**Metro Interstate Traffic Volume Project**

Simon Gamboa 300353695,

Taweepong Hirikul 300356615

CSIS 3360-001: FUNDAMENTALS OF DATA ANALYTICS

Summer 2022

Dr. Samuel Otim

July 6, 2022

**Part 1: Discovery**

**Learning the Business Domain and Framing the Problem**

The public transportation is one of the most important components for any society. Millions of people are moving per minute around the world. An efficient transportation system depends on multiple variables such as logistic transportation, transport demand, urban rail/bus system construction, and so forth. All those variables can be manipulated by humans (anthropogenic factor) to offer an efficient transportation network. However, there are factors or events that that human can not handle and adversely affect the human mobility such as extreme weather or better known as climate change (Ji et al., 2022). According to United States Environmental Protection Agency (EPA)[[1]](#footnote-1) extreme weather has a negative effect on transportation network performance. For example, in transportation system as railways high temperature (hot waves) cause rail tracks to expand and buckle. These phenomena cause a delay in the schedule and decrease speed limit due to rail conditions to devoid derailments.

The focus on our data analysis is analyzes the impact of temperature, rain, and snow on traffic volume and the US national holidays in Minneapolis and St Paul Minnesota from 2012-2018. We will make a series of predictions under three different climate change sceneries: slight, moderate, and extreme.

**Initial Hypotheses**

This data analytics project work focused on analyzing and examining the following questions:

* Do temperature, raining time, snowing time have a significant impact on traffic volume?
* If so, does the impact is negative which holiday days would be more affected?
* Which climate change scenarios have a negative effect on the traffic volume?
* Which weather conditions determine the increase or decrease of traffic volume?

**Identifying Potential Data Sources**

This Metro Interstate Traffic Volume dataset[[2]](#footnote-2) has 9 columns and shows the information on various weather, temperature, and a holiday which contains as following:

* holiday,
* temp,
* rain\_1h,
* snow\_1h,
* clouds\_all,
* weather\_main,
* weather\_description,
* date\_time,
* traffic\_volume.

**Part 2: Data Preparation**

**Data inventory**

The data and variables are related, and this data set does not contain empty or nullใ Moreover, the data and variables of this dataset can be used to predict objectives through data preparation, process, and model implementation.

**Data processing**

Since the dataset consists of variable categories (holiday, weather), we decided to group those variables and convert them into dummy variables:

* holiday variable - This data set includes a variety of holidays. So, we grouped the data into only 2 groups, holiday (Yes) and normal day (No).
* weather variables - weather conditions are very diverse in this data set, such as rain, drizzle, fog, thunderstorm, etc. We group the data into 4 groups as follows:
  + bad weather
  + moderately bad weather
  + moderately good weather
  + good weather

**Part 3: Model Planning and Implementation**

We plan to implement a multilinear regression analysis to look at how variables are correlated with each other. Also checking for the outliner value of the dataset then we will use the log function to compare the result that which model can provide the best result for this dataset.

Since this dataset contains numerical and categorical data, it is important to find a model that best fits this dataset based on the purpose of this project which is Traffic Volume prediction. so, we're focusing on the regression model. therefore, we must prepare and transform the data to be used for regression and prediction models. We will also compare results from various techniques to select the best model and the best score to implement and predict.

As the assumption of this project is that temperature, holidays, time and others affect the traffic volumes. The analysis requires a large amount of data and processes to analyze the correlation of each variable. It is difficult and can make a lot of mistakes when using people to do this task. Therefore, it is necessary to use analytical techniques to help analyze the data and the correlation of various variables. To reduce errors and increase the efficiency of work. It also saves a lot of time as the analytical technique takes much less time to analyze.

**Part 4. Results Interpretation and Business Implications**

**1. Assess the validity of the results. Questions to consider when interpreting the results:**

**• Does the model/analysis appear valid?**

The best model does not appear to be valid since their Multiple R-squared and Adjusted R-squared was 0.4201 and 0.42 respectively. We assess several models even with data log transformation. In the best model most of independent variables were statistically significant. Although our best model got a statistically significant p-value this only explains a 42% of variation. Despite a low R-squared, it is possible to makes predictions based on actual data set as shown in **figure 1**. Due to the thousand of data point from data set a slight variation of some independent variables can impacting over response variable.

|  |  |
| --- | --- |
|  |  |

**Figure 1.** Predicted values from training data set which was get from the best multilinear regression model

**• Does the model/analysis output make sense to the domain experts?**

Yes, it does. The best model explains interesting variation in the variable response from independent variables. In our best model when the temperature increases one unit the response variable (traffic volume) has a positive effect. It makes sense since the temperature in seasonal places can change the life dynamics. For example, in winter seasons the traffic volume is reduced because environmental conditions make harder driving on road. So, as the seasons change with an increase in temperature more people are going out to enjoy the sunshine and hot days therefore the volume of traffic increases. The best model adjusted a predicted scenario such as raining = 10%, Clouds = 20%, daytime = 09:00 am modelling different temperature values. We can see that an increase in 10 points of the temperature (°K), the traffic volume predicted values were increased proportionally as shown in **Table 1**. On the contrary, the rainy periods variable showed a negative effect on traffic volume. Our model predicted that there would be 1.5 times less traffic volume when raining period raising around one hour as shown **Table 2**.

Table 1. Predicted temperature values from the best model

|  |  |
| --- | --- |
|  |  |

Table 2. Predicted Raining values from the best model

|  |  |
| --- | --- |
|  |  |

The data analytics project work focused on analyzing and examining the following questions:

* **Does temperature, raining time, snowing time have a significant impact on traffic volume?**

The best model showed that the temperature has a positive effect over traffic volume, While the raining variables has a negative effect over traffic volume. Moreover, we found that the snowing time variable was not statistically significant and was removed by the stepwise model.

* **If so, does the impact is negative which holiday days would be more affected?**

In general, the best model showed that when there is a holiday date it has a positive effect on traffic volume.

* **Which climate change scenarios have a negative effect on the traffic volume?**

Modelling the raining variable in different daytime periods would be the most public interesting because its impact could reduce the human mobility.

**• Do the parameter values make sense in the context of the domain?**

According to the alpha and beta parameters our model make sense in the climate change context. The predicted model shown that temperature, raining periods, clouds, and weather conditions variables affect the traffic volume variable (response variable). the temperature, clouds, and weather conditions have a positive effect while raining periods has a negative effect as shown in the **table 3**.

**Table 3.** Slopes of independent variables of the best multilinear regression model

|  |  |
| --- | --- |
|  |  |
| **Temperature (°K)** | **Raining range 1 hour** |
|  |  |
| **Clouds (%)** | **Weather condition (moderately bad)** |
|  |  |
| **Weather condition (moderately good)** | **Weather condition (good)** |

**• Is the model/analysis sufficiently accurate to meet the goal?**

The model must be improved adding another meteorological condition that consider variations in the context of climate change if we want to get an accuracy model. For example, percentage of visibility, wind speed, night-time (fall-winter season) and so forth. Despite the model has data points enough, probably, these independent variables have not the direction and strength enough of the association between variables. This is the reason why our model only would explain a variation of 42%.

**• Are more data or inputs needed?**

We consider that the data set has data point enough. Since the data was collected in a short time range, for too long time period (years). If we want to explain variation in the climate change context more meteorological variables should be included than data points.

**• Is a different form of the model/analysis required to address the business problem?**

We can explore scaling or another types of transformations. But also, fitting a nonlinear model would be an option when the linear model is inappropriate. Some authors suggest implementing parametric and nonparametric test, but it depends upon the complexity of data if we decide adding new variables. We know that there exists a linearity between independent variables with response variable. It was demonstrated in our preview best model. One possible resource could be using penalized linear regression (i.e., Lasso or Ridge regression models).

**Part 5. Application of the Model for Predictive Analysis (if applicable)**

The volume of traffic on the roads represents enormous challenges for the city, such as mobility, transport of supplies, public transport, among others. The increase in traffic volume produces vehicular congestion and a serious impact on the environment due to pollution, which escalating to a global context produces greenhouse gases causing changes in the climate such as temperature increases, rainfall, and drought. The modeling of traffic volume as a function of meteorological variables makes it possible to predict the dynamics of mobility in the city. Changes in temperature or rainfall forecasts help determine the flow of vehicles on the main roads.

One of the advantages of predictive model applications is the possibility of regulating peak hour traffic volume based on weather conditions. Since the weather forecast provides prior information, it can be implemented in the model to make decisions about traffic volume for the following days. Another advantage offered by the predictive model is that the combination of weather variables with daily time is that in hours with low traffic volume it is possible to implement road maintenance activities.

The company can create a traffic volume estimation system to allocate trains to meet the demand of passengers during different times and weather conditions of the day. such as

* Temperature: 311.15 (kelvin)
* Clouds: 75 (percentage)
* Moderately good weather
* 7 a.m.

The system can predict the number of users is 3571 people.

Based on this information, the company knows the results of the traffic volume prediction and can prepare various work plans such as

* Decide on increasing the number of trains in service
* reducing customer waiting time for each train

To reduce problems and increase the service to be sufficient to meet the needs of customers effectively.

**Part 6. Conclusion**

1. United States Environmental Protection Agency EPA. (2022). Climate Impacts on Transportation. <https://www.epa.gov/transportation-air-pollution-and-climate-change> [↑](#footnote-ref-1)
2. Metro Interstate Traffic Volume Data Set. Recovery from <https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume> [↑](#footnote-ref-2)