Optimizing Train Services: Predictive Modeling and Capacity Planning Amidst COVID-19 Dynamics

Student Name: [Your Name]
Student ID: [Your ID]
Student Email: [Your Email]

Spring 2024

1 Objective

Summarize the main project goal/aim in 3-4 sentences.

The goal of this project is to develop a predictive model for train station ridership, leveraging historical data spanning from 2019 to 2022. By analyzing various factors, including time, station location, and the impact of COVID-19, the objective is to create accurate predictions for the number of passengers at each station. The project aims to provide insights into ridership patterns, assess the influence of external events, such as the COVID-19 pandemic, and offer recommendations for optimizing train services based on forecasted demand.

2 Motivation

Explain what interests you about this topic and why you want to pursue this project.

The dynamic nature of train ridership, influenced by temporal, geographical, and external factors such as the COVID-19 pandemic, presents a captivating challenge. This project provides a unique opportunity to explore the intricate patterns within ridership data and contribute valuable insights to the domain of public transportation. Understanding the underlying trends and predicting passenger demand not only aligns with the current challenges faced by transportation systems but also has the potential to optimize resource allocation and enhance overall service efficiency. This endeavor is driven by the curiosity to unravel the complexities of ridership dynamics and contribute to the improvement of public transit systems.

3 Related Work

Summarize prior work on the topic you have chosen. A critical review of the available publications and industry practices is recommended. [1-2 paragraphs]

Recent studies in public transportation have emphasized the importance of accurate short-term ridership prediction to enhance service efficiency. One such study explores the relationship between short-term subway ridership and influential factors, incorporating bus transfer activities and temporal features into the prediction model [1]. Utilizing gradient boosting decision trees (GBDT), the study demonstrates the model's capability to identify and rank the influences of bus transfer activities and temporal features on short-term subway ridership. This approach offers advantages in capturing complex nonlinear relationships and automatically handling multicollinearity effects, showcasing its potential in improving prediction accuracy within a multimodal public transportation system.

Additionally, research in bus passenger volume forecasting addresses challenges posed by weather and temperature fluctuations, which significantly impact short-term passenger volume. Leveraging bus smart card data and integrating weather and temperature factors, another study proposes an integrated learning algorithm based on XGBoost for forecasting total and commuter line passenger volumes [2]. The model, optimized through hyperparameter tuning, achieves high accuracy in predicting passenger volume every 3 minutes, showcasing its potential for accurate and real-time bus passenger volume forecasting.

4 Methodology

Explain your planned approach, methods, and steps to achieve the objective, and provide rationale for proposing these plans.

The primary goal of this project is to predict subway ridership accurately, considering various influential factors such as temporal and spatial features, and external factors. To achieve this objective, a comprehensive methodology leveraging multiple regression models and advanced predictive techniques is proposed.

4.1 Data Preprocessing

The first step involves meticulous data preprocessing to ensure the quality and relevance of the dataset. This includes handling missing values, encoding categorical variables, and extracting relevant features. Moreover, the dataset will be examined for outliers, and appropriate strategies will be applied for their treatment.

4.2 Feature Engineering

Feature engineering plays a crucial role in enhancing model performance. In this phase, relevant features such as temporal attributes, and external factors will be carefully selected and engineered to capture their impact on subway ridership.

4.3 Model Selection

Multiple regression models will be employed to predict short-term subway ridership. The selected models include:

1. **Linear Regression:** A fundamental regression model to establish a baseline prediction.
2. **Random Forest Regressor:** A robust ensemble model capable of capturing complex relationships within the data.
3. **XGBoost Regressor:** Utilizing the gradient boosting algorithm for improved predictive accuracy.
4. **GLM (Generalized Linear Model) with Poisson Distribution:** A statistical approach considering the nature of ridership data, leveraging the Poisson distribution to model count data.
5. **XGBoost Regressor with Poisson Objective:** Tailoring XGBoost for count data by using the Poisson objective.
6. **Poisson Regression with Exposure:** A specialized regression model accounting for exposure, providing a nuanced perspective on short-term subway ridership.

4.4 Model Training and Evaluation

The models will be trained on a portion of the dataset and evaluated using appropriate metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared. Cross-validation techniques will be employed to assess the models' generalization capabilities and mitigate overfitting.

4.5 Probabilistic Analysis

The prediction outcomes will be subjected to a probabilistic analysis using the Poisson distribution. Probability density functions will be employed to estimate the likelihood of observing specific ridership levels, aiding in decision-making regarding train capacity.

4.6 Optimization Strategies

To further enhance the practical utility of the models, optimization strategies will be explored. This may involve adjusting hyperparameters, fine-tuning model configurations, and assessing the trade-offs between model complexity and predictive accuracy.

The proposed methodology combines traditional regression techniques, machine learning algorithms, and statistical models to provide a comprehensive and accurate framework for short-term subway ridership prediction. By considering a diverse set of factors and leveraging the strengths of various models, the project aims to contribute valuable insights to the field of public transportation management.

5 Deliverables

List any tangible outcomes like code, designs, etc. Or state if mainly report/presentation.

The project will yield several tangible outcomes and deliverables, contributing to both the academic understanding of subway ridership prediction and practical applications in the field of public transportation management. The main deliverables include:

1. **Codebase:** A comprehensive and well-documented codebase containing implementations of the selected regression models, data preprocessing procedures, feature engineering, and model evaluation scripts. The code will be organized and annotated for clarity and reproducibility.

2. **Model Artifacts:** The trained models, including the optimized configurations, will be saved as artifacts. These artifacts can be directly used for predictions or further analysis, facilitating seamless integration into operational systems.

3. **Report:** A detailed technical report summarizing the methodology, data analysis, feature engineering, model selection, training, evaluation, and findings. The report will include visualizations, statistical analyses, and a discussion of the implications of the results.

4. **Presentation Slides:** A set of presentation slides suitable for academic and professional settings. These slides will cover the project's background, objectives, methodology, results, and potential applications.

5. **Documentation:** Thorough documentation accompanying the codebase, providing insights into the structure of the project, dependencies, and instructions for running and deploying the models.

The deliverables aim to provide a comprehensive package, enabling both the replication of the study and practical implementation of the predictive models in real-world scenarios.

6 Resources

List required resources like lab space, software, data access, etc. And note if available.

The successful execution of this project relies on access to the following resources:

- **Hardware:** A personal computer or access to a computing environment with sufficient computational resources to train and evaluate machine learning models. A machine with a multicore CPU and a dedicated GPU (optional but beneficial) is recommended.
- **Software:** The project will utilize the following software tools and libraries:
  - Python programming language (version 3.x)
  - Jupyter Notebooks for interactive development and documentation
  - Scikit-learn, XGBoost, and Statsmodels for machine learning models
  - Pandas and NumPy for data manipulation and analysis
  - Matplotlib and Seaborn for data visualization
  - LaTeX for document preparation
- **Data Access:** Access to the train rides dataset, including historical ridership information, train schedules, and the COVID-19 lockdown status. The dataset will be obtained from [Source Name/URL].
- **Internet Access:** Reliable internet access is required for literature review, accessing online resources, and potential collaboration with external datasets or research findings.

All the mentioned resources are available and accessible, ensuring the feasibility of the project within a typical academic setting.

7 Impact

Describe potential real-world applications or knowledge contributions.

The proposed project has the potential to make significant contributions to both academic research and real-world applications in the following ways:

- **Public Transportation Optimization:** By accurately predicting subway ridership, the project aims to contribute to the optimization of public transportation systems. This can lead to improved planning, resource allocation, and overall efficiency in the operation of subway services.

- **Data-Driven Decision-Making:** The utilization of advanced machine learning models for predicting ridership can empower transportation authorities with valuable insights. This project's findings can facilitate data-driven decision-making, enabling authorities to respond dynamically to changing ridership patterns and external factors like COVID-19 lockdowns.
- **Generalizability to Other Domains:** The methodology developed for predicting public transportation ridership is not limited to subway systems. The same principles can be adapted and applied to forecast passenger volumes in other public transportation modes, such as buses or trains, contributing to a broader understanding of urban mobility.
- **Scientific Contribution:** The application of different regression models, including linear regression, random forest, XGBoost, and Poisson regression, contributes to the scientific understanding of the strengths and limitations of these methods in predicting ridership. The comparative analysis will provide insights into the most suitable models for such forecasting tasks.

The impact of this project extends beyond theoretical advancements to practical applications, fostering improvements in urban transportation planning and management.

8 Milestones

List key project milestones and target dates.

The successful completion of the project will involve achieving several key milestones. The proposed timeline for these milestones is outlined below:

1. **Literature Review:** Conduct an extensive review of literature and related works on short-term ridership prediction and regression modeling.
   Target Date: MM/YYYY
2. **Data Collection and Preprocessing:** Gather and preprocess the train rides dataset, ensuring data quality and compatibility with machine learning models.
   Target Date: MM/YYYY
3. **Exploratory Data Analysis:** Perform exploratory data analysis to understand the characteristics of the dataset and identify relevant features.
   Target Date: MM/YYYY
4. **Model Development:** Implement and train multiple regression models, including Linear Regression, Random Forest, XGBoost, and Poisson Regression. Fine-tune model parameters for optimal performance.
   Target Date: MM/YYYY

5. **Model Evaluation:** Evaluate the performance of each model using appropriate metrics such as Mean Squared Error, R-squared, and Poisson deviance.
   Target Date: MM/YYYY

6. **Probability Calculation:** Implement the probability calculation based on the Poisson distribution for estimating the need for additional trains.
   Target Date: MM/YYYY

7. **Report and Documentation:** Prepare a comprehensive report documenting the methodology, findings, and recommendations.
   Target Date: MM/YYYY

8. **Presentation:** Create a presentation to deliver the project findings and insights to the audience.
   Target Date: MM/YYYY

9. **Submission:** Submit the final project report and presentation materials.
   Target Date: MM/YYYY

The milestones are subject to adjustment based on project progress and unforeseen challenges.

References

[1] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees," *Sustainability*, vol. 8, no. 11, 2016. [Online]. Available: https://www.mdpi.com/2071-1050/8/11/1100

[2] Z. Mei, J. Yu, W. Ding, L. Kong, and J. Zhao, "Bus passenger volume forecasting model based on xgboost integrated learning algorithm," in *CICTP 2020*, pp. 3100–3113. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/9780784483053.261