UPF – Gestió de projectes

GP2425 Challenge Final Report

EPEAK:
Mobility Prediction Application

Document Control Information

Settings	Value
Document Title:	Challenge Final Report
Project Title:	EPEAK
Team Number:	102.G
Project Owner:	Elena Díaz
Project Manager:	Telmo Linacisoro
Project Core Team:	Paula Ceprián, Judit Viladecans, Eric Berenguer, Elena Barrio, Telmo Linacisoro

Revision	Date	Created by	Short Description of Changes
v0.0	30/11/2024	Judit	Creation of Template and Development of Section 1
v0.1	01/12/2024	Elena, Judit, Paula	Completion of Section 2, 3, 4, 5
v0.2	01/12/2024	Eric, Telmo	Revision and completion of Section 2

The latest version of this controlled document is stored in Google Drive\GP2024-25\Students' deliverables\3. Executing - Final Iteration

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Executive Summary

This project uses Telefónica's extensive mobility data to optimize public transportation services in Spain. By analyzing high-traffic regions, we identified key factors such as weather conditions, special events, and recurring traffic patterns that significantly influence mobility patterns.

We developed a predictive model that incorporates this data with meteorological information, holiday calendars, and day-of-week data to accurately forecast future mobility demand. This model enables us to anticipate periods of peak and off-peak demand, allowing transportation companies to strategically adjust their fleet sizes and service schedules.

By optimizing resource allocation and service offerings, we aim to improve public transportation efficiency, reduce congestion, and promote sustainable mobility. In the end, this project will enhance the passenger experience and contribute to a more sustainable transportation landscape in Spain.

1.2. Project Background

The growing need for efficient and sustainable transportation systems in Spain inspired this project. Traditional methods of planning and managing public transportation often struggle to adapt to dynamic changes in mobility demand, resulting in overcrowded vehicles, underutilized resources, and a suboptimal passenger experience.

By utilizing advanced data analytics and predictive modeling techniques, this project aims to address these challenges. By forecasting future demand, transportation operators can optimize their service offerings, ensuring sufficient capacity during peak periods and avoiding unnecessary resource allocation during low-demand times. This approach not only enhances the overall efficiency of public transportation but also contributes to a more sustainable and environmentally friendly transportation landscape.

The collaboration with Telefónica, with its vast amount of mobility data, provided a unique opportunity to develop a data-driven solution that can significantly impact the way public transportation is planned and operated in Spain.

2. Key Deliverables

Work Breakdown Structure	
1.0	**EPEAK**
1.1	**Initiating Phase**
1.1.1	Project Charter (ID: 01)
1.1.1.1	First PCT meeting
1.1.1.2	Video Pitch recording
1.2	**Planning Phase**
1.2.1	Project Work Plan (ID:02)
1.2.1.1	PCT meeting for work breakdown structure
1.2.1.2	Requirements identification
1.2.2	Risk Management Plan (ID: 03)
1.2.3	Project Stakeholders Matrix (ID: 04)
1.3	**Execution Phase**
1.3.1	Exploration Data Analysis (ID: 05)
1.3.1.1	Kickoff session and first mentoring
1.3.1.2	Source code for Initial Analysis
1.3.2	Processed Dataset (ID: 06)
1.3.2.1	Data integration
1.3.2.2	Data filtering
1.3.3	Streamlit Web App for Data Visualization and Prediction (ID: 07)
1.3.4	Machine Learning Prediction Model (ID: 08)
1.3.4.1	Code Testing and Improvement
1.3.5	Transportation Management Proposal (ID: 09)
1.3.5.1	PCT meeting
1.3.6	Project Status Report (Sprint 1 & 2) (ID: 10)
1.3.7	Challenge Final Report (ID: 11)
1.4	**Closing Phase**
1.4.1	Final Presentation (ID: 12)
1.4.2	Project End Report (ID: 13)

2.1. Data Exploration

Initially, we worked with Telefónica's mobility dataset, which provided information on the number of trips and travelers for each journey (origin-destination pair) and date. To analyze travel trends and identify traffic peaks, we developed the notebook ExploratoryDataAnalysis.v1.0.ipynb, which offered an initial exploration of mobility fluctuations.

Our first step was to construct a unified DataFrame containing mobility data from three individual files, each corresponding to a specific year. This allowed us to perform exploratory analysis by visualizing the number of trips and travelers across regions and over different time periods (yearly, monthly, and weekly). Additionally, we identified the most frequently traveled connections. However, these were excluded from the scope of our project as we focused only on the seven provinces with the largest populations.

After identifying key mobility patterns, we conducted a simple forecasting, which we used as the foundation for the development of our predictive model.

2.2. Processed Dataset

We have created a notebook called DataFilteringIntegration.v1.0.ipynb, which focuses on cleaning and integrating various datasets essential for the project.

We began by loading the data provided by Telefónica, including mobility, weather, and holiday data. Next, we selected the top 7 provinces in Spain by applying a population threshold of 1.5 million inhabitants.

The datasets were then merged into a single dataframe, with modifications and additions to facilitate its use in the predictive model, such as encoding holiday and weather information into binary (True/False) values and adding the weekday as a feature. Additionally, we created a dataset of the most relevant recurring events, such as congresses, festivals, and fairs celebrated annually on approximately the same dates, using a threshold of 700,000 attendees to determine relevance. This events dataset was also merged with the main dataframe and encoded into binary format.

Finally, the notebook outputs a unified and clean dataset that combines all relevant variables, ready to be used in the predictive modeling phase.

2.3. Streamlit Web App for Data Visualization and Prediction

We have created a web application in which we include some visual analysis that helps understand the before and after of applying our model to the given data, so we can explain the key factors that influence it, or oppositely, that don't.

To **run the app**, you must enter the directory where *epeak.py* is saved, that is, open the folder we submitted named *streamlit,* and run on your coding application (such as Visual Studio Code) terminal the command: ***streamlit run epeak.py***

This Streamlit code contains some different files, each one with different approaches and functionalities:

- ***epeak.py***
  Is the main page, the first one displaying. This main page of the app is configured with a page title, "EPEAK," to make it clear and branded. Our full dataset, *processed_data.csv*, is

loaded so it serves as the basis for analysis and modeling, and is also a tool of visualization for Telefonicas' team and for Upf professors of our data.

- ***Pages/Data Exploration.py***

  In this part of the app we want to get the users a clear understanding of the data so they understand some key insights. We empower transport planners to make data-driven decisions, optimize resource allocation, and improve public transport services:

  - ➢ Total Travellers per Province (Origin): bar plot to display the total number of travelers from each province (as an origin). By grouping the data by the provincia_origen_name column and aggregating the total travelers, we ensure that users can quickly see where the most travel activity originates.
  - ➢ Daily Trips Over Time: We use a line plot to show the total number of trips over time, aggregated by day. If necessary, we adjust the date format to ensure the data is displayed accurately. This visualization helps us and our users spot patterns in daily trips, such as increases during specific periods or abrupt changes.
  - ➢ Top Travel Connections Between Provinces: heatmap for the travel intensity between provinces. By grouping the data by provincia_origen_name and provincia_destino_name and aggregating it, we show the strength of connections using color intensity. Lighter shades indicate fewer trips, while darker shades highlight more trips. The user can quickly see that the most popular trajectory is between Alicante and Murcia.
  - ➢ Average Peak Detection: we let users select a specific month, origin, and destination to explore weekly and monthly trip patterns. This section is designed to help us identify both peak and trough periods for mobility of a specific trajectory in a more clear and interactive way.
    - Weekly Peaks and Troughs: We calculate the day-wise averages for the selected origin, destination, and month and plot the data. Using find_peaks with adjustable prominence, we identify peaks (high traffic) and troughs (low traffic). Peaks are marked in red, and troughs are marked in blue on the plot, allowing us to clearly see the busiest and least busy days of the week.
    - Monthly Peaks and Troughs: Similarly, we calculate the monthly averages for the selected origin-destination pair and visualize them. Peaks and troughs in monthly traffic are identified and marked, helping us uncover seasonal trends, such as busier months with higher traffic.

- ***Pages/Prediction.py*:**

  We aim to provide accurate traffic peak predictions by combining a regression model and the Prophet time series forecasting model. By incorporating user inputs and contextual knowledge (like holidays and events), our model adjusts its predictions to reflect real-world variability more accurately.

  We start by loading the .pkl file, which contains:

  1. The Trained Models (*modDict*): These are specific to origin-destination pairs and have been trained on historical data to forecast travel peaks.
  2. Processed Dataframes (*database*): These include the data used for predictions and adjustments.

  We use Streamlit's widgets to collect necessary inputs from the user:

  1. Basic Inputs: Users specify the travel date, origin province, and destination province.
  2. Contextual Adjustments: Users can choose whether holidays and events should be considered. If enabled, they can rate the significance (relevance) of these factors using sliders.

At the core of our prediction logic is the `predict_and_apriori_knowledge` function. This function:

1. Retrieves Raw Predictions: It extracts Prophet-based predictions for the specified date.
2. Applies Trend Adjustments: Using a Linear Regression model, it adjusts the prediction by de-trending the data and compensating for any shifts in patterns.
3. Incorporates Contextual Corrections:
    - Weekday effects.
    - Event or holiday influences at both origin and destination provinces.
    - Each correction is derived statistically from historical data based on user-defined parameters.

These adjustments are combined into a final correction factor, which is applied to the raw prediction to produce the final traffic forecast. This interactive approach ensures that the predictions are personalized and account for user-perceived importance of external factors.

Once the user fills out the required inputs and presses the "Predict Traffic Peaks" button:

1. The prediction function runs for the specified origin-destination pair on the chosen date.
2. The final prediction, representing the expected number of travels, is displayed, giving users actionable insights.

This application bridges the gap between predictive modeling and user interaction, enabling transport planners or stakeholders to:

- ➢ Anticipate peak travel times with greater accuracy.
- ➢ Adjust resources dynamically based on expected demand.
- ➢ Make informed decisions to optimize traffic and public transport management.

By combining historical patterns, machine learning, and contextual knowledge, we ensure that our predictions are both data-driven and contextually relevant.

2.4. Machine Learning Prediction Model

We decided to train a separate model for each pair of cities. For example, one model is trained for Madrid-to-Barcelona, another for Alicante-to-Murcia, and so on. This approach allows us to achieve better predictions for each individual pair of cities. Otherwise, the variations in mobility volume and the unique characteristics of each type of travel would not be effectively captured by the model.

The model is divided into two parts:

1. The first part of the model focuses on general forecasting, using Prophet developed by Meta implemented in Stan. We decided to use this model due to its robustness and fine-tuning possibilities. We found that a multiplicative weekly and yearly seasonality along with a changepoint prior scale worked the best for our dataset.
2. Once we have an array of trained forecasting models for each journey, we add prior knowledge about festivities, the day of the week and events using linear regression on these variables. For this purpose, we differentiated between various types of festivities, days of the week and events, considering their percentile rank relative to others. For example, the initial forecasting model does not really capture the importance of

Christmas, as it is encoded as a holiday just like National Labor Day. However, the mobility on Christmas is much higher than on Labor Day, for obvious reasons. By integrating this prior knowledge, we can improve the model's predictions, making them more context-aware.

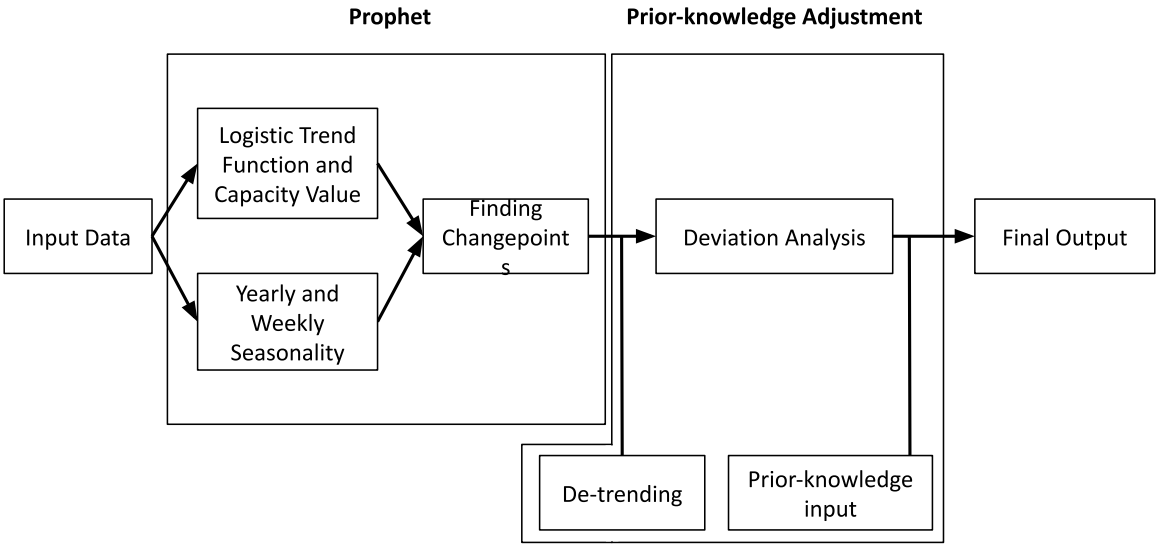

Fig. I Model Architecture

In the figure above, we can see the architecture of the model with the two parts that compose it. Bear in mind that de-trending is applied to the output of the Prophet model. This is done to remove long-term trends and isolate short-term variations, making the adjustments more precise based on recent data and contextual knowledge. Once this deviation analysis is done we add the prior-knowledge that has been introduced by the user as described above.

3. Foundations and Assumptions

3.1. Methodology and Justification

The methodology for this project was carefully chosen to match the main goal of predicting mobility demand between provinces in Spain and providing useful insights for transportation management. The approach combined data exploration, machine learning, and visualization techniques to ensure accuracy, ease of use, and clear decision-making. Below, we explain why we selected specific deliverables and tools to achieve these goals.

3.1.1. Deliverables

- **Data Exploration Notebook:** The development of a data exploration notebook was critical for understanding the Telefonica dataset. This step allowed us to identify mobility patterns and peaks, which informed the subsequent stages of data preprocessing and modeling. Exploring the dataset also helped identify relevant features, such as dates, holidays, weather, and events, that influence mobility demand.
- **Processed Dataset:** Filtering and integrating the data ensured that our analysis focused on significant mobility peaks and targeted the largest regions, where demand prediction is most impactful. This step also allowed us to add additional features that might be key to mobility trends, creating a robust foundation for the machine learning model.
- **Streamlit Web App:** A Streamlit web application was developed to visualize and present the project results interactively. This tool helped make the analysis and predictions accessible to stakeholders, enabling them to understand the impact of applying predictive models to optimize transportation resources.
- **Machine Learning Prediction Model:** A machine learning model was developed to predict future mobility peaks based on historical data. Prophet, a time series forecasting library, was used for its ability to handle seasonality and holidays, while linear regression complemented this by identifying and adjusting trends in the data. Together, these techniques provided a robust framework for forecasting mobility demand.

3.1.2. Tools and Techniques

- **Python:** Python was chosen as the primary programming language due to its versatility and extensive libraries for data analysis and machine learning. Its ease of use and integration with tools like Streamlit and Prophet made it ideal for this project.
- **Google Colab:** Google Colab provided an efficient, collaborative, and cloud-based environment for developing and testing code. It enabled access to necessary computing resources, as well as the ability to quickly share progress among team members.
- **Prophet:** Prophet, developed by Meta, was selected for its time series forecasting capabilities, particularly its ability to handle holidays, seasonality, and irregular data. This made it well-suited for predicting mobility demand, which is influenced by these factors.
- **Linear Regression:** Linear regression was used to detrend the data and identify underlying trends in mobility demand. This complemented Prophet by providing insights into long-term patterns and deviations from expected behavior.
- **Streamlit:** Streamlit was chosen for its simplicity in creating interactive web applications for data visualization. It allowed us to present the results of the analysis and predictions in a way that stakeholders could easily understand.

3.2. Assumptions

For the successful execution of this project, several key assumptions have been made regarding the resources and methodology to achieve our objectives:

- **Continuous Data Availability:** The mobility data from Telefónica will be accessible throughout the project without interruptions. This assumption includes both the availability of historical data, as any disruptions could impact our ability to perform timely analyses and develop the predictive model accurately.
- **Technical Support:** The Telefónica team will provide the necessary technical assistance for the integration and processing of mobility data. This support encompasses guidance on data formats, clarifications on data attributes, and balances help if technical issues arise with data accessibility or compatibility with our operative systems/applications to handle the data.
- **Mentorship:** Our assigned mentor will offer guidance and support, addressing any questions or unforeseen challenges that may emerge during the project. This includes regular feedback on progress, direction for technical development, and insights into practical applications of the predictive model within public transportation planning. Also thoughts on if the team is facing a realistic project in terms of amount of work and time
- **Team Commitment and Collaboration:** Strong communication and teamwork will be maintained throughout the project. All team members will remain fully engaged and dedicated to achieving project goals, which involves adhering to project timelines, participating in regular meetings, and ensuring effective task distribution across the team

4. Challenges and Lessons Learned

Throughout the project, we faced several challenges that required flexibility and teamwork to solve them.

- ➢ One of the biggest hurdles was the lack of available datasets for events like concerts, congresses, and festivals, which meant we had to manually gather data from various online sources, a process that was both time-consuming and limited.
- ➢ There were also delays in getting the complete meteorological and holiday datasets, which temporarily slowed down our progress in the integration phase.
- ➢ The process of building the predictive model involved a lot of trial and error, with extensive testing and adjustments needed to get the best results. This made the modeling phase take longer than we initially expected

This project taught us valuable lessons in both project management and data analysis.

- On the management side, we learned the importance of solid planning and clear communication, especially when working with stakeholders like Telefónica, coordinating with the PCT team, and handling delays in data delivery. Flexibility was key in overcoming challenges, like manually collecting event data and going through multiple iterations to fine-tune the predictive model.
- From a technical perspective, we discovered the value of bringing together diverse datasets—like weather, holidays, and events—into a unified model. This process emphasized how important it is to ensure data consistency and effective feature engineering for accurate forecasting. We also saw how much external factors can influence mobility, highlighting the need for comprehensive data in planning.

In the end, this project showed us how data-driven solutions can tackle real-world problems while supporting more sustainable transportation

5. Recommendations and Future Directions

5.1. Recommendations

- **Data Enhancement**: Collaborate with additional partners to access more datasets, such as real-time public transportation data or anonymized individual mobility patterns. This could improve the predictive accuracy of the model.
- **Model Optimization**: Do further development of the predictive model by exploring advanced machine learning techniques, such as neural networks, to improve prediction accuracy for more complex mobility trends.
- **Broader Application**: Expand the model to include additional provinces or smaller regions to provide more localized mobility insights.

5.2. Future Directions

- **Dataset improvement:** Delete the columns of our final dataset that finally have not been used for the predictive model.
- **Integration with Public Policy**: Work with governmental bodies to align the predictive model with public transportation policies and broader sustainability goals, such as reducing carbon emissions or promoting public transport over private vehicles.
- **Impact Analysis**: Conduct detailed studies to assess the real-world impact of implementing the model's recommendations, focusing on metrics such as passenger satisfaction, cost savings, and environmental benefits.
- **Incorporation of Emerging Technologies**: Investigate the use of AI-driven optimization for scheduling and resource allocation in public transport systems.
- **Long-Term Collaboration**: Establish ongoing partnerships with data providers, academic institutions, and technology firms to continuously refine the model and address emerging transportation challenges.