Humana

Humana-Mays Healthcare Analytics Case Competition

Transportation Issue Descriptive & Predictive Project

By Deseret Analytics

Table of Contents

Executive Summary

This project aims to provide Humana Healthcare with more profound and broader insights into what traits and characteristics most associate with patients' likelihood of having transportation issues that prevent them from getting to medical appointments, meetings, work, or getting necessities for daily living. The second aim is to provide accurate predictions on which patients will most likely have issues with transportation.

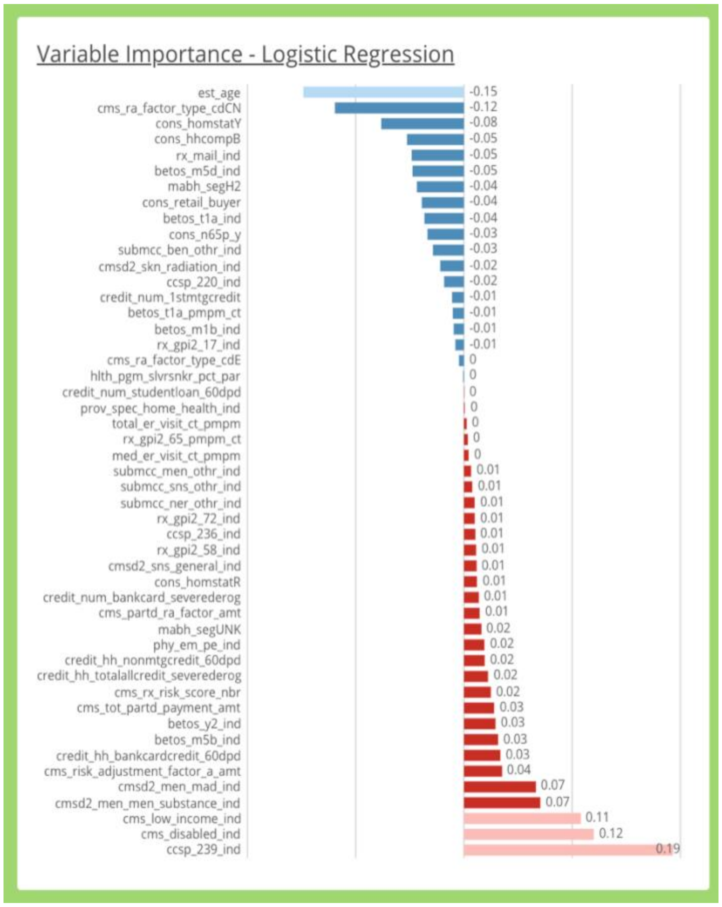
The generalized linear model will determine the most influential variables from 826 available variables provided by the competition committee. The model identifies 49 variables that are most associative with the patients that have transportation issues. Several of them give fascinating insights that lead to the discovery of a subgroup that has a distinct trait that separates it from the rest of the observations. This subgroup plays an essential role in the process of formulating recommendations for this project.

This project's predictive part will involve various data cleaning and preparation methods to prepare it for training several machine learning algorithms on RStudio. The Gradient Boosting Machine model outperforms other models when verified against the testing dataset derived from the original data. The Receiver Operating Characteristic (ROC) score of the best model is 75%, which is not very high for most classification modeling. However, because the source of the competition data is a survey, it is generally acceptable to have a ROC that is lower than normally accepted. Due to the nature of the survey, which is often incomplete and prone to human error, having a high ROC score can lead to questions about the data's validity or issues with bias.

Descriptive Analysis

Determining Most Influential Variables

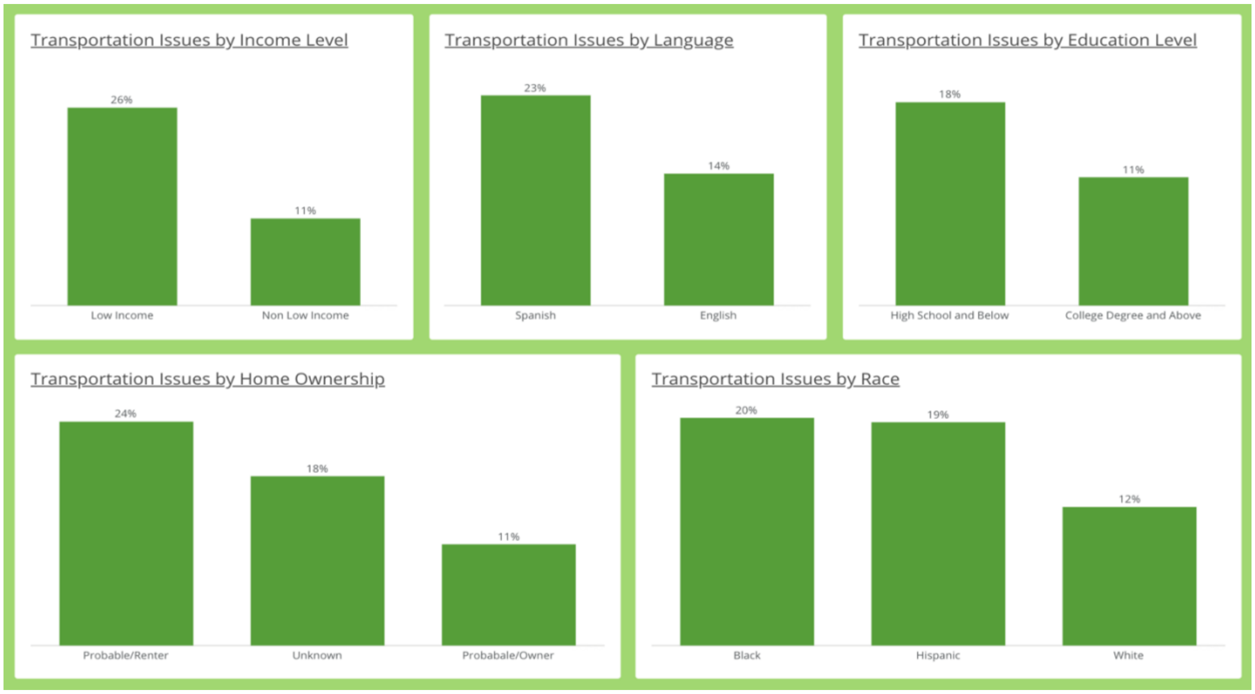
The generalized linear model with lasso regularization provides an accurate and reliable method to determine which variables are most influential in the patient's likelihood of having transportation issues. In addition to providing the statistical significance of each variable through its p-value, the model also provides the magnitude and direction of each variable in affecting the odds of having transportation issues. Here is a graph that visualizes the 49 essential variables and their direction of influence (the negative and positive signs illustrate the direction).



The red bars represent positive influence, which means that the higher the variable's number gets, the larger impact it has on whether or not the patient has transportation issues. The blue bar is the opposite of the red bars meaning there is an inverse relationship between that variable and the patient having transportation issues. The light-red and light-blue bars represent est_age, low_income_indicator, disable_indicator, and ccsp_239_ind variables that have meaningful insights and will be discussed shortly.

Low Income and Other Demographics

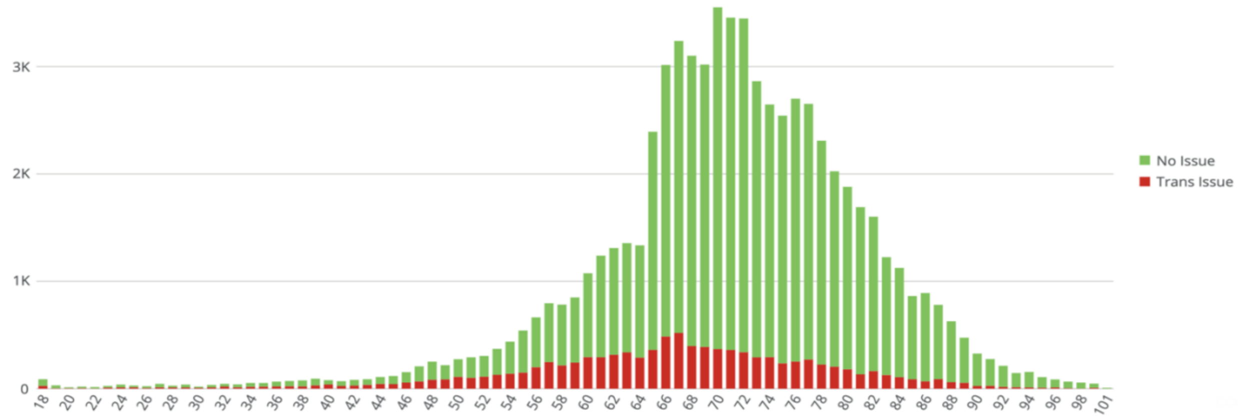
There are certain characteristics or traits of the patients that have a higher percentage of having transportation issues such as being considered low income, speaking Spanish, not having a college degree, not owning a house, and not being White.



All of the characteristics listed above are closely related to each other; for example, those struggling financially (low income) are less likely to have a college degree because of its expensive tuition. Furthermore, not having a college degree may limit their career prospects, thus keeping their income levels below those of college graduates. Other relationships can easily be drawn from any combinations of the traits above. The data show that those communities and demographics are more likely to have issues with transportation to fulfill their daily needs, such as going to work or going to a medical appointment.

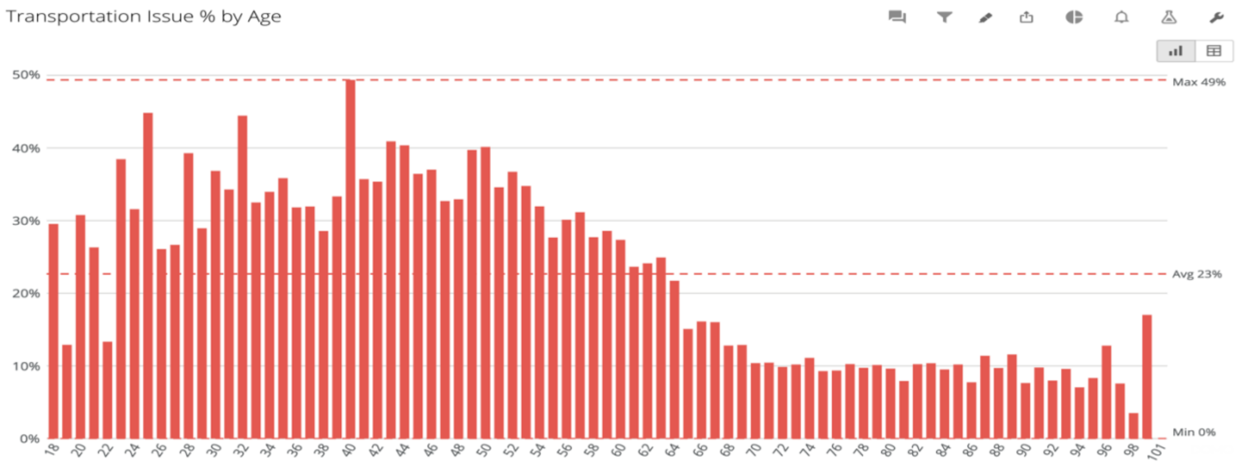
Estimated Age and Subgroup

Patients with transportation difficulties are somewhat evenly distributed when laid out against their estimated age, as shown in the graph below:



The first and third quartiles of the patient's estimated age are 66 and 77. This means that patients between 66 and 77 years old make up 50% of the total number of patients in the original data. It can safely be concluded that most observations or data points represent older individuals.

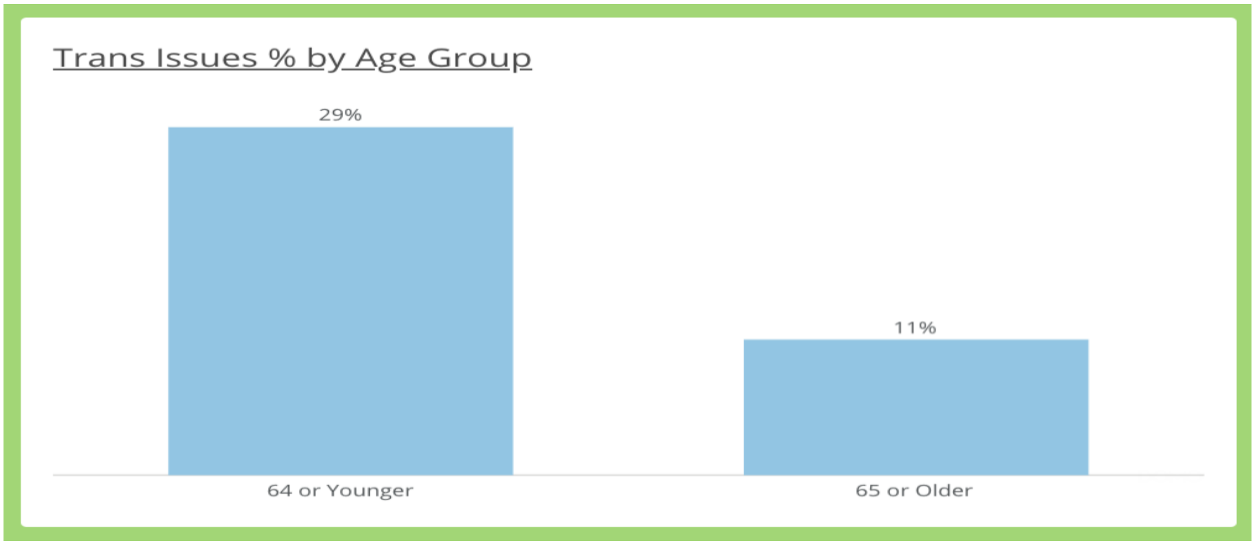
However, when patients with transportation issues are compared to the total number and grouped by their estimated age, we find a very insightful pattern.



Patients who are younger have a higher chance of having transportation issues compared to older patients. One might ordinarily expect the opposite to be true. This suggests that there is a particular group within the original data that needs to be treated differently.

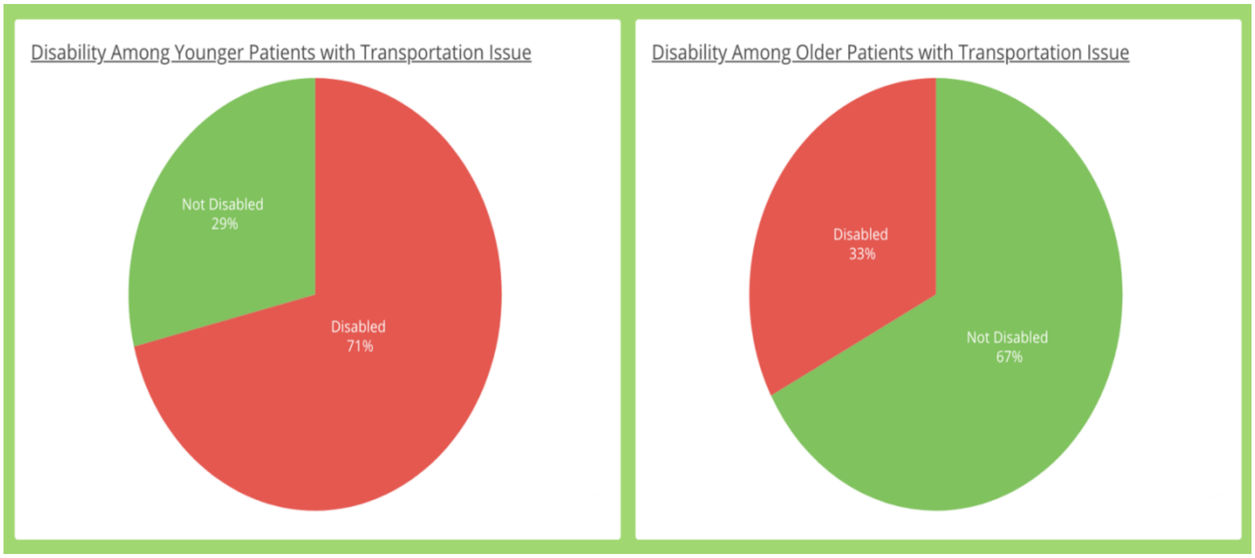
A simple decision tree model will be used to accurately determine the cutoff point to separate the subgroup from the rest of the group. A C.50 decision tree algorithm determines that age 64 should be the age cutoff between the subgroup and the main group. There are nearly 14,000 patients who fall into the subgroup, and slightly over 55,000 patients who fall into the main group.

The subgroup with patients who are 64 years old or younger will be labeled as "Younger," and the main group with patients who are older than 64 will be labeled as "Older." Here is a simple visualization of the percentage of transportation issues by age group.



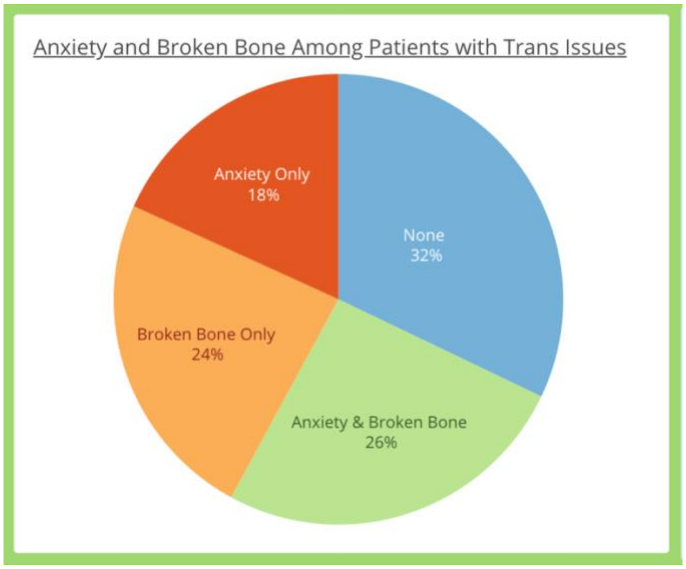
Disability and Transportation Issue

In reviewing the "younger" group, there is a distinct trait that is very prevalent among them that will help us gain a deeper understanding of why this group has a much higher percentage of having transportation issues than the "older" group.



The pie chart on the left represents the comparison of total patients in the "younger" group by their disability status; red means there is a disability and green means there is no disability. The pie chart on the right represents the comparison of total patients in the "older" group by their disability status with red still meaning there is a disability and green still meaning there is no disability. It is clearly visualized on the graphs above that the majority of those "younger" patients who have transportation issues have a disability.

Overall Common Health Condition



The majority (68%) of those patients with transportation issues have some kind of anxiety disorder and/or a broken bone. These groups of patients require different approaches from Humana Healthcare to help them overcome transportation issues, especially when they're related to going to medical appointments. For example, patients with an anxiety disorder may not require in-person care, while patients with broken bones will likely need in-person care at a clinic or hospital for X-Ray or for setting the bone

For the breakdown of how anxiety and broken bone categories are calculated, please refer to the Appendix 3.

Conclusion and Recommendation

The patients in the original data can be divided into two groups. The younger group is comprised of those patients who are 64 years old or younger. The older group is comprised of patients who are over age 65. The main difference between these groups is that the majority of younger patients with transportation issues have a disability while the majority of older patients with transportation issues do not have a disability. It can be concluded that disability is a deciding factor as to why younger patients have issues with transportation.

In general, regardless of age and disability, having transportation issues is closely related to financial and racial backgrounds. The less financially stable a patient is, the more likely it is that he or she will have problems with transportation. Non-white patients will be more likely to have transportation issues than white patients.

Helping patients struggling with transportation issues to get the necessary healthcare they need will require a different approach for each of the groups. In general, Humana healthcare should provide a phone/video call service for treatment that does not require in-person care. This will help at least 18% of those patients with transportation issues to get the care they need without having them travel to the hospital or clinic. This is also a great option for those patients who have anxiety issues or who are at high risk when contracting COVID-19, such as people with a pre-existing health condition and/or older people. Those patients who have gone through the phone/video service and have determined that additional in-person care is needed will need additional help from the Humana Healthcare to take them to the nearest hospital or clinic.

Since most of the patients struggling with transportation have low income, it is understandable that they probably do not have easy access to a personal car. If they do have a personal car, it may be difficult for them to step out from their hourly-paid job because of various reasons such as being off the clock without any paid sick time. Transportation assistance in the forms of free bus passes or discounted ride-sharing prices will decrease or eliminate transportation issues for low-income patients.

By implementing the recommendations above, those patients with transportation issues who have anxiety, broken bone, or both (approximately 68% of total patients who have transportation issues) will get their needed healthcare either through phone/video call or through in-person care with some form of transportation assistance.

Although some recommendations regarding transportation assistance include an additional cost to Humana, there will be more patients going in for their needed care. When every patient gets their needed care as soon as possible, it will reduce the risk of them needing more serious and intensive care in the future. This reduced risk will also substantially reduce the potential future cost for Humana Healthcare.

Instead of calling the transportation assistance an additional short-term cost, it will be more accurate and appropriate to call it a long-term investment.

Predictive Modeling

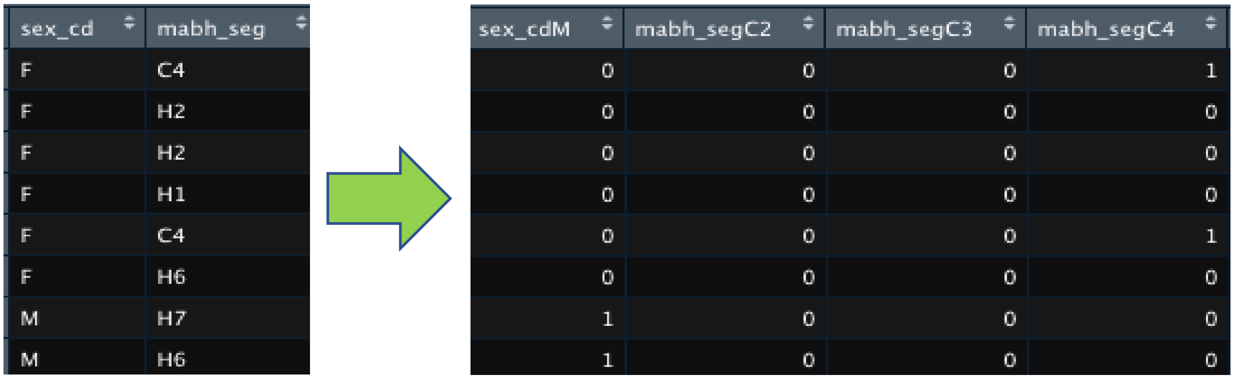
<u>Original Dataset</u>

- The dimensions of the original data are 69,572 rows and 826 variables
- “Transportation Issues” is the target variable
- 14.6% of the total patients (represented by rows) have issues with transportation
- 54,185 rows or 78% of the total number of rows have at least one missing value

<u>Data Cleaning & Preprocessing</u>

- Dummy Variable

  Create dummy variables to convert categorical variables into numerical variables. This technique produces more than 2,000 variables from the original 826 variables because some categorical variables have hundreds of different values, such as city and county variables.



sex_cd	mabh_seg
F	C4
F	H2
F	H2
F	H1
F	C4
F	H6
M	H7
M	H6

sex_cdM	mabh_segC2	mabh_segC3	mabh_segC4
0	0	0	1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	1
0	0	0	0
1	0	0	0
1	0	0	0

- Missing Value Imputation

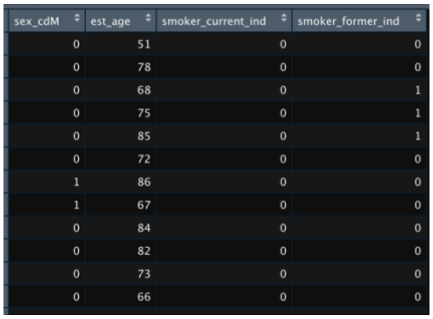
  Impute missing values from each variable with the median of its respective variable.

- Remove Near-zero-variance Variables

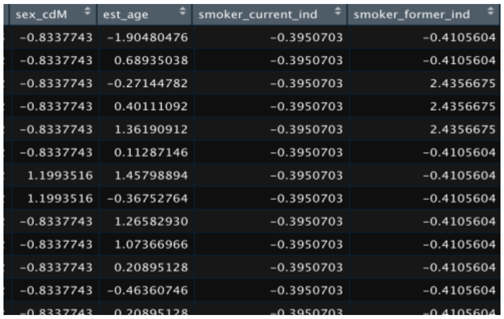
  Remove near-zero-variance variables to reduce the dimensionality and noise of the data. This technique brings down the number of variables from 2,000 to 438.

- Normalize the Data

  Center and scale preprocessing will convert each variable to have zero average and one standard deviation.



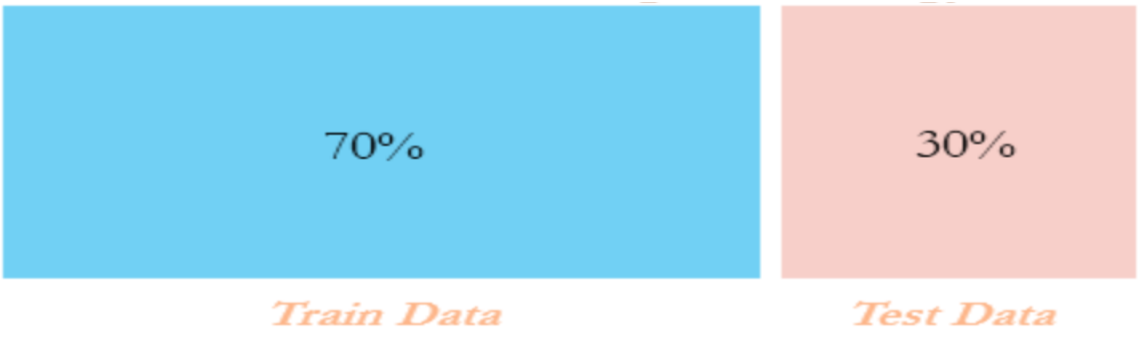
sex_cdM	est_age	smoker_current_ind	smoker_former_ind
0	51	0	0
0	78	0	0
0	68	0	1
0	75	0	1
0	85	0	1
0	72	0	0
1	86	0	0
1	67	0	0
0	84	0	0
0	82	0	0
0	73	0	0
0	66	0	0



sex_cdM	est_age	smoker_current_ind	smoker_former_ind
-0.8337743	-1.90480476	-0.3950703	-0.4105604
-0.8337743	0.68935038	-0.3950703	-0.4105604
-0.8337743	-0.27144782	-0.3950703	2.4356675
-0.8337743	0.40111092	-0.3950703	2.4356675
-0.8337743	1.36190912	-0.3950703	2.4356675
-0.8337743	0.11287146	-0.3950703	-0.4105604
1.1993516	1.45798894	-0.3950703	-0.4105604
1.1993516	-0.36752764	-0.3950703	-0.4105604
-0.8337743	1.26582930	-0.3950703	-0.4105604
-0.8337743	1.07366966	-0.3950703	-0.4105604
-0.8337743	0.20895128	-0.3950703	-0.4105604
-0.8337743	-0.46360746	-0.3950703	-0.4105604
-0.8337743	0.20895128	-0.3950703	-0.4105604

- Train and Test Data

  Randomly divide the original data into two groups for validation purposes. One subset is with 70% of the original data to train the machine learning models. The other is with 30% of the original data to test the accuracy of the trained model.



- Balance the Dataset

  The proportion of the target variable (transportation issues) is 85% with no issues and 15% for with issues. This imbalanced target proportion may cause a challenge for most machine learning algorithms because they assume that the proportion of the target is somewhat balanced. Undersampling method is used to balance the dataset into 50 : 50 proprotion. The resulting dataset has 14,276 rows with 7,138 rows from each category in the transportation issues variable.

- Penalized Regression with Lasso

  Penalized regression or generalized linear regression with lasso regularization model is an algorithm that aims to bring the coefficient of each variable to zero in respect of the target variable. The remaining variables that are not eliminated by the algorithm will be those that are most likely to have influence on the target variables. The resulting dataset of this method is a dataset with 50 variables.

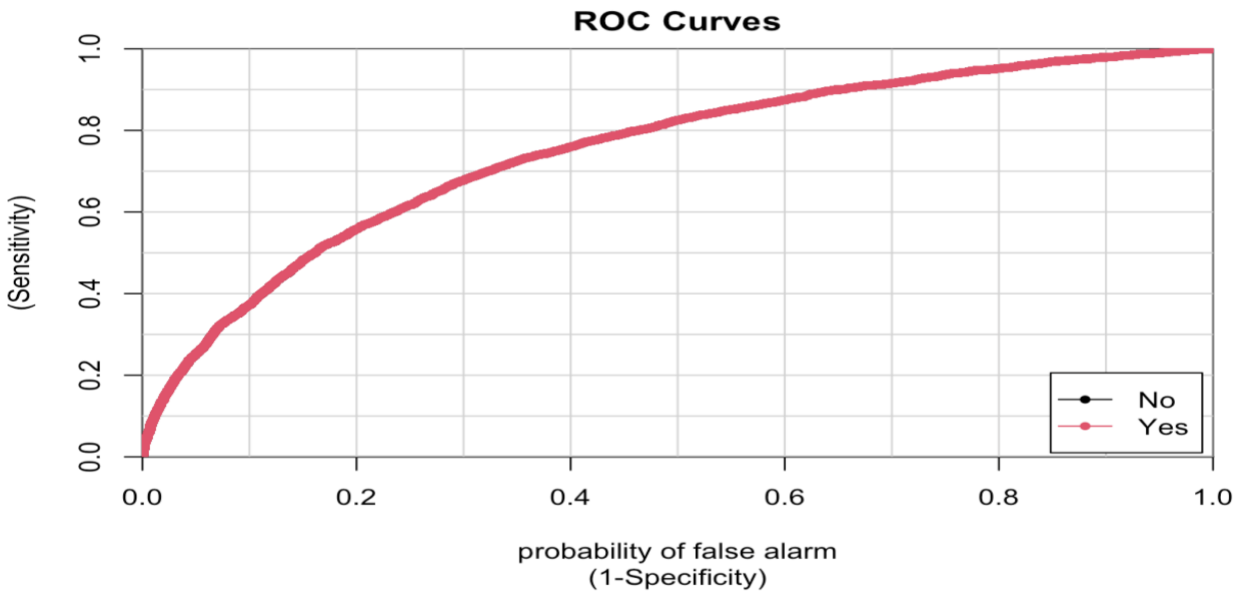
The above process and methods ultimately shrink the original dataset into a smaller data with more predictive power. The original dataset has almost 70,000 rows and more than 800 variables, the new dataset only has 14,000 rows and 50 variables (12% of the original size).

Training the Models

For this project, we trained various models with six different machine learning algorithms, such as Gradient Boosting Machine, K-Nearest Neighbors, Single-layer Neural Network, Support Vector Machine (Linear, Radial, Polynomial), Random Forest, and XGBoosting.

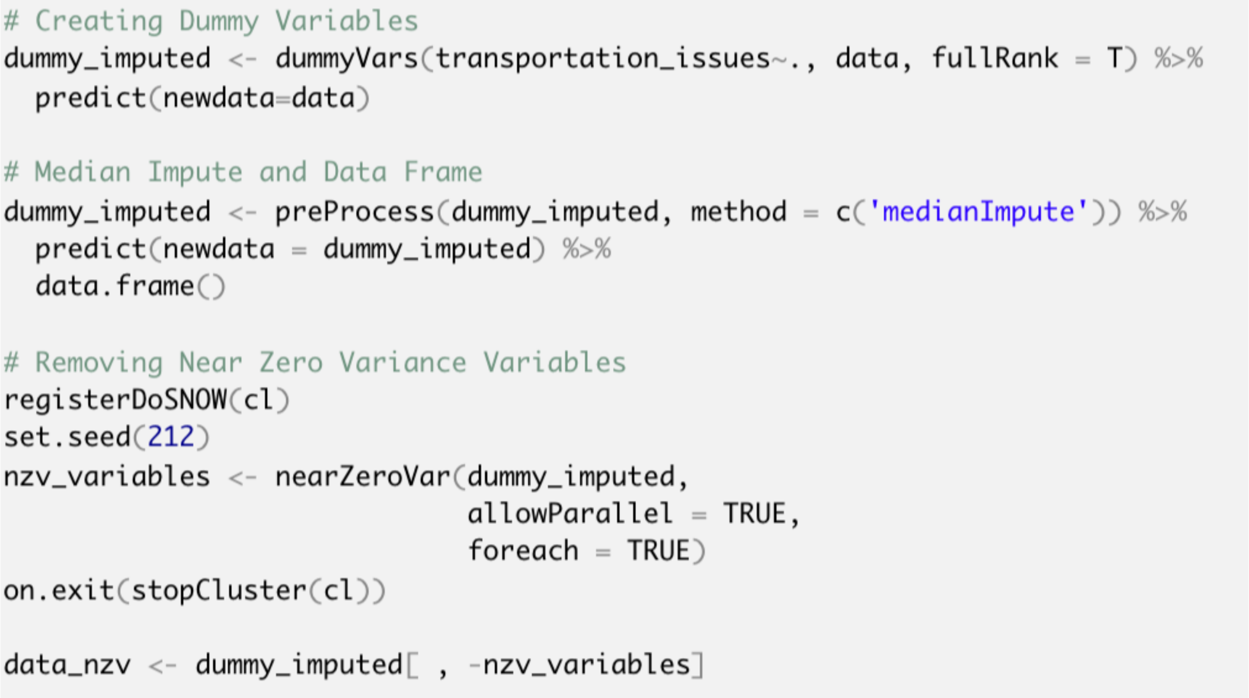
Each of the models is trained using the same train dataset with a 7-fold validation method to increase the performance. Each of the models is then used to predict the test dataset and the result was compared against the available target variable in the test dataset.

The model that was trained using Gradient Boosting Machine (Stochastic Gradient Boosting) algorithm turned out to be the best performer compared to the other models against the test data. The ROC score of this model is 75%, as shown in the graph below:

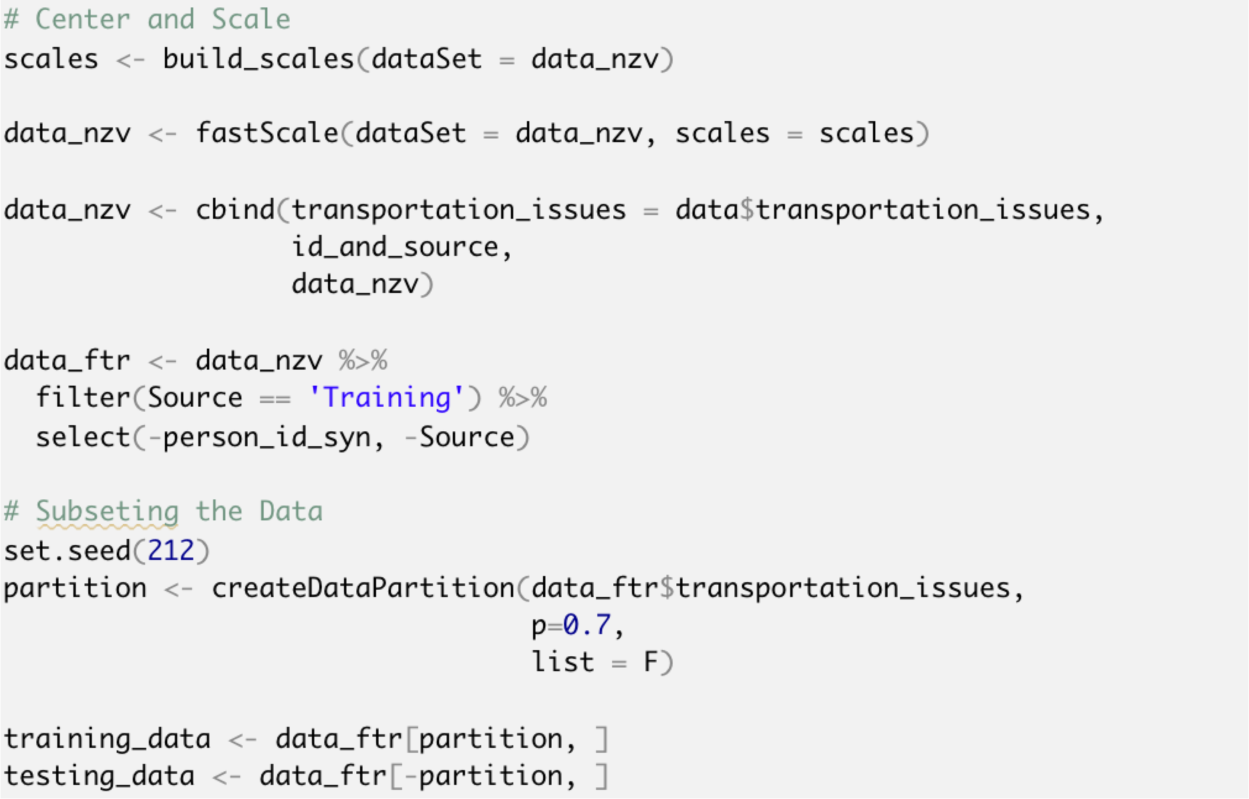


Appendix

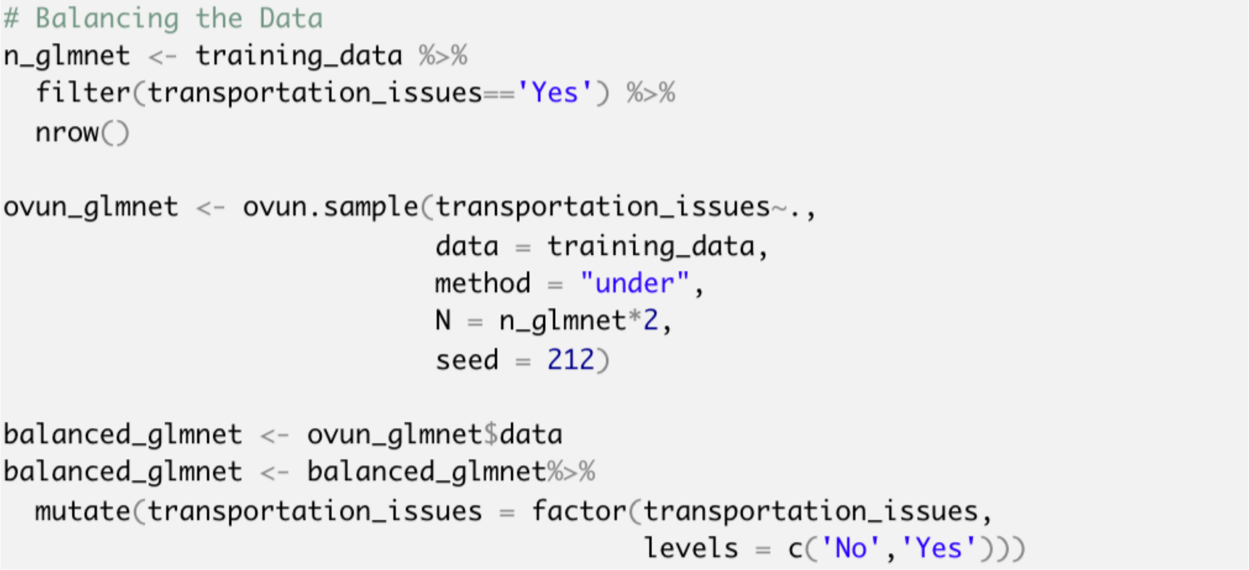
Appendix 1: R Codes



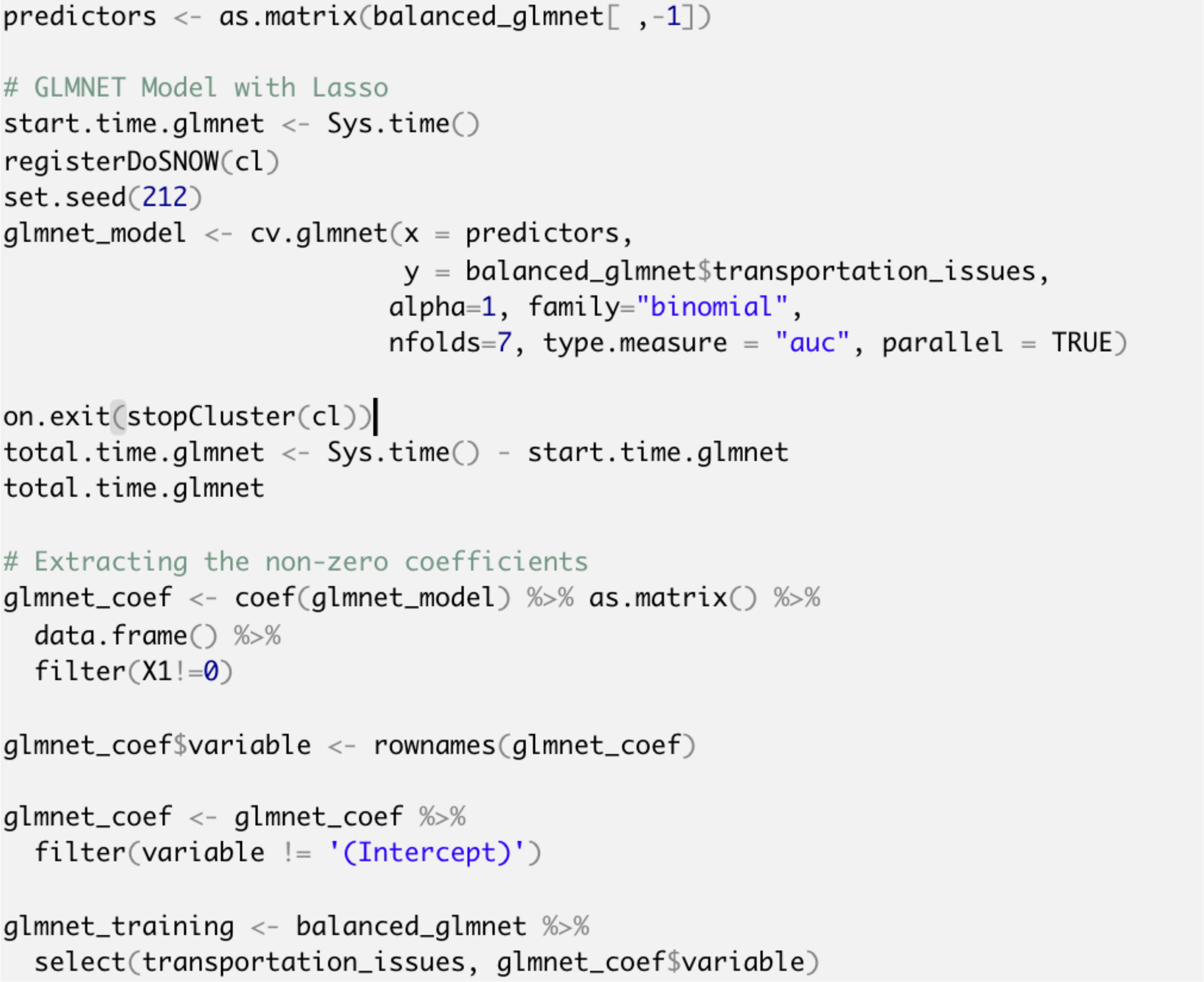
```
# Creating Dummy Variables
dummy_imputed <- dummyVars(transportation_issues~., data, fullRank = T) %>%
  predict(newdata=data)

# Median Impute and Data Frame
dummy_imputed <- preProcess(dummy_imputed, method = c('medianImpute')) %>%
  predict(newdata = dummy_imputed) %>%
  data.frame()

# Removing Near Zero Variance Variables
registerDoSNOW(cl)
set.seed(212)
nzv_variables <- nearZeroVar(dummy_imputed,
                             allowParallel = TRUE,
                             foreach = TRUE)
on.exit(stopCluster(cl))

data_nzv <- dummy_imputed[ , -nzv_variables]
```



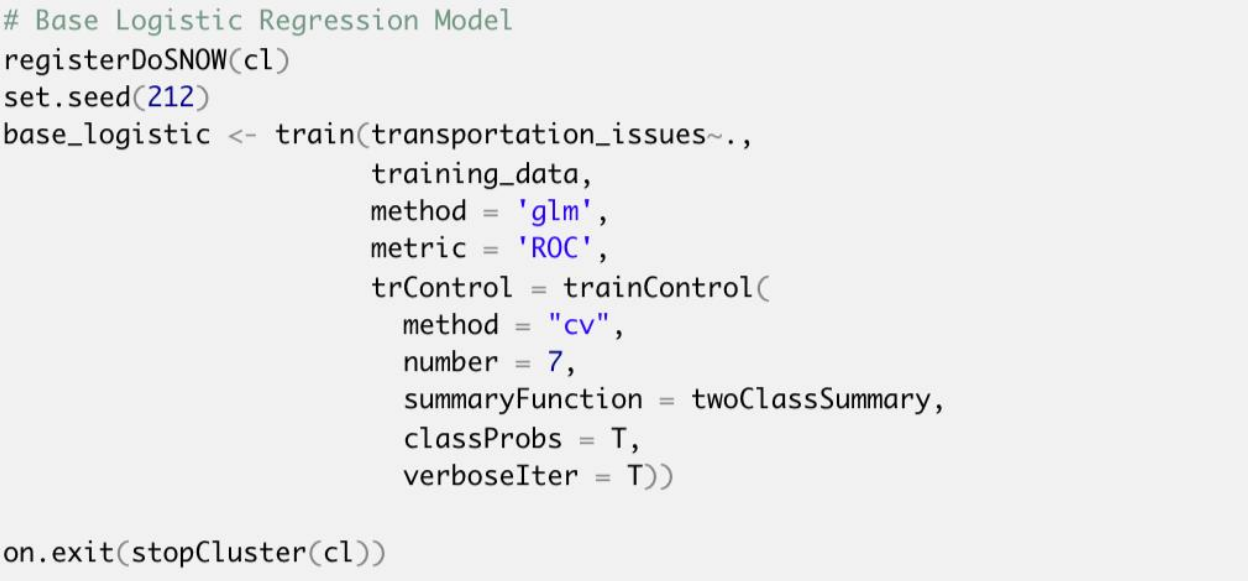
```
# Center and Scale
scales <- build_scales(dataSet = data_nzv)

data_nzv <- fastScale(dataSet = data_nzv, scales = scales)

data_nzv <- cbind(transportation_issues = data$transportation_issues,
                  id_and_source,
                  data_nzv)

data_ftr <- data_nzv %>%
  filter(Source == 'Training') %>%
  select(-person_id_syn, -Source)

# Subseting the Data
set.seed(212)
partition <- createDataPartition(data_ftr$transportation_issues,
                                 p=0.7,
                                 list = F)

training_data <- data_ftr[partition, ]
testing_data <- data_ftr[-partition, ]
```



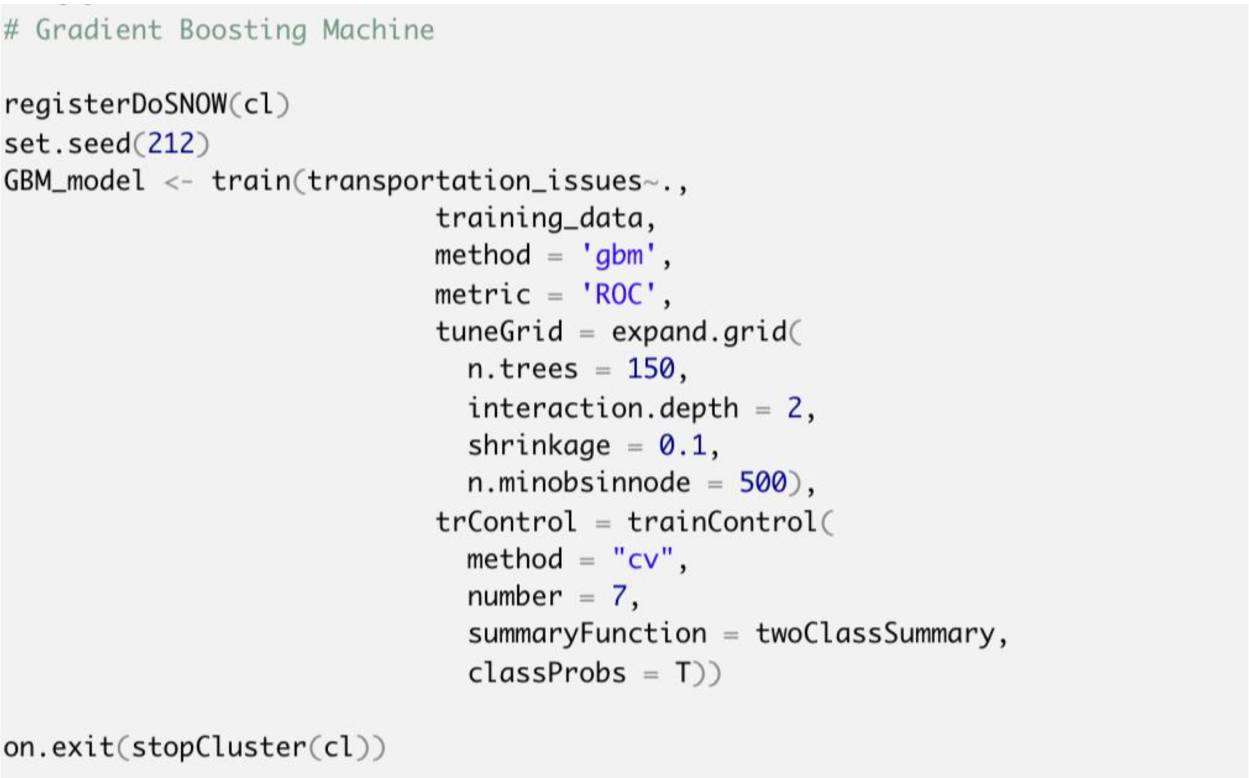
```r
# Balancing the Data
n_glmnet <- training_data %>%
  filter(transportation_issues=='Yes') %>%
  nrow()

ovun_glmnet <- ovun.sample(transportation_issues~.,
                           data = training_data,
                           method = "under",
                           N = n_glmnet*2,
                           seed = 212)

balanced_glmnet <- ovun_glmnet$data
balanced_glmnet <- balanced_glmnet%>%
  mutate(transportation_issues = factor(transportation_issues,
                                        levels = c('No','Yes')))
```



```r
predictors <- as.matrix(balanced_glmnet[ ,-1])

# GLMNET Model with Lasso
start.time.glmnet <- Sys.time()
registerDoSNOW(cl)
set.seed(212)
glmnet_model <- cv.glmnet(x = predictors,
                           y = balanced_glmnet$transportation_issues,
                         alpha=1, family="binomial",
                         nfolds=7, type.measure = "auc", parallel = TRUE)

on.exit(stopCluster(cl))
total.time.glmnet <- Sys.time() - start.time.glmnet
total.time.glmnet

# Extracting the non-zero coefficients
glmnet_coef <- coef(glmnet_model) %>% as.matrix() %>%
  data.frame() %>%
  filter(X1!=0)

glmnet_coef$variable <- rownames(glmnet_coef)

glmnet_coef <- glmnet_coef %>%
  filter(variable != '(Intercept)')

glmnet_training <- balanced_glmnet %>%
  select(transportation_issues, glmnet_coef$variable)
```



```
# Base Logistic Regression Model
registerDoSNOW(cl)
set.seed(212)
base_logistic <- train(transportation_issues~.,
                       training_data,
                       method = 'glm',
                       metric = 'ROC',
                       trControl = trainControl(
                         method = "cv",
                         number = 7,
                         summaryFunction = twoClassSummary,
                         classProbs = T,
                         verboseIter = T))

on.exit(stopCluster(cl))
```



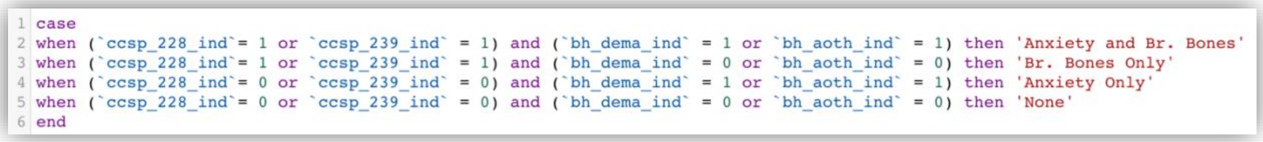
```
# Gradient Boosting Machine

registerDoSNOW(cl)
set.seed(212)
GBM_model <- train(transportation_issues~.,
                           training_data,
                           method = 'gbm',
                           metric = 'ROC',
                           tuneGrid = expand.grid(
                             n.trees = 150,
                             interaction.depth = 2,
                             shrinkage = 0.1,
                             n.minobsinnode = 500),
                           trControl = trainControl(
                             method = "cv",
                             number = 7,
                             summaryFunction = twoClassSummary,
                             classProbs = T))

on.exit(stopCluster(cl))
```

Appendix 2: Variables from Generalized Linear Model – Lasso Regularization

Variable	Coefficient
ccsp_239_ind	0.1933
cms_disabled_ind	0.1200
cms_low_income_ind	0.1081
cmsd2_men_men_substance_ind	0.0706
cmsd2_men_mad_ind	0.0665
cms_risk_adjustment_factor_a_amt	0.0353
credit_hh_bankcardcredit_60dpd	0.0337
betos_m5b_ind	0.0316
betos_y2_ind	0.0292
cms_tot_partd_payment_amt	0.0279
cms_rx_risk_score_nbr	0.0249
credit_hh_totalallcredit_severederog	0.0222
credit_hh_nonmtgcredit_60dpd	0.0192
phy_em_pe_ind	0.0189
mabh_segUNK	0.0162
cms_partd_ra_factor_amt	0.0145
credit_num_bankcard_severederog	0.0138
cons_homstatR	0.0119
cmsd2_sns_general_ind	0.0117
rx_gpi2_58_ind	0.0112
ccsp_236_ind	0.0106
rx_gpi2_72_ind	0.0100
submcc_ner_othr_ind	0.0099
submcc_sns_othr_ind	0.0077
submcc_men_othr_ind	0.0066
med_er_visit_ct_pmpm	0.0043
rx_gpi2_65_pmpm_ct	0.0036
total_er_visit_ct_pmpm	0.0026
prov_spec_home_health_ind	0.0007
credit_num_studentloan_60dpd	0.0004
hlth_pgm_slvrsnkr_pct_par	-0.0008
cms_ra_factor_type_cdE	-0.0046
rx_gpi2_17_ind	-0.0079
betos_m1b_ind	-0.0095
betos_t1a_pmpm_ct	-0.0103
credit_num_1stmtgcredit	-0.0110

ccsp_220_ind	-0.0183
cmsd2_skn_radiation_ind	-0.0219
submcc_ben_othr_ind	-0.0286
cons_n65p_y	-0.0337
betos_t1a_ind	-0.0365
cons_retail_buyer	-0.0389
mabh_segH2	-0.0435
betos_m5d_ind	-0.0475
rx_mail_ind	-0.0483
cons_hhcompB	-0.0526
cons_homstatY	-0.0764
cms_ra_factor_type_cdCN	-0.1193
est_age	-0.1484

Appendix 3: Anxiety and Broken Calculation



```
case
when (`ccsp_228_ind`= 1 or `ccsp_239_ind` = 1) and (`bh_dema_ind` = 1 or `bh_aoth_ind` = 1) then 'Anxiety and Br. Bones'
when (`ccsp_228_ind`= 1 or `ccsp_239_ind` = 1) and (`bh_dema_ind` = 0 or `bh_aoth_ind` = 0) then 'Br. Bones Only'
when (`ccsp_228_ind`= 0 or `ccsp_239_ind` = 0) and (`bh_dema_ind` = 1 or `bh_aoth_ind` = 1) then 'Anxiety Only'
when (`ccsp_228_ind`= 0 or `ccsp_239_ind` = 0) and (`bh_dema_ind` = 0 or `bh_aoth_ind` = 0) then 'None'
end
```