Noé Flores
Predict Blood Donations

Introduction:
There are few commodities that impact humanity the way blood does. Human blood is life for mankind, and while some luckily never have a need for this commodity in excess of what we naturally produce, others are not as fortunate, and the availability of blood is literally the difference between life and death. This is what makes blood so precious and why blood donations are so important to have good modeling behind them.

The problem presented for us is to build a predictive model that would enable a mobile blood transfusion service to predict whether or not a previous blood donor will be willing to donate blood when the vehicle returns to their location. Solving this problem is significant for several reasons. As we previously mentioned, blood is a precious commodity most of us take for granted until we need it and when it becomes a necessity it will either be available or not. Blood isn't something that can be produced in a lab, and therefore we depend on the good nature of blood donors to provide the supply.

We will perform an exploratory analysis of our data set to gather information regarding the variables we have available to incorporate into this research and follow up with any data preparation needed to ensure that we are working with optimal data. Finally, we will build and compare several models to determine which has the most significant predictive value, interpretability, and practicality with respect to our goal of determining likely repeat blood donors for our mobile collection team.

Data Analysis:
Our dataset contains 576 observations of 6 different variables. The table provides a breakdown and description of those variables along with the abbreviations we will use for each variable in an effort to simplify the interpretability of future analysis.

Figure 1: Variables and Descriptions.

Blood donation Data	**Contents**
Total Observations	576
Variables	**Abbreviation**
X	Tag
Months Since Last Donation	MonLD
Number of Donations	NumD
Total Volume Donated	TotVol
Months Since First Donation	MonFD
Made Donation in March 2007	MadeDM

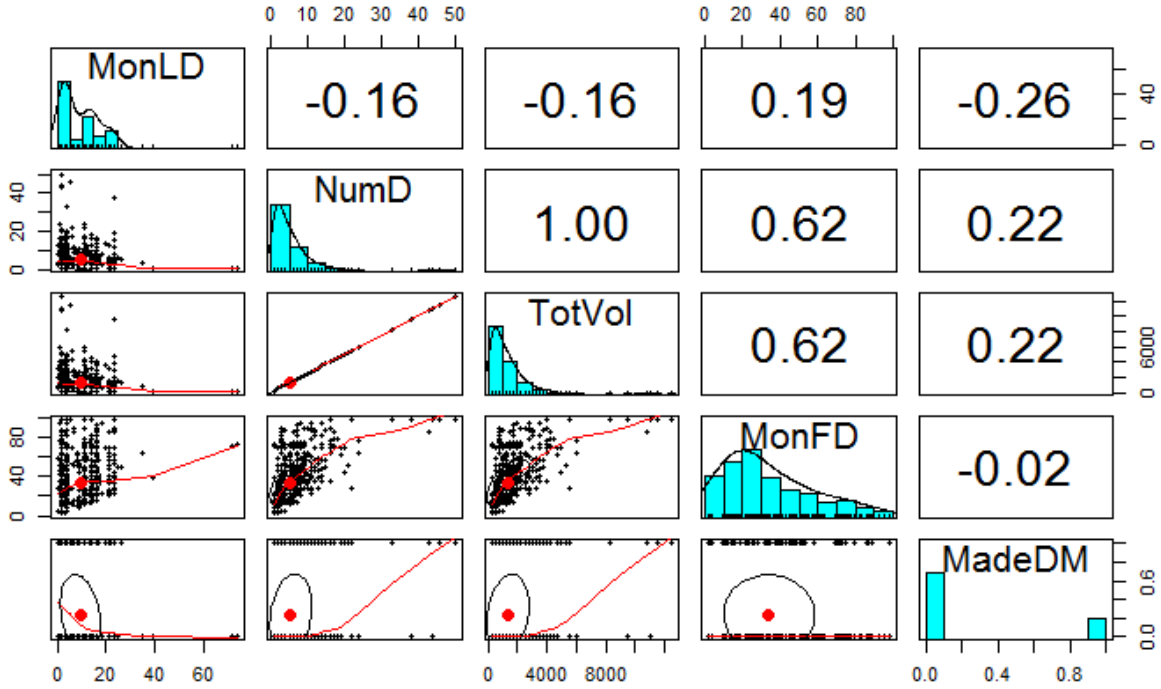
We provide a breakdown of the measures of central tendency for our data. Immediately present in this table is the min and max range of each variable. In some instances, there appears to be a relatively large range which could be an indication of potential outliers that we may need to address

Figure 2: Measures of central Tendency

MonLD	NumD	TotVol	MonFD
Min. : 0.000	Min. : 1.000	Min. : 250	Min. : 2.00
1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.: 500	1st Qu.:16.00
Median : 7.000	Median : 4.000	Median : 1000	Median :28.00
Mean : 9.439	Mean : 5.427	Mean : 1357	Mean :34.05
3rd Qu.:14.000	3rd Qu.: 7.000	3rd Qu.: 1750	3rd Qu.:49.25
Max. :74.000	Max. :50.000	Max. :12500	Max. :98.00

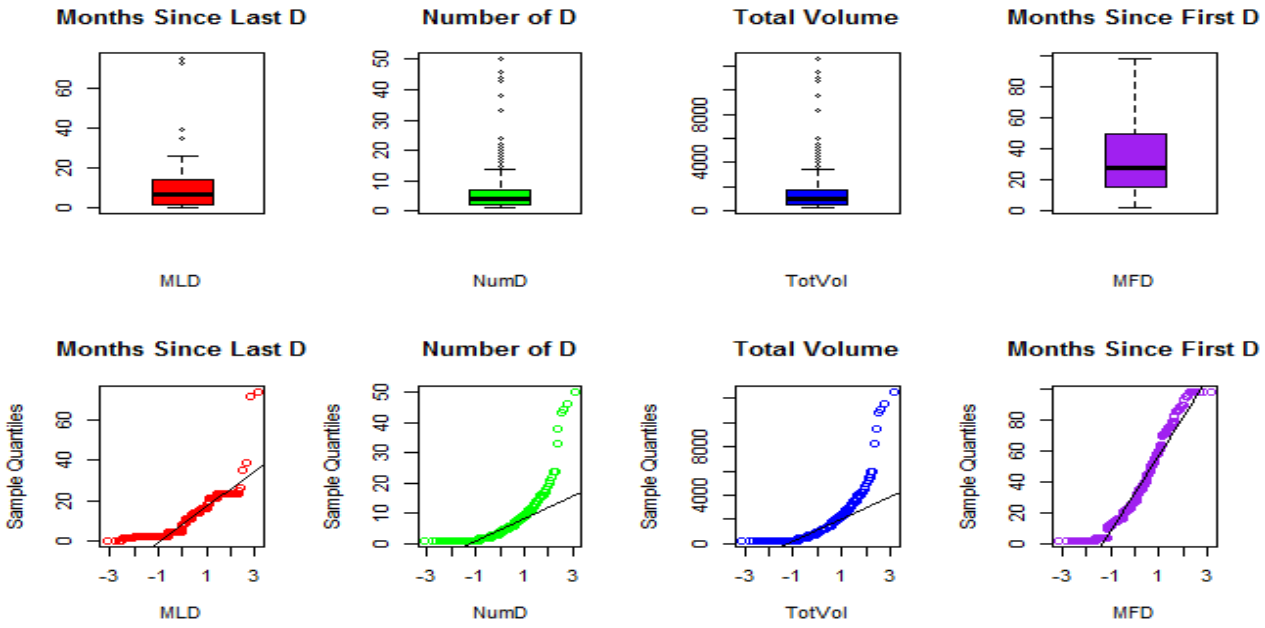
The correlation matrix below provides us with some interesting information about our variables. The first observation presented to us is that Number of Donations (NumD) and Total Volume Donated (TotVol) are perfectly correlated to one another. We can also observe from the histograms present that some of our variables don't appear to be normally distributed and deserve additional exploration before proceeding with our model building.

Figure 3: Correlation Matrix.



The box-plots and Q-Q plots provide a clear view of the variable distributions. and possible outliers. It's clear we have some outliers present in Months since last Donation (MonLD) and Number of Donations (NumD), and this is something we suspected from our measures of Central Tendency. We can also confirm from the Q-Q plot that out variables aren't normally distributed and could possibly benefit from variable transformation.

Figure 4: Box-plots and Q-Q plot of variables.



Data Preparation and Processing:
We have a strong aversion towards removing or treating outlying data points, solely for the sake of attempting to force a model to fit better. We believe those data points are there for a reason. However, after going through our exploratory analysis, we decided to cap the upper observations of Months Since last Donation (MonLD) and Number of Donations (NumD) at their respective 95% values rather than remove the more extreme data points.

We also incorporated BoxCox transformations on some of the variables in an effort to increase normality and created and added two new variables to the data, donations per month (DperM) and donation ratio (Dratio) in an attempt to get a more robust look at the data through the addition of two new independent variables.

Figure 5: Data Preparation Summary

Existing Variable Transformations	Capped at 95%	BoxCox Variable transformation and addition
Months Since Last Donation (MonLD)	Yes	Yes
Months Since First Donation (MonFD)	No	Yes
Number of Donations	Yes	Yes
Total Volume Donated	No	Yes
New Variables Added		
Donations Per Month (DperM)		
Donation Ratio (Dratio)		

Model Research and Literature:
We are dealing with a situation where we have to predict a binary outcome. The literature listed below contains information relative to our task of predicting blood donations in the future. Each piece includes examples of the models we elected to use in developing predictive models for binary outcomes. After researching the literature, we decided to focus on simple linear and logistic regression techniques that were intuitive to users but powerful enough to produce statistically significant results.

1. Alaeddini, A., Yang, K, Reddy, C, Yu, S (2010). *A probabilistic Model for predicting the probability of no-show in hospital appointments.*
2. Chatla, S. B., & Shmueli, G. (2017). *An extensive examination of regression models with a binary outcome variable.*
3. Ge, W., & Whitmore, G. A. (2010). *Binary response and logistic regression in recent accounting research publications:*
4. Youn, H., & Gu, Z. (2010). *Predict US restaurant firm failures: The artificial neural network model versus logistic regression model.*
5. Bonney, G. (1987). *Logistic Regression for Dependent Binary Observations. ,*

Modeling Building:
Our work will center on building four distinct models. Two model will be simple linear regression models, and the other two will be logistic regression models. These models will also differ in their composition and variable selection.

Model 1 linear Regression
model1 <- lm(formula = MadeDM ~ MonLD + NumD + MonFD + TotVol+ DperM + Dratio, data = train)

Model 1 is a simple linear regression model incorporating all of the original non-transformed predictor variables from our data set as well as the two new variables we added to the analysis.

Model 2 linear Regression
model2 <- lm(formula = MadeDM ~ BMonLD + BTotVol + MonFD + BNumD + DperM + Dratio, data=train)

Model 2 is a logistic regression model incorporating the BoxCox transformed variables from our dataset as well as the two new variables we added to the analysis.

Model 3 logistic Regression
model3 <- glm(formula = MadeDM ~ MonLD + NumD + MonFD + TotVol + DperM + Dratio, data = train, family = binomial(logit))

Model 3 is a logistic regression model incorporating all of the non-transformed predictor variables from our data set as well as the two new variables we added to the analysis.

Model 4 logistic Regression
model4 <- glm(formula = MadeDM ~ BMonLD + BNumD + MonFD + BTotVol + DperM + Dratio, data = train, family = binomial(logit))

Model 4 is a logistic regression model incorporating the BoxCox transformed variables from our dataset as well as the two new variables we added to the analysis.

Modeling Selection:
Before submitting our models for judgment, we used the Akaike information criterion (AIC) to choose our top three models for submission. The results are below.

Figure 6: Model AIC

Model	AIC
Model 1	577.13
Model 2	579.37
Model 3	557.35
Model 4	554.36

Based on the AIC metric, Model 1, 3, and 4 should be our top models coming out of the driven data blood donation challenge, with Model 4 as the winner.

Figure 6: Driven Data Results

The results of the drivendata.org assessment coincide with my expectations and Model 4 was the better choice albeit not a flattering choice.

Limitations:
The main drawback for me with this analysis is the limited amount of data points. Perhaps certain categorical variables associated with education level, employment, relationship status, etc can assist with this prediction analysis. Even with the additional variables, we would certainly benefit from a more robust set of observations greater than the 576 we worked with in this project.

Future work and learning:
In future models of this nature where there are limited data points, I would look to experiment with more sophisticated modeling techniques, variable transformations, and perhaps create a more dynamic variable to assist in our model building. In this particular case, we made use of time tested regression models. Incorporating Neural Networks or Decision Tree's could have helped us produce more statistically significant results but I didn't feel confident producing a model of that nature without fully grasping its functionality. Even without the use of more advanced techniques, we learned that we should build many models and have a set criteria for model selection. I was happy to see that the AIC was a good indicator of the results I should have expected from driven data. Finally, I learned that modeling is tough. I am not satisfied with the results, but I think this drawback is simply a limitation of the modeling techniques I incorporated during this specific analysis and it is something that I will look to improved upon in the future.

References:

Alaeddini, A., Yang, K, Reddy, C, Yu, S (2010). A probabilistic Model for predicting the probability of no-show in hospital appointments. *Health Care Management Science, 14*(2), 146-57. doi:http://dx.doi.org.turing.library.northwestern.edu/10.1007/s10729-011-9148-9.

Chatla, S. B., & Shmueli, G. (2017). An extensive examination of regression models with a binary outcome variable. *Journal of the Association for Information Systems, 18*(4), 340-371. Retrieved from http://turing.library.northwestern.edu/login?url=https://search-proquest-com.turing.library.northwestern.edu/docview/1897788483?accountid=12861

Ge, W., & Whitmore, G. A. (2010). Binary response and logistic regression in recent accounting research publications: A methodological note. *Review of Quantitative Finance and Accounting, 34*(1), 81-93. doi:http://dx.doi.org.turing.library.northwestern.edu/10.1007/s11156-009-0123-1

Youn, H., & Gu, Z. (2010). Predict US restaurant firm failures: The artificial neural network model versus logistic regression model. *Tourism and Hospitality Research, 10*(3), 171-187. doi:http://dx.doi.org.turing.library.northwestern.edu/10.1057/thr.2010.2

Lahiri, K., & Liu, Y. (2012). *Forecasting binary outcomes.* St. Louis: Federal Reserve Bank of St Louis. Retrieved from http://turing.library.northwestern.edu/login?url=https://search-proquest-com.turing.library.northwestern.edu/docview/1698841071?accountid=12861

Bonney, G. (1987). Logistic Regression for Dependent Binary Observations. *Biometrics, 43*(4), 951-973. doi:10.2307/2531548

Appendix: R-Code

```
###413 Midterm DrivenData

# Loading the data

train <- read.csv(file.path("C:/Users/Noe/Downloads/blood_Donations_Train.csv"),sep=",")

test <- read.csv(file.path("C:/Users/Noe/Downloads/blood_Donations_Test.csv"),sep=",")

###Start###

require(ggplot2)

require(forecast)

require(psych)

require(readr)

require(dplyr)

require(corrplot)

require(outliers)

require(rpart)

require(MASS)

###check my data train

str(train)

head(train)

tail(train)

###check my data train

str(test)

head(test)
```

```
tail(test)

###Exploratory Data Analysis and descriptive Stata

###Measures of central tendency and Summary stats

summary(train)

###Rename Variables for simplicity

colnames(train) <- c("Tag","MonLD", "NumD", "TotVol",  "MonFD", "MadeDM")

colnames(test)  <- c("Tag","MonLD", "NumD", "TotVol", "MonFD")

###Simple Correlations

cor(train[,2:6])

pairs.panels(train[c("MonLD", "NumD", "TotVol",  "MonFD", "MadeDM")])

### boxplots, and q-q plots of MLD, NumD, MFD

par(mfrow = c(2,4))

boxplot(train$MonLD, col = "red",main ="Months Since Last D" , xlab = "MLD")

boxplot(train$NumD, col = "green",main ="Number of D", xlab = "NumD")

boxplot(train$TotVol, col = "blue",main ="Total Volume", xlab = "TotVol")

boxplot(train$MonFD, col = "purple",main ="Months Since First D", xlab="MFD")

qqnorm(train$MonLD, col = "red",main = "Months Since Last D" , xlab = "MLD")

qqline(train$MonLD, col = "black",main ="")

qqnorm(train$NumD, col = "green",main = "Number of D", xlab = "NumD")

qqline(train$NumD, col = "black",main ="")

qqnorm(train$TotVol, col = "blue",main = "Total Volume", xlab = "TotVol")

qqline(train$TotVol, col = "black",main ="")
```

```
qqnorm(train$MonFD, col = "purple",main ="Months Since First D", xlab = "MFD")

qqline(train$MonFD, col = "black",main ="")

###Data Preparation

fun <- function(x){

  quantiles <- quantile( x, c(.00, .95 ) )

  x[ x < quantiles[1] ] <- quantiles[1]

  x[ x > quantiles[2] ] <- quantiles[2]

  x

}


fun(train$MonLD)

fun(train$NumD)

###Add to train dataset

train <- train %>% mutate(MonLD = fun(train$MonLD))

train <- train %>% mutate(NumD = fun(train$NumD))

###correct and add to test dataset

test <- test %>% mutate(MonLD = fun(test$MonLD))

test <- test %>% mutate(NumD = fun(test$NumD))

###Add new variables to train and test data

train <- train %>% mutate(DperM = NumD/MonFD)

train <- train %>% mutate(Dratio = MonLD/MonFD)

test <- test %>% mutate(DperM = NumD/MonFD)
```

```
test <- test %>% mutate(Dratio = MonLD/MonFD)

###BoxCox Transformations

###lambda value for BoxCox

lambda1 <- BoxCox.lambda(train$MonLD)

lambda1

###lambda value for BoxCox

lambda2 <- BoxCox.lambda(train$NumD)

lambda2

###lambda value for BoxCox

lambda3 <- BoxCox.lambda(train$TotVol)

lambda3

###Add new variables to train and test data

train$BMonLD <- BoxCox(train$MonLD,lambda1)

train$BNumD <- BoxCox(train$NumD,lambda2)

train$BTotVol <- BoxCox(train$TotVol,lambda3)

test$BMonLD <- BoxCox(test$MonLD,lambda1)

test$BNumD <- BoxCox(test$NumD,lambda2)

test$BTotVol <- BoxCox(test$TotVol,lambda3)

###check Data

summary(train)

summary(test)

###Build Models
```

```
################################################################################
          ##
###Model 1 Multiple Regression
model1 <- lm(formula = MadeDM ~ MonLD + NumD  + MonFD + TotVol+ DperM +
             Dratio, data = train)
summary(model1)
####Predict
predict1<-predict(model1, newdata = test, type = "response")
predict1<-data.frame(Tag=test$Tag,donate = predict1)
predict1<-select(predict1, -donate.0)
write.csv(predict1, file = "model1.csv", row.names = FALSE)
###Model 2
model2 <- lm(formula = MadeDM ~ BMonLD + BTotVol + MonFD + BNumD +
             DperM + Dratio, data=train)
summary(model2)
###predict
predict2<-predict(model2, newdata = test, type = "response")
predict2<-data.frame(Tag=test$Tag,donate = predict2)
predict2<-select(predict2, -donate.0)
write.csv(predict2, file = "model2.csv", row.names = FALSE)
###Model 3 logistic
model3 <- glm(formula = MadeDM ~ MonLD + NumD  + MonFD + TotVol + DperM + Dratio,
```

```
        data = train, family = binomial(logit))

summary(model3)

###Predict

predict3<-predict(model3, newdata = test, type = "response")

predict3<-data.frame(Tag=test$Tag,donate = predict3)

predict3<-select(predict3, -donate.0)

write.csv(predict3, file = "model3.csv", row.names = FALSE)

###Model 4

model4 <- glm(formula = MadeDM ~ BMonLD + BNumD  + MonFD + BTotVol + DperM +

        Dratio,

         data = train, family = binomial(logit))


summary(model4)

predict4<-predict(model4, newdata = test, type = "response")

predict4<-data.frame(Tag=test$Tag,donate = predict4)

predict4<-select(predict4, -donate.0)

write.csv(predict4, file = "model4.csv", row.names = FALSE)

###Model Analysis for selection of top 3

AIC(model1,model2,model3,model4)
```