

A Web Based Application For Tracking Public Transport And Predicting Usage

Rekha. R¹, Abirami A P², Aishvarya G³, Akshaya B⁴, Annapoorna K Athreya⁵,
Sanchana S⁶

^{1,2,3,4,5,6}Department of IT, PSG College of Technology, Coimbatore.

E-Mail: rra.it@psgtech.ac.in¹

Abstract: *The application aims to address the problem of pattern analysis and predictions of business assets usage. The main goal is to provide business insights and recognize patterns of utilization of assets—whether the assets are employees, books, bicycles, so on—against the behavior of a variety of external factors, and to predict the usage by way of a Linear Regression algorithm with multiple variables to predict usage patterns for better clarity, and better financial decision making in areas like budget, marketing, etc. In the proposed work, the business is considered as a public bus company and the assets are buses, since it is a socioeconomically beneficial use case. Using a system of stops and 3 overlapping routes, The application reports on bus journeys and predicts passengers, according to several parameters. This can be generalized for any business with assets that need monitoring and have varying levels of availability. For example, restaurants (assets here could be the wait-staff or tables, and so on), furniture rental companies, libraries, etc. Based on collected data, and performed correlations between dependent and independent factors, a linear regression algorithm will project the frequency of bus usage for the next duration of time (week), in accordance with various factors like routes, capacity, etc, and also external factors like weather (precipitation), month of the year, etc. Specifically, this work would help make the public transport system more efficient, and therefore environment-friendly, economically more viable, more accessible and easier to use for the general population.*

Keywords: *Linear Regression, Data visualization, Report generation, Prediction*

1. INTRODUCTION

Transportation infrastructure is essential to respond to an increasing travel demand, with greater strength and better regulation of the public transport system. In terms of both efficiency, and interest of public sector growth and environmentally-friendly practices, urban public transit systems require more support, to reduce dependency on private transport contractors and vehicles[1].

Accessibility is also an important factor in assessing a public transport system. This work attempts to help organize, analyze and generate reports, based on several factors to improve the public transportation system, and predict the number of passengers travelling on a day of the week, for given parameters.

Based on the Global bus survey of 2019, 90% of people are standard bus users and 10% of people use other transportation modes. Out of this Almost 68% of the bus fleet are standard buses (as regular 12-meter buses). This showcases a global average, including countries that

are well past their “developing” stages and contain smaller populations and lesser per-head expenditure. When it comes to India, there is great potential for sustainable development and improvement on many fronts, one of which is the public (bus) transportation sector with limited resources [2 -4].

The application, by tracking and monitoring the usage of buses, aids the bus authority (in this case, the public bus system is part of the government) make important decisions in the future- like in marketing strategies, finance, budgeting, product development, surveying, manufacturing and better customer service. It also serves as documentation and offers transparency in business decision-making.

The objective of this work is to generate comprehensive reports using data visualization techniques and a multi-variable linear regression algorithm to help predict the usage of buses in terms of passenger count in the next week/month.

2. Literature Survey

Table 1 shows the studies that were carried out in various industry sectors such as the healthcare, the financial and transportation. The limitations of these papers have been included to understand what improvements can be made while applying and modifying similar concepts. In the current scenario, Artificial Intelligence is used in several applications such as prediction of travel demand, environmental pollution, CO2 emission and so on.[5]. Prediction models for public bus transport usage were developed by [6] and achieved over 78% of accuracy.

Table 1: Literature Survey

S.NO.	ARTICLE	DESCRIPTION	RESULTS & LIMITATIONS
1.	Multiple Linear Regression Model to Assess the Effects of Macroeconomic Factors on Small and Medium-Sized Enterprise.[7]	This paper focuses on the impact of macroeconomic conditions on small and medium sized enterprises (SMEs) in Sweden where a data set of over a 10-year duration (2009- 2019)	A multiple regression analysis had been performed to examine the significance of the relationship between macroeconomic variables which resulted with a coefficient of determination of 98%. However, the 10-year time span used is relatively short and as often the case, interpolation and prediction of the future under these circumstances is usually quite unreliable and inadequate.

2.	Predicting Apartment Prices with Multiple Linear Regression Model.[8]	This paper uses factors like area, number of rooms, monthly fee, etc to predict the prices of apartments in Stockholm City Centre.	Using multiple linear regression approach, the project determines apartment prices, having various external factors (dependent variables).It is limited to sporadic and varied external independent parameters.
3.	An Analysis of the Relationship Between the Utilization of Physical Therapy Services and Outcomes for Patients with Acute Stroke.[9]	In this study, the examination of acute care of patients with stroke who were treated in US academic health center (AHC) hospitals was conducted. Here, multiple linear regression analysis was conducted for the examination of the relationship between utilization of physical therapy and total cost of care.	The results indicate that physical therapy utilization was directly related to a total cost of care that was less expensive than expected and to an increased probability of the patient being discharged back home. Nevertheless, the regression analysis in this study did not explain a high percentage of the variation in the dependent measures and this suggests that additional variables are needed in the models.
4.	LiRCUP: Linear regression-based CPU usage prediction algorithm for live migration of virtual machines in data centers [10]	This paper focuses upon prediction of CPU usage using linear regression model.	The advantage of this method is that using the predictions made by the model, the under loaded and over loaded hosts can be identified.

Proposed method

The system design of the application is shown in figure 1.

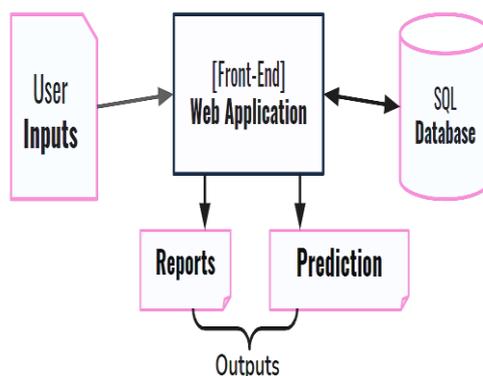


Fig 1. Structure of the System

The system has 4 modules: The inputs, the storage (the Database), the outputs (Reports and prediction). They work in tandem behind the front-end system, to produce a well-rounded tool that can be adapted for any business and any asset. Since a Machine learning linear regression model is being used, future is predicted in a more accurate way based on the usage of asset

The flow of the system has three sections to it and they are as shown in figure 2.

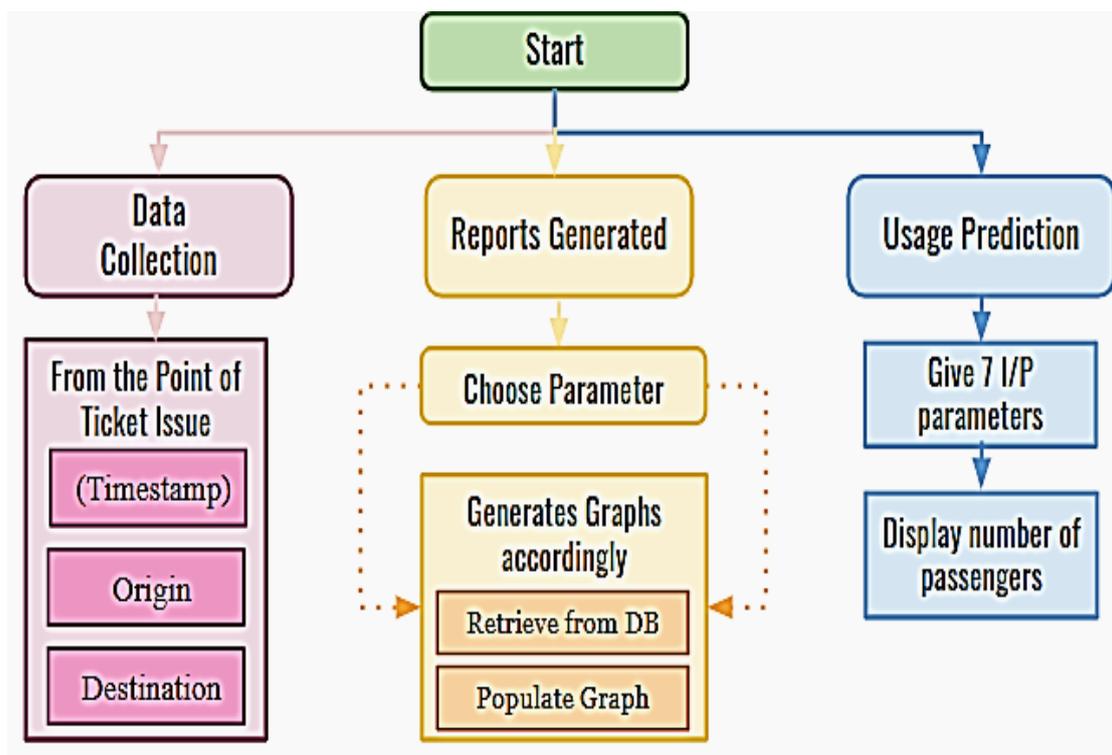


Fig 2. Flow of the System

2.1 Routes and Nodes

In the proposed work, it is assumed that the system has a fixed number of bus stops and routes of the buses that will pass through these points, in order to increase the reporting

capabilities and for compatibility of the system with the format of data collected. A total of 14 possible bus stops, spread over 3 possible routes, as visualized in the image below is considered.

The 3 Routes are:

Route 1 (20A): Gandhipuram, Lakshmi Mills, PSG (Peelameedu), CIT, Sitra/Airport, Karumathampatti.

Route 2 (8): Singanallur, Ramanathapuram, Town Hall, Ukkadam, Chettipalayam.

Route 3 (3R): Saibaba Colony, Gandhipuram, VOC Park, Town Hall, Ukkadam, Kuniamuthur.

For further reference, the terminologies used will be:

Node: Bus stop/ Point

Source: Any of the 14 nodes as boarding point

Destination: Any of the 14 nodes as a destination/ deboarding point for passenger

Segment: Path between two nodes (a single straight path with just a source and a destination)

Route: Any of the 3 fixed routes leading from each starting source to Destination

Since there is one Source and one Destination per ticket issue/ passenger journey, one can form ${}^n\text{C}_2$ combinations of Origin-Destination where 'n' is the number of nodes. The fixed number of nodes for this system is 14, (with 3 routes) and hence the total number of unique Origin-Destination pair combinations possible is ${}^{14}\text{C}_3 = 91$.

The overlapping routes would look similar to the map represented in the figure 3.

2.2 Dataset Description

A generated dataset has been used for the analysis and prediction of data for which the parameters include date, day, time slot, precipitation(in), temperature, passengers, whether it is a working day or not (0 or 1, boolean), the origin and destination of the bus journey. This dataset consists of information about the parameters that have been mentioned above for a month's record. All the reports and predictions are made using this dataset, and it has 2877 entries/rows, which amount to a month's recorded data. A sample of the dataset has been in figure 4 to illustrate the structure of the dataset.

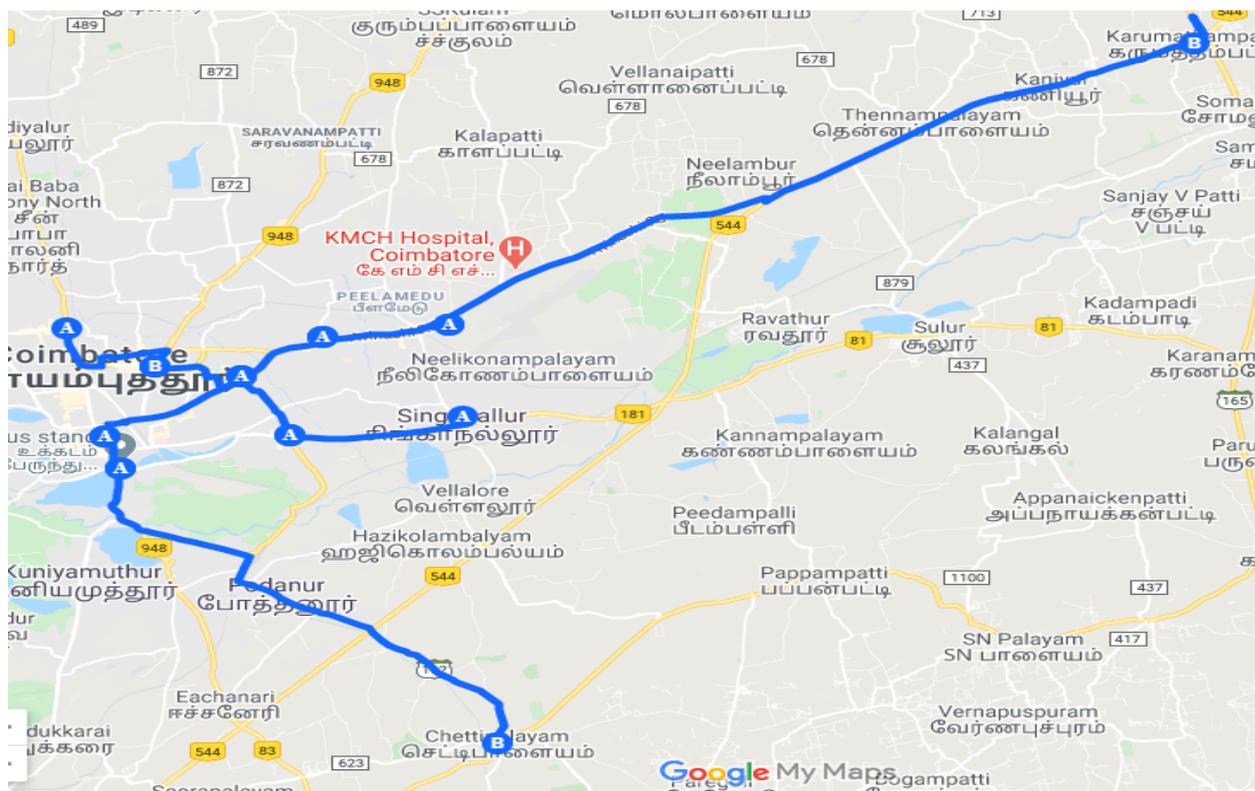


Fig 3. Visualization of Routes

2.3 Storage and Data Report System

Data Reports are generated on the basis of the user's choice of parameters for pattern analysis. One important note is that the data retrieved is taken from the point of ticket-issue, and for purposes of this work, a code stub has been added to act as the ticket-issue portal. However, in a real-life implementation scenario, data collection is outside the box of the software and this data will be received from the system that the current bus authority already uses- perhaps through an API or sharing of database access, etc. In ticket-issue, the point of boarding and deboarding of the passenger is taken, along with a timestamp. Now, this raw ticket issue data is processed to form another table which splits the time into time slots of every hour, starting from the first working hour to the last, of the bus service, and also aggregates all the passenger counts on an hourly basis, along with the other parameters.

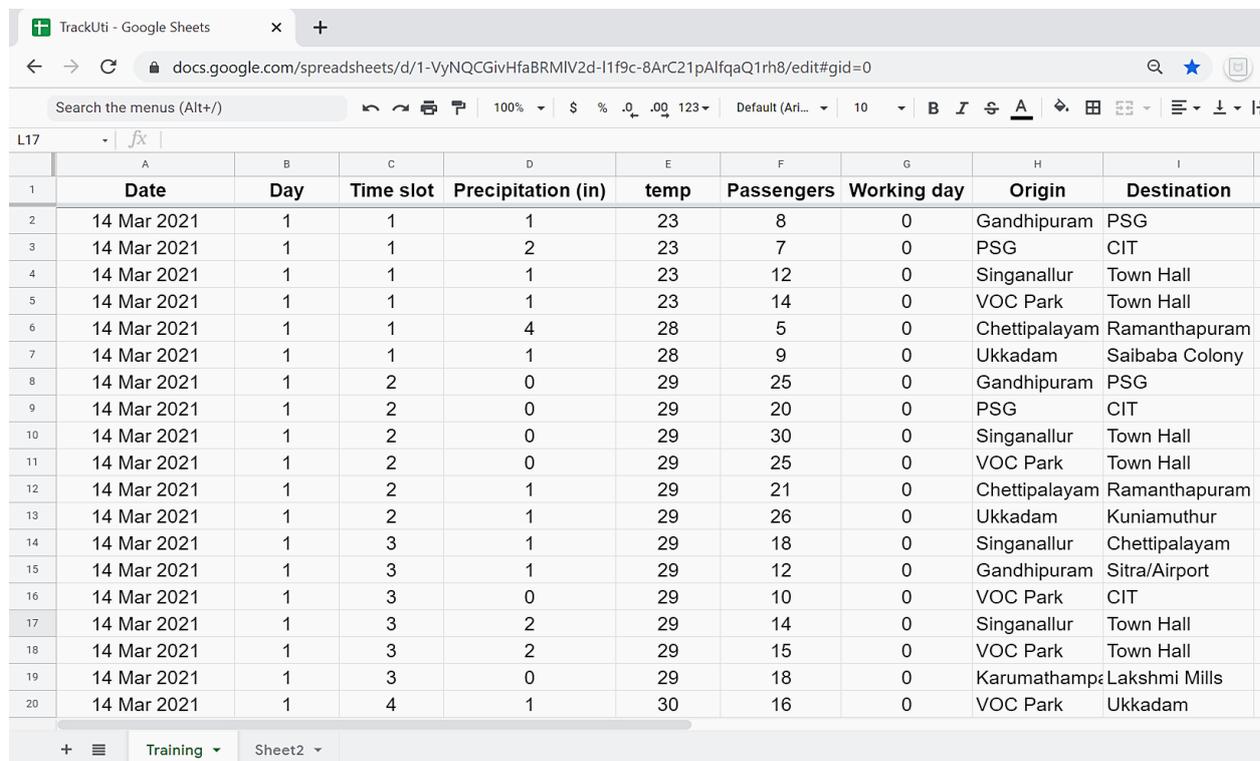
The data is stored in a Relational database, since there is no need for ambiguity in data storage and all values are similarly formatted, with the same fields. In this work, Google sheets, google's cloud service DB provided is being utilized. The front end UI is developed in HTML plus JavaScript. While making reports, the data taken from the front end and stored in the database, is retrieved and represented as graphs with plots of Passenger counts (y-axis) against every parameter (x-axis), using D3js- which is a JavaScript framework for data manipulation and representation.

2.4 Prediction System (ML Model)

The usage of the public bus system- by measure of number of passengers expected- is predicted (Dependent variable), in relation to multiple factors (Independent variables), through a multivariate **Linear Regression** model. This supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a linear regression model to generate reasonable predictions for the response to new data.

The project uses a training dataset with varying values of the nine inputs:

- **Date** (Specified in DD/MM/YYYY format)
- **Day** (Represents the number of days in a week ranging from 1 (Sunday) to 7 (Saturday))



	A	B	C	D	E	F	G	H	I
1	Date	Day	Time slot	Precipitation (in)	temp	Passengers	Working day	Origin	Destination
2	14 Mar 2021	1	1	1	23	8	0	Gandhipuram	PSG
3	14 Mar 2021	1	1	2	23	7	0	PSG	CIT
4	14 Mar 2021	1	1	1	23	12	0	Singanallur	Town Hall
5	14 Mar 2021	1	1	1	23	14	0	VOC Park	Town Hall
6	14 Mar 2021	1	1	4	28	5	0	Chettipalayam	Ramanthapuram
7	14 Mar 2021	1	1	1	28	9	0	Ukkadam	Saibaba Colony
8	14 Mar 2021	1	2	0	29	25	0	Gandhipuram	PSG
9	14 Mar 2021	1	2	0	29	20	0	PSG	CIT
10	14 Mar 2021	1	2	0	29	30	0	Singanallur	Town Hall
11	14 Mar 2021	1	2	0	29	25	0	VOC Park	Town Hall
12	14 Mar 2021	1	2	1	29	21	0	Chettipalayam	Ramanthapuram
13	14 Mar 2021	1	2	1	29	26	0	Ukkadam	Kuniamuthur
14	14 Mar 2021	1	3	1	29	18	0	Singanallur	Chettipalayam
15	14 Mar 2021	1	3	1	29	12	0	Gandhipuram	Sitra/Airport
16	14 Mar 2021	1	3	0	29	10	0	VOC Park	CIT
17	14 Mar 2021	1	3	2	29	14	0	Singanallur	Town Hall
18	14 Mar 2021	1	3	2	29	15	0	VOC Park	Town Hall
19	14 Mar 2021	1	3	0	29	18	0	Karumathamp	Lakshmi Mills
20	14 Mar 2021	1	4	1	30	16	0	VOC Park	Ukkadam

Fig 4. Dataset

- **Time Slot** (Indicates the timing when buses travel which ranges from 1st to the 6th working hour)
- **Precipitation (%)** (the weather plays a role in willingness of citizens to use public transport)
- **Temperature** (Another measure of the Weather of the place at the time)
- **Working Day** (Working days see a lot more usage of public transport. 1 represents that it is a working day and 0 represents that it is a holiday)
- **Origin** (Boarding point for passenger)
- **Destination** (De-boarding point for passenger)
- **Passengers** (Number of passengers per journey)

Out of which, seven parameters (Day of the week, time slot, precipitation, temperature, origin, destination) are independent variables on which the dependent variable, Passengers, depends. These follow a pattern that will help the model gain insight into which factors affect

the usage of public transport buses the most, etc. The system, through training, uses the given dataset to form a mathematical model in a format similar to following simplified version:

$$Y = b_0 + m_1x_1 + m_2x_2 + \dots + m_nx_n \text{ (Eq. 1)}$$

Where, Y is the Output to be predicted/ Dependent variable, and each x_i is an independent variable on which Y depends with intercept b_0 . For this work, the equation specifically would represent,

$$\text{Passengers} = \text{Intercept} + \text{coeff1}(\text{Day}) + \text{coeff2}(\text{Timeslot}) + \dots \text{ (Eq. 2)}$$

2.5 Data Pre-Processing: String Hot Encoding

Linear Regression algorithm uses all real number or Boolean inputs for the model. Therefore, when using String parameters such as Origin (Boarding point) and Destination (De-

	CIT Gandhipuram	CIT PSG	Chettipalayam Ramanthapuram	Chettipalayam Singanallur	Gandhipuram CIT	Gandhipuram Karumathampatti	Gandhipuram Kuniamuthur	Gandhipuram PSG	Gandhipuram Sitra/Airport
0	0	0	0	0	0	0	1	0	
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	
0	0	1	0	0	0	0	0	0	
...	
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	1	0	
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	

ows x 52 columns

Fig. 5. Hot encoded Origin-Destination pairs

boarding point) in the model, one must employ some pre-processing methods for representing these strings as number values. Strings cannot directly be replaced with a number to represent it, since the model will make assumptions about these values which will not be true and therefore misleading for the string values it represents. This will lead to the model being either entirely invalid or abysmally inaccurate.

For example, one cannot replace Nodes such as “Singanallur” with a 0, and “Gandhipuram” with a 1, since it will assume that $0 < 1$, which is true for numerical values, but not for the strings that they represent. Hence, a technique for String encoding method known as Hot encoding to be employed. This method essentially takes each unique String value in the (selected) column and creates a column each with Boolean values. If row i has the string value S_j in the string column, then for m unique strings, m columns are added where element at index i, j will have the Boolean 1 value and the rest of the columns in row i will have the value 0.

String encoding is complex enough as it is, without including two separate String parameter columns as in the application (Origin and Destination), so the two columns are merged and separated with a single space in order to form unique Origin-destination pairs. The Origin-

Destination pairs are hot-encoded, and shown in figure 5. There will be ${}^n\text{C}_2$ pairs possible (where n is the number of nodes/ bus stops), as explained before. [Here, since $n = 14$, as fixed, there will be 91 possible combinations.]

The linear regression algorithm will project the bus usage in terms of number of passengers in a particular route for the next week, and will help the bus transport authorities to better evaluate and distribute resources, in terms of manpower, budget, and time. This way, the demand for a particular route or for a range of buses will make itself known, and can be met accordingly.

2.6 Performance Evaluation of the Model

2.6.1 Mean Squared Error

It is simply the average of the squared difference between the target value and the value predicted by the regression model

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (\text{Eq. 3})$$

MSE = mean squared error

n = number of data points

Y_i = observed values

\hat{Y}_i = predicted values

2.6.2 R² Error

The metric helps us to compare the current model with a constant baseline and tells us how much the model is better. The constant baseline is chosen by drawing a line at the mean of the data.

$$R^2 = 1 - \frac{RSS}{TSS} \quad (\text{Eq. 4})$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

3. Experimental Results

The web-based Application, written in HTML and JS, opens up to the starting page as shown, with the two distinct columns representing page navigation and the opening main page with the content are created with frames.

When the “utilization” option on the left column is clicked, the utilization stub for ticket-issue is brought to the webpage using an iframe and a google form that links to a sheet as the SQL database.

When the left-side “Prediction” option is clicked, the page loads and the user choose the independent parameters for which the system predicts the number of passengers. This is given in figure 6.

Usage Prediction

Usage prediction and reports are the two results obtained. Seven parameters (day, time slot, boarding point, drop point, temperature, precipitation and whether it is a working day or not) are collected from the user for prediction. The number of passengers that can be expected for the given input set is predicted using a simple linear regression algorithm with R square of 0.9 and displayed in an iframe, when the predict button is clicked. The predicted output is shown in figure 7.

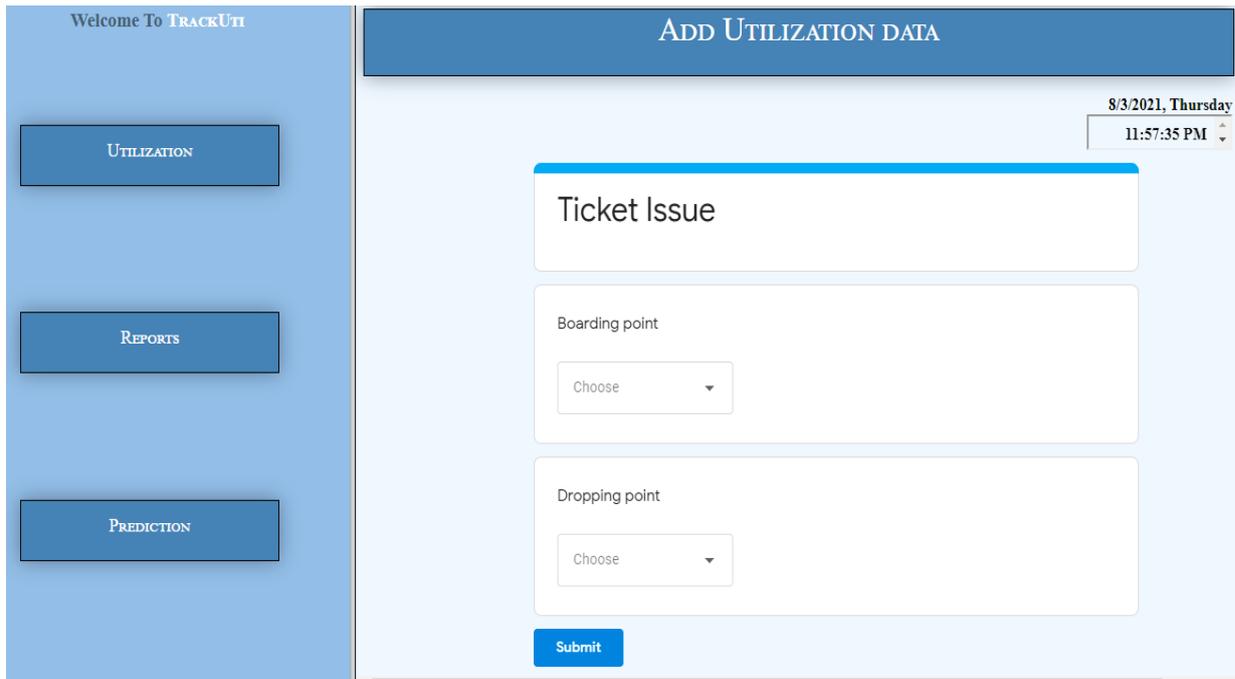


Fig. 6 Utilization Page

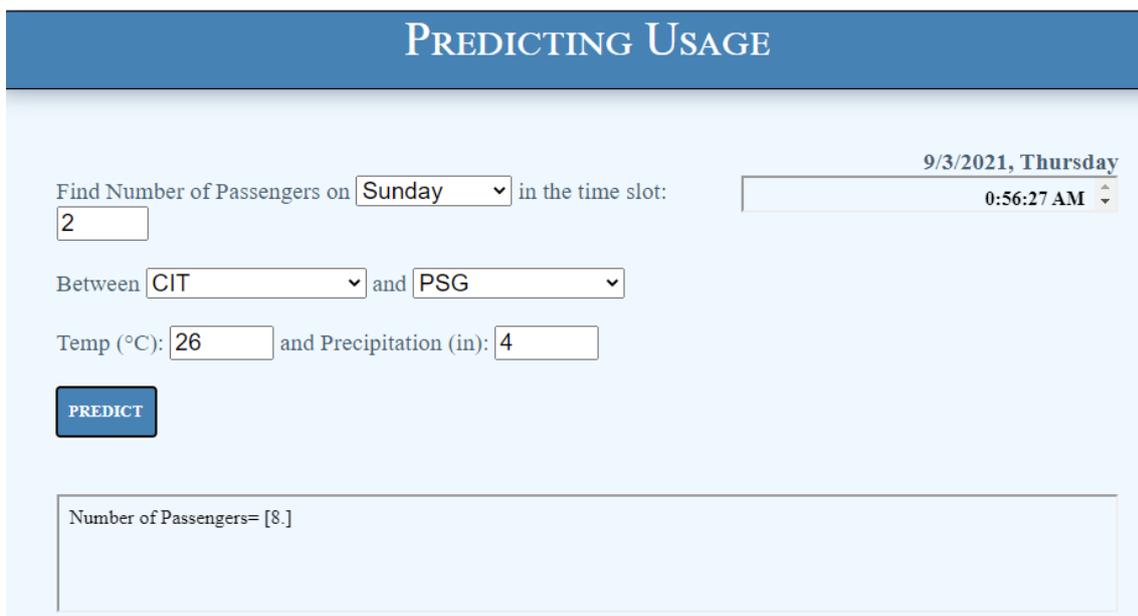


Fig 7. Prediction output

Reports

The relation between the number of passengers per journey and parameters such as precipitation, temperature, working day, etc is represented graphically. The data used for reports is collected from the point of ticket issue. The outputs are in the form of a scatter plot graph and provide a comparison between the number of passengers per journey and the other parameters. D3.js, a javascript library which is used for producing dynamic and interactive data visualizations has been used for the generation of these reports.

Passengers Vs Precipitation:

The independent variable which is precipitation in inches is plotted on the X axis and the number of passengers is plotted on the Y axis. The obtained graph indicates that the number of passengers is inversely proportional to the independent variable- precipitation. It can be seen that the number of passengers is less when the precipitation is high, as in figure 8.

Passengers Vs Temperature:

The X axis has the independent variable, temperature in degree Celsius and the Y axis has the number of passengers on it. The graph indicates that the number of passengers is higher when the temperature is moderate, as given in figure 9.

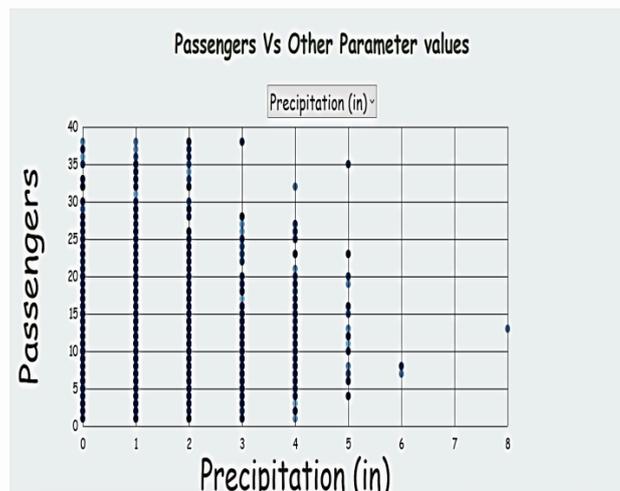


Fig 8. Passengers Vs Precipitation

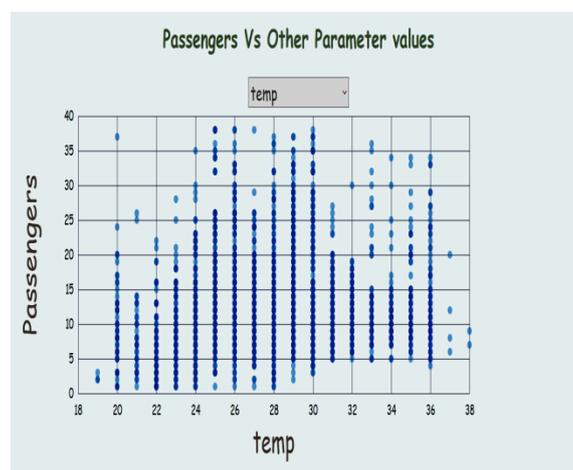


Fig 9. Passengers Vs Temperature

Passengers Vs Time Slot:

The independent variable represented on the X axis of the graph is the time slot. Each slot is of the duration of one hour. 7AM - 10PM have been split into 15 time slots. Bus usage is maximum in the second and third time slot, indicating that these are peak hours, as shown in figure 10.

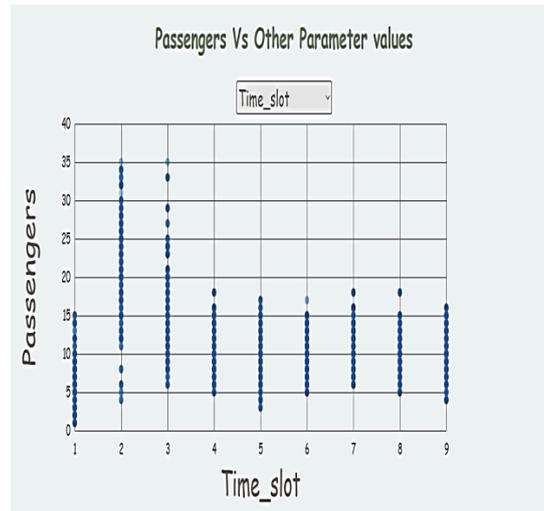


Fig 10. Passengers Vs Temperature

Passengers Vs Working day:

The independent variable taken in this graph is whether it is a working day or not. 0 indicates that it is not a working day whereas 1 indicates that it is a working day. It can be observed that there are relatively a smaller number of passengers on working days, as shown in figure 11.

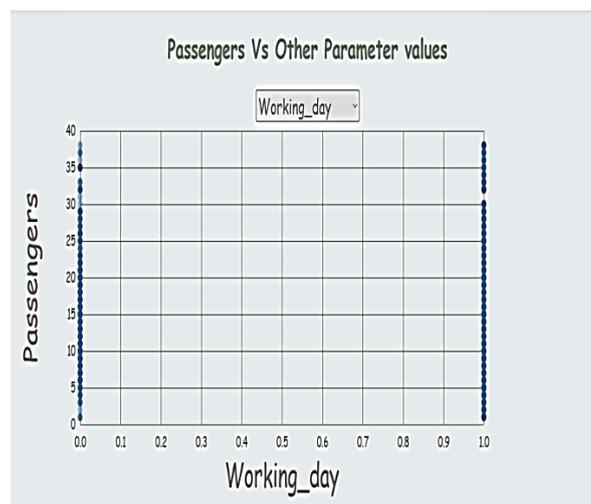


Fig 11. Passengers Vs Working Day

Passengers Vs Days of the week:

The independent variable plotted here is the day of the week. Each day has been represented by a number - 1 for Sunday, 2 for Monday and so on. The inference made is that the number

of passengers is higher on Sundays and Saturdays when compared to the other days of the week, as illustrated in figure 12.

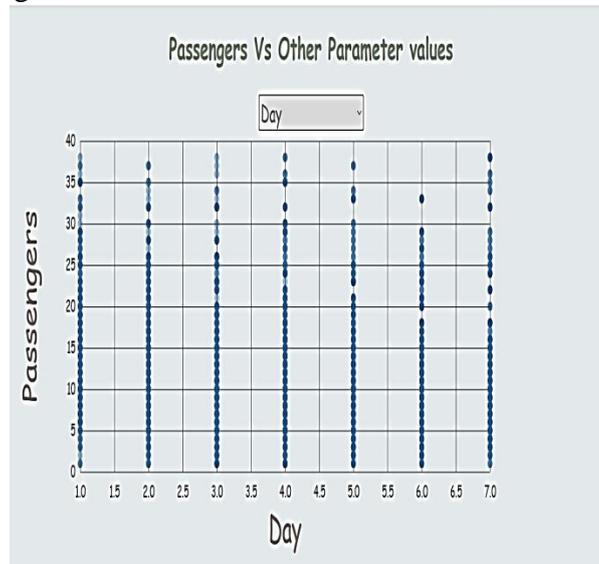


Fig 12. Passengers Vs Working Day

4. CONCLUSION AND FUTURE SCOPE

A lot of factors affect the demand for a service. Analyzing these dependency patterns allows businesses to recognize potential improvements, perform risk analysis, etc. Specifically in the public sector, there are many areas of improvement, one of the dominant ones being the transportation sector.

The application is one method of improving the efficacy of the public bus system in Indian cities, from the coherent analyses of patterns of usage, and the simple prediction of passenger count, thus making it more environmentally friendly, economically viable, and accessible and easy to use for the general public. The project can also be tweaked to suit a variety of businesses with assets that require usage pattern analysis and prediction.

5. REFERENCES

- [1] Fatima, E., & Kumar, R. (2014). Introduction of public bus transit in Indian cities. *International Journal of Sustainable Built Environment*, 3(1), 27-34.
- [2] Patankar, V. M., Kumar, R., & Tiwari, G. (2007). Impacts of bus rapid transit lanes on traffic and commuter mobility. *Journal of urban planning and development*, 133(2), 99-106.
- [3] [Litman, T., & Burwell, D. (2006). Issues in sustainable transportation. *International Journal of Global Environmental Issues*, 6(4), 331-347.
- [4] Pucher, J., Korattyswaroopam, N., & Ittyerah, N. (2004). The crisis of public transport in India: overwhelming needs but limited resources. *Journal of public transportation*, 7(4), 1.
- [5] Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of artificial intelligence in transport: An overview. *Sustainability*, 11(1), 189.
- [6] Zhou, C., Dai, P., & Zhang, Z. (2015, October). Passenger demand prediction on bus services. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 1430-1435). IEEE.

- [7] Book, E., & Ekelöf, L. (2019). A Multiple Linear Regression Model To Assess The Effects of Macroeconomic Factors On Small and Medium-Sized Enterprises.
- [8] Gustafsson A, Wogenius S (2014) Modelling apartment prices with the multiple linear regression model. Thesis, KTH Royal Institute of Technology SCI School of Engineering Sciences
- [9] Freburger, J. K. (1999). An analysis of the relationship between the utilization of physical therapy services and outcomes for patients with acute stroke. *Physical therapy*, 79(10), 906-918.
- [10] Farahnakian, F., Liljeberg, P., & Plosila, J. (2013, September). LiRCUP: Linear regression based CPU usage prediction algorithm for live migration of virtual machines in data centers. In 2013 39th Euromicro conference on software engineering and advanced applications (pp. 357-364). IEEE.
- [11] Ahmadi, A.; Sajadian, N.; Jalaliyan, H.; Naghibirokni, N. Study And Analysis of Optimized Site-selection for Urban Green Space by Using Fuzzy logic: Case Study: Seventh Region of Ahvaz Municipality. *IARS' International Research Journal*, Vic. Australia, v. 2, n. 2, 2012. DOI: 10.51611/iars.irj.v2i2.2012.23.