Weather Data Analysis: Impact on Road Accident Frequency and Classification Model Building

Teng Yun, *16098399D*

Abstract—Road accident frequency may be influenced by several factors. This report investigates the relationship between weather and road accident severity which measured by vehicle collisions in New York, USA. The weather obtained using weather API on the OpenWeatherMap website and vehicle collisions data acquired in NYC OpenData website. In this article, we will try to find the crucial weather feature which is the principal cause of accidents. Also, we will try to build an good classification model to predict and prevent the future accidents efficiently.

1 Introduction

Road accidents has long been recognized the consequence of the combined effects of people behavioral, technological, and environmental factors. In recent years increased attention has been directed at the effects of weather on road accidents' frequency and severity [1].

However, the weather's impact on accidents is not simple and obvious. It seems that frog or heavily rain may lead to a high accident rate. But it may not be always true because people may prefer walk or stay home in a bad weather. Therefore, digging the data and finding the real significant feature is needed and important.

In this article, we will use the weather data obtained using weather API on the OpenWeatherMap website and vehicle collisions data acquired in NYC OpenData website. First, this article will introduce the data preprocessing and data exploration procedures and get an overview knowledge with our data. Then this article will introduce some simple regression analysis to find important features influencing the accidents frequency. At last, we will try to build a classification model basing on several algorithm.

The aim of this article is to analysis how much attributes each weather features may cause to the accidents and try to find some insights about weather impact on traffic accidents. Also, based on the knowledge we found, we will build a simple classification model to help people predict potential accidents and preventing them.

2 Data Preprocessing

2.1 Data Preview

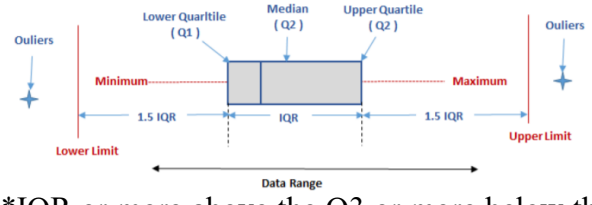
In this article, we mainly used two datasets for our research. The first is the Historical Hourly Weather Data from 2012-2017. It records the temperature, humidity, pressure, wind direction, wind speed, and weather description hourly. Because wind direction may have poor extensibility and the dependence on city's geographic information have been too strong. So, in this article, we will ignore this feature. The dataset includes 30 US and Canadian Cities with 6 Israeli Cities. In this article, we mainly focus on the data in New York, USA. The second dataset is NYPD Motor Vehicle Collisions data which records every vehicle collision accident in New York from 2012-2018. For this dataset, we mainly used the datetime and count the total accidents number in a given time period to measure the frequency of traffic accidents.

2.2 Data Cleaning

Data cleaning for both datasets are mainly using Python.

First, the missing values should be handled. In this research, we fill the missing data with the last valid observation. Especially for weather data, the data is sorted with time. Weather data will be similar when time is close. So, fill missing data with the closest data should get better performance.

Second, the noise data should be handled too. For this task, this research using Graphical (Box Plot) approaches to identify the outlier. are either



1.5*IQR or more above the Q3 or more below the Q1 which we should delete.

After detection the missing value and outlier, we can test our data's quality with SPSS modeler Data Audit Node.

- *Weather data can be found at: https://openweathermap.org*
- *Vehicle collisions data can be found at: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95*
- *Cleaning data will be attached in this article with the related code.*

2.3 Data Integration

For weather data, we have five csv files, and each represents a factor of weather. We merged these six csv files on date time and filter it by New York. After we merge the data, we can aggregate them in unit of 4 hours. We use average value in each unit for temperature, humidity, pressure, wind speed. We used last value to in each unit for weather description.

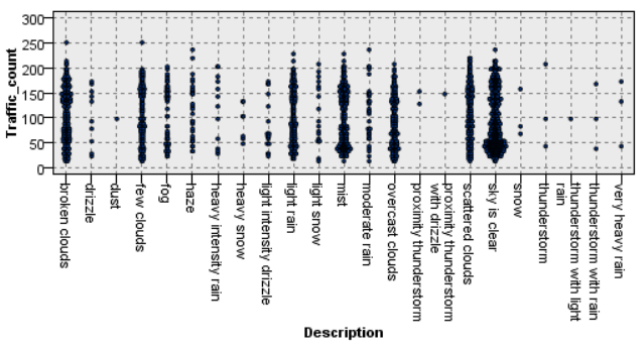
For traffic data, we will count the traffic accident number in every 4 hours period. And we will merge the number of the traffic accidents with our weather features data.

3 Data Exploration and Virtualization

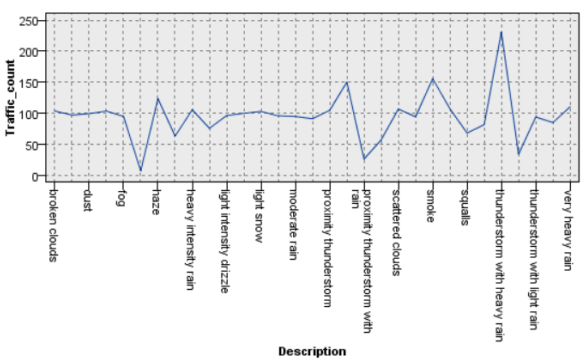
We use Data Audio and GraphBoard Node in SPSS modeler to explore our data. First, we make an overview for our data:

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Datetime		Continuous	2012-10-0...	2017-11-30 00:00:00	...	...	...	...	10201
Description		Nominal	...	...	...	...	...	31	10201
Temperature		Continuous	254.251	309.483	285.906	10.046	-0.272	...	10201
Humidity		Continuous	13.750	100.000	66.837	18.317	-0.226	...	10201
Pressure		Continuous	997.000	1037.750	1017.912	7.501	0.069	...	10201
Wind_speed		Continuous	0.000	6.750	2.900	1.491	0.577	...	10201
Traffic_count		Continuous	6	284	98.398	54.401	0.283	...	10201

Also, we try to find the relation between weather description and traffic accident count using 2-D Dot Plot graph:



Through this graph, we can find that there is no significance relation between weather description and traffic accident counts. In each weather description label, we found that the traffic counts distribution is similar. So, we try to draw the average count plot to find more interesting information. In the graph, we could find that the plot is stable at most weather description labels. The interesting point is that the weather of haze and frog tend to a lower accidents rate and thunderstorm with heavily rain tend to a higher accidents rate



We could get a conclusion that the bad weather like heavily rain or snow doesn't always indicate the high traffic frequency except the very extremely weather condition like thunderstorm with heavily rain. This conclusion holds the same point with Edwards [4]. Sometimes, the bad weather, like fog or haze, can even decreasing the accidents rate. Codling [7] points out that fog weather condition can reduced traffic flow thus decreasing the serious accident rate which can be explain and support our conclusion.

In addition, we try to find the Pearson Correlation between traffic accident count and other 4 weather factors.

Traffic_count

Pearson Correlations

Temperature	0.108	Strong
Humidity	-0.039	Strong
Pressure	0.012	Weak
Wind_speed	0.007	Weak

In our correlation analysis, we assume that temperature, humidity, pressure, and wind speed are continuous variable. Also, we assume that correlation strength is strong when p value is in range of (0,0.05). Correlation strength is labeled medium when p value is in range of (0.05,0.10) and labeled weak when p value is in range of (0.10,1).

Through the Pearson Correlation values, we know that correlation strength (measure by p value) between Traffic counts and wind speed is week. Also, we can get the similar conclusion that correlation strength between Traffic counts and pressure is week too. Traffic count and Temperature's correlation strength is strong and Cor (Traffic count, Temperature)>0 which points out that Traffic count and Temperature are positive correlated. In other hand, humidity also has strong correlation strength with traffic count and Cor (Traffic count, Humidity) < 0 which indicate negatively correlation.

The result gives us the knowledge that pressure and wind speed are not significantly influenced the traffic accident frequency. Also, we can get more valuable insights that the higher the temperature, the more likely the accident will occur. Similarly, we can get the knowledge that the lower the humidity, the more likely the accident will occur. This pattern and trendy were seldom mentioned in related works [1][4][5][6][7]. We conjecture the reason for this result may be related to the mood and mind of

driver. High temperature and low humidity may influence people's mood and mind [8] which is the key factor of accidents frequency.

4 Regression

Regression model is one of the most widely used techniques for analyzing multifactor data [2] In this article, we also applied traditional regression model on our data to find the relation between weather factors and traffic accident counts. For this task, we mainly used Regression node in SPSS modeler and measured the model accuracy by adjusted R square.

First, we normalized the continuous data. We rescaled the range of features by linear function below:

$$x_{normalization} = \frac{x - Min}{Max - Min}$$

After we do the correlation analysis, we know that the high correlation strength features with traffic accident are humidity and temperature. So, we will remove the weak features and just use this two feature to do the regression.

Then we set the Traffic count as our target feature and temperature, humidity as our input features. We can get a regression model:

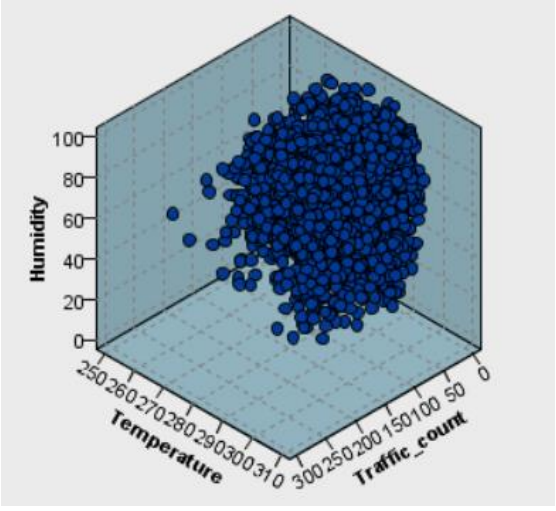
$$\text{Traffic\_count} = \text{Temperature} \times 0.1145 + \text{Humidity} \times (-0.02698) - 0.000358$$

We analysis the performance of this model:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.123[a]	.015	.015	.194311

Unfortunately, we found that the adjusted R squared value is extremely low which means the performance of this model is terrible. The reason we purpose is that the relation between two weather features and traffic accidents is not linear. To prove this, we also draw 3-D scatterplot.



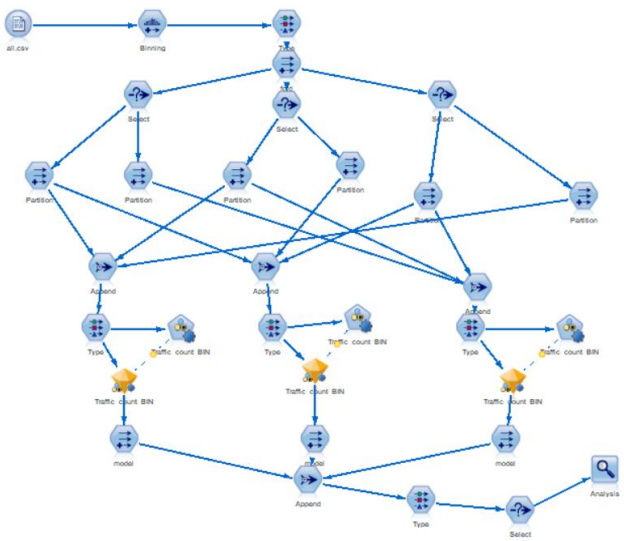
We can find that the data are not approximate located in a surface which we called regression surface. So, through the data virtualization, we enhance the evidence that the relation between two weather features and traffic accidents is not linear.

So, we try to use other model to model the relation between weather factors and traffic accidents

5 Classification

For this task, we used SPSS modeler to build an Auto Classifier node. The Auto Classifier node creates and compares several different models, allowing user to choose the best approach for a given analysis. Also, we used 3-fold cross validation to test our data.

The model we build is as follows:



After running 14 classification models, we choose XGBoost Tree model because of its high accuracy.

Results for output field Traffic_count_BIN
- Overall Results
  - Comparing $XS-Traffic_count_BIN with Traffic_count_BIN

Correct	7,443	72.96%
Wrong	2,758	27.04%
Total	10,201	

Output field Traffic_count_BIN, splitting by field model
- model = 1
  - Comparing $XS-Traffic_count_BIN with Traffic_count_BIN

Correct	2,511	74.51%
Wrong	859	25.49%
Total	3,370	

- model = 2
  - Comparing $XS-Traffic_count_BIN with Traffic_count_BIN

Correct	2,531	73.81%
Wrong	898	26.19%
Total	3,429	

- model = 3
  - Comparing $XS-Traffic_count_BIN with Traffic_count_BIN

Correct	2,401	70.58%
Wrong	1,001	29.42%
Total	3,402	

After we use 3-fold cross validation to test our

data, we get the result of model overall accuracy: **72.96%**

6 CONCLUSIONS

The analysis of traffic accident data and weather data reveals some undetectable and consistent pattern to weather-related accidents. In common sense, we usually hold the opinion that the bad weather like heavily rain or snow may lead accidents increasing. [6] Also, we are easy to believe that high wind speed will lead accidents increasing too. [4] However, after we analysis the correlation between weather factors and traffic accidents counts, we should believe that there is no significance relevant between bad weather (except extremely weather condition) and accident frequency, neither do wind speed. Sometimes, the bad weather may even lead to a low accidents rate. One of the reasons may be the reduced traffic flow. [7] In other hand, we reveal that temperature and humidity is the key weather-related factor of accidents frequency. We speculate that the reason behind may be related to driver's mood and mind.

Also, based on the analysis we make, we have built a classification model which overall accuracy reach to 72.96% to help people predict potential accidents and preventing them.

REFERENCES

[1] Edwards, J. B. (1998). The relationship between road accident severity and recorded weather. Journal of Safety Research, 29(4), 249-262.

[2] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

[3] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[4] Edwards, J. B. (1996). Weather-related road accidents in England and Wales: a spatial analysis. Journal of transport geography, 4(3), 201-212.

[5] Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., & Thomsen, L. K. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis & Prevention, 27(1), 1-20.

[6] Brodsky, H., & Hakkert, A. S. (1988). Risk of a road accident in rainy weather. Accident Analysis & Prevention, 20(3), 161-176.

[7] Codling, P. J. (1971). Thick fog and its effect on traffic flow and accidents.

[8] Keller, M. C., Fredrickson, B. L., Ybarra, O., Côté, S., Johnson, K., Mikels, J., ... & Wager, T. (2005). A warm heart and a clear head: The contingent effects of weather on mood and cognition. Psychological science, 16(9), 724-731.