MGT 6203 Group 47 Project Final Report

S. Adachi K. Colasono E. Maldonado

Bike Accidents in Madrid: How Bike Sharing, Traffic, and Weather Influence Incidents

Github repository: https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-47

Abstract

In recent years, bike-sharing services have gained increasing popularity as an eco-friendly and efficient means of transportation around the world. As such a service quickly evolves, however, the concerns for riders' safety naturally arise. This paper delves into the case of BiciMAD, the municipalized bike-sharing system in Madrid, Spain, which has grown to encompass over 7500 electric bikes serving 21 districts. This study seeks to uncover the potential impact of traffic, weather, and shared bike use on bike-related accidents. Our initial hypothesis was that traffic volume and the amount of rainfall are substantial factors in predicting the number of bike-related accidents. A logistic regression model using the hourly traffic volume, the shared bike rental use, and precipitation as the predictors reveals the significant positive effect of the first two predictors on the bike-related accident. Upon further analysis, a temporal factor considering the time of the day seems to provide better results when predicting the likelihood of cycling accidents.The model could further be improved by integrating additional variables, such as geographical locations and seasonality.

Literature Review

Bike-sharing services have exploded in popularity all over the world. One example from the United Kingdom's Department of Transport is that the average number of miles cycled per person increased from 39 miles in 2002 to 54 miles in 2019 - a 40% increase - and it continues to grow at a steady pace (Yang et al, 2021). Bike-forward cities like Amsterdam and Copenhagen currently have 40% of trips being completed by bikes (Yang et al, 2021).

The growth in the popularity of biking can partly be explained by the substantial increase in the desire for an environmentally friendly way to commute. Not only does biking reduce air pollution but it increases physical fitness, improves mental health, reduces time commuting, and reduces the financial burden of parking fees. Bicimad, Madrid's popular bike rental company, even notes that the bicycle has the highest average speed in the city, among all other types of vehicles.

Bicimad has grown substantially in the popularity of commuting by bike. They recently reached all 21 districts in Spain, with 7500 rentable bikes and 611 stations (Bicimad, 2023).

However, with every growth spurt comes growing pains. In Madrid, bicycling accidents increased from 339 in 2010 to 891 in 2019 (Mobility, 2023) - that's a 163% increase in nine years! In Italy, the mortality index for cyclists is 1.42, which is double that for car users (Prati et al, 2017). Across the pond, the US Department of Transportation cited that cyclists are 12 times more likely than passengers in cars to be killed (Yang, 2021).

Yang et al. examined roughly 49,000 road crashes where at least one cyclist was injured or killed from 2011 to 2013 in Italy (2021). They found that 31% of the fatalities happened at junctions - and that rear-end collisions were the most dangerous. Through their Bayesian network analysis, they identified three key predictors for the severity of bicycle crashes including crash type (rear-end vs angle collisions), road type (urban vs rural), and type of opponent vehicle (car, truck, motorcycle, etc.). A different paper by Prati et al., which analyzed cyclist accidents in Liverpool, cited that the most important factors in determining the severity of an accident are the cyclist's age, district, day, and the type of opponent car (2017). Other researchers have used speed limits as a way to measure traffic to predict accident severity (Yang et al., 2021).

Within this paper, we explore the importance of traffic density in predicting the frequency of bicycle accidents. Daraei et al includes traffic as part of their model, but it is represented by a proxy variable that is calculated as a network feature of the set of intersections linked to the street (2021). Vandenbulcke et al. (2014) also account for traffic in their model to assess the risk of cycling accidents in Brussels. In that research, authors segment traffic volume by low vs high passenger car traffic.

Several studies have researched the potential increase in cycling accidents because of bike-sharing service (see Kim et al[7] for a comprehensive list), but there are not any studies on the influence of bike-sharing services on bicycle accidents. However, the purpose of this study is to examine this relationship directly. Our final model delivers on the likelihood of a cycling accident occurring, while also measuring the impact of weather, traffic density and time of day on the likelihood of a cycling accident occurring.

Study Framework

The data for this research comes in an hourly fashion, therefore our approach adopts an hourly granularity for analysis. Our working hypothesis is that traffic volume and the amount of rainfall are substantial factors in predicting the number of bike-related accidents. It is our assumption that traffic volume and weather will have a substantial positive impact on the number of accidents.

Data

The city of Madrid, Spain's capital, is one the most populated cities of the European Union. Mardrid, specifically, has had a well-established bike-sharing system since 2014. Most of the data collected for this paper was provided in an open source format by the Madrid government. This study aims to predict the likelihood of a bicycle accident based on the weather condition, bike rental service demand, traffic density, and time of day. We obtained the open source data on each variable and processed following the steps below to be best utilized in the experiment. One issue we encountered with our data is the lack of actual accidents - only 10% of the hours in 2022 had a bicycle accident. We address the possible implications and solutions to this problem in the Model Selection section.

1. **Weather data:**
    a. The time/date format was changed from 12 to 24 hour time to coincide with the other data sources.
    b. Several grouping variables were created to potentially use as dummy variables in the model including: seasons (i.e., winter, spring, summer, fall), temperature (i.e., hot (>= 75F), moderate/mild (56-75F), cold (<= 55F)), condition (i.e., fair, cloudy, precipitation, fog/haze, windy)
    c. The dataset originally reported the weather on a bi-hourly basis, but the data was reduced to only an hourly. This was done by selecting the rows associated with the whole hour (i.e., 12, 13, 14, etc.). It was unlikely that the weather would change dramatically in 30 mins, so the top of the hour was taken for analysis.
    d. A new column rain, a binary variable that represents whether it rains, was added based on the aforementioned condition.
2. **Bike accidents:**
    a. The format of each accident location was converted from UTM to latitude/longitude format.
    b. A new column "hour" was created based on the accident time.
    c. A new column "cluster" was added to indicate the nearest traffic station ID based on the geographical location of each accident.
    d. The columns that are included in the cleaned dataset are date, hour, and cluster.
3. **Bike rentals from "BiciMAD":**
    a. Incorrect inputs and outliers were removed by identifying the inputs with negative trip duration.
    b. A new column "hour" was created based on the rental starting time
    c. A new column "cluster" was added in the following ways, which results in two datasets to be compared in the model:
        i. Only rental starting location was used to be assigned the nearest traffic station for each trip
        ii. The nearest traffic station was assigned to both rental starting location and ending location for each trip, and two entries were created for each trip.
    d. The columns that are included in the cleaned dataset are date, hour, and cluster.

4. **Traffic:**
    a. The 0 values in the dataset were not removed since we don't have certainty of their meaning. A 0 value may indicate that the instrument is not measuring correctly, or that due to some road issue there is not traffic at all.
    b. Original data included 12 columns numerated HOR1, HOR2, …, HOR12 indicating the hours, while a column "FSEN" indicated the direction of the traffic and the period of the day (first 12 hours or second 12 hours). With these 13 columns the table described the number of cars passing by the respective station (column FEST) in a given day (column FDIA). Twenty four (24) new columns were added replacing the original 13. The 24 columns (Hour1,Hour2,..., HOUR24) indicate the total traffic (both directions) for the station (Station) on a given day (Date).

Exploration/Initial Discoveries

After aggregating the data by hour, some clear temporal patterns emerge, as illustrated in Figure 2. Peak demand for bike-sharing resources occurs between 7-9 am, 1-3 pm, 6-8 pm and 11 pm - 12 am. This aligns with the usual morning, lunch and dinner "rush" hours. A similar pattern appears in time series data for bicycle accidents. Accidents occur most often at 10 am, 2 pm, 7 pm and midnight. Again, the density of traffic increases around the same times as the bike-sharing demand and accidents increase. From these charts, the relationship between bike-sharing demand, traffic density and bicycle accidents can be seen.

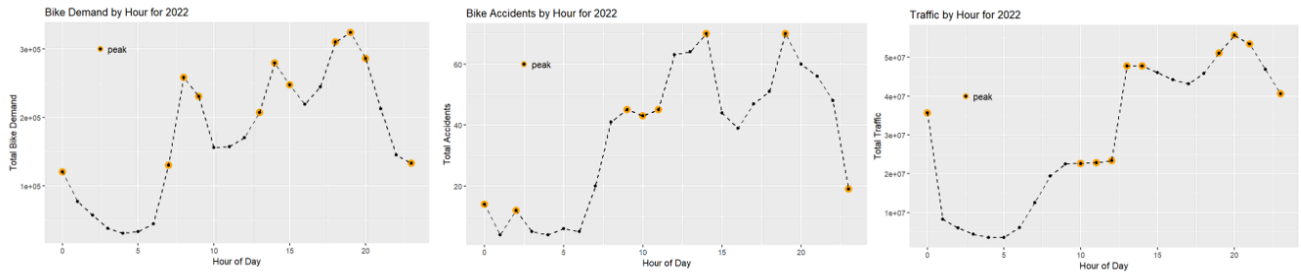

Figure 2. Total Number of Bike Accidents, Bike Demand, and Traffic in Madrid in 2022 aggregated by hour

Furthermore, the correlation coefficients for the total number of accidents and the predicting variables (i.e., total traffic, bike demand and hour of the day) are the following, respectively: 0.14, 0.16, 0.12. The correlation coefficients are all positive but weak.

Additionally, the skewness in the traffic data poses a concern in modeling (Figure 3). Numerous instances of minimal to no traffic, likely occurring during nighttime, contribute to significant variance in the data. This skewness has to be dealt with in the modeling process.

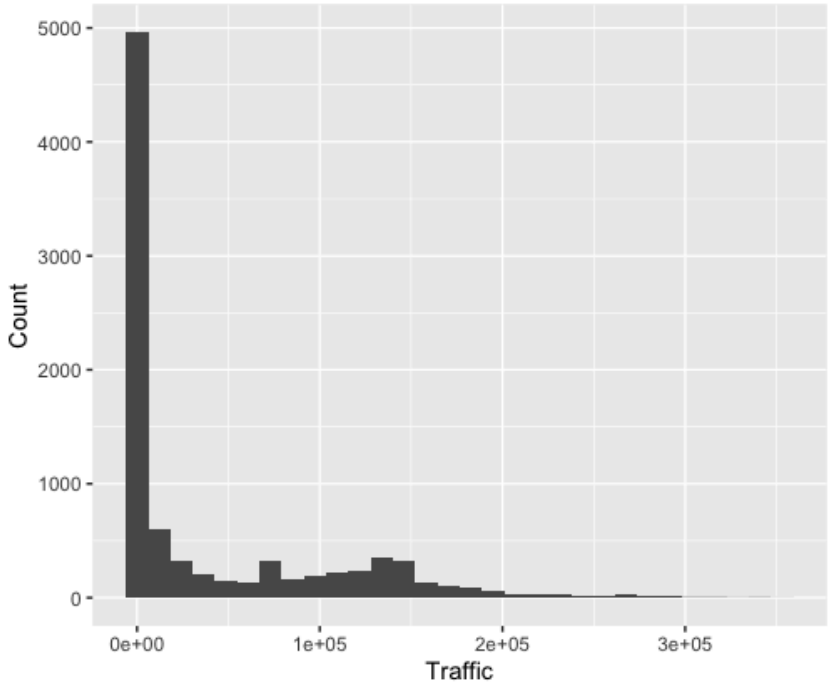

Figure 3. Traffic volume distribution aggregated by hour

Model Selection

Given the aforementioned concern regarding the small cases of accidents, a logistic regression model was deemed most appropriate to predict the odds of an accident. Additionally, based on the peak hours discussed above, aggregating the data by a six-hour time period (morning, afternoon, evening, and night) rather than each hour allowed us to build a more robust model.

Our initial goal was to use logistic regression to predict the probability of a bike accident based on bike rental activity, weather, and traffic volume. The data was split into a training and testing data set. The training set was obtained through random selection of 85% of the data and 15% was selected to be the testing set. 10-fold cross validation was run as a part of the model selection process.

To start off, we built a all-factor model to understand the significance of each predictor:

$$log(accident) = \beta_0 + \beta_1 \cdot bike\_rental\_count + \beta_2 rain + \beta_3 \cdot traffic + \beta_4 \cdot Morning + \beta_5 \cdot Afternoon + \beta_6 Evening + \beta_7 \cdot Weekend$$

Where *traffic* represents the number of cars that pass by all the stations, *bike use* is the number of bike rentals, and *rain* is whether it rains or not during each time period. *Morning* is from 6 am to 12 pm, *afternoon* is from 12pm to 6pm, *evening* is from 6pm to 12am, and the base case is *night* which is from 12am to 6am. The predictor variable *Weekend* is a binary variable to indicate whether it is weekend or not. The response variable, accident, is the total number of accidents

per time period for each day in 2022. The predictor variables are also on a per time period/per day basis. When running the model with the aggregated data, the following results were achieved:

```
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.334e+00  1.797e-01 -12.986  < 2e-16 ***
rental_count          2.432e-04  4.659e-05   5.220 1.78e-07 ***
rain1                -1.957e-01  1.775e-01  -1.102    0.270
traffic              -2.385e-07  1.513e-07  -1.576    0.115
time_periodMorning    1.023e+00  2.035e-01   5.027 4.98e-07 ***
time_periodAfternoon  1.684e+00  2.268e-01   7.425 1.13e-13 ***
time_periodEvening    1.455e+00  2.323e-01   6.266 3.70e-10 ***
weekend1              1.626e-01  1.314e-01   1.238    0.216
```

Figure 4. Summary of all-factor logistic regression model with aggregate data

All of the predictor variables, except for the bike rental and the time periods, are not statistically significant at the 0.1 level (Figure 4). The coefficient for rental bike count was positive, implying a potential effect on accidents. Similarly, the time periods of morning, afternoon, and evening all have positive coefficients, which also suggest an impact on accidents relative to the base case of the night time.

Multiple models were compared based on their performance metrics, such as AIC and BIC. The predictor variables were selected through a stepwise selection process. We also attempted transforming the traffic predictor to mitigate its skewness. Specifically, we substituted the traffic predictors in the all-factor model above with the natural logarithm-transformed traffic volume. The transformations involved employing both the natural logarithm (ln) with the formula ln(x+1) to address instances of zero values, and logarithm base 10. Nevertheless,untransformeted data performed better for the model. Subsequently, we conducted stepwise variable selection to refine the model. Following a thorough selection procedure, it was confirmed that the chosen model below yielded the most favorable results:

$$log(accident) = \beta_0 + \beta_1 \cdot bike\_rental\_count + \beta_2 \cdot Morning + \beta_3 \cdot Afternoon + \beta_4 \cdot Evening + \beta_5 \cdot Weekend$$

The predictors for rain and traffic got eliminated while the others were selected. The bike rental amount, all the time periods, and the binary variable for weekend are all significant at the 0.05 level. (Figure 5) The model's BIC was 1404, which was the smallest among the ones that were experimented.

```
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.334e+00  2.034e-01 -11.477  < 2e-16 ***
rental_count         2.561e-04  5.358e-05   4.780 1.75e-06 ***
time_periodMorning   9.265e-01  2.350e-01   3.942 8.07e-05 ***
time_periodAfternoon 1.530e+00  2.557e-01   5.984 2.18e-09 ***
time_periodEvening   1.112e+00  2.584e-01   4.303 1.68e-05 ***
weekend1             3.159e-01  1.533e-01   2.060   0.0394 *
```

Figure 5. Summary of selected logistic regression model with aggregate data

We moved on to use the testing set to evaluate the model performance. The testing showed the following model performance with the optimal cut-off value of 0.366 selected based on the ROC curve (Figure 6):

AUC	0.768
Accuracy	0.722
Precision	0.848
Sensitivity	0.679
F1 Score	0.754

Table 1. Selected model performance

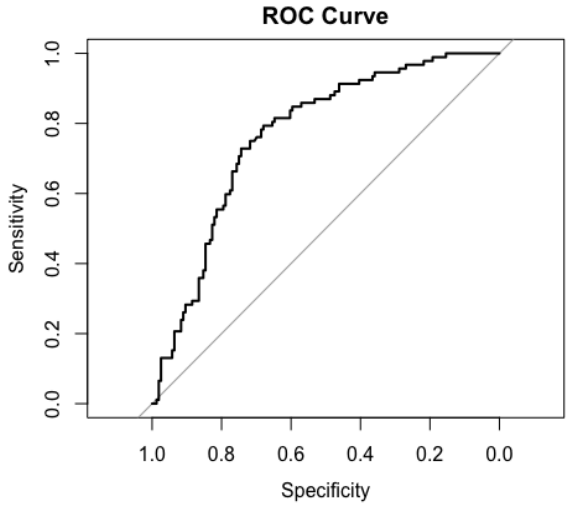

Figure 6. ROC curve of the selected model

Discussion

Given this result, we can get the following probability function on whether an accident happens during a six-hour time period:

$$P(accident) = \frac{1}{1 + e^{-(-2.33 + 0.000256 \cdot rental\_amount + 0.926 \cdot Morning + 1.53 \cdot Afternoon + 1.11 \cdot Evening + 0.316 \cdot Weekend)}}$$

The selected model indicates the significant positive impact on the odds of an accident by all the predictors. In other words, as more people use the bike rental service, the odds of a bike-related accident increases. Therefore, it can be reasonably argued that bike rental services contribute to the occurrence of bike-related accidents. Similarly, the odds of bike-related accidents are significantly higher in the morning, the afternoon, and the evening relative to the night time. The afternoon has the largest coefficient, meaning the most significant impact on the odds of an accident, followed by the evening and the morning. This could be intuitively explained by the amount of people who are at risk of being involved in an accident. It is also shown that the odds are higher on the weekend relative to the odds on the weekdays. An example probability may be that 150,000 bikes were rented on a weekend morning. The probability that a cyclist may get into an accident is almost a 100% chance. This model can easily be used to determine which times of the day where more precaution be taken by the government, pedestrians and cyclists to avoid accidents.

Traffic and precipitation, on the other hand, were ruled out as the predictors in the selected model. Contrary to our initial hypothesis, the model suggests no effect of traffic and precipitation on bike-related accidents. Combined with the increasing risk during the daytime due to a larger number of individuals, the insignificant effect by the amount of traffic implies that bike-related accidents could be largely attributed to pedestrians or bikers rather than cars.

Future work

We started this study with a regression model using the traffic volume, the number of bike rentals, and the precipitation as the predictors to predict the bike-related accidents. As discussed above, building a similar model with the number of pedestrians rather than the car traffic volume will give us a better understanding of the true nature of bike accidents. This approach is likely to generate valuable insights for crafting policies and regulations aimed at reducing bike-related accidents.

As the next step, additionally, integrating additional variables, such as geographical areas and seasonality, will allow us to gain insight into the underlying cause of bike related accidents. As we discussed above, we revealed the different traffic volume and the bike rental activities inside and outside the city loop. The use of the traffic stations as “clusters” for the bike sharing and accident data, for example, could illustrate the effect of geographical locations when predicting the bike accident. The additional variable to be considered in the model is seasonality. The traffic volume and the bike rental usage are very different between weekdays and weekends. The patterns of the weather and bike rental are also different between seasons, too.

Considering the seasonality factor will thus help us better understand the effect of each predictor on bike accidents.

The infrastructure across the world is lacking to accommodate the demand for cycle-friendly cities. Cities need more information on where the demand for cycling lies and how to provide proper infrastructure that reduces bicycle accidents. Madrid's city officials could use the result of the present research to create pilot programs focused on specific times of the day, afternoon hours, where the likelihood of cycling accidents is high. Such pilot programs might implement temporary bike lanes, increasing the presence of traffic authorities, or adding signs to increase awareness of the risk factor.

Citations

Bicimad. (2023). La Revolucion Bicimad. https://www.bicimad.com/index.php/blog/5-motivos-para-unirte-la-revolucionbicimad. Accessed 1 November 2023.

Daraei, S., Pelechrinis, K. & Quercia, D. A data-driven approach for assessing biking safety in cities. EPJ Data Sci. 10, 11 (2021).

Kim, Tae You, et al. "Prediction of Bike Share Demand by Machine Learning: Role of Vehicle Accident as the New Feature." International Journal of Business Analytics (IJBAN) 9.1 (2022): 1-16.

Mobility (2023). Evolution bike accidents in Madrid. https://mobility-friendly.com/en/evolution-bike-accidents-in-madrid/. Accessed 1 November 2023.

Prati, G., Pietrantoni, L., & Fraboni, F. (2017). Using data mining techniques to predict the severity of bicycle crashes. Accident Analysis & Prevention, *101*, 44-54.

Yang, Z., Smith, J., & Robert, B. A. P. (2021). Risk analysis of bicycle accidents: A Bayesian approach. Reliability Engineering & System Safety, 209.

Vandenbulcke, G., Thomas, I., & Panis, L. I. (2014). Predicting cycling accident risk in Brussels: a spatial case–control approach. Accident Analysis & Prevention, 62, 341-357.