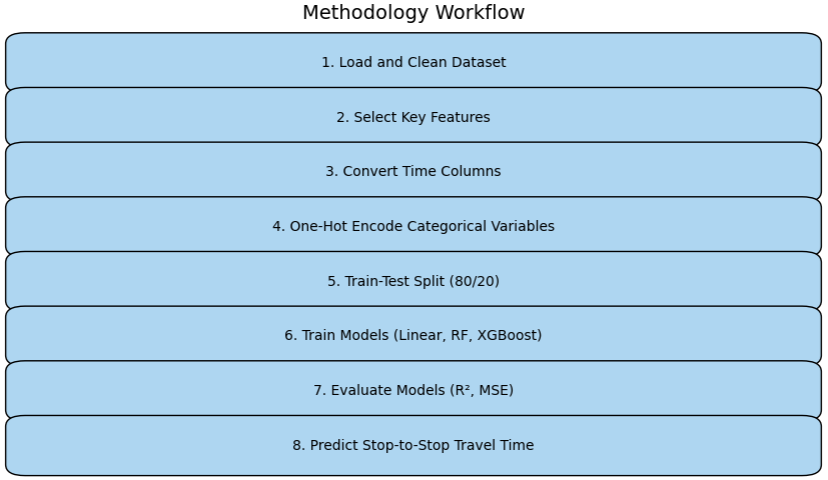
Predicting WRTA Bus Travel Time

Final Project – Data Science 2010

By: Daniel Kwan



Goal:

To predict actual WRTA bus travel time between stops using historical data found on Kaggle.

What are the questions?:

Can Travel time be predicted using route, time of day, speed, and stop data?

How do different regression models compare in accuracy?

Dataset Overview

- Rows: 115,147
- Features: 44
- Target Variable: Running Time Act

17. TIME_PERIOD_SORT: Numeric code for sorting the time periods.
18. SORT_ORDER_1: Sort index for sequence of observations.
19. SORT_ORDER_2: Secondary sort order (usually 0).
20. TIMEPOINT_ID_1: ID for the origin stop of this segment.
21. TIMEPOINT_ID_2: ID for the destination stop of this segment.
22. TIMEPOINT_NAME_1: Text name of the origin stop.
23. TIMEPOINT_NAME_2: Text name of the destination stop.
24. STOP_ID_1: Alternate ID for the origin stop.
25. STOP_ID_2: Alternate ID for the destination stop.
26. STOP_KEY_1: Another stop key for the origin stop.
27. STOP_KEY_2: Another stop key for the destination stop.
28. TIME_SCHEDULED: Scheduled time of arrival at the destination stop.
29. TIMEPOINT_MILES: Distance in miles between the two stops.
30. READ_TIME: Time the data collection system recorded the bus.
31. READ_DATE: Date the reading was captured.
32. RUNNING_TIME_SCH: Scheduled duration for this segment (in minutes).
33. RUNNING_TIME_ACT: Actual duration the bus took.
34. RUNNING_TIME_DIFF: Difference = Actual - Scheduled (positive = late).
35. TIMEPOINT_DWELL_1: Time bus spent waiting at the first stop.
36. TIMEPOINT_DWELL_2: Time spent at the second stop.
37. SPEED_SCHEDULED: Planned average speed between the two stops (mph).
38. SPEED_ACTUAL: Actual average speed.
39. FIRST_LAST_STOP: Flag (1 or 2) indicating if it's the start or end of the route.
40. TIME_ACTUAL_ARRIVE: Actual arrival time at the destination.
41. TIME_ACTUAL_DEPART: Actual departure time.
42. ONTIME_METHOD_1: On-time performance category (1 = on time, 2 = late, etc.).
43. ONTIME_METHOD_2: Alternative method of determining on-time status.
44. TRIPS_COUNT: Number of trips aggregated into this row (often 0 = individual trip).

- Key features:

```
columns_to_keep = [
    'TIMEPOINT_MILES', 'TRIP_START_TIME', 'ROUTE_NAME', 'DIRECTION_NAME',
    'SERVICE_PERIOD', 'TIME_SCHEDULED', 'SPEED_SCHEDULED', 'SPEED_ACTUAL', 'FIRST_LAST_STOP',
    'STOP_ID_1', 'STOP_ID_2', 'RUNNING_TIME_ACT', 'TIME_ACTUAL_DEPART'
]
```

Data Cleaning

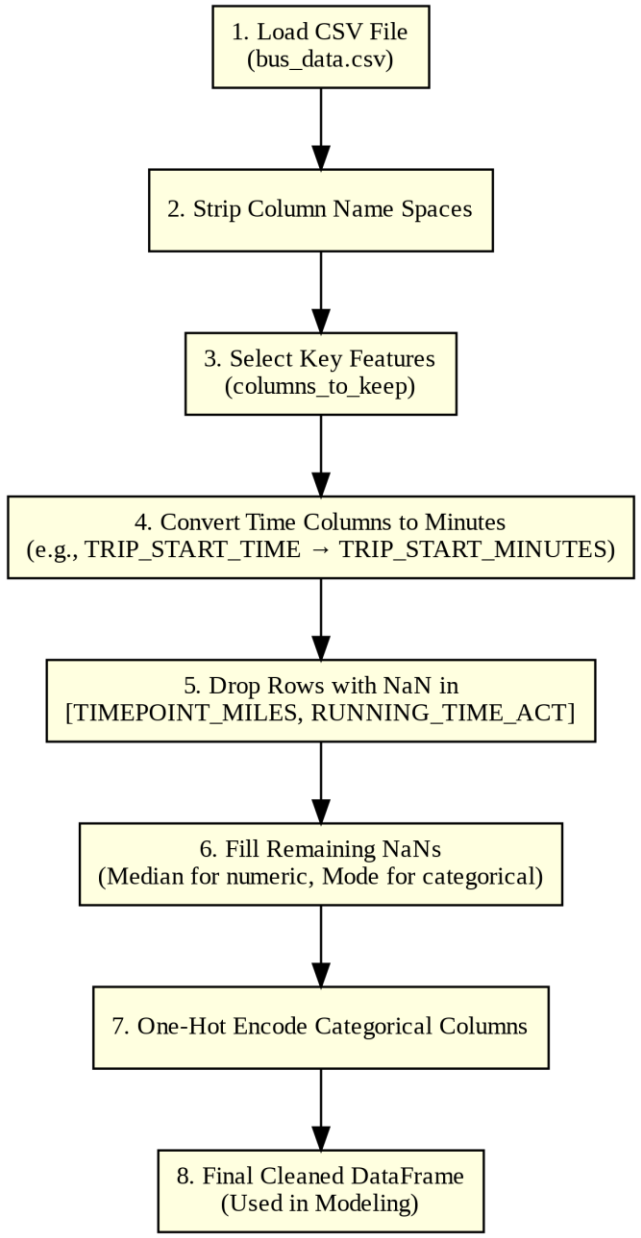
Dropped unnecessary columns, retained 13 key features.

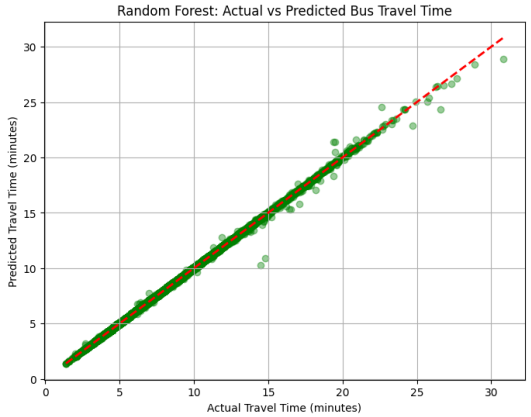
Converted time fields to minutes after midnight.

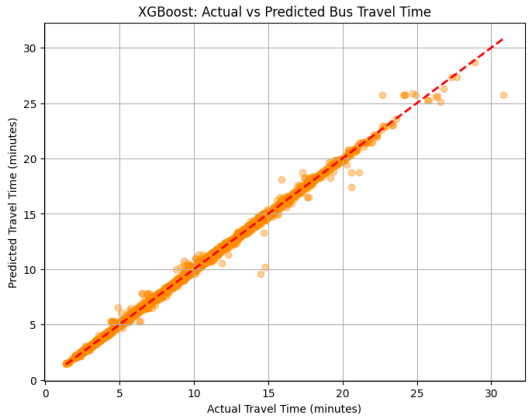
One-hot encoded categorical columns (ROUTE_NAME, DIRECTION_NAME, SERVICE_PERIOD)

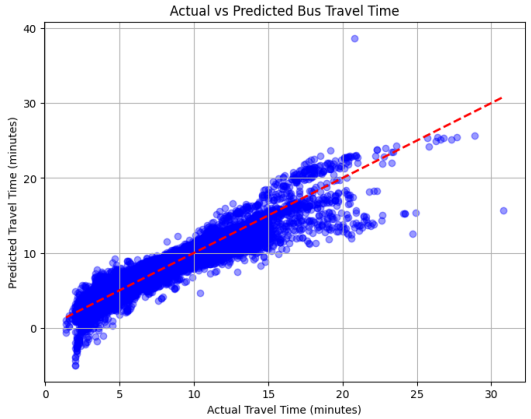
Removed rows with NaNs in key fields, filled others with median/mode.

Verified all data was numeric and ready for modeling.







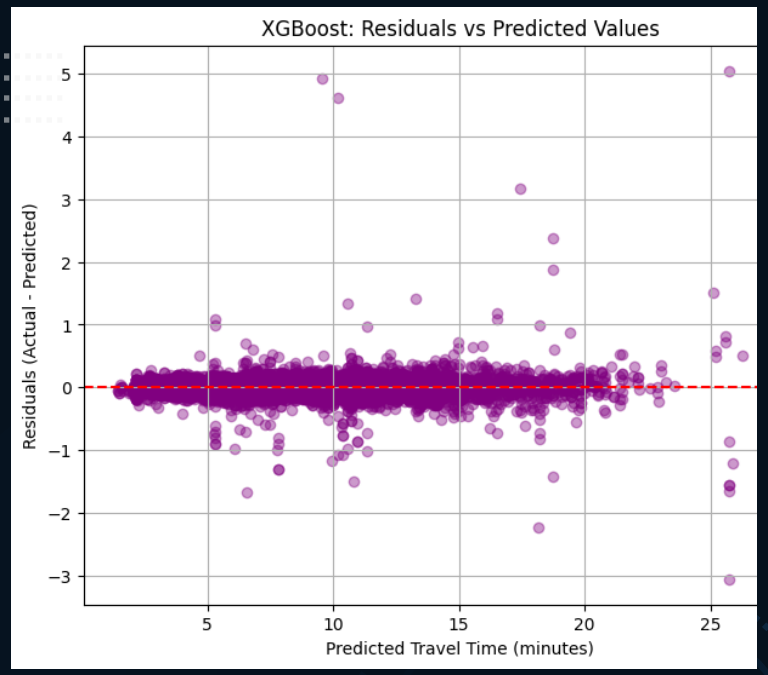


Models Tried:

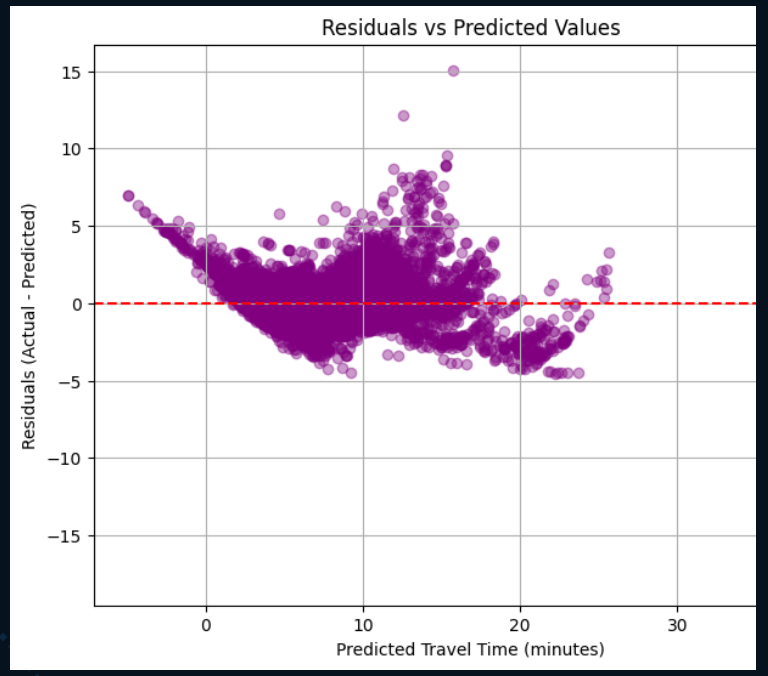
1. Linear Regression
- Simple, fast, interpretable
- Struggled with non-linearity
- R^2 = 0.89

2. Random Forest Regressor
- Handles non-linearity well
- Best performer (R^2 = 0.9995)

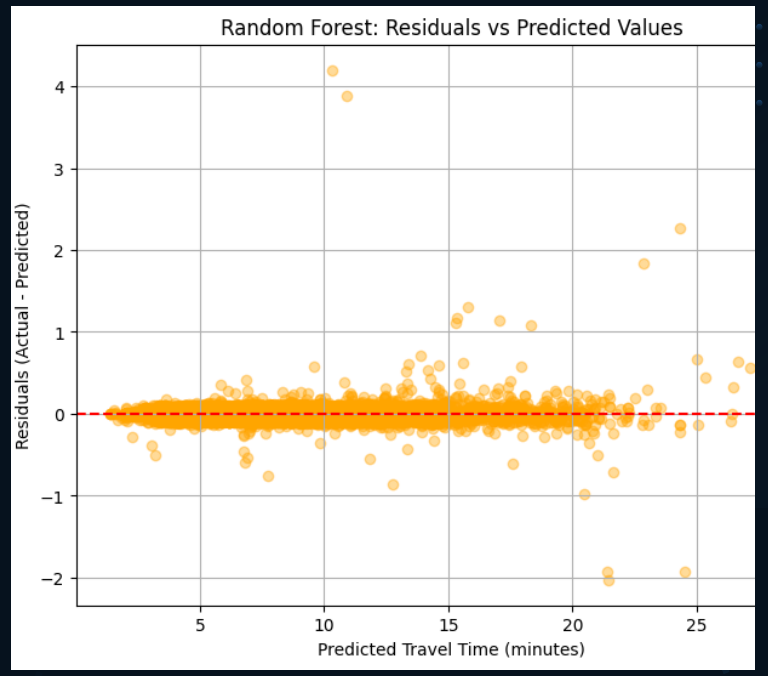
3. XGBoost Regressor
- Gradient boosting, slightly below RF
- R^2 = 0.9986



Linear:



Random Forest



XGBoost:

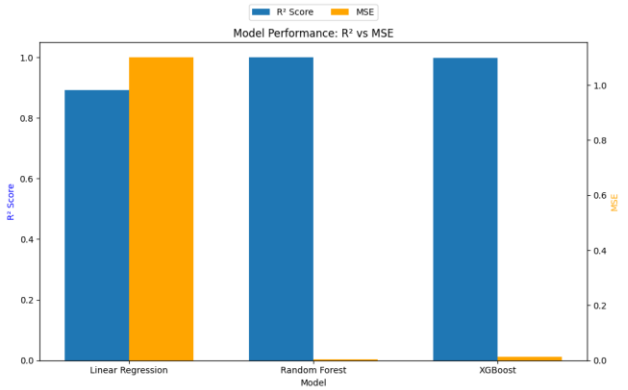
Residual Graphs

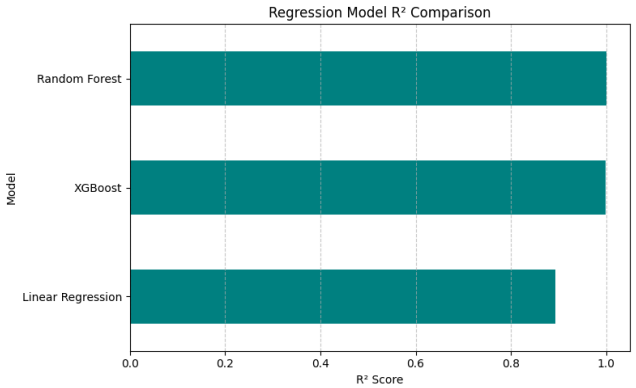
- Linear: High Variance
- Random Forest and XGBoost: Low, mostly centered around 0

Results Summary

Model	R^2	MSE
Linear Regression	0.8928	1.10
Random Forest	0.9995	0.0048
XGBoost	0.9986	0.0137

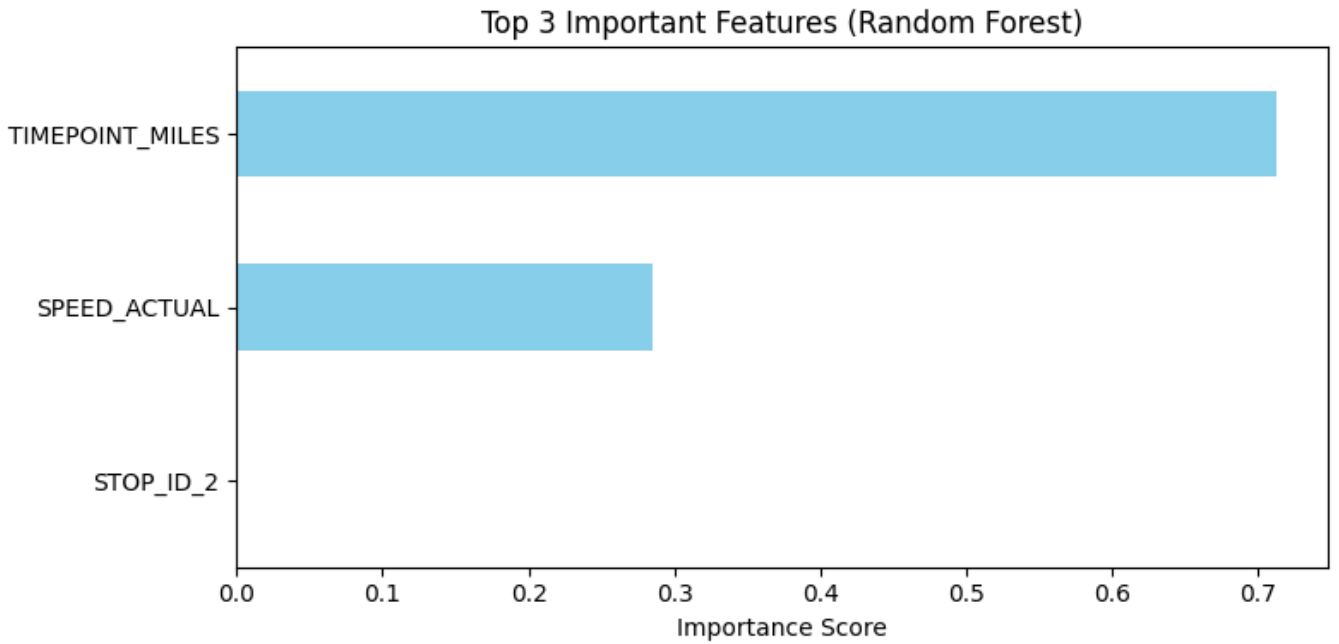
- Three regression models were evaluated: Linear Regression, Random Forest, and XGBoost.
- Random Forest achieved the best performance with an R^2 score of 0.9995 and an MSE of 0.0048, indicating highly accurate predictions.
- XGBoost also performed exceptionally well ($R^2 = 0.9986$, MSE = 0.0137), only slightly less accurate than Random Forest.
- Linear Regression, while still useful, showed significantly lower performance ($R^2 = 0.8928$, MSE = 1.10), suggesting that linear assumptions were too simplistic for the data.
- The chart clearly shows that tree-based models significantly outperform linear methods in predicting WRTA bus travel time.



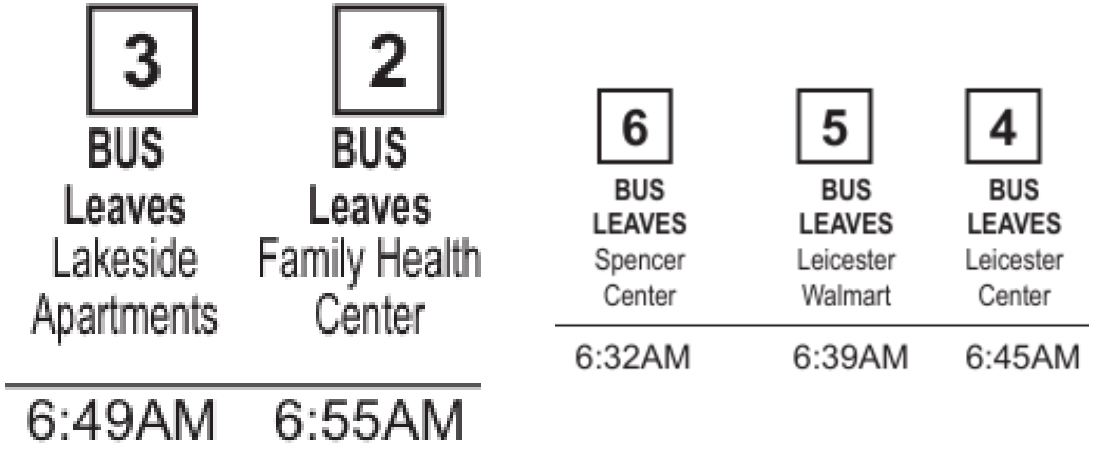


Feature Importance

- Most important features:
- TIMEPOINT_MILES
- SPEED_ACTUAL
- STOP_ID_2



Prediction Examples



- Comparison of predicted travel times between key WRTA bus stop pairs using three regression models. Random Forest provides the closest estimates to real-world delays so that is the focus.
- Important observation: the predicted bus time is between two stops is greater than actual scheduled

Stop Pair	Linear Regression	Random Forest	XGBoost
Spencer Center → Leicester Walmart	10.56 min	10.61 min	10.03 min
Leicester Walmart → Leicester Ctr	7.47 min	7.31 min	7.27 min
Lakeside Apartments → Family Health Center	7.58 min	7.40 min	7.41 min

Conclusions

Buses are consistently late by 1-4 minutes.

Random Forest is the best model.

Scheduled times underestimate actual delays.

Model Could help WRTA improve scheduling

THANK YOU!