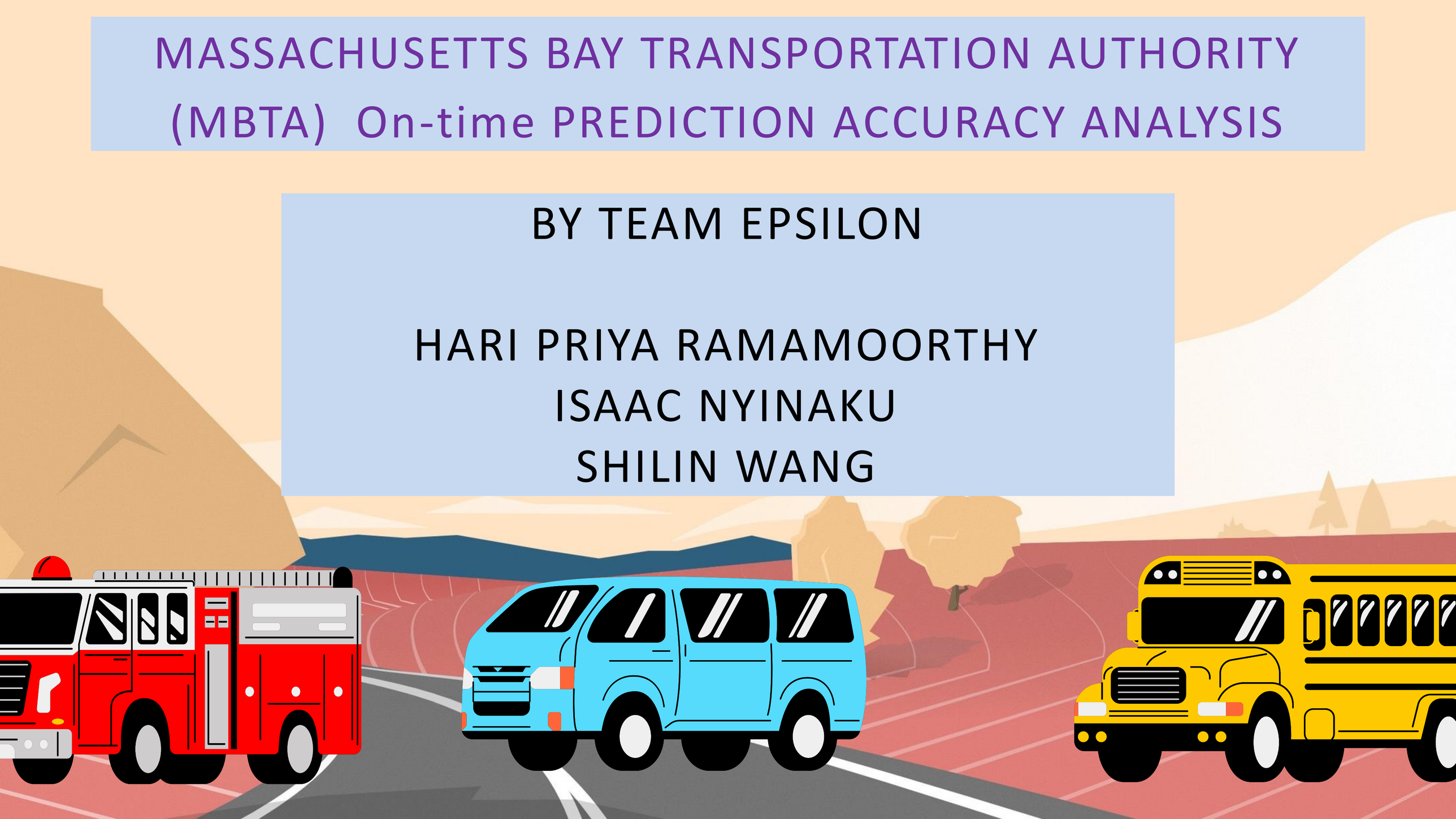
MASSACHUSETTS BAY TRANSPORTATION AUTHORITY (MBTA) On-time PREDICTION ACCURACY ANALYSIS

BY TEAM EPSILON

HARI PRIYA RAMAMOORTHY
ISAAC NYINAKU
SHILIN WANG



OBJECTIVE

Business Problem Statement :

- MBTA operates many transport lines (Bus, Red, Green, Orange, & Blue lines).
- Aims to improve MBTA On-Time prediction model to provide accurate predictions to improve operational efficiency and ridership.

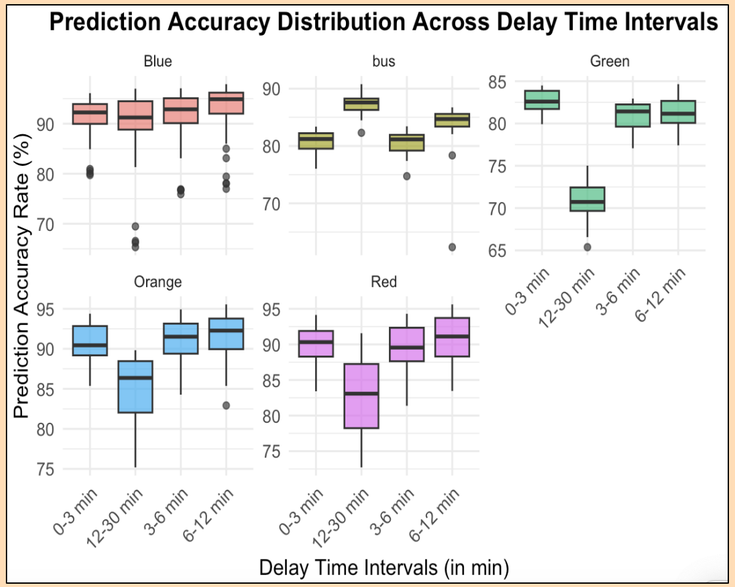
Objective of the Analysis: Recommend significant factors to add in MBTA prediction model to improve it's accuracy.

Datasets Used: MBTA Prediction Accuracy, Seasons, and Ridership Volume from MBTA Opendata portal.

Methodology:

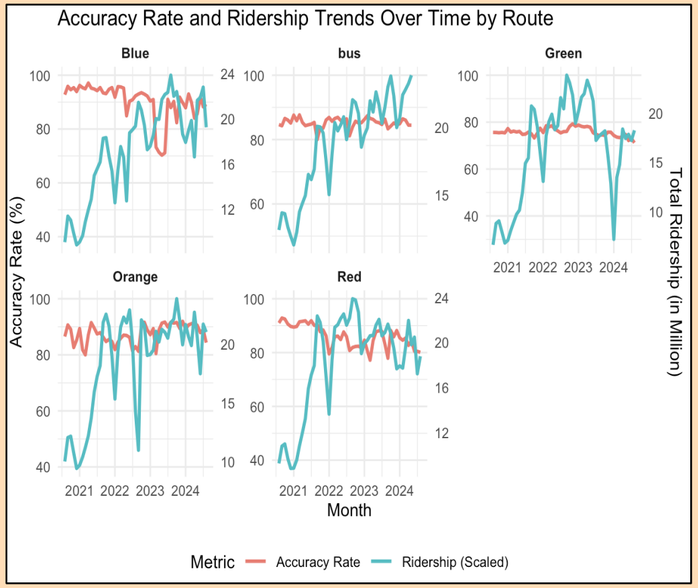
1. Data Manipulation & Integration
2. EDA & Hypothesis Testing
3. Linear Accuracy Prediction Model
4. Feature Selection – Lasso & Step-wise Regression

EDA & Hypothesis Testing



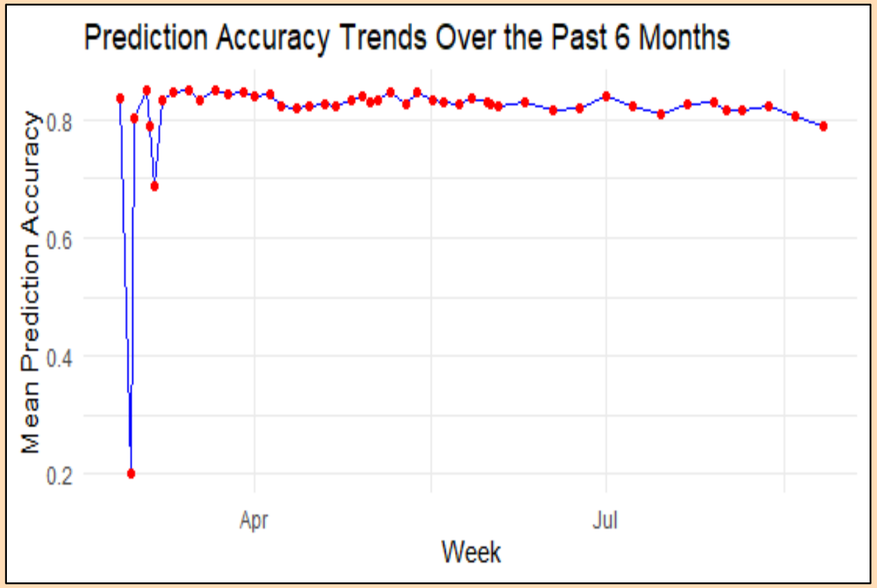
EDA Insight : Prediction accuracy varies across Route
Statistical Test : ANOVA Test
H0: No difference between routes
H1: Significant difference in prediction accuracy across different transportation routes.

Result: Prediction Accuracy varies across routes



EDA Insight : Ridership and Prediction Accuracy exhibits similar trend relationship .
Statistical Test : Pearson's Correlation Test between Accuracy rate and ridership volume.
H0: Correlation=0
H1: Correlation is not equal to 0

Result: Significant negative correlation of -0.298 exists.



EDA Insight : No variation in Prediction accuracy across past 6 Months
Statistical Test : Chi-Square Independence Test on Accuracy Vs Time period
H0: Prediction accuracy is not independent of Time.
H1: Prediction accuracy is independent of Time.

Result: Prediction accuracy is independent of Time.

Linear model

- Linear regression model:
  - **Y-Var:** Accuracy
  - **X-Var: R**oute, Ridership, and Month
- **Identified Significant Predictors:** Route , Ridership
- **Identified Insignificant Predictor:** Month
- **Adjusted R-Squared:** 0.6384 = 64%

Table-1 : Linear Model Summary

```
> summary(lm_model)

Call:
lm(formula = accuracy_rate ~ route_id + month_numeric + total_ridership,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.682  -1.738   0.366   2.484   8.234

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.204e+01  1.077e+00  85.482  < 2e-16 ***
route_idbus     -2.568e+00  1.912e+00  -1.343 0.180604
route_idGreen   -1.472e+01  9.279e-01 -15.862  < 2e-16 ***
route_idOrange  -3.341e+00  9.080e-01  -3.679 0.000289 ***
route_idRed     -4.646e+00  1.012e+00  -4.591 7.13e-06 ***
month_numeric02  7.801e-01  1.328e+00   0.588 0.557341
month_numeric03 -4.302e-01  1.210e+00  -0.355 0.722560
month_numeric04  1.772e+00  1.431e+00   1.239 0.216733
month_numeric05  1.186e+00  1.335e+00   0.889 0.375122
month_numeric06  9.745e-01  1.320e+00   0.738 0.461234
month_numeric07  3.352e-01  1.306e+00   0.257 0.797559
month_numeric08 -9.599e-01  1.159e+00  -0.828 0.408420
month_numeric09  1.966e-01  1.386e+00   0.142 0.887303
month_numeric10  1.168e+00  1.407e+00   0.830 0.407422
month_numeric11  1.650e+00  1.436e+00   1.150 0.251453
month_numeric12  8.660e-01  1.177e+00   0.736 0.462598
total_ridership -2.530e-05  1.128e-05  -2.243 0.025794 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.902 on 239 degrees of freedom
Multiple R-squared:  0.6611,    Adjusted R-squared:  0.6384
F-statistic: 29.14 on 16 and 239 DF,  p-value: < 2.2e-16
```

LASSO & STEPWISE : Feature Selection

Table-2 : LASSO Regularization : Shrink coefficients to 0.

```
> ## PLOT COEFFICIENTS
> print(coef(lasso_model))
17 x 1 sparse Matrix of class "dgCMatrix"
                            s1
(Intercept)       8.899792e+01
route_idbus       .
route_idGreen    -1.057279e+01
route_idOrange    .
route_idRed      -5.280684e-01
month_numeric02   .
month_numeric03   .
month_numeric04   .
month_numeric05   .
month_numeric06   .
month_numeric07   .
month_numeric08   .
month_numeric09   .
month_numeric10   .
month_numeric11   .
month_numeric12   .
total_ridership -1.913739e-05
>
```

Table-3: Step-wise Model: Optimal AIC

Table: Stepwise Regression: Steps, AIC, and Variables Added/Removed

Step	AIC	Variable Added/Removed
1	713.4887	
2	703.6829	- month_numeric

LASSO & Stepwise Results (Table-2,3):
month/seasonal pattern does not influence prediction accuracy.

Model Comparison (Table-4):
Linear regression **Adj.R.Squared remains unchanged** after removing "Month," confirming its redundancy.

Table-4: Model Comparison

Table: Comparison of Models: Adjusted R-squared and RMSE

Model	Adj..R.squared	Train.RMSE	Test.RMSE
Linear Regression	0.638431932912848	3.770193	5.842114
Stepwise Regression	0.637477500404615	3.861065	5.561925
Lasso Regression	NA	3.819269	5.546328

Recommendations & NEXT STEPS

Key Insights from Hypothesis and Feature Selection Tests:

- **Influential Factors:** Route and ridership volume significantly influence MBTA's ETA, explaining ~64% of the variance.
- **Seasonality Impact:** No discernible seasonal patterns affect the ETA of transport lines.

Recommendations for MBTA Improvement:

- **Route-Specific Prediction Models:** MBTA should have different On-Time prediction models across transport lines.
- **Ridership Volume as a Predictor:** Incorporate ridership volume as a core variable in On-time prediction models to improve accuracy and reliability.

Further Research:

- **Traffic-Aware Prediction Algorithms:** Explore models that factor in real-time traffic conditions for enhanced On-time predictions.

References

1. Bluman, A. G. (2014). Elementary statistics: A step by step approach (9th ed.). McGraw-Hill.
2. MBTA. (n.d.). MBTA Blue Book Open Data Portal. Retrieved from https://mbtahttps://mbta-massdot.opendata.arcgis.com/massdot.opendata.arcgis.com/
3. Massachusetts Bay Transportation Authority (MBTA). (n.d.). MBTA Rapid Transit and Bus Prediction Accuracy Data. Retrieved from https://mbta-massdot.opendata.arcgis.com/
4. Massachusetts Bay Transportation Authority (MBTA). (n.d.). MBTA Seasonality Data. Retrieved from https://mbta-massdot.opendata.arcgis.com/datasets/a2d15ddd86b34867a31cd4b8e0a83932_0/explore
5. Chen, M., Liu, X., Xia, J., & Chien, S. (2004). Predicting bus arrival time on the basis of global positioning system data. *Journal of Transportation Research Board*, 1885(1), 98-106. Retrieved from Researchgate

Q&A

Thank You !!