IE434 – Final Project Presentation

NYC Citi Bike Rentals Demand Prediction Model

Group: Deep Dive 11

- Amarthya Kuchana,
- Kibae Kim,
- Nithin Balaji,
- Safin Akash,
- Surya Vasanth

Problem Statement:

Objective:

The project aims to develop a Deep Learning model that **classifies** the **daily demand** for Lyft Bikes in Jersey City, NY, into three categories: **High, Medium, and Low.**

Significance:

- Optimize bike distribution to meet customer demand efficiently.
- Increase resource utilization and operational efficiency.
- Enhance customer satisfaction by reducing potential wait times.

Methodology:

Multi label classification.

Raw Dataset:

Data: The dataset consists of Bike rental data for Jersey City from January 2022 – September 2023

License: https://ride.citibikenyc.com/data-sharing-policy

Column Name	Description	Data Type	Example Value
Ride ID	Unique identifier for each ride	String	4D7C2514E8852AF7
Rideable Type	Type of rideable used	String	classic_bike
Started At	Start date and time of the ride	Datetime	2022-02-03 19:29:10
Ended At	End date and time of the ride	Datetime	2022-02-03 19:39:35
Start Station Name	Name of the start station	String	Marshall St & 2 St
Start Station ID	Unique identifier for start station	String	HB408
End Station Name	Name of the end station	String	Grand St & 14 St
End Station ID	Unique identifier for end station	String	HB506
Start Latitude	Latitude of the start location	Float	40.739804
Start Longitude	Longitude of the start location	Float	-74.064198
End Latitude	Latitude of the end location	Float	40.743139
End Longitude	Longitude of the end location	Float	74.043959
Member or Casual	Type of rider	String	Member

Additional Data Integration:

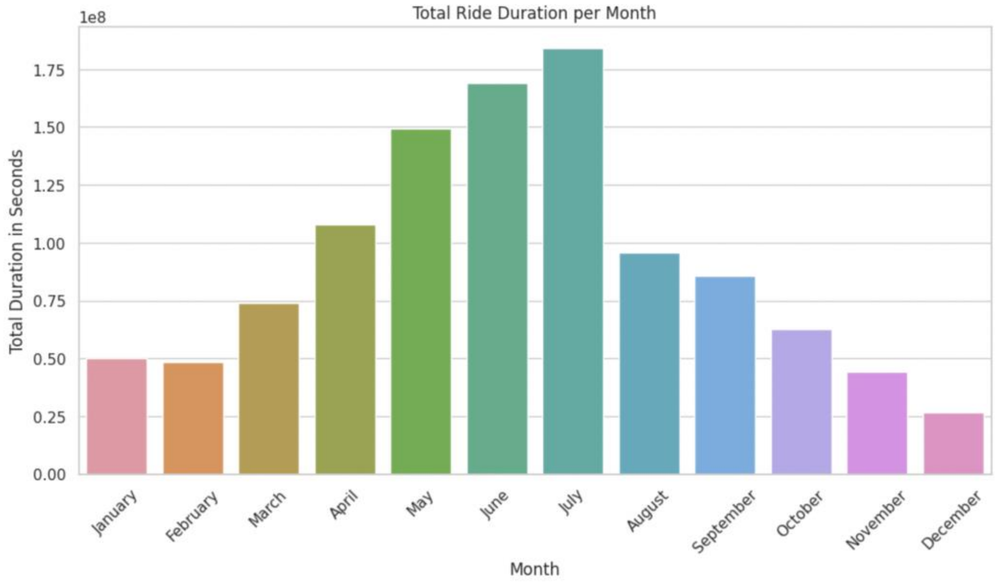
Weather Data: Daily climate data has been taken for the Jersey City, NY and integrated with original data

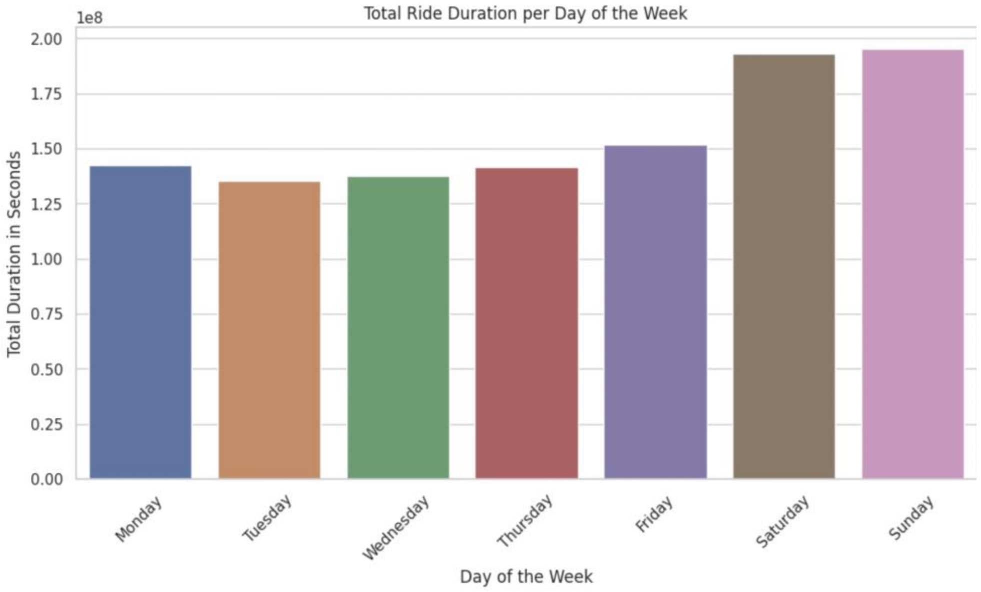
https://www.ncdc.noaa.gov/cdo-web/datasets

Column Name	Description	Data Type	Example Value
Date	Date	Date	1/1/2022
PRCP	Precipitation value for the Day	Float	0.89
SNOW	Snow Fall	Float	0.0
TMAX	Maximum temperature of the Day	Integer	60
TMIN	Minimum temperature of the Day	Integer	30

Exploratory Data Analysis: Uncovering Patterns and Trends

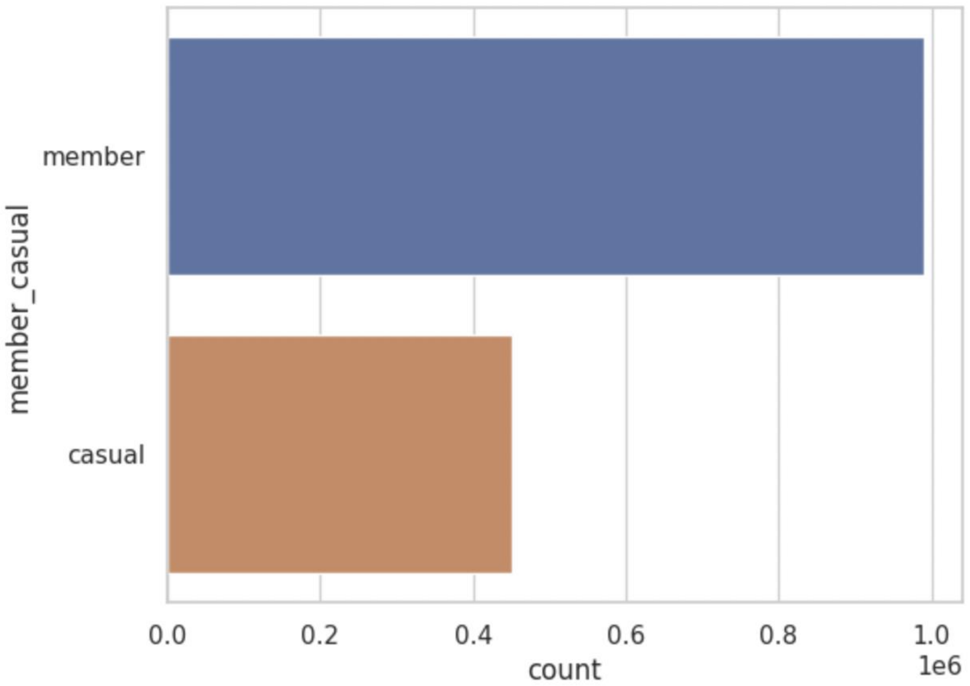
Trend Analysis: Investigated rental patterns, revealing seasonal and weekly demand fluctuations critical for demand forecasting.

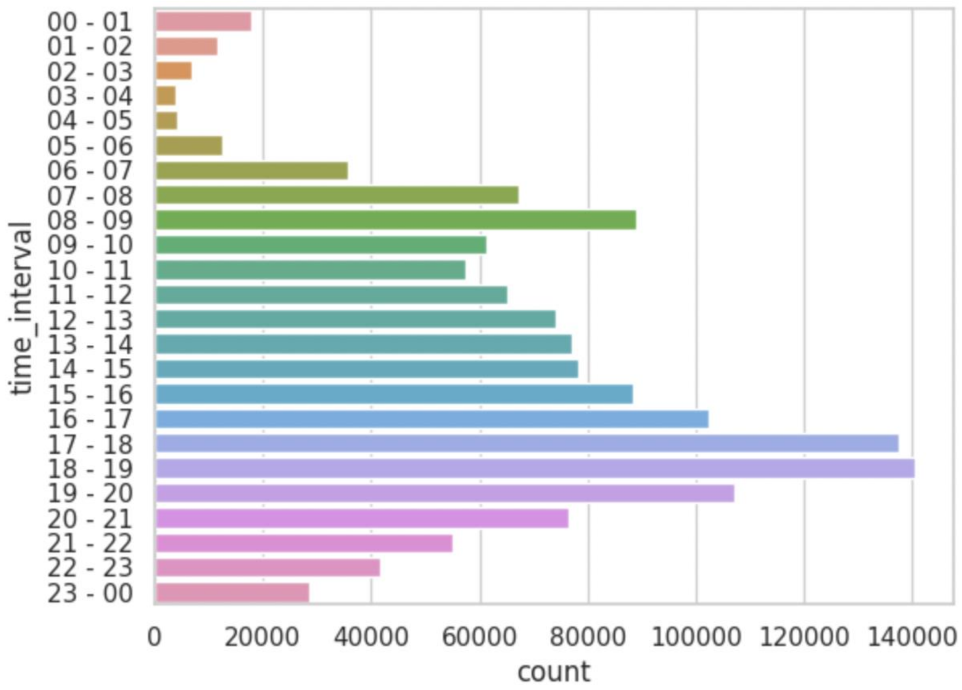




Exploratory Data Analysis: Uncovering Patterns and Trends

User Behaviour: Analysed ride durations, time interval and user types, discerning distinct usage trends between members and casual riders.





Data Preprocessing: From Raw Data to Insightful Features

- Dropped missing values in the dataset.

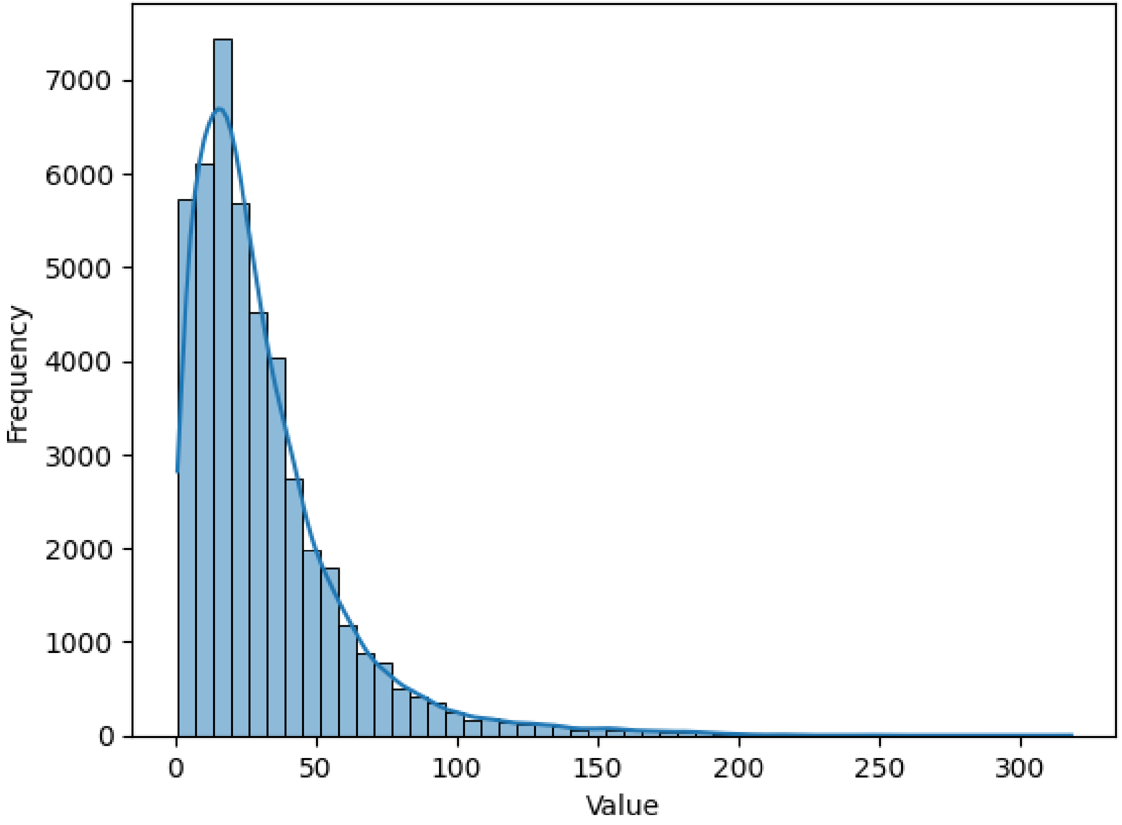
- Spatial Information from Latitude and Longitude values of Start Stations is obtained by scaling the latitude and longitude values of that station.

Target Vector (Daily demand) Pre-processing

Frequency Distribution of target vector (Daily Demand)

We converted the numerical Demand variable into Categorical variable (High, Medium, Low) based on the frequency distribution curve.

Demand less than 25 percentile (13) is considered as low and Demand greater than 75 percentile (41) is considered as high demand and the rest as medium demand.



Final Data for Model:

Feature : 23	Target Variable : 1
Start Latitude	Demand (Low, Medium, High)
Start Longitude	
PRCP	
SNOW	
TMAX	
TMIN	
Day of the Week (6 features)	
Day of the Month (11 features)	

Parameter	Dimension
X_train	(36576, 23)
Y_train	(36576,1)
X_Val (subset of training)	(7315, 23)
Y_Val (subset of training)	(7315, 1)
X_test	(9144, 23)
Y_test	(9144,1)

Final Data for Model:

Rephrasing Problem Statement: By inputting specific data points such as a Day of the week, Month of the year at a specific start station location(Latitude and Longitude), with precipitation, snow, and temperature our model can classify whether the demand for bikes will be low, medium, or high.

Sample Data:

Start_lat	Start_lng	PRCP	Snow	TMAX	TMIN	Monday	Saturday	...	October	September	Daily Demand
0.1979	0.2312	0.01	0	72	59	0	0	…	0	0	med
0.969396	0.5722	0.48	0	40	23	0	1	…	0	0	med

Baseline Learning:

Model Used: Logistic Regression (multi label classification problem)

Setting: Multinomial, to handle multiple classes (Low, Medium, High demand)

Solver Used: 'lbfgs', the algorithm for optimizing the models.

Data: Used the pre-processed X_train data to fit the model and to evaluate the model's performance used X_test.

Accuracy: 0.6075 (ie) **60.75%**

Confusion Matrix:

Actual \ Predicted	High Prediction	Mid Prediction	Low Prediction
Actual High	785	35	1433
Actual Mid	55	882	1190
Actual Low	445	431	3888

Deep Learning Model: GRU (Gated Recurrent Unit) based Recurring Neural Network

Architecture:

GRU Layer:

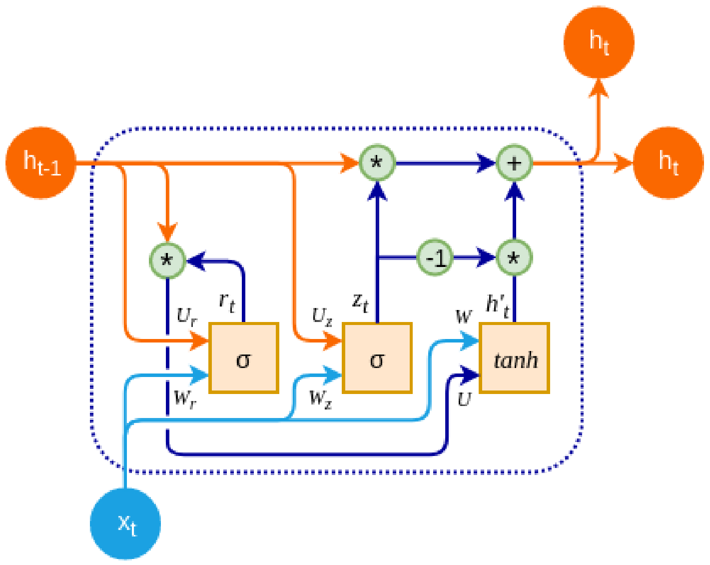
- Effectively handle time-series predictions for bike demand through sequential data processing .
- Its gating mechanism memorize and utilize important information throughout the sequence while discarding the irrelevant things.
- Has two Sigmoid and one Tanh activation functions.

Dropout Layer: A dropout is used for regularization, reducing the overfitting.

Linear Layer: A linear layer at the end is used to map the output of GRU to desired output(3 (ie) Low, Mid, High)

Accuracy: 77.3% at 600 Epoch

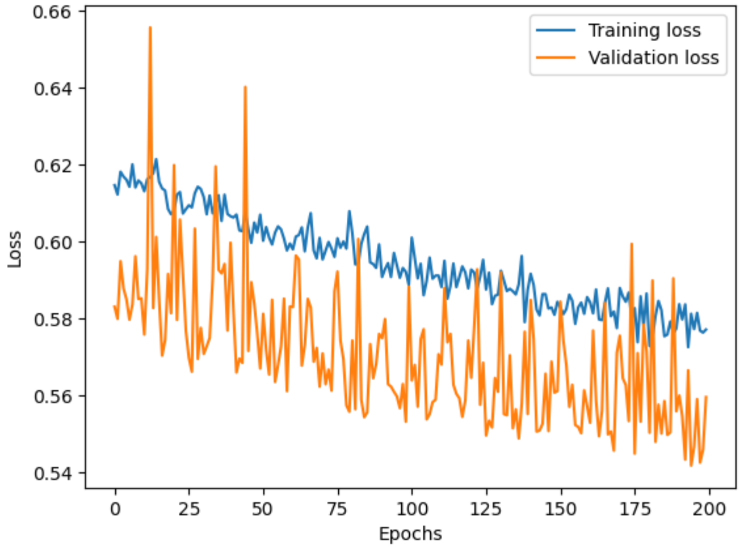
Parameter Name	Description	Value
input_size	The number of features	23
hidden_size	Number of features in the hidden state of the GRU.	50
output_size	The size of the output.	3
num_layers	Number of GRU layers in the network.	2
dropout_rate	Dropout rate to prevent overfitting	0.2
batch_size	Batch size for each epoch	64
Optimizer	Adam, SGD	
Loss function	Cross Entropy Loss	



Hyperparameter Tuning:

Batch Size : 64							
Adam Optimizer				SGD Optimizer			
Learning rate : 0.001		Learning rate : 0.01		Learning rate : 0.001		Learning rate : 0.01	
Epochs	Test Accuracy	Epochs	Test Accuracy	Epochs	Test Accuracy	Epochs	Test Accuracy
100	72%	100	66.6%	100	67.9%	100	55.3%
300	75%	300	65.1%	300	54.0%	300	56.8%
600	77.3%	500	65.1%	500	55.0%	500	45.1%

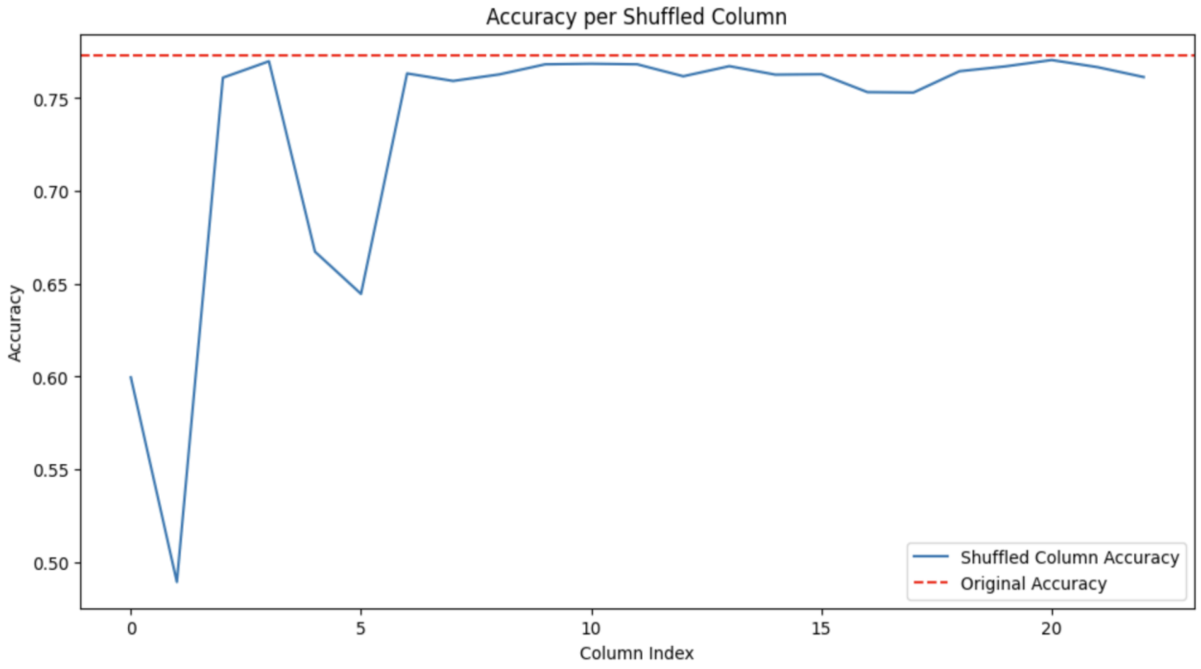
Different Batch Size (Adam Optimizer)			
Batch Size : 32		Batch Size : 128	
Learning rate : 0.001		Learning rate : 0.001	
Epochs	Test Accuracy	Epochs	Test Accuracy
100	72%	100	69.6%
300	76.69%	300	74.6%
500	76.35%	500	75.78%



Feature Importance

Permutation Feature Importance

- Shuffle each of 23 column and calculate their confusion matrix and accuracy of already trained model over test dataset.
- If shuffling a feature significantly decreases accuracy of trained model, that feature is considered important for the model.

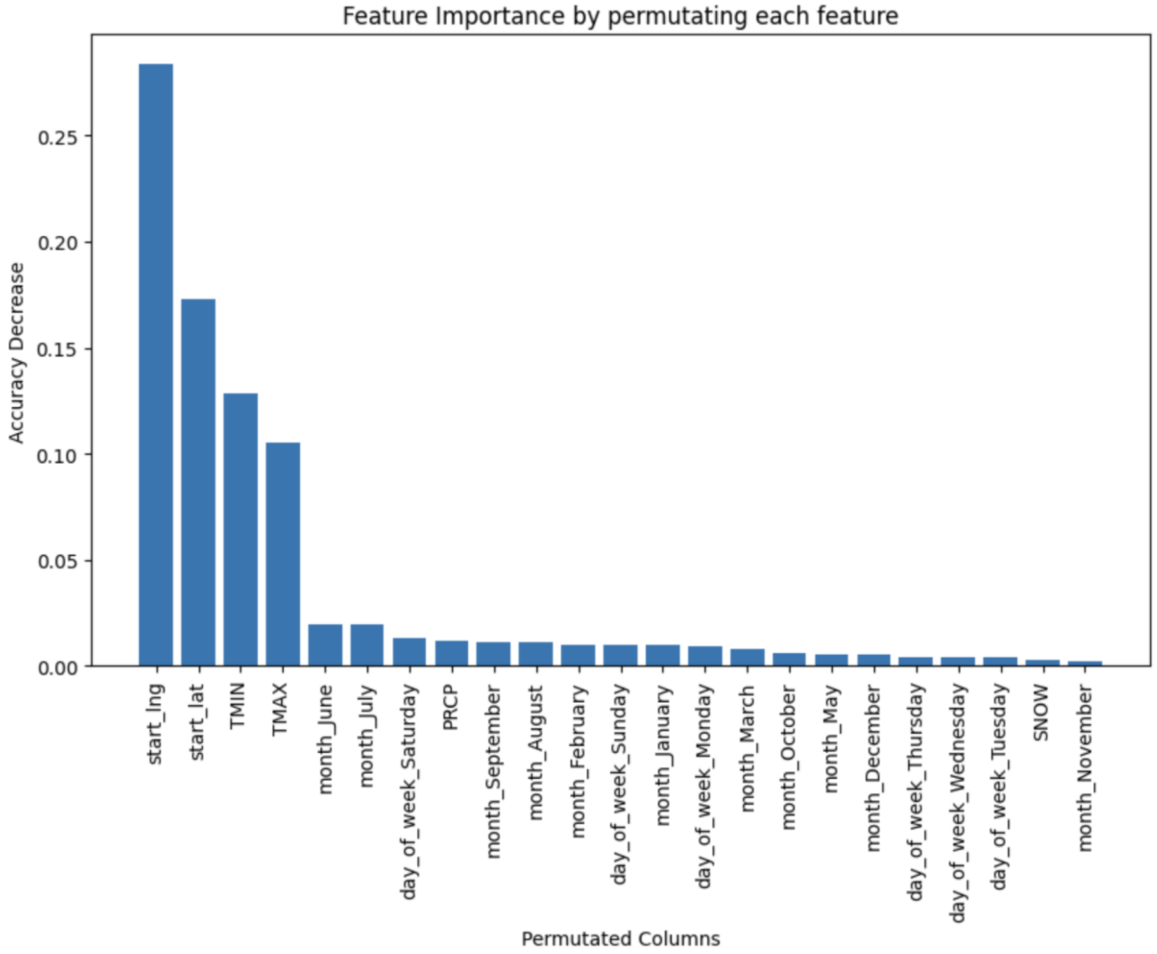


Feature Importance:

Visualize the decrease in accuracy as the measurement of feature importance.

Identification of Important features:

- Longitude of Starting point
- Latitude of Starting point
- Minimum of Temperature
- Maximum of Temperature



Challenges faced:

1)Limited Computational Resources:

Constrained resources restricted our ability to thoroughly explore hyperparameter settings, potentially limiting model optimization.

2)Model Sensitivity to Station Proximity and Outliers:

The model was notably sensitive to the station's geographic layout and required the exclusion of outliers, posing a challenge in terms of data robustness and generalizability.

3)Categorization of Skewed Demand Data:

The decision to categorize skewed demand data into three levels—Low, Medium, and High—posed a challenge, as it oversimplified the complex distribution.

Business Value for possible stakeholders

- **Optimized Fleet Management:** By accurately predicting bike demand, companies can efficiently manage their fleet. For instance, during periods of high demand, additional bikes can be allocated to the station, and during low demand, the excess bikes can be redistributed to other locations.
- **Maintenance and Staffing Schedules:** Predicting demand allows for better planning of maintenance and staffing. High-demand periods might require more staff for customer service and bike maintenance, while lower-demand periods could see reduced staffing, optimizing labor costs.
- **Dynamic Pricing Strategy:** Companies could use demand predictions to implement dynamic pricing strategies. During peak demand times, prices could be slightly increased to manage demand and optimize revenue. Conversely, lower prices during off-peak times could attract more users, balancing overall usage and maintaining consistent revenue streams.

Conclusion and Improvements

We were able to successfully develop a model to estimate the category of the daily demand for a given station, on a given day. (Example: The Daily demand on 12/11/2023(December, Sunday) at Astor Place is – High). This model is going to be very important and crucial for the stakeholders mentioned previously.

Best Model: GRU, Adam optimizer, lr = 0.001, Epochs = 600, Batch Size = 64

Despite the unconventional approach of using only latitude and longitude to capture spatial information of stations, our model demonstrated effective predictive performance of 77.3% Accuracy.

Improvements:

Exploring GNNs could offer a more sophisticated method to integrate spatial data, potentially enhancing model accuracy.

THANK YOU!