STA304 Final Project: To Evaluate Effects of Bike Features and Parking Locations on Bike Thefts in Toronto

Ziyi Qu

12/21/2020

Code and data supporting this analysis is available at: https://github.com/Ziyi-Qu/STA304-Final-Project

Abstract

As cycling, a sustainable and economical means of transportation, has become increasingly popular in Toronto in recent years, bike theft has become a worsening social issue there. It is in the interest of Toronto cyclists to understand what factors are associated with bike thefts. Mainly, I explore the effect of independent factors such as the cost of bike, color of bike, type of bike, whether the bike is parked in downtown Toronto and type of building near which the bike is parked. The reasons why I am interested in those factors are based on the common expectations that properties with a high cost, in dark colour and of a popular type tends to be stolen more frequently, and that densely-populated places such as downtown area are more likely to suffer from property losses. After getting an observational dataset called "Bicycle Thefts" from Toronto Open Data Portal website, I do some data cleaning to get binary variable as the outcome of interest: whether the bike is stolen, and predictors including both numerical and categorical variables: the cost of bike (numerical), whether the bike is in dark colour (binary), whether the bike is of a popular type (binary), whether the bike is parked in downtown Toronto (binary) and the type of building where the bike is parked (categorical with three subgroups: "residential"; "social"; "educational"). I use R to run a logistic regression on whether the bike is stolen with five predictors above. From the summary of the model, all five predictors show a positive correlation with bike theft risk: one can also find that, holding all other predictors the same, an expensive bike has a higher theft risk; a bike in dark colour has a higher theft risk; a bike of a popular type has a higher theft risk; a bike parked in downtown has a higher theft risk; a bike parked near a residential or social building has a higher theft risk. Hence, corresponding precautionary actions can be considered by cyclists in Toronto, including but not limited to using more reliable bicycle locks, personalizing bikes with distinctive colours or components, parking bicycles in a more visible area in residential neighbourhoods or in crowded areas such as downtown. I also compare my model with an alternative logistic model with all other components the same except for excluding the factor on the type of building where bike is parked, where I use model diagnostics techniques including AIC (Akaike information criterion), BIC (Bayesian information criterion) and AUC (Area Under the ROC Curve). Two of the three model evaluation tools, AIC and AUC, show a preference for the original model, and thus I conclude my original model is better. It is in the interest of of students, like me, who study and live in downtown Toronto, to know if there is a causality between downtown and bike thefts or not. Thus I use propensity score matching on my observational data to infer causality. The propensity score matching performs as a quasi-experimental technique, where I select whether the bike is parked in downtown Toronto as the treatment and whether the bike is stolen as the outcome of interest. I calculate the propensity score based on the treatment using logistic regression in R, and the create a new matched data based on matching pairs from treatment group and untreatment group with similar propensity score. I also evaluate the quality of propensity score matching by t-test on the means of the numerical variable: cost of bike, under treated and untreated groups, where the result shows that the quality of propensity score matching is good. Finally, I run a new logistic regression on the outcome of interest on five predictors including the treatment based on the new matched dataset. One can find that the treatment is not significant in the new model, and thus I conclude that whether the bike is parked in downtown Toronto

does not have a causal relationship with whether the bike will be stolen, and thus people study and live there do not have to worry too much on their bikes just because they are parked in downtown Toronto.

Key words

Bike Thefts, Property Security Issue, Greater Toronto Area, Observational Data, Logistic Regression, Model Diagnostics, Causal Inference, Propensity Score Matching

Introduction

Unlike driving where people will find it easy to get involved in a traffic jam with byproduct in the form of car exhausts, cycling provides with a more efficient, more economical, more sustainable and healthier means of transportation. According to the paper on cycling, there has been a significant increase in the number of cyclists in both small and large communities across Canada, with Toronto has witnessed a percentage increase in cycling of 141% from 1996 to 2016. However, with the growing use of bikes, bike thefts have become a common property security issue there for a long time (Beth, 2019). Last year, there were around twenty thousand bikes reported stolen in Toronto, of which only one percent were recovered (Canadian Cycling Magazine, 2020). Especially, during the most recent COVID-19 pandemic this year, biking as a social distant means of exercise and transit further arouses renewed interest and triggers a bike boom in Canada, where bike thefts has gone up around 30% in 2020 (Lily, 2020).

It is of the public interest to understand what possible factors, such as parking locations (type of area and type of building where the bike is parked) and bicycle features (color, cost and type), are responsible for a higher likelihood of bike theft in Toronto, and Toronto cyclists could therefore take corresponding precautionary actions towards potential theft risks. In analyzing bike thefts in Toronto, an observational data is not only more feasible since it will be financially costly and time-consuming to conduct an experiment with control groups for property thefts, but also more reliable since it provides with a natural setting without any artificial intervention. To conduct an analysis on the likelihood of bike theft, a proper analysis tool to use is logistic regression model, which was first introduced by Joseph Berkson in 1944 (Tinbergen Institute, 2002). Compared to linear regression used on a continuous and numerical dependent variable, the logistic regression builds a model on a binary dependent variable (in the form of log-odds) as a linear combination of predictors. Besides using logistic model to infer any correlation from the observational data, using propensity score matching will be helpful to infer possible causality, a quasi-experimental method that uses a treatment group on the observational dataset (Peter, 2011).

To consider what factors are potentially associated with bike thefts, here could be some shared ideas and assumptions on property loss: expensive properties tend to face higher theft risks for a high resold price; properties with dark colours are easier to be taken away without attracting much attention; popular properties with a high market demand easily become the target of theft; crowded places are more likely to suffer from thefts. To understand whether the stolen bikes in Toronto share the similar patterns, relevant variables can be chosen from the observational dataset. Specifically, five independent variables are selected for analysis: one numerical predictor (cost of bike) and four categorical predictors (whether the bike has a dark colour; whether the bike is of a popular type; whether the bike is parked in downtown Toronto; the type of building near which the bike is parked) to run a logistic regression on the likelihood of theft, to discover any significant correlations among them. For students, like me, who study in and live in downtown Toronto, it will be of their interests to further investigate whether a bike parked in downtown has a causal relationship with bike theft, through propensity score matching.

In this report, an observational dataset, updated annually from 2014 to 2019, called "Bicycle Thefts" from the Toronto Open Data Portal (Toronto Police Services, 2020), will be used to infer correlation and causality between five independent variables (cost of bike; bike color; bike type; whether bike is parked in downtown Toronto; type of building where bike is parked) and the outcome of interest (whether the bike is stolen). After the Abstract and Introduction sections, information on the dataset and steps for the constructing the model, involving logistic regression for inferring correlation and propensity score matching for inferring causality,

will be discussed in the Methodology section. Then, the results from the constructed model will be shown in a proper and understandable way in the Results section. Next, any summary, significance, implication and conclusion derived from the results will be elaborately discussed in the Discussion section, along with weakness and next steps proposed. The final part will be the References section where sources of data used in this report and developed by others are provided.

Methodology

Data

I download the observational dataset called "Bicycle Thefts", updated annually by the Toronto Police Services, from the Toronto Open Data Portal website. This data contains bike thefts occurrences in the Greater Toronto Area from 2014 to 2019. There are in total 26 columns (characteristics of bikes reported lost) and 21584 rows (observations), where the 26 columns include both numerical variables (such as cost of bike, speed of bike, year/month/day/time of occurrence, identifier, longitude, latitude and some other less distinguishable characteristics) and categorical variables (such as offence of occurrence, bike features including make/model/type/colour, current status, neighborhood, location type and some other less distinguishable characteristics).

Here are the columns in the raw dataset:

```
##  [1] "_id"              "Index_"           "event_unique_id"  "Primary_Offence"
##  [5] "Occurrence_Date"  "Occurrence_Year"  "Occurrence_Month" "Occurrence_Day"
##  [9] "Occurrence_Time"  "Division"         "City"             "Location_Type"
## [13] "Premise_Type"     "Bike_Make"        "Bike_Model"       "Bike_Type"
## [17] "Bike_Speed"       "Bike_Colour"      "Cost_of_Bike"     "Status"
## [21] "Hood_ID"          "Neighbourhood"    "Lat"              "Long"
## [25] "ObjectId"         "geometry"
```

There are both advantages and disadvantages of this dataset. Advantages can include that the dataset provides with both numeric and categorical features of observations, that there are abundant observations (with a number 21584) which can reduce the result bias, and that the data is provided by an authority. There are also, however, disadvantages. Disadvantages of this dataset is that the time frame is between 2014 and 2019 and there is no updated information on 2020 (which is expected to be updated with at the beginning of 2021), that the dataset only includes information on bikes that is reported lost due to various reasons (not limited to theft) and does not cover information on all bikes in Toronto, that there are many N/A information that will reduce the number of useful observations, and that there are sometimes too many sub-groups under one category (for example, 140 subgroups under "Neighbourhood") which make the real analysis challenging and daunting. All weaknesses will be discussed in detailed in the Weakness&Next Steps part under the Discussion section later.

The population of my analysis is all the bikes that are actually theft in the Greater Toronto Area, the frame is the potential police official list including all stolen bikes that are reported in the Greater Toronto Area, and the sample is the observations from actual dataset uploaded by the Toronto Police Services on the Toronto Open Data Portal, which I use to conduct the analysis.

For analysis purpose, I choose one column from the original dataset as the dependent binary variable: if the stolen bike is attributed to "theft" offence or not ("Primary_Offence"). Subsequently, I select five columns as five predictors in the logistic model. Specifically, I choose one numeric variable: the cost of bike ("Cost_of_Bike") and four categorical variables: the colour of bike ("Bike_Colour"), the type of bike ("Bike_Type"), the distance from where the bike is parked to the city center ("Neighbourhood") and the type of building near which the bike is parked ("Location_Type").

After deciding on the necessary components in the logistic model, I do data cleaning. I first filter the original dataset by the six variables mentioned above. I use "clean_name()" function to transform any upper case letters to lower case letters, and rename four variables whose original names are not literally reflective of my intention, and make them more readable and understandable, with keeping names of all other

variables unchanged. For example, I rename "primary_offence" to "if_stolen", which will be used to indicate whether the bike is actually stolen instead of being robbed or other offences; I rename "bike_colour" to "if_dark_colour", which will be used to indicate whether the bike is in dark colour; I rename "neighbourhood" to "if_near_city_center", which will be used to indicate whether the bike is parked near the Toronto city center; I rename "bike_type" to "if_popular_type", which will be used to indicate whether the bike is of a popular type.

Second, I filter the reduced data again with condition where cost of bike is greater or equal to 1, since there are some observations reporting a cost of 0 and taking these into considerations does not make scientific or analytical sense.

Next, I transform "if_stolen", 'if_near_city_center","if_dark_colour" and "if_popular_type" into binary predictors by "mutate_at()" function: for "if_stolen", among all sub-groups related to lost bikes, I assign sub-groups related to theft and break&enter offence with a value of 1, with all other a value of 0, since some sub-groups, such as robbery and careless driving, are intrinsically different from theft (the outcome of interest); for "if_near_city_center", among all 140 sub-groups, I select sub-groups with high frequencies, and assign each of them with a value of 1 if this location is within 10 kilometers of the Toronto city center (using benchmark of Toronto City Hall as the very city center of Toronto), while assigning a value of 0 if this location is 10 kilometers far from the Toronto city center; for "if_dark_colour", among all the 216 sub-groups, I select ones with highest frequencies, and assign each of them with a value of 1 if the colour is associated with dark colors such as black, brown and gray, while assigning a value of 0 if the colour is associated with bright colors such as white, red and yellow; for "if_popular_type", I assign each sub-group that belongs to a mountain bike, touring bike, road bike, trail bike or BMX bike, all the most popular bike types based on research (Jetsetta, n.d.), with a value of 1, and assign any other type with a value of 0.

I also use "mutate_at()" to divide 43 sub-groups under the categorical variable indicating the type of building where the bike is parked ("location_type") into three main sub-categories: "residential" for places where people live; "social" for both commercial (such as restaurants and corporate places) and non-commercial (such as parks, TTC stations and religious facilities) places; "educational" for places providing with educational services, including schools, colleges and universities, and I want an "educational" type since it is of interests of students like me to know whether educational places tend to suffer higher bike thefts. Finally, I omit all the observations with N/As by "na.omit()".

Here is a baseline characteristics summary of my reduced data:

```
##
##                                  Overall
##   n                                4305
##   if_stolen (mean (SD))            0.52 (0.50)
##   if_dark_colour (mean (SD))       0.55 (0.50)
##   cost_of_bike (mean (SD))       977.74 (1733.57)
##   if_near_city_center (mean (SD))  0.89 (0.31)
##   location_type (%)
##      educational                    173 ( 4.0)
##      residential                   2587 (60.1)
##      social                        1545 (35.9)
##   if_popular_type (mean (SD))      0.79 (0.41)
##   bike_type (mean (SD))            0.79 (0.41)
```

The dependent variable is: if_stolen(binary): =1 if the bike is stolen; = 0 if the bike is not stolen

The independent variables are: cost_of_bike(integer): indicating the cost of bike if_dark_colour(binary): =1 if the bike is in a dark colour; = 0 if the bike is in a bright colour if_popular_type(binary): = 1 if the bike is of a popular type; = 0 if the bike is not of a popular type if_near_city_center(binary): = 1 if the bike is parked near the Toronto city center; = 0 if the bike is parked far from the Toronto city center location_type(categorical): = "social" if the place is either commercial or non-commercial; = "residential" if the place is where people reside and live; = "educational" if the place is where educational services are

offered.

Model

To analyze what factors will contribute to a higher or lower probability of bike thefts, the logistic model stands out among other options. Compared to other models such as multiple linear regression model that yields prediction on a numeric variable, the logistic regression model yields prediction on a binary variable and this fits my goal, which is to test if the bike is stolen or not. Basically, the logistic regression method builds a model on a binary dependent variable (in the form of log-odds) as a linear combination of predictors. As I mention before, I will use the whether the bike is stolen as the binary dependent variable, and use bike cost, whether bike has a dark colour, whether bike is of a popular type, whether the bike is parked near city center, and the type of building where bike is parked as predictors.

First, I believe that there is such a true relationship in the population:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 cost\_of\_bike + \beta_2 if\_dark\_colour + \beta_3 if\_popular\_type + \beta_4 if\_near\_city\_center + \beta_5 location\_typeresidential + \beta_6 location\_typesocial + \epsilon$

where β_0 is the intercept parameter, $\beta_1 \sim \beta_6$ are the slope parameters, and ϵ is the error term.

Then I build a logistic model, using R, to estimate the true model by "glm()" function, and the model built based on the reduced dataset looks like:

$log(\frac{p}{1-p}) = \hat{\beta}_0 + \hat{\beta}_1 cost\_of\_bike + \hat{\beta}_2 if\_dark\_colour + \hat{\beta}_3 if\_popular\_type + \hat{\beta}_4 if\_near\_city\_center + \hat{\beta}_5 location\_typeresidential + \hat{\beta}_6 location\_typesocial$

where p stands for the probability of theft for a certain bike, $log(\frac{p}{1-p})$ stands for the log odds of bike theft, $\hat{\beta}_0$ is the intercept estimate of β_0, with $\hat{\beta}_1 \sim \hat{\beta}_6$ the slope estimates of $\beta_1 \sim \beta_6$.

Here are reasons for choosing each variable. The reason why I choose the cost of bike as one predictor is that expensive properties tend to face higher theft risks since they can be resold at a good price and thus there is a financial motivation for thieves to steal expensive properties. It is therefore assumed that expensive bikes have a higher risk of being stolen than less expensive ones. The reason why I choose the colour of bike as one predictor is that, usually properties with dark colors (in some cases redeemed as common colours for certain properties) are easier to be taken away without drawing too much attention by either the property owner or others in public. It is thus assumed that dark-coloured bikes will be stolen more often than bright-colored ones. The reason why I choose the type of bike as one predictor is that properties of a popular type tend to become the target for theft since they are in high demand and thus can be resold at a good price. It is thus assumed that bikes of popular types are will be stolen more frequently than less popular ones. The reason why I choose whether the bike is parked in downtown as one predictor is that crowded places are more likely to suffer property loss since it may be hard to visually track items there, and city center is usually most densely populated place in a city. It is thus assumed that Toronto city center tends to have more frequent bike thefts. The reason why I choose types of location where the bike is parked as one predictor is that different places, such as social, educational and residential, may have different theft patterns due to different levels of surveillance, and it is also of students' interest to know does educational places (including schools, colleges and universities) have more bike thefts than any other place in Toronto.

To see if this model is good, I can make a comparison between the model and its alternative. There is actually an alternative logistic model, where I can omit the predictor indicating the type of building where bike loss is reported, since it is my curiosity to discover if educational places tend to have higher or lower bike theft risks, and it is possible that this variable is not significant in my model. To construct the alternative model, I use the same dependent variable: whether the bike is stolen and four predictors: cost of bike, whether the bike is in dark colour, whether the bike is of a popular type, and whether the bike is parked in downtown Toronto.

I can check which logistic model is better by model diagnostics tools, mainly AIC (Akaike information criterion), BIC (Bayesian information criterion) and AUC (Area Under the ROC Curve). AIC and BIC are penalized-likelihood criteria and perform as a means for model selection, estimating the quality of one model relative to each other. Preferably, a better model (more close to the true model) will have a lower AIC and

a BIC score. The AUC serves as a global measure of diagnostic accuracy, which estimates how high the discriminate power of a test is. Ideally, a higher AUC is better and an AUC at least above 0.5 will prove the model's accuracy. For logistic model, there is no model convergence.

By conducting AIC and BIC on the original model and on the alternative model, one can find that the original model has a lower AIC (5920.108<5928.837), but a higher BIC (5964.681>5960.674) than alternative model. I need to further conduct AUC to decide which one is better. The AUC result shows that the original model has a higher value of 0.559 than the alternative with a value of 0.551. Based on three model diagnostics techniques, I can conclude that my original model that additionally includes the variable on type of location where bike is parked is generally better than the alternative model, based on AIC and AUC, though only BIC score shows a slight preference for the alternative.

Cross validation shall also be used and this technique assesses the prediction capability of logistic regression on the relevant dataset as expected, and helps to identify any potential limitations with their impacts. Specifically, cross-validation would be conducted by dividing the original dataset into "training" and "test" subsets. The algorithm is, taking my situation as an example, to randomly divide the reduced dataset into 100 subsets using R, then randomly use 99 of subsets as "training" dataset while the remaining 1 become the "test" dataset. Each time, by fitting a new model based on the different combinations of 99 "training" subsets and making prediction on the 1 "test" subset, I could check the mean prediction error (MPE) to assess the prediction accuracy of our model. Calibration plot is useful for checking the accuracy of regression estimates by visually observing whether the prediction line align with the observed line.

After constructing the model and conducting model diagnostics, I also use propensity score matching to infer any causal relationship between the binary predictor, whether the bike is parked in downtown Toronto, and the dependent binary variable, whether bike is stolen. The reason why I am interested in exploring possible causality between these two variables is that, as a student studying and living in downtown Toronto, it is of my and my peers' interests to understand whether downtown location has a significant (causal) relationship with bike thefts. That is, it is important to detect whether the most common transportation properties (bikes) for students are in high risk of theft as we study and live in downtown.

Here is how the propensity score matching works and why it is suitable in my case. Basically, regression analysis results from observational data can only be used to infer correlation between predictors and the dependent variable, where observations are observed and recorded in a natural setting without any artificial interventions. Regression analysis from experimental data can be used to infer causality between the treatment (controlled) variable and the dependent variable, where observations are intentionally controlled according to certain characteristics. It is important to explore potential causal relationship. The propensity score matching performs as a quasi-experimental technique, selecting a predictor as the treatment (controlled) group, and is useful to infer causality between the treatment variable and the outcome of interest on an observational data.

Here is how I apply the propensity score matching on my dataset, where the treatment variable is whether the bike is parked in downtown Toronto and the outcome of interest is whether the bike is stolen. I first calculates the propensity score by logistic regression, regressing all other 4 predictors (the cost of bike; whether the bike is in dark colour; whether the bike is of a popular type; the type of building near which the bike is parked) on the treatment variable. Next, I create a new variable called "treated" based on the treatment variable. After this, I use "matching()" function from arm package in R to match pairs from the treated group ("treated"=1) and untreated group ("treated"=0) with similar propensity score by nearest neighbour matching. Finally, I reduce the dataset to a new matched dataset where I only keep matched pairs from the last step. After propensity score matching, I get a new matched dataset where whether the bike is parked in downtown Toronto is the controlled variable. I also examine the quality of propensity score matching by doing t-test on the means of the numerical predictor (the cost of bike) under treated group and untreated group. Last, by running a new logistic regression on the outcome of interest with five predictors from the new matched dataset, I can observe the level of significance of the treatment variable in the model to determine whether parking a bike in downtown has a causal relationship with bike thefts.

The propensity score method also has its drawbacks (Rohan, 2020). For example, propensity score cannot be applied on unobserved variables, which can cause problems in a more realistic setting. The result from propensity score tends to be specific to the model used. Another problem is that, statistically, propensity

score uses the whole dataset twice. What is more, when reducing to a new matched dataset based on matched pairs, there is a significant reduction in the number of observations in my dataset (from 4305 to 914), and this may cause bias problems when I run a new regression with a smaller sample.

Results

The logistic model on whether the bike is stolen with five predictors (the cost of bike; whether the bike is in a dark colour; whether the bike is of a popular type; whether the bike is parked in downtown; the type of building where the bike is parked) has the following model summary:

```
## # A tibble: 7 x 5
##   term                              estimate std.error statistic     p.value
##   <chr>                                <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)                       -0.663    0.188        -3.52  0.000434
## 2 cost_of_bike                       0.000173 0.0000347     4.99  0.000000612
## 3 as.factor(if_dark_colour)1         0.0109   0.0618        0.177 0.860
## 4 as.factor(if_popular_type)1        0.142    0.0761        1.86  0.0628
## 5 as.factor(if_near_city_center)1    0.273    0.101         2.70  0.00696
## 6 location_typeresidential           0.292    0.160         1.83  0.0668
## 7 location_typesocial                0.0771   0.163         0.472 0.637
```

Thus, the model will looks like: $log(\frac{p}{1-p})$ = -0.663 + 0.000173$cost_of_bike$ + 0.0109if_dark_colour + 0.142$if_popular_type$ + 0.273$if_near_city_center$ + 0.292$location_typeresidential$ + 0.0771$location_typesocial$

where $\hat{\beta_0}$ = -0.663 means that, for a bike with a cost of one dollar, in bright color, not of a popular type, parked far from the city center and near an educational building, the log odds of that bike to be stolen is -0.663.

$\hat{\beta_1}$= 0.000173 means that a bike that costs one dollar more will have a 0.000173 higher log odds of being stolen than a bike that costs one dollar less, holding everything else the same.

$\hat{\beta_2}$= 0.0109 means that a bike in a dark colour (black-related colour) will have a 0.0109 higher log odds of being stolen than a bright-coloured bike, holding everything else the same.

$\hat{\beta_3}$= 0.142 means that a bike of a popular type (if a mountain /touring /road /trail /BMX bike) will have a 0.142 higher log odds of being stolen than a bike not of a popular type, holding everything else the same.

$\hat{\beta_4}$= 0.273 means that a bike parked in downtown Toronto will have a 0.273 higher log odds of being stolen than a bike parked far from downtown Toronto, holding everything else the same.

$\hat{\beta_5}$= 0.292 means that a bike parked near a residential place (where people live) will have a 0.292 higher log odds of being stolen than a bike parked near an educational place (school/college/university), holding everything else the same.

$\hat{\beta_6}$= 0.0771 means that a bike parked near a social place, including both commercial and non-commercial places, will have a 0.0771 higher log odds of being stolen than a bike parked near an educational place (schools/colleges/universities), holding everything else the same.

Since log-odds may be found hard to read and understand, I transform log-odds into probability:

After getting the model, I make a comparison between the original model and an alternative model without the variable on type of location where the bike is parked by comparing their AIC, BIC and AUC values, and here is the summary of the model diagnostics results:

```
##                          AIC      BIC      AUC
## original model    5920.108 5964.681    0.559
## alternative model 5928.837 5960.674    0.551
```

Here is the Area Under Curve graphs for the original model and the alternative model:

Table 1: AUC of original model

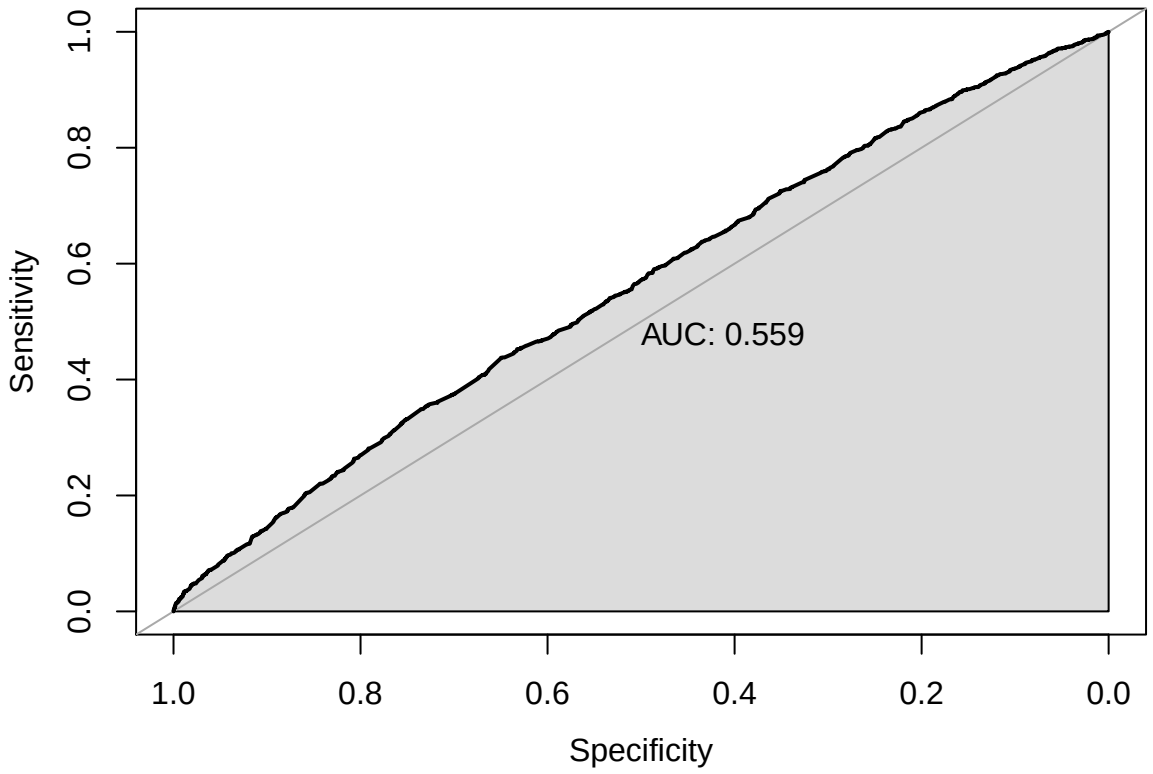
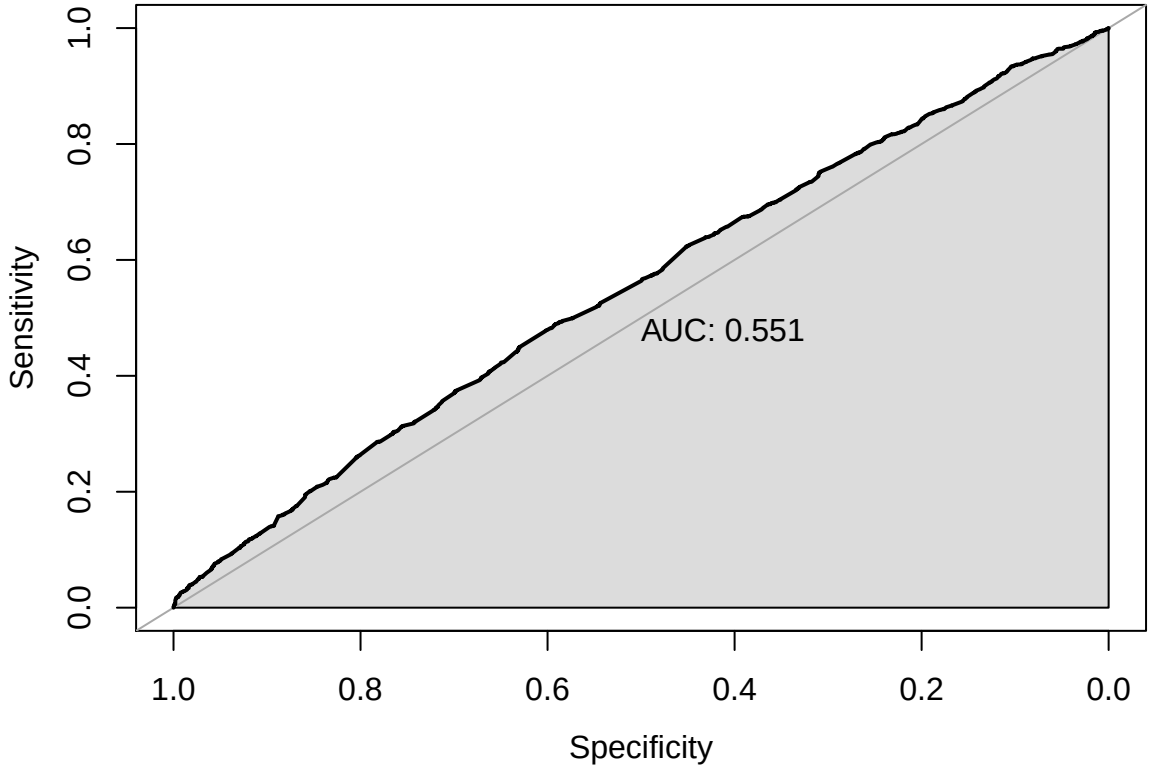


Table 2: AUC of alternative model



From the results of model diagnostics tools (Habshah Midi, S.K. Sarkar & Sohel Rana, 2010), I can conclude

that my original model is better than the alternative since two (AIC and AUC) of the three diagnostics techniques show a preference for the original model.

After model comparisons and diagnostics, I also conduct propensity score matching on the observational dataset to infer causality between whether the bike is parked in downtown Toronto and whether the bike is stolen, with the former as the treatment (controlled) variable and the latter as the outcome of interest.

Having reduced the dataset into a new matched data based on similar propensity scores, I evaluate the quality of propensity score matching by conducting t-test on means of the numeric variable: cost of bike, from the treated group and untreated group.

Here is a summary of the t-test:

```
##
##  Welch Two Sample t-test
##
## data:  near_group$cost_of_bike and far_group$cost_of_bike
## t = -8.0537, df = 543.92, p-value = 5.1e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -419.5521 -255.0212
## sample estimates:
## mean of x mean of y
##  416.4967  753.7834
```

The t-test result shows that I fail to reject the null hypothesis which states that there is no significant difference between means of the cost of bike between two groups, since the difference of means -337.2867 (= 416.4967-753.7834)lies within the 95% confidence interval, from which I can conclude that the propensity score matching quality is good, given a 5 percent significance level.

Last, here is a summary of the new logistic regression on the outcome of interest on 5 predictors including the treatment variable ("if_near_city_center"):

```
## # A tibble: 7 x 5
##   term                       estimate std.error statistic p.value
##   <chr>                         <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)              -0.682     0.322       -2.12   0.0344
## 2 if_dark_colour            0.0167    0.164        0.102  0.919
## 3 if_near_city_center       0.189     0.166        1.14   0.254
## 4 cost_of_bike              0.000314  0.000124     2.52   0.0116
## 5 if_popular_type           0.139     0.231        0.599  0.549
## 6 location_typeresidential  0.257     0.191        1.35   0.177
## 7 location_typesocial      -0.223     0.293       -0.760  0.447
```

The treatment is not significant in the model, even at 10 percent significance level, from which I conclude that there is no causal relationship between downtown location and bike thefts.

Discussion

Summary

So far, I have decided on a research topic on how bike features and parking locations ar associated with bike thefts in the Greater Toronto Area, and mainly I use five independent factors including the cost of bike, color of bike, type of bike, whether the bike is parked in downtown Toronto and type of building near which the bike is parked. Getting an observational dataset called "Bicycle Thefts" from Toronto Open Data Portal website, I use R to run a logistic regression on the binary outcome of interest: whether the bike is stolen, based on five predictors mentioned above. From the summary of the model, one can find that, holding all other predictors the same, an expensive bike has a higher theft risk; a bike in a dark colour has a higher

theft risk; a bike of a popular type has a higher theft risk; a bike parked in downtown has a higher theft risk; a bike parked near a residential or social building has a higher theft risk, compared to one parked in an educational place. I also evaluate the quality of the model through comparing my model with an alternative logistic model excluding the predictor on the type of building where bike is parked. I use model diagnostics techniques such as AIC (Akaike information criterion), BIC (Bayesian information criterion) and AUC (Area Under the ROC Curve), where two (AIC and AUC) of the three model evaluation tools show a preference for the original model. I also use propensity score matching on my observational data to infer possible causality between whether the bike is parked in downtown Toronto (the treatment) and whether the bike is stolen (the outcome of interest). I calculate the propensity score based on the treatment, using logistic regression in R, and match pairs from the treatment group and the untreatment group with similar scores to create a new matched dataset, by matching function from "arm" package in R. Next, I evaluate the quality of propensity score matching through t-test on the means of the numerical variable, cost of bike, under treated and untreated groups, where the result shows that the quality of propensity score matching is good. Finally, I run a new logistic regression on the outcome of interest with five predictors including the treatment based on the new matched dataset. Since the treatment is not significant in the new model, I conclude that whether the bike is parked in downtown Toronto does not have a causal relationship with whether the bike will be stolen.

Results Analysis

Based on the P-values from the model summary, one can observe that four of the five predictors are significant to predict the log odds of bike theft: the cost of bike and whether the bike is parked in downtown Toronto are significant at a 5 percent significance level, while whether the bike is of a popular type and the type of location where the bike is parked are significant at a 10 percent significance level, with whether the bike is in dark color not that statistically meaningful in this model.

By analyzing the log-odds estimates or probability estimates of each predictor, it can be found that all predictors in the logistic regression have positive correlations with bike thefts. That is, if keeping all other predictors the same, a bike with a higher cost will face a higher bike theft risk than a bike cost less; a bike in a dark colour will face a higher theft risk than a bright-coloured bike; a bike of a popular type (including mountain, touring, road, trail and BMX types) will face a higher theft risk than a bike not of a popular type; a bike parked in downtown Toronto will face a higher theft risk than a bike parked not in downtown Toronto; a bike parked near a residential building (where people reside and live) or a social building (including both for-profit and not-for-profit) will have a higher theft risk than one parked near an educational building (schools /colleges /universities). To make the summary of the logistic regression more readable and understandable, I also convert the intercept and coefficient estimates from log-odds to probability, which makes more numerical sense.

Besides, I also compare the original model with an alternative model excluding the variable on the type of building where the bike is parked. I choose this variable based on my and my peers' interests to see whether universities (categorized as "residential" type under this variable) tend to suffer from more bike thefts than other types of locations ("social" and "residential"), and I am therefore not sure whether this variable will play a role in predicting the outcome of interest: whether the bike is stolen. By conducting model diagnostics techniques on two models, including AIC, BIC and AUC, where a better model will have a lower AIC, a lower BIC and a higher AUC at least above 0.5. The result from diagnostics shows that my original model has a lower AIC ($5920.108 < 5928.837$), a higher BIC ($5964.681 > 5960.674$), and a higher AUC ($0.559 > 0.551$). Since two (AIC and AUC) of the three criteria show a preference for the original model, I conclude that my original model is better and more suitable than the alternative one, from which I can infer that the location type of building near which the bike is parked is important in predicting whether the bike will be stolen in Toronto.

Overall, the coefficient estimates of the logistic model make sense in a real-life setting. Here are the useful information provided to the cyclists in Toronto from the model results, and corresponding precautionary actions can be considered:

For the cost of bike (the most significant predictor in the model), the model shows that, keeping others constant, a bike with a higher cost will have a higher theft risk compared to a lower-cost one, and this meets

the common sense that expensive items tend to be stolen than less expensive ones and satisfies my assumption that bikes also share such a property. The reason behind is that stolen bikes will not be used for thieves' personal purposes but will be resold for a financial incentive. Research has shown that where the stolen bikes eventually go will depend on the level of sophistication and professionalism of thieves. The most amateur ones, such as the homeless, may want to resell the stolen bikes on the street for a smaller fraction of the high intrinsic value of the bike, or exchange them directly for what they want to satisfy their survival needs, such as food and clothing. However, the more professional thieves usually steal more bikes and resell them online for a better price (Michael, 2018). Therefore, cyclists whose bikes are expensive should be more concerned when they park their bikes outside their eye reach, and they can double up bike security by using a smart lock. Though a smart lock costs more than a usual lock, compared to the high cost of the bike, it worth a consideration. There is one kind of smart lock that can only be controlled on the owner's smartphone and can alert the owner any motion and the location of the bike. Other types of smart locks can include that, when there is deliberate destruction on the lock, the lock makes an alarming sound or emits smelly gases to repel thieves.

For the color of bike, the model shows that, keeping others constant, a bike in a dark colour has a higher theft risk than a bright-coloured one, and this also meets the common expectation that less observable (usually in a dark colour) properties are more likely to be stolen. From research, the most popular and common bike colour is black (Socaltrailriders Organization, 2008) and bikes with such a dark colour can be stolen without too much attention in public. What is more, black bikes are also theft the most frequently from a Chicago research (Chicago Stolen Bike Registry, 2012), which can support our model results in a global context. Therefore, given the information on the bike colour, cyclists whose bikes are in dark colours should be more vigilant on the bike theft issue. They can recolor their bikes with a bright and conspicuous covering such as red and yellow, or cyclists can customize their bikes with personalized elements that are distinguishable and identifiable, such that even the bikes are stolen, bikes can be easily tracked or thieves would find it hard to resell such a distinctive bike in the markets, making thefts less possible.

For the type of bike, the model demonstrates that, keeping others constant, a bike of a popular type (if a mountain /touring /road /trail /BMX bike) faces a higher theft risk than one not of a popular type, and this also meets the initial assumption that popular items are more likely to be stolen. Popular items usually have a high demand in the market, and, from the perspective of economics, a high demand is associated with a higher market price based on the demand-supply relationship. Thus, it is intuitive to deduce that popular items have a financial incentive for thieves. The previously mentioned research in Chicago also shows that road and mountain bikes (the popular types) experience the most frequent theft (Chicago Stolen Bike Registry, 2012). Therefore, cyclists whose bikes are of a popular type should be more cautious about their bikes, and precautionary actions such as increase the reliability of bicycle locks and decorating with less generic painting or components will decrease the likelihood of theft.

For whether the bike is parked in downtown Toronto, the model shows that, keeping others constant, a bike parked in downtown will have a higher theft risk than one parked far away from the city center, and this result meets the initial expectation that busy and crowded areas such as downtown tends to suffer from property loss than less-populated areas. One CBC news report also supports this result by revealing that most bikes are stolen in downtown Toronto in the last few years, with the Waterfront Communities, Bay Street Corridor, Church-Yonge Corridor, Annex and University area being the five biggest victims of bike thefts (Nicole, 2017). Hence it should be given consideration when cyclists park their bikes in Toronto downtown area, where people can reduce the unnecessary time of leaving their bikes alone, and can bring two bicycle locks or take easily removed items with them rather than leave them on the bike. The Toronto Government can also help to alleviate this issue, by providing cyclists in downtown with better bike racks (for example, with at least two points of contact and allows the frame and wheels to both be locked) and more visible racks in busy areas, such as on the adjacent streets to the ground-level restaurants near the CF Toronto Eaton Center, to further deter bike thefts in downtown.

For the location type near which the bike is parked, it is interesting to find that, keeping others constant and compared to a bike parked near an educational place (school, college and university), a bike parked near a residential place (where people live) or parked near a social place (including both commercial and non-commercial building) has a a higher theft risk, from which I can infer that bikes parked on university campus

should be safer. A residential place suffers from bike thefts because, for example for house neighbourhoods, residents may park their bike outside the house or in the garage. Study has revealed that, of all the places in a house, the garage entrance is one of the easiest to access for thieves (Garage Door Repair, 2018). There are also two surveys, one is in Canada the other is in America, show the same conclusion as my model result. A survey conducted in Montreal supports the idea that bike theft occurs near residential place, where many bike thefts can occur two miles from home (Eric, 2014). The survey from Chicago also shows that the top two types of location that suffer most bike thefts are "sidewalk in front of a open business", corresponding to the "social" sub-group, and "front of building in residential area", corresponding to the "residential" sub-group (Chicago Stolen Bike Registry, 2012). A social place also tends to face theft risks since in public places such as crowded parks and business districts, the visibility and traceability of a bike parked on the streets shall be questioned. Here are possible precautionary actions towards the information on location types. People should be careful even in their own neighbourhoods, where they can always close the garage door, have a quality and functioning garage door that is supposed to be secure, solid and well-built as the front door, also have good surveillance system for areas that people not often go such as backyard and basement where the bikes and other valuable properties may be stored. For cyclists parking their bikes in business areas or public areas such as parks, they should also be careful, using more secured locks, personalizing bikes as mentioned above, and paying sufficient attentions to the parking location.

Besides the protection techniques previously proposed, there are also some other tips for the readers, such as joining a bike recovery group of your community or city on a social media platform, using locks on the wheels when parking the bike and recording the serial number of the bike (if applicable) since even the bike is resold the owner can ask help from local authorities to look for the stolen bike based on the number.

After analyzing each predictor with correspondingly precautionary actions provided, based on the model summary table, I infer possible causality between whether the bike is parked in downtown (as treatment) and whether the bike is stolen (as outcome of interest), by applying propensity matching score on the observational dataset. I regress all other predictors on whether the bike is parked in downtown, using logistic regression in R. Then I match pairs, one from treated group (if parked in downtown) and one from untreated group (if not parked in downtown) with similar propensity score to create a new matched data. After, I evaluate the quality of propensity score matching by t-test on means of cost of bike under each group, where the null hypothesis is that the difference in means equals to 0, and the alternative hypothesis is that the differences in means is not equal to 0. From the t-test summary, one can find that the difference in means is (416.4967-753.7834=)-337.2867 which lies within the acceptance region (from -419.5521 to 753.7834) given a 5 percent significance level, from which I can conclude that I fail to reject the null hypothesis given such a significance level and that there is no significant difference in means of cost of bike under two groups. Thus, the quality of propensity score matching is good.

Finally, I run a new logistic regression on the outcome of interest with five predictors including the treatment, and the model summary shows that the treatment is not significant in the model. Thus, I can conclude that there is no causal relationship between whether the bike is parked in downtown and whether the bike is stolen, and thus people study and live in downtown do not have to worry too much on their bikes just because they are parked in downtown Toronto.

Conclusion

Bicycle theft has become a worsening issue on property security in Toronto and it worth time to analyze and to understand what factors are correlated to a higher bicycle theft risk. By using logistic regression model in R, one can find that the cost of bike, whether the bike is in a dark colour, whether the bike is of a popular type, whether the bike is parked in downtown Toronto and type of locations near which the bike is parked are correlated with whether the bike is stolen. Further, by propensity score matching, there is no causal relationship between downtown location and bike theft. Corresponding precautionary actions, including using more reliable bicycle locks especially in busy areas, personalizing bikes with distinctive components and being cautious about bike thefts even in the neighbourhood, should be considered by cyclists who face the potential theft risks predicted by the model.

Weaknesses

There are some weaknesses related to the original observational data, "Bicycle Thefts", from the Toronto Open Data Portal website, though I have made adjustments to clean the data, such as filtering with useful information and omitting N/As. First, the dataset has limited coverage. The "Bicycle Thefts" only gives information on bikes reported lost in Toronto due to various reasons, including theft, robbery, mischief, careless driving and others, and does not provide with information on all bikes in Toronto. This may cause bias since I analyze effects on bike thefts based on bikes that are actually lost and reported to the Toronto Police Services, not all bikes in Toronto. Second, the dataset only covers a period from 2014 to 2019, where the most recent information in 2020 is to be updated at the beginning of 2021. This may also cause the model results less reflective and representative of the current situation, since research has shown that bike theft has increased in 2020 as cycling as a transportation tool of social distance has attracted renewed interests and the bike theft has increased correspondingly. Including data from 2020 would better to indicate which variable is more significant on the probability of bicycle theft. Third, though the original data has abundant observations (21548), there are many N/As in information categories for many observations, and when I omit N/As to make the data more useful, based on the predictors, there is a big size reduction in the dataset, only 4305 observations kept. This could also affect the accuracy and precision of the intercept and coefficient estimates of the logistic model. Another weakness of the dataset can be limited number of variables (columns), for example, there may be other more relevant information categories that are significant in predicting the probability of bicycle theft than the five predictors I select based on the current available dataset, such as whether the bike is locked before theft and the type of the bicycle lock used.

There are also weaknesses related to the methods I use. First, when I make a comparison between my original logistic model and an alternative logistic model by model diagnostics, I use AIC, BIC and AUC, where only two (AIC and AUC) of the three show a preference for the original model, and this means that the original model is not strictly (100 percent) better. Second, I only use three model diagnostics techniques on the logistics model, but there are actually more applicable methods, such as cross-validation and confusion matrix, that are beyond my reach. Other diagnostics may yield a different preference between two models. Third, when I calculate propensity score, I use logistic regression on the binary treatment variable. There are other models, such as probit model and multinomial models for variables with more than two outcomes, that may be more reflective of the current situation where I can use a variable with three or more outcomes, but again those methods are beyond my reach. Fourth, the propensity score matching has its own drawbacks (Rohan, 2020), for example, propensity score cannot be applied on unobserved variables, which can cause problems in a more realistic setting. The result from propensity score tends to be specific to the model used. Another problem associated with propensity score is that, statistically, propensity score uses the whole dataset twice. Last, when reducing to a new matched dataset based on matched pairs, there is another significant reduction in the number of observations in my dataset (from 4305 to 914), and this may cause bias problems when I run a new regression on the outcome of interest based on a much smaller sample.

Next Steps

Here are what can be done next with respect to problems of dataset. To deal with the limited coverage in the outcome of interest, another census (if applicable) or survey can be conducted to estimate information on all bikes in Toronto. To encourage more cyclists in Toronto to participate in the survey on their bikes, one can use incentives, both financial (such as payments) and non-financial (such as charities), and various mode of contact, such as personalized survey with personal contact, which are proposed by Groves, Cialdini, and Couper (1992), to increase the response rate. To deal with the limited time frame, one can submit requests for getting the most up-to-date data from the Toronto Police Service (if allowed). To tackle the problem of incomplete information collected (with many N/As), there are three general strategies for analyzing incomplete data, summarized by Little (1988), Rubin (1987) and others (more recently): direct analysis of the incomplete data, weighting, and imputation. Taking imputation as an example, people can use R or STATA to conduct such a technique in machine learning. What is more, more potentially relevant factors, such as whether the bike is locked before theft or the type of lock used, can be included in the dataset to make the model more comprehensive.

Here are what can be done next with respect to problems of methods used. More advanced model diagnostics

methods on logistic regression, such as cross validation (which is to test the prediction power of the model by dividing the dataset into training and test subsets) and confusion matrix (whose basic is to compare the actual result with predicted result from a model in a visualized table) can be used after I have acquired sufficient knowledge on them. I shall also apply multinomial model to calculate propensity score on the treatment after diving deeper into the propensity score matching.

References

Toronto Police Services. (2020). BICYCLE THEFTS. Retrieved from https://open.toronto.ca/dataset/bicycle-thefts/

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Sam, F. (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

David, R. and Alex, H. (2020). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.5.https://CRAN.R-project.org/package=broom

Kazuki, Y. and Alexander, B. (2020). tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.12.0. https://CRAN.R-project.org/package=tableone

Xavier, B., Natacha, T., Alexandre, H., Natalia, T., Frédérique, L., Jean-Charles, S., and Markus, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/

Canadian Cycling Magazine. (2020, August 27). Reports of bold bike thieves in Toronto and Vancouver as used bike prices increase. Retrieved from https://cyclingmagazine.ca/sections/news/reports-of-bold-bike-thieves-in-toronto-and-vancouver-as-used-bike-prices-increase/

Lily, H. (2020, November 4). Bike thefts are increasing in Canada: Here's what you can do to protect your bike. Retrieved from https://cyclingmagazine.ca/sections/news/bike-theft-canada/

J.S., C. (2002, November). The Origins of Logistic Regression. Retrieved from https://papers.tinbergen.nl/02119.pdf

Jetsetta. (n.d.). 6 Most Popular Types of Bicycles for Your Next Adventure. Retrieved fromhttps://jetsetta.com/6-most-popular-types-of-bicycles-for-your-next-adventure/

Rohan, A. (2020, November 5). Matching. Retrieved from https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html

Michael, R. (2018, October 11). Underworld Economics: Why Are So Many Bikes Stolen? What Happens to Them? Retrieved from https://www.treehugger.com/underworld-economics-what-happens-stolen-bikes-4858256

Socaltrailriders Organization. (2008, Oct 30). What's the most popular bike color? Retrieved from https://www.socaltrailriders.org/threads/whats-the-most-popular-bike-color.24050/

Archive Organization. (2012, September 4). Chicago Stolen Bike Registry. Retrieved from http://web.archive.org/web/20120904165633/http://chicago.stolenbike.org:80/report-statistics

Nicole, B. (2017, November 28). Here's where your bike is most likely to get stolen in Toronto. Retrieved from https://www.cbc.ca/news/canada/toronto/worst-toronto-neighbourhoods-bike-theft-1.4421633

Garage Door Repair. (2018, June 25). 5 tips on keeping your garage safe from theft. Retrieved from https://wtop.com/overhead-door/2018/06/5-tips-on-keeping-your-garage-safe- from-theft/#:~:text=Garages%20are%20a%20vulnerable%20place,your%20garage%20safe%20from%2 0theft.&text=Of%20all%20th

Eric, J. (2014, April 16). These 8 Depressing Bike Theft Statistics Show Just How Bad the Problem Is. Retrieved from https://www.bloomberg.com/news/articles/2014-04-16/these-8-depressing-bike-theft-statistics-show-just-how-bad-the-problem-is

Habshah, M., S.K., Sarkar & Sohel, R. (2010) Collinearity diagnostics of binary logistic regression model, Journal of Interdisciplinary Mathematics, 13:3, 253-267, DOI: 10.1080/09720502.2010.10700699 https://www.tandfonline.com/doi/abs/10.1080/09720502.2010.10700699?journalCode=tjim20