Choose Your Own Project: Interstate Traffic Volume

edX HarvardX: PH125.9x - Data Science: Capstone

Maria Eugenia Fonseca

May 2019

Contents

1 Overview

This project is the second assignment of the 'Data Science: Capstone' course (PH125.9x), offered by edX HarvardX. The aim of this project is to use a publicly available dataset to apply machine learning techniques that go beyond standard linear regression and to clearly communicate the process and insights gained from the analysis.

1.1 Introduction

Traffic volume on a road is defined as the number of vehicles passing the measurement point per unit time. The traffic counts can be used by local councils to identify which routes are used most, and to either improve that road or provide an alternative if there is an excessive amount of traffic [1].

1.2 Project Description

The objective of this project is to use machine learning models to predict the traffic volume at an american interstate and understand what features are important to explain the transit. Data transformation and feature engineering are included to improve the predictions and, as various modeling approaches are presented, the best model will be selected based on metrics such as the RMSE.

1.3 Dataset

The present project studies the Metro Interstate Traffic Volume Dataset, available at the UCI Machine Learning Repository [2]. This dataset is composed with hourly traffic volumes for westbound Interstate 94 (I-94), including weather and holiday features from 2012 to 2018.

The I-94 is an east–west Interstate Highway connecting the Great Lakes and northern Great Plains regions of the United States. Its western terminus is in Billings, Montana and its eastern terminus is in Port Huron, Michigan [3]. The measuring point for the traffic volume is roughly midway between Minneapolis and St Paul, in the state of Minnesota, as shown in the figure bellow [4].

Figure 1: The measuring point for the traffic volume at I-94

The dataset is downloaded directly from the UCI Machine Learning Repository. The traffic data is provided by the MN Department of Transportation, while the weather data source is OpenWeatherMap. The dataset includes the following variables:

- Response variable:
  - Traffic volume: numeric hourly traffic volume
- Features:
  - Holiday: categorical US National holidays plus regional holiday
  - Temperature: average temperature in kelvin
  - Rain: amount in mm of rain that occurred in the hour
  - Snow: amount in mm of snow that occurred in the hour
  - Clouds: percentage of cloud cover
  - Weather main: short textual description of the current weather
  - Weather description: longer textual description of the current weather
  - Date time: hour of the data collected in local CST time

Data exploration and visualization, as well as transformation and feature engineering will be presented in the next section.

2 Methods and Analysis

2.1 Exploratory Data Analysis

The dataset contains 48204 hourly registers of traffic volume, weather and holiday features. The first observation is dated 2012-10-02 and the last 2018-09-30, but between August 2014 and June 2015 there are no registers.

```
## Observations: 48,204
## Variables: 9
## $ holiday             <fct> None, None, None, None, None, None, None, ...
## $ temp                <dbl> 288.28, 289.36, 289.58, 290.13, 291.14, 29...
## $ rain_1h             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ snow_1h             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ clouds_all          <int> 40, 75, 90, 90, 75, 1, 1, 1, 20, 20, 20, 1...
## $ weather_main        <fct> Clouds, Clouds, Clouds, Clouds, Clouds, Cl...
## $ weather_description <fct> scattered clouds, broken clouds, overcast ...
## $ date_time           <fct> 2012-10-02 09:00:00, 2012-10-02 10:00:00, ...
## $ traffic_volume      <int> 5545, 4516, 4767, 5026, 4918, 5181, 5584, ...
```

The dataset presents duplicated problems. 17 observations are recorded twice, and 7629 observations are duplicated per date time (the only different features are the weather descriptions). We also have some observations described as *thunderstorm*, but with 0 mm of rain that occurred in the hour. The problem occurred similarly with the snow feature.

holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume
None	275.66	0	0	90	Rain	light rain	2012-10-26 09:00:00	5234
None	275.66	0	0	90	Mist	mist	2012-10-26 09:00:00	5234
None	275.66	0	0	90	Snow	heavy snow	2012-10-26 09:00:00	5234

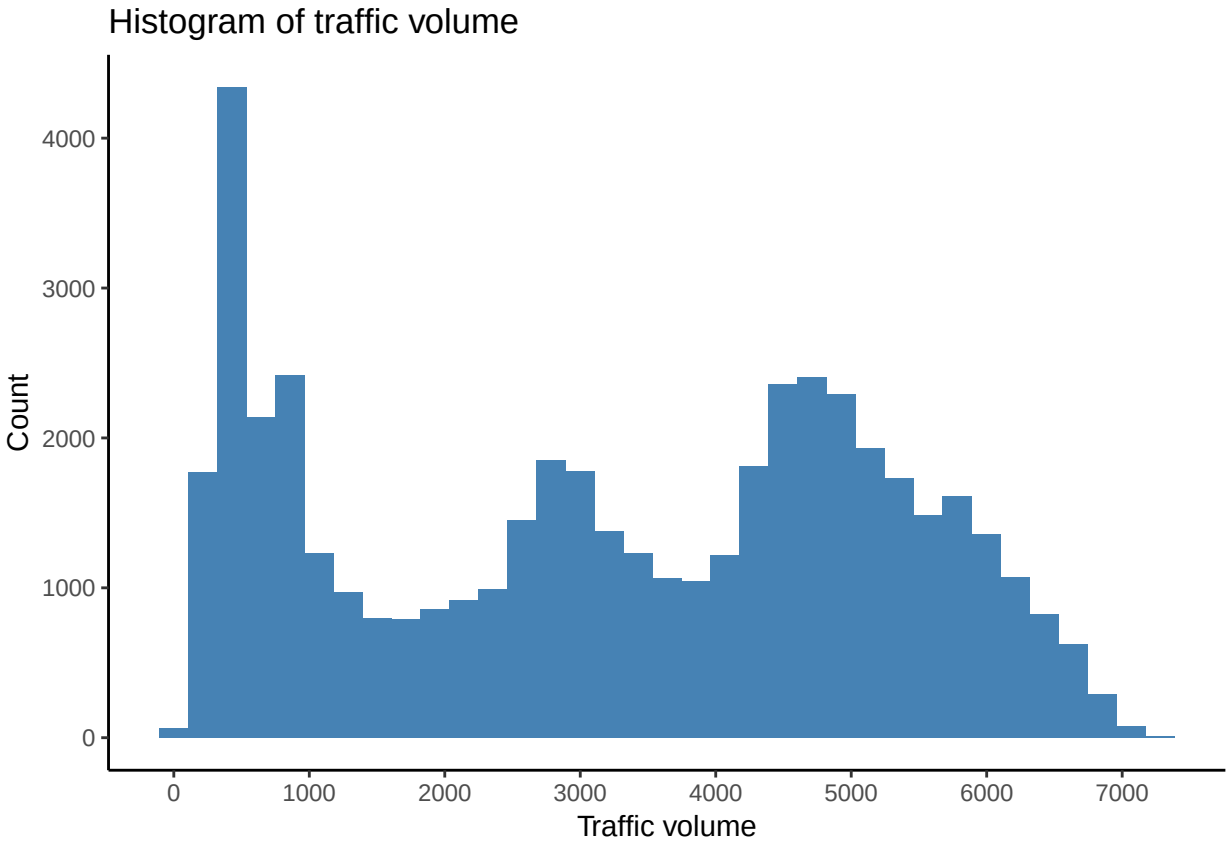
Considering the above mentioned problems, two actions were taken. First, the duplicated lines were removed. Second, for the duplicated dates with different weather condition:

TEST IF KEEPING THE WEATHER CONDITION IS WORTH IT

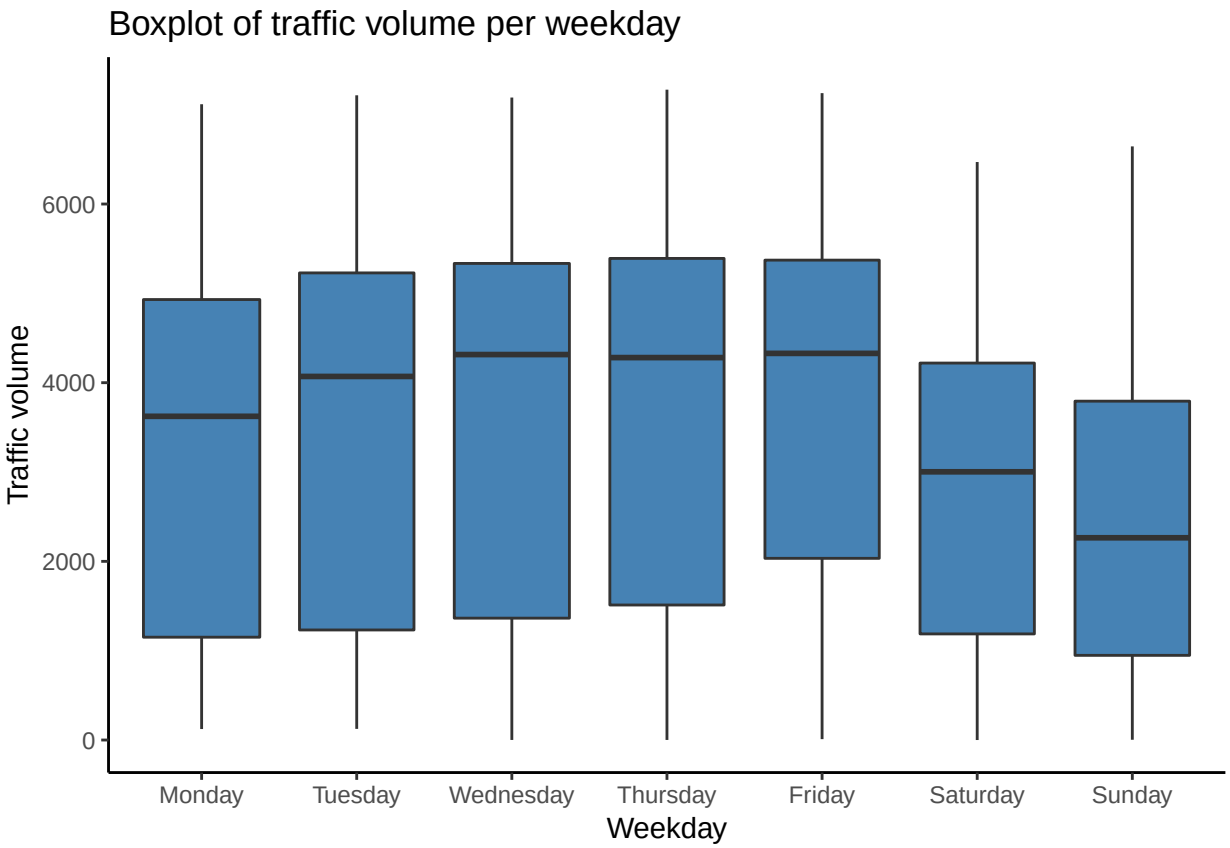
- If the world “rain”, “drizzle” or “thunderstorm” is present in the weather description features, but there is no amount in mm of rain that occurred in the hour, this will not be the weather description considered
- If the world “snow” is present in the weather description features, but there is no amount in mm of snow that occurred in the hour, this will not be the weather description considered

For example, four observations were dated *2012-10-24 06:00:00*, with the descriptions as: mist, thunderstorm with light rain, moderate rain and proximity thunderstorm. As no amout of rain in mm was recorded for the period, *mist* was the weather description considered.

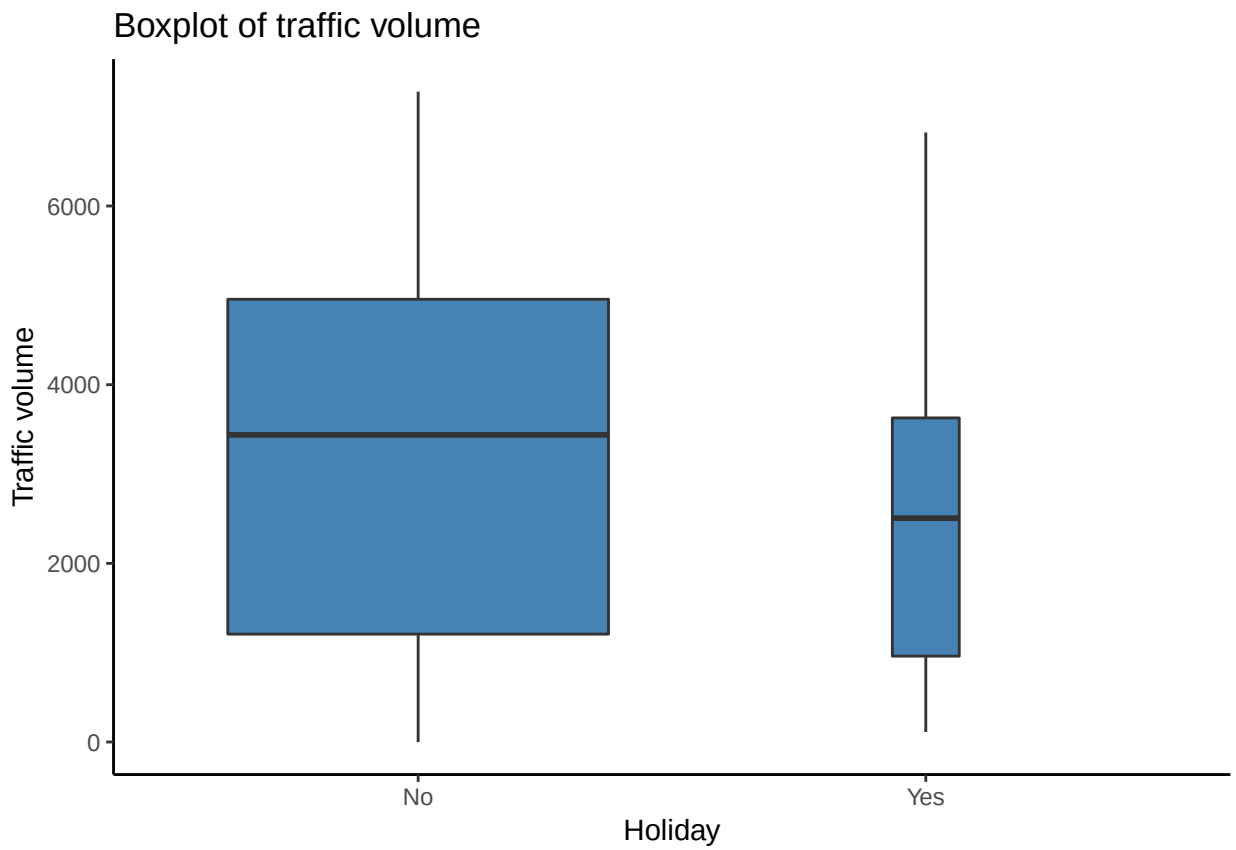
The traffic volume is a multimodal distribution with three peaks. The first peak contains the highest frequency of traffic value, below 1000 vehicles per hour. The second peak occours around 3000 vehicles/hour and the third peak, around 4500.



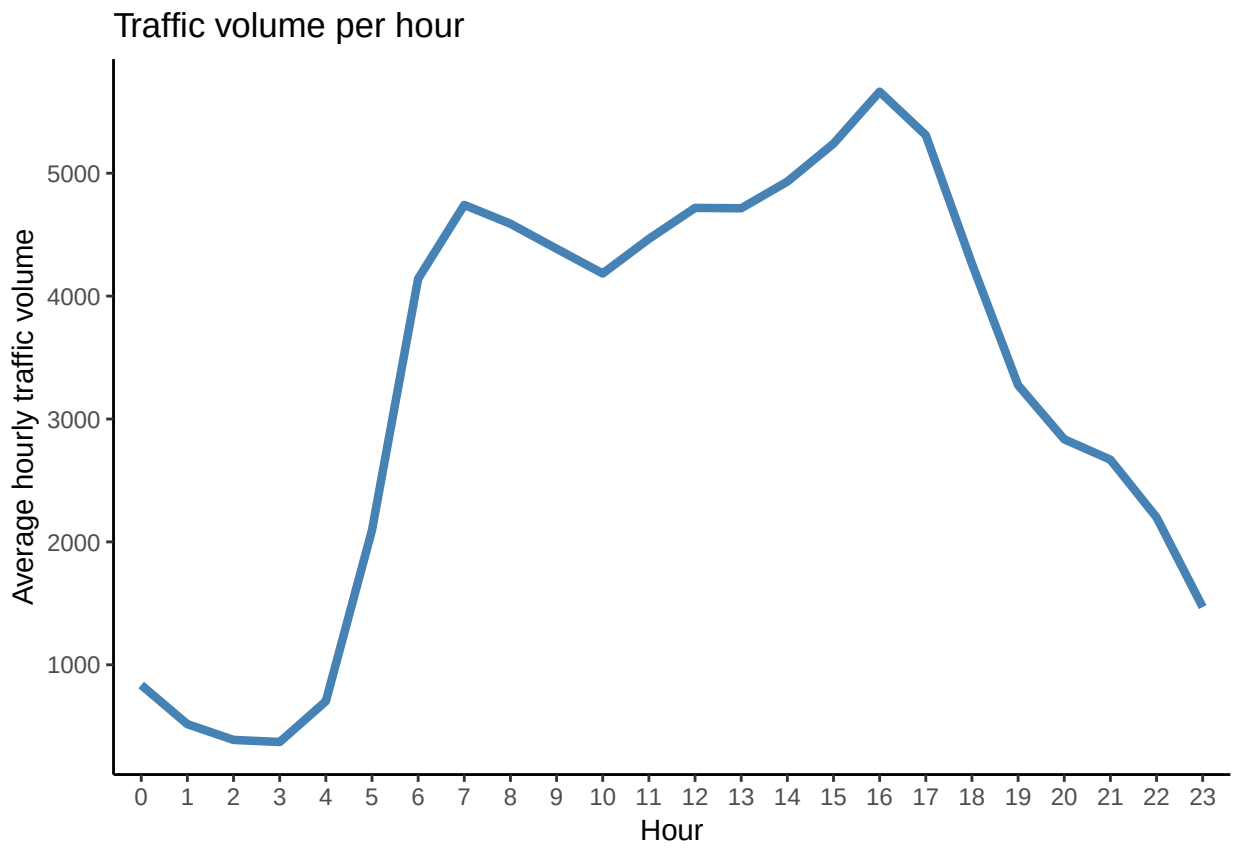
Some new features were created from the original dataset, such as the weekday. As shown in the boxplot below, the traffic volume appears to increase slowly over the weekdays and is considerably lower on weekends.



When analyzing the holiday variable, we noted that a holiday is only labeled as such on its first hour. For instance, 2012-12-25 00:00:00 is labeled *Christmas Day*, but 2012-12-25 01:00:00 and all the other following hours of the mentioned day are labeled as *none*. We corrected this in the dataset and moved from 61 days labeled as holidays to 1443.



The traffic volume varies per hour of the day, which is an indicative that it will be a good feature to the predictive model. The first big peak in traffic volume of the day is in the early in the morning, from 6 to 7 am. The traffic decreases a little in the late hours of the morning, but increases again after lunch, reaching is maximum between 4 and 5pm.



Only 63 observations have amount of snow registered. The amount of rain in mm per hour is heavily right skewed, a transformation will be considered in the following sections.

2.2 Data Transformation and Feature Engineering

2.3 Modeling Approaches

3 Results

4 Conclusion