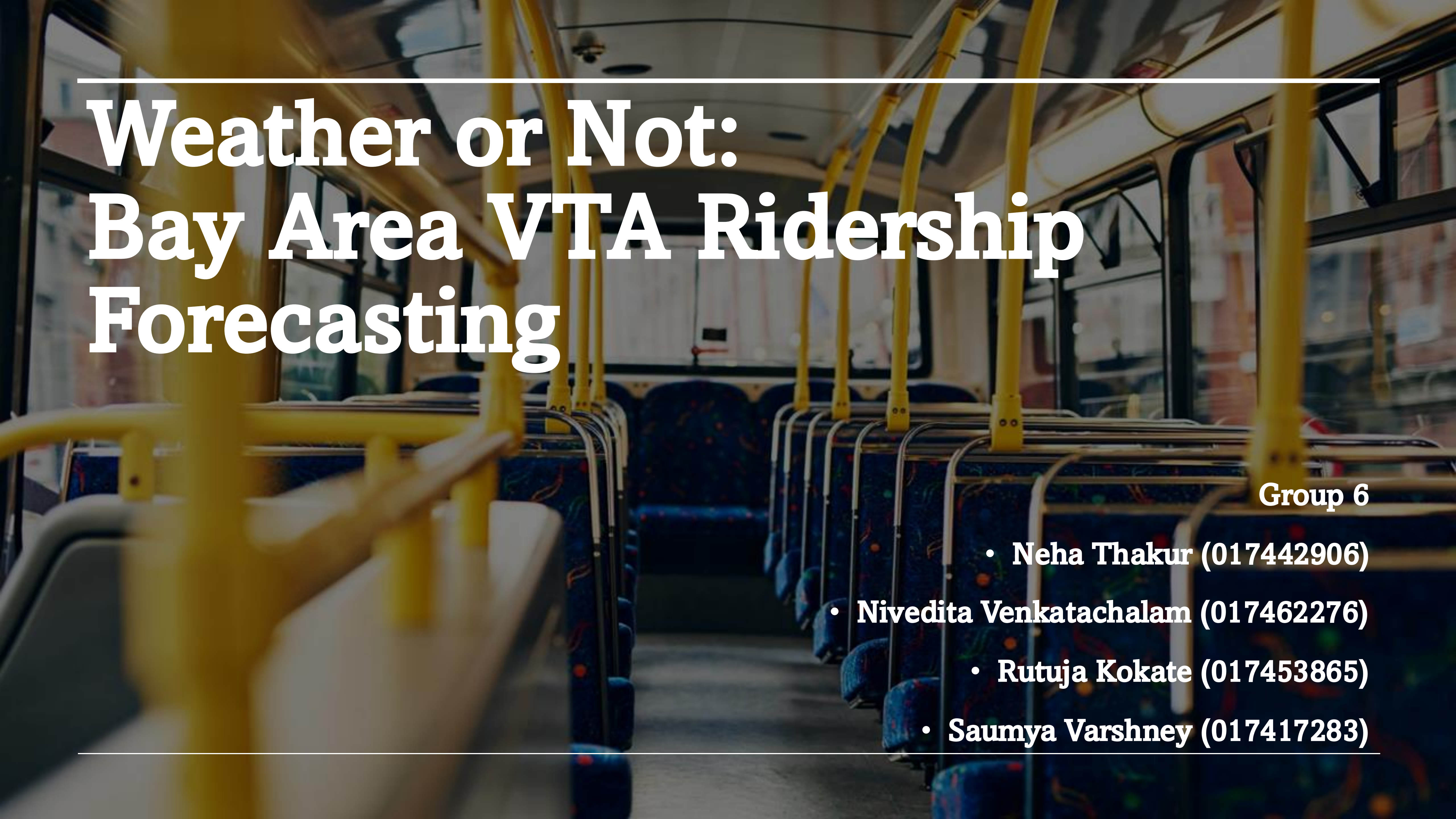


Weather or Not: Bay Area VTA Ridership Forecasting

Group 6

- Neha Thakur (017442906)
- Nivedita Venkatachalam (017462276)
- Rutuja Kokate (017453865)
- Saumya Varshney (017417283)

What is VTA?

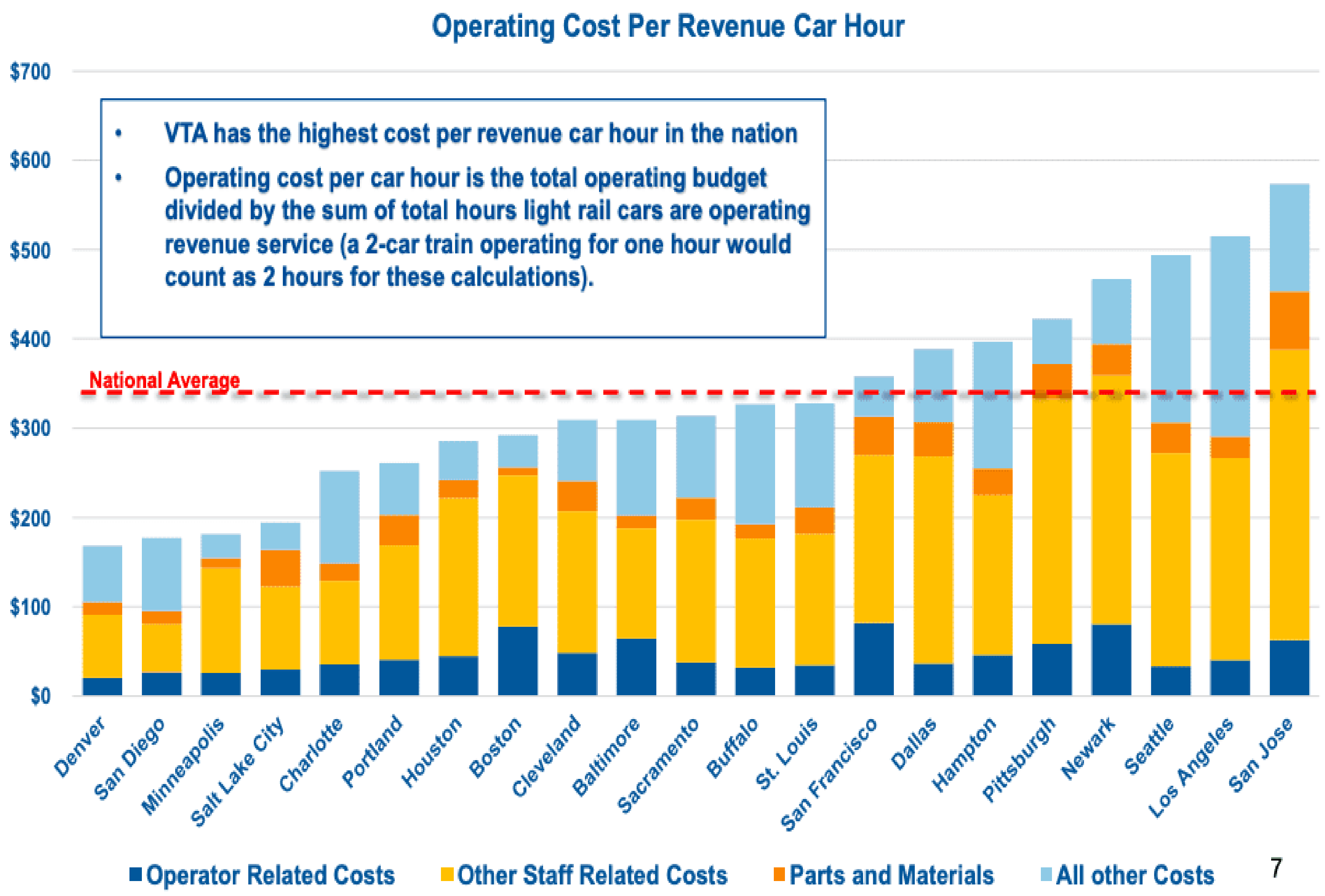
Valley Transportation Authority, is responsible for managing public transportation in the Bay Area, and has observed a notable surge in ridership in recent years.

Weather conditions can fluctuate ridership demand.

Source: https://www.vta.org/blog/second-year-strong-ridership-growth#:~:text=VTA%20bus%20ridership%20increased%20by,growth%20was%2015%25%20in%202023.

Why VTA?

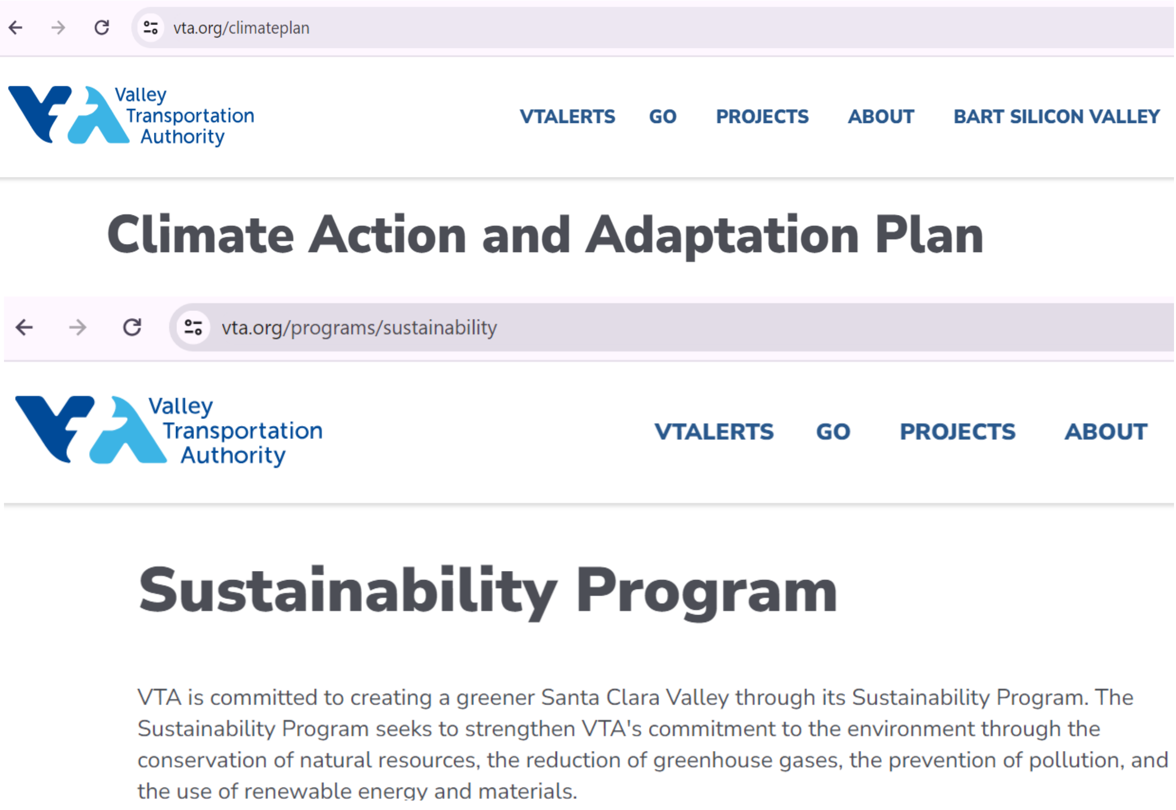
- VTA's operating costs rank fifth highest nationally, 35% above the average.
- An average passenger vehicle emits 4.6 metric tons of CO_2 annually.



Source: https://sanjosespotlight.com/san-jose-silicon-valley-transit-agency-vta-future-may-be-falling-off-the-rails-light-rail-train/ https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle#:~:text=typical%20passenger%20vehicle%3F-,A%20typical%20passenger%20vehicle%20emits%20about%204.6%20metric%20tons%20of,8%2C887%20grams%20of%20CO2.

Goal is to..

- Leveraging predictive power of Machine learning to
  - Identify how weather conditions impact ridership.
  - Suggesting optimization of VTA's/Transport Schedules thereby reducing greenhouse emissions.
  - Saving cost, energy and CO_2 emissions



Source - https://www.vta.org/climateplan https://www.vta.org/programs/sustainability

Dataset

Ridership Dataset **is** sourced from https://data.vta.org/

Weather Data: https://noaa-ghcn-pds.s3.amazonaws.com/

Source: https://dot.ca.gov/cal-itp/california-minimum-general-transit-feed-specification-gtfs-guidelines-v2_0#:~:text=The%20General%20Transit%20Feed%20Specification,produced%20by%20U.S.%20transit%20agencies.

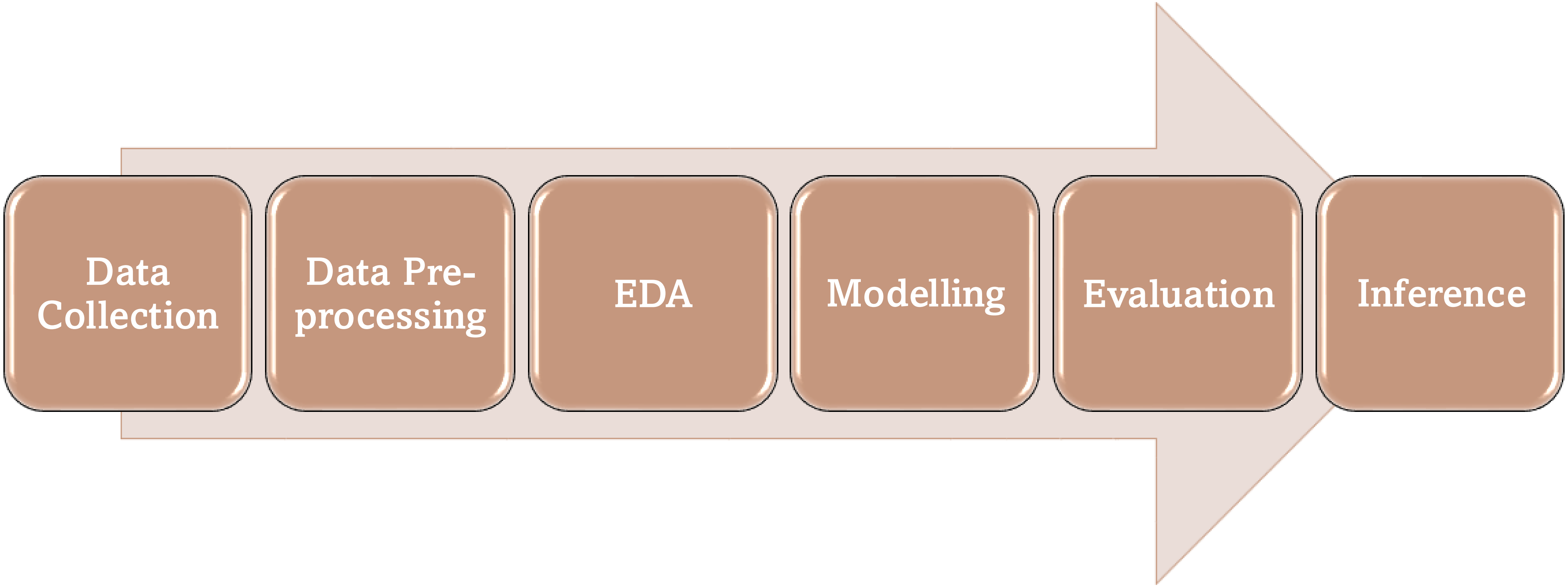
Innovation

Overcame dataset inconsistencies through feature aggregation and Weather data Integration.

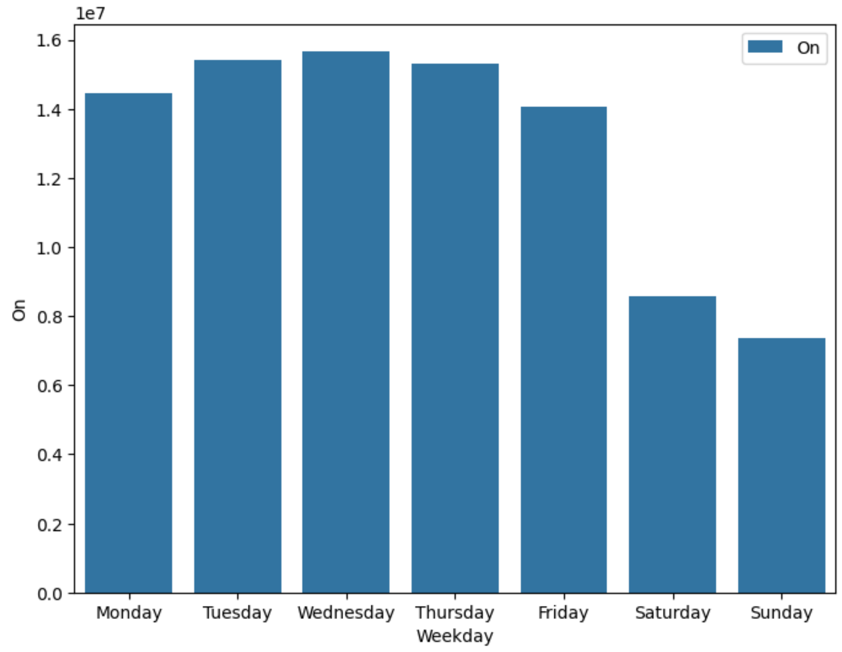
Employed diverse machine learning models to enhance prediction accuracy.

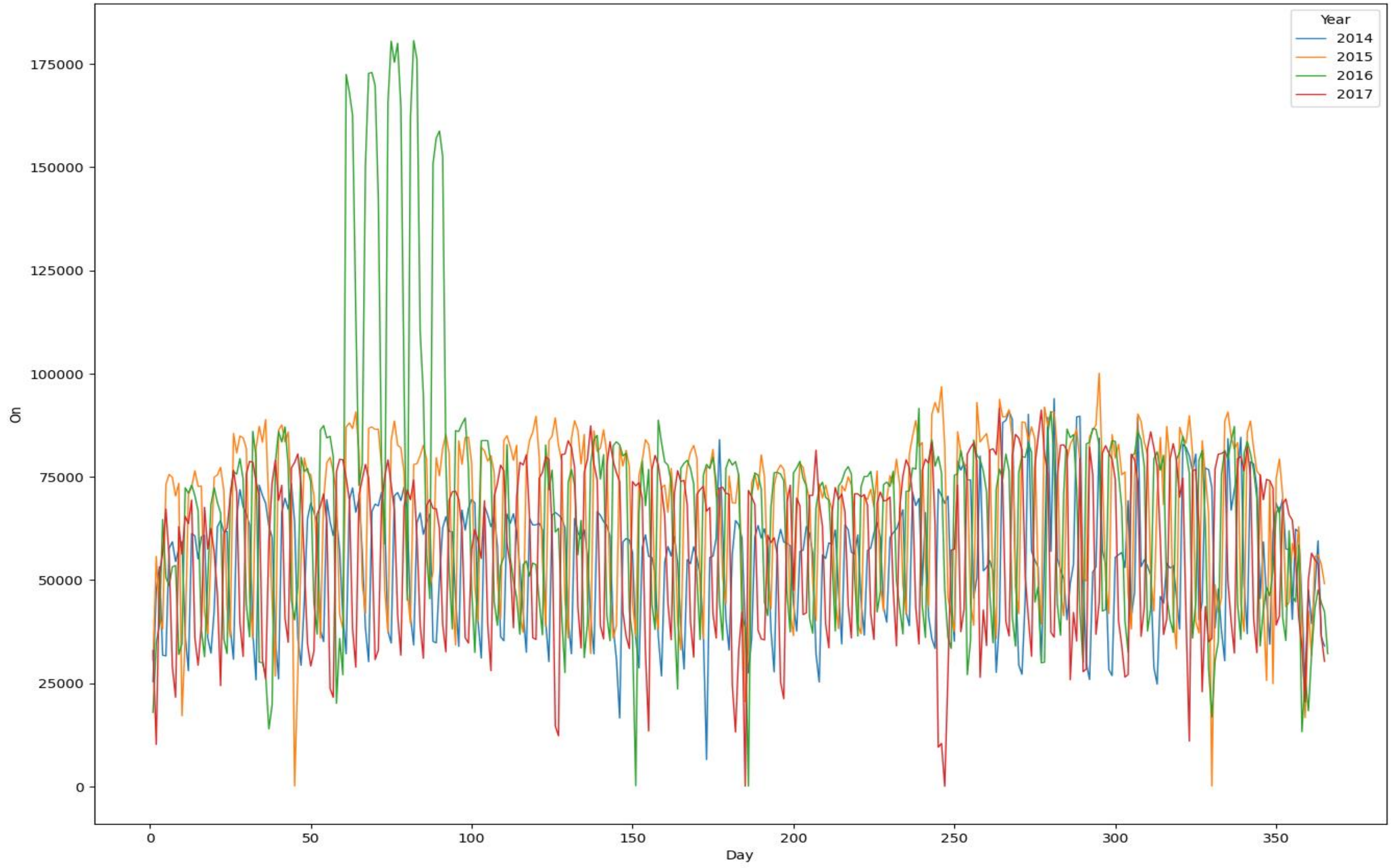
Analyzed qualitative data to discern weather's impact on travel behavior.

Methodology



Exploratory Data Analysis





Correlation Matrix

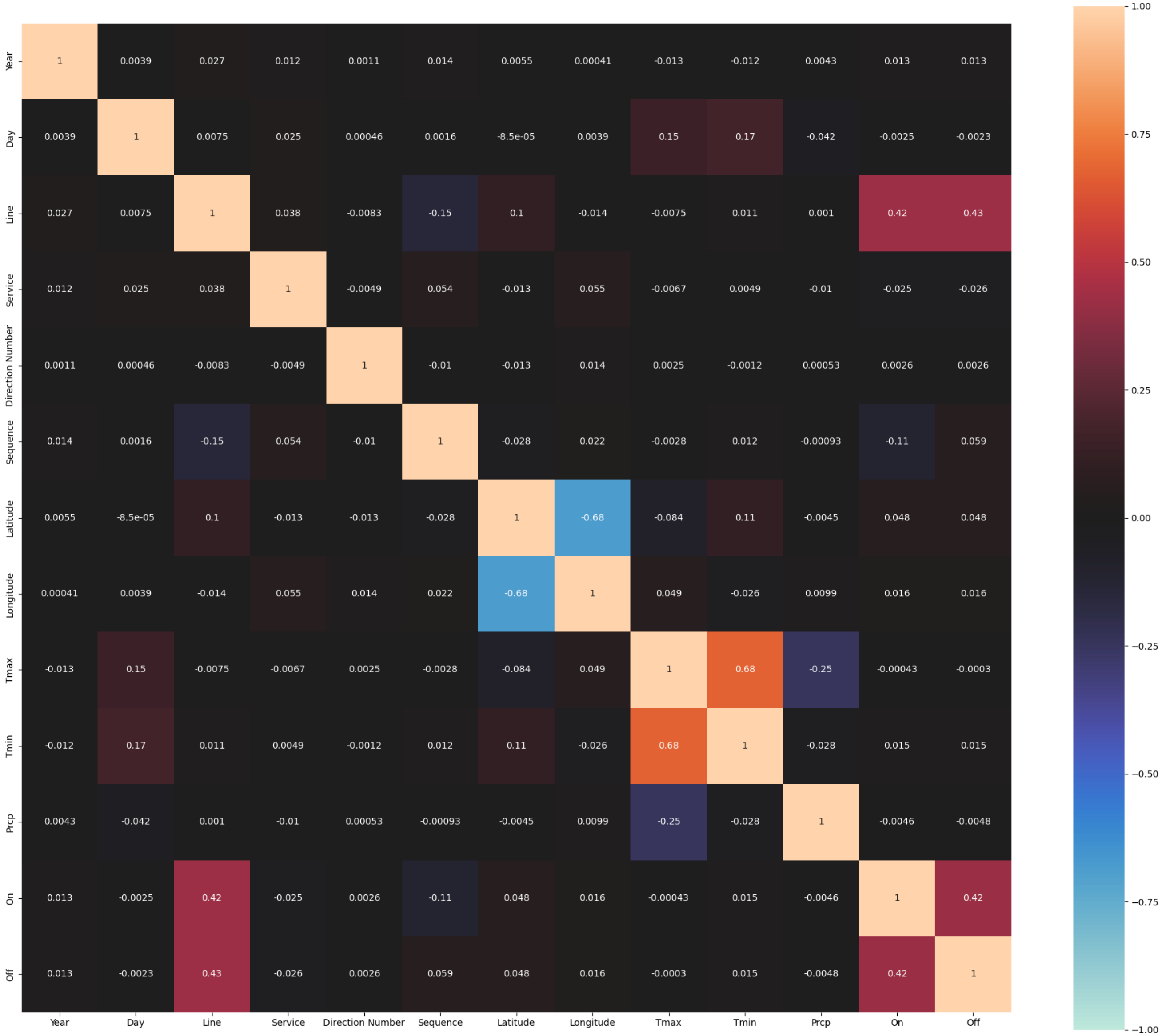
Negative correlation:
latitude vs. Longitude

Positive correlation:
Tmax vs. Tmin

Slight negative correlation:
Tmax vs. Precipitation

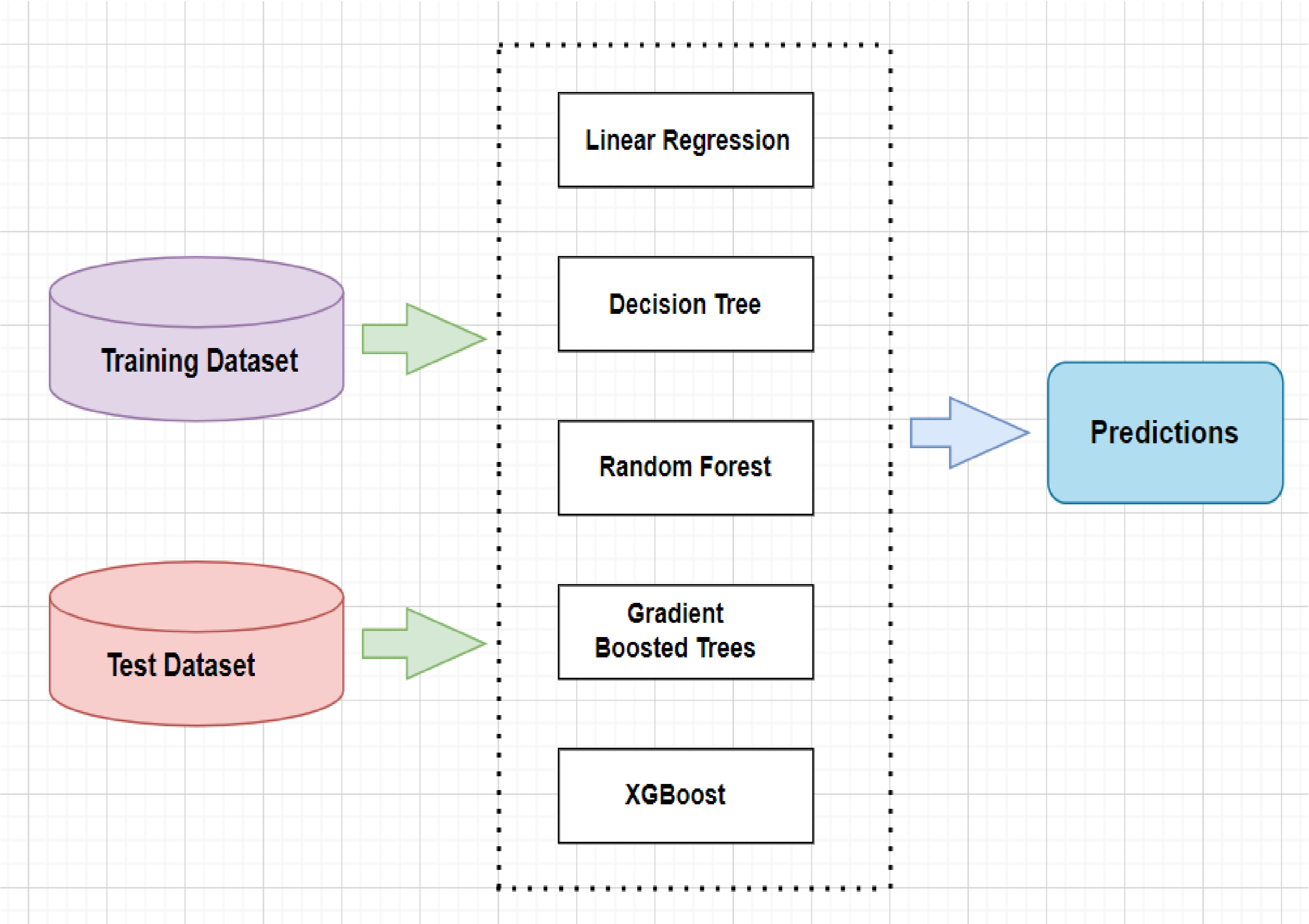
Positive correlation:
on vs. Off

Significant correlation:
line vs. on/off



Modelling

To predict the impact of weather on VTA ridership, we utilized a diverse array of regression models spanning from basic to sophisticated techniques.



Results and Comparison of Model performance
(Without Weather)

	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

Results and Comparison of Model performance

(With Weather)

	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

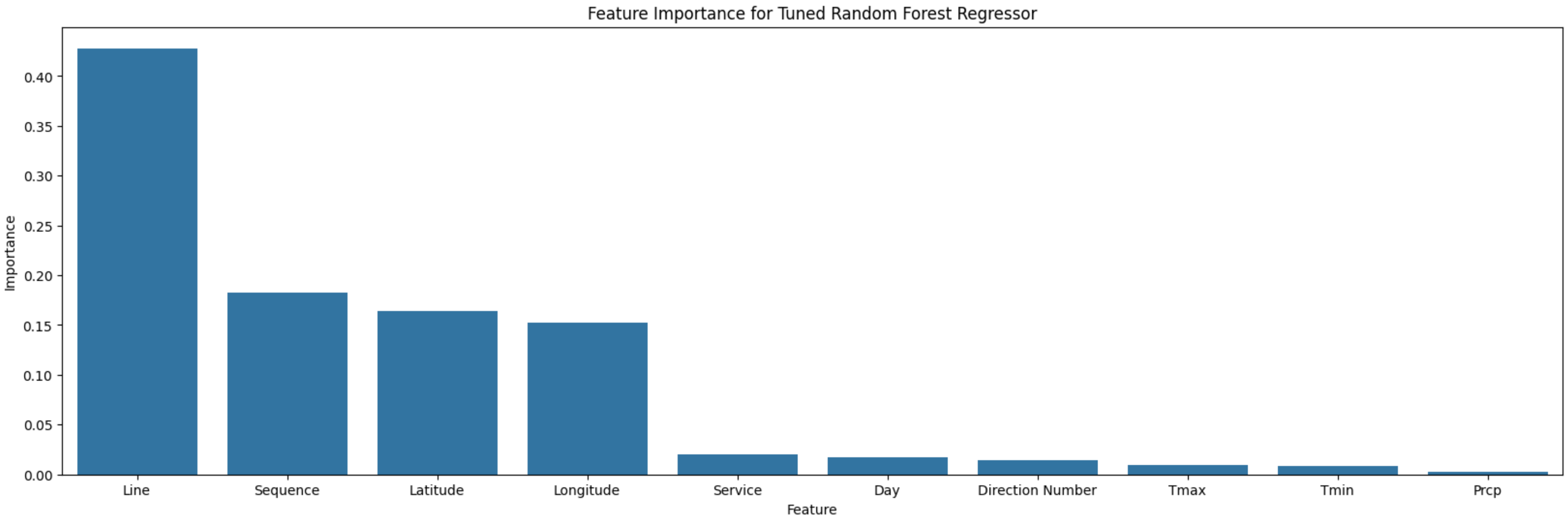
Best Performing Model is…..

	Train	Test
MAE	5.75	7.03
RMSE	21.9	29.28
EVS	0.80	0.72
R2	0.80	0.72

- Source: https://medium.com/almabetter/xgboost-a-boosting-ensemble-b273a71de7a8

Final Model - Random forest with tuned hyperparameters

RMSE	21.7712
MAE	5.7359
EVS	0.8051
R2 Score	0.8050




Code Walkthrough

Demo Time!

Lesson's Learnt

Key learnings:

1. Dealing with Huge datasets and compute resource requirements.
2. Integrating Multiple Datasets.
3. Data Imputations and its impact on model's performance.
4. Comparing model performance across various metrics.
5. Ensemble methods are sophisticated and yield much better results.

Conclusion

- VTA Ridership prediction model 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.

Thank You