WalkSense

Accelerating Human Mobility Prediction

(https://colab.research.google.com/drive/1OIMrpOOZbp8eBzD6_zIkrbDmh6zPLFnr?usp=sharing)

Project Report

by Arush Saxena, Intern, UST Product Engineering

Table of Contents

Business problem

Optimal Route Planning for Commuters

Description: The analysis of GPS data from the Geolife Trajectories dataset can be leveraged to solve the business problem of optimizing route planning for commuters. By predicting whether a person is using a means of transport or walking based on their GPS data, we can provide valuable insights and recommendations to commuters to help them choose the most efficient and time-saving routes for their daily travel.

Benefits and Use Cases:

1. Commute Time Optimization: By accurately identifying periods of time when a person is using a means of transport, we can analyze historical data to identify recurring traffic patterns and congestion areas. This information can be used to suggest alternative routes or departure times to minimize commute time.
2. Public Transportation Planning: By understanding the means of transport used by individuals in specific areas and at different times, transportation authorities can optimize public transportation routes and schedules. This can lead to improved efficiency, reduced congestion, and enhanced public transportation services.
3. Infrastructure Planning: Analysis of the GPS data can help identify areas with high pedestrian activity, which can aid urban planners and city officials in making informed decisions about infrastructure development. This information can be utilized to plan pedestrian-friendly zones, improve walkways, and enhance safety measures in high-footfall areas.
4. Ride-Sharing and Carpooling Services: With insights into the usage of private vehicles and means of transport, ride-sharing platforms and carpooling services can optimize their offerings. They can identify high-demand areas and peak travel times to efficiently match commuters with similar routes, leading to reduced traffic congestion, lower carbon emissions, and cost savings for users.
5. Personalized Travel Recommendations: Leveraging the predicted means of transport, personalized travel recommendations can be provided to individual commuters. This can include suggestions for alternative routes, public transportation options, or even incentives for adopting sustainable modes of travel like walking or cycling.

By incorporating the analysis of GPS data from the Geolife Trajectories dataset into route planning and transportation optimization, the business can provide valuable insights and services that enhance the commuting experience, reduce travel time, and contribute to a more sustainable and efficient transportation ecosystem.

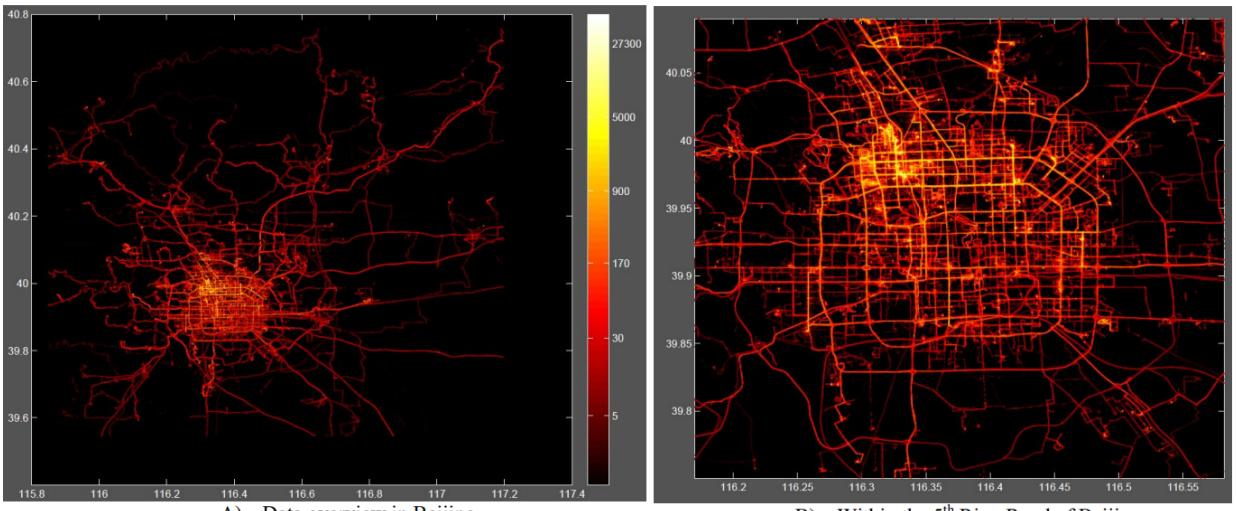
Dataset

Data Description

The dataset used for the project was Geolife Trajectories 1.3. This GPS trajectory dataset was collected in (Microsoft Research Asia) Geolife project by 182 users in a period of over five years (from April 2007 to August 2012). A GPS trajectory of this dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude and altitude. This dataset contains 17,621 trajectories with a total distance of 1,292,951 kilometers and a total duration of 50,176 hours. These trajectories were recorded by different GPS loggers and GPS phones, and have a variety of sampling rates. 91.5 percent of the trajectories are logged in a dense representation, e.g. every 1~5 seconds or every 5~10 meters per point.

This dataset recorded a broad range of users' outdoor movements, including not only life routines like go home and go to work but also some entertainments and sports activities, such as shopping, sightseeing, dining, hiking, and cycling. This trajectory dataset can be used in many research fields, such as mobility pattern mining, user activity recognition, location-based social networks, location privacy, and location recommendation.

Although this dataset is widely distributed in over 30 cities of China and even in some cities located in the USA and Europe, the majority of the data was created in Beijing, China. Figure 1 plots the distribution (heat map) of this dataset in Beijing. The figures standing on the right side of the heat bar denote the number of points generated in a location.



A) Data overview in Beijing

B) Within the 5th Ring Road of Beijing

Figure 1 Distribution of the dataset in Beijing city

The distributions of distance and duration of the trajectories are presented in Figure 2 and Figure 3.

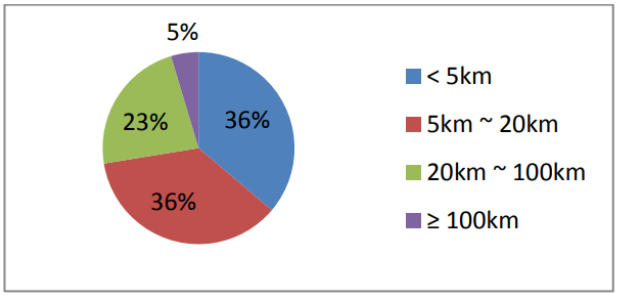


Figure 2 Distribution of trajectories by distance

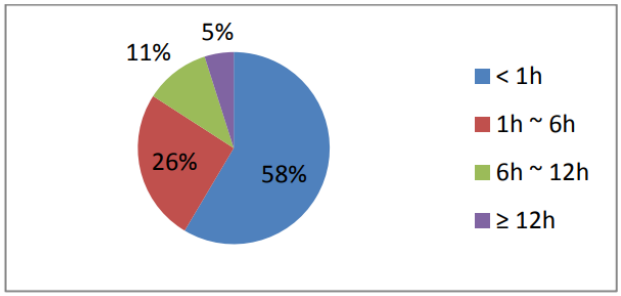


Figure 3 Distribution of trajectories by effective duration

In the data collection program, a portion of users have carried a GPS logger for years, while some of the others only have a trajectory dataset of a few weeks. This distribution is presented in Figure 4, and the distribution of the number of trajectories collected by each user is shown in Figure 5.

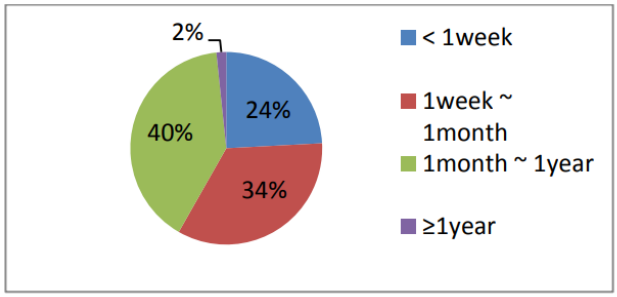


Figure 4 Distribution of users by data collection period

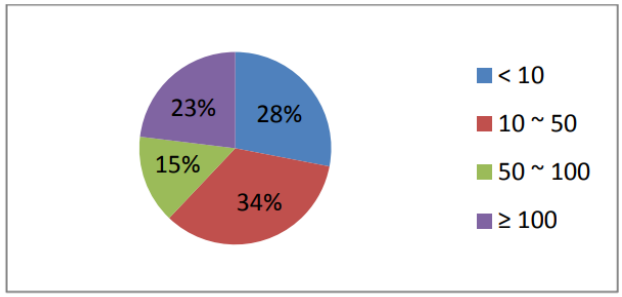


Figure 5 Distribution of users by trajectories

73 users have labeled their trajectories with transportation mode, such as driving, taking a bus, riding a bike and walking. There is a label file storing the transportation mode labels in each user's folder.

Data Format

Trajectory file

Every single folder of this dataset stores a user's GPS log files, which were converted to PLT format. Each PLT file contains a single trajectory and is named by its starting time. To avoid potential confusion of time zones, GMT is used in the date/time property of each point.

PLT format:
Line 1…6 are useless in this dataset, and can be ignored. Points are described in following lines, one for each line.
Field 1: Latitude in decimal degrees.
Field 2: Longitude in decimal degrees.
Field 3: All set to 0 for this dataset.
Field 4: Altitude in feet (-777 if not valid).
Field 5: Date - number of days (with fractional part) that have passed since 12/30/1899.
Field 6: Date as a string.
Field 7: Time as a string.

Example:
39.906631,116.385564,0,492,40097.5864583333,2009-10-11,14:04:30
39.906554,116.385625,0,492,40097.5865162037,2009-10-11,14:04:35

Transportation mode labels

Possible transportation modes are: walk, bike, bus, car, subway, train, airplane, boat, run and motorcycle.

Example:

Start Time	End Time	Transportation Mode
2007/06/26 11:32:29	2007/06/26 11:40:29	bus
2008/03/28 14:52:54	2008/03/28 15:59:59	train
2008/03/28 16:00:00	2008/03/28 22:02:00	train
2008/03/29 01:27:50	2008/03/29 15:59:59	train

Exploratory Data Analysis

Analyzing the data was a crucial step due to its unconventional format. The following steps were undertaken to gain insights from the data:

1. **Initial Data Exploration**: Basic print statements were utilized to understand the data format and explore the various features associated with it. This allowed for a preliminary understanding of the data structure.
2. **Statistical Analysis**: The .describe() function was employed to obtain key statistical information about the data, including its size and other relevant characteristics. This aided in gaining a comprehensive overview of the dataset.
3. **Heatmap Visualization**: To derive meaningful visualizations from the data, the folium library was utilized. Heatmaps were generated to depict the trajectories of two randomly chosen individuals, differentiating between their weekday and weekend activities.

   Weekday Trajectory Analysis: Figure 6 illustrates the heatmap of person 1's weekday trajectory. Notably, a prominent red region within the heatmap suggests their workplace. Assuming the individual is employed, this indicates a significant time spent at their workplace during weekdays.

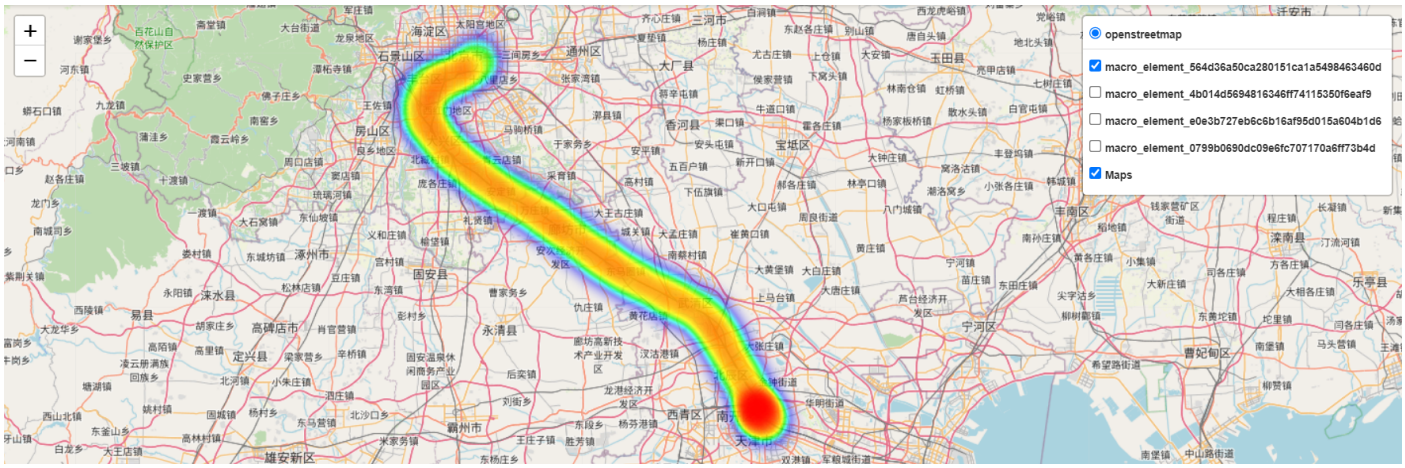


Figure 6

Weekend Trajectory Analysis: Figure 7 showcases person 1's weekend trajectory, which aligns with the starting point of their weekday trajectory (figure 6). This observation suggests that person 1 remained at home during that particular weekend.

Figure 7

Person 2's trajectory analysis yielded similar results, albeit on a smaller scale, suggesting that their workplace or school is located within their local area. The corresponding heatmaps can be observed in Figure 8 (weekday trajectory) and Figure 9 (weekend trajectory).

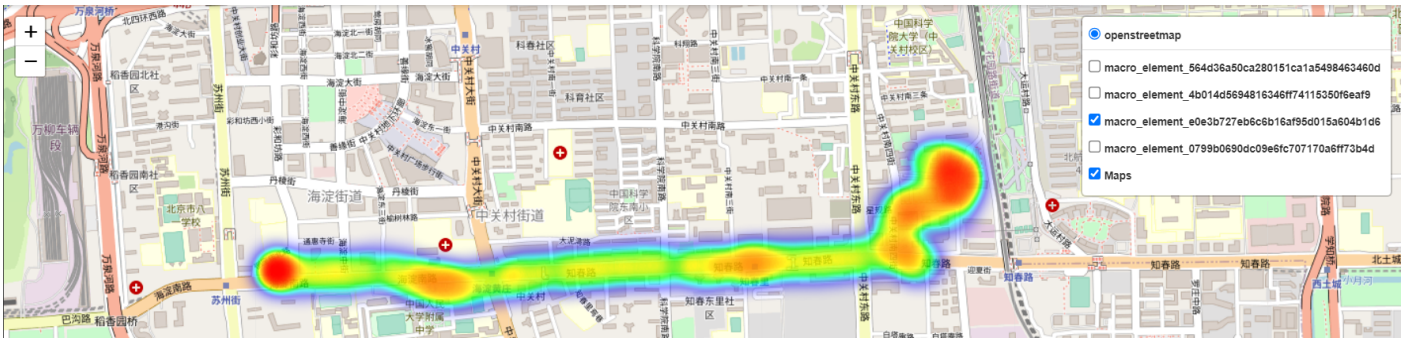


Figure 8

Figure 9

These analyses provide initial insights into the daily and weekend activities of the selected individuals, highlighting patterns and potential routines in their trajectories.

Data Cleaning

Since, the data, which was stored on the Google Drive, was not a single file but a number of folders, which contained multiple files, the data was read using a walkthrough by using the OS package. Once the data was read, some data cleaning techniques were performed on the data and further operations were performed on the refined data obtained. Following were the data cleaning steps:

1. Only those folders/users which had a labels file were chosen for further use in the project.
2. Unnecessary features from all the accessed files were removed, which included the 'Zero' and 'Days_Elapsed' features.
3. A 'Timestamp' feature was introduced, which was made by the combination of 'Date' and 'Time' features. This Timestamp feature would later be used for further data cleaning and feature engineering steps.
4. The sampling was reduced, as per the 'Timestamp' feature, to be per minute.

Feature Engineering

The feature engineering steps carried out for further modeling were:

1. **Data Cleaning**: The .plt files and labels.txt files were processed to handle any missing values or inconsistencies. Missing data points were either imputed or removed, ensuring the dataset was clean and ready for analysis.
2. **Feature Extraction**: Relevant features were extracted from the .plt files to capture important information about the mode of transport at each timestamp. This involved extracting features such as latitude, longitude, and timestamp from the .plt files.
3. **Feature Engineering**: Additional features were created or derived from the existing variables to enhance the predictive power of the model. This could include creating features like distance traveled between timestamps, speed and acceleration of travel, or time of day (extracting hour/minute information from the timestamp).
4. **Encoding Categorical Variables**: The labels.txt files were used to assign appropriate labels for the mode of transport. This involved encoding the categorical variable representing the mode of transport into numerical values that can be used for model training.
5. **Feature Scaling**: The numerical features in the dataset were scaled to a similar range to prevent any bias in the model. Techniques like standardization or normalization were used to ensure all features had a comparable impact.
6. **Feature Selection**: The most relevant features, which were distance covered, which had a significant impact on predicting the mode of transport were selected.
7. **Model Building and Evaluation**: After the feature engineering steps, machine learning models were built using the transformed and selected features. The performance of the models was evaluated using appropriate metrics, such as accuracy, precision, recall, or F1 score.

Modeling

Once the feature selection process was completed, an additional step was taken to prepare the data for modeling. This involved adding a new feature called 'Encoded_Mode' to the main data. To accomplish this, the labels.txt file in each folder was utilized, and the timestamps in the refined data were compared with those in the labels.txt file. When a match was found, the corresponding mode of transport assigned to that timestamp in the labels.txt file was assigned to the 'Encoded_Mode' feature in the main data file. The assignment condition was such that if the mode of transport was 'walk', the 'Encoded_Mode' in the data file was updated as 'walk'; otherwise, it was updated as 'transport'. If a particular timestamp in the main data file had no corresponding match in the label.txt file, indicating the absence of mode of transport labels, that timestamp's record was dropped.

Once this data preparation process was completed, the data consisted of timestamps, distance covered, speed, acceleration, mode of transport, and some other features for potential future usage.
Subsequently, the prepared data was fed into several machine learning (ML) models for training and testing. The following ML models were employed:

- Logistic Regression
- KNN
- SVM
- Bagging
- XGBoost
- XGBoost with GridSearchCV

To ensure appropriate evaluation of the time-series data, TimeSeriesSplit with 5 randomly chosen splits was utilized for training and testing the models. The highest accuracies achieved among the 5 splits were as follows:

- Logistic Regression: 0.7744286561767282
- KNN: 0.8306731828582729
- SVM: 0.8311816717986383
- Bagging: 0.8356450747196249
- XGBoost: 0.8742902341874065

Furthermore, to examine the impact of hyperparameter tuning on the accuracies, the model with the highest accuracy (XGBoost) was selected, and Grid Search was performed in conjunction with XGBoost. This experimentation was conducted using test data. The average accuracies obtained from 5 time-series splits were as follows:

- XGBoost on test data: 0.8547719509744826
- XGBoost on test data with GridSearchCV: 0.864657424151095

Consequently, it can be observed that Grid Search with XGBoost yielded higher accuracies when predicting the mode of transport based on the provided features, including distance, speed, and acceleration.

Experiments Done

Several experiments were conducted to ensure smooth execution of the code and resolve any encountered errors. The following experiments were performed:

1. **Validation of File Access**: The first experiment involved checking if the OS loops successfully accessed the files in the data folder on Google Drive. This was done by verifying that the file paths were correctly specified and that the desired files were being read.
2. **Data Reading Verification**: To ensure the data was properly read, print statements and for loops were used to inspect the data at different stages. This allowed verification of data reading, cleaning, and feature engineering steps, ensuring the data was processed correctly.
3. **Verification of Encoded Mode Assignment**: It was essential to validate that the encoded mode of transport was correctly appended to the data. This experiment involved checking if the timestamps in the labels.txt file matched the timestamps in the main data file, and verifying that the corresponding mode of transport was assigned accurately.
4. **Data Quality Assessment**: Conducted a thorough assessment of the data quality, including checking for missing values, outliers, and inconsistencies. This experiment helped identify any data issues that may affect the modeling process.
5. **Feature Importance Analysis**: Performed a feature importance analysis to determine the relative importance of each feature in predicting the mode of transport accurately. This experiment provided insights into which features have the most significant impact on the model's performance.
6. **Model Performance Comparison**: Compared the performance of the different machine learning models used in the project (Logistic Regression, KNN, SVM, Bagging, and XGBoost). Evaluated their accuracy, precision, recall, and F1 score to determine which model performs best for the given task.

Please note that the experiments conducted may vary based on the specific requirements and challenges of the project.

Conclusion

In conclusion, the analysis of GPS data from the Geolife Trajectories dataset proved to be instrumental in addressing the business problem of optimizing route planning for commuters. By leveraging the predictive capabilities of machine learning models, valuable insights and recommendations can be provided to commuters, enabling them to make informed decisions and choose the most efficient and time-saving routes for their daily travel.

Through the initial analysis of the dataset and exploration of potential use cases, several benefits and possibilities for application emerged. These include the potential for commute time optimization by identifying recurring traffic patterns and congestion areas, offering alternative routes or departure times to minimize travel time. Additionally, there is the potential for improved public transportation planning by gaining a better understanding of individual means of transport in specific areas and at different times, resulting in optimized routes and schedules. Furthermore, the identification of areas with high pedestrian activity could inform infrastructure planning and the development of pedestrian-friendly zones. Another potential application is the optimization of ride-sharing and carpooling services by leveraging insights into private vehicle usage and different means of transport, leading to reduced traffic congestion, lower carbon emissions, and potential cost savings for users. Finally, personalized travel recommendations promoting sustainable modes of travel like walking or cycling could be provided to individual commuters. It is important to note that while these possibilities represent potential benefits, they were not directly explored or implemented in the project. They serve to highlight the broader impact and potential applications of the project's findings and insights.

The project made use of the Geolife Trajectories dataset, which recorded outdoor movements and provided a rich source of data for research in various fields, including mobility pattern mining, user activity recognition, location-based social networks, location privacy, and location recommendation.

The project encompassed different stages, including exploratory data analysis, data cleaning, feature engineering, and model building. By incorporating various machine learning algorithms such as Logistic Regression, KNN, SVM, Bagging, and XGBoost, accurate predictions of the mode of transport were achieved. Evaluation metrics, including accuracy, precision, recall, and F1 score, were utilized to assess the performance of the models. The XGBoost model consistently exhibited high accuracy in predicting the mode of transport.

In summary, this project successfully addressed the business problem of optimal route planning for commuters through the analysis of GPS data. The findings and methodologies presented in this report lay the foundation for future research and development in the field of transportation optimization, contributing to a more efficient, sustainable, and user-centric commuting experience.

Future Steps

Here are several future steps that can be incorporated into this project moving forward:

- **Multiclass Classification**: Expand the classification task to handle multiple classes or transportation modes, allowing for more granular predictions beyond just "walk" and "transport." This will provide a more comprehensive understanding of commuters' transportation choices.
- **Feature Engineering**: Further explore and engineer additional features that can capture relevant information and enhance the predictive power of the models. This could involve extracting more insights from the GPS data, such as location-based features, time-based patterns, or aggregating data at different temporal resolutions. These enriched features will provide more context and improve the accuracy of the predictions.
- **Traditional and Deep Learning (LSTM)**: Extend the modeling approach by incorporating both traditional machine learning algorithms and deep learning techniques like Long Short-Term Memory (LSTM) networks. Traditional algorithms such as Random Forest, Gradient Boosting, or Support Vector Machines can provide alternative modeling perspectives, while LSTM can capture temporal dependencies in the data. This will allow for a comprehensive evaluation of various modeling approaches.
- **Notebook to .py Format**: Convert the project code from Jupyter notebooks to a modular Python script format (.py) for improved organization, reproducibility, and seamless integration into a larger codebase. This will facilitate code maintenance, collaboration, and deployment readiness.
- **Wandb**: Integrate the use of Weights & Biases (wandb), a powerful experiment tracking and visualization tool. Wandb will enable the monitoring and comparison of different models, hyperparameters, and experiments, providing valuable insights and facilitating informed decision-making throughout the project.
- **DVC (Data Version Control)**: Utilize Data Version Control (DVC) to effectively manage and version control large datasets. DVC will enable the tracking of changes to data files, collaboration among team members, and ensure the reproducibility of the analysis. This will enhance the project's transparency and facilitate efficient data management.
- **Production (MLOps)**: Implement MLOps (Machine Learning Operations) practices to establish a robust and scalable production pipeline. This includes automating the training, testing, and deployment processes, monitoring model performance, and ensuring reproducibility. MLOps will enable seamless integration of the models into a production environment and facilitate efficient model maintenance.
- **UI (Streamlit, Gradio)**: Develop a user interface using tools like Streamlit or Gradio to create interactive dashboards or web applications. This will provide an intuitive and user-friendly interface for users to interact with the models and obtain predictions based on their input data. The UI will enhance the accessibility and usability of the models, making them more valuable to end-users.
- **Ensemble**: Explore ensemble modeling techniques to combine the predictions from multiple models. By leveraging techniques like model averaging, stacking, or boosting, ensemble models can improve overall performance and robustness. This will provide a

more accurate and reliable prediction mechanism, enhancing the effectiveness of the models.

By incorporating these future steps into the project, the aim is to enhance the classification accuracy, explore different modeling approaches, improve code organization, enable experiment tracking, ensure data version control, streamline production deployment, provide a user-friendly interface, and potentially leverage the benefits of ensemble modeling. These steps will contribute to the overall success and impact of the project, enabling the delivery of valuable insights and recommendations for optimizing route planning and transportation efficiency for commuters.

References

- Geolife Trajectories Dataset: https://www.microsoft.com/en-us/download/details.aspx?id=52367
- Scikit-learn Documentation: https://scikit-learn.org/stable/documentation.html
- Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W. Y. (2008, November). Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems* (pp. 1-10).
- Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008, April). Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web* (pp. 247-256).
- Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A. L. (2011, August). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1100-1108).