Creating a Predictive Model for COVID Patients

Sara Hamdy, 1005295734

December 22, 2020

GitHub Repo Link: https://github.com/Saraahamdy/STA304FINAL.git

Abstract:

Currently Corona Virus - COVID - is a present issue that society is facing and creating more efficient methods to treat and prioritize patients is necessary. In this report, the process of creating a predictive model to analyze the probability of a COVID patient being alive is explored. To create the model a Multiple Linear Regression model is calculated using closed COVID cases in Toronto, where the data is gathered from Open Data Toronto (Open Data Dataset. (n.d.)). The model is then run through a backwards step-wise regression with Akaike information criterion, AIC, and with Bayesian information criterion, BIC, to optimize the model's performance and lower the AIC. The model run with the lowest AIC is used to predict the outcome of the current 5127 active COVID cases. In the original 37151 closed cases, 4% of cases resulted in deaths, and the model predicts that 1% of patients have less than a 50% chance of surviving. We conclude that the model is performing at an adequate level and can be used to predict the level of care a patient might need, and an increase of data and constant updates of closed COVID cases can further improve results.

Keywords:

COVID, Observational Study, Multiple Logistic Regression, Predictive Model, Akaike information criterion, Bayesian information criterion

Introduction:

COVID entered our world around December 2019, its numbers multiplied at an uncontrollable rate that resulted in many lives being lost. Having 76,013,074 cases worldwide, 52191 in Toronto and 1,681,249 deaths worldwide, 1770 in Toronto, the virus made itself present and demanded the attention of the world. COVID is very contagious and presents itself differently in patients, this made treating it very hard in the beginning. Not being aware of symptoms and not being able to identify whether a patient was a higher risk than others resulted in misclassification of a patient's potential needs. The virus has been present for almost a year and the world has been in a state of a pandemic for nine months, while deaths have significantly decreased when compared to earlier months, there are still cases of patients dying. Thus, being able to predict a patient's probability of surviving COVID can highlight where assets should be relocated.

Multiple Logistic Regression allows us to create a model the will give the probability of a dependent variable occurring, given the conditions of the independent variables that correlate with it. In this case, we are predicting the probability of a patient living after COVID. The model will be trained on the 37151 COVID cases in Toronto that have been closed, the outcomes being Resolved, the patient lived, or fatal. The model will depend on variables such as the patient's age, and gender. This model will depend on multiple variables and will be tuned based on its AIC. The AIC is a criterion used in statistics to determine the best model for the given data. Models with too many predictor variables tend to over-fit the data, the model should be parsimonious, simple, and informative. AIC allows us to control this by keeping variables with the lower AIC and maintaining a high maximum likelihood estimate, which tells us how well the model can reproduce the data. To analyze the AIC of the model backward step-wise regression will be applied to the original model,

with all variables, to locate different models. The model with the lowest AIC will be used to predict the outcomes of the 5129 pending cases.

In Methodology, the original data with the 37151 closed cases will be used to train the model that will predict the outcome of the 5129 pending cases. The process used to achieve the final model, backward stepwise regression with AIC and BIC, alongside the model itself is described. Then in the Results section, the log-odds and probabilities that the model computes for the remaining cases are displayed and explained. Lastly in the Discussion section, the results are summarized and the next steps for the model are provided.

Methodology:

Data:

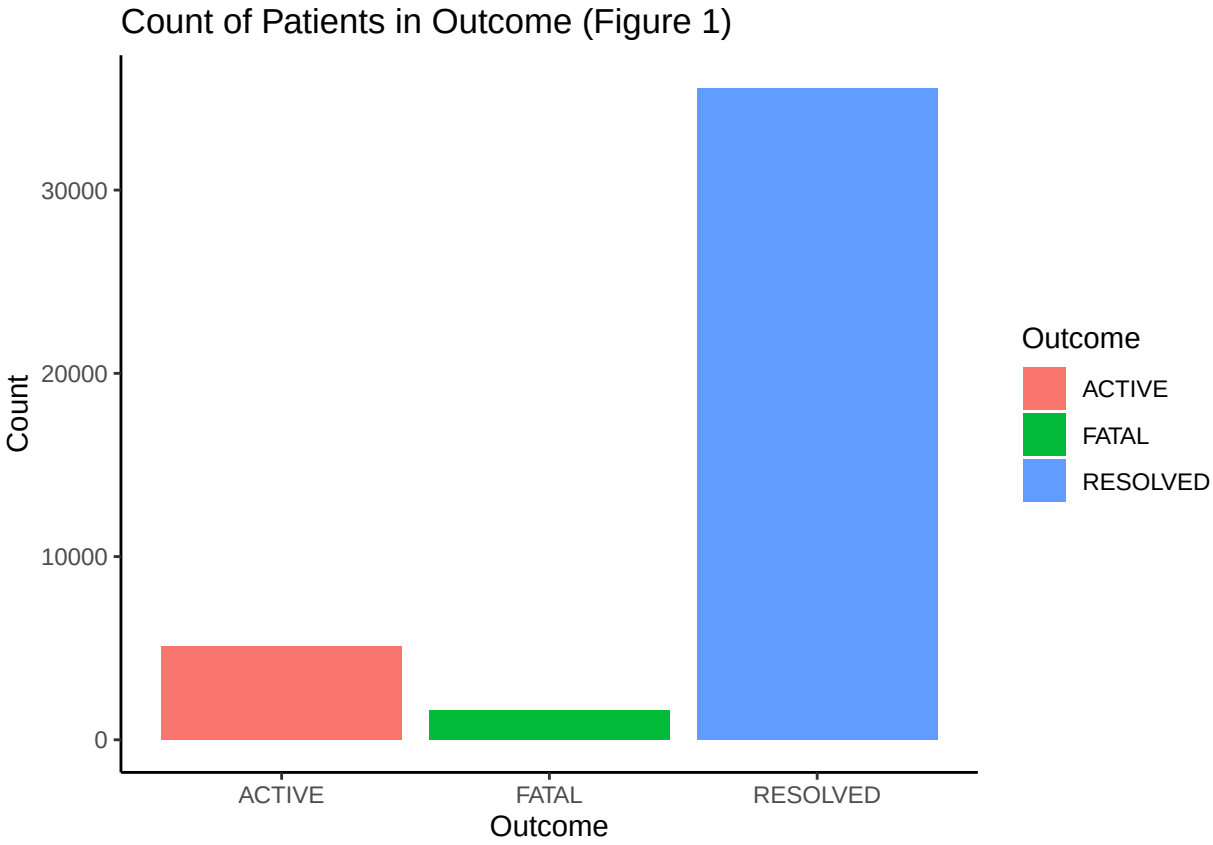
The data used for this analysis was gathered from the Toronto Open Data website on COVID cases in Toronto from January to December. the data is collected from a combination of sources, including the provincial communicable disease reporting system (iPHIS) and Toronto's custom COVID-19 case management system (CORES). The target population of the data are all COVID patients, the frame population are COVID patients in Toronto, and the sample is collected from patients that have reported their case. Very few cases in the data have missing observations, the few that are missing are removed. Overall this data set is very thorough accounting for many factors of a patient and most observations have no missing elements. While the data set is strong some variables recorded are not very useful overall such as Neighborhood Name that is removed before analysis. The data includes variables such as Assigned_ID, to track specific cases, Age.Group, that is later manipulated to highlight cases in older age groups, Source.of.Infection, whether it is from an Outbreak or hospital and other sources, Gender, Male or Female or Other, Outcome, Resolved or Fatal or Active, and more that can be seen in Table 1 below. Of the original 18 variables, 9 are used in the first model created to predict the variable Outcome. Those variables are Age.Group, Source.of.Infection, Classification, Episode.Data, Reported.Date, Client.Gender, Ever.in.ICU, Ever.Intubated, and Ever.Hospitalized. diff_dates is a variable that is created from Episode.Date and Reported.Data is the difference between the two dates.

The variables not included in the model development are X_id, Assigned_ID, Neighborhood.Name since it did not seem relevant to the Outcome of the patient. The FSA variable was another sorting variable that accounted for the postal code of the patient's home. Lastly, whether the patient was Currently.Hospitalized, Currently.in.ICU, and Currently.Intubated since the goal of the model is to account for the patient's Outcome the variables Ever.in.ICU, Ever.Intubated, and Ever.Hospitalized are included instead.

Table 1: First two observations in data

Variables	First Observation	Second Observation
X_id	484597	484598
Assigned_ID	1	2
Outbreak.Associated	Sporadic	Sporadic
Age.Group	50 to 59 Years	50 to 59 Years
Neighbourhood.Name	Willowdale East	Willowdale East
FSA	M2N	M2N
Source.of.Infection	Travel	Travel
Classification	CONFIRMED	CONFIRMED
Episode.Date	2020-01-22	2020-01-21
Reported.Date	2020-01-23	2020-01-23
Client.Gender	FEMALE	MALE
Outcome	RESOLVED	RESOLVED
Currently.Hospitalized	No	No
Currently.in.ICU	No	No
Currently.Intubated	No	No
Ever.Hospitalized	No	Yes
Ever.in.ICU	No	No

Variables	First Observation	Second Observation
Ever.Intubated	No	No



The variables that are selected are based on the potential correlation with the variable Outcome to predict the patient's outcome. Firstly, Age.Group categorizes each patient as less than 19, 20 to 29 years, 30 to 39 years, up to more than 90 years. With most deaths being present in patients older than 60 we regrouped the variable with patients 59 and younger in the same category, the remaining categories are left untouched, and the 34 observations with blank results are removed. Source.of.Infection indicates where the patient most likely came into contact with COVID such as from and Outbreak or travel. We grouped Unknown/Missing, Pending, and Travel sources of infection due to the lack of presence of the categories in the data gathered. Furthermore, we used the variable Classification which differentiated whether the case was CONFIRMED or PROBABLE. Also, diff_dates replaces Episode.Date and Reported.Date, it represents the number of days that passed from where the patient got infected and when they reported to the government. Higher days are expected to have less severe cases since the patients did not feel symptomatic or obligated to go to the hospital. We will also incorporate the patient's gender, which originally was categorized as FEMALE, MALE, OTHER, TRANSGENDER, and UNKOWN. Since TRANSGENDER, OTHER, and UNKNOWN reported very low numbers, 343 collectively, they were grouped as one category OTHER. Finally, we included whether a patient was Ever.Hospitalized, Ever.in.ICU, or Ever.Intubated. This can indicate the severity of the case. These variables are expected to influence the Outcome, the dependent variable in the model, of the patients. The count of patients in each category of Outcome can be seen above in Figure 1. It should be noted that Resolved heavily dominates the data, only 4% of closed cases were FATAL.

Model:

After selecting the variables that would initially be included in the model an initial Multiple Logistic Regression model was computed with all the variables present. The model can be seen below (Model 1). $\hat{p}$ in the model represents the chance of a patient being alive after COVID. Each β indicates the log odds and slope coefficient.

The intercept of the model, β_0, indicates the log odds of a patient that is 59 or under, was infected from close contact, classified as CONFIRMED, has a difference of dates of zero, is FEMALE, was never in the ICU, Intubated, or Hospitalized. Furthermore, β_1 is the slope of the predictor of the age predictor for patients who are 60 to 69 years old. β_2 is the slope of the predictor of the age predictor for patients who are 70 to 79 years old. β_3 is the slope of the predictor of the age predictor for patients who are 80 to 89 years old. β_4 is the slope of the predictor of the age predictor for patients who are 90 years or older. β_5, β_6, β_7, β_8, and β_9 represent the slope coefficients for the Source of Infection variable. Respectively, they represent the slope coefficient for Community, Healthcare, Institutional, N/A - Outbreak Associated, and Unkown/Missing, Travel. $beta_{10}$ is the slope coefficient if the patient was classified as PROBABLE. β_{11} is the slope coefficient for the variable diff_dates, for each one-unit change in diff dates the log odds are affected by -0.001684 units. β_12 and β_13 are for the variable Client.Gender, MALE for the former and OTHER for the latter. β_{14}, β_{15}, and β_{16} are for the last three variables Ever.in.ICU, Ever.Intubated, and Ever.Hospitalized. If a patient was ever in the ICU, Intubated, or Hospitalized the intercept changes with respect to their slope coefficient. Lastly, ϵ represents the error term in the function.

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = 7.512917 - 2.306783x_{age60-69} - 3.529714x_{age70-79} - 4.170225x_{age80-89} - 4.723182x_{age90+} \\ -0.246571x_{SOIcommunity} - 0.745137x_{SOIhealthcare} - 0.919889x_{SOIinstitutional} - 1.782164x_{SOIoutbreak} \\ +0.172207x_{SOIumt} - 0.767192x_{CLASSIFICATIONprobable} - 0.001684x_{dates} - 0.488358x_{GENDERmale} \\ -0.274357x_{GENDERother} - 1.568365x_{ICU} - 1.464162x_{Intubated} - 1.573555x_{Hospitalized} + \epsilon \end{aligned}$$

Model 1: All variables are included in this model.

After running backward step-wise regression with AIC and BIC. Running the regression with AIC and then BIC presented more variations of the model since BIC is similar to AIC but tends to set stricter guidelines for the variables resulting in fewer variables in the model. The model run with AIC, Model 2 seen below, had the lowest AIC of the three models. The model has similar coefficient estimates for the variables present in Model 1, and the variable diff_dates was removed from the model. $beta_0$ represents the same thing but no longer taking into account the diff_dates variable. β_1 to β_{10} represent the same thing as the original model. β_{11} and β_{12} now are the slope coefficients for Client.Gender, with 11 for MALE and 12 for OTHER. β_{13}, β_{14}, and β_{15} are for the last three variables Ever.in.ICU, Ever.Intubated, and Ever.Hospitalized. Lastly, ϵ represents the error term in the function. This model is used for the analysis of the remaining data. The remaining data from the original data set had 5129 cases that were still open, 2 had missing observations and were removed, and the remaining were used to test the model. The cases were plugged into the model, Model 2, and the probability each case has of being alive after COVID was computed.

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = 7.50265 - 2.30716x_{age60-69} - 3.52851x_{age70-79} - 4.16966x_{age80-89} - 4.72214x_{age90+} \\ -0.24687x_{SOIcommunity} - 0.74540x_{SOIhealthcare} - 0.91787x_{SOIinstitutional} - 1.77865x_{SOIoutbreak} \\ +0.17336x_{SOIumt} - 0.78075x_{CLASSIFICATIONprobable} - 0.48797x_{GENDERmale} \\ -0.27422x_{GENDERother} - 1.56860x_{ICU} - 1.46533x_{Intubated} - 1.57413x_{Hospitalized} + \epsilon \end{aligned}$$

Model 2: Backwards Step-wise Regression with AIC

Results

The first model calculated is Model 1 seen above, this model included all variables originally chosen with the intent that they could influence the outcome of the patient. The table below (Table 2) displays the model's Coefficient estimates and corresponding p-values.

Table 2: Summary output for Model 1 (All variables)

Coefficients in Model 1	Estimates in Model 1	P-values in Model 1
Intercept	7.512917	< 2e-16
Age.Group 60-69	-2.306783	< 2e-16
Age.Group 70-79	-3.529714	< 2e-16
Age.Group 80-89	-4.170225	< 2e-16
Age.Group 90+	-4.723182	< 2e-16
Source.of.Infection Community	-0.246571	0.19427
Source.of.Infection Healthcare	-0.745137	0.00017
Source.of.Infection Institutional	-0.919889	0.00199
Source.of.Infection N/A-Outbreak	-1.782164	< 2e-16
Source.of.Infection U/M/T	0.172207	0.32778
Classification probable	-0.767192	0.00392
diff_dates	-0.001684	0.67072
Client.Gender MALE	-0.488358	1.95e-13
Client.Gender OTHER	-0.274357	0.25706
Ever.Hospitalized Yes	-1.568365	1.64e-14
Ever.in.ICU Yes	-1.464162	2.68e-10
Ever.Intubated Yes	-1.573555	< 2e-16

The second model calculated, Model 2, seen above is the outcome of the backwards step-wise regression run with AIC on the original model. This method analyzes the model and removes variables that are insignificant with respect to the dependent variable and produces a model with the lower AIC while simultaneously considering the predictive capabilities of the model. Table 3 below displays the estimated coefficients and p-values.

Table 3: Summary output for Model 2, AIC (Final) Model

Coefficients in Model 2	Estimate in Model 2	P-values in Model 2
Intercept	7.50265	< 2e-16
Age.Group 60-69	-2.30716	< 2e-16
Age.Group 70-79	-3.52851	< 2e-16
Age.Group 80-89	-4.16966	< 2e-16
Age.Group 90+	-4.72214	< 2e-16
Source.of.Infection Community	-0.24687	0.193676
Source.of.Infection Healthcare	-0.74540	0.000169
Source.of.Infection Institutional	-0.91787	0.002023
Source.of.Infection N/A-Outbreak	-1.77865	< 2e-16
Source.of.Infection U/M/T	0.17336	0.324543
Classification probable	-0.78075	0.003044
Client.Gender MALE	-0.48797	2.02e-13
Client.Gender OTHER	-0.27422	0.257291
Ever.Hospitalized Yes	-1.56860	1.64e-14
Ever.in.ICU Yes	-1.46533	2.61e-10
Ever.Intubated Yes	-1.57413	< 2e-16

The last model calculated is Model 3, seen below. This model is the result of a backwards step-wise regression with BIC run on Model 1. Models run with BIC tend to have fewer variables since there are stricter restrictions placed on the variables. This method also aims to remove insignificant predictors as well as decreasing AIC. Table 4 depicts the coefficient estimates and p-values of Model 3.

$$log(\frac{\hat{p}}{1-\hat{p}}) = 7.42603 - 2.29069x_{age60-69} - 3.51038x_{age70-79} - 4.14952x_{age80-89}$$
$$-4.69683x_{age90+} - 0.20586x_{SOIcommunity} - 0.71082x_{SOIhealthcare} - 0.90226x_{SOIinstitutional}$$
$$-1.73090x_{SOIoutbreak} + 0.20971x_{SOIumt} - 0.48911x_{GENDERmale} - 0.26981x_{GENDERother}$$
$$-1.55611x_{ICU} - 1.46250x_{Intubated} - 1.56300x_{Hospitalized} + \epsilon$$

Model 3: Backwards Step-wise Regression with BIC

Table 4: Summary output for Model 3, BIC Model

Coefficients in Model 3	Estimate in Model 3	P-values in Model 3
Intercept	7.42603	< 2e-16
Age.Group 60-69	-2.29069	< 2e-16
Age.Group 70-79	-3.51038	< 2e-16
Age.Group 80-89	-4.14952	< 2e-16
Age.Group 90+	-4.69683	< 2e-16
Source.of.Infection Community	-0.20586	0.276671
Source.of.Infection Healthcare	-0.71082	0.000325
Source.of.Infection Institutional	-0.90226	0.002466
Source.of.Infection N/A-Outbreak	-1.73090	< 2e-16
Source.of.Infection U/M/T	0.20971	0.231651
Client.Gender MALE	-0.48911	1.72e-13
Client.Gender OTHER	-0.26981	0.265012
Ever.Hospitalized Yes	-1.55611	2.43e-14
Ever.in.ICU Yes	-1.46250	2.71e-10
Ever.Intubated Yes	-1.56300	< 2e-16

From the three models, their AICs were compared and Model 2 had the lowest AIC since lower AIC indicates better models, this is presented in Table 5 below. Model 2 is used to calculate each patient's log odds and the probability of survival. The model takes into account the patient's age group, source of infection, classification, gender, and whether they have ever been hospitalized, in the ICU, or Intubated.

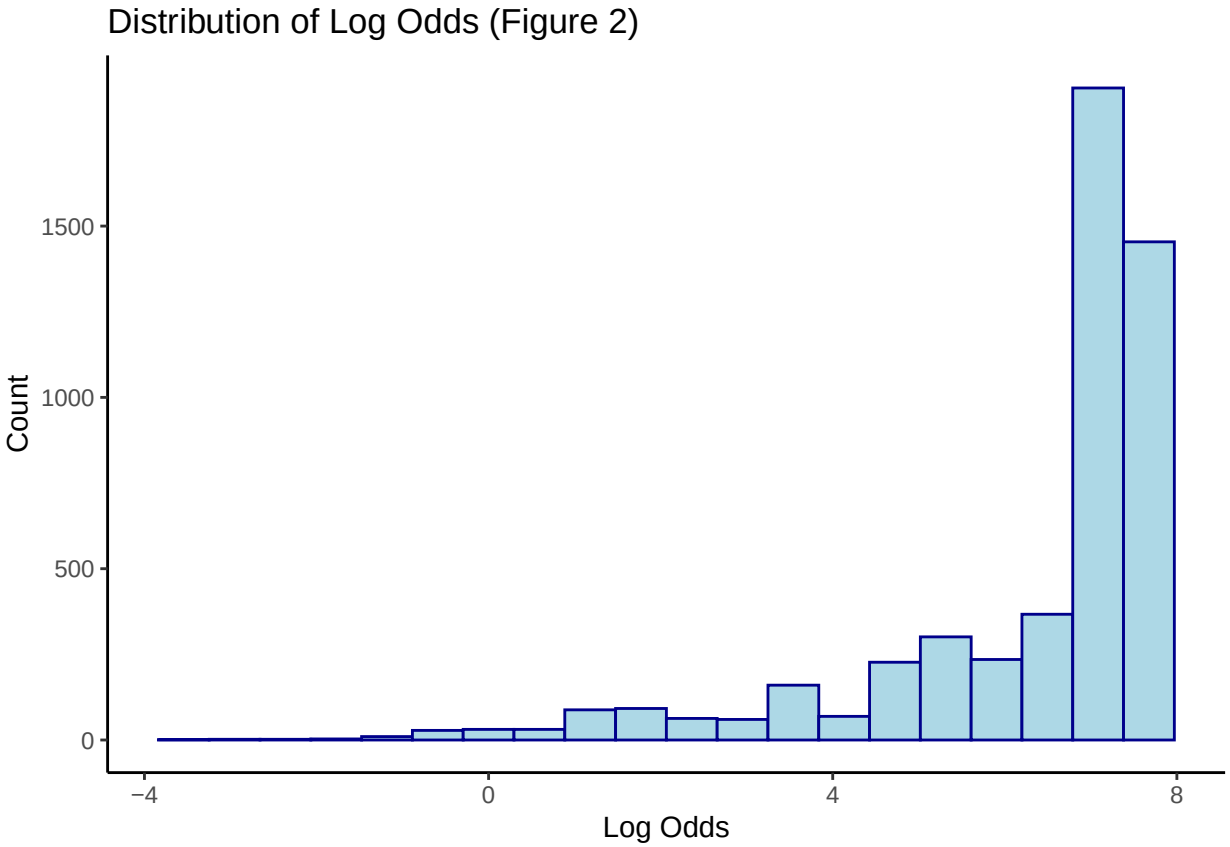
Table 5: Models' AIC

Models	AIC
Model 1	6857.5
Model 2	6855.6
Model 3	6861.4

The distribution of the log odds is presented in Figure 3 below. From the graph, it can be noted that most patients presented high log odds. This indicates that the model predicts most patients have higher chances of surviving.

```
##
## Call:
## glm(formula = as.factor(Outcome) ~ as.factor(Age.Group) + as.factor(Source.of.Infection) +
##     as.factor(Classification) + as.factor(Client.Gender) + as.factor(Ever.in.ICU) +
##     as.factor(Ever.Intubated) + as.factor(Ever.Hospitalized),
##     family = "binomial", data = model_data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.8738   0.0332   0.0424   0.1030   2.6722
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                               7.50265    0.18766
## as.factor(Age.Group)60 to 69 Years                       -2.30716    0.16634
## as.factor(Age.Group)70 to 79 Years                       -3.52851    0.15618
## as.factor(Age.Group)80 to 89 Years                       -4.16966    0.15081
## as.factor(Age.Group)90 and older                         -4.72214    0.15411
## as.factor(Source.of.Infection)Community                  -0.24687    0.18993
## as.factor(Source.of.Infection)Healthcare                 -0.74540    0.19817
## as.factor(Source.of.Infection)Institutional              -0.91787    0.29736
## as.factor(Source.of.Infection)N/A - Outbreak associated -1.77865    0.13899
## as.factor(Source.of.Infection)Unknown/Missing/Travel      0.17336    0.17597
## as.factor(Classification)PROBABLE                        -0.78075    0.26347
## as.factor(Client.Gender)MALE                             -0.48797    0.06641
## as.factor(Client.Gender)OTHER                            -0.27422    0.24207
## as.factor(Ever.in.ICU)Yes                                -1.56860    0.20435
## as.factor(Ever.Intubated)Yes                             -1.46533    0.23184
## as.factor(Ever.Hospitalized)Yes                          -1.57413    0.07557
##                                                          z value Pr(>|z|)
## (Intercept)                                               39.979  < 2e-16 ***
## as.factor(Age.Group)60 to 69 Years                       -13.870  < 2e-16 ***
## as.factor(Age.Group)70 to 79 Years                       -22.592  < 2e-16 ***
## as.factor(Age.Group)80 to 89 Years                       -27.648  < 2e-16 ***
## as.factor(Age.Group)90 and older                         -30.641  < 2e-16 ***
## as.factor(Source.of.Infection)Community                   -1.300 0.193676
## as.factor(Source.of.Infection)Healthcare                  -3.762 0.000169 ***
## as.factor(Source.of.Infection)Institutional               -3.087 0.002023 **
## as.factor(Source.of.Infection)N/A - Outbreak associated -12.797  < 2e-16 ***
## as.factor(Source.of.Infection)Unknown/Missing/Travel       0.985 0.324543
## as.factor(Classification)PROBABLE                         -2.963 0.003044 **
## as.factor(Client.Gender)MALE                              -7.347 2.02e-13 ***
## as.factor(Client.Gender)OTHER                             -1.133 0.257291
## as.factor(Ever.in.ICU)Yes                                 -7.676 1.64e-14 ***
## as.factor(Ever.Intubated)Yes                              -6.320 2.61e-10 ***
## as.factor(Ever.Hospitalized)Yes                          -20.831  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13212.6  on 37150  degrees of freedom
## Residual deviance:  6823.6  on 37135  degrees of freedom
## AIC: 6855.6
##
## Number of Fisher Scoring iterations: 9
```



The probability was computed from the log odds, the distribution of the number of patients in each percentile can be seen below in Table 6. From the table it should be noted that only 62 patients have less than a 50% chance of survival, that is 1.209% of the patients. In the original data set, 4% of patients died. The model predicts that around 200 patients, 4%, have an 80% chance of survival. While this model is not producing identical ratios of the original data it does not mean that it is performing poorly, and reasons for these results are explored in the Discussion below.

Table 6: Represents the number of patients in each quantile for chance of living

p<=0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9<p<=1.0
4 (0.078%)	6 (0.117%)	9 (0.176%)	25 (0.488%)	18 (0.351%)	20 (0.390%)	26 (0.507%)	82 (1.599%)	124 (2.419%)	4813 (93.876%)

Discussion

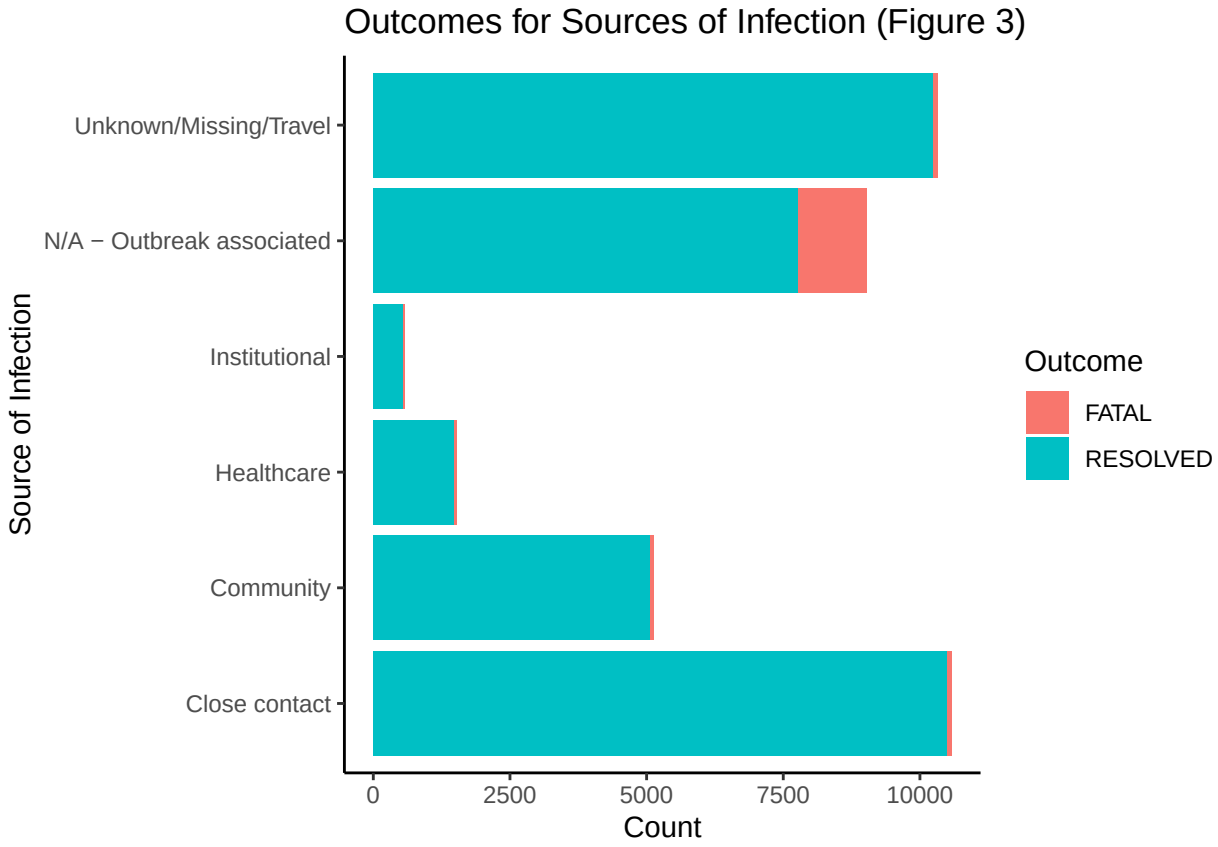
Summary:

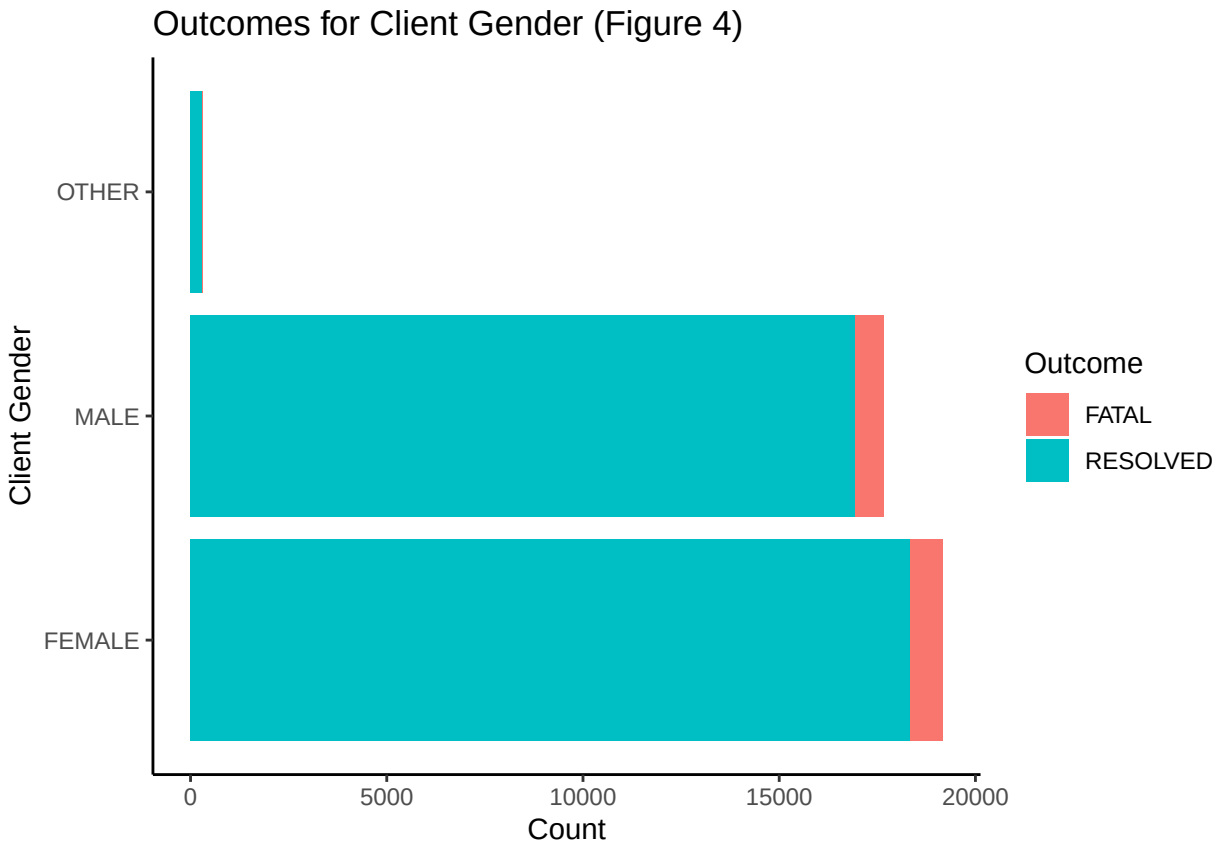
To conclude, a few steps were taken towards predicting the outcomes of the current ACTIVE cases in the data. The goal was to predict the outcomes of the ACTIVE COVID cases in the Toronto data. To achieve this a model was calculated using the closed cases in the data. The original variables were selected based on the assumption that they could correlate towards the patient's outcome. 12 of the 18 variables were selected, two of them, Episode.Data and Reported.Data were used to calculate the difference between them and were replaced by the new variable diff_dates. With the 11 variables, the first model was computed. Then a backwards step-wise regression with AIC then with BIC were run on the model, this method results in the removal of variables. From the three models, Model 2 had the lowest AIC and no longer took diff_dates into

account for the prediction of the outcome. It was used to calculate the log odds and probability of survival for the remaining patients. The model predicted that around 1.2% of the patients have less than a 50% chance of survival.

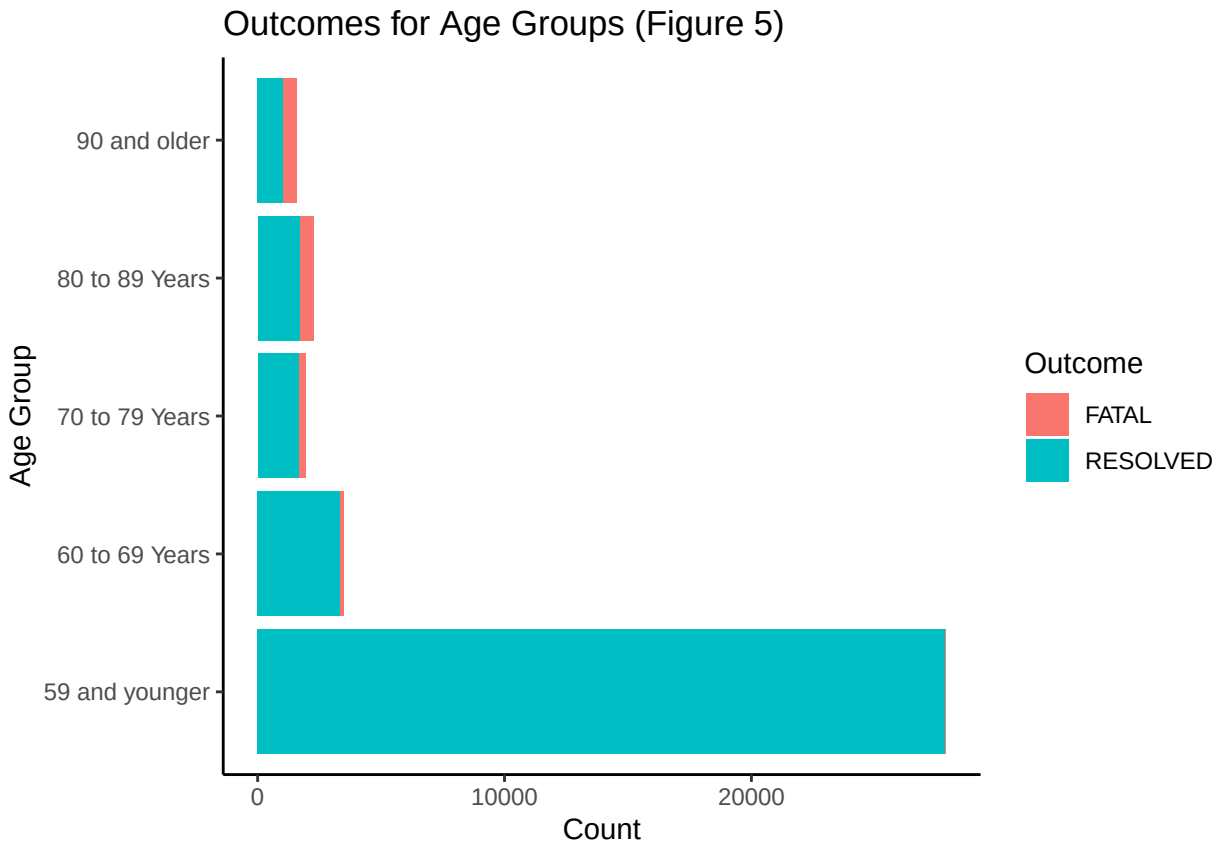
Conclusions:

After analyzing the results displayed above a few summaries can be made. The final model, Model 2, that is used to predict the Outcomes of the remaining patients included the majority of its coefficient estimates that were significant, had a p-value less than 0.05. This can be seen in Table 3 above. Three coefficients had p-values larger than 0.05. This means that we cannot reject the hypothesis that the estimate is zero. Since these coefficients are categories in variables, Source.of.Infection Healthcare and Unknown/Missing/Travel, and Client.Gender OTHER, it could be because of the lack of variability in the outcomes. As can be seen from Figure 3 below both Healthcare and Unknown/Missing/Travel have very few fatal cases which can prevent the model from making more conclusive results. Since only two sub-categories of the variable were insignificant it was decided to be kept in the model to provide insight into the patient's Outcome. Furthermore, Figure 4 displays the Outcomes grouped by Gender and it can also be noted that other has very few reported cases and even less variability present. This can also factor in the estimate being classified as insignificant. But it is also kept since it is only a subcategory of the variable and the removal of the gender OTHER ignores a proportion of people that must be acknowledged.





With the insignificant estimates kept in mind, the predictions provided by the model can be classified as useful. Only three β that are taken into account are insignificant the remaining coefficients can be classified as influencing the Outcome. If we classify that a patient with less than a 50% chance of survival has low odds of survival then only around 1% of the active cases can be classified as FATAL. This is not equal to the original 4% of fatalities found in the data but does not mean that the model is performing poorly. Many factors can influence this result including the conditions of these patients and procedures that are now in place now to treat compared to the original patients. The model can be used to analyze the conditions of patients and their chance of survival. With this model, fewer patients can be hospitalized and resources can be allocated towards more needy patients. For example Figure 5 below displays, the Outcomes grouped by age groups, and it should be noted older patients, over 80, have more fatalities when compared to other age groups. The model takes this into account by decreasing the log odds and simultaneously decreasing the chances of survival. Doctors can classify direct resources towards patients with lower chances of survival. With this model, more patients could be saved and better procedures can be placed in hospitals.



Weakness & Next Steps:

Some issues that can be found include the lack of variability in some predictors. As mentioned above some categories have limited Outcomes that can be preventing the model from performing at its highest potential. One solution for this is data being gathered from other locations outside of Toronto to train the model. Furthermore, the data does not take into account the patients' ethnicity/background, this variable may also correlate to the outcome of the patients. Lastly, the model does not reproduce identical results when compared to the original data, this can be because hospitals are currently more prepared to deal with COVID patients compared to initial cases. Also, older patients were more susceptible to aggressive cases of COVID resulting in more deaths within that category. Yet the new cases have less elderly compared to the original data, this can be seen in Table 7 and Table 9 below. Only 5.5% of the new cases are over 90 while 10.4% of the original data were over 80. Since the virus has been around for almost a year better procedures were put in place to treat patients, this can mean that newer patients may no longer share stronger similarities with previous patients. This can mean the model will predict lower survival for some patients. This suggests that the model is updated regularly with patient outcomes to maintain stable predictor capabilities.

Table 7: Number of patients grouped by Age in new patients

Age Group	Count
59 and younger	4132
60 to 69 Years	483
70 to 79 Years	231
80 to 89 Years	185
90 and older	96

Table 8: Number of patients grouped by Age in old patients

Age Group	Count
59 and younger	27887
60 to 69 Years	3482
70 to 79 Years	1935
80 to 89 Years	2273
90 and older	1574

In conclusion future steps that can be taken to further enhance the predictive capabilities of the model are as follows. Expand the region of data collection, to by province rather than by city. Then analyze whether race can contribute towards outcome prediction. Furthermore, update data used to train the model at least biweekly to keep the model up to current standards.

References

Anders. (2018, December 10). How to combine values within a variable. Retrieved December 22, 2020, from https://community.rstudio.com/t/how-to-combine-values-within-a-variable/19679/9

Bevans, R. (2020, March 27). Akaike Information Criterion: When & How to Use It. Retrieved December 22, 2020, from https://www.scribbr.com/statistics/akaike-information-criterion/

City of Toronto. (2020, December 16). COVID-19: Status of Cases in Toronto. Retrieved December 22, 2020, from https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/

Coronavirus Cases:. (n.d.). Retrieved December 22, 2020, from https://www.worldometers.info/coronaviruses/?utm_campaign=homeAdvegas1%3F

DQdlMDQdlM 8, Joshua UlrichJoshua Ulrich 159k2929 gold badges316316 silver badges394394 bronze badges, & RamnathRamnath 49.8k1313 gold badges113113 silver badges147147 bronze badges. (1960, May 01). Subset a data frame using OR when the column contains a factor. Retrieved December 22, 2020, from https://stackoverflow.com/questions/5680819/subset-a-data-frame-using-or-when-the-column-contains-a-factor

Ggplot2 histogram plot : Quick start guide - R software and data visualization. (n.d.). Retrieved December 22, 2020, from http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

How to Make a Histogram with ggplot2. (n.d.). Retrieved December 22, 2020, from https://www.datacamp.com/community/tutorials/make-histogram-ggplot2

Kumar, A. (2019, September 09). Learn R: How to Extract Rows and Columns From Data Frame - DZone Big Data. Retrieved December 22, 2020, from https://dzone.com/articles/learn-r-how-extract-rows

Learnerlearner 67533 gold badges88 silver badges1313 bronze badges, DWinDWin 6, NoLongerRandomnoLongerRandom 25122 silver badges33 bronze badges, & Dylanjmdylanjm 33422 silver badges1717 bronze badges. (1962, January 01). How to calculate goodness of fit in glm (R). Retrieved December 22, 2020, from https://stats.stackexchange.com/questions/46345/how-to-calculate-goodness-of-fit-in-glm-r

Make Beautiful Tables with the Formattable Package. (2020, December 07). Retrieved December 22, 2020, from https://www.displayr.com/formattable/

Open Data Dataset. (n.d.). Retrieved December 22, 2020, from https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

Philippe RemyPhilippe Remy 1, Konvaskonvas 12.7k22 gold badges3333 silver badges4343 bronze badges, & JasonJason 1. (1963, November 01). How to add elements to a list in R (loop). Retrieved December 22, 2020, from https://stackoverflow.com/questions/26508519/how-to-add-elements-to-a-list-in-r-loop

Plotting with ggplot: : Adding titles and axis names. (n.d.). Retrieved December 22, 2020, from http://environmentalcomputing.net/plotting-with-ggplot-adding-titles-and-axis-names/

R for Loop (With Examples). (2018, October 08). Retrieved December 22, 2020, from https://www.datamentor.io/r-programming/for-loop/

R if...else Statement (With Examples). (2018, October 08). Retrieved December 22, 2020, from https://www.datamentor.io/r-programming/if-else-statement/

Robk@statmethods.net, R. (n.d.). Creating new variables. Retrieved December 22, 2020, from https://www.statmethods.net/management/variables.html

Shaxi LiverShaxi Liver 2, & Akrunakrun 621k2222 gold badges352352 silver badges450450 bronze badges. (1964, September 01). How to remove rows from data frame based on subset function? Retrieved December 22, 2020, from https://stackoverflow.com/questions/32011244/how-to-remove-rows-from-data-frame-based-on-subset-function

Stephanie. (2020, December 14). Log Odds: Definition and Worked Statistics Problems. Retrieved December 22, 2020, from https://www.statisticshowto.com/log-odds/

Subtraction of dates. (n.d.). Retrieved December 22, 2020, from https://campus.datacamp.com/courses/intermediate-r-for-finance/dates?ex=8

Yihui Xie, C. (2020, November 23). R Markdown Cookbook. Retrieved December 22, 2020, from https://bookdown.org/yihui/rmarkdown-cookbook/kable.html