TRAFFIC FLOW PREDICTION IN A U.S. METROPOLIS

JINGBAO LUO

Contents

Date: 2023-01-26.

1. Problem Definition

The source of data for forecasting traffic flows in metropolitan areas in the USA is mainly based on kaggle competitions. The data contains characteristics such as time, coordinates, direction, etc. The aim of the article is to predict the amount of congestion at a particular time.

The files and data attrubitions are described in Table 1:

Table 1. DATA

File	Description	Attribution
train.csv	traffic congestion from April through September of 1991	row˙id,time,x,y, direction,congestion
test.csv	hourly predictions on the day of 1991-09-30	row˙id,time,x,y, direction

2. Data Processing

Since the training data has both segmentable and mergeable data, we segmented the temporal data and merged the x, y and direction data.

First, we divide the 24 hours into 6 time periods, as in Table 2:

Table 2. period

period	Description	period	Description
Late Night	0:00-4:00	Noon	12:00-16:00
Early Morning	4:00-8:00	Evening	16:00-20:00
Morning	8:00-12:00	Night	20:00-24:00

Second, we split the time into the following features:

- month
- weekday
- minute
- is˙month˙start
- is˙month˙end
- is˙weekday
- is˙Monday
- is˙Friday
- period
- road:x+y+direction(00EB)

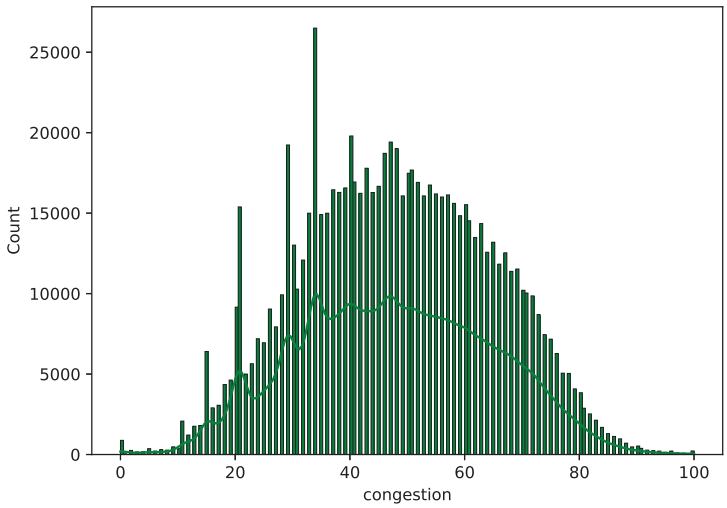

FIGURE 1. Congestion data

3. DATA DESCRIPTION

After the data has been processed, the data is explored for congestion, time, road, etc. to find more potential relationships between them.

Firstly, the histogram(figure 1) shows the congestion data, which is normalised as can be seen from the graph.

Secondly, we use various graphs to show the relationship between features and congestion.(fig.2-fig.9)

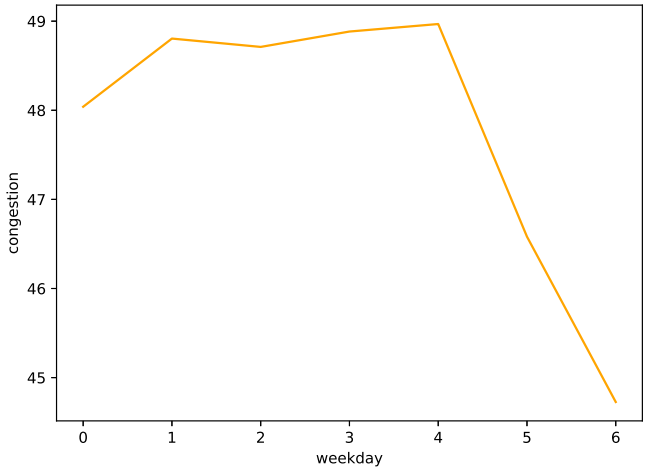

FIGURE 2. The effect of weekday on congestion

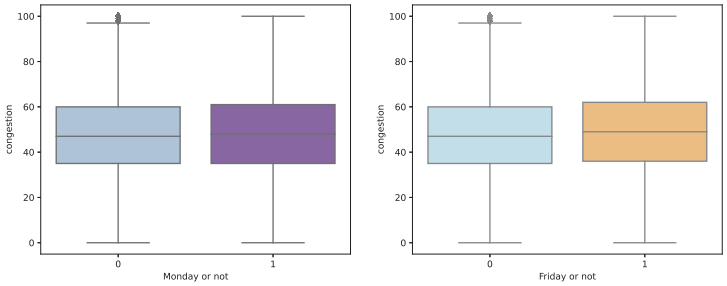

FIGURE 3. Congestion in special day or not

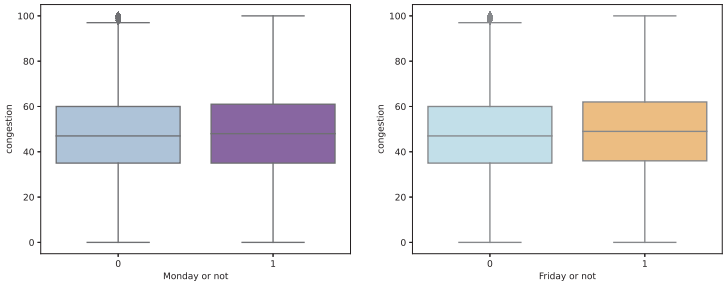

FIGURE 4. Congestion in special day or not

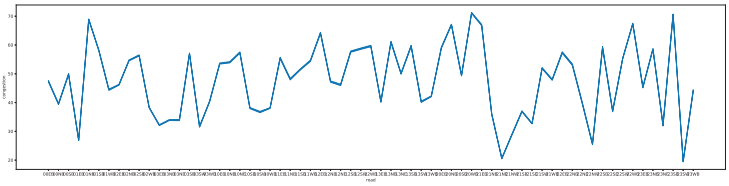

FIGURE 5. The effect of road on congestion

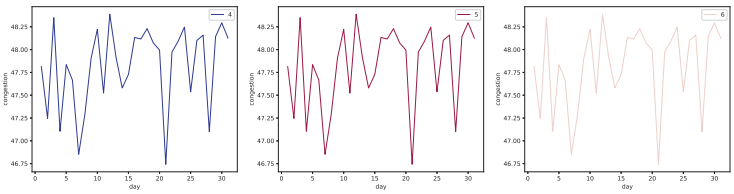

FIGURE 6. The effect of day on congestion group by month

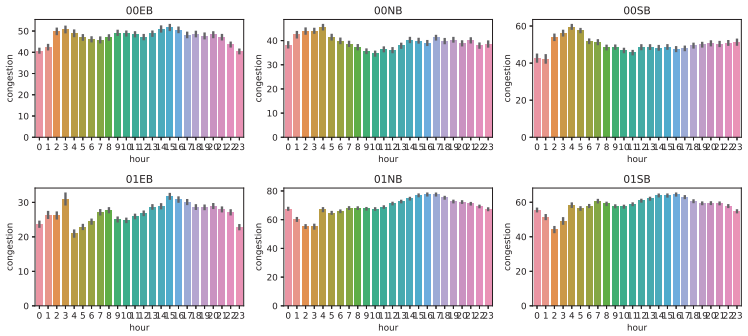

FIGURE 7. The effect of hour on congestion group by road

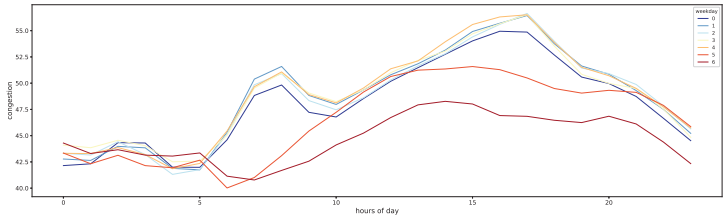

FIGURE 8. The effect of weekday on congestion group by hour

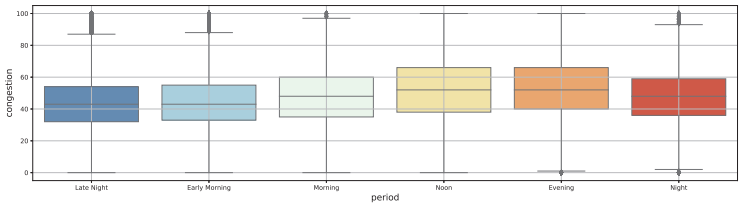

FIGURE 9. The effect of period on congestion

4. Model Train and Evaluation

Before training the model, we carried out the following steps:

(1) Divide the training and validation sets using sklearn's library.
(2) Process the non-integer data from the training set, validation set and test set

In this paper, a lightgbm model is used to train to predict congestion in the US metropolitan area with the following parameter settings:

'objective' : 'regression'
'metric': 'mae'
'learning˙rate': 0.25
'num˙iteration': 200
'num˙leaves':250
'device':'gpu'

After setting up the model parameters, we encapsulated the data into the format required by lightgbm for training.

As the training was completed, we extracted the following features.(figure 10)

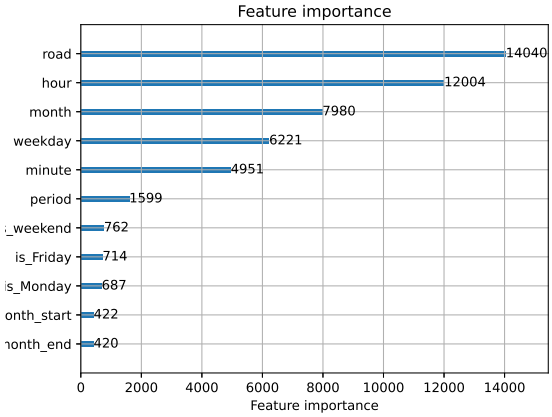

FIGURE 10. Feature Importance

In the evaluation phase of the model, we evaluated the model using the regression model evaluation metrics. The evaluation indicators are listed in the table 3 below:

TABLE 3. Model evaluation indexes

Index	Eexplication
explained˙variance˙score	Explain the variance score of the regression model.
mean˙absolute˙error	Assess the proximity of the predicted results to the real data set.
Mean squared error	Calculate the mean value of the square sum of the errors of the corresponding sample points of the fitting data and the original data
r2˙score	Judge the fitting degree of prediction model and real data

Three of these indicators were used for the assessment,the results of the assessment are as follows:

TABLE 4. The Evaluation results

Index	Result
explained˙variance˙score	0.7277243544483329
mean˙absolute˙error	6.167491947603395
r2˙score	0.7277251135484366

From Table 4, it can be seen that both explained˙variance˙score and r2˙score are 0.73, so the model is well trained without optimization.

5. Result

Based on the data from the test set, some of the test results are shown in the table 5 below:

Table 5. The prediction results

Row˙id	Congestion	Row˙id	Congestion
848835	47	848836	33
848837	39	848838	54
848839	64	848840	23
848841	28	848842	70
848843	25	848844	47
848845	46	848846	25
848847	69	848848	60

(A. 1) School of Economics and Management, Nanjing University of Science and Technology, Jiangsu 210094, China

Email address, A. 1: jluo@tulip.academy