Estimating free capacities in a transportation system

Mhyar Kousa 1, Tarcsi Ádám 2, and Dr. Tomas Horvath 3
1 MSc. Data Science, ELTE University - Faculty of Informatics
mhyarov.k@gmail.com
2 PHD. Data Science, ELTE University - Faculty of Informatics
ade@inf.elte.hu
3 Head of Data Science and Energeeing Department, ELTE University - Faculty of Informatics. 1117 Budapest, Pázmány Péter str. 1/A Nortern Building
tomas.horvath@inf.elte.hu
https://www.elte.hu/en/

Abstract. In this report,there are many statistical operations on traffic data stations in Hungary ,we use the XGBoost as a training model , Gaussian NB and Random Forest as a base line to predict the volume of traffic around the three year 2006,2007 and 2008 .

Keywords: Stations Traffic XGBoost, Random Forest, Gaussian NB.

1 Introduction

Traffic itself can be a huge challenge for most commuters regardless of the transportation method of their choice. For example, it is inevitable to experience delays and congestion during rush hours. All commute methods have their own specific characteristics when it comes to delays - cars and buses suffer from traffic jams and similar principles apply to railways as well. However, the causes of railway delays are not that straightforward and they need further investigation. According to our personal experiences most passengers are not aware of the reasons behind train delays even though they are usually encountered multiple times a day. In this theisis I will present possible answers based on the data collected from the publicly available APIs of Hungarian State Railways over the past 1.5 years. The analysis and the machine learning models could be useful for the betterment of railway services in Hungary and they may also increase the satisfaction of the passengers. Hungarian State Railways also expressed their interest in the continuation of the research project in cooperation with our university

2 Exploratory Data Analysis

2.1 About Dataset:

The files contains information about transportation stations system from 2016 to 2018 in each file there are many features like Date,Station Name,Number of passenger,Ticket Type and Quantity of tickets

2.2 Dataset Structure

Table .1. Dataset Description.

Features		Description
VAS_DATUM	Date	The values are purchase date for customers
CSATORNA	Channel	The channel where the ticket was bought, if its online ticket or other sources for tickets
ALLOMAS	Station	Station name where the ticket were bought from
VONATSZAM	Train Number	Many Null Values
JEGYKOD	Ticket No	I think its ticket number
INDALL	Start Station Code	
INDAL ALLOMAS	Start Station	Start station - From
ERKALL	Destination Station Code	Destination - To
ERK_ALLOMAS	Destination Station	
UTVONAL	Route	Route which the train will take
VAL_TAV		
DIJSZ_TAV		
DIJSZ_TAV_BB		
VAL_TAV_BB		
MENNYISEG	Quantity	#of tickets I think
SZALLITO	Supplier	
UZEMELTETO	Operated	
TAV_ROVID_BB		
UTASFO	Passengers Number	Number of passengers
JEGYTIPUS	Ticket Type1	If the ticket has a discount, if its for students etc

JEGYFAJTA	Ticket Type2	if its a season ticket, one time ticket or etc
UTAZAS FAJTA	Travel Type	if the travel is one way, two way, etc
BB CSATLAKOZÓ		
UTASKM_STAT		

2.3 Merging File Of Data Set:

Data merging is the process of combining two or more data sets into a single data set. Most often, this process is necessary because there are raw data for three different years (2016,2017 and 2018) stored in multiple files and worksheets, that in this case is important to analyze all in one .

3 Cleaning The Data:

3.1 Rename columns (translate from Hungarian to English language) as Table 1 shown.

3.2 Handle misdesignated missing values ('(-)', ' (=)',):

3.3 Convert the date column to date time:
In the next step after finish process missing values the floating point number was occured ,to handle this problem I use df_data.astype(int) and return the data to the integer.

3.4 Convert the quantity column to numeric:

3.5 Convert the passengers number column to numeric:

4 Data Analyses && Visualisation:

4.1 What was the metro station with the most traffic in Hungary?

Through statistical operations on the stations for the number of passengers we observe that Menetjegyiroda Szolnok station has the highest number of passengers as the **Table .2.** shown

Table.2. **The metro station with the most traffic**

Station	Passengers Number
Menetjegyiroda Szolnok	195
Kismaros	35
Dabas	13

4.2 What was the metro station with the least traffic in Hungary?

Through statistical operations on the stations for the number of passengers we observe that Menetjegyiroda Szolnok station has the highest number of passengersas the **Table.3.** shown

Table.3. **The metro station with the least traffic**

Station	Passengers Number
Abony automata	0
Ajka	0
Albertirsa	0
Badacsony	**0**

4.3 What was the day with the most traffic?

```
The answer is 'July 01, 2016'.
```

4.4 What was the day with the least traffic?

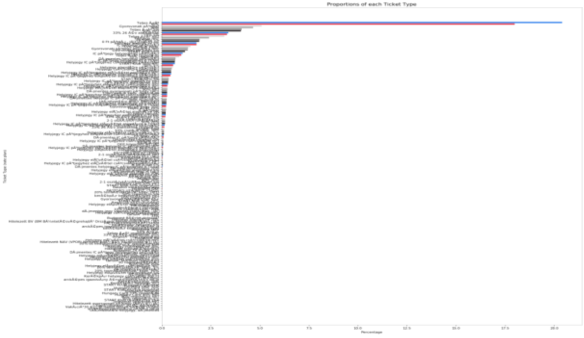
```
The answer is 'February 01, 2016'.
```

4.5 What was the most frequent ticket type?

4.5.1 Ticket Type1:

Menetjegy tickets represent more than 70% of all checkins, followed by Helyjegy with only 26% of all checkins. Other ticket types represent less than 2% of all checkins as shown in **Figure1**.

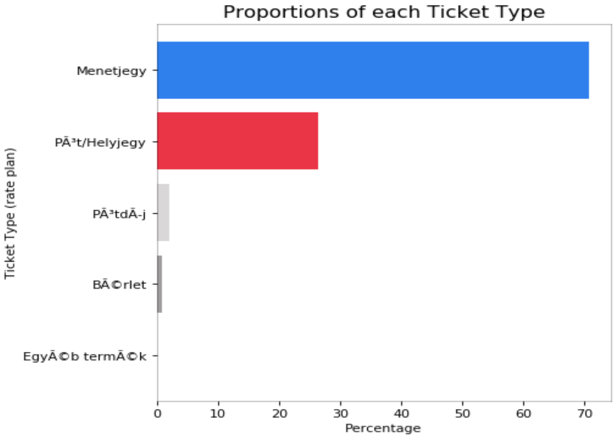
Fig .1.Ticket Type1



4.5.2 Ticket Type2:

Teljes Ã¡rÃ tickets represent more than 20% of all chickens, followed by 50% with only 18% of all checkens,followed by Other ticket types represent less than 6% of all checking as shown in **Figure2**.

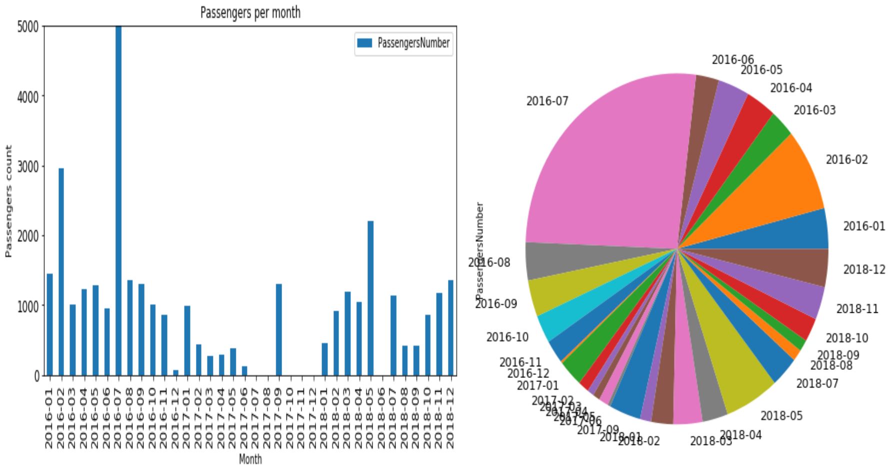
Fig.2.Ticket Type2



4.6 Number of passengers per ticket type, per month:

The **Figure3** describe the number of passengers per month the max number of passengers is in 07-2016 .

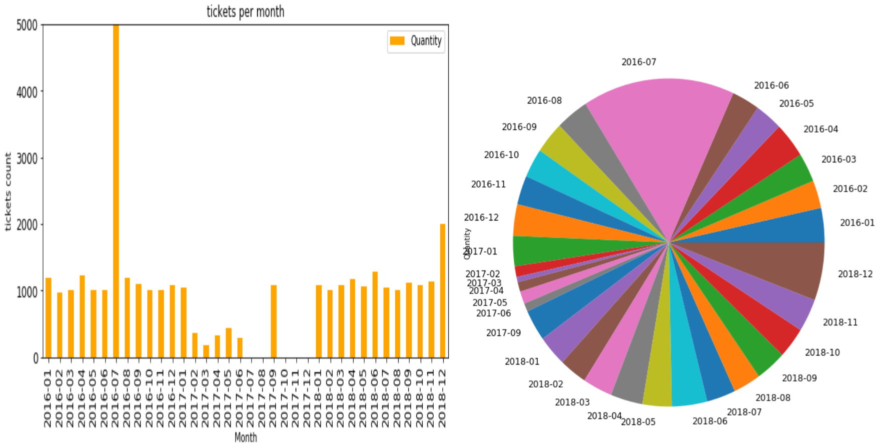
Fig.3.Number of passengers per month



4.7 Quantity of ticket , per month:

The **Figure4** describe the quantity of ticket per month the max quantity ticket is in 07-2016 .

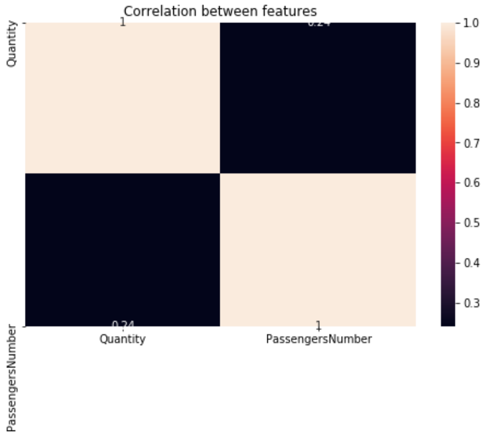
Fig.4.Quantity of ticket per month



4.8 Correlation between Features:

There is a strong correlation between number of passengers and the Quantity of ticket as shown in **Figure5** .

Fig.5.Correlation between features



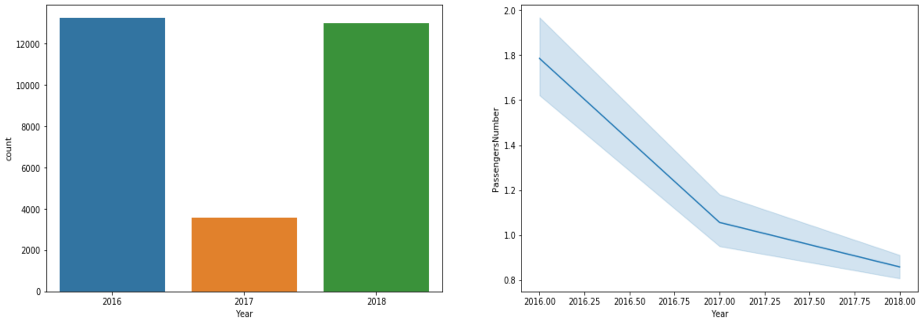
5 Preprocessing The Data

In this section there is a new shape of the dataset that contain the number of passengers and the splitting date into day , month ,year, weekend and day of week to do some important statistical operations that depend on date and time

5.1 Number of Passenger Per Year:

The Figure6 show the number of passengers around three years 2016,2017 and 2018 the maximum number is at 2016 year as shown in **Figure6**.

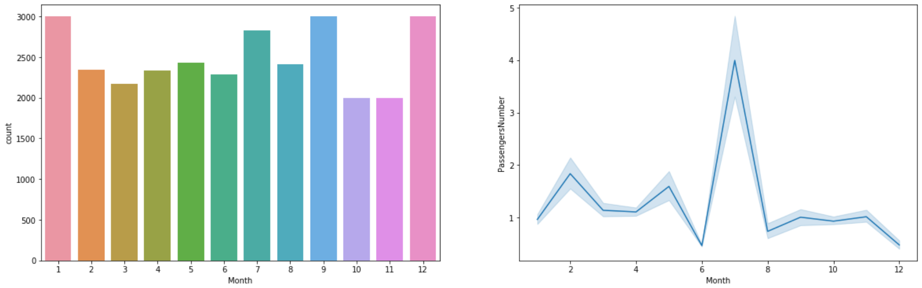
Fig.6.Number of passengers per year



5.2 Number Of Passenger Per Month:

The fig show the number of passengers around the months of year and the maximum number of passengers is at january as shown in **Figure 7**.

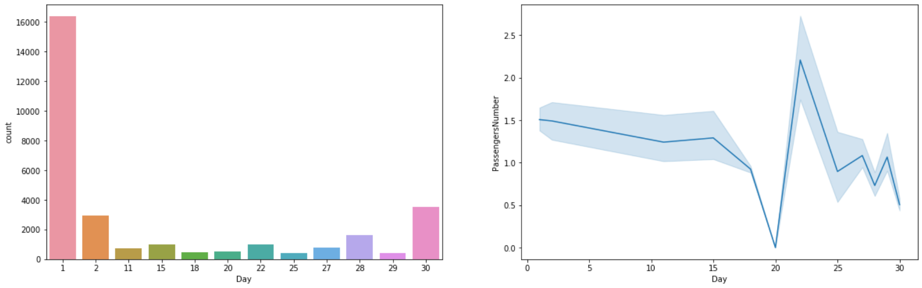
Fig.7.Number of passengers per month



5.3 Number Of Passenger Per Day Of Month:

The fig show the number of passengers around the day of month and the maximum number of passengers is in the first day of month as shown in **Figure 8.**

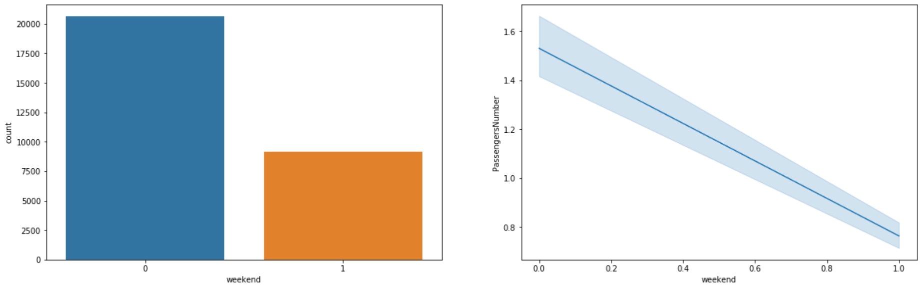
Fig.8.Number of passengers per day of month



5.4 Number Of Passenger Per Week End:

The **Figure 9** show the number of passengers between holiday and non holiday.

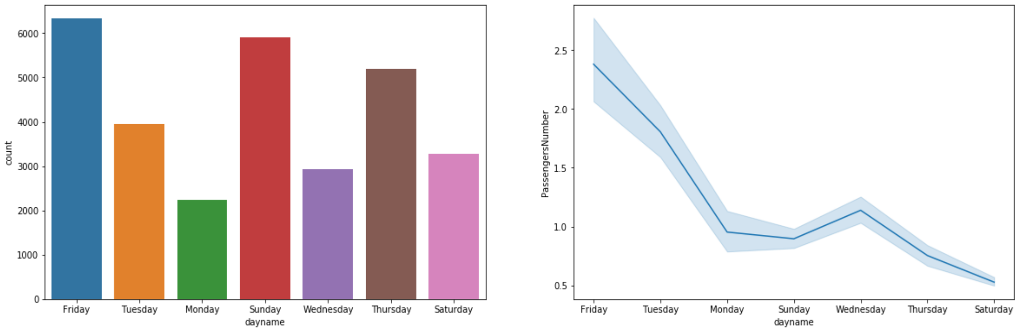
Fig.9.Number of passengers per weekend



5.5 Number Of Passenger Per Day Of Week:

The **Figure 10** show the number of passengers around the day of week and the maximum number of passengers is in the first day of month .

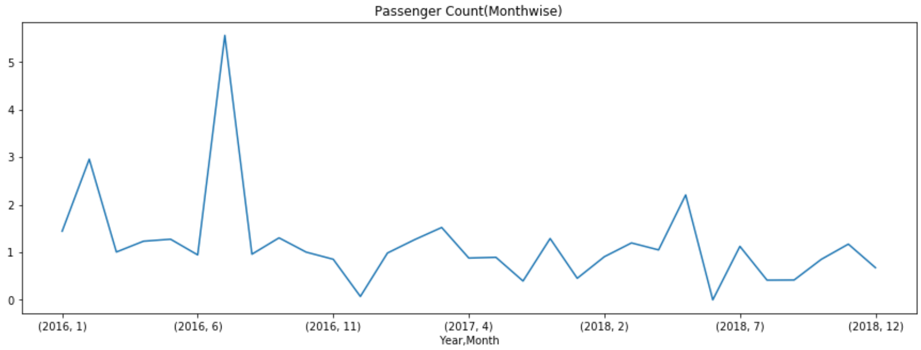
Fig.10.Number of passengers per day of week.



5.6 Year and Month wise Count:

The Figure 11 show the number of passengers wise month and years.

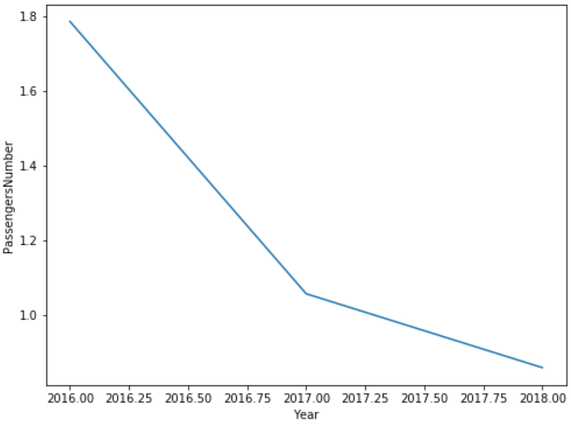
Fig.11.Passengers count month wise year



5.7 Traffic Over The Year:

The Figure 12 show the number of passengers around the three years 2016,2017 and 2018.

Fig.12.Passengers Count Years



6 Splitting dataset into train and test data

As we work with datasets, a **machine learning algorithm** works in two stages. We usually split the data around 20%-80% between testing and training stages. Under supervised learning, we split a dataset into a training data and test data in Python ML.

7 Apply Label Encoding:

Applying Label Encoding on the dataset is benefit to convert all categorical data into numerical that effect on the performance of the model that increases because when we have a numerical data is better for model to understand.

8 XGBoost

XGBoost showed the following result:

MSE = 0:21699002415640098

After tuning hyperparameters, the following result was achieved:

MSE = 0:16778714457090063

9 Baselin:

In order to compare XGBoost Model result, the following models were tried as baseline:

_ Random Forest Regressor
_ GaussianNB

Table.4.Base Line

Model	MSE
XGBoost	**0:16778714457090063**
Random Forest Regressor	**0:25699002415640098**
GaussianNB	**0:31699002415640098**

10 Conclusion:

In this paper I do statistics operational on the transpot station dataset and also visualising the data to have better understanding how the traffic act arount the three year at the stations in hungary using machine learning model to predict traffic inside the stations like Xgboost,RandomeForest and GaussianNB.

11 References

[1] Roland Krisztián Szabó. Smart alarm clock based on traffic and weather information, 2018.
[2] MÁV Szolgáltató Központ Zrt. MÁV-START térkép, 2020.

[Online; accessed 16-January-2020].