Prediction of result of Polls to establish the opinions of TORONTO residents on various topics covered by a City by-law*

Paper 2

HaoCheng Xu, Jing Li

Feb 27, 2022

Abstract

The Polls are usually conducted to represent the idea of a population about a specific problem or a suggestion. This analysis is based on such Polls which are conducted by the City government in Toronto. If it can be predicted the idea of residents, businesses and property owners who are in the affected area about particular changes that going to be made in Toronto City, then the City government can be applied these changes to the city. Other than using all most all the variables collected during polls, this predictive model will be constructed based on a few interesting and important variables. Therefore, it will not be needed to collect all most all the variables that are collected during polls seasons to predict the final result of polls.

Introduction

The City conducts polls to determine the opinions of property owners, residents and businesses that could be affected by a change in their neighbourhood regarding some applications. Because, the changes made by the city government can be affected the residents, businesses or property owners in a positive way or a negative way. Therefore, it is more successful to conduct polls before making a decision about the applications which would be affected peoples' daily lives.

If it can be predicted the idea of residents, businesses and property owners who are in the affected area about particular changes that going to be made in Toronto City, then the City government can be applied these changes to the city. Therefore, the objective of this analysis is to predict the final results of such polls based on a predictive model. Since the analysis is based on the polls data which are collected from 2015 to 2021, it will be able to construct a more accurate predictive model for the prediction of final results of specific kinds of polls. The final result will be either residents, businesses and property owners would like or not for the new changes which are suggested by the Toronto city government. Other than using all most all the variables collected during polls, this predictive model will be constructed based on a few interesting and important variables. Therefore, it will not be needed to collect all most all the variables that are collected during polls seasons to predict the final result of polls.

Predictive Modelling: Predictive modelling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Carew, J. M., & Burns, E. (2020, December 2).

The objective of this analysis is to predict the final results of polls that are conducted by the city government of Toronto for make specific application in the city based on specific factors.

*Code and data are available at: https://github.com/Jasonfallen/Prediction-of-result-of-Polls-to-establish-the-opinions-of-TORONTO-residents-on-various-topics-cover.git

Data

Data Collection Process

The City conducts polls to determine the opinions of property owners, residents and businesses that could be affected by a change in their neighbourhood regarding Boulevard cafe, Off-street parking (front yard parking and commercial boulevard parking), Permit parking, Traffic calming and Business Improvement Area. When an application for a proposal is submitted, the City conducts a poll of people in the affected area. In order for a poll to be considered positive, it must meet benchmarks as determined by specific by-law or city policy. If the result of the poll is positive, the application may proceed through the approval process. The data were collected from 2015 to the current time regarding various types of applications from the residents, businesses and property owners. Sometimes, the data are collected through ePolls. As examples, FYP 2016-208: 1 Astoria Ave (Type: Front Yard Parking), FYP 2021-029: 1 Hurndale Ave (Type: Front Yard Parking) are conducted as ePolls and data are collected through these ePolls in 2016. As well as some Polls are conducted through postage-paid return envelopes and these data are collected in a manual way. As well as there is another type of Polls which are conducted for Business Improvement Areas (BIA). In there Registry Services mails a BIA Poll Notice to each business property owner in the proposed BIA area. A business property owner must give a copy to each commercial and industrial tenant of the property within 30 days of receipt. Likewise, the data collection methods are varied according to the type of the Poll.

The dataset which is used for this analysis was taken from the **Toronto Open Data portal**. The dataset and more information about the dataset can be accessed through the following URL. Since the dataset is refreshed day by day, so it is worth using the updated dataset if doing further analysis.
https://open.toronto.ca/dataset/polls-conducted-by-the-city/

Data Summary

As introduced previously, the dataset is taken from the Toronto Open Data Portal and the dataset is refreshed day by day. The dataset which is used for this predictive analysis is corresponding to the Polls which were conducted from 2015-May-01 to 2021-October-16. Table 1 is illustrated the variables of the dataset and the description of each attribute in the dataset.

```
# setting variables
Variable <- c("X_id","ADDRESS","APPLICATION_FOR","BALLOTS_BLANK","BALLOTS_CAST",
              "BALLOTS_DISTRIBUTED","BALLOTS_IN_FAVOUR","BALLOTS_NEEDED_TO_PROCEED",
              "BALLOTS_NEEDED_TO_PROCEED_LBL","BALLOTS_OPPOSED","BALLOTS_RECEIVED_BY_VOTERS","BALLOTS_R
              "POLL_RESULT","POTENTIAL_VOTERS","RESPONSE_RATE_MET")
Description <- c("Unique row identifier for Open Data database","Street address of the application","Ty
The number of ballots returned has met the required response rate")
des <- data.frame(Variable, Description)
```

As Table 1 shows, there are 25 variables in the Polls dataset. As well as the data corresponding to 1013 Polls. Before carrying out further analysis it is required to clean the dataset. Because there may have missing values, outliers, unnecessary variables in the dataset. Before checking for the missing values the variables such as x_id, addresses,ballots_need_to_be_proceed, clode_date,moratorium_date , open_date,pass_rate_label,poll_cd and poll_id variable can be removed from the dataset. It can be removed x_id, poll_cd and poll_id from the dataset because these variables represent just unique identifiers for polls and ballots. As well as in this analysis we are not aware about the date that Poll is carried out. Therefore, close_date, moratorium_date and open_date are removed. The *rm* function can be used to remove these unnecessary variables from the model. Then *na.omit* function is applied to remove all of the missing values from the dataset. The winzoring techniques was applied to treat to the outliers. It can not be removed outliers because then the size of the dataset will be decreased. After that it can be checking for the data type of each variable in the data set. Since there are altogether two variables namely apply_for and respondent_rate_met are in character format, these variables are converted to factor variables by *as.factor* function. It's not needed to create dummy variables here because R already created dummy variables in modelling. As mentioned in the Introduction, to release the final results of a particular Poll respondent rate should be passed a specific

Table 1: Description of ths Dataset

Variable	Description
X_id	Unique row identifier for Open Data database
ADDRESS	Street address of the application
APPLICATION_FOR	Type of application
BALLOTS_BLANK	Number of ballots received with no mark to identify in favour or opposed
BALLOTS_CAST	Number of ballots returned
BALLOTS_DISTRIBUTED	Number of ballots distributed
BALLOTS_IN_FAVOUR	Number of ballots received and marked favour
BALLOTS_NEEDED_TO_PROCEED	The number of ballots needed to proceed based on percentage of return
BALLOTS_NEEDED_TO_PROCEED_LBL	Percentage of returned ballots needed to consider poll valid
BALLOTS_OPPOSED	Number of ballots received and marked opposed
BALLOTS_RECEIVED_BY_VOTERS	Number of ballots returned to City Clerk's Office
BALLOTS_RETURNED_TO_SENDER	Number of ballot returned by Canada Post as not delivered
BALLOTS_SPOILED	Number of ballots received and were not clearly marked as either in favour or opposed
CLOSE_DATE	Date the poll has been closed
DECLARATIONS_ADDED	Number of individuals added to the poll list after the poll has open
FINAL_VOTER_COUNT	Number of total voters on the final poll list
MORATORIUM_DATE	The date to which this poll can be conducted again
OPEN_DATE	Date the poll is open to public
PASS_RATE	Number of returned ballots needed for a positive poll result
PASS_RATE_LABEL	Percentage of returned ballots needed to determine poll result
POLL_CD	Poll Identification Number (Public)
POLL_ID	Poll Application Number
POLL_RESULT	Final result of poll
POTENTIAL_VOTERS	Number of people residing within poll boundary range
RESPONSE_RATE_MET	The number of ballots returned has met the required response rate

benchmark value. Therefore, for this analysis, it is selected polls data only if the required Respondent rate is met(RESPONSE_RATE_MET ="Yes"). To select those polls data, *filter* function can be applied. This dataset can be used for further analysis because the dataset is completely clean now. Then it's needed to split the whole dataset into training and testing set with a 3:1 ratio. The whole descriptive analysis and model constructions were performed based on the training dataset and the test set was sided to evaluate the model performances.

Table 2: Summary Statistics of variables in Dataset

APPLICATION_FOR	BALLOTS_BLANK	BALLOTS_CAST	BALLOTS_DISTRIBUTED	BALLOTS_IN_FAVOUR	BALLOTS_NEEDED_TO_PROCEED	BALLOTS_OPPOSED	BALLOTS_RECEIVED_BY_VOTERS
Length:616	Min. :0.00000	Min. : 2.00	Min. : 2.00	Min. : 0.00	Min. : 1.0	Min. : 0.000	Min. : 2.00
Class :character	1st Qu.:0.00000	1st Qu.: 25.00	1st Qu.: 53.00	1st Qu.: 16.00	1st Qu.: 13.0	1st Qu.: 1.000	1st Qu.: 50.00
Mode :character	Median :0.00000	Median : 38.00	Median : 78.00	Median : 27.00	Median : 20.0	Median : 4.000	Median : 76.00
NA	Mean :0.04545	Mean : 42.19	Mean : 88.97	Mean : 30.18	Mean : 25.2	Mean : 8.563	Mean : 84.98
NA	3rd Qu.:0.00000	3rd Qu.: 51.25	3rd Qu.:108.25	3rd Qu.: 39.00	3rd Qu.: 28.0	3rd Qu.:12.000	3rd Qu.:104.00
NA	Max. :3.00000	Max. :297.00	Max. :645.00	Max. :197.00	Max. :264.0	Max. :84.000	Max. :598.00

BALLOTS_RETURNED_TO_SENDER	BALLOTS_SPOILED	DECLARATIONS_ADDED	FINAL_VOTER_COUNT	PASS_RATE	POLL_RESULT	POTENTIAL_VOTERS	RESPONSE_RATE_MET
Min. : 0.000	Min. : 0.000	Min. : 0.0000	Min. : 2.00	Min. : 2.00	Min. :0.0000	Min. : 2.0	Length:821
1st Qu.: 0.000	1st Qu.: 1.000	1st Qu.: 0.0000	1st Qu.: 55.00	1st Qu.: 14.00	1st Qu.:1.0000	1st Qu.: 68.0	Class :character
Median : 2.000	Median : 3.000	Median : 0.0000	Median : 80.00	Median : 20.00	Median :1.0000	Median : 99.0	Mode :character
Mean : 4.206	Mean : 3.514	Mean : 0.8611	Mean : 91.03	Mean : 23.59	Mean :0.8124	Mean : 160.7	NA
3rd Qu.: 6.000	3rd Qu.: 5.000	3rd Qu.: 0.0000	3rd Qu.:111.00	3rd Qu.: 28.00	3rd Qu.:1.0000	3rd Qu.: 144.0	NA
Max. :56.000	Max. :31.000	Max. :78.0000	Max. :645.00	Max. :186.00	Max. :1.0000	Max. :13458.0	NA

After removing the unnecessary variables in the model remaining variables are used to carry out further analysis. These remaining variables are ballots_blank,ballots_cast,ballots_distributed,ballots_in favour, ballots_needed_to_proceed,ballots_opposed,ballots_received_by_voters, ballets_return_to_senders, ballot_spoiled,decleration_added, final_voter_count, pass_rate,potential_voters and application_for variables. But it may have highly correlated variables among these variables. These variables can be identified in later analysis.

Table 2 represented the summary statistics for the dataset. There are altogether 16 variables in the final dataset which is used for analysis and three variables of them are categorical variables. These are application_for, respondent_rate_met and the response variable of this analysis which is poll_result. All other variables are in this dataset are numerical variables. Here it is not needed to create dummy variables for categorical variables because R automatically created dummy variables for modelling. It is very important to visualize how the variables are related to each other and as well as how each variable is distributed in the sample. Therefore, in this part, it is discussed the most important descriptive analysis results correspond to this analysis.

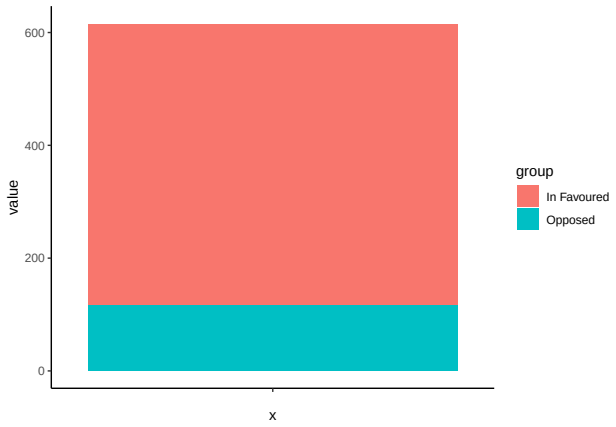

Figure 1: The Distribution of Polls Results

Figure 1 shows the distribution of Polls results in the collected sample. It is clear that more than 50% of the residents/ businesses or property owners in the affected areas are voted in favour of the new applications which are suggested by the city government of Toronto. By seeing this variation, it can be decided that most of the new applications which are suggested by the city government of Toronto on the city are better for residents as well as businesses and property owners in affected areas. When carrying out the modelling for predicting the final results, this distribution might be influenced by the final model performances. Because there is a class imbalanced problem in the dataset.

According to Figure 2, it can be seen that the positive rate of the polls across the different kinds of applications are varying in different levels. The positive rate is highest for the applications that are making for Traffic

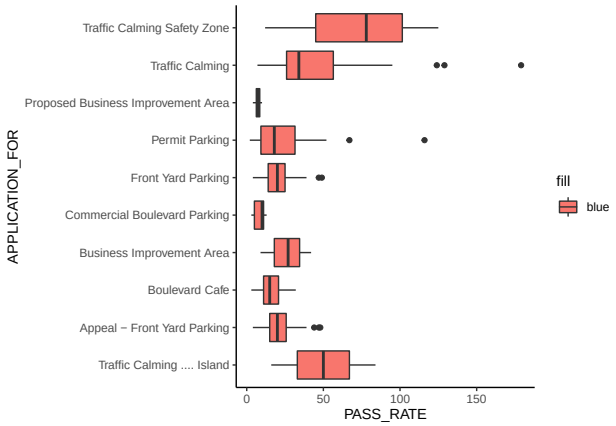

Figure 2: Suggested Application with Number of Positive Results

Calming. As well according to this plot, the positive rate of the polls is very fewer for the application that is made for Commercial Boulevard Parking, Business Improvement Area and Boulevard Cafe. This implies that most of the residents and property/ businesses owners do not like making new applications regarding commercial applications. But the positive rates of Polls are higher for the application those are regarding Traffic Calming.

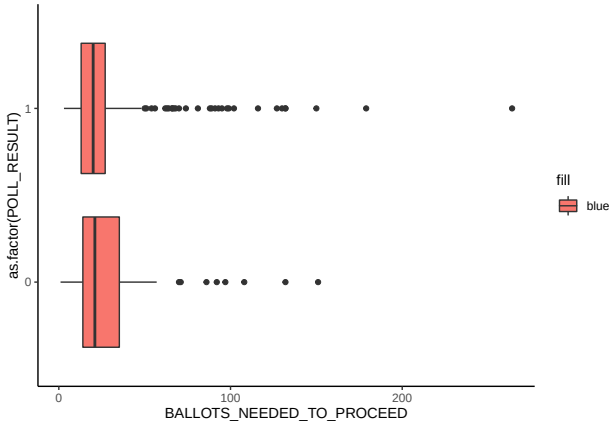

Figure 3: Poll Results with Ballots needed to be Proceed

According to figure 3, the number of ballots needed to proceed based on a percentage of return is higher for those who are voted as "Opposed" in the polls. By this variation, it can get an idea that we can not say the residents and businesses/property owners affected are most likely to be voted as "Opposed" in the Polls. That is, it might be most of the suggestions made by the city government of Toronto are not better for the residents and property/ businesses owners in the affected area.

Figure 4 shows the correlation matrix of numerical variables. The correlation matrix has illustrated the strengths of linear relationships between numerical variables. The dark blue coloured cell shows a relatively higher association between corresponding variables. Therefore, it is multicollinearity will be present in the model which is going to be fitted for predicting poll results if all of these variables are considered. Therefore highly associated variables will be removed before carrying out the predictive modelling.

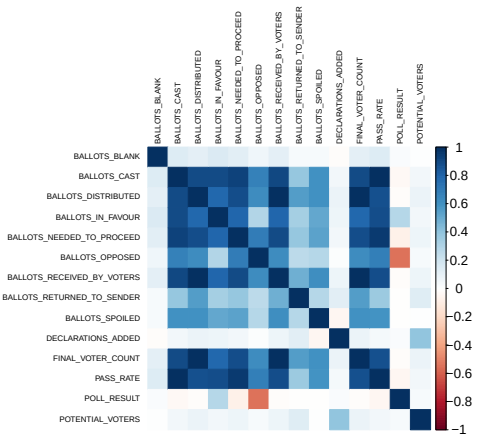

Figure 4: Correlation Plot for Numerical Variables in the dataset

Methods

The main objective of this analysis is to predict the final results of polls that are conducted by the city government of Toronto for making a specific application in the city based on specific factors. Since the response variable of this analysis is a categorical variable, Logistic regression is applied to the dataset to fulfill the objective. Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. In this scenario, the outcome is In Favour and Opposing. Logistic Regression is actually the sigmoid transformation of a linear regression model. It will be used the probability of getting a particular outcome as the response variable in Logistic Regression.

The outcome in logistic regression analysis is often coded as 0 or 1, where 1 indicates that the outcome of interest is present, and 0 indicates that the outcome of interest is absent. If we define p as the probability that the outcome is 1, the multiple logistic regression model can be written as

$$\ln(\frac{\hat{p}}{(1-\hat{p})}) = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$$

$\hat{p}$ is the expected probability that the outcome is present; X_1 through X_p are distinct independent variables; and b_0 through bp are the regression coefficients. The multiple logistic regression model is sometimes written differently. In the following form, the outcome is the expected log of the odds that the outcome is present,

$$\hat{p} = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}{1 + (\exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}$$

p = 1,2,. . . ,16

Notice that the right-hand side of the equation above looks like the multiple linear regression equation. However, the technique for estimating the regression coefficients in a logistic regression model is different from that used to estimate the regression coefficients in a multiple linear regression model. In logistic regression the coefficients derived from the model (e.g., b_1) indicate the change in the expected log-odds relative to a one-unit change in X_1, holding all other predictors constant. There are some assumptions of the Logistic Regression. These are independence of errors, linearity in the logit for continuous variables, absence of multicollinearity and lack of strongly influential outliers.

In this analysis, the AIC method is applied to select the best model for predicting the Final Result of the Poll. The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and

determine which one is the best fit for the data. AIC is calculated from the number of independent variables used to build the model and the maximum likelihood estimate of the model (how well the model reproduces the data). The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables. In statistics, AIC is most often used for model selection. By calculating and comparing the AIC scores of several possible models, we can choose the one that is the best fit for the data. Nath, R. (2020, February 26). Logistic Regression- the Theory and Code - Rajwrita Nath. Medium. https://medium.com/@rajwrita/logistic-regression-the-the-e8ed646e6a29

It's very difficult to select a model with the best variables when there are more variables in the dataset. We want to know which of the independent variables we have measured explain the variation in our dependent variable. A good way to find out is to create a set of models, each containing a different combination of the independent variables we have measured. Once we've created several possible models, it can be used AIC to compare them. Lower AIC scores are better, and AIC penalizes models that use more parameters. So if two models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model. AIC determines the relative information value of the model using the maximum likelihood estimate and the number of parameters (independent variables) in the model. The formula for AIC is:

$$AIC = 2K - 2\ln(L)$$

K is the number of independent variables used and L is the log-likelihood estimate. To compare models using AIC, we need to calculate the AIC of each model. If a model is more than 2 AIC units lower than another, then it is considered significantly better than that model.

Results

As mentioned in the methodology section the variables for the Logistic Regression model are selected based on the AIC method. The lowest AIC score is obtained when the variables application_for, ballots_cast, ballots_opposed, ballots_recieved_by_voters, pass_rate, ballots_needed_to_proceed and ballots_in_favour are in the model and observed the minimum AIC score corresponding to this model as 78.95. Except for the application_for variable, all other variables are numerical in type and the application_for variables is categorical variable with 10 levels.

After fitting the logistic regression with selected variables the following estimated coefficients were observed

Parameter	Coefficient
$\hat{\beta}_0$	6.16294
$\hat{\beta}_1$	-2.56811
$\hat{\beta}_2$	5.67348
$\hat{\beta}_3$	-18.14404
$\hat{\beta}_4$	-2.42293
$\hat{\beta}_5$	-2.55335
$\hat{\beta}_6$	9.12779
$\hat{\beta}_7$	-9.03117
$\hat{\beta}_8$	-29.58771
$\hat{\beta}_9$	2.94098
$\hat{\beta}_10$	1.15200
$\hat{\beta}_11$	0.70619
$\hat{\beta}_12$	0.77785
$\hat{\beta}_13$	-3.3330
$\hat{\beta}_14$	-0.18018

Table 4: Deviance Analysis

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	227	316.0049353	NA
BALLOTS_CAST	1	0.4108011	226	315.5941342	0.5215632
BALLOTS_IN_FAVOUR	1	277.8200272	225	37.7741070	0.0000000
BALLOTS_NEEDED_TO_PROCEED	1	15.0434632	224	22.7306437	0.0001051
BALLOTS_RECEIVED_BY_VOTERS	1	22.7306436	223	0.0000002	0.0000019

Parameter	Coefficient
$\hat{\beta}_15$	-2.99484

Variables in the order corresponding to the above table are, Intercept, Boulevard Cafe, Improvement Area, Boulevard Parking, Yard Parking, Permit Parking, Proposed Business Improvement Area, Traffic Calming, Traffic Calming â€“ Island, Traffic Calming Safety Zone, ballots_cast, ballots_in_favour, ballots_needed_to_proceed, ballots_opposed, ballots_recieved_by_voters, pass_rate.

After fitting the model the accuracy of the classification or the model performance can be evaluated on the testing dataset. The predicting accuracy of Final Polls results by the fitted logistic model is 92.19%.

Conclusions

The objective of this analysis was to predict the final results of polls that are conducted by the city government of Toronto for making specific applications in the city based on specific factors. Since the Logistic Regression was fitted on the dataset after cleaning the dataset. Under the cleaning process, it was removed all the missing values from the dataset and treated to the outliers with the wizarding technique. For this analysis, it was selected only the Polls which have met the benchmark of response rate. Before doing model selection based on the AIC method, the correlation analysis was done. According to the correlation plot, it was observed that there are numerical variables that are highly correlated with each other. Since these variables definitely cause multicollinearity the highly correlated variables were removed from the model before carrying out AIC for model selection. The initially removed perfectly related covariates are Ballots_return_to_sender, ballots_spoiled and Potential_voters variables.

Then was applied logistic regression and the best model was selected under the AIC criterion. From the anova table, it can be seen that the best model is the combination model that includes Intercept, application_for,ballots_cast, ballots_in_favour, ballots_needed_to_proceed, ballots_opposed, ballots_recieved_by_voters and pass_rate. According to the coefficient output table, the p-value corresponding to traffic calming is too small indicating that for the probability of the final poll result is to be In Favour, the application of Traffic Calming is highly affected. That is most residents and property owners say Yes to the application of Trafic Calming. For the unit increment of ballots_cast, the log(odds) is increased by 1.15. But for unit increments of ballots_opposed, ballots_recieved_by_voters and pass_rate the log(odds) or the probability of saying Yes for polls are decreases.

The anova output table shows the deviance results. The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Analyzing the table we can see the drop in deviance when adding each variable one at a time. Again, adding application type, ballots_in_favour, ballots_opposed and pass_rate significantly reduces the residual deviance. A large p-value indicates that the model without the variable explains more or less the same amount of variation.

Finally, the model performances are evaluated based on the testing dataset and it was observed that the accuracy of correctly predicting the final results is 92.19%. Which is relatively higher accuracy. Based on these results finally, it can be concluded that to make specific applications like Traffic Clamings, the Toronto

City government does not need to conduct polls and it may be wasting of money. Therefore city government can do such applications without conducting polls. As well as to predict the final results of a poll, the city government will not be needed to collect all most all data as currently done. They needed only information about specific variables that discusses previously.

Weaknesses

The size of the dataset is relatively small. If can be applied this predictive model on a relatively large dataset, then it will be able to increase the accuracy of prediction furthermore. As well as there were some missinf values in the datase and one can be apllied imputaion method on missing values. But in this analysis, it was removed those missing values because it's wanted to work with original data.

Next Steps

As a further analysis, it can be suggested to construct a predictive model with interaction effects between significant variables and this method will be improved the final accuracy many more.

Discussion

Throughout this report, it was discussed how to perform data preprocessing and predictive model building. At end of this analysis, it was observed that only application_for,ballots_cast, ballots_in_favour, ballots_needed_to_proceed, ballots_opposed, ballots_recieved_by_voters and pass_rate predicted the final results of Polls. Therefore, based on only these variables, one can predict the final result of a poll earlier without waiting for the last moment of the poll.

Bibliography

1. Boston University School of Public Health. (2013, January 17). *Multiple Logistic Regression Analysis.* Https://Sphweb.Bumc.Bu.Edu/. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_multivariable/bs704_multivariable8.html

2. Carew, J. M., & Burns, E. (2020, December 2). *Predictive modeling.* SearchEnterpriseAI. https://searchenterpriseai.techtarget.com/definition/predictive-modeling

3. Nath, R. (2020, February 26). *Logistic Regression- the Theory and Code* - Rajwrita Nath. Medium. https://medium.com/@rajwrita/logistic-regression-the-the-e8ed646e6a29

4. R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

5. Mervisiano, M. (2021, January 3). *How to do Logistic Regression in R* - Towards Data Science. Medium. https://towardsdatascience.com/how-to-do-logistic-regression-in-r-456e9cfec7cd

6. Vidhya, A. (2020, June 26). *Logistic Regression R | Introduction to Logistic Regression.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/