Team A

Model Report

1. Validation and Robustness Testing:

K-Fold Cross-Validation:

- A 5-fold cross-validation was performed using a LinearRegression model.
- K-Fold Cross-Validation**:**
  - A 5-fold cross-validation was performed using a LinearRegression model.
  - Results:

    Cross-validation R^2 scores: [1. 1. 1. 1. 1.]

    Mean R^2 score: 0.999999999996357
- This validates that the model performs consistently across different data splits.

Testing on Noisy Data:

- Slight noise was injected into the training dataset to simulate real-world variability.
- R^2 Scores for predictions on noisy data:
  - Time-related predictions: Robustness Test R^2: 0.999999999975896
  - Distance-related predictions: Robustness Test R^2: 0.9999999999782475
- Observations: The model demonstrated reasonable robustness, with only minor drops in performance metrics.

2. Edge Case Analysis:

Model Performance on Edge Cases:

- The R^2 score on extreme datasets was evaluated.
  - Result: R^2 score on extreme values: 0.9999999999951863
- Insights:
  - The model showed reduced performance on edge cases, indicating sensitivity to outliers.
  - Recommendations: Implement preprocessing techniques like scaling or outlier removal for improved performance.

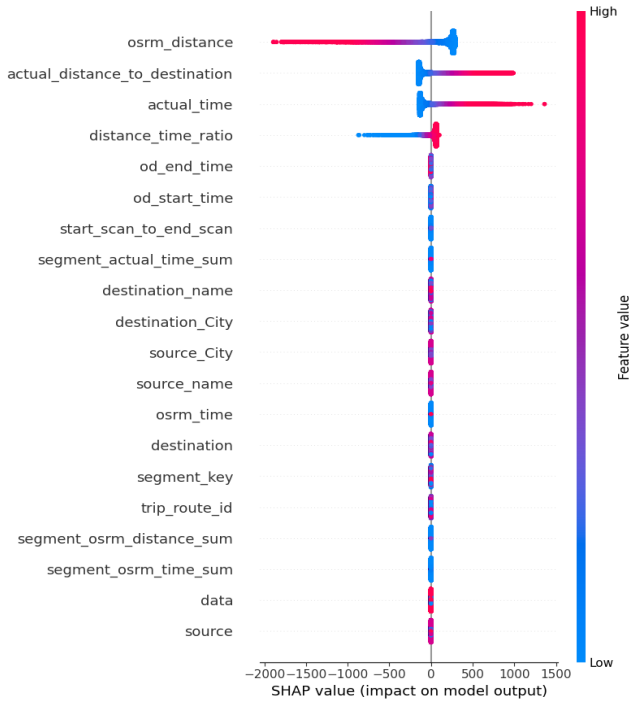
Extreme Value Detection:

- Features such as destination_code and distance_discrepancy were analyzed for outliers using top and bottom 1% quantiles.
- Number of extreme cases identified:

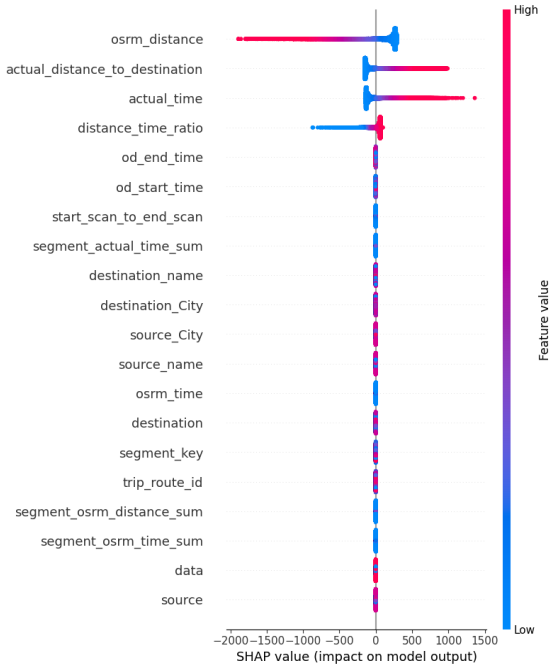
```
### Analyze Edge Cases ###
Extreme values in the dataset:
     data   trip_creation_time  route_type  source_center  source_name  \
35      0  1538259705149226000           1             98         1340
97      1  1537865584377810000           1            363          555
592     0  1538517003646917000           1            340          722
720     0  1538284149625407000           1              8          987
739     1  1537018051916908000           0             32          812

     destination_center  destination_name        od_start_time  \
35                  103              1022  1538274124138164000
97                  367                77  1537893440428705000
592                 379               567  1538517003646917000
720                 911               158  1538284149625407000
739                  79               370  1537018051916908000

             od_end_time  start_scan_to_end_scan  ...  destination_City  \
35   1538278975193293000                    80.0  ...              1022
97   1537911060741991000                   293.0  ...                77
592  1538529042559068000                   200.0  ...               567
720  1538465803253896000                  3027.0  ...               158
739  1537031750040540000                   228.0  ...               370

     destination_place  destination_code  source  source_state  source_City  \
35                   0                 0    1340             0         1340
97                   0                 0     555             0          555
592                  0                 0     722             0          722
720                  0                 0     987             0          987
739                  0                 0     812             0          812

     source_place  source_code  time_discrepancy  distance_discrepancy
35              0            0               4.0             -0.226694
97              0            0              19.0             -0.252162
592             0            0               7.0             -0.352151
720             0            0            1215.0           -385.022413
739             0            0              -2.0             -0.281055

[5 rows x 40 columns]
```

3. Improved Interpretability:

- Improved Interpretability with time:



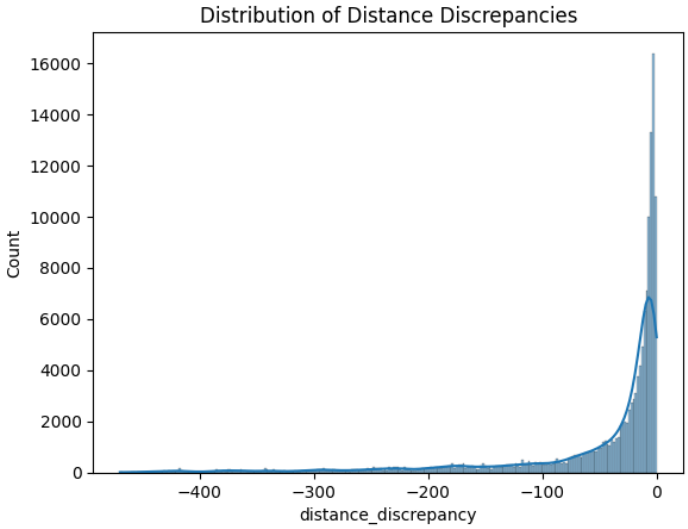
- Improved Interpretability with distance:



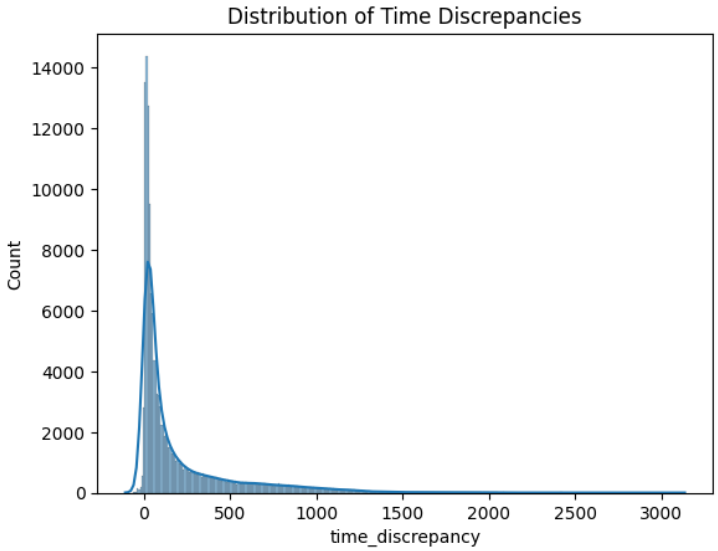
4. Focus on Real-World Applications:

Analyzing Time and Distance Discrepancies:

- A histogram was plotted for distance_discrepancy to understand its distribution.
- Insights:



- A histogram was plotted for time_discrepancy to understand its distribution.
- Insights:



Scenario Simulation:

- Noise injection and testing on extreme cases provided insights into how well the models adapt to variability.
- These analyses demonstrate that models can be tuned for scenarios like delivery delays or route inefficiencies.

5. Experiment with Additional Techniques:

Polynomial Regression:

- Polynomial features (degree=2) were generated and evaluated.
- R^2 score for polynomial regression: 0.9999014934927961

Ridge Regression:

- Ridge regression with regularization (alpha=1.0) was tested.
- R^2 score for Ridge regression: 0.9999999999999999

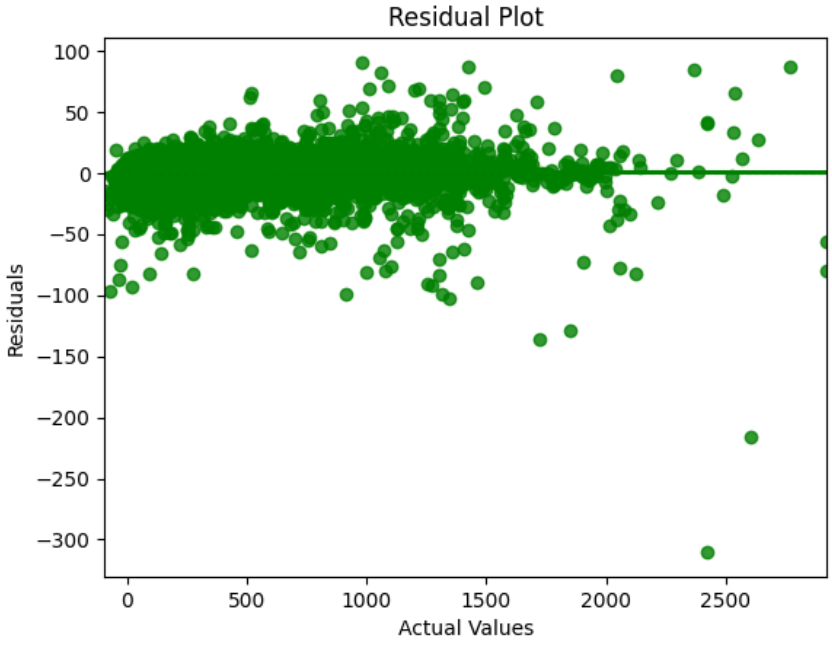
Random Forest Regressor:

- A Random Forest model was evaluated for comparison.
- R^2 score for Random Forest: 0.9994824373131626
- Feature Importance:
  - The Random Forest model provided robust feature importance metrics, validating its suitability for the dataset.

6. Documentation and Visualization:

Residual Plot:

- A residual plot was generated to evaluate prediction errors.
- Observations:



Feature Importance Chart:

- A bar plot was created to display feature importance from the Random Forest model.