BUDAPEST SCHOLARSHIP PROGRAM SUMMARY

A PREPRINT

Do Chau Tuan
BKK - Budapesti Közlekedési Központ
tuanb2610@gmail.com

TÓTH Patrik
BKK - Budapesti Közlekedési Központ
Patrik.Toth@bkk.hu

Max von Muenster
BKK - Budapesti Közlekedési Központ
Theodor.Maximilian@bkk.hu

Galiger Gergő
BKK - Budapesti Közlekedési Központ
Gergo.Galiger@bkk.hu

Mark Andras Nyerges
BKK - Budapesti Közlekedési Központ
markandras.nyerges@bkk.hu

January 31, 2024

ABSTRACT

This report encapsulates the comprehensive research undertaken over a five-month period of the Budapest Scholarship Program. My work at BKK aimed at enhancing traffic system analysis through the integration of Artificial Intelligence (AI), Machine Learning (ML), and Mathematical methodologies. The study focused on addressing contemporary challenges in traffic management by leveraging advanced computational techniques.
During my period in BKK, I was fortunate to participate in many data science projects to solve traffic system problems in Budapest. However, there are two big projects that I mainly contribute to which are "LineString" and "Budapest Metric". Both of them are purposely to present a more efficient and optimized way to estimate the distance and the angle of a plethora of points and lines in the Budapest map.

Keywords AI · Machine Learning · BKK

1 LineString Project

The purpose of this project is to analyze the traffic congestion in Budapest and retrieve some useful information about the congestion problem in Budapest then find a clever solution to handle it.

1.1 Data Collection

First, we extracted the real-time dataset about the delayed journey from API. We scraped the dataset from API by the MySQL method in the form of Excel file. The dataset we scraped from API contained a bunch of features related to a journey which are described in Table 1.

1.2 Data Processing

As We might see the dataset contained a bunch of unnecessary and overlapping features have the same meaning like *Kerulet, Egyedi azonosito, Év, óra, Perc, Hét, napjai, Hónap, ...*
Including these unnecessary features will crucially bring many disadvantages to our future analysis work.

Features	Descriptions
Időbélyeg	Timestamp
Kerulet	District
Egyedi azonosito	Coordinate code
Földrajzi point	LineString
Átlagos időtöbblet (Perc)	Time delay in min
Átlagos torlódási hossz (Méter)	Length in m
Átlagos torlódási szint	Congestion level
Év	Year
Óra	Hour
Perc	Minute
Negyedóra	Second

Table 1: Table of features in our jam traffic dataset

- **Increased Dimensionality**: Including unnecessary features in machine learning models can lead to higher dimensionality of the dataset. This can exacerbate the curse of dimensionality, making algorithms computationally expensive and more prone to overfitting.
- **Overfitting**: Unnecessary features can introduce noise into the dataset, causing the model to fit too closely to the training data and perform poorly on unseen data. This phenomenon, known as overfitting, compromises the generalization ability of the model.
- **Reduced Model Interpretability**: Extraneous features can obscure the underlying relationships between input variables and the target variable, making it harder to interpret the model's predictions. This lack of interpretability can impede the understanding of the decision-making process, particularly in sensitive domains like healthcare or finance.

That's why we decided to eliminate all those abnormal features that play the same role of meaning like *Év, Óra, Perc, Negyedóra* or *Kerulet, Egyedi azonosito*.
After removing all unnecessary features our dataset became more elegant than before.

1.3 Mathematical Methodology

Our question on this topic is to distinguish whether the jammed traffic happens and how it interacts with the city center. We focus mainly on the direction of the delay with the center of the city by calculating the angle of the LineString with the center city. In our project, we defined the center of the city will located at Astoria station with coordinate $(19.060010, 47.494320)$

1.3.1 Cosine Similarity

In data analysis, cosine similarity is a measure of similarity between two non-zero vectors defined in an inner product space. Cosine similarity is the cosine of the angle between the vectors

$$cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

1.3.2 Define Direction Rule

After we successfully calculated the angle of all LineString with the center city we need to find out their direction and relation with the center of the city.
After some research, we deliver a function to decide the direction of the jammed path in Table 2.
The below picture 1 will demonstrate how the Jam journey relates to its direction to the center city

Angle	Direction
$-60° \leq x \leq 60°$	IN
$x \leq -130° \vee 130° \leq x$	OUT
Otherwise	PARALLEL

Table 2: Direction function

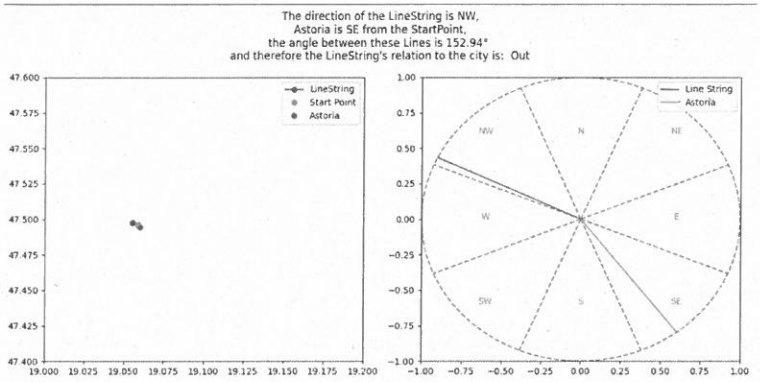

Figure 1: Visualization of the direction of the Jam Path with Astoria

2 Budapest Metric Project

This project aims to deliver a more accurate and optimized method to calculate the distance of different points(station, home, etc) in the Budapest map. Besides that, we also delivered a dashboard website to help the user have a better visualization of the Map in Budapest and the distance between the points they choose.

2.1 Metrics

There are two famous metric functions that data scientists mostly use to calculate the distance between two points the Euclidean Distance and the Manhattan Distance.

- Euclidean distance, also known as straight-line distance or L2 distance, is a measure of the straight-line distance between two points in an Euclidean space.

$$Distance = \sqrt{(x_2 - x_2)^2 + (y_2 - y_1)^2}$$

- Manhattan distance, also known as city block distance, taxicab distance, or L1 distance, is a measure of the distance between two points in a grid-based system. It is calculated as the sum of the absolute differences between the corresponding coordinates of the two points

$$Distance = |x_2 - x_1| + |y_2 - y_1|$$

2.2 Our Approach

Our goal is to calculate the distance in the drive network. That is why Euclidean performed badly in estimating the distance between two points since the car can not go straight through a mountain or building. Besides this, the distance between two points is also not symmetric due to the direction of the road.
Because of this, we have to develop a new technique to address the direction of the road and also find the shortest distance in the directed graph. Our approach divides into 3 steps:

- Project the new point to the existing edges on the graph
- We split it into 4 different situations starting from the projected point to left or right.
- We took the minimum of 4 result above to achieve the most optimal one.

2.3 Dashboard

Next, We will build an interactive dashboard to help the users have a better visualization of the Budapest Map.
First, the user has to upload the Excel file which contains the points they wish to calculate the distance between them by clicking the Upload button in the web interface.
After that, our dashboard will add the following points in the file to the map and visualize it by the Marker on the map. Furthermore, we also return the download option so the user can check the distance of all pairs of points they upload.

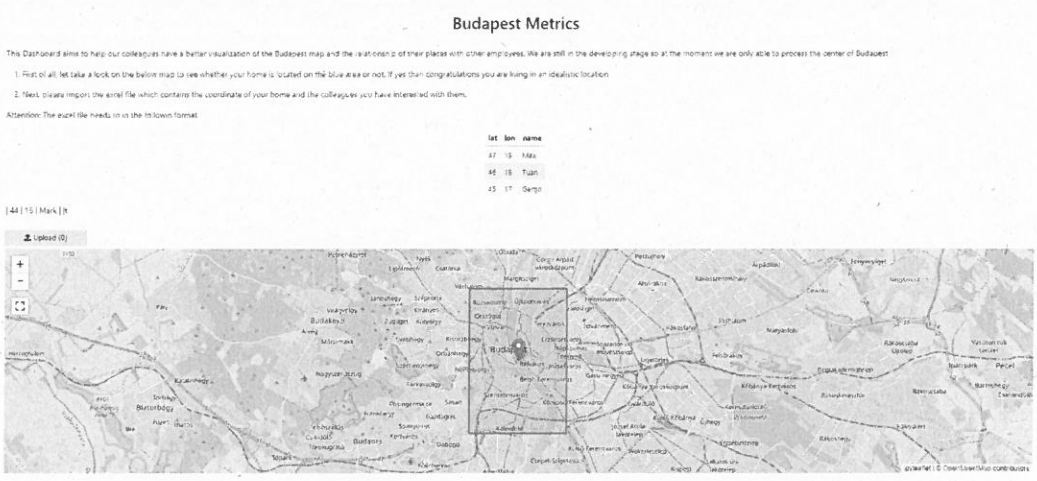

Figure 2: The original interface of our dashboard

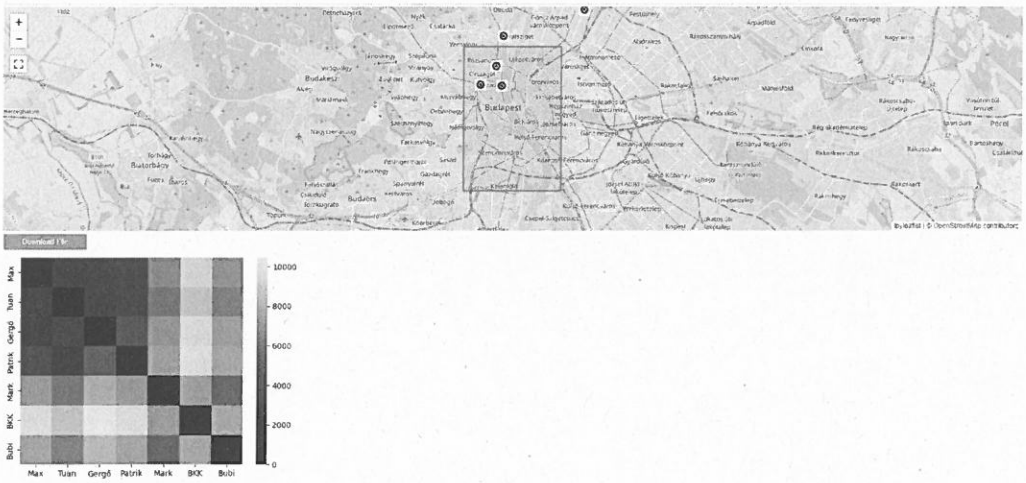

Figure 3: The interface after the user upload the excel file

3 Conclusion

First of all, I want to express my heartfelt gratitude for the scholarship awarded to me. It is with immense appreciation that I acknowledge the generosity and support extended by the Municipality. Receiving this scholarship has not only alleviated financial burdens but has also empowered me to pursue my academic and career aspirations with renewed vigor and determination.

Last but not least, my whole work in BKK could not have been done without the help of my mentor TÓTH Patrik and my data science colleagues Max, Gergő, Mark. Working under your mentorship has not only enhanced my understanding about Data Science and Traffic but has also significantly contributed to my personal and professional growth. Your insights and feedback have been invaluable, and I deeply appreciate the time and effort you have dedicated to our project.