



Deep Dive – 9
Project 14 - NYC Citi Bike Rentals
Industrial Engineering
12-10-2023



Project Members

Deepanshu Malhotra

Prannoy Kathiresan

Selvamani Subramaniam

Vishal Dayalan



Milestone 1

- Data Extraction

Milestone 1



PROBLEM DESCRIPTION:

- To predict the demand at Lyft Bike destination stations in JC of NY based on the history of trip details. In our case, we are predicting the demand at a specific station for a given day and time.

DATA EXTRACTION:

- Dataset: 24 months of historical data (JC 2021 and JC 2022)
- Extraction process : Zip file download → Un Zip → Convert to CSVs → Merge to single CSV
→ Convert to pickle file → Split into train pkl and test pkl
- We have used two years data alone to understand the effect of public commute in the post-COVID period.



Milestone 2

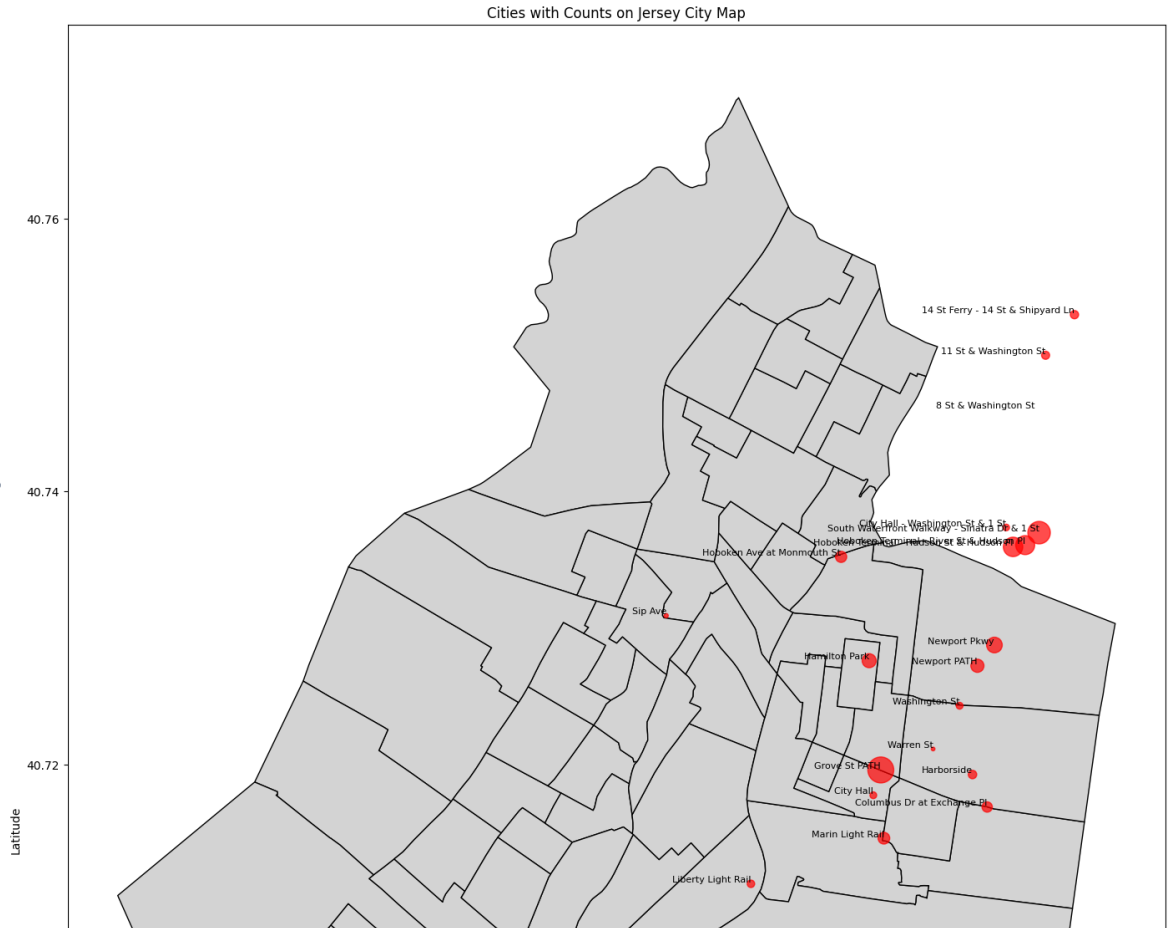
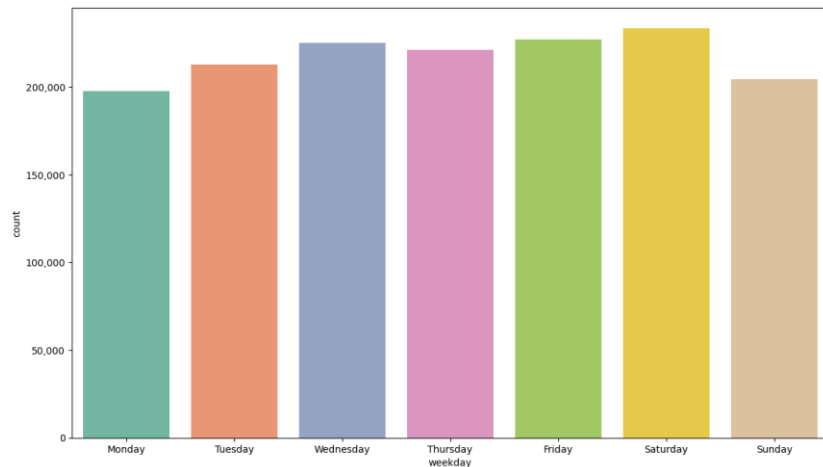
- Data Exploration & Visualization
- Baseline Learning

Milestone 2



DATA EXPLORATION & VISUALIZATION

- Total number of trips : 1,521,600
- First most frequent trips (6-minute) : ~140,000
- Second most frequent trips (7-minute) : ~139,900
- Third most frequent trips (8-minute) : ~127,000
- Start time of most frequent trips : 1800 – 1900 hrs



Milestone 2 (Contd...)



BASELINE LEARNING:

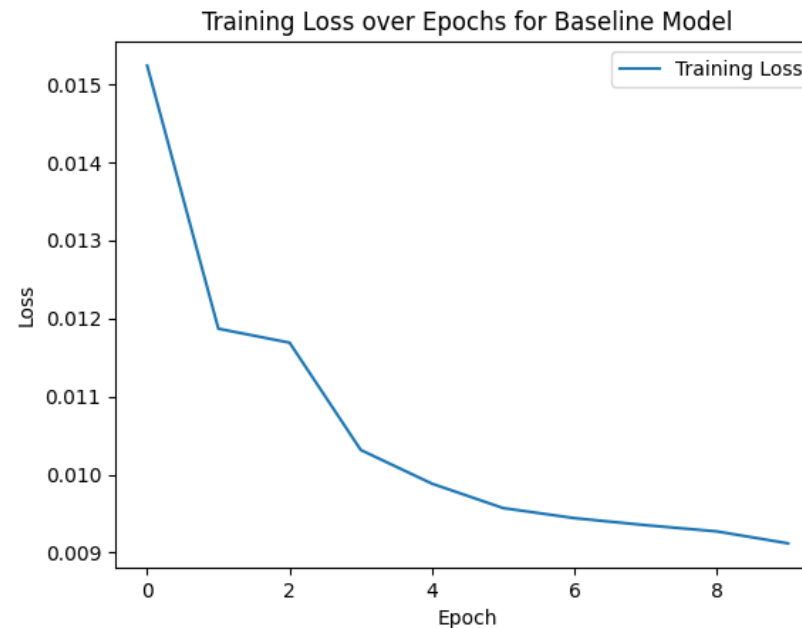
- Data preprocessing:
 - Categorical features (rideable_type, member_casual) are one-hot encoded.
 - Numerical features are standardized using Standard_Scaler.
 - Data grouped by start_day_of_the_week, start_hour, end_station_name.
 - Demand is a calculated field based on the groupby count.
- Features {'rideable_type', 'start_lat', 'start_long', 'end_lat', 'end_long', 'member_casual', 'start_month', 'stop_month', 'start_day_of_the_week', 'stop_day_of_the_week', 'start_hour', 'stop_hour', and 'trip_duration_minute'}
- Labels {demand}
- Two-layer Simple neural network
 - Fully connected linear layer
 - ReLU activation
 - Fully connected linear layer
- Hyperparameters → {lr, epochs, batch_size, optimizer} = {0.001, 10, 32, ADAM}
- Loss → MSELoss

Milestone 2 (Contd...)



BASELINE LEARNING PERFORMANCE:

- MSE on Test set : 0.0015 (i.e Squared residual between actual demand and predicted demand)
- Loss value on Train set reduced from 0.0152 to 0.0091 over 10 epochs





Milestone 3

- Deep Learning

Milestone 3



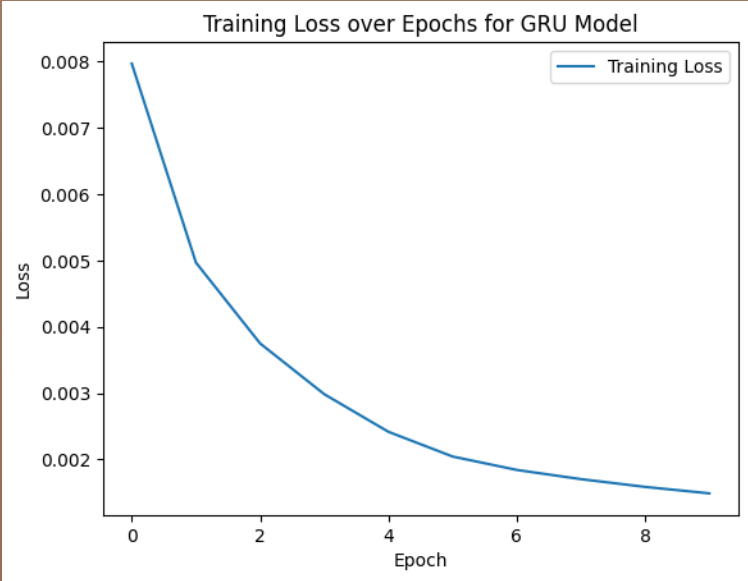
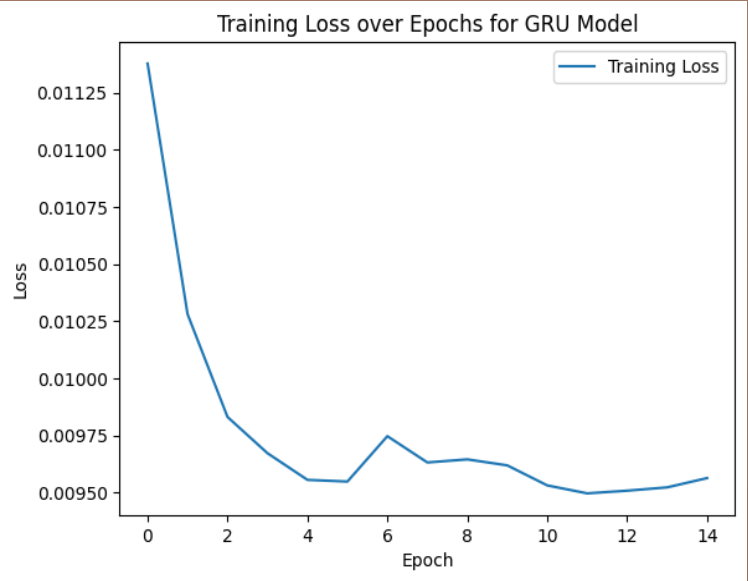
Deep learning model:

- Model : GRU (Gated Recurrent Unit)
- Reason for this model : Sequential data handling, Effective capturing of temporal dependencies, learning pattern from historical sequences, Long-range dependencies within sequential data
- Model Architecture: GRU Layer & Linear layer
- Hyperparameters → {lr, epochs, batch_size, hidden_layer}
- Loss → MSELoss

Milestone 3 (Contd...)



DEEP LEARNING MODEL (GRU) PERFORMANCE:

Hyperparameters	{0.001, 10, 32, 50}	{0.1, 15, 64, 100}
Training loss (MSE)		
Test loss (MSE)	0.0015	0.0090



Milestone 4

- Feature Importance

Milestone 4



FEATURE IMPORTANCE:

- GRU Challenges : Sequential data processing nature of GRU, spread of input features across multiple time steps make it more complicated to perform feature importance.
- Even feedforward-based techniques such as Gradient based techniques or permutation importance are not directly applicable to GRUs
- Alternatively, linear regression is used to perform feature importance.
- Coefficients of linear regression are used to calculate weights for each feature.
- Weights are sign-invariant.
- Higher the weight (magnitude) \rightarrow Higher the correlation \rightarrow More important feature.

Milestone 4 (Contd...)



FEATURE IMPORTANCE:

- Following is the ranking of the features based on their importance on target variable.

```
Rank 1: end_lng : 0.8471573956039891
Rank 2: start_lat : 0.027009764661260177
Rank 3: member_casual_member : 0.0045341548531796574
Rank 4: end_hour : 0.002035528124128154
Rank 5: rideable_type_classic_bike : 0.0018088129201581354
Rank 6: stop_day_of_week : 0.0011883416346306647
Rank 7: rideable_type_electric_bike : 0.0008314827805634868
Rank 8: stop_month : 0.0005206808500088228
Rank 9: start_month : -0.0007151188407709952
Rank 10: tripduration_minute : -0.0010775361407049779
Rank 11: start_hour : -0.0022935633074943652
Rank 12: rideable_type_docked_bike : -0.002640295700721454
Rank 13: start_day_of_week : -0.0031733774370082827
Rank 14: member_casual_Customer : -0.004534154853177073
Rank 15: start_lng : -0.06880580901285406
Rank 16: end_lat : -0.6807360669038712
```

Explicit Sample data



	rideable_type	start_station_name	end_station_name	start_lat	start_lng	end_lat	end_lng	member_casual	start_month	stop_month	start_day_of_week	stop_day_of_week	start_hour	end_hour	tripduration_minute	demand
0	classic_bike	Mama Johnson Field - 4 St & Jackson St	South Waterfront Walkway - Sinatra Dr & 1 St	0.755607	0.747216	0.400846	0.304860	Customer	3	3	4	4	15	15	-0.033168	537
1	electric_bike	Baldwin at Montgomery	Grove St PATH	0.353913	0.378314	0.323884	0.228441	member	3	3	4	4	16	16	-0.037164	631
2	electric_bike	Baldwin at Montgomery	Grove St PATH	0.353913	0.378314	0.323884	0.228441	member	3	3	6	6	17	17	-0.021181	565
3	classic_bike	Baldwin at Montgomery	Grove St PATH	0.353913	0.378314	0.323884	0.228441	member	3	3	6	6	15	15	-0.029172	605
4	classic_bike	Baldwin at Montgomery	Grove St PATH	0.353913	0.378314	0.323884	0.228441	member	3	3	4	4	12	12	-0.013190	414
...
1521595	electric_bike	Madison St & 1 St	Columbus Dr at Exchange Pl	0.665912	0.758533	0.311868	0.279800	Customer	9	9	4	4	21	22	-0.013190	147
1521596	classic_bike	Monmouth and 6th	Bergen Ave & Stegman St	0.394022	0.612992	0.266324	0.011277	member	9	9	6	6	17	18	-0.009194	24
1521597	electric_bike	4 St & Grand St	Madison St & 10 St	0.737421	0.822512	0.458187	0.264579	member	9	9	2	2	16	16	-0.041159	147
1521598	classic_bike	4 St & Grand St	Madison St & 10 St	0.737421	0.822512	0.458187	0.264579	member	9	9	1	1	11	11	-0.045155	82
1521599	classic_bike	4 St & Grand St	Madison St & 10 St	0.737421	0.822512	0.458187	0.264579	Customer	9	9	4	4	20	21	-0.037164	115

1521600 rows × 16 columns

Conclusion



- Motivation behind choosing this problem is to predict the demand at various bike stations, enabling bike companies to plan ahead, increase revenue, and enhance operational efficiency. This, in turn, contributes to an improved user experience.
- To meet these objectives, we initially constructed a simple neural network as a baseline model. Subsequently, we implemented a deep learning model, specifically a Gated Recurrent Unit (GRU), to forecast the demand at a given station, considering the specific time and day of the week.
- One of the challenges encountered during the deployment of the GRU model was extracting feature importance. The complexity arose from the interaction of each input feature across multiple time steps, making it challenging to isolate the significance of individual features.
- To address this challenge, we employed Linear Regression as an alternative model. This allowed us to calculate the importance of each feature based on the feature coefficients in relation to the target variable.



I ILLINOIS