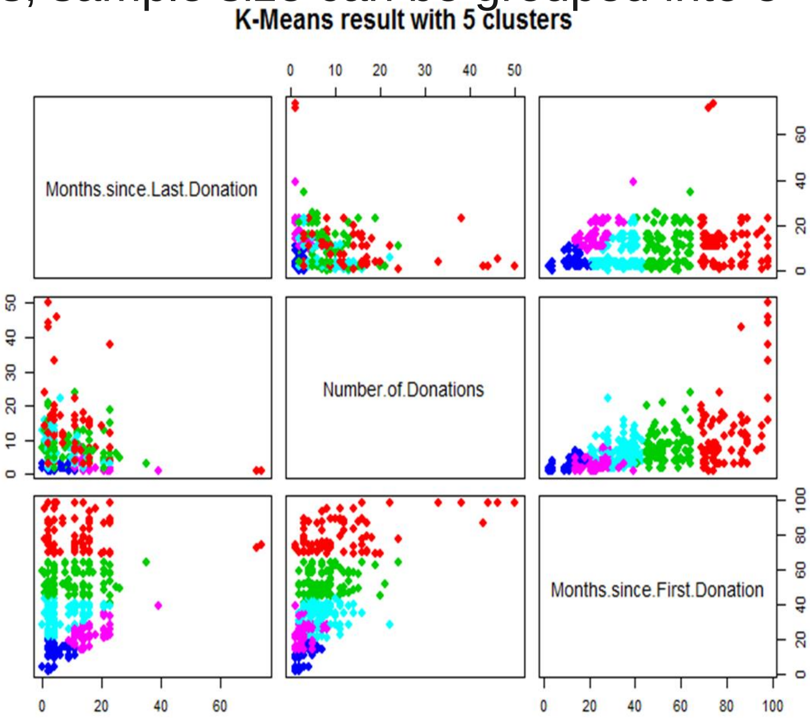
		Training				Testing			
		Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
No Clustering	ANN	0.8610	0.9348	0.6220	0.7635	0.8372	0.8931	0.6585	0.7190
	C5.0	0.8836	0.9576	0.6494	0.7688	0.8837	0.9236	0.7560	0.6809
	CART	0.8143	0.9218	0.5054	0.7629	0.7965	0.8625	0.5853	0.6937
	Logistic 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

Logistic Regression model has the highest AUC in non clustered model.

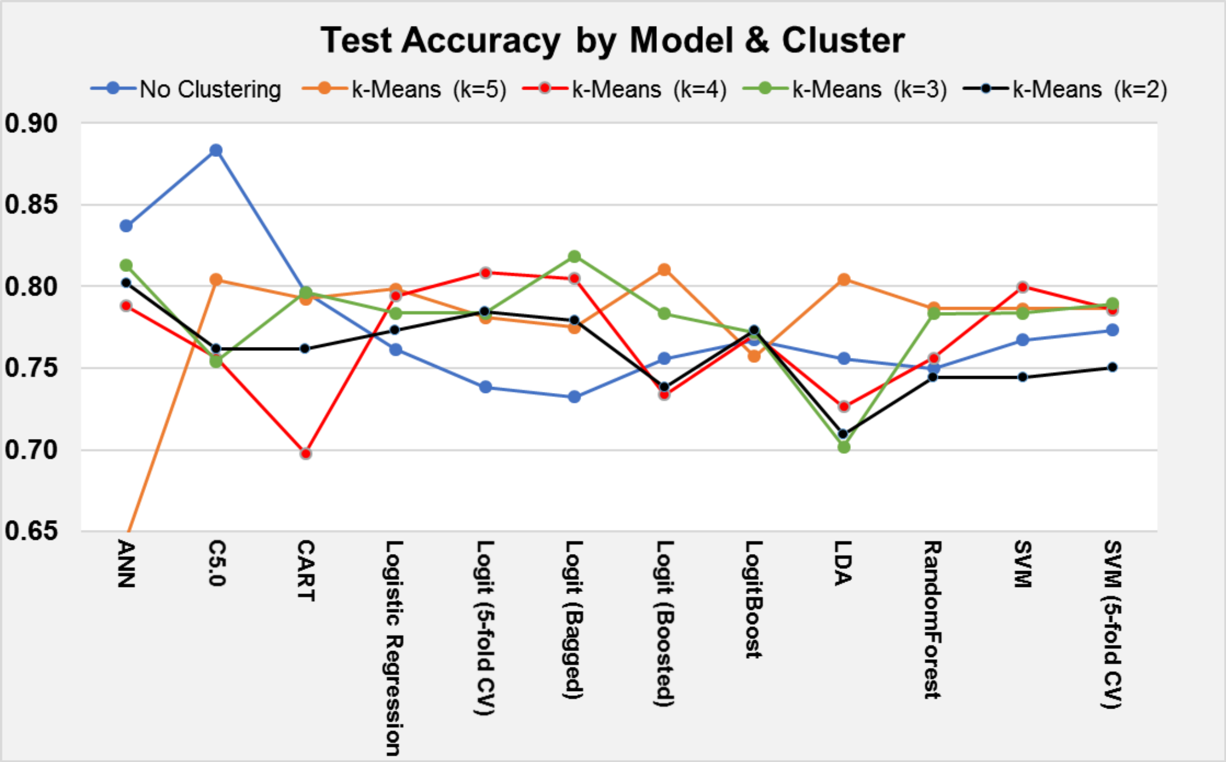


Based on associate between features, sample size can be grouped into 5 clusters.

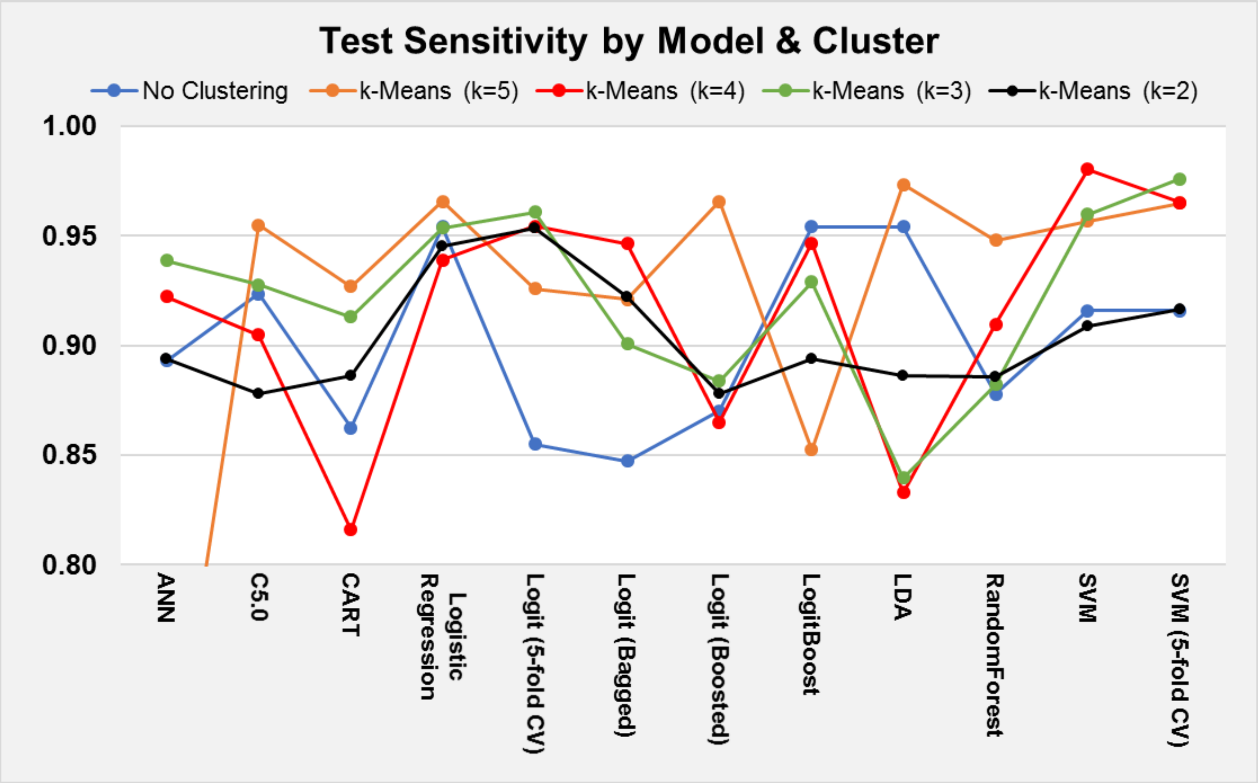




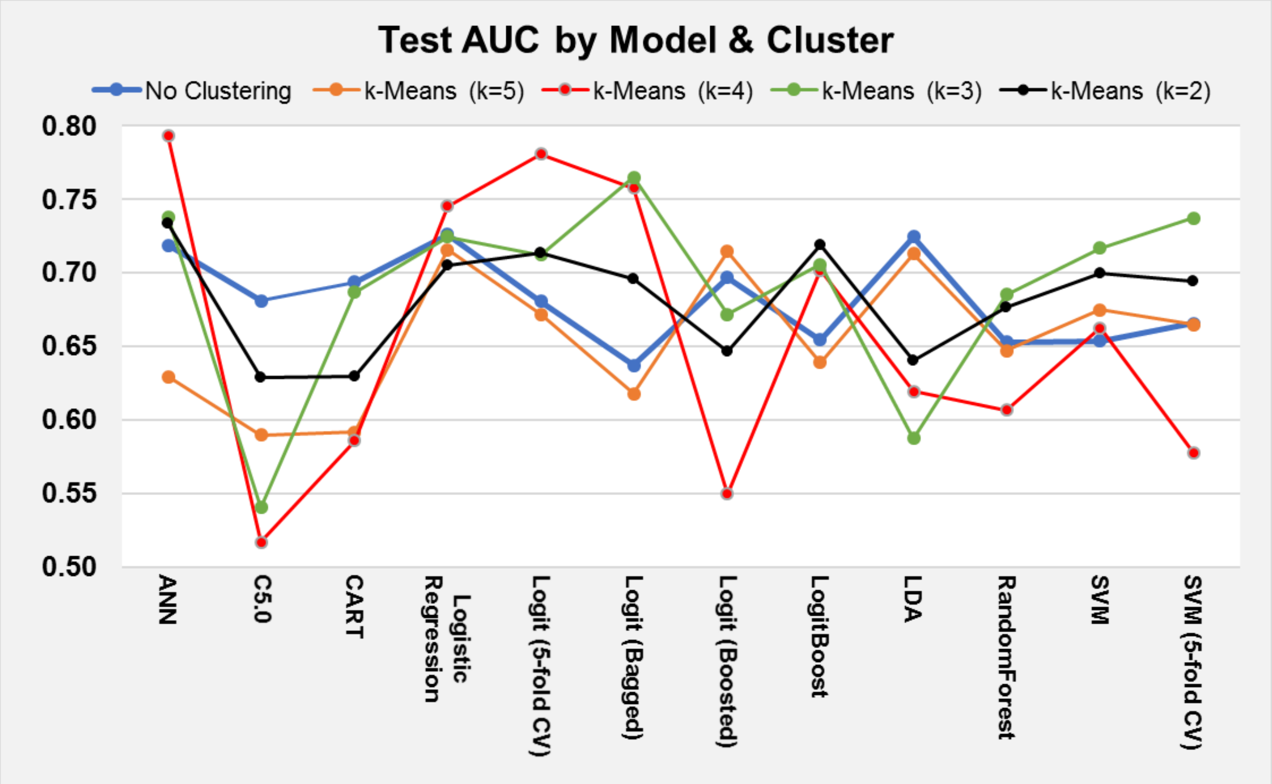
C5.0 model gives the highest accuracy in non clustered model



SVM model gives the highest sensitivity in clustered model with K=4



ANN model gives the highest AUC in clustered model with K=5



- 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.
- AUC alone may not be the best measure with respect to likelihood to predict blood.
- Focusing on targeted donors leads to using a clustered (k=4) SVM model.

Next steps

- More variables such as age, gender will help in improving the model.
- Evaluate the models on the basis of cost associated with each model and find the expected value of if somebody donates or not donates.

Deepti Bahel
(dbahel@purdue.edu)