Holly Cohan, Yesh Onipede, Yu Xia

Week 10 Report: Data Centric AI

The objective of this week's assignment is to apply data-centric AI principles to our modeling project by iterating over the data and using the winning model, the pruned decision tree, to improve prediction performance. We focused on enhancing our dataset by making several key changes aimed at improving the model's predictive capability on accurately predicting rideshare prices in Boston.

In our error analysis, we identified several areas where the model could benefit from improved data quality and feature enhancements. One key improvement was correcting the coordinates for the Theatre District location. By refining these coordinates, we provided the model with more precise spatial information. Additionally, we introduced two new features, eta_minutes and eta_minutes_rh, to address gaps in our model's ability to accurately capture travel duration and its impact on fares. The eta_minutes variable estimated the travel duration based solely on distance, while eta_minutes_rh further adjusted this estimate by factoring in rush hour conditions to account for expected delays during peak times. These enhancements aimed to address the model's lack of detailed temporal context. By providing a more nuanced understanding of travel time, particularly during high-traffic periods, we sought to improve the model's ability to predict fare variations due to traffic congestion and time-of-day effects

Data Improvement

- This week, we made three slight changes to our model
    - **Improved coordinate accuracy of the Theatre District**
        - Initially, we used Geopy for acquiring the coordinates or our source and destination locations. However, Geopy incorrectly used a coordinate outside of Boston for the location of the Theatre District. We updated this coordinate to now be accurate.
    - **Addition of eta_minutes variable:** The estimated duration of the ride in minutes, calculated based on the straight-line distance between the trip's starting point and destination. It provides a baseline estimate of travel time but doesn't take into account factors like traffic or time of day.
        - We suspected that adding eta_minutes might be valuable for predicting rideshare prices as it provides a standard estimate of how long a ride should take under ideal conditions. This baseline is useful as it reflects the fundamental relationship between distance and time, which offers the model a sense of expected duration based solely on the distance of the trip. This can help the model better understand how distance generally translates into time, which is still a meaningful factor in pricing.
        - While eta_minutes does not consider real-time traffic or demand conditions, it serves as a neutral reference point. For example, if actual prices deviate significantly from what would be expected based on eta_minutes alone, this could indicate that surge pricing, traffic delays, or other factors are influencing the cost. This baseline estimate allows the model to distinguish between the base cost of a ride and adjustments made for more dynamic, real-world conditions.

    - **Addition of eta_minutes_rh**: This enhanced version of the eta_minutes variable takes into account not only the distance but also the specific departure time, which helps account for rush hour conditions. By factoring in peak travel times, eta_minutes_rh aims to provide a more accurate estimate of travel duration, especially during periods of heavy traffic, which are known to impact rideshare prices.
        - We hypothesize that incorporating the eta_minutes_rh variable could improve the model's predictive accuracy by offering a time-aware estimate of travel duration. During peak times, such as rush hour, rides tend to take longer, potentially resulting in higher fares due to extended travel times and surge pricing. By incorporating the departure time into the calculation, eta_minutes_rh enables the model to capture these temporal fluctuations more effectively, leading to more accurate predictions of rideshare prices.
- In previous weeks, we combined several variables into one feature in order to reduce dimensionality. We used a custom weighting system which assigned scores to each type of weather event based on severity or relevance. The weighting system will assess the weather condition by considering several weather severity related variables, creating a single new variable for weather severity. In addition to reducing dimensionality, this variable should improve interpretability.
- We had also previously added variables indicating rush hour, whether it is a weekend day, and whether there is a Bruins or Celtics home game that day. While none of these variables reduce the dimensionality of the dataset, they could have a critical impact on the prices of rideshare, and therefore are imperative to include in the model.

Data Preprocessing

After adding this additional feature eta_minutes, we ran the following preprocessing pipeline again before training the pruned decision tree model again.

Imputing Missing Values:

- We imputed missing values of the Uber Taxi rides using the Taxi fare breakdowns in the Uber app, as explained in the previous report.

Encoding Categorical Variables:

- We encoded the categorical variable using dummy coding, one-hot encoding, and ordinal encoding based on the type and context of the variable.
- After encoding, we were left with 84 features, which are significantly less than 110 variables without feature engineering.

Standardization:

- We also standardized the features using StandardScaler to ensure that each feature contributes equally to the distance calculations involved in PCA. By combining PCA with standardization, we can effectively reduce dimensionality while preserving the most significant variance in our dataset, making it more suitable for predictive modeling techniques.

Principal Component Analysis:

- We conducted PCA again to address high dimensionality and correlation identified. We set a threshold of 0.95 for the cumulative explained variance ratio, which resulted in retaining **N** principal components. This threshold was chosen to ensure that we capture at least 95% of the variance in the dataset while reducing the dimensionality of our feature space. By selecting these components, we aim to minimize noise and redundancy in the data, thus enhancing the model's performance and interpretability. As the result, the PCA reduce the features to 45 components, which is about half of the original training dataset after feature engineering and categorical variable encoding.

Re-training the Pruned Decision Tree Model with the Enhanced Data

This week we focused on applying data-centric AI principles to enhance our dataset with additional information in hopes of improving our model's predictive accuracy. We added an additional feature called "eta_minutes" to our dataset which represents the estimated length of the ride in minutes, calculated based on the distance from the trip's source to its destination. By integrating this variable into our data enhancement function, we retrained our winning pruned decision tree model with the enriched dataset. This retraining allows the model to leverage the eta_minutes variable to make more informed predictions. After retraining the model, we re-acquired the performance metrics and compared them against those of our original model.

Performance Metrics for Original Data and Enhanced Data

Model Type	**Metrics**	**Training Set Values**	**Validation Set Values**	**Testing Set Values**
Pruned Decision Tree Model	**MSE**	2.6747	3.0939	**3.1037**
	RMSE	1.6354	1.7589	**1.7617**
	R^2	0.9997	0.9996	**0.9997**
Pruned Decision Tree Model on Enhanced Data (with updated coordinates and eta_minutes)	**MSE**	2.6491	3.0202	**3.0452**
	RMSE	1.6276	1.7379	**1.7450**
	R^2	0.9679	0.9631	**0.9635**
Pruned Decision Tree Model on Enhanced Data (with updated coordinates and	**MSE**	2.6874	3.0487	**3.0047**
	RMSE	1.6393	1.7460	**1.7334**
	R^2	0.9675	0.9628	**0.9641**

eta_minutes_rh)				

Results

Summary of Percentage Changes

Validation Set Changes

- **MSE**
  - Adding eta_minutes: MSE decreased by 2.38%
  - Adding eta_minutes_rh: MSE decreased by 1.46%
- **RMSE**
  - Adding eta_minutes: RMSE decreased by 1.19%
  - Adding eta_minutes_rh: RMSE decreased by 0.73%
- **R^2**
  - Adding eta_minutes: R^2 decreased by 3.67%
  - Adding eta_minutes_rh: R^2 decreased by 3.68%

Test Set Changes

- **MSE**
  - Adding eta_minutes: MSE decreased by 1.89%
  - Adding eta_minutes_rh: MSE decreased by 3.19%
- **RMSE**
  - Adding eta_minutes: RMSE decreased by 0.95%
  - Adding eta_minutes_rh: RMSE decreased by 1.61%
- **R^2**
  - Adding eta_minutes: R^2 decreased by 3.62%
  - Adding eta_minutes_rh: R^2 decreased by 3.56%

Note on R^2: In this context, R^2 is less critical because we are focusing on predictive performance (as indicated by MSE and RMSE), rather than strictly on the model's variance explanation. When adding new variables, slight reductions in R^2 can occur if the new features don't add substantial new information, as seen here. Therefore, we prioritize MSE and RMSE improvements as indicators of model accuracy over R^2 in this case.

Based on our analysis, we decided to select the pruned decision tree model enhanced with eta_minutes_rh and updated coordinates as our final model. This choice is driven by the improvements in MSE and RMSE, especially on the test set:

- **Test Set Changes**:
  - **MSE**: With eta_minutes_rh, the MSE decreased by **3.19%**, compared to a **1.89% decrease** with only eta_minutes.
  - **RMSE**: The addition of eta_minutes_rh reduced RMSE by **1.61%**, compared to **0.95%** with eta_minutes.

These metrics indicate that the model with eta_minutes_rh and improved coordinates provides better predictive accuracy, likely due to the additional nuance in travel duration estimates during rush hour. While there was a slight decrease in R^2, we prioritized MSE and RMSE as key indicators of prediction performance, given our objective to reduce raw error.

The eta_minutes_rh feature, by accounting for peak travel times, enables the model to capture variations in ride duration and pricing more effectively, particularly in high-traffic conditions. Thus, we concluded that the model incorporating eta_minutes_rh and updated coordinates offers the best balance between enhanced accuracy and contextual relevance, making it our final choice. However, there is definitely room for further improvements.