Comparative Analysis of Machine Learning Models of Delivery Time Prediction

Group 6

Sai Soumya Aloor
Sri Charan Desetty
Venkata Naga Anirudh Chaganti
Anantha Prahlada Kurudi
Hemanth Varma Pericherla

CONTENTS

Introduction

- For our final capstone, our team applied the complete workflow—data sourcing, cleaning, exploratory data analysis, modeling, and interpretation to a real-world problem: predicting food delivery times.

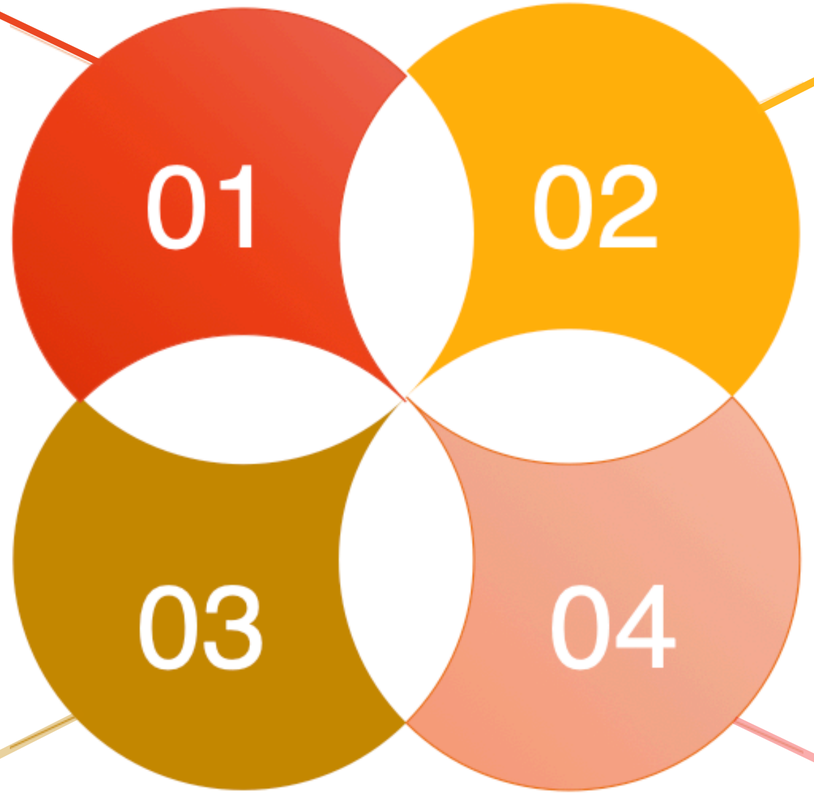
- Timely deliveries are critical for customer satisfaction and operational efficiency in today's logistics-driven economy.

- Our goal was to extract insights and build predictive models that can help optimize delivery operations, demonstrating both our technical skills and the power of data-driven decision making.

Reasons for Dataset Selection

Relevant Variables

The dataset contains several factors that directly influence delivery time.



Real World Application

The dataset closely mirrors real- world delivery scenarios, making it highly applicable to industries like logistics, food delivery, and e- commerce.

Variety of Data Types

It includes both numerical features (e.g., temperature, delivery distance, delivery time) and categorical features (e.g., type of vehicle, type of order, weather conditions).

Predictive Modeling Potential

Since the target variable delivery time is continuous, we can apply various regression algorithms and explore time series patterns.

LIST OF ATTRIBUTES

1) ID : A unique identifier for each delivery.
2) Delivery_person_ID : A unique identifier assigned to each delivery person for tracking purposes.
3) Delivery_person_Age : Age of the delivery person.
4) Delivery_person_Ratings : Customer ratings of the delivery person.
5) Restaurant_latitude : Geographical latitude coordinate of the restaurant's location.
6) Restaurant_longitude : Geographical longitude coordinate of the restaurant's location.
7) Delivery_location_latitude : Latitude coordinate of the delivery location where the order is delivered.
8) Delivery_location_longitude : Longitude coordinate of the delivery location for the order.
9) Type_of_order : Category of food ordered (e.g., meal, snacks ,drinks,buffet) to analyze preparation times.
10) Type_of_vehicle : The vehicle used for delivery (e.g., scooter , motor cycyle , cycle ,ev scooter), which affects speed and travel time.
11) Temperature : Ambient temperature during the delivery time, potentially impacting delivery efficiency.
12) Humidity : Level of atmospheric moisture during delivery, affecting conditions for travel.
13) Precipitation : Amount of rainfall or snow, indicating weather disruptions during delivery.
14) Weather_description : Textual description of the weather (e.g., sunny, cloudy, stormy) for context in travel conditions.
15) Traffic_Level : Severity of traffic congestion during the delivery (e.g., low, medium, high).
16) Distance (km) : The calculated distance between the restaurant and the delivery location in kilometers.
17) TARGET : The target variable representing the delivery time in minutes for model predictions.

Data collection, preprocessing and data cleaning

DATA COLLECTION

1 **.info() :** We applied this to understand the datatype of each attribute.

```
Data columns (total 18 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   ID                           10000 non-null  object
 1   Delivery_person_ID           10000 non-null  object
 2   Delivery_person_Age          10000 non-null  float64
 3   Delivery_person_Ratings      10000 non-null  float64
 4   Restaurant_latitude          10000 non-null  float64
 5   Restaurant_longitude         10000 non-null  float64
 6   Delivery_location_latitude   10000 non-null  float64
 7   Delivery_location_longitude  10000 non-null  float64
 8   Type_of_order                10000 non-null  object
 9   Type_of_vehicle              10000 non-null  object
 10  temperature                  9995 non-null   float64
 11  humidity                     9995 non-null   float64
 12  precipitation                9995 non-null   float64
 13  weather_description          9995 non-null   object
 14  Unnamed: 14                  0 non-null      float64
 15  Traffic_Level                9085 non-null   object
 16  Distance (km)                9080 non-null   float64
 17  TARGET                       9459 non-null   object
dtypes: float64(11), object(7)
memory usage: 1.4+ MB
None
```

Inference:

Numerical columns: 11 columns are of type float64, indicating continuous or numeric data such as age, ratings, weather conditions, and distance.

Categorical columns: 7 columns are of type object, representing identifiers and categorical variables like Type_of_order, Type_of_vehicle, and Traffic_Level.

2 **.describe() :** We applied this to understand the stats of each attribute.

```
       Delivery_person_Age  Delivery_person_Ratings  Restaurant_latitude  \
count         10000.000000             10000.000000         10000.000000
mean             29.522000                 4.629370            16.893418
std               5.700348                 0.322941             8.330948
min              15.000000                 1.000000           -30.902872
25%              25.000000                 4.500000            12.913041
50%              29.000000                 4.700000            18.546258
75%              34.000000                 4.800000            22.727021
max              50.000000                 6.000000            30.914057

       Restaurant_longitude  Delivery_location_latitude  \
count          10000.000000                10000.000000
mean              70.177749                   17.412655
std               23.203352                    7.336846
min              -88.352885                    0.010000
25%               73.170937                   12.983959
50%               75.902847                   18.626216
75%               78.047717                   22.785089
max               88.433452                   31.054057

       Delivery_location_longitude  temperature     humidity  precipitation  \
count                 10000.000000  9995.000000  9995.000000    9995.000000
mean                     70.880072    22.936907    66.164882       0.016233
std                      21.174585     3.379448    15.602939       0.074911
min                       0.010000     6.770000    27.000000       0.000000
...
25%                            NaN     7.620000
50%                            NaN    13.400000
75%                            NaN    19.610000
max                            NaN    59.840000
```

Inference:

Age & Ratings: Most delivery persons are aged 25–34 with ratings between 4.5–5. Some outliers exist.

Weather: Avg temp ~23°C, humidity ~66%, and low precipitation overall.

Location: Most lat/long values are normal, but some are outliers.

Distance: Typical delivery is around 13 km, max ~60 km.

3 **.value_columns() :** We applied this to understand what is the number of unique values in each attribute.

```
data['Type_of_vehicle'].value_counts()
✓ 0.0s

motorcycle          5862
scooter             3304
electric_scooter     814
bicycle               20
Name: Type_of_vehicle, dtype: int64
```

```
data['Type_of_order'].value_counts()
✓ 0.0s

Snack      2551
Meal       2530
Drinks     2507
Buffet     2412
Name: Type_of_order, dtype: int64
```

Inference :

Top Delivery Persons: Some IDs appear up to 22 times, indicating frequent assignments.

Order IDs: Mostly unique, but a few duplicates and anomalies like scientific notation or malformed IDs (e.g., "BEF 1.00").

Order Type: Fairly balanced — Snacks (2551), Meals (2530), Drinks (2507), Buffet (2412).

Vehicle Type: Motorcycles dominate (5862), followed by scooters; very few bicycles.

Weather: Mostly clear sky, haze, and mist. Rare cases of rain or fog.

Traffic Levels: Majority face high to moderate traffic; very low traffic is least common.

TARGET Issues: 419 values are invalid (`#VALUE!`). The rest vary widely from ~9 to 99 minutes, with many repeated durations.

PREPROCESSING AND DATA CLEANING

1 Clearing NaN values

```
ID                           1
Delivery_person_ID           1
Delivery_person_Age          1
Delivery_person_Ratings      1
Restaurant_latitude          1
Restaurant_longitude         1
Delivery_location_latitude   1
Delivery_location_longitude  1
Type_of_order                1
Type_of_vehicle              1
temperature                  6
humidity                     6
precipitation                6
weather_description          6
Unnamed: 14                  10001
Traffic_Level                916
Distance (km)                921
TARGET                       542
dtype: int64
```

Dropping NaN values

.dropna(inplace=True)

```
ID                           0
Delivery_person_ID           0
Delivery_person_Age          0
Delivery_person_Ratings      0
Restaurant_latitude          0
Restaurant_longitude         0
Delivery_location_latitude   0
Delivery_location_longitude  0
Type_of_order                0
Type_of_vehicle              0
temperature                  0
humidity                     0
precipitation                0
weather_description          0
Traffic_Level                0
Distance (km)                0
TARGET                       0
dtype: int64
```

PREPROCESSING AND DATA CLEANING

2

Next step we have taken to calculate the distance between the Restaurant and the delvery address using longitude and latitude.

temperature	humidity	precipitation	weather_description	Traffic_Level	Distance (km)	TARGET	Calculated_Distance
19.50	93.0	0.0	mist	Very High	37.17	85.26666667	20.183530
20.45	91.0	0.0	mist	Low	3.34	28.58333333	1.552758
23.86	78.0	0.0	mist	Moderate	10.05	35.18333333	7.790401
26.55	87.0	0.0	mist	High	9.89	43.45	6.210138
21.43	65.0	0.0	broken clouds	Moderate	11.30	30.6	4.610365
...	...	...	...	...	...	...	...
28.03	57.0	0.0	smoke	Low	3.78	18.2	1.529877
23.96	64.0	0.0	haze	High	18.92	32.61666667	13.631344
22.94	60.0	0.0	haze	Low	2.64	12.01666667	1.536621
23.72	31.0	0.0	clear sky	Very High	28.80	51.06666667	20.851557
28.01	57.0	0.0	smoke	High	17.63	43.8	13.771133

PREPROCESSING AND DATA CLEANING

3

Next we have removed 5 columns and created a new Data frame as df. Since "ID","Delivery_person_ID", "Restaurant_latitude", "Restaurant_longitude", “Delivery_location_latitude”, “Delivery_location_longitude" are no longer needed.

```
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Delivery_person_Age    9035 non-null   float64
 1   Delivery_person_Ratings  9035 non-null   float64
 2   Type_of_order          9035 non-null   object
 3   Type_of_vehicle        9035 non-null   object
 4   temperature            9035 non-null   float64
 5   humidity               9035 non-null   float64
 6   precipitation          9035 non-null   float64
 7   weather_description    9035 non-null   object
 8   Traffic_Level          9035 non-null   object
 9   Distance (km)          9035 non-null   float64
 10  TARGET                 9035 non-null   object
 11  Calculated_Distance    9035 non-null   float64
dtypes: float64(7), object(5)
```

```
Data columns (total 24 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Delivery_person_Age                   9035 non-null   float64
 1   Delivery_person_Ratings               9035 non-null   float64
 2   temperature                           9035 non-null   float64
 3   humidity                              9035 non-null   float64
 4   precipitation                         9035 non-null   float64
 5   Traffic_Level                         9035 non-null   int64
 6   Distance (km)                         9035 non-null   float64
 7   TARGET                                9035 non-null   float64
 8   Calculated_Distance                   9035 non-null   float64
 9   Type_of_order_Drinks                  9035 non-null   uint8
 10  Type_of_order_Meal                    9035 non-null   uint8
 11  Type_of_order_Snack                   9035 non-null   uint8
 12  Type_of_vehicle_electric_scooter      9035 non-null   uint8
 13  Type_of_vehicle_motorcycle            9035 non-null   uint8
 14  Type_of_vehicle_scooter               9035 non-null   uint8
 15  weather_description_clear sky         9035 non-null   uint8
 16  weather_description_few clouds        9035 non-null   uint8
 17  weather_description_fog               9035 non-null   uint8
 18  weather_description_haze              9035 non-null   uint8
 19  weather_description_mist              9035 non-null   uint8
...
 22  weather_description_scattered clouds  9035 non-null   uint8
 23  weather_description_smoke             9035 non-null   uint8
dtypes: float64(8), int64(1), uint8(15)
```

We converted them to numerical or structured format using One-hot and ordinal encoding. values in TARGET column must be converted to numeric as the values are in numeric values we did it by using (df['TARGET'] = pd.to_numeric(df['TARGET'], errors=‘coerce')

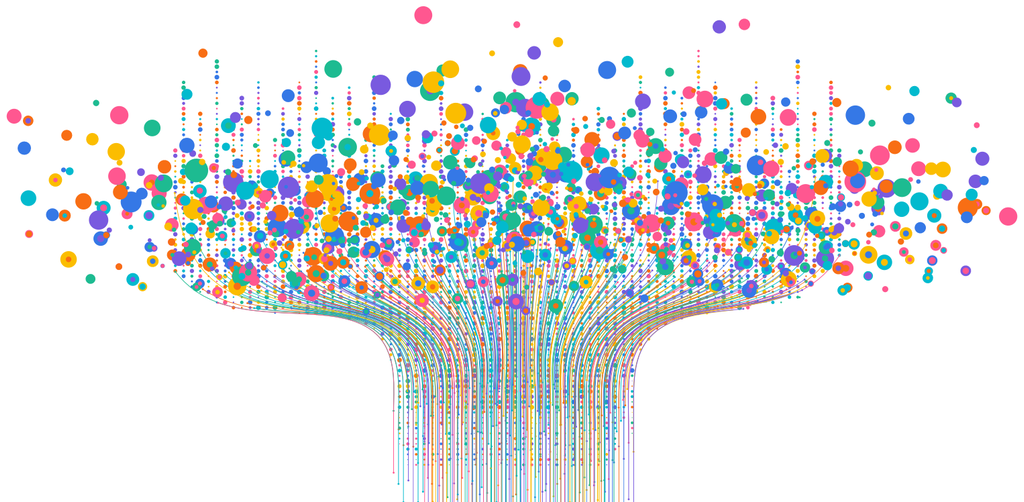
PREPROCESSING AND DATA CLEANING

4

At the end of Preprocessing and cleaning of data we get the below attributes with No NaN values and all the attributes containing numeric values.

```
Delivery_person_Age                          0
Delivery_person_Ratings                      0
temperature                                  0
humidity                                     0
precipitation                                0
Traffic_Level                                0
Distance (km)                                0
TARGET                                       0
Calculated_Distance                          0
Type_of_order_Drinks                         0
Type_of_order_Meal                           0
Type_of_order_Snack                          0
Type_of_vehicle_electric_scooter             0
Type_of_vehicle_motorcycle                   0
Type_of_vehicle_scooter                      0
weather_description_clear sky                0
weather_description_few clouds               0
weather_description_fog                      0
weather_description_haze                     0
weather_description_mist                     0
weather_description_moderate rain            0
weather_description_overcast clouds          0
weather_description_scattered clouds         0
weather_description_smoke                    0
dtype: int64
```

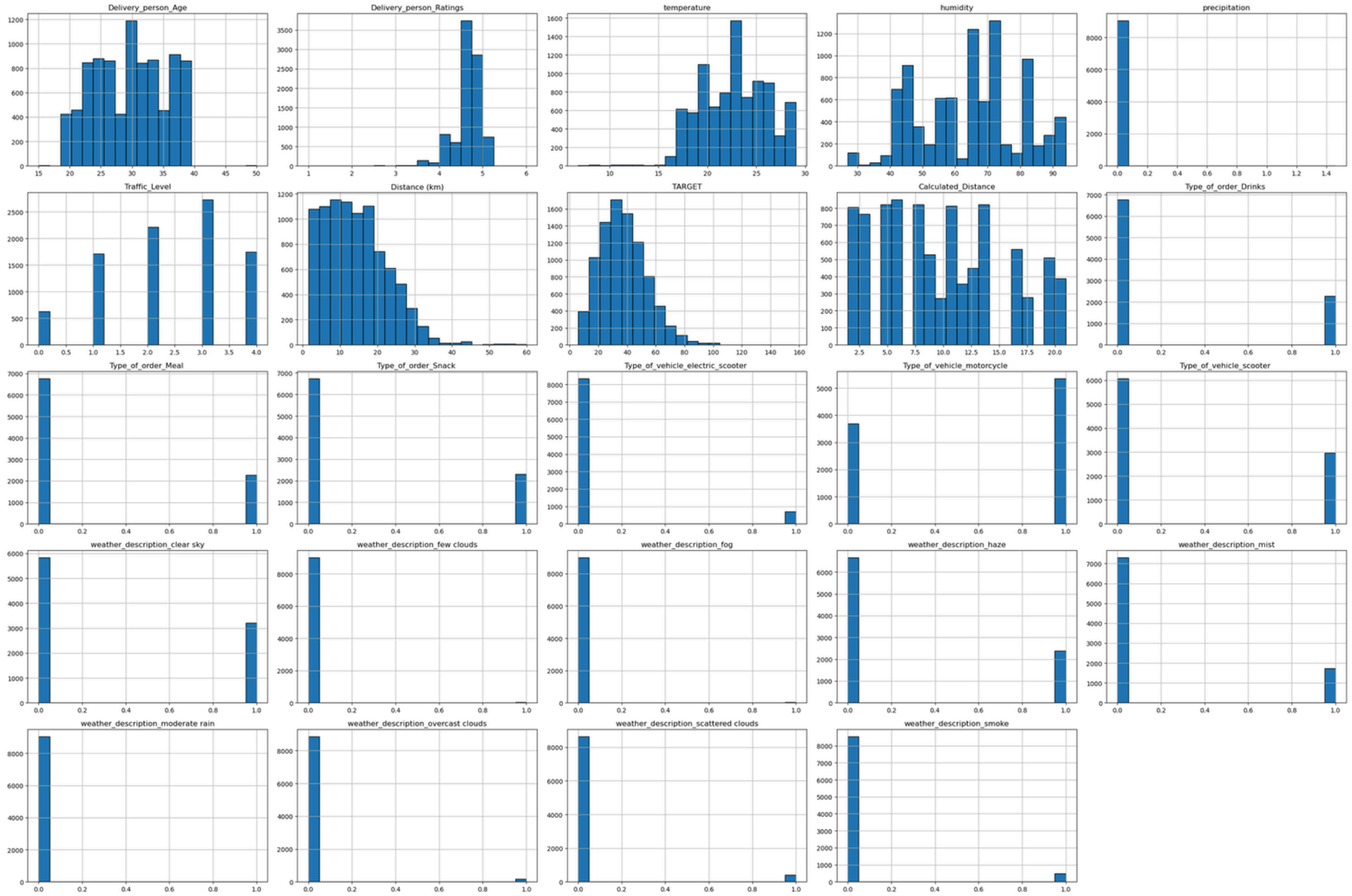
Summarizing Data and Constructing Data Visualizations



Summarising the Data

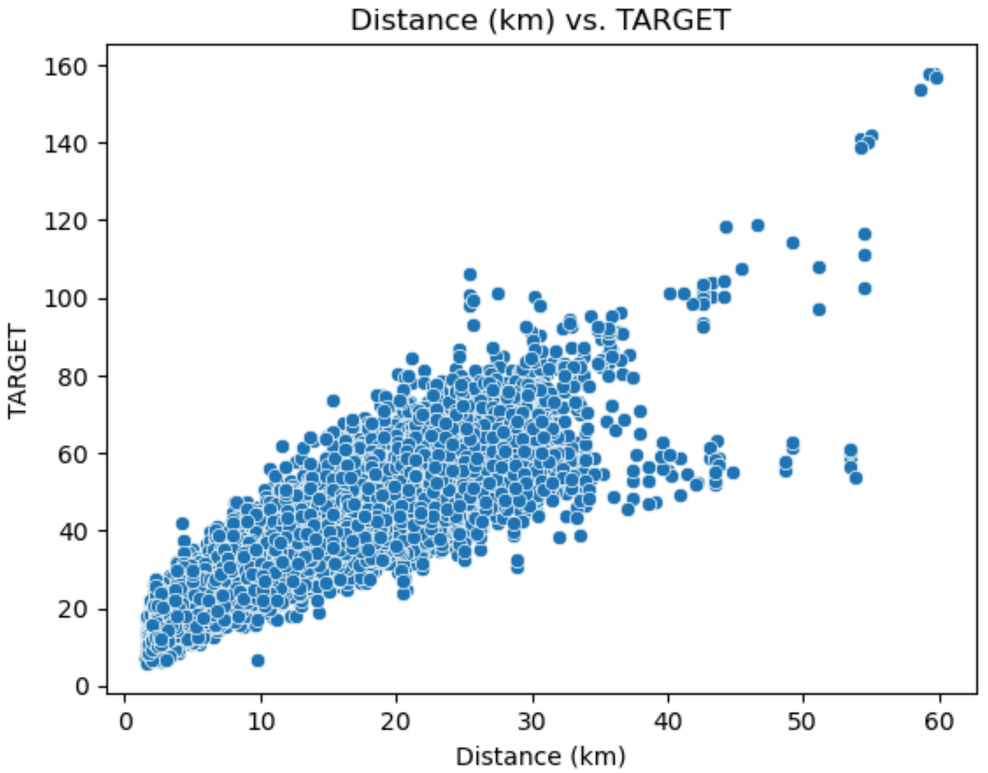
- The dataset provides useful details on various factors affecting delivery times and conditions, such as delivery person attributes, traffic levels, weather conditions, and order types.
- The delivery persons are relatively young and highly rated. Orders are more commonly meals and snacks, while motorcycles are the most frequently used vehicles.
- The weather data indicates moderate temperature and humidity, with clear skies being the most common weather condition.
- The dataset’s wide range of delivery distances and traffic levels highlights the variability in delivery conditions.
- This summary provides an overview of the data's structure and key variables, which could guide further analysis or model building.

Visualizations: Univariate Visualzations and Key Insights



- Delivery_person_Age: Most delivery persons are aged between 25 and 34, with a mean age of 29.5. The distribution is symmetric with a few outliers (ages between 11.5 and 47.5)
- Delivery_person_Ratings: Ratings are concentrated between 4.5 and 5.0, with a peak at the higher end. The distribution is skewed left, indicating more higher ratings and some outliers.
- Temperature: Temperatures mostly range from 19°C to 25°C, with a mean of 22.6°C. The distribution is roughly normal, with some outliers in the higher range.
- Humidity: Humidity values range from 27% to 94%, with a peak around 65%. It has a slightly negative skew, indicating lower values are more frequent.
- Precipitation: Most values are 0 (no precipitation), with a very small proportion showing higher values. The data is highly skewed to the right.
- Traffic_Level: Traffic levels mostly range from 1 to 3, with a peak at 2 (moderate traffic). The distribution is slightly left skewed.
- Distance (km): Distances mostly range from 1.5 km to 19.6 km, with a mean of 14.3 km. The distribution is slightly right skewed, with some long distance deliveries as outliers.
- TARGET: Delivery times (TARGET) range from 5.8 to 157.75 minutes, with a mean of 37.65 minutes. It has a right skewed distribution, indicating most deliveries take less time but some outliers take much longer.
- Calculated_Distance: The calculated distances range from 1.47 km to 20.97 km, with a mean of 9.7 km. The distribution is normal with no significant outliers.

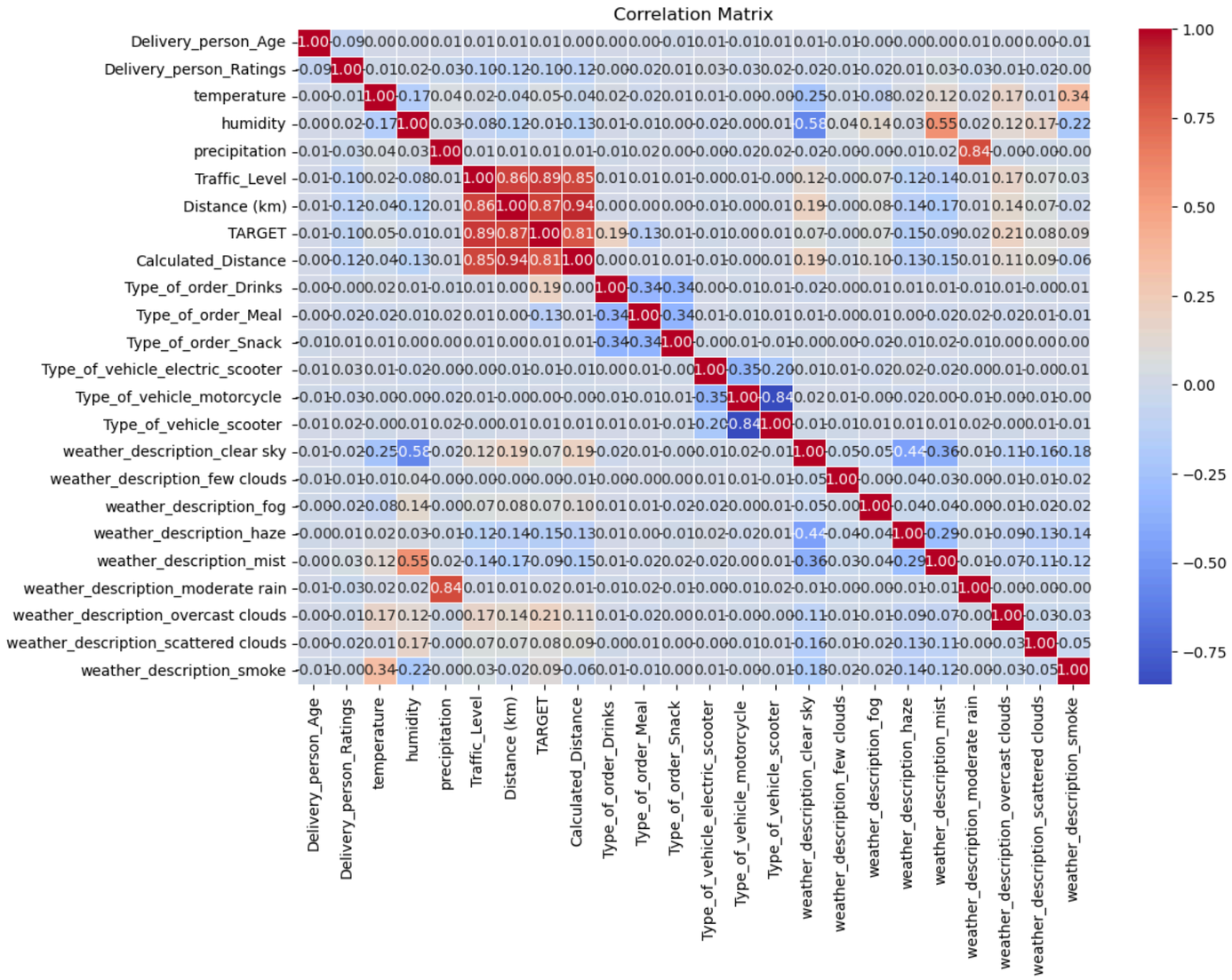
Visualizations: Bivariate Visualzations and Key Insights



1. Delivery_person_Age: Weak correlation (0.01) with `TARGET`. Regression shows it's not a significant predictor (Pvalue = 0.583), with a negligible effect on `TARGET` (coefficient = 0.0168).

2. Distance (km): Strong correlation (0.87) with `TARGET`. Regression indicates a significant relationship (Pvalue = 0.000), with each additional kilometer increasing `TARGET` by 1.72 units.

Visualizations: Correlation Matrix



Delivery_person_Age: Slight negative correlation with delivery ratings, indicating older or younger delivery persons may have slightly lower ratings. Mild relationships with Traffic_Level and Distance, suggesting age influences delivery conditions to some extent.

Delivery_person_Ratings: Negative correlation with age, which could imply younger delivery persons tend to receive higher ratings. Slight correlations with Traffic_Level and Distance, suggesting performance may be affected by traffic conditions.

Temperature: Strong positive correlation with Humidity, and mild correlation with Traffic_Level. Negative correlation with clear skies and certain weather conditions, which may impact delivery performance.

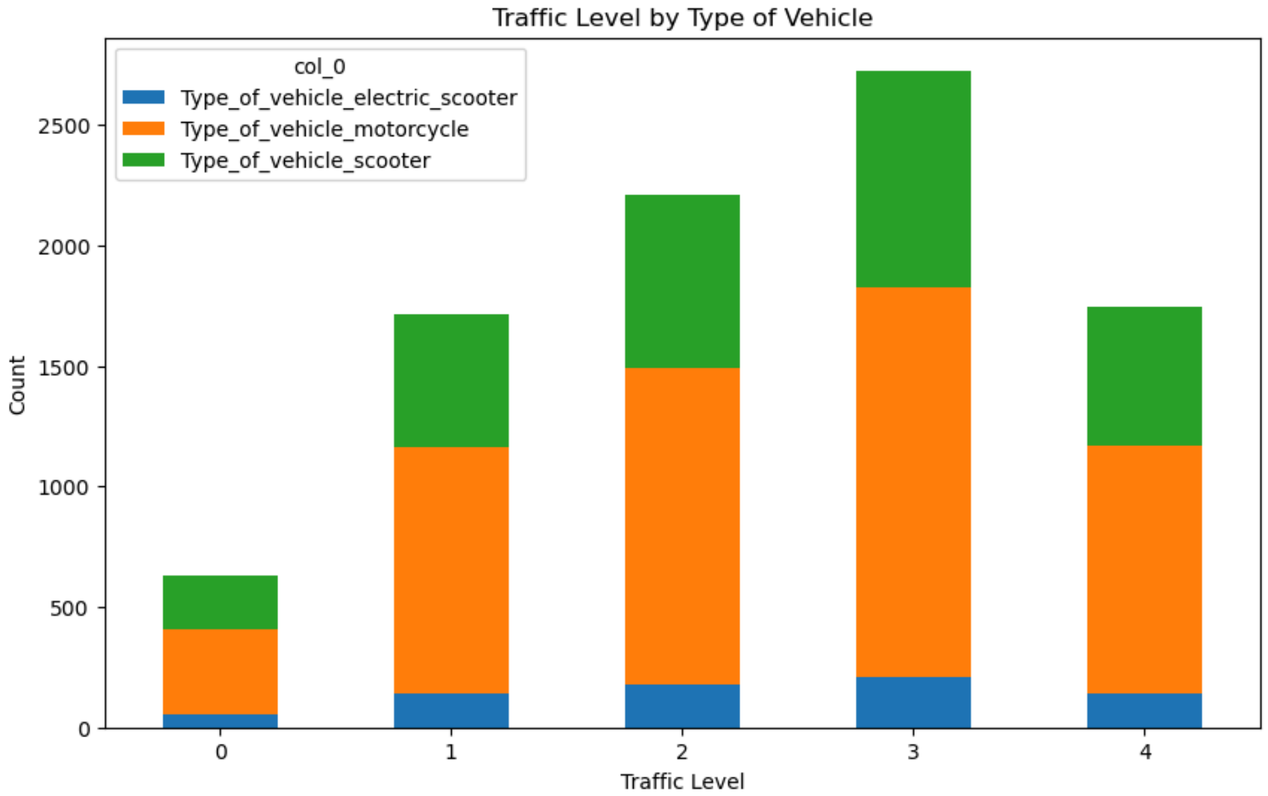
Precipitation: Strong positive correlation with Moderate Rain, and links to other weather types, indicating it affects delivery performance.

Traffic_Level: Strongly correlated with Distance (km), suggesting longer distances usually experience more traffic. Affects TARGET, which may indicate longer trips or delays.

Distance (km): Correlates positively with Traffic_Level and Calculated Distance, impacting delivery time or efficiency.

TARGET: Correlates with Distance and Traffic_Level, showing that delivery times increase with longer distances and higher traffic. Weather conditions, especially fog and haze, also impact this outcome.

Visualizations: Categorical Visualizations



- Motorcycles dominate across all traffic levels, especially in higher traffic conditions.
- Scooters show increased usage at moderate to high traffic levels, peaking at level 3.
- Electric scooters remain the least used across all traffic levels.

Delivery services prefer motorcycles, likely for their speed and flexibility. Electric scooters may face limitations like lower range or availability.

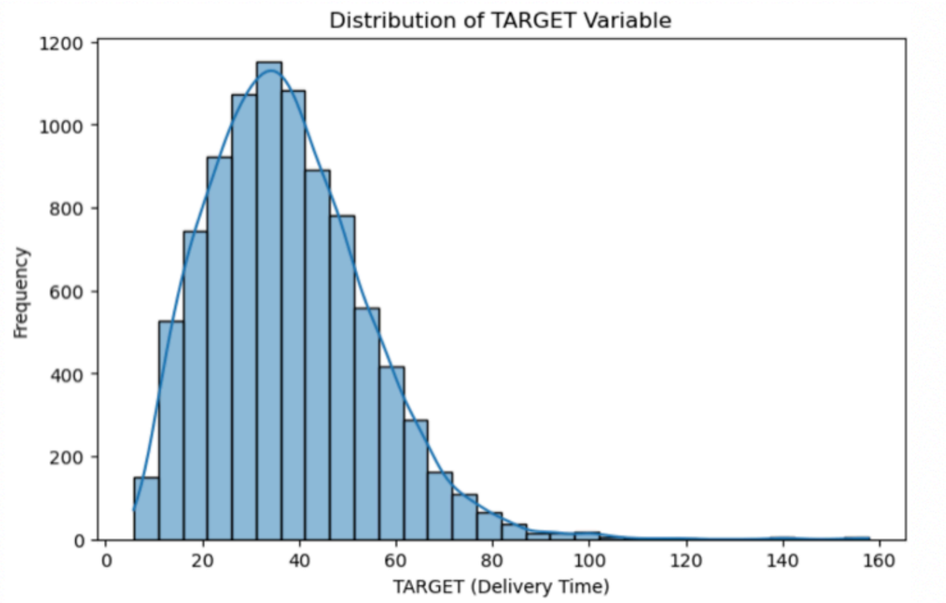
Various methods, tools to analyse the data to develop data models

1 Classification Analysis

- Since TARGET is continuous, we eliminated classification methods (e.g., Logistic Regression, Naive Bayes, Decision Trees), which are best suited for categorical targets.
- TARGET is continuous float64, confirming that classification is not suitable.
- Mean: ~37.65 min, Median: ~35.98 min, Std Dev: ~16.55 min.
- Right skewed distribution with most deliveries between 2050 min.

Graph Analysis :

- The histogram shows a peak around 3040 min, suggesting common delivery times.
- The right skew confirms that longer deliveries are less frequent.
- Classification models (e.g., Logistic Regression, Naïve Bayes, Decision Trees) are designed for categorical targets (e.g., "Late" vs. "Ontime"), not for predicting continuous values.
- Our goal is to predict exact delivery times, which requires regression models instead of assigning labels.
- Regression models (Linear, SVR, KNN) are more appropriate for prediction.



2 Clustering Analysis (K Means VS DBSCAN)

	K Means	DBSCAN
Silhouette Score	0.3429 (Moderate Cluster Separation)	**0.2021** (lower separation, more noise handling)
No of Clusters	3 (predefined using Elbow Method, k=3)	**11** (automatically identified, includes small clusters)
Outlier Detection	NO	Yes (e.g., cluster label -1 for noise points)
Shape Flexibility	Assumes spherical clusters	Detects irregular shapes and densities
Best Use Case	Structured, dense data	Noisy data with outliers or unknown structure

- **KMeans is ideal for structured, dense data with a known number of clusters, but struggles with noise and irregular shapes.**
- **DBSCAN, on the other hand, excels with noisy data, irregular cluster shapes, and outlier detection, though it may show slightly lower cluster separation.**

3 Forecast and Time Series Analysis

KNN Regression	CART	SVR & GPR
• **Lowest MAE (4.39) → Smallest absolute errors.** • **Lowest MSE (35.19) & RMSE (5.93)→ Smallest squared errors.** • **Highest R^2 Score (0.86) → Explains 86% of variance, meaning strong predictive power.**	• RMSE is slightly higher (6.18), but still a decent model.	• SVR has a negative R^2 Score (0.25) → Means it performs worse than a simple mean predication. • GPR is the least performing model (5.41 R^2, 1666.43 MSE) → Huge prediction errors, likely due to overfitting or inappropriate kernel selection.

	Model	MAE	MSE	RMSE	R² Score
0	KNN Regression	4.388297	35.185505	5.931737	0.864743
1	CART Regression	4.749953	38.172757	6.178411	0.853260
2	SVR	14.969733	325.813190	18.050296	-0.252462
3	GPR	37.500507	1666.426312	40.821885	-5.405925

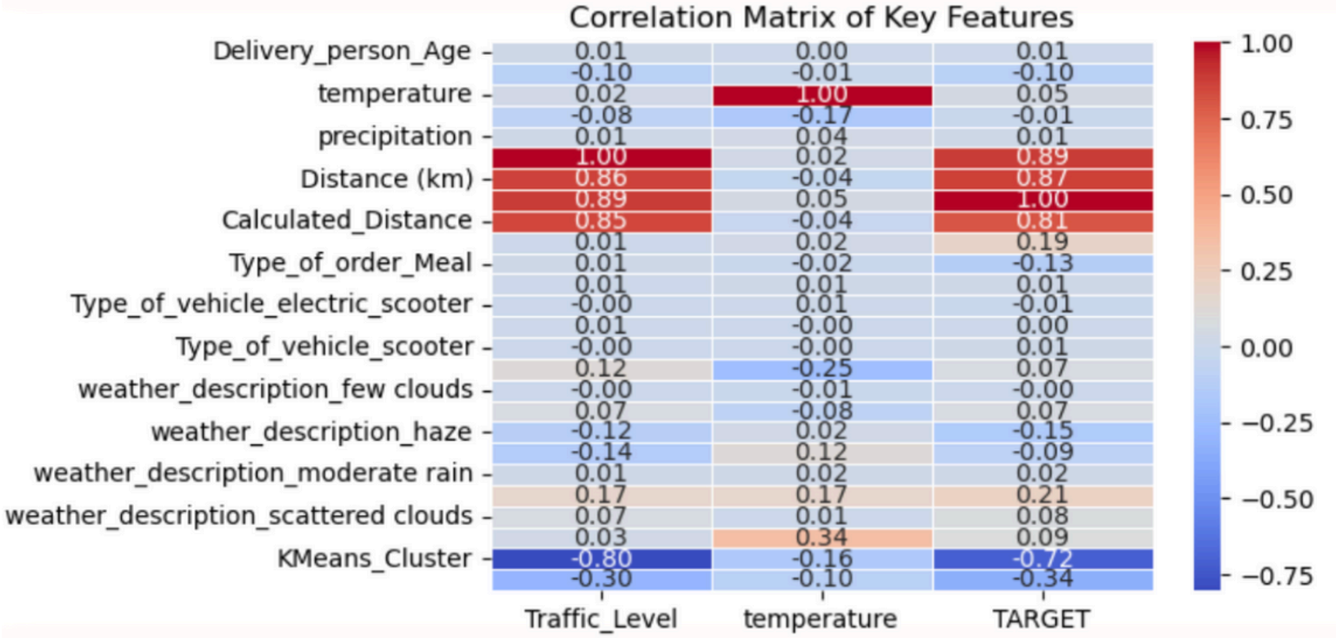
KNN Regression is the best model for predicting TARGET.

SVR and GPR should be eliminated or optimised.

4 Association Analysis Methods (Market Basket Analysis)

To find relationships between features:

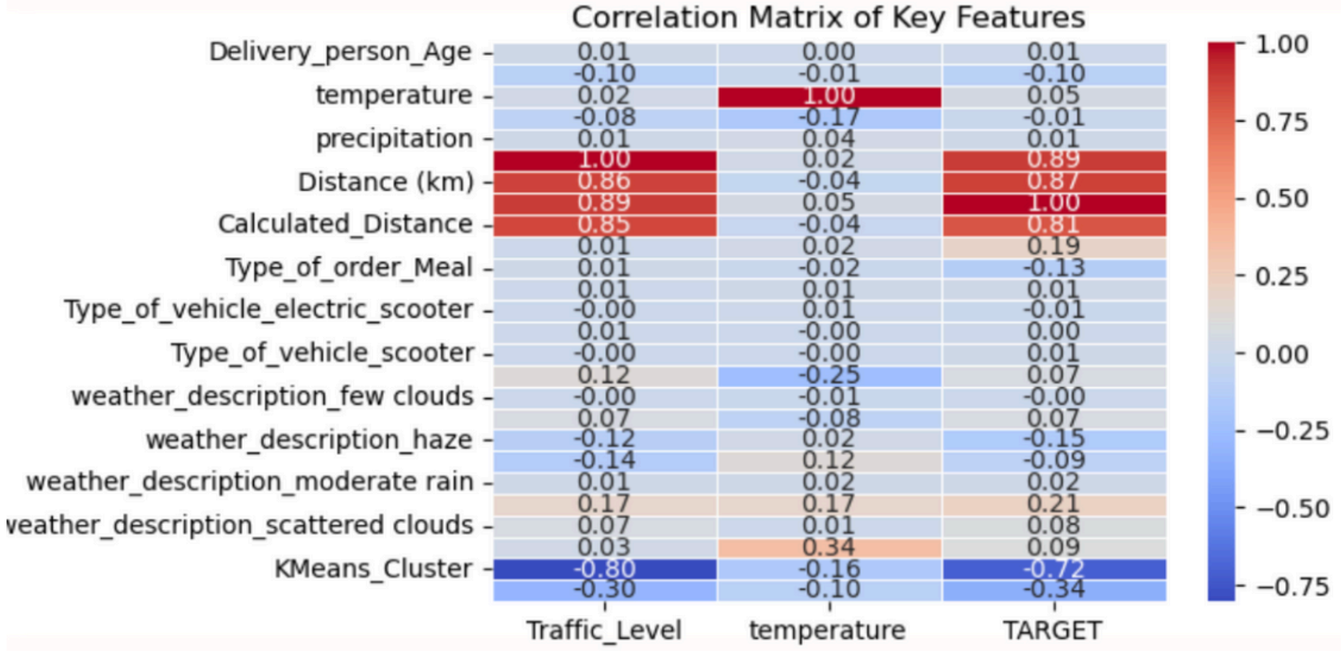
- **We want to use : Correlation Matrix to analyze how features like Traffic_Level and Temperature affect TARGET.**
- **We want to eliminate : NLP and Apriori, as they are not suitable for our dataset.**



Inference from the Correlation Matrix:

1. **Traffic_Level vs. TARGET (0.12)**
   - A weak negative correlation suggests that as `Traffic_Level` increases, `TARGET` (Delivery Time) slightly decreases.
   - This is unexpected—higher traffic is usually expected to increase delivery time. You might want to check data consistency.
2. **Temperature vs. TARGET (0.10)**
   - A weak positive correlation means that higher `Temperature` might slightly increase `TARGET`, but the effect is minimal.
   - This suggests weather conditions don't strongly affect delivery time in this dataset.
3. **Distance (km) vs. TARGET (0.87)**
   - A strong positive correlation shows that longer distances significantly increase delivery time, which is expected.

4 Association Analysis Methods (Market Basket Analysis)



4. **KMeans_Cluster vs. Traffic_Level (0.80)**

- Very strong negative correlation implies that the clustering algorithm has grouped data where higher traffic levels are in one cluster and lower in another.
- This indicates that `KMeans_Cluster` effectively separates traffic levels.

5. **KMeans_Cluster vs. TARGET (0.34)**

- A moderate negative correlation suggests that deliveries in certain clusters (possibly high traffic areas) tend to have lower `TARGET` values.
- This contradicts intuition, as one would expect higher traffic to delay deliveries.

The weak correlations between `Traffic_Level`, `Temperature`, and `TARGET` suggest other features (like `Distance (km)`) play a more significant role in determining delivery time.

Conclusion

- The goal of this project was to predict delivery time using four regression models: K-Nearest Neighbors (KNN), Decision Tree (CART), Support Vector Regressor (SVR), and Gaussian Process Regressor (GPR). We evaluated each model using Mean Squared Error (MSE) and R^2 Score.

- KNN performed best, with the lowest MSE (35.18) and highest R^2 score (0.8647), explaining about 86.5% of the variance. CART and SVR also gave good results but slightly underperformed compared to KNN. GPR performed poorly, with a high MSE (1069.02) and negative R^2 (–3.11), making it unsuitable for this task.

- In conclusion, KNN proved to be the most accurate and reliable model for predicting delivery time in our dataset.