Predicting a Real Time Passenger Occupancy Using Historical Ticketing Data:

A Case Study of Varanasi

Paper ID: 9707

Authors

- Aadil M Moopan
- Shreepati Jha
- Rahul Kumar Jha
- Dr Agnivesh Pani

IIT (BHU) Varanasi – Transportation Engineering Department

Introduction

- Developing countries lack reliable crowding measures, reducing passenger comfort and ridership – idea of travel itinerary planning is not yet present
- Lack of occupancy data leads to poor transit management, longer wait times, and substantially lesser transit satisfaction
- This study offers a novel method to derive occupancy from ticketing data in tier-2 and tier-3 cities using GTFS and census data
- Improved prediction accuracy helps optimize bus routes and schedules, enhancing service quality and ridership

Figure - Varanasi EV Buses

Extent of Available Literature

- **Passenger Willingness**: Willingness to take longer routes or pay extra to avoid crowds.
- **Improved Distribution**: Crowding data enhances traveler distribution, reducing extreme crowding.
- **Increased Comfort and Reduced Risks**: Better information boosts comfort and mitigates issues like bus bunching and health risks.*(Drabicki et al., 2022)(Thomas et al., 2022)*
- **Effective Resource Allocation**: Knowing crowding patterns enables better resource allocation and scheduling of extra services.*(Marra et al., 2022; Shelat et al., 2022)*

Research Background

Aim of Research

Developing real-time transit occupancy prediction models are a critical imperative for crowd management and pre-empting the service level changes required at the level of a transit system. Transit agencies' experiences evident in the literature underline that occupancy information enhances the passenger satisfaction and overall system reliability.

Expected Outcomes

Enhancing Operational Efficiency: Accurate occupancy predictions allow transit agencies to optimize route planning and schedules for EV buses, reducing trips with low ridership and extending battery lifespan.

Improving Passenger Comfort: Predicting occupancy helps alleviate overcrowding, enhancing passenger comfort and satisfaction while encouraging more people to use public transportation.

Sustainable Urban Mobility: Occupancy prediction aids in optimizing electric bus operations, contributing to a sustainable transportation system and reducing the carbon footprint.

Leveraging Advanced Technologies: Utilizing AI/ML and big data analytics, along with a real-time dashboard, enables predictive analysis and data-driven decision-making for transit agencies.

Data Sources – Ticketing Data

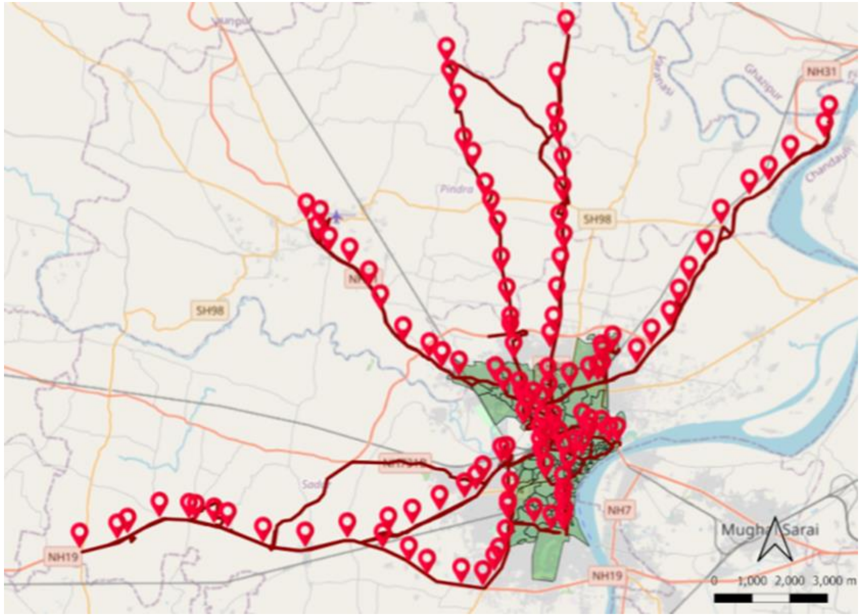


Figure - Bus routes and stop location in Varanasi district

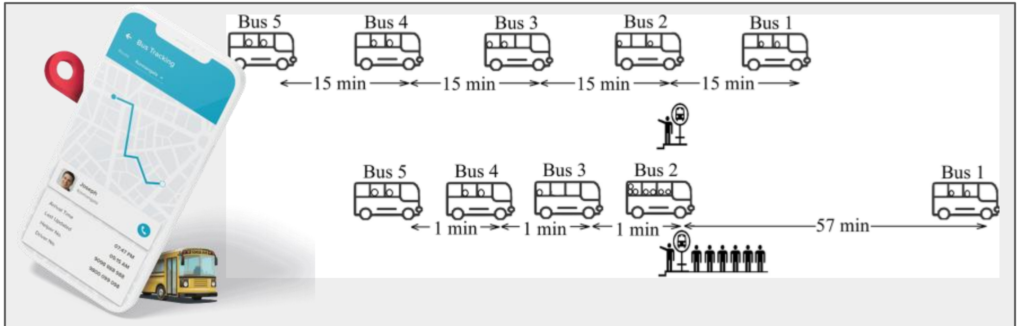
Table - Key metrics related to Varanasi Bus

Metric	***Value***
Average Daily Ridership	*Approximately 7,658 passengers*
Number of Routes Covered	*26 routes*
Total Number of Buses	*56 buses*
Average Ticket Price	*₹ 28.84*
Average Daily Revenue from Ticket Sales	*₹ 2,20,836*
Average Ridership per Bus, per Route, per Trip	*29.35 passengers*
Total Unique Stops Served	*114 stops*
Number of Stops Within City Limits	*45 stops*
Average Speed of Buses	*14.3 km/h*

Specific Research Problem Addressed

Occupancy of a bus stop = Occupancy of Previous stop + Boarding of the current stop – Alighting of the current stop

Based on the observation and Preliminary study of the ticketing data, 22% of the passenger trips experience overloading in the peak hours of the day



- Data-driven decisions enhance transit efficiency and service quality by prioritizing passenger comfort over convenience in crowded situations.
- Consistent bus spacing reduces crowding and health concerns
- Effective crowd management enhances comfort for all passengers

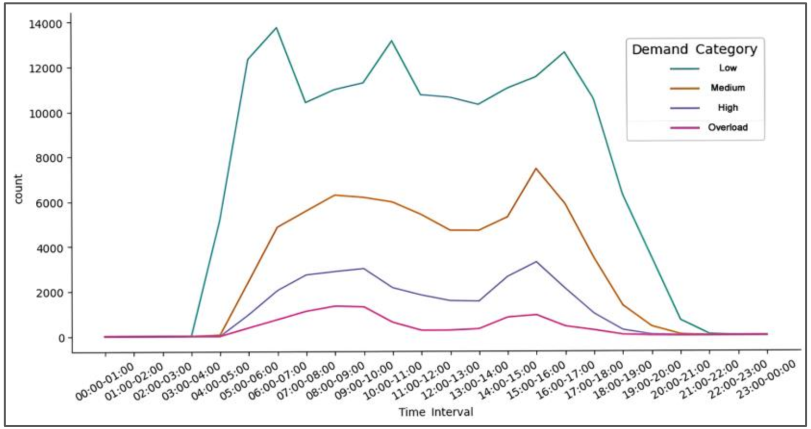


Figure -Demand Category v/s Count

- Low: 0% to 33% of total vehicle capacity.
- Medium: 33% to 66% of total vehicle capacity.
- High: 66% to 100% of total vehicle capacity.
- Overload: 100%+ of total vehicle capacity.

Data Preprocessing

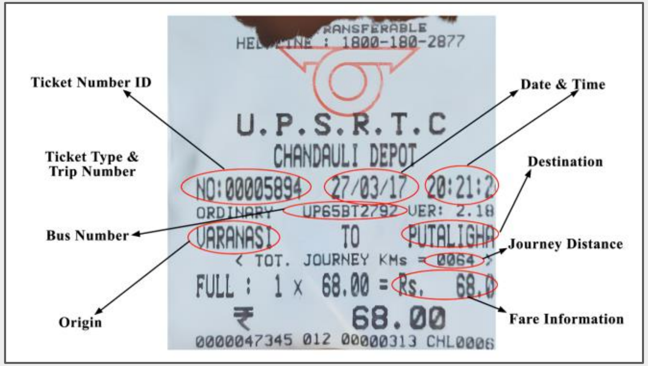


Figure - ETM data for Varanasi Buses

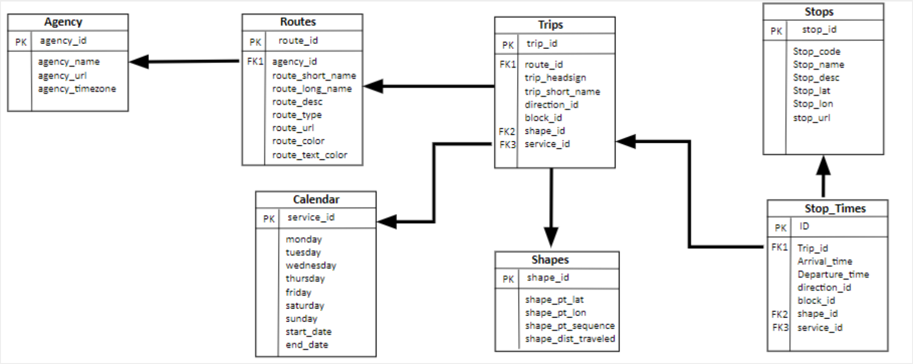
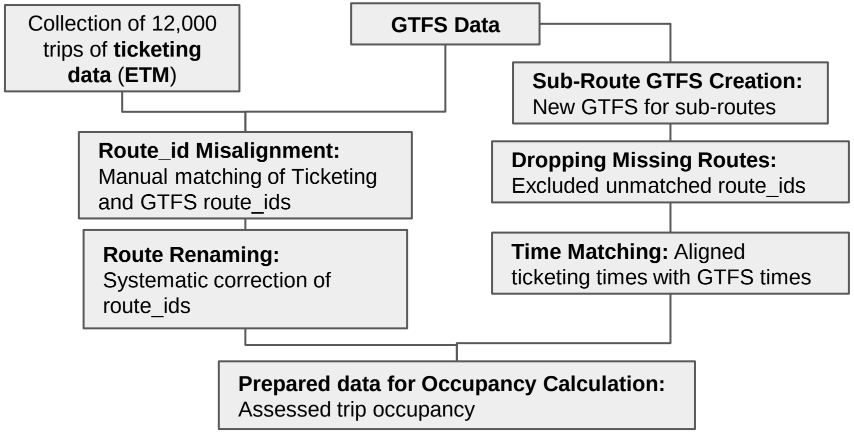


Figure - GTFS and its components

Missing Values	Problem: Missing values can bias results. Solution: Used **KNN Imputation to fill gaps**, using similar data points to estimate missing values.
Categorical Variables	Problem: Categorical variables aren't directly usable. Solution: Applied **One-Hot Encoding to convert categories into a binary format** for machine learning.
Continuous Variables	Problem: Differing scales in continuous variables can skew results. Solution: Used **scaling techniques (Min-Max and Standardization) to normalize data**.



Data Preprocessing

Socio Demographics & Economic Indicators - 2011 Census Data

- Household Density
- Population Density
- Percent of SC or ST Population
- Sex Ratio
- Literacy Rate
- Percent of Workers
- Share of Main Work in Agriculture
- Share of Main Work in Industry
- Share of Main Work in Services
- Road Density
- Rail Density
- Water Density
- Intersection Density
- Share of Marginal Work in Agriculture
- Share of Marginal Work in Industry
- Share of Marginal Work in Services

Preparing Night Time Light(NTL) Dataset

Data Access: VIIRS collected and raster clipping

Data Pre-Processing: Perform noise reduction and calibration

Data Alignment: Conduct georeferencing, projection, and resampling

GIS Integration: Overlay spatial data and perform spatial queries

Extraction of Mean NTL: Apply masking and spatial aggregation (mean, median)

Additional Indicators: Compute SD, max/min intensity, and temporal change

Mean Night-Time Light (NTL) for each zone/ward in Varanasi

Geospatial Dataset

Village boundary Shapefile downloaded from OVSF/-/10(SOI,2023) ,

Gathered ward dataset at the urban level.

Merged shapefile with the ward dataset.(114 Zones)

Mapped 16 variable from Census 2011 and NTL to zones.

Validated accuracy and completeness of the dataset.

Dataset ready for the analysis

Summary of Preprocessed Dataset

Descriptive statistics of all Socio economic +NTL dataset

	count	mean	min	25%	50%	75%	max	std
Trip_ID	271220.0	6265.848798	1.0	3124.0	6216.0	9388.0	84690.0	3758.263769
Demand	271220.0	11.58398	0.0	3.0	9.0	18.0	97.0	10.477135
Date	271220	2023-06-16 14:19:08.198510848	2023-05-24 00:00:00	2023-06-04 00:00:00	2023-06-16 00:00:00	2023-06-28 00:00:00	2023-07-11 00:00:00	NaN
Zone_ID	271220.0	135091.217974	3.0	32.0	209039.0	209465.0	249531.0	100480.841623
HH_DEN	271220.0	1356.079276	0.0	365.0311	421.9084	1998.909	11884.60629	1711.992308
POP_DEN	269423.0	8945.209164	228.4317	2460.817	2842.751	11578.25	89315.41129	11378.633821
SCST_CENT	269423.0	11.246826	0.0	7.277377	10.447603	14.672708	40.265487	7.191406
SEX_RATIO	269423.0	893.2146	737.963265	871.665133	897.042607	928.5547	1065.594059	48.883303
LIT_RATE	269423.0	50.251712	16.866709	30.438312	60.477723	62.93882	78.80214	18.624042
WORK_CENT	269423.0	32.889839	19.911504	29.66198	30.778703	34.612993	51.348113	6.56053
MAINWORK_CENT	269423.0	24.645112	9.538003	21.74177	25.274725	27.92544	49.191132	5.959256
MAINWORK_SHARE	269423.0	75.207389	22.87234	68.073879	79.38428	84.513591	97.658402	12.849296
ROAD_DEN	271220.0	16.372026	3.611519	10.204369	12.746835	18.86737	54.898419	10.411958
RAIL_WATER_DEN	271220.0	109.251331	0.0	0.0	0.086957	9.730227	1357.877313	266.107748
INT_DEN	271220.0	239.957286	16.171117	47.69689	64.85643	334.5355	1251.751169	303.939973
ntl_mean	271220.0	20.874675	1.683888	7.009539	12.324284	34.677212	60.939944	17.794448
MARGWORK_CENT	269423.0	8.244728	0.622939	4.74318	6.87865	10.176991	33.105023	5.069658
MARGWORK_SHARE	269423.0	24.79261	2.341598	15.486409	20.615723	31.926121	77.12766	12.849296

Figure - Descriptive statistics of all Socio economic +NTL dataset

Occupancy over 45 days of Collected data

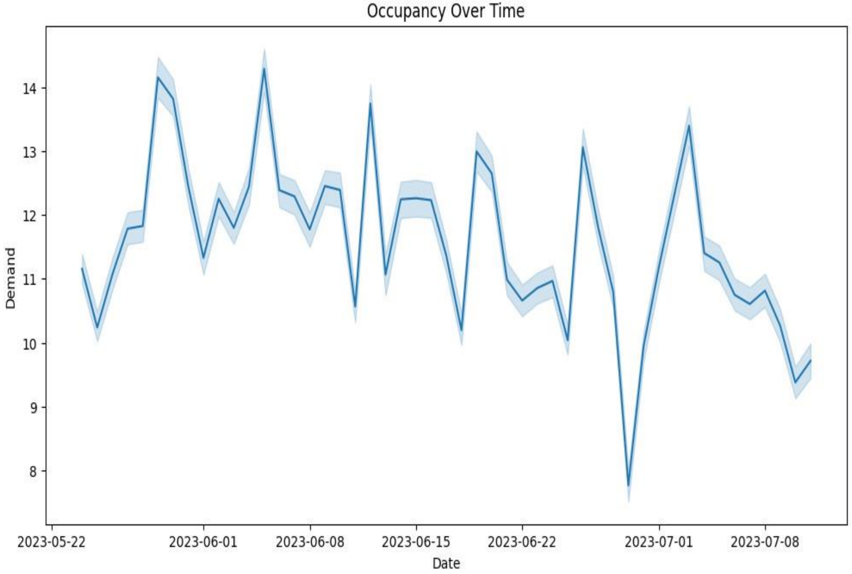
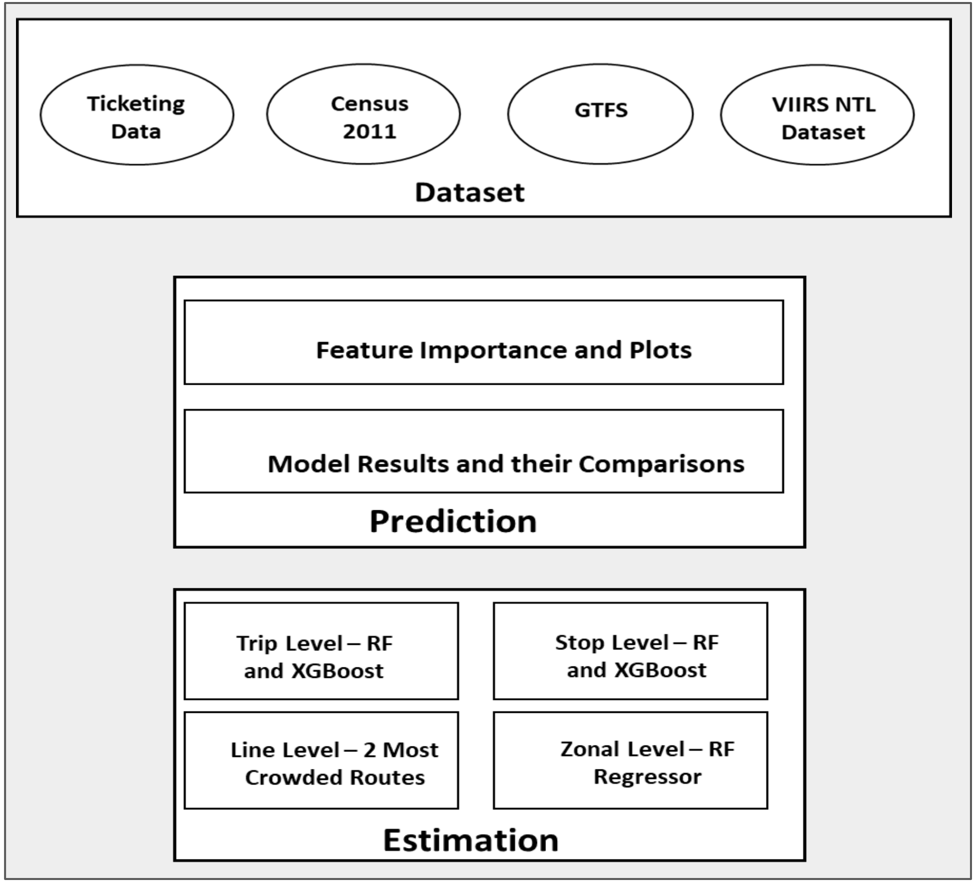
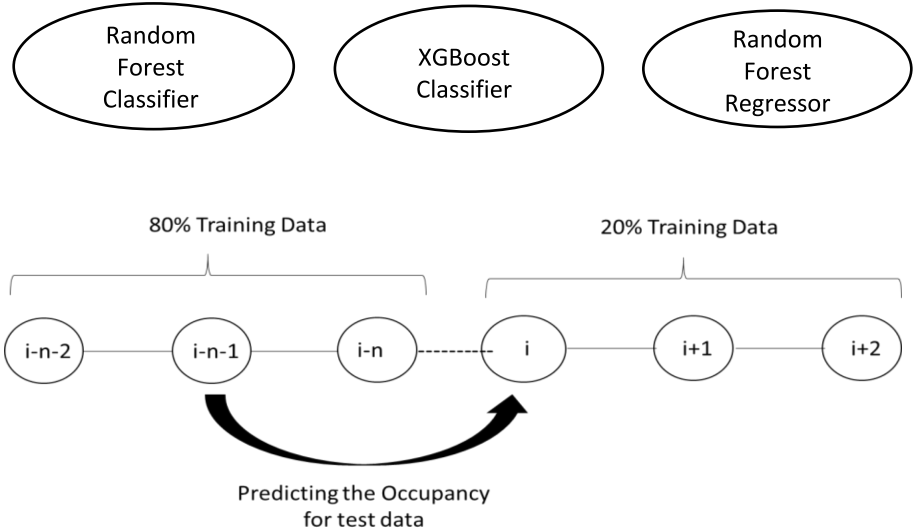


Figure - Occupancy distribution of Bus network in a 45 days period

Methodological Framework





Stop Level	Random Forest Classifier
Trip Level	XGBoost Classifier
Zone Level	Random Forest Regressor

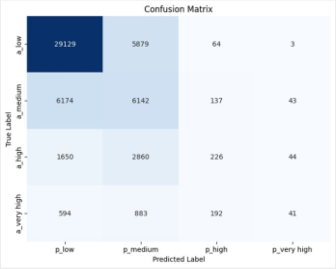
Trip Level Analysis

Model	Accuracy	Precision	Recall	F1-score	Custom Accuracy Metric
RF without stop characteristics	0.553	0.6662	0.553	0.5944	0.553
XGB without stop characteristics	0.6582	0.5999	0.6582	0.6032	0.6582
RF with stop characteristics	0.5441	0.6698	0.5441	0.5834	0.5441
XGB with stop characteristics	**0.6573**	**0.6355**	**0.6573**	**0.6297**	**0.6573**

Table - Trip Level Model Comparisons

Feature Importance: The total effect of each variable on transit trip occupancy is analyzed, highlighting their influence on the final outcome.

Key Predictors: The time interval is identified as the most significant variable for predicting occupancy, followed by Route ID, Weekday, and stop station.



A confusion matrix is a 4x4 table that assesses a classification model's accuracy by comparing actual and predicted values across four categories (0, 1, 2, 3).

Figure - Confusion Matrix for best XGBoost Model

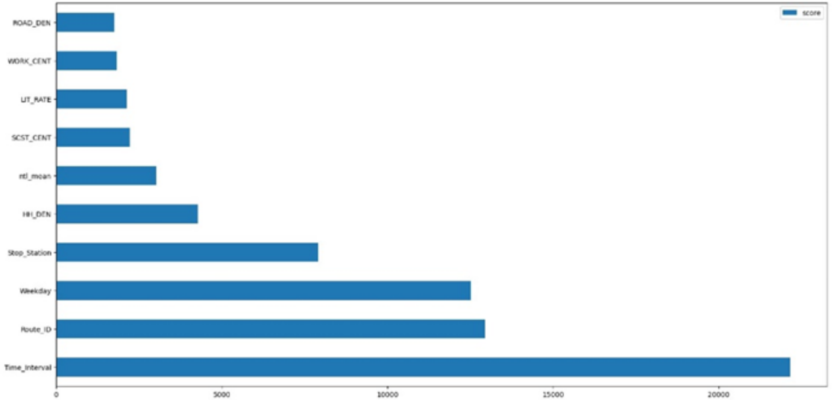


Figure - Feature Importance of XGBoost Model

Additional Contributors: Household Density, Mean Night Time Light, Percent of SC/ST, and Literacy Rate also significantly contribute to the model's predictions.

Stop Level Analysis

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Custom Accuracy Metric*
XGB with stop characteristics	***0.6527***	***0.6136***	***0.6527***	***0.6168***	***0.6527***
Random Forest with stop characteristics	*0.5706*	*0.6676*	*0.5706*	*0.6056*	*0.5706*

Table - Stop Level Model Comparisons

Permutation Importance: A model-agnostic technique that estimates feature importance by shuffling feature values and observing the impact on model performance.

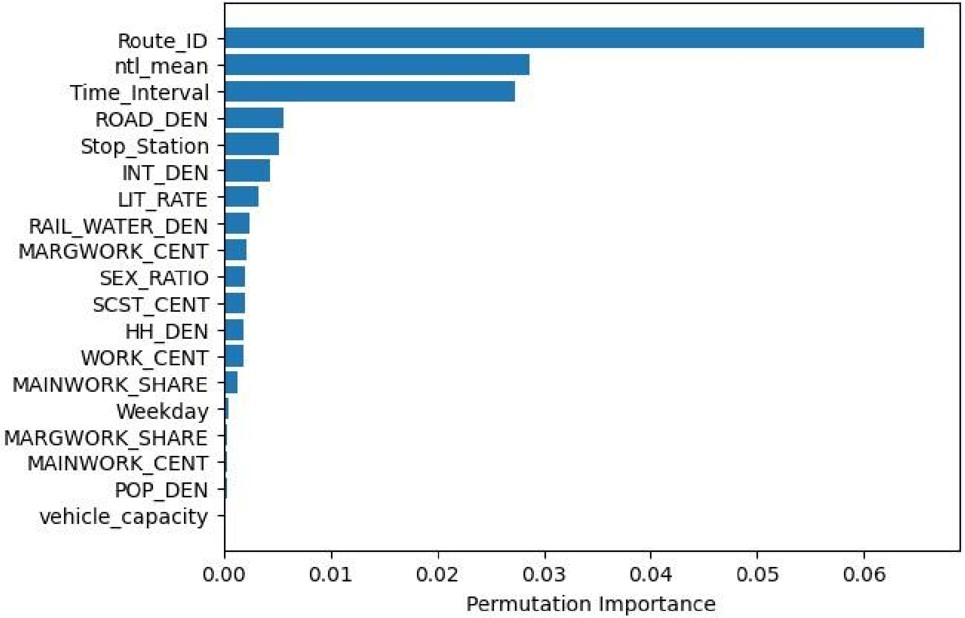


Figure- Permutation Importance of Best model

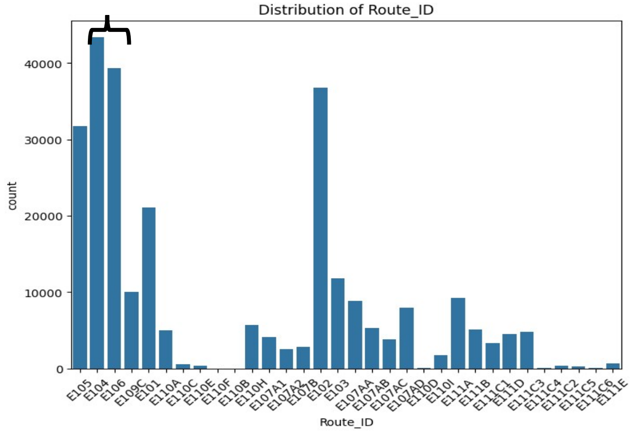
Key Features in Occupancy Prediction: In predicting stop-level occupancy, Route ID is critical, while the average level of Night Time Light (NTL_MEAN) emerges as the most influential factor. Higher NTL values correlate with increased occupancy, and additional important features include Road Density and Intersection Density, indicating that well-developed road networks also contribute to higher occupancy rates.

Line Level Analysis

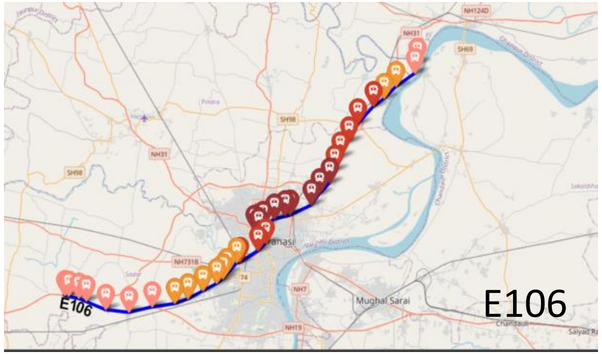
Most Crowded Lines of Varanasi

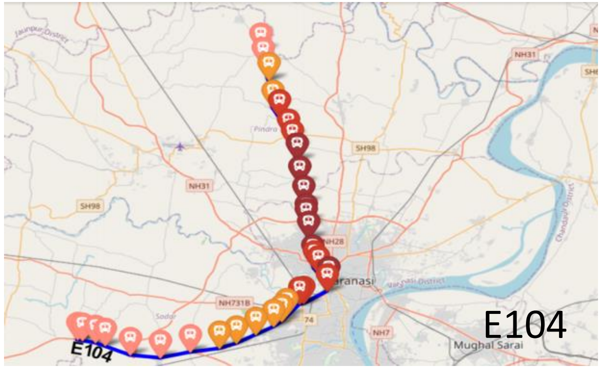
Route ID	*Demand*
E106	*636405*
E104	*589078*
E102	*503558*
E105	*350249*
E101	*201032*
E103	*129804*

2 most crowded routes



Comparison study of two routes – E106 and E104



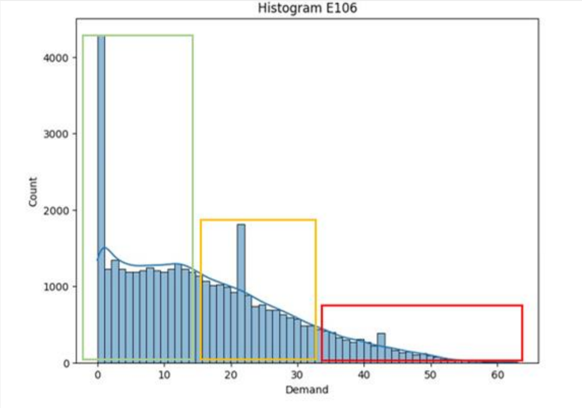


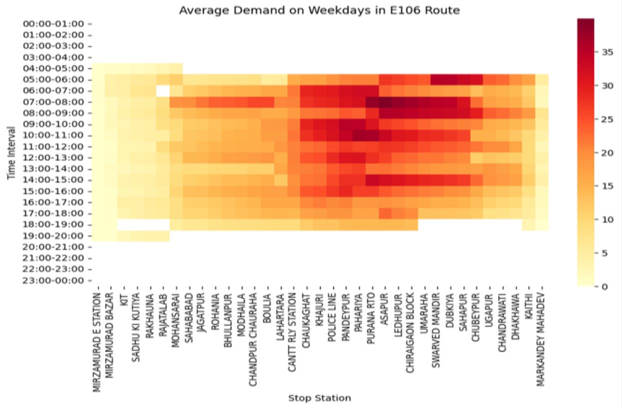
Bus Line	Number of Stops	Key Features
E106	30	Starts in the outskirts, passes through the city center, ends in another outskirt area; includes Cantt Railway Station.
E104	34	Starts in the outskirts, passes through the city center, ends in another outskirt area; includes Cantt Railway Station.

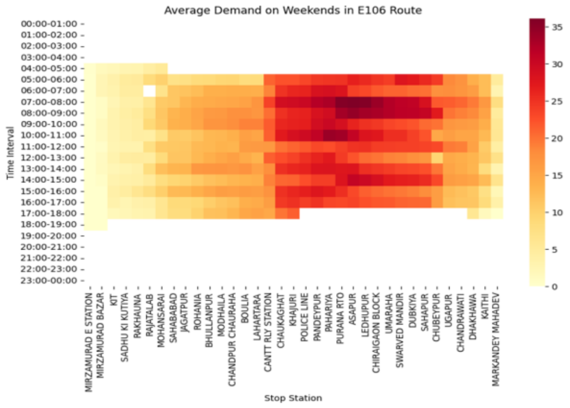
Line Level Analysis

Comparison study of two routes Occupancy vs Count Plots	Comparison study of two crowded routes: Heatmap for Weekdays	Comparison study of two crowded routes: Heatmap for Weekends

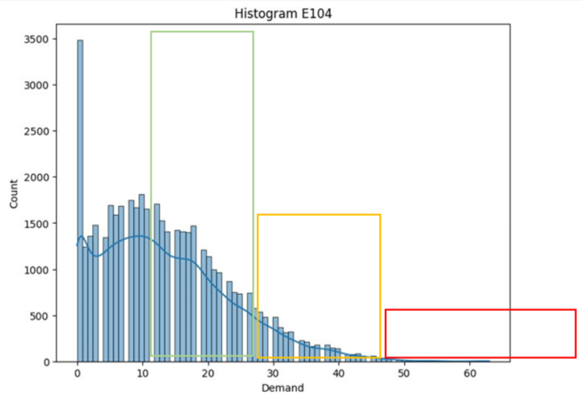
E106

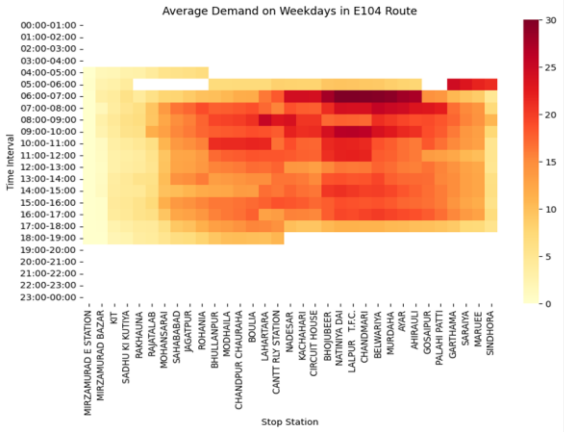


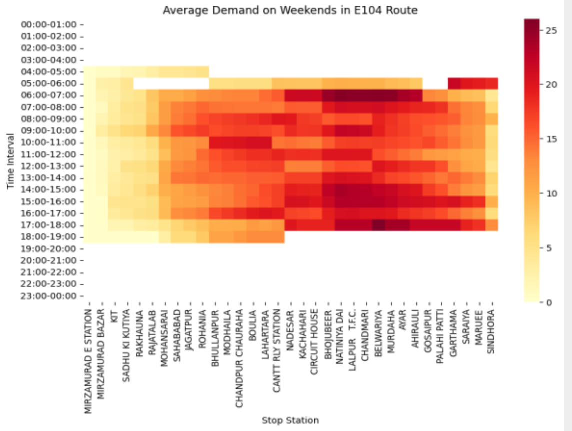




E104







Zone Level Analysis

At the zonal level, we used a Random Forest Regressor to predict occupancy, evaluating its performance with two key metrics:

- **Mean Squared Error (MSE)**: Measures the average of squared differences between predicted and actual values. Our model's MSE is 0.04384, indicating low error and sensitivity to outliers.
- **Mean Absolute Error (MAE)**: Assesses the average absolute differences between predicted and actual values. The MAE for our model is 0.15688, suggesting a modest deviation without emphasizing larger errors.

Together, these metrics indicate the model's accuracy, with low MSE showing strong performance and MAE reflecting robustness against extreme deviations.

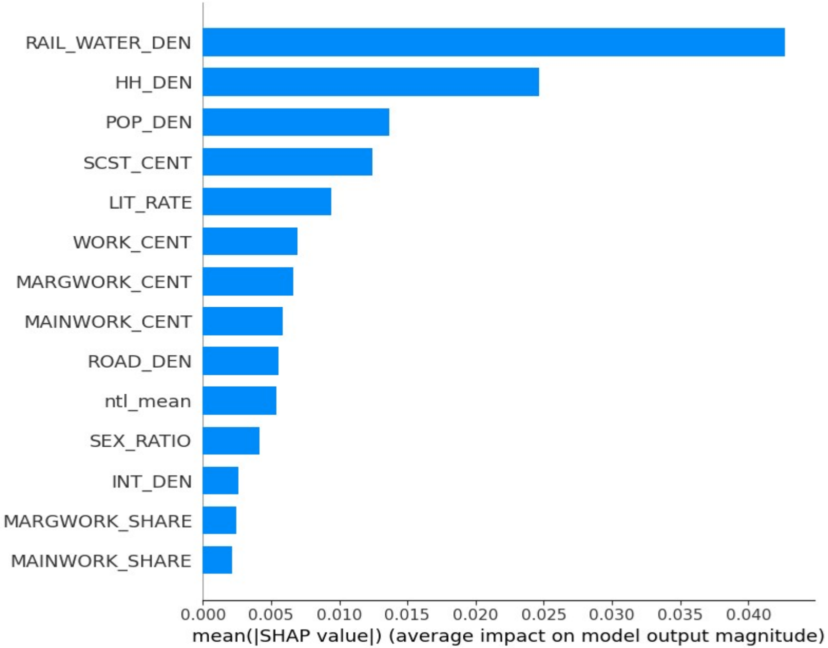
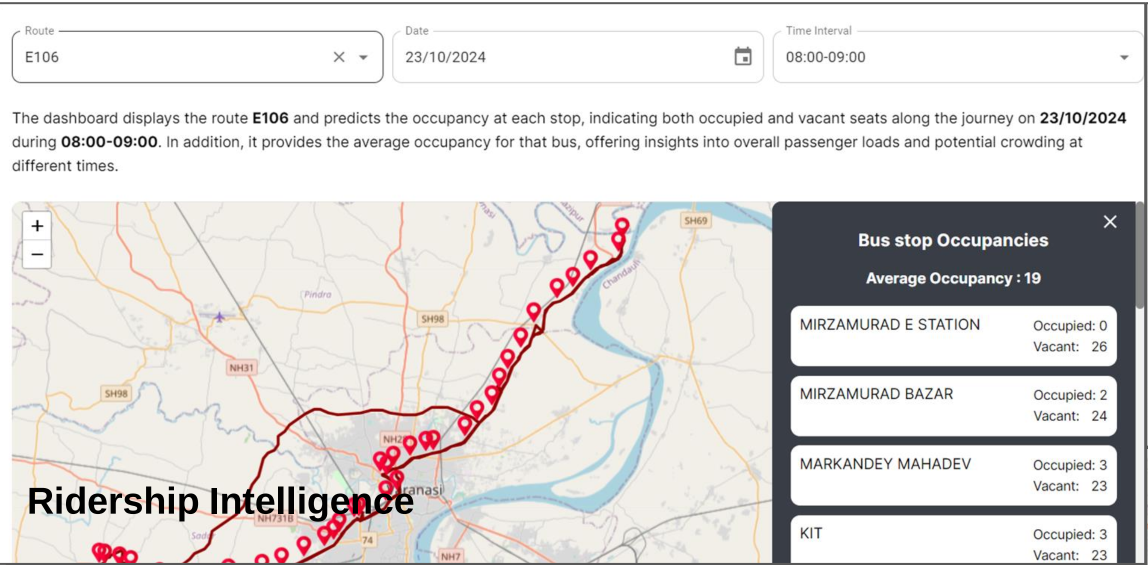
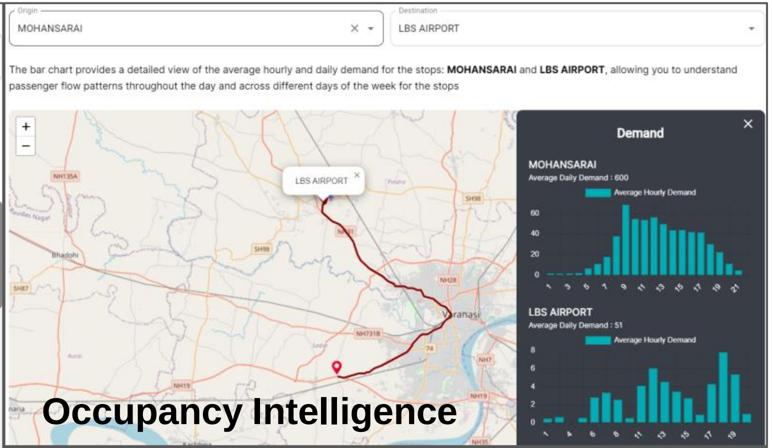
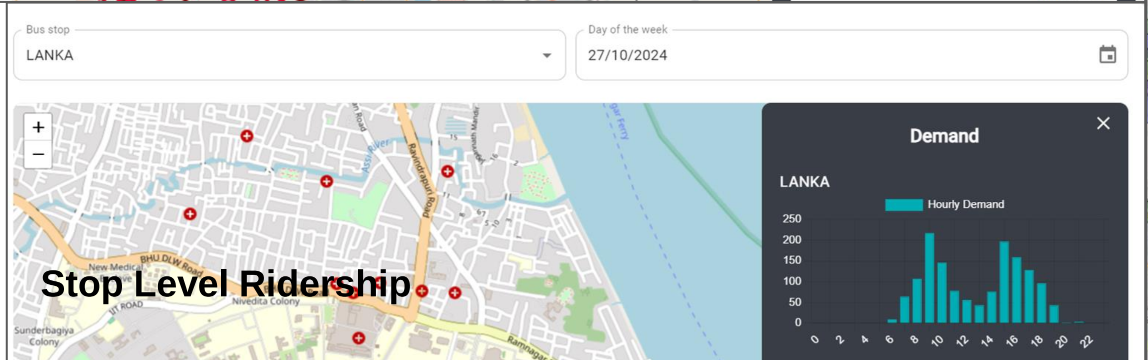


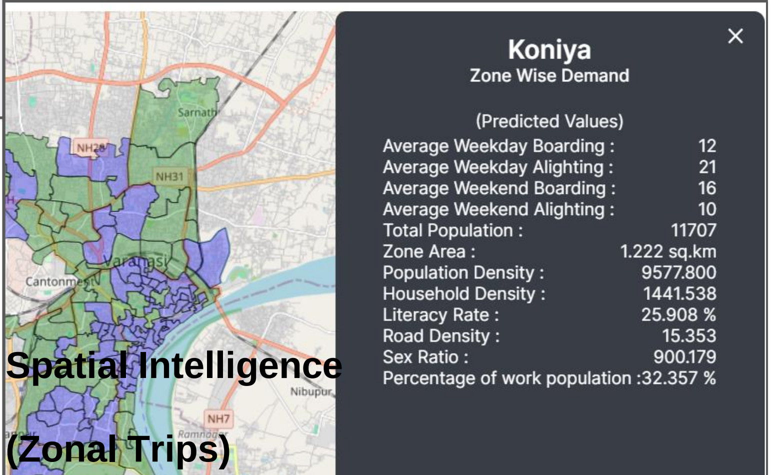
Figure - SHAP Analysis for Random Forest Regressor

Final Dashboard Developed for Transit Management









Thank You

For More Information:
Shreepati Jha
Incoming PhD Student
University of Alabama at Birmingham
Project Fellow at IIT BHU Varanasi
Email: shreepatijha777@gmail.com

Research Lab Website: Dr. Agnivesh Pani's SCULPT{Lab} http://sculptlab.in/ (agnivesh.civ@iitbhu.ac.in)