**Model Building and Testing**

**Student: *Srivishnu Maganti***

**Date Submitted: December 6th, 2020**

**Task Selected:** ***Response time.***

Now that we have completed our exploration of the data through basic descriptive statistics, we focus on the final effort for the class: model building and testing. All semester, you have been working to better understand what factors may influence ambulance demand and/or response time, exploring a number of datasets and assessing these collections for potential biases. Through the descriptive statistics, you were able to identify potential patterns in the data that may affect your prediction model, or outliers you need to account for that could skew your ultimate results.

For this assignment, you will be building and testing your model, and then testing it against a set of your data. Your model, in short, must do one of two things:

1. Predict the **aggregate demand** for ambulance for each day from 2015-2016 according to your specific geographic unit of analysis (borough, community district, etc.);

OR

1. Predict the **incident response time** (in seconds) for each incident in 2015-2016.

In both of these cases, you will have relevant parameters to use in your final prediction model (i.e. the temperature, precipitation, community district info, call type, etc.) that are attached to the observations from 2015-2016. All you will be doing is predicting either demand or response time, and comparing that prediction with the actual value to identify the precision of your model.

This template is one of two items that needs to be submitted for your final project. The other one is your R code / additional data sets, annotated and ready for replication by the professor (see assignment note for that information). You will build on the efforts you have already completed for this class so far, including the bias checks and descriptive statistics assignments.

**Expectations:** You are expected to:

* Be through and clear in your descriptions, using complete sentences and checking for punctuation and grammar;
* Format table and charts for profession presentation (no R output from the console; you must either create a graphic using R or export/copy-paste the output to Excel and create a professional table with headings, formatting, and notation); and,
* Complete the entirety of the assignment. Any portions missing will be docked from the final grade of the assignment.

**Part 1: Dataset Pre-Processing**

The first step in this effort is to prepare your final data for analysis. In this step, you will process your data to remove variables you conclude would be theoretically irrelevant, join your calls for service data with any other dataset you plan on using, and create a final version with the appropriate unit of analysis ready for modeling.

1. To begin, please identify what variable you believe will be your response variable (y); the variable that you are trying to predict:

Here we are going to predict incident response time, using 9 different columns. All these columns which are used eighter directly or indirectly responsible for EMS response time.

The target variable here we are going to predict is “incident\_response\_seconds\_qy”.

1. Reviewing your Initial Project Plan and Descriptive Statistics, please list the final set of variables you intend on using for your analysis, along with their data type (please list all variables and add more rows to the table where needed):

**Incident Response Time**

|  |  |  |  |
| --- | --- | --- | --- |
| **Notation** | **Variable** | **Data Type** | **Source** |
| Y | incident\_response\_seconds\_qy | Numeric | Calls for Service |
| X1 | initial\_call\_type | Text | Calls for Service |
| X2 | initial\_severity\_level\_code | Numeric | Calls for service |
| X3 | Poilice Precint | Numeric | Calls for service |
| X4 | month | Numeric | Calls for service |
| X6 | Incident\_year | Numeric | Calls for service |
| X7 | PRCP | Numeric(float) | NYC weather data |
| X8 | SNOW | Numeric | NYC weather data |
| X9 | SNWD | Numeric | NYC weather data |
| X10 | TMIN | Numeric | NYC weather data |

1. Please remove the variables you will not be using from the Calls for Service final dataset, and provide the code for removal that you use here:

I have considered the merged data and from the merged data, I have selected few columns which are required for my analysis.

And I named the data frame used for my analysis as DataFrame.

DataFrame=select(merged\_data,initial\_call\_type, incident\_response\_seconds\_qy,policeprecinct, month, initial\_severity\_level\_code, incident\_year, PRCP, SNOW, SNWD,TMIN)

1. Please join or attach any additional variables from other datasets you identified earlier in your table, and provide your code here:

#create date variable for merging

a$date\_m <- as.Date(a$incident\_dt, "%m/%d/%Y")

#reformat date variable in weather data from character to date

b$date\_m <- as.Date(b$DATE, "%m/%d/%Y")

#merge data at the incident level

merged\_data <- left\_join(a,b,by="date\_m")

1. Please provide any additional code you need to modify or transform the variables into the final dataset for use in creating the initial model. This may include creating a month or day of week variable from the incident\_dt, an hour of day variable, transforming one of the precipitation variables into a binary “rain/no rain”, etc. It is expected that you will need to do some transformation of variables from at least one of your datasets to get them ready for analysis.

DataFrame$month = as.numeric(DataFrame$month)

**Part 2: Initial Correlations**

In this section, you will construct initial correlations to identify potential relationships between one of your predictors (x) and your response variable (y), as well as potential issues of collinearity between your predictors.

1. Please run a correlation analysis between all predictor and response variables, and place the output here:



1. What predictors are correlated highly (>0.6) with your response variable? Are they in the direction you expected?

According to results of correlation matrix, there is no such predicter which have correlation value >0.6, but relatively police precinct, severity level code and incident year are more correlated than that of any other predicters.

1. What predictors [x] are correlated highly with other predictors [x]?
2. Relatively, Month is highly correlated with SNOW and minimum temperature.
3. Precipitation is correlated with snow.
4. SNWD is also better correlated with SNOW.
5. Minimum temperature is correlated with month.
6. Minimum temperature is correlated with SNWD.

**Part 3: Initial Regression Model**

In this section, you are going to construct an initial ordinary least squares (OLS) regression model using your predictors (x) and response variable (y). Coding for running an OLS model can be found in the lab examples in Module 3, and please refer to the videos for Module 3 for more information on interpreting regression coefficients.

1. Based on the results from the correlations, and from the descriptive statistics in Assignment 3, please identify what predictors you would like to select as your initial regression model:

Incident year, initial call type, month, initial severity level code, police precinct, PRCP, SNOW, SNWD, TMIN.

1. Please provide the code that you will run to generate your initial regression model:

#regression

library(MASS)

library(ISLR)

lm.fit=lm(Dataset1$incident\_response\_seconds\_qy~

Dataset1$incident\_year +Dataset1$initial\_call\_type+Dataset1$month+Dataset1$initial\_severity\_level\_code+Dataset1$policeprecinct+ Dataset1$PRCP+Dataset1$SNOW+Dataset1$SNWD+Dataset1$TMIN,data=Dataset1)

summary(lm.fit)

AIC(lm.fit)

BIC(lm.fit)

1. Using the format provided below, please provide a table of results for your initial regression model. Please ensure to list all the variables in your dataset, and where there are no results (or results are missing in the model), please note by putting “—” in the respective column:



(Intercept) 3.598e+04

AIC (lm.fit): 63206594

BIC (lm.fit): 63206739

Adjusted R-squared: 0.061

1. Were any of the predictors significantly correlated with your response variable, while holding the other predictors at their mean values (i.e. were any predictor coefficients significant at a p<0.05)? If so, which ones?

After performing analysis and looking at the summary of regression and its coefficients, we can go to a conclusion that initial severity level relatively has the highest correlation. Then follows with precipitation by coefficient as 27.4. And also, incident year by -17.7 this means, response time decreases with every year.

1. Please describe each significant relationship between your predictors and response variable (i.e. if hour of day was significantly correlated with incident response time, and the unstandardized effect was 3.46, then we can state “For every one unit increase in hour of day, there was a corresponding 3.46 second increase in incident response time.”)
2. For incident year, every year response time is affected by -17.7, negative indicate that for every year incident response is decreased by 17.7 seconds
3. For each call type response time is getting decreased by 0.2 seconds.
4. For each month the incident response time is decreased by 4.5 seconds.
5. Depending upon severity level code response time is increased by 91.7 seconds. For each level of severity response time is varied by 91 seconds.
6. For each police precinct response time is changed by -0.2, it is actually getting decreased.
7. Depending upon precipitation, it is increasing by 27 seconds. For every unit of increase in precipitation response time is increased by 27 seconds.
8. For every unit of increase in snow, response time is increased by 19 seconds.
9. For every unit increase in SNWD, response time is increased by 6.1 seconds.
10. For drop in unit of temperature, response time is increased by 0.6 seconds.
11. Please select two separate significant relationships between one predictor (x) and your response variable (y). Construct a scatterplot chart for each relationship with the x-axis being the predictor and the y-axis being the response variable, and insert those two charts in here:

Chart, scatter chart

Description automatically generated

A picture containing text

Description automatically generated

1. What are the patterns that you see with those scatterplots? Just because relationship is significant within the model, does the scatterplot visualization also suggest the presence of a relationship between predictor and response?

I have considered police precinct and initial severity level, I have considered these two because they are most correlated than other variables, so let us consider the scatter plot of initial severity level, with increase in severity level response time Is increasing.

And, police precinct scatter plot explains the variation of response time with police precinct. There is a wide distribution of response time with police precinct.

**Part 4: Cross-Validation and Model Testing**

In this section, we will now focus on how well the predictors work across different samples of our training data. Predictors can fluctuate quite dramatically from one sample to another, and the purpose of this section is to examine the stability of predictors to identify those which are consistently significantly correlated with the dependent variable, and those whose correlation depends on the constitution of the training sample. As a reference, this section follows closely the lab examples provided in Module 6, so if there are any questions on the methods run, please also check those videos for more information on how to run these functions.

1. First, please provide the dimensions of the final dataset that you are using (# of observations by # of predictors; i.e. 10,000 calls by 19 predictors):

4241668 \*10

1. Following the Best Subset Selection process outlined in Module 6, please enter the code that you will be using to run regression models on the subset of all possible combinations of your predictors:

library(ISLR)

names(trainingdataset)

dim(trainingdataset)

sum(is.na(trainingdataset$incident\_response\_seconds\_qy))

trainingdataset=na.omit(trainingdataset)

dim(trainingdataset)

sum(is.na(trainingdataset))

library(leaps)

regfit.full=regsubsets(incident\_response\_seconds\_qy~.,trainingdataset)

summary(regfit.full)

regfit.full=regsubsets(incident\_response\_seconds\_qy~.,data=trainingdataset,nvmax=9)

reg.summary=summary(regfit.full)

names(reg.summary)

reg.summary$rsq

par(mfrow=c(2,2))

which.max(reg.summary$rsq)

plot(reg.summary$rsq,xlab="Number of Variables",ylab="RSQ",type="l")

points(9,reg.summary$rss[9], col="red",cex=2,pch=25)

which.min(reg.summary$rss)

plot(reg.summary$rss,xlab="Number of Variables",ylab="RSS",type="l")

points(9,reg.summary$rss[9], col="red",cex=2,pch=25)

plot(reg.summary$adjr2,xlab="Number of Variables",ylab="Adjusted RSq",type="l")

which.max(reg.summary$adjr2)

points(9,reg.summary$adjr2[9], col="red",cex=2,pch=25)

plot(reg.summary$cp,xlab="Number of Variables",ylab="Cp",type='l')

which.min(reg.summary$cp)

points(9,reg.summary$cp[9],col="red",cex=2,pch=20)

which.min(reg.summary$bic)

plot(reg.summary$bic,xlab="Number of Variables",ylab="BIC",type='l')

points(9,reg.summary$bic[9],col="red",cex=2,pch=20)

1. Please provide the following line plots using best subset selection, with “Number of Variables” along the x-axis and the specific parameter on the y-axis. These plots should also have a red-dot identifying the point with the maximum / minimum value for those parameters, depending on the parameter:
   1. *r2* (dot on maximum)

Chart, line chart

Description automatically generated

* 1. Residual Sum of Squares (RSS) - (dot on minimum)

Chart, line chart

Description automatically generated

* 1. Adjusted *r2* - (dot on maximum)

Chart, line chart

Description automatically generated

* 1. Mallows *Cp* – (dot on minimum)

Chart, line chart

Description automatically generated

* 1. Bayesian Information Criteria (BIC) – (dot on minimum)

Chart, line chart

Description automatically generated

1. Please compare and contrast the results from the different plots, interpreting the results in alignment with the meaning of each test statistic. What do these results suggest about the number of predictors that should be in each model? This should be at least 4 sentences long.

So basically here, the summary function returns some metrics – R-squared, Adjusted R2, Cp and BIC these matrices allow to interpret the best overall model, where best is defined as the model that maximize the adjusted R2 and minimize the prediction error. Prediction error is nothing but RSS, BIC and cp.

1. R-Squared: It is the coefficient of multiple determination for multiple regression. This evaluates the scatter of the data points around the fitted regression line. Higher R2 represents that data is being scatter around the dataset. So this chart has the maximum at nice (not dotted because it exceeded the limit of Y axis), and this explains that considering 9 predicters will provide us the best model for prediction.
2. RSS: If the value of RSS is more then we can tell that, our model is not working efficiently and out predicters are not good to fit, but if the value of RSS is less it indicates that the model is best to fit, so if you can see in the chart the model with 8,9 predicters are nearly zero, so this explains that 8 or predicters are good for fitting in this model.
3. Adjusted R-squared: The adjusted R2 will increase you for adding independent variables. Adjusted R-squared will increase if the added predicters improves the model performance. The adjusted R^2 will increase if the added predicters make sense to the performance. So here in this chart, after 8 it is being constant. So, here 8,9 are the best predicters.
4. Mallow cp: Well, in short, Mallow Cp statistic estimates the size of the bias that is introduced into the predicted responses by having an underspecified model. So mallow cp that is close to number of predicters indicates that model is precise, and mallow cp greater than number of predicters indicates that it is not precise. If you see in our chart 2,3 have highest mallow cp value, so this indicates that the model with 2 predicters are unprecise.
5. Bayesian Information Criteria (BIC): lower BIC indicates that model has lower penalty and the model will have better performance. Here lower BIC is at 9, so it is better to have 9 predicters.

So after these tests, based on this information we can tell that having 9 predicters are good for performing the regression.

1. We are going to split the data into two samples: 2008-2014, and 2015-2016. Please post the code you will use to split the dataset into two samples: a “train” sample (2008-2014) and a “test” sample (2015-2016).

#Filter data for training data i.e data from 2014 to 2015

trainingdataset<- subset(DataFrame, DataFrame$incident\_year >=2014 & DataFrame$incident\_year <=2015)

#Filter fata for test data i.e data from the year 2016

test<- subset(DataFrame, DataFrame$incident\_year == 2016)

1. Run *n* regression models (based on the number of your predictors) regressing your response variable [y] on your predictors [x], using the training dataset.
2. Create a matrix of model parameters (*model.matrix* function) using the test data, as well as a matrix to capture the prediction error between your training and test models.
3. Run a *for* loop to extract the coefficients from the trained models, the predicted values for your test data, and the Mean Squared Error (MSE) in prediction between the two.
4. Plot the Root Mean Squared Error (MSE) as a function of the number of variables, and use the *which.min* command to identify which model minimizes the prediction error most effectively.

Chart, line chart

Description automatically generated

1. In two sentences, what does this plot tell you about your models?

Basically, the smaller the mean square error, the closer you are finding the line of best fit. So, for our model

Above chart explains that when the number of variables are two then mean square error is less, so this

explains that with 2 variables we can have minimum mean squared error.

After performing this model tell that police precinct and severity level code are the two models that we are

Suppose to consider for prediction.

1. Based on the model which minimizes the prediction error, please provide a table of those model results, with their respective coefficients. Please note that, with the reg command, your results may only show those coefficients that were ultimately kept in the model (unlike your original regression model). Please ensure that you also provide the standard errors and p-values for the coefficients in the specific model.

(Intercept) policeprecinct initial\_severity\_level\_code

219.795865 -1.070933 101.755825

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Coefficient** | **Standard Error** | **p-value** |
| intitial\_severity\_level\_code  policeprecinct  *Intercept* | 91.196025  -0.907992  *233.543953* | 0.189539  0.009346 | 2e-16  2e-16 |
| *AIC* | *63224312* |  |  |
| *BIC* | *63224365* |  |  |
| *Adjusted R-squared:* | *0.05684* |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. Now we are going to conduct the cross-validation using a *k*-folds approach. Using data from the entire time period, please execute the following steps:
   1. Run the predict.regsubsets function provided in the sample code for the lab work in Module 6. This code will be used in the cross-validation.
   2. Run the same sets of best subset selection models that you used for the training vs. test data, but rather than splitting it across those two datasets, please use a 13-fold approach.
   3. Calculate the mean prediction error across the folds for each model, and provide a plot of mean prediction error by model.

Chart, line chart

Description automatically generated

* 1. In two sentences, what do these results say about your modeling approach?

This graph depicts that mean prediction error decreases with addition of each predicters, so this concludes that model with 9 predicters are the best fit.

* 1. Using the which.min function, identify which model minimizes mean prediction error, and provide a table of those model’s final results.



(Intercept) 3.598e+04

AIC (lm.fit): 63206594

BIC (lm.fit): 63206739

Adjusted R-squared: 0.061

**Part 5: Discussion and Conclusion**

This semester, you have spent many hours trying to identify those factors that are both highly correlated with and generally predictive of either incident response time or ambulance demand in NYC. You have reviewed datasets, identified potential sources of bias, and conducted initial descriptive analysis to build towards your predictive modeling. In this section, we want to answer some of these questions and interpret our overall modeling approach as well as areas where it may have fallen short.

1. **Comparative Analysis** – in the chart below, please provide the regression coefficients for your three key models (initial regression - Part 3-3; Training vs. Test modeling – Part 4-11; and k-folds – Part 4-12e). For the initial regression model, please provide all coefficients and bold those which were significantly correlated with the response variable [y]. For the other two models, please report the coefficients that were reported in the final models (which may or may not overlap with each other). Where there are no coefficients reported for a specific variable in a specific model, please put “—“. Also, please be sure to report out the AIC, BIC, and R2 for all three models reported here, and make sure your variables align across models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Model 1**  Initial Regression | **Model 2**  Training vs. Test | **Model 3**  *k*-folds |
| *Dataset1incident\_year*  *Dataset1$initial\_call\_type*  *Dataset1$month*  *Dataset1$initial\_severity\_level\_code*  *Dataset1$policeprecinct*  *Dataset1$PRCP*  *Dataset1$SNOW*  *Dataset1$SNWD*  *Dataset1$TMIN* | *-.77E+01*  *-2.65E-01*  *-.50E+00*  *9.17E+01*  *-9.15E-01*  *2.74E+01*  *1.90E+01*  *6.15E+00*  *6.63E-01* | *-*  *-*  *-*  *91.196025*  *-0.907992*  *-*  *-*  *-*  *-*  *-* | *-1.77E+01*  *-2.65E-01*  *-4.50E+00*  *9.17E+01*  *-9.15E-01*  *2.74E+01*  *1.90E+01*  *6.15E+00*  *6.63E-01* |
| *Intercept* | 3.598e+04 | *233.543953* | 3.598e+04 |
|  |  |  |  |
| *AIC* | 63206594 | *63224312* | 63206594 |
| *BIC* | 63206739 | *63224365* | 63206739 |
| *R2* | 0.061 | *0.05684* | 0.061 |

1. Based on these results, how well did your initial regression model align with the other two models regarding which predictors were significant in their predictive capability for your specific response variable? What differences were there? Why do you think those differences exist?

In Initial regression model 9 predicters were involved, which are physically thought as best predicters for prediction. After that best subset selection model was performed and it draws to conclusion that with using 2 predicters models works more efficient. So, two predicters were used. Furthermore, k fold cross validation has been performed and that suggests that model predicts well if the model has 9 variables. So, in k fold 9 variables were used and prediction is similar to initial regression, so here k fold suggests initial regression as the best model. Also, if we look at the intercepts in part 2 and other parts, the intercepts are almost similar. For initial severity level in model 2 it has intercept of 91.1 seconds, same goes with other models with little variation.

1. Comparing across models, are there variables whose relationship in models 2 and 3 seem strange or unexplainable from a theoretical perspective?

Comparing model 2 and model 3 we can tell that there is no significant change in the values, if we compare the severity level the value in model 2 is 91.1 and in model 3 is 91.7, and police precinct has

-0.9 in model 2 and -0.915 in model 3. So this difference is negligible in large datasets. Dealing with huge dataset might be a reason for that.

1. Looking at the predictive efficacy of your models, how effective were they? What proportion of the variance in your dependent variable were you ultimately able to predict (i.e. R2)?

As per the models that we have done, prediction is not effective. Because R-squared will reflect how good data points are scattered around the best fit line. So here for three models R-squared values are nearly same and there is no specific difference among them. The R squared value for model 2 is 0.05, which means that, the model is 5% effective and other two models are 6% effective. But, as the models have nearly same values we cannot depict the differences in three models.

1. Based on these results, imagine you were the Fire Chief for FDNY. Would you use one of your models to inform decision making about how fast someone will respond or how much demand there will be for ambulances on a specific day? Why or why not?

We can predict the response time of ambulance data based on all the factors that I considered. Based on k-fold model we can predict the response time, by initial severity level code, police precinct, snow, precipitation. If the severity level code increases, then the response time is less. We can also tell based on police precinct, amount of snow fall: response time increases with increases in snow fall, response time increases with increase in precipitation.

For a typical day, first I will look at the snowfall, precipitation and then look at the severity(initial) of situation, and police precinct.