Weather or Not: Bay Area VTA Ridership Forecasting

Neha Thakur
San Jose State University
neha.thakur@sjsu.edu

Nivedita Venkatachalam
San Jose State University
nivedita.venkatachalam@sjsu.edu

Rutuja Kokate
San Jose State University
rutuja.kokate@sjsu.edu

Saumya Varshney
San Jose State University
saumya.varshney@sjsu.edu

***Abstract*—Combining the data on weather and transit can be an excellent way to develop sustainable public transport systems that are effective. This project proposes a novel way of combining NOAA weather data and VTA transit data to effectively model and predict ridership on VTA stops. The above problem was formulated as a regression problem where the ridership on a particular stop will be predicted using VTA route features, geographical features of the stop, and the weather conditions of that stop on any day. Based on this description regression models like ElasticNet, Decision trees, Random Forest, gradient boosted trees and, XGBoost were selected. Evaluation metrics like Root Mean Squared Error, Mean Absolute Error, Evaluation Variance Score and R2 score were selected to appropriately compare the model. After thoroughly training and evaluating the above models on this data using K-Fold Cross-Validation and hyperparameter tuning the project delivers a tuned Random Forest model with an R2 score of 0.8050. Using this model VTA authorities can effectively predict the ridership on a particular station or stop and make effective route and transit plans.**

***Index Terms*—machine learning, weather data, ridership, cross-validation, ensemble models, hyperparameter tuning**

I. Introduction

The amalgamation of machine learning methodologies with transit data offers an prospect for transforming public transportation infrastructure. We conduct a thorough investigation into the use of machine learning, encompassing basic to sophisticated approaches, including ensemble methods, to evaluate the effect of weather on ridership in the Valley Transportation Authority (VTA) network. Enhancing transportation operations and advancing sustainability requires an understanding of the complex link between weather patterns and user demands. We use a rich dataset covering four years to accomplish this, understanding that larger datasets provide more opportunity for insightful analysis and enhanced model performance. A structured machine learning pipeline that includes several stages, including feature engineering, model construction, exploratory data analysis (EDA), data integration and processing, and assessment, will be used in this work. We want to determine the most efficient methods for forecasting ridership fluctuations in response to weather variations through extensive research with various machine learning models.

II. Problem Statement

In the Bay Area public transportation serves as a backbone for sustainable urban mobility by significantly contributing to the reduction of carbon footprints by limiting private vehicles. However, the system's reliability is compromised by frequent delays because of unpredictable weather patterns, which in turn leads to higher emissions. This is because idle vehicles consume energy without serving transportation needs. The goal of this is to create a predictive model which can forecast transportation ridership by analyzing impacts caused by weather with a goal of strengthening service reliability and sustainability.

III. Literature Survey

The primary goal is to utilize machine learning to predict transportation ridership variations based on weather conditions, aiming to enhance the operational efficiency and sustainability of Bay Area's transport systems.

A. Weather and Public Transportation

Torvela, T. (2024) highlighted the significant impact of weather on public transportation ridership and demonstrated a machine learning approach to predict changes in passenger load. This study underlines the sensitivity of ridership to weather variations and its implications for route planning and schedule optimization. [2]

Pleisch, A. (2023) analyzed automated count data to evaluate urban transport demand changes due to weather conditions in Zurich. The findings suggest similar patterns could be observed in the Bay Area, supporting the relevance of integrating weather data into ridership forecasting models. [3]

B. Particulate Matter and Traffic Data

Yang, J., et al. (2024) utilized spatiotemporal data to predict particulate matter concentrations, offering insights into the relationship between air quality, traffic, and meteorological conditions. This research provides a foundational understanding that can be leveraged to assess the environmental impacts of fluctuating ridership. [4]

C. Machine Learning in Transportation

Kumar, B.N.S., & Swetha Shri, K. (2023) applied machine learning to predict real-time passenger train delays. The methodology and insights from this study are particularly pertinent for addressing similar challenges within bus and light rail systems in the Bay Area. [5]

D. Computational Fluid Dynamics and Health Safety

Yoo, S.J., et al. (2023) explored the risks associated with airborne infections in buses through computational fluid dynamics. This research is important for developing strategies to improve public health safety in public transportation, especially in response to change in ridership due to weather conditions. [6]

IV. Data Collection

The project leverages data from the following sources:

A. VTA Ridership Data

- Source URL:VTA Open Data - Ridership Data [1]
- Data Range: 2014 - 2017
- Data Description: This dataset consists of ridership numbers across various VTA transit services such as VTA bus and light rail. Specific data structure and attributes are described below (Refer Fig 1).

Field	Data Type	Description
Date	Date	Date when data was collected, also referred to as the Transit Day.
Trip ID	Integer	Unique ID number for a given trip. These numbers change with each sign-up cycle (about every three months). The same Trip Id represents the same trip during a given sign-up.
Block	Integer	The Block Number that operates a given trip. The block number represents one train or bus from pull-out to pull-in and is used for all trips operated by that one bus.
Line	Text (7)	The line this data is from. Normally this will be a number
Service	Integer	Abbreviation for the service operated the day this data was collected. 1 = weekday, 2 = Saturday, 3 = Sunday/Holiday, 4 = July 4, 5 = Special Weekday, 6 = Special Saturday, 7 = Special Sunday/Holiday. Only services 1, 2, and 3 are normally used.
Direction Number	Integer	Number assigned to the direction the trip is operating. 0 = North, East, Loop 1 = South, West, Reverse
Direction	Text (15)	Name of the direction the given trip operates.
Pattern	Text (10)	The name of the pattern on the line this trip operates. Patterns are used to identify different variations in trips, such as trips that do not operate the entire length of the line or have some variation id routing from other tips. There are also other reasons for having different patterns.
From Time	Time	The scheduled start time of the trip. Based on the first timepoint observed on the trip. It is possible this time is not on the public schedules.
To Time	Time	The scheduled end time of the trip. Based on the last timepoint observed on the trip. It is possible this time is not on the public schedules.
Start Location	Text (8)	Abbreviation used to identify the first timepoint of the trip. This timepoint may not always be on the public time Guides.
End Location	Text (8)	Abbreviation used to identify the Last timepoint of the trip. This timepoint may not always be on the public time guides
On	Integer	The number of people counted boarding at the stop.
Off	Integer	Number of people counted alighting at the stop.
Stop Id	Integer	Unique ID for each stop. The same ID is used for all lines using the same bus stop. Some locations such as transit centers have multiple stops, each stop will have a unique ID. All Light Rail Stations will have at least two IDs, one is used for each direction.
IVR Number	Integer	This number is used to identify the stop for the various Real Time systems, such as the Bay Area 511 system.
Stop Name	Text (75)	The name of the bus stop.
Seq	Integer	The sequence of the stop along the route in question. This allows the data for a given trip to be sorted in order.

Fig. 1. Ridership Data Description

B. NOAA Weather and Station Data

- Source URL:NOAA Weather and Station Data [7]
- Data Range: 2014 - 2017
- Data Description: The NOAA Weather Data comprises meteorological observations collected across various weather stations globally. It includes detailed records of daily weather conditions such as temperature, precipitation, snowfall, and wind speeds. Additionally, the NOAA Station Data provides crucial metadata for each weather station, including geographic coordinates, elevation, and network affiliations (Refer Table I and II).

TABLE I
NOAA WEATHER DATA

Field	Data Type	Description
Date	Integer	Date in format YYYYMMDD
Tmax	Float	Maximum temperature for a day in degree Celsius
Tmin	Float	Minimum temperature for a day in degree Celsius
Prcp	Float	Precipitation in mm
Station ID	String	ID of the weather station

TABLE II
NOAA STATION DATA

Field	Data Type	Description
Station ID	String	ID of the weather station
Latitude	Float	Latitude of the weather station
Longitude	Float	Longitude of the weather station

V. EXPLORATORY DATA ANALYSIS

This section discusses all the unique features and trends present in the data. This includes doing a statistical analysis on each feature of the dataset. Visualizations are generated that help in visualizing the distributions and understanding the relationship between the features of the data.

Fig. 2 represents a bar graph was plotted to analyze the distribution of data on a week day basis for the years 2014 to 2017.

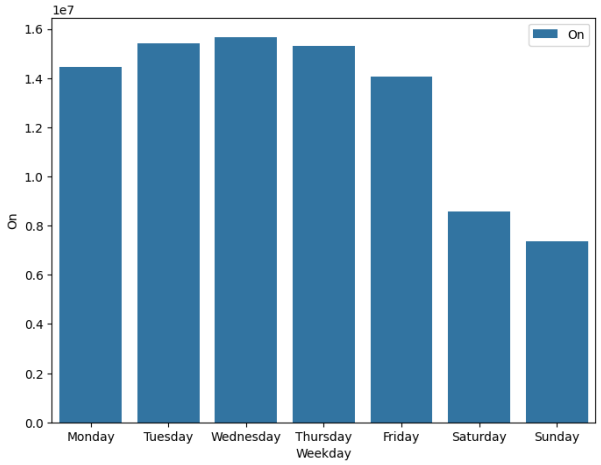

Fig. 2. Weekly Distribution of Ridership

Fig. 3 represents "Daily distribution of ridership categorized by year" graph which suggests consistent patterns in transit use with variances that may be attributed to specific events or service changes underscoring the importance of temporal factors in ridership prediction.

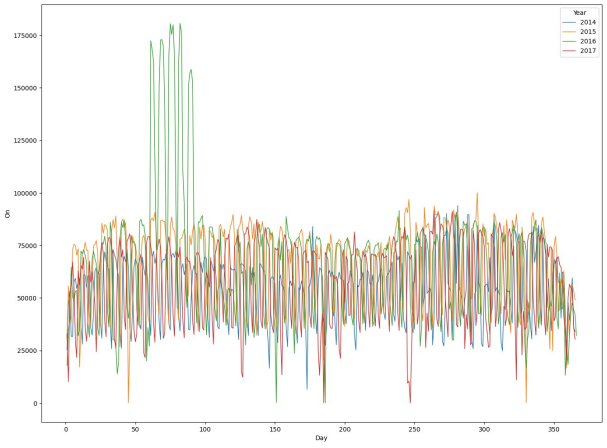

Fig. 3. Daily Distribution of Ridership Categorized by Year

A. Correlation Heat Map

Fig. 4 represents a correlation heat map,

- There is a negative correlation between latitude and longitude. This is due to the geographical structure of the Bay area and the distribution of stations and stops in the VTA network.
- There is a positive correlation between Tmax and Tmin. As the climate gets warmer or cooler throughout the year Tmax and Tmin also increase and decrease together.
- There is a slightly negative correlation between Tmax and precipitation because precipitation often lowers the maximum temperature of that day. The effect on minimum temperature is little.

- There is a positive correlation between on and off. If more people get on at a station, more people also get off at another station.
- The only correlation between feature and target variable is between line and on and line and off. This correlation was not apparent in the original dataset. After doing aggregations on the data, this correlation became more significant.

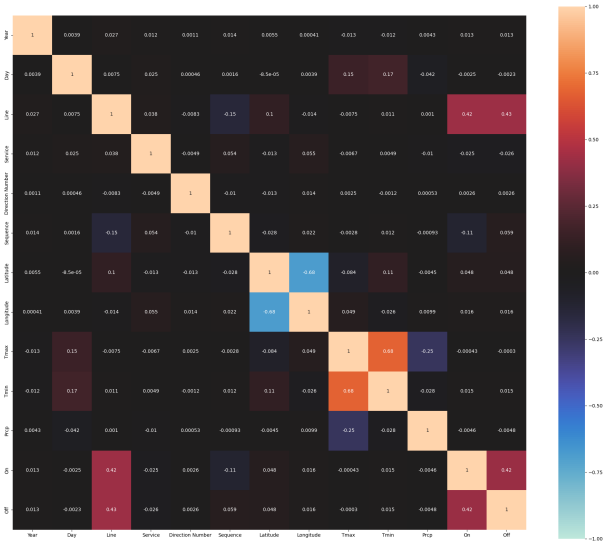

Fig. 4. Correlation Heat Map

VI. Data Preprocessing

- **Ridership Data:** There were many textual features in the original dataset provided by VTA. Some of the features were Stop Name, Start Location and End Location. These features didn't provide much information about the daily ridership based on the documentation provided by VTA for this dataset.
  Some fields like Block, Line, Stop Id were considered string-like fields since these IDs contained numbers with commas. These were converted to integer data types after removing the comma. The date field was a timestamp at 12AM every day. This date field was converted to a date field, since our target is to predict ridership on a day-to-day basis.
  There were some data samples where the value of "On" was less than zero and "Off" was zero. Those rows were effectively capturing net flow on these days. Net flow is essentially the difference between "On" and "Off". Since it is difficult to estimate the correct On and Off values for these samples, these rows were removed from the dataset.
  The ridership data was also aggregated on Date, Line, Service, Direction Number, Sequence and Stop Id, and "On" and "Off" were summed together to provide a concise dataset.
- **Station Data:** The station data extracted from NOAA contained weather stations from across the globe. Since the ridership data provided by VTA contains stops from only the Bay Area, weather stations that lie within the Bay Area vicinity were picked from the NOAA stations dataset. The latitudes were filtered to lie between 36.974922 and 37.558388, and the longitudes were filtered to only lie between –122.17364 and –121.54903.
- **Weather Data:** Data samples from the Weather Data provided by NOAA were filtered to only include recording from weather stations that lie in the Bay Area (from the above step). After this, the weather stations were filtered to only include stations that provided weather data for more than 95% of all the days in the period between 2014 to 2017. This was done to ensure that high quality weather data was available for all the stops in the VTA network. This also helps in reducing the need to fill the null values with either median or mean.
- **Combining all the datasets:** Based on the latitude and longitude of each stop in the VTA network, nearest weather stations were identified. This weather station would provide the weather features for that station on all the dataset days. The weather features were integrated into the main dataset by joining on Date and Stop Id.

VII. Modeling

To effectively investigate the association between weather and transit ridership, this project uses a variety of machine learning models. The train set consists of data from 2014 to 2016 and test set contains data of year 2017. Each model was first trained with default hyperparameters using the dataset. After reporting the results for this model, the hyper parameters of that model were tuned using k-fold cross validation and grid search. For simple models like linear regression and decision tree, 5-fold cross validation was used to determine the best hyperparameters. For ensemble models like random forest, gradient boosted trees and XGBoost 3-fold cross validation was used to determine the best hyperparameters.

A. Linear Regression (Elastic Net with PCA)

Elastic Net with parameters in conjunction with Linear Regression was used because of their ease of interpretation and simplicity. Model performance was optimized by reducing dimensionality in the presence of weather data by combining PCA with Linear Regression and Elastic Net regularization, with n_components set to 7, 8, 9 (Fig 4). In contrast, PCA was used in the absence of weather data, with n_components adjusted to 5, 6, 7 (Fig 5), to better capture pertinent aspects and improve model interpretability. The best hyper-parameters are highlighted in **bold** (Refer Table III and IV).

TABLE III
Hyperparameter with weather data

n_components (PCA)	7, 8, **9**
alpha (elastic net)	**0.5**, 1.0
l1_ratio (elastic net)	0.3, 0.5, **0.7**

TABLE IV
Hyperparameter without weather data

n_components (PCA)	5, 6, **7**
alpha (elastic net)	**0.5**, 1.0
l1_ratio (elastic net)	0.3, 0.5, **0.7**

B. Decision Tree Regression

Non-linear correlations and interactions between weather variables were captured using decision trees, which shed light on intricate feature relevance and decision bounds. Decision Trees

intrinsic interpretability gives us access to the model's decision-making process, which makes them useful for figuring out how different weather conditions affect ridership. Two sets of hyper-parameters are used. Best hyperparameters were from set 2 and are highlighted in **bold**.(Refer Table V and VI)

TABLE V
HYPERPARAMETERS WITH WEATHER DATA

Set	criterion	max_depth	min_samples split	min_samples leaf
Set 1	squared_error, friedman_mse, poisson	20, 40, None	2	1
Set 2	**squared_error**, friedman_mse, poisson	**20**, 40, None	**14**	**7**

TABLE VI
HYPERPARAMETERS WITHOUT WEATHER DATA

Set	criterion	max_depth	min_samples split	min_samples leaf
Set 1	squared_error, friedman_mse, poisson	20, 40, None	2	1
Set 2	squared_error, **friedman_mse**, poisson	**20**, 40, None	**14**	**7**

C. Random Forest Regression

Random forest can handle non-linear correlations and interactions between weather variables while reducing overfitting and improving forecast accuracy. This model improves generalization and reduces variance by averaging the predictions of many decision trees trained on bootstrapped subsets of the data. The model also shows which weather factors most affect ridership changes. Four sets of hyper-parameters are used. Best hyper-parameters were from set 2 and are highlighted in **bold**. (Refer Table VII and VIII)

TABLE VII
HYPERPARAMETERS WITH WEATHER DATA

Set	criterion	n_estimator	max_depth	max_features
Set 1	squared_error, poisson	50	20	1.0
Set 2	squared_error, **poisson**	**50**	**100**	**sqrt**
Set 3	squared_error, poisson	10	70	1.0
Set 4	squared_error, poisson	10	100	sqrt

TABLE VIII
HYPERPARAMETERS WITHOUT WEATHER DATA

Set	criterion	n_estimator	max_depth	max_features
Set 1	squared_error, poisson	50	20	1.0
Set 2	squared_error, **poisson**	**50**	**100**	**sqrt**
Set 3	squared_error, poisson	10	70	1.0
Set 4	squared_error, poisson	10	100	sqrt

To maximize Random Forest's ability to capture intricate weather-transit interactions while minimizing overfitting, its hyper-parameters were chosen to achieve a compromise between model complexity and generalization capacity. Through testing various parameters such as criterion, number of estimators, maximum depth, and maximum features, we were able to determine configurations that enhanced prediction accuracy and emphasized the significance of meteorological conditions in changes in transit ridership.

D. Gradient Boosted Trees

Because of their capacity to progressively tune weak learners to reduce prediction errors and enhance model performance, gradient boosted trees were chosen. Gradient Boosted Trees successfully capture the complex linkages and interactions between meteorological variables and ridership patterns by concentrating on minimizing residual errors. By prioritizing the most informative features, the model's adaptive learning technique improves prediction accuracy and robustness Four sets of hyper-parameters are used. Best hyper-parameters were from set 1 and are highlighted in **bold**. (Refer Table IX and X)

TABLE IX
HYPERPARAMETERS WITH WEATHER DATA

Set	learning rate	subsample	n_estimator	max_depth	max_features
Set 1	0.001, **0.1**	**0.8**, 1.0	**50**	**20**	**1.0**
Set 2	0.001, 0.1	0.8, 1.0	50	100	sqrt
Set 3	0.001, 0.1	0.8, 1.0	10	70	1.0
Set 4	0.001, 0.1	0.8, 1.0	10	100	sqrt

TABLE X
HYPERPARAMETERS WITHOUT WEATHER DATA

Set	learning rate	subsample	n_estimator	max_depth	max_features
Set 1	0.001, **0.1**	0.8, **1.0**	**50**	**20**	**1.0**
Set 2	0.001, 0.1	0.8, 1.0	50	100	sqrt
Set 3	0.001, 0.1	0.8, 1.0	10	70	1.0
Set 4	0.001, 0.1	0.8, 1.0	10	100	sqrt

E. XGBoost

Because of its cutting-edge functionality, scalability, and effectiveness in managing big datasets and intricate interactions, XGBoost Regression was selected. XGBoost is an efficient implementation of gradient boosting that produces extremely accurate predictions by iteratively creating a sequence of decision trees to minimize a given loss function. The model is particularly suited for large-scale predictive modeling jobs because it uses parallelized computation and sophisticated regularization approaches to reduce overfitting and increase computational efficiency. Furthermore, XGBoost provides flexibility in hyperparameter optimization and model tuning, enabling us to customize the model's functionality and meet project-specific needs. Our goal is to improve prediction accuracy and obtain a deeper understanding of how weather affects transit passengers by utilizing the capabilities of XGBoost Regression. This will enable data-driven decision-making and resource allocation in transit planning and operations. The best hyperparameters are highlighted in **bold** (Refer Table XI and XII).

TABLE XI
HYPERPARAMETERS WITH WEATHER DATA

max_depth	3, 5, **7**
learning_rate	**0.1**, 0.01, 0.001
subsample	0.5, **0.7**, 1

TABLE XII
HYPERPARAMETERS WITHOUT WEATHER DATA

max_depth	3, 5, **7**
learning_rate	**0.1**, 0.01, 0.001
subsample	0.5, **0.7**, 1

VIII. EVALUATION

To evaluate the performance of models, we computed the RMSE, MAE, EVS, and R2 score for both Train and Test datasets. In this section, we describe the performance of each of the models.

A. Evaluation Metrics

- RMSE (Root Mean Squared Error): It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} w_i (o_i - p_i)^2}{\sum_{i=1}^{N} w_i}}$$

- MAE (Mean Absolute Error): It is a statistical metric that measures the average size of errors in a set of predictions by calculating the arithmetic average of the absolute differences between predicted and actual values.

$$\text{MAE} = \frac{\sum_{i=1}^{N} |o_i - p_i|}{N}$$

- EVS (Explained Variance Score) : In a regression model, the explained variance is summarized by R2. This value represents the proportion of the variance in the response variable that can be explained by the predictor variable(s) in the model.

$$\text{EVS} = 1 - \frac{Var\{y_{true} - y_{pred}\}}{Var\{y_{true}\}}$$

- R2 Score : An R-Squared value shows how well the model predicts the outcome of the dependent variable. R-Squared values range from 0 to 1. An R-Squared value of 0 means that the model explains or predicts 0% of the relationship between the dependent and independent variables.

$$\text{R2} = 1 - \frac{SS_{\text{RES}}}{SS_{\text{TOT}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

IX. Results

A. Metrics evaluation without weather data integration

Decision Tree model without any hyper-parameter tuning had the highest R2 score on the train set, with a MAE of 3.3869. But the model over-fitted on the training data, and had a very low R2 score on the test set. Random Forest model with tuned hyper-parameters generalized really well on the test, achieving an R2 score of 0.7028. In fact, that model had the best EVS score, RMSE and MAE score on the test set among all the model. Only XGBoost with its default parameters and a tuned XGBoost model had a better R2 score than the tuned Random Forest model. The agreement between the R2 score on train and test data was quite high for the XGBoost models. This means that the training metrics reported for this model were pretty good representation of its predictive power on unseen data. ElasticNet and its hyper-parameter tuned version had the lowest EVS and R2 score than all the other models, indicating that this model was not capable enough to learn the complex relationships present in this dataset. (Refer Fig. 5).

B. Metrics evaluation with weather data integration

Decision Tree model without any hyper-parameter tuning had the highest R2 score on the train set, with a MAE of 0.2688. But the model over-fitted on the training data, and had a very low R2 score on the test set. XGBoost on the other hand with its default parameters had the second highest R2 score on the test set, and tuning the hyper-parameters didn't improve its performance further. Random Forest model with tuned hyper-parameters had the best R2 score on the test set, and was better than all the other models for RMSE, MAE and EVS metrics on the test set. ElasticNet had the worst performance on the train set and test across all the evaluation metrics, even after tuning. (Refer Fig. 6)

	Train				Test			
	RMSE	MAE	EVS	R2	RMSE	MAE	EVS	R2
XGBoost	26.523086	8.759610	0.714517	0.714416	25.600350	9.023668	0.719192	0.719070
XGBoost (tuned)	27.365872	9.540672	0.696081	0.695979	25.946126	9.408564	0.711575	0.711430
Random Forest (tuned)	22.720250	5.953004	0.790532	0.790438	26.330031	7.258850	0.703317	0.702827
Random Forest	30.934791	10.948782	0.611621	0.611510	28.081744	10.524957	0.662087	0.661970
Decision Tree (tuned)	22.986859	6.575651	0.785585	0.785491	28.711754	7.653365	0.647169	0.646633
Gradient Boosted Trees (tuned)	20.424988	5.685599	0.830740	0.830641	29.326463	7.677398	0.631653	0.631340
Gradient Boosted Trees	36.005872	12.988031	0.473813	0.473701	32.339984	12.464130	0.551808	0.551682
Decision Tree	16.310750	3.386988	0.892012	0.891998	37.581580	9.581290	0.395364	0.394580
ElasticNet (tuned)	45.302237	17.412751	0.166949	0.166847	42.778246	17.029905	0.215704	0.215573
ElasticNet	45.772322	17.472761	0.149568	0.149467	43.452505	17.146198	0.190722	0.190650

Fig. 5. Evaluation Results without Weather data

	Train				Test			
	RMSE	MAE	EVS	R2	RMSE	MAE	EVS	R2
Random Forest (tuned)	21.921486	5.753937	0.805010	0.804914	25.282108	7.034397	0.726423	0.726011
XGBoost	26.297237	8.858497	0.719361	0.719259	25.880109	9.397562	0.712901	0.712896
XGBoost (tuned)	27.074637	9.531467	0.702517	0.702415	26.069883	9.491517	0.708766	0.708670
Random Forest	30.815540	10.932296	0.614610	0.614499	28.111466	10.522375	0.661372	0.661254
Gradient Boosted Trees (tuned)	14.952138	4.751572	0.909340	0.909241	28.511713	7.577102	0.651773	0.651540
Decision Tree (tuned)	21.516920	6.321250	0.812143	0.812049	28.993712	7.754166	0.640138	0.639659
Gradient Boosted Trees	36.044313	12.996771	0.472675	0.472577	32.329988	12.469461	0.552098	0.551959
Decision Tree	4.004799	0.268897	0.993490	0.993489	38.072360	10.010802	0.379542	0.378665
ElasticNet (tuned)	45.300691	17.420682	0.167005	0.166904	42.781109	17.040605	0.215589	0.215468
ElasticNet	45.784973	17.482478	0.149098	0.148996	43.472548	17.163334	0.189971	0.189904

Fig. 6. Evaluation Results with Weather data

C. Evaluating the Final Model

Based on the results obtained from the above sections, a tuned Random forest model trained on ridership data with weather features provided the best R2 score on the test dataset compared to all the other models. The final step was to train this tuned random forest model using the entire dataset from year 2014 to year 2017. Below fig. are the results from training this final model (Refer Table XIII, Fig. 7, 8 and 9).).

TABLE XIII
EVALATION METRICS FOR FINAL MODEL

Root Mean Squared Error	21.7712
Mean Absolute Error	5.7359
Explained Variance Score	0.8051
R2 score	0.8050

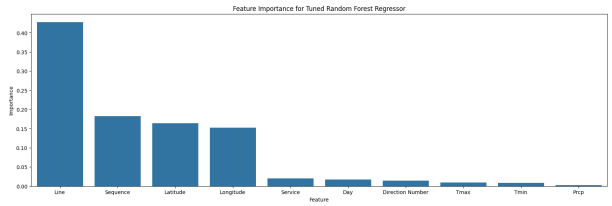

Fig. 7. Feature importance for the final Random Forest model

X. Innovation

By aggregating the data on features like date, line, service, direction number, sequence and stop id, some correlations between features and target features were enhanced. This increased the predictive capacity of the various machine learning models, thereby improving the prediction of ridership on a particular stop on a particular day. By training various models, tuning their

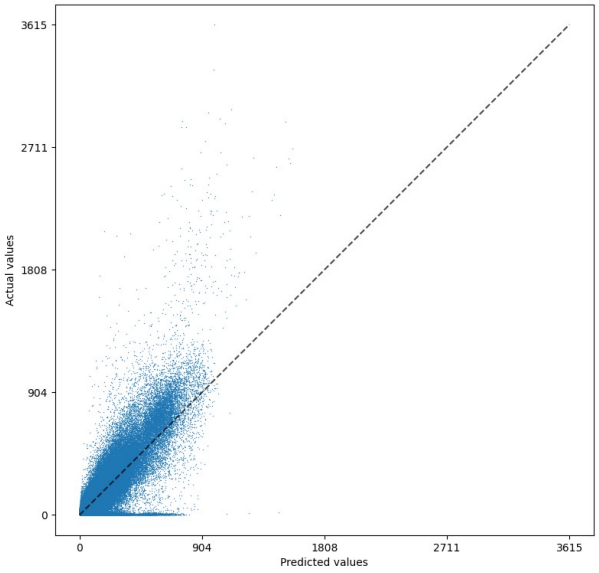

Fig. 8. Actual vs Predicted scatter plot for the final model

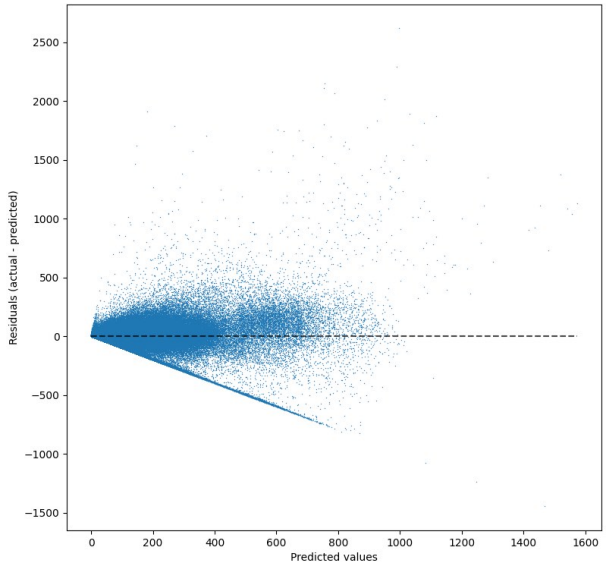

Fig. 9. Scatter plot comparing Residuals with Predicted values

hyperparameters, and comparing various metrics, a pretty robust model was selected that generalized really well on the test set. Integrating the local weather conditions of stops with this dataset improved the predictive power of some machine learning models, and helped factor in location specific features in the model. Using these innovative methods, VTA can analyze alternative routes and schedules to improve their carbon footprint and lower energy costs while being mindful of the local weather conditions.

XI. Sustainability

The Bay Area VTA Ridership Forecasting Project helps to increase the sustainability and efficiency of public transportation, which supports the Sustainable Development Goals (SDGs) of the United Nations. This project helps in making cities inclusive, safe and sustainable by improving public transportation systems. The project helps address climate change by improving the predictability of public transportation. Reduced dependency on fossil fuels and decreased carbon emissions are results of efficient public transportation networks. Improving energy efficiency of public transport systems by optimizing operations and decreasing unnecessary trips helps in ensuring access to affordable and clean energy.

XII. Technical Difficulty

In our project on Bay Area VTA Ridership Forecasting, we encountered significant technical difficulties due to large size of the VTA dataset, which was further expanded by the integration of extensive weather data. The initial dataset alone was approximately 8 GB, and with additional weather data, it became unmanageable on our local system equipped with only 8-16 GB of RAM. To address this challenge, we leveraged Google Colab Pro which provided access to higher CPU and Memory, enabling us to handle the larger data efficiently. Furthermore, we utilized our lab's resources to perform necessary data transformations, modeling and exploratory data analysis (EDA). These adjustments were crucial for managing the computational demands of our project ensuring that we could continue our analysis without compromising on the depth and quality of our insights.

XIII. Lessons Learned

A. Management of Large Datasets and Computational Resources

Working with big data brings about substantial problems concerning storage, processing, and computer resources. This project has served as a good opportunity to learn about scalable ways of dealing with big datasets. Cloud computing services were utilized while managing or analyzing large volumes of information.

B. Integration of Different Data Sets

Bringing together different pieces of information was indeed difficult, and prepossessing the data needs to be done with care. Each dataset needed careful cleaning and prepossessing, which took time but ensured accuracy later in the project. We were able generate useful insights by ensuring that all our data sets were compatible hence reliable models could be developed thereafter without any challenges along the way.

C. Effects of Data Imputation on Model Performance

The performance of machine learning models is greatly affected by imputation techniques used to address missing or incomplete data. The selection of imputation method alongside its parameters greatly impacted the model's accuracy and generalization ability, bringing the need for different strategies in evaluating them against predictive quality.

D. Models Selection and Performances

While it may seem that the most complex models are always better than simple ones, we found that ensemble methods outperform simplistic models. This shows us how important it is to select a model according to its appropriateness for the task, considering aspects like model complexity, interpretability and computational cost.

XIV. Conclusion and Future work

This project effectively integrates NOAA weather data with VTA's ridership data and as such has significant implications for optimizing public transit practices in the era of global warming. This project can be a helpful tool for VTA authorities to save costs and make an ecological impact. Ensemble Models provide

the best results when this problem is modeled as a Regression task.

In future scope, these models still have a lot of scope of improvement. There is still some correlation between the residuals and predicted values in our best model. There is a lot of scope in improving the R2 score for such models. More data can help improve the model performance, but VTA has not published extensive data for ridership on the VTA network after 2017. Upon visualizing the feature importance of many ensemble models in this project, it can be seen that the weather features had a meaningful but weak predictive power. Neural Networks can possibly capture the latent features in this dataset, and utilize them further to provide an even better prediction of ridership of VTA stops.

REFERENCES

[1] https://s3-us-west-2.amazonaws.com/gisopendataportal/Ridership+Data+Description.pdf

[2] Torvela, T. (2024). Weather impact on public transport ridership: Predicting passenger load using machine learning. Retrieved from Theseus.fi.

[3] Pleisch, A. (2023). The impact of weather on urban transport demand: An analysis of automated count data in Zurich. Retrieved from ETH Zurich Research Collection.

[4] Yang, J., Shi, L., Lee, J., & Ryu, I. (2024). Spatiotemporal prediction of particulate matter concentration based on traffic and meteorological data. Transportation Research Part D: Transport and Environment, Elsevier. Retrieved from ScienceDirect.

[5] Kumar, B.N.S., & Swetha Shri, K. (2023). Real-time passenger train delay prediction using machine learning: a case study with amtrak passenger train routes. Retrieved from IRJMETS.

[6] Yoo, S.J., Yamauchi, S., Park, H., & Ito, K. (2023). Computational Fluid and Particle Dynamics Analyses for Prediction of Airborne Infection/Spread Risks in Highway Buses: A Parametric Study. Fluids. Retrieved from MDPI.

[7] https://noaa-ghcn-pds.s3.amazonaws.com/index.html#csv/

XV. APPENDIX

A. CRediT statement

The following table shows the CRediT statement (Contributor Roles Taxonomy) highlighting the roles and responsibilities shared by team members for the success of this project (Refer Table XIV).

TABLE XIV
TEAM ROLES AND RESPONSIBILITIES

Name	Roles and Responsibility
Neha Thakur	Data collection and EDA, Feature Engineering, Model Development and Hyperparameter tuning – Decision trees and Random Forest.
Nivedita Venkatachalam	Data Cleaning, Performance metrics, Model Development and testing – Linear regression with Elastic Net, EDA on weather data, Report Documentation.
Rutuja Kokate	Data Transformation, Model Development and testing- XGBoost, Model Evaluation, AR video, Elevator pitch Video, Presentation slides.
Saumya Varshney	Hyperparameter Tuning, Model Evaluation, Model Development and testing- Gradient Boosted Trees, Maintaining Agile Board.

B. Term Project Rubric

- **Code Walkthrough** - In project demonstration the team will provide a detail code walkthrough.
- **Google Drive link:** - The code files are shared in this Drive link to meet the criteria. Since the file size is large. Few code files are around 4 MB and the pickle file size for the final model is around 2.26 GB. - Google Drive
- **Presentation Skills** - In the project demonstration, the team will efficiently present proposed ideas and results with a structured PowerPoint, carefully planning time allocation through rehearsal.
- **Discussion QA** - The team will address questions and discussions
- **Demo** - A concise and comprehensive demonstration will be provided, supplemented with an inference script for clarity and depth though pre-saved best performing model.
- **Visualization** - Visualizations will be strategically used in the presentation slides and report to offer enhanced insights and understanding into the data and results to the audience and readers.
- **Report** - The report will primarily focus on detailed textual explanations of the ideas and solutions, with limited use of images for visual representations of results. It adheres to IEEE format with grammar checked for conciseness using Grammarly.
- **Version Control** - All code and versions are saved on GitHub for version control purposes and is publicly available. GitHub
- **Relates to sustainability** - The objective of this project is to promote sustainability by testing the hypothesis for sustainable public transport. The hypothesis aims to align with below Sustainable Development Goals (SDGs) of the United Nations: Climate Action, Sustainable Cities and Communities, Affordable and Clean Energy
- **Lessons learned** - Lessons learnt are captured in Presentation slides and in-depth in the report.
- **Prospects of winning competition / publication** - VTA Ridership dataset integrated with weather data is not available online on any competition platform. Planning to upload it on Kaggle and when there is competition, confidently this project will rank quite high.
- **Innovation** - The key learnings through this project are: Overcame dataset inconsistencies through feature aggregation, Employed diverse machine learning models to enhance prediction accuracy, Analyzed qualitative data to discern weather's impact on travel behavior.
- **Evaluation of performance** - The project extensively makes use of Evalution metric related for Regression problem to measure the performances of models.
- **Teamwork** - All tasks were divided between team members for effective project management using JIRA tasks.
- Technical difficulty - Since the Data is big in size storing the data and running models need huge compute units to run.
- **Practiced pair programming** - Pared programming was implemented between team members using Google collab and Google SharePoint with shared access. Git for version control and collaboration, Zoom meetings from team meeting.
- **Practiced agile / scrum (1-week sprints)** - 1-week Sprints were carried out through scheduled meetings to ensure high quality and manageable work. The minutes of meetings are captured for each sprint. Epics and tasks were assigned to team members for efficient use of resources and time. Jira Board
- **Used Grammarly / other tools for language?** - Grammarly

was used to ensure the content is grammatically accurate and clear.

- **Slides** - The slides are a comprehensive pitch of the idea and employed to demonstrate and provide effective presentation to the audience capturing concise details of the project.
- **Saving the model for quick demo** - The model is saved using python pickle package. The pickle file with trained model will be loaded in the inference script to provide demonstration of model performance and prediction.
- **Used LaTeX** - Employed Overleaf to comply with the report format with IEEE standards. Latex file is provided with the submission
- **Used creative presentation techniques** - Invideo AI was used to create Elevator Pitch video. This is an AI text-to-video platform.https://invideo.io/ , https://youtu.be/LdZe8xCfNbA
- **Literature Survey** - The Literature survey comprehensively covers the existing work related to sustainable VTA transportation using weather data. They are mentioned in the report's literature survey section to enable the reader with background. The references were cited wherever necessary.