

|  |
| --- |
| DFL Donation Predictions |
|  |
| May 5  Authored by: Dylan Zetah |

# Introduction and Background

In recent history, politicians and political parties have increasingly relied on statistical models for both inference and prediction (Patterson, 2018; Garrison, 2020). This paper will focus on the prediction of donations made to the DFL party with an emphasis on the accuracy of the model. Furthermore, the goal is to use this model to choose which constituents to solicit donations from via mail. The data used to build this model was gathered from the 2020 election, so all information should be up to date to ensure the highest degree of accuracy. As requested, two models were fitted: a zero-inflated Gaussian model with a logarithmic transformation on the response variable and a random forest regression model.

The zero-inflated Gaussian model with a logarithmic transformation on the response variable is a special type linear model from the gaussian, or normal, family. Furthermore, it differs from a typical linear model because it accounts for an abnormal number of zeros in the response variable (Mills, 2013). This means that other than the zeros, the response variable is approximately normally distributed after the logarithmic transformation. However, the transformation was only applied to the non-zero values because the logarithm of zero is undefined. Therefore, this model is the perfect fit for a response variable that is approximately normally distributed with an excessive number of zeros after a logarithmic transformation on the non-zero responses.

For the second model, I used random forest regression, which utilizes regression trees and a type of bootstrapping called bagging. One significant advantage of regression trees is that it is non-parametric and makes no assumptions on the distribution of the response variable (James et al., 2014). This is especially helpful because it can handle abnormally distributed data without requiring a transformation. Additionally, bagging is a way to lower the variance of regression trees by using a bootstrap (Algeri 2021). This is accomplished by taking repeated samples from the data set and constructing a decision tree for each one, then averaging the resulting predictions (Algeri 2021). Furthermore, random forest regression improves bagging by building trees that are decorrelated. Specifically, each tree that is built from the bootstrap only considers a fraction of the predictors, typically the square root of the total predictors (Algeri 2021). This decorrelates the trees by considering more than just the strongest predictors, so every tree does not have the same predictors at the first, second, or third split (Algeri 2021). Hence, random forest regression is a good fit for the data because it does not assume a distribution for the response variable, and it can reliably make predictions while maintaining low variance.

# Methods and Materials

The data used for this study was collected by Governor Walz's staff in the form of a mailed survey. There was a total of 1000 participants that were chosen at random, with the non-response rate not recorded. The survey contained a total of 9 questions which were all answered by each participant. Each survey recorded the participant's political party affiliation, age, level of education (bachelor's degree or higher or not), yearly income, gender, race, ethnicity(Hispanic or non-Hispanic), location(urban, suburban or rural), and the amount of money donated to DFL candidates in the past year.

The zero-inflated Gaussian model with a logarithmic transformation on the response variable produces a model of the form:

The assumptions of this model are independence, non-zero responses following the normal distribution after transformation, and large sample size. The assumption of independence is verified by the fact that the survey participants were not related to each other in any way. Additionally, since there were 1000 participants, the assumption of a large sample size is verified. Finally, the QQ plot (fig. 1) and histogram (fig. 2) of log-transformed responses confirm that the data is approximately normally distributed, albeit slightly skewed at the tails.

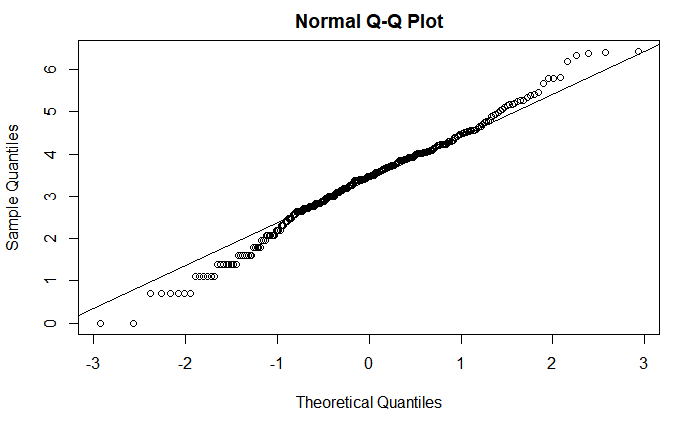


fig. 1

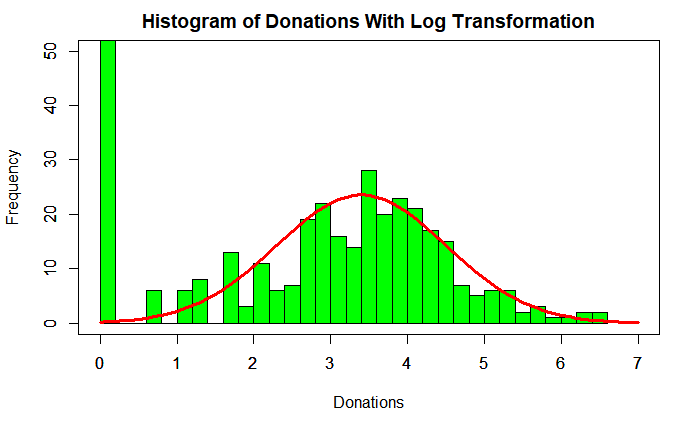


Fig. 2

~Normal(mu = 3.4,sd = 1)

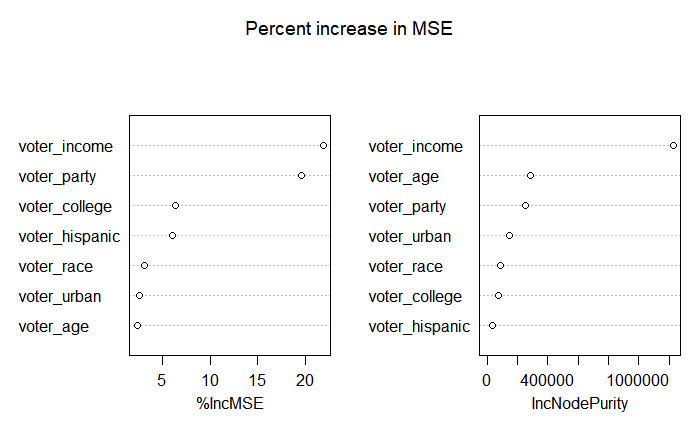
Random forest regression is a non-parametric model, so it has no assumptions to verify. Additionally, random forest regression model cannot be quantified or shown in a simple way because it is an average of 500 trees, so there is no formula to show. However, we can see how influential each variable is by looking at its importance plot. To produce this plot, R randomly permutes each predictor while keeping the other predictors the same and measures how much the MSE increases. Therefore this plot is an effective way to quantify and rank how significant each predictor is.

To compare both of these models, I will be using 10-fold cross-validation, a type of validation set approach (Algeri 2021). When performing 10-fold cross-validation, the first step is to split the data into 10 "folds" or sets. Next, designate one of the folds as the validation set and the rest of the data as the training set. Then construct a model using the training set and calculate the MSE using the validation set. Finally, repeat this procedure for the nine other folds and average the resulting MSEs to obtain the 10-fold cross-validated MSE (Algeri 2021).

# Results

The final model produced by the linear model with logarithmic transformation on the response variable was:

Predicted Donation = exp(.882 + 1.688 \*(Independent) + 2.204 \*(Republican) + .0000473\*(income) + .485\*(Suburban) + .632\*(Urban) + .203\*(college) - .0000434\*(Independent\*Income) - .0000529(Republican\*Income) -.743\*(Independent\*Suburban)-. 0.5317 \*(Republican\*Suburban) - 0.4886 \*(Independent\*Urban) - 1.0682 \*(Republican\*Urban))

Where Independent, Republican, Urban, and Suburban are all dummy variables. This model had a cross-validated MSE of 1629.55, which means each prediction was off by an average of about 40.

As previously stated, there is not a good way to show the random forest regression model. Still, we can see how important each variable is in determining the predicted donation with the importance plot (fig. 3). This shows that voter income and voter party are the most influential variables. Also, while less important, college education and ethnicity play a small role, and race, location, and age play a tiny role in predicting donations with the random forest regression model. Furthermore, gender was taken out of the model entirely because it actually caused a decrease in MSE when it was randomly permuted. The random forest model ended up performing better than the zero-inflated Gaussian model with a log transformation with a cross-validated MSE of 1425.579, which means each prediction was off by about 38 on average.

fig. 3

# Discussion and Summary

The results of the zero-inflated Gaussian model with a logarithmic transformation on the response variable are very encouraging. Even though it had a higher cross-validated MSE than the random forest model, it is easier to interpret and visualize. This model used party, location, income, and college education and had interactions between party and income as well as party and location. This model would be better for finding characteristics of someone who donates a lot instead of trying to predict the exact donation value.

The results of the random forest regression model were also beneficial. While hard to visualize and interpret, this model was more accurate and valued income, party, college education, and ethnicity. This model would perform better if you are directly imputing data into the model to see what the predicted donation would be.

There were a few models that I strongly considered in my analysis that I ended up not using. At first, the zero-inflated Poisson model seemed like a logical choice, but the data was too overdispersed, so I ended up abandoning that approach for the zero-inflated negative binomial model. The zero-inflated negative binomial model performed almost as well in terms of cross-validated MSE and used the same predictors as the zero-inflated Gaussian model with a logarithmic transformation. However, I think with a little more tweaking and maybe adding more predictors, a zero-inflated negative binomial model could end up performing as good, if not better, than the zero-inflated Gaussian model with a logarithmic transformation. One other model to consider in the future would be boosting, a type of regression tree that grows trees sequentially instead of taking a bootstrap. Therefore, the models I chose to fit are not the only options for analyzing this data.

Overall, both models provide valuable insight into who to target with donation requests for Governor Walz's team. However, both models could be improved with a larger sample size and more predictors such as number of children, occupation, and multiple years of donation history. These improvements would make the models more accurate and allow for better interpretation. In conclusion, I hope these models help Governor Walz secure as many donations as possible, and I am excited to see what's next for his campaign.

# References

Algeri, S. (2021). *Handout 9 - Treess*. *STAT 4052*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *An Introduction to Statistical Learning with Applications in R*. *An introduction to Statistical Learning* (Vol. 7). https://www.statlearning.com/.

Garrison, S. (2020, December 17). *Using Statistical Inference in the Practice of Politics*. Swash Labs. https://swashlabs.com/blog/using-statistical-inference-in-the-practice-of-politics.

Mills, E. (2013). *Adjusting for co Adjusting for covariates in z ariates in zero-inflated gamma and z o-inflated gamma and zero-inflated o-inflated log-normal models for semicontinuous data* (dissertation).

Patterson, D. (2018, November 7). *How campaigns use big data tools to micro-target voters*. CBS News. https://www.cbsnews.com/news/election-campaigns-big-data-analytics/.

# Appendix

## R Code

data = read.csv("~/STAT 4052/2020 DFL Data v2.csv")

donodata = data[,-c(9,10)]

yesdono = donodata[donodata$donations!=0,]

logdons = ifelse(donodata[,9]==0, 0,log(donodata[,9]))

logdono = donodata

logdono[,9] = logdons

qqnorm(log(yesdono$donations))

qqline(log(yesdono$donations))

hist(logdono$donations,

breaks = 25,

main = "Histogram of Donations With Log Transformation",

xlab = "Donations",

col = "green",

ylim = c(0,50),xlim = c(0,7))

x = seq(0,7,.1)

y = 65\*dnorm(x, mean = 3.4,sd = 1.1)

par(new=TRUE)

plot(x,y,ylim = c(0,50),xlim = c(0,7),type = "l",col = "red", lwd = 3,xlab = "",ylab = "")

library("NBZIMM")

gzig = lme.zig(donations~voter\_party\*voter\_income+voter\_party\*voter\_urban+voter\_college,~1|ones,zi.fixed = ~voter\_income^2,data = logdono)

summary(gzig)

library(tree)

library(randomForest)

set.seed(1234)

n=nrow(donodata)

train.id=sample(1:n,(9\*n)/10)

train=donodata[train.id,]

val=donodata[-train.id,]

m4 = randomForest(donations~.-voter\_female,data=train,mtry=3,importance=TRUE)

varImpPlot(m4, main = "Percent increase in MSE")

library(tree)

library("NBZIMM")

library(randomForest)

require("randomForest")

library(rpart)

library(rattle)

library(rpart.plot)

library(DAAG)

nfolds = 10

KCVerr1=0

KCVerr2=0

KCVerr3=0

library(caret)

fold = createFolds(1:1000,k = nfolds,list = F)

for(i in 1:nfolds)

{

m1 = zeroinfl(donations~voter\_party+(voter\_income)|poly(voter\_income,2),data = donodata[fold!=i,],dist = "negbin",link = "log")

m2 = lme.zig(donations~voter\_party\*voter\_income+voter\_party\*voter\_urban+voter\_college,~1|ones,zi.fixed = ~poly(voter\_income,2),data = logdono[fold!=i,],verbose = F)

m3 = randomForest(donations~.-voter\_female,data=donodata[fold!=i,],mtry=3,importance=TRUE)

pred1 = predict(m1,donodata[fold==i,])

pred2 = predict(m2,logdono[fold==i,])

pred3 = predict(m3,donodata[fold==i,])

KCVerr1 = KCVerr1 + mean((donodata$donations[fold==i]-pred1)^2)

KCVerr2 = KCVerr2 + mean((donodata$donations[fold==i]-exp(pred2))^2)

KCVerr3 = KCVerr3 + mean((donodata$donations[fold==i]-pred3)^2)

}

KCVerr1= KCVerr1/nfolds

KCVerr2= KCVerr2/nfolds

KCVerr3= KCVerr3/nfolds

data.frame(KCVerr1,KCVerr2,KCVerr3)

## R Output

Random effects:

Formula: ~1 | ones

(Intercept) Residual

StdDev: 4.585208e-05 0.760508

Variance function:

Structure: fixed weights

Formula: ~invwt

Fixed effects: donations ~ voter\_party \* voter\_income + voter\_party \* voter\_urban + voter\_college

Correlation:

(Intr) vtr\_pI vtr\_pR vtr\_nc vtr\_rS vtr\_rU vtr\_cl vt\_I:\_ vt\_R:\_

voter\_partyIndependent -0.682

voter\_partyRepublican -0.748 0.515

voter\_income -0.741 0.473 0.538

voter\_urbanSuburban -0.456 0.313 0.342 -0.031

voter\_urbanUrban -0.411 0.278 0.306 -0.074 0.597

voter\_college 0.187 -0.030 -0.091 -0.462 -0.069 -0.100

voter\_partyIndependent:voter\_income 0.454 -0.754 -0.343 -0.553 0.041 0.079 0.032

voter\_partyRepublican:voter\_income 0.489 -0.337 -0.706 -0.601 0.042 0.082 0.058 0.395

voter\_partyIndependent:voter\_urbanSuburban 0.303 -0.470 -0.228 0.027 -0.670 -0.399 0.034 -0.010 -0.029

voter\_partyRepublican:voter\_urbanSuburban 0.319 -0.216 -0.468 0.009 -0.691 -0.414 0.074 -0.027 -0.071

voter\_partyIndependent:voter\_urbanUrban 0.267 -0.435 -0.202 0.079 -0.401 -0.672 0.005 -0.056 -0.059

voter\_partyRepublican:voter\_urbanUrban 0.290 -0.196 -0.405 0.050 -0.421 -0.704 0.075 -0.056 -0.178

v\_I:\_S v\_R:\_S v\_I:\_U

voter\_partyIndependent

voter\_partyRepublican

voter\_income

voter\_urbanSuburban

voter\_urbanUrban

voter\_college

voter\_partyIndependent:voter\_income

voter\_partyRepublican:voter\_income

voter\_partyIndependent:voter\_urbanSuburban

voter\_partyRepublican:voter\_urbanSuburban 0.462

voter\_partyIndependent:voter\_urbanUrban 0.550 0.276

voter\_partyRepublican:voter\_urbanUrban 0.281 0.580 0.473

Standardized Within-Group Residuals:

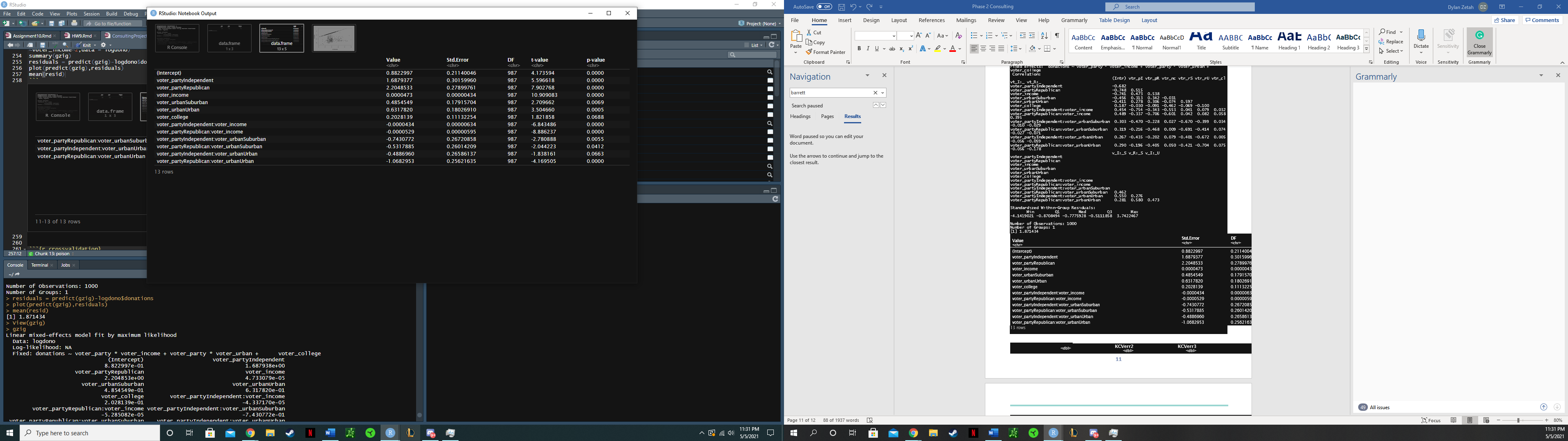
Min Q1 Med Q3 Max

-4.1419021 -0.8708494 -0.7775928 -0.5111858 3.7422467

Number of Observations: 1000

Number of Groups: 1

[1] 1.871434



KCVerr1

<dbl>

KCVerr2

<dbl>

KCVerr3

<dbl>

2196.727 1622.851 1524.251