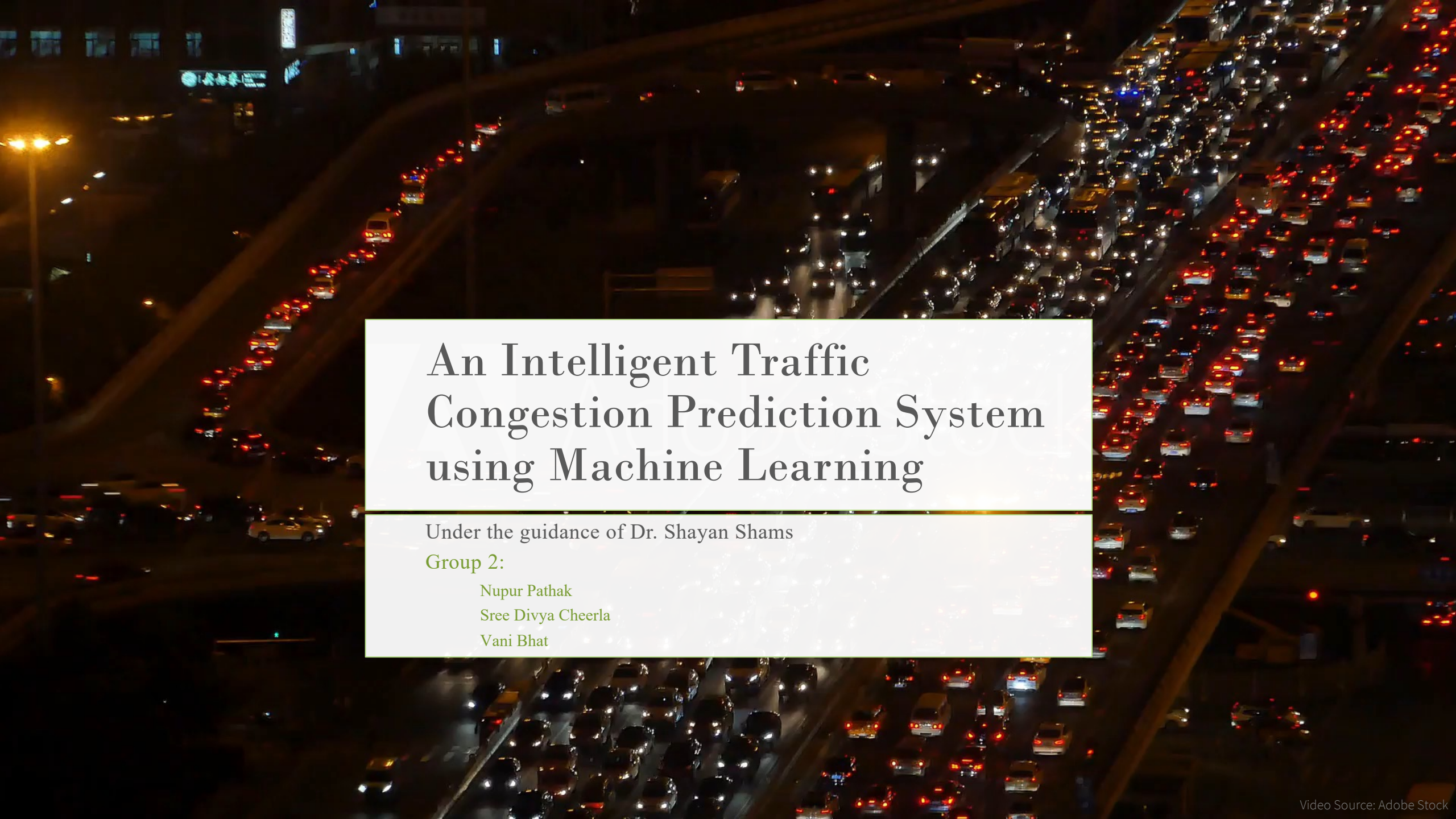


An Intelligent Traffic Congestion Prediction System using Machine Learning

Under the guidance of Dr. Shayan Shams

Group 2:

Nupur Pathak

Sree Divya Cheerla

Vani Bhat

Video Source: Adobe Stock

AGENDA

I. Introduction
II. Motivation and Background
III. Literature Review
IV. Methodology
V. Experimental Results
VI. Team Member and Contribution
VII. Conclusion and Future Work

INTRODUCTION

- Traffic congestion has become a pressing issue affecting urban areas worldwide.
- Leading to congestion, delays, and frustration for commuters and travelers.
- It is crucial to address this problem efficiently to enhance transportation systems and improve the quality of life for residents.

MOTIVATION AND BACKGROUND

- Traffic congestion is a complex phenomenon influenced by various factors
  1. Road infrastructure and Traffic volume
  2. Weather conditions, special events, and accidents.

- Traditionally, traffic management relied on static traffic models and manual data collection, which often resulted in inefficient responses to changing traffic patterns.

- Recent advancements in data mining and analytics have opened up the opportunity to revolutionize traffic prediction and management.

- Alleviate adverse effects of traffic congestion on society by predicting traffic congestion levels

- By accurately predicting traffic congestion in advance, can help develop proactive measures to optimize traffic flow, reroute vehicles, and provide real-time information to commuters, empowering them to make informed decisions.

LITERATURE REVIEW

Research Paper	Authors	Business Objective	Models Used	Performance Evaluation
Traffic Congestion Prediction Using Machine Learning Techniques	Yasir et al.(2022)	Proposed a prediction model for the traffic congestion that can predict congestion based on day, time and several weather data (e.g., temperature, humidity).	SVR (Support Vector Regressor)	RMSE
Gradient Boosting Approach for Traffic Flow Prediction using CatBoost	Singh et at. (2021)	Proposed an approach which considers important factors such as no. of intersections on the street, no. of commercial places near the street, and structure of the street.	BPNN, XGBoost, CatBoost, CatBoost with hyperparameter tuning	Accuracy Precision F1 score Recall
Prediction of Road Traffic Congestion Based on Random Forest	Liu1 et al.(2017)	Weather conditions, time period, special conditions of road, road quality and holiday are used as model input variables to establish road traffic forecasting model	Random Forest	Accuracy
Short term traffic flow prediction based on combination model of xgboost-lightgbm	Mei et al.(2018)	Proposed a combined prediction model where xgboost and lightgbm are constructed individually and later merged to generate final model.	Xgboost and lightgbm	MAPE (mean absolute percentage error)

METHODOLOGY

I. Data Understanding
   a. Data Quality Report – Continuous and Categorical Features
   b. Visualizations
II. Data Preparation
III. Data Modeling
   a. Random Forest
   b. XGBoost
   c. LightGBM
   d. CatBoost
IV. Model Evaluation

DATA UNDERSTANDING

Data Collection

- Traffic intersection congestion dataset consists of aggregated trip logging metrics from commercial vehicles.
- It contains 856k observations aggregating stopped vehicle information and intersection wait times.
- This information is captured at intersections in 4 major US cities: Atlanta, Boston, Chicago & Philadelphia.
- Weather information on temperature and rainfall of each city by month is collected from NCDC.

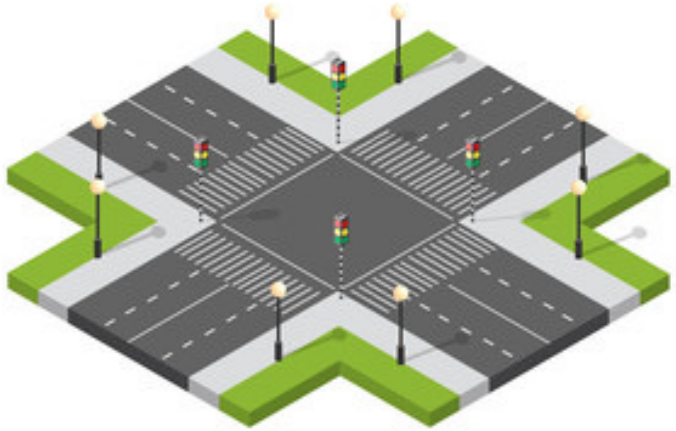
Data Import and ABT

- The traffic congestion dataset and weather dataset are appended
- Features:
  - Raw Features (28)
  - Derived Features (8)

Exploratory Data Analysis

- Generate Data Quality Report for Categorical and Continuous features
- Histogram/ Bar plot for the feature distribution
- Correlation between continuous features

3V's OF BIG DATA



VOLUME

```
# Volume: To print the number of rows and columns in the dataset
print('Volume of data:',df.shape)
```

```
Volume of data: (856387, 28)
```

- Comprises of 856,387 objects and 28 attributes

VARIETY

- Comprises of the historical traffic data, such as entry street, exit street, intersection details, direction driven at the intersection, the time required to travel the street, month and hour of the day, and weekend indicator.

VELOCITY

- We have collected the dataset from Kaggle competition. This is a static one-time data.

```
df.dtypes
```

```
RowId                      int64
IntersectionId             int64
Latitude                 float64
Longitude                float64
EntryStreetName           object
ExitStreetName            object
EntryHeading              object
ExitHeading               object
Hour                       int64
Weekend                    int64
Month                      int64
Path                      object
TotalTimeStopped_p20     float64
TotalTimeStopped_p40     float64
TotalTimeStopped_p50     float64
TotalTimeStopped_p60     float64
TotalTimeStopped_p80     float64
TimeFromFirstStop_p20    float64
TimeFromFirstStop_p40    float64
TimeFromFirstStop_p50    float64
TimeFromFirstStop_p60    float64
TimeFromFirstStop_p80    float64
DistanceToFirstStop_p20  float64
DistanceToFirstStop_p40  float64
DistanceToFirstStop_p50  float64
DistanceToFirstStop_p60  float64
DistanceToFirstStop_p80  float64
City                      object
```

DATA QUALITY REPORT - CATEGORICAL FIELDS

Data Quality Report - Categorical Features
Total records: 8

	Data Type	Count	%Miss	Cardinality	Mode	Mode Freq	Mode Perc	Second Mode	Second Mode Freq	Second Mode Perc
RowId	object	856387	0.000000	856387	1921357	1	0.000117	2492248	1	0.000117
IntersectionId	object	856387	0.000000	2559	84	3048	0.355914	112	2771	0.323569
EntryStreetName	object	848239	0.951439	1723	North Broad Street	14228	1.661398	South Broad Street	12045	1.406490
ExitStreetName	object	850100	0.734131	1703	North Broad Street	15339	1.791129	South Broad Street	12269	1.432647
EntryHeading	object	856387	0.000000	8	E	172398	20.130852	W	169738	19.820245
ExitHeading	object	856387	0.000000	8	W	171588	20.036269	E	169204	19.757890
Path	object	856387	0.000000	15075	Walnut Street_W_Walnut Street_W	5388	0.629155	North Broad Street_S_North Broad Street_S	5353	0.625068
City	object	856387	0.000000	4	Philadelphia	390237	45.567833	Boston	178617	20.857042

- % Miss: There are missing values for EntryStreetName and ExitStreetName. These observations have been dropped in the data pre-processing stage
- Cardinality: EntryStreetName and ExitStreetName has high cardinality. These have been encoded based on road encoding (Road, Street, Drive, etc.).
- Cardinality: RowId, IntersectionId, and Path feature has high cardinality.
- Mode: Most of the observations are for Philadelphia city followed by Boston.

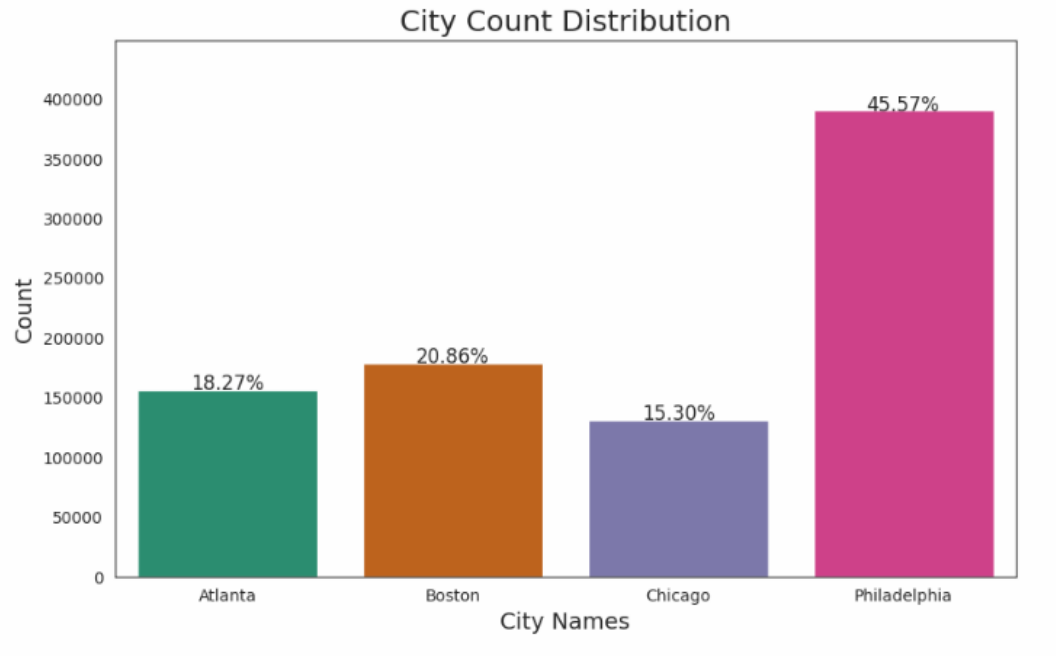
DATA QUALITY REPORT - CONTINUOUS FIELDS

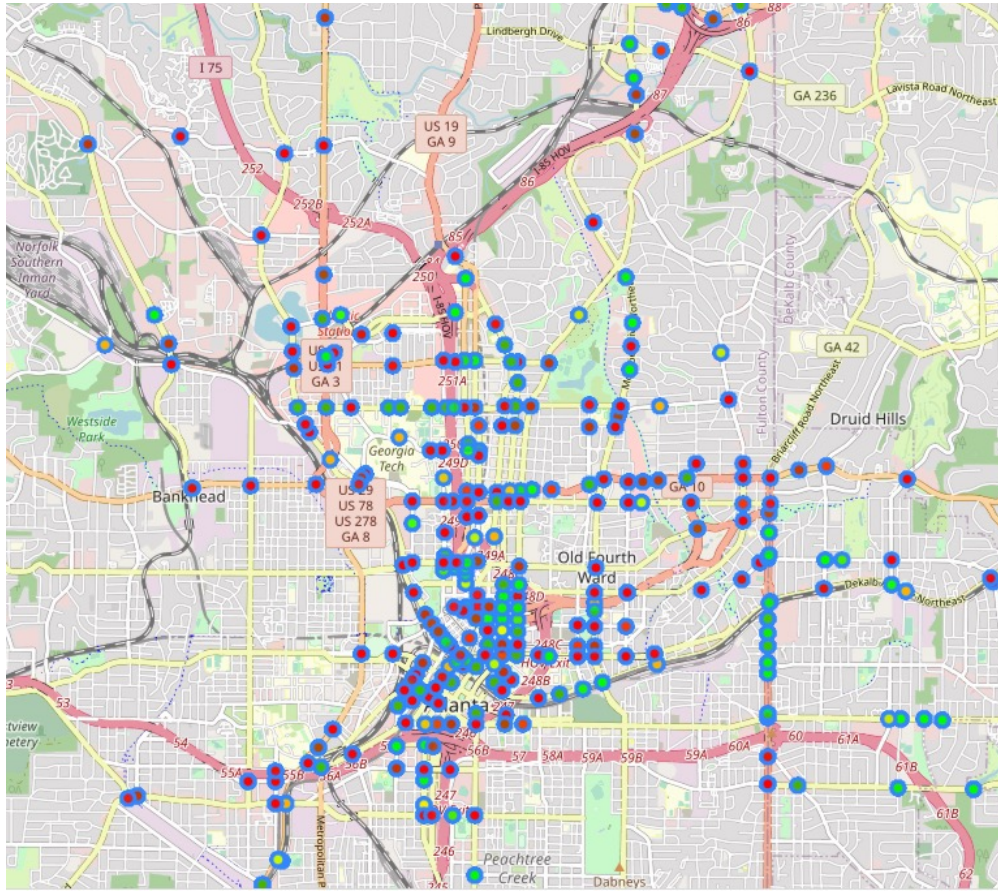
Data Quality Report for continuous varaiables
Total records: 20

	Data Type	Count	%Miss	Cardinality	Min	1st Qrt	Mean	Median	3rd Qrt	Max	Std_Dev
Latitude	float64	856387	0.0	4799	33.65	39.94	39.62	39.98	41.91	42.38	2.94
Longitude	float64	856387	0.0	4804	-87.86	-84.39	-77.92	-75.18	-75.10	-71.03	5.95
Hour	int64	856387	0.0	24	0.00	8.00	12.43	13.00	17.00	23.00	6.07
Weekend	int64	856387	0.0	2	0.00	0.00	0.28	0.00	1.00	1.00	0.45
Month	int64	856387	0.0	9	1.00	7.00	9.10	9.00	11.00	12.00	1.99
TotalTimeStopped_p20	float64	856387	0.0	171	0.00	0.00	1.76	0.00	0.00	298.00	7.15
TotalTimeStopped_p40	float64	856387	0.0	238	0.00	0.00	5.40	0.00	0.00	375.00	12.98
TotalTimeStopped_p50	float64	856387	0.0	262	0.00	0.00	7.72	0.00	10.00	375.00	15.69
TotalTimeStopped_p60	float64	856387	0.0	306	0.00	0.00	11.93	0.00	18.00	377.00	19.76
TotalTimeStopped_p80	float64	856387	0.0	403	0.00	0.00	22.95	16.00	34.00	763.00	28.27
TimeFromFirstStop_p20	float64	856387	0.0	244	0.00	0.00	3.18	0.00	0.00	337.00	11.84
TimeFromFirstStop_p40	float64	856387	0.0	316	0.00	0.00	9.16	0.00	0.00	356.00	20.45
TimeFromFirstStop_p50	float64	856387	0.0	336	0.00	0.00	12.72	0.00	22.00	356.00	24.22
TimeFromFirstStop_p60	float64	856387	0.0	353	0.00	0.00	18.93	0.00	31.00	357.00	29.85
TimeFromFirstStop_p80	float64	856387	0.0	355	0.00	0.00	34.20	27.00	49.00	359.00	41.13
DistanceToFirstStop_p20	float64	856387	0.0	3631	0.00	0.00	6.77	0.00	0.00	1901.90	29.54
DistanceToFirstStop_p40	float64	856387	0.0	6415	0.00	0.00	20.29	0.00	0.00	2844.40	59.20
DistanceToFirstStop_p50	float64	856387	0.0	7751	0.00	0.00	28.84	0.00	53.10	2851.10	75.22
DistanceToFirstStop_p60	float64	856387	0.0	9826	0.00	0.00	44.27	0.00	64.20	3282.40	102.03
DistanceToFirstStop_p80	float64	856387	0.0	13689	0.00	0.00	83.99	60.40	85.95	4079.20	160.71

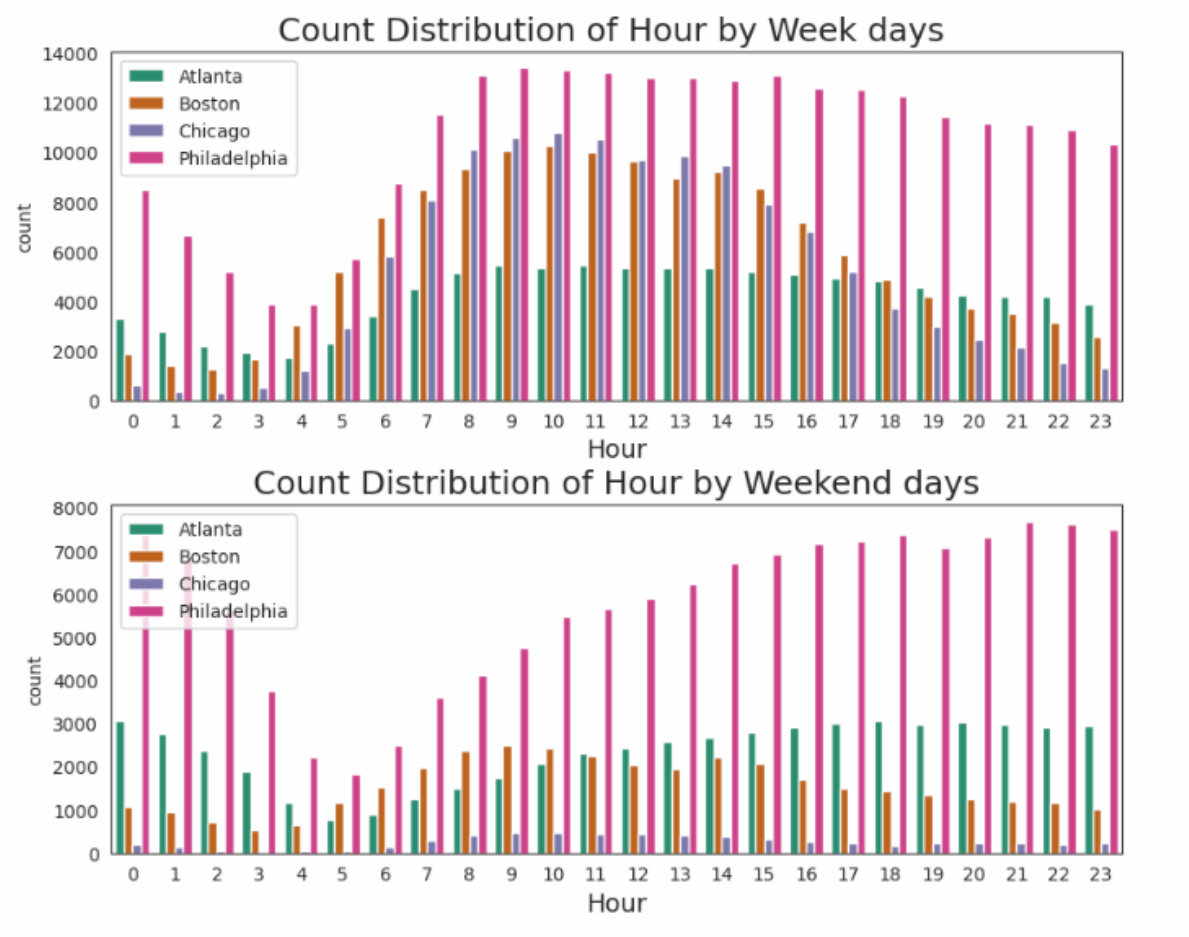
DATA EXPLORATION

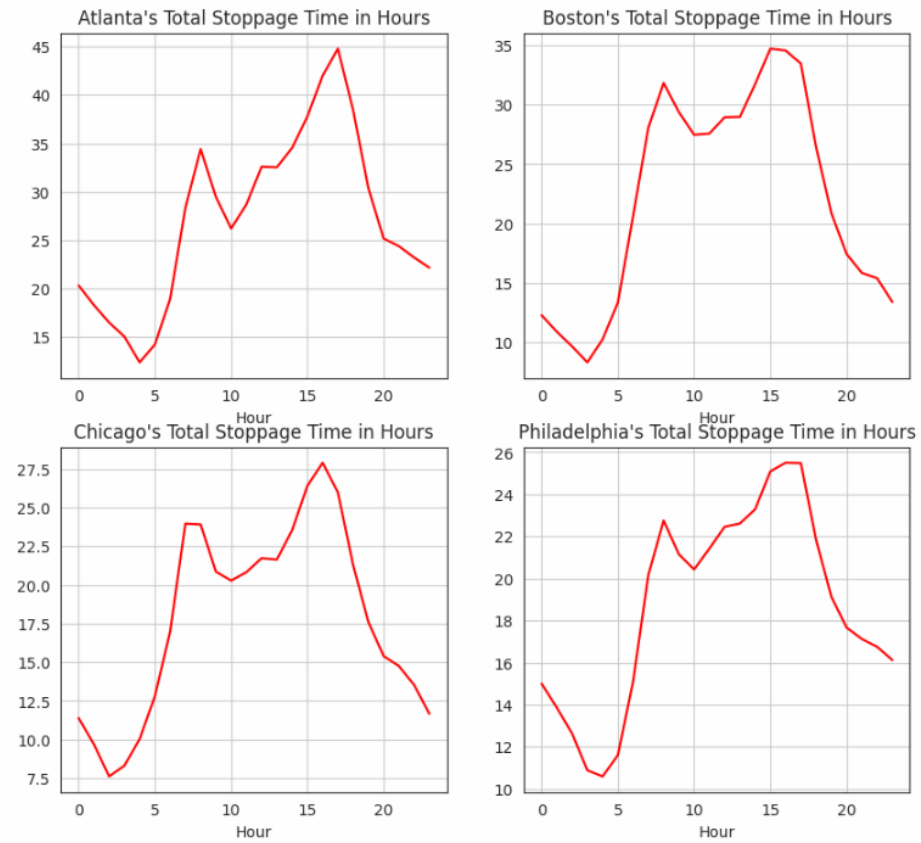
SJSU SAN JOSÉ STATE UNIVERSITY





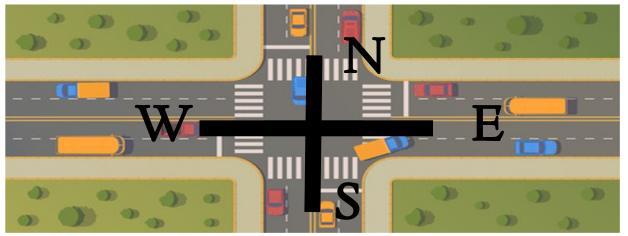
DATA EXPLORATION (CONTD.)





DATA PRE-PROCESSING

Data Cleaning → Feature Engineering → Data Scaling → Data Splitting

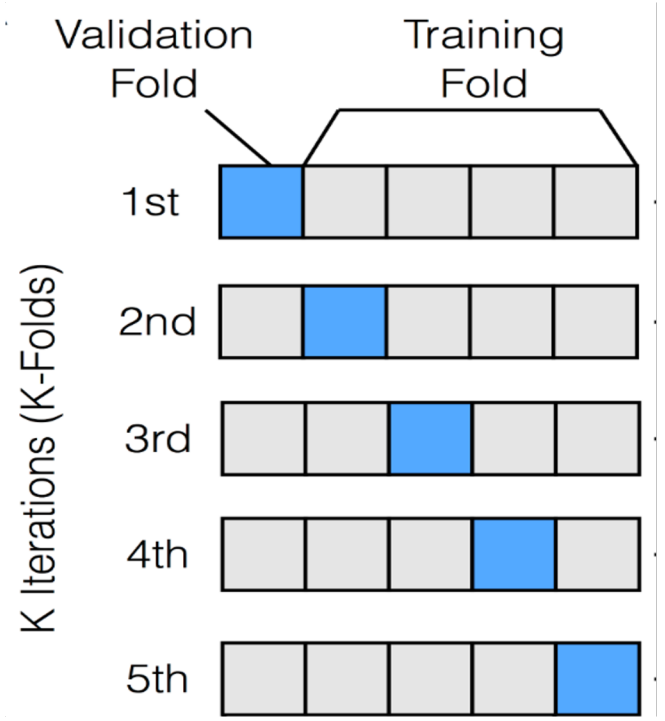


Transformation	Transformed Column
Mapping directions: 'N': 0, 'NE': 1/4, 'E': 1/2, 'SE': 3/4,'S': 1, 'SW': 5/4, 'W': 3/2, 'NW': 7/4	EntryHeading , ExitHeading
Encoding road types: 'Road': 1, 'Street': 2, 'Avenue': 2, 'Drive': 3, 'Broad': 3, 'Boulevard': 4	EntryTypeStreet ExitTypeStreet
Difference in heading: Subtracting ExitHeading from EntryHeading to give the difference in the heading of the car.	diffHeading
Label Encoding: IntersectionId and City columns are combined to form a unique identifier for each intersection	Intersection
Temperature: Monthly average temperature of each city is added as a feature by mapping the city-month variable to its corresponding average monthly temperature.	average_temp
Rainfall: Monthly average rainfall of each city is added as a feature by mapping the city-month variable to its corresponding average monthly rainfall.	average_rainfall

SJSU SAN JOSÉ STATE UNIVERSITY

DATA PREPARATION

- The final preprocessed data is then split for training and testing in the 80-20 ratio.
- **K- fold Cross Validation** :
  - It involves dividing the dataset into k-folds and training the model k times, with each fold used as a testing set once.
  - This technique helps us to evaluate the model's performance on multiple splits of the data and provides a better estimate of the model's performance.
  - The average performance across all the k-folds is used as the final estimate of the model's performance.
  - It helps to reduce overfitting, which occurs when a model is trained on a specific set of data and performs poorly on new, unseen data.
  - Furthermore, K-Fold cross-validation helps us to identify any data-specific issues, such as bias or variance, that might impact the model's performance.

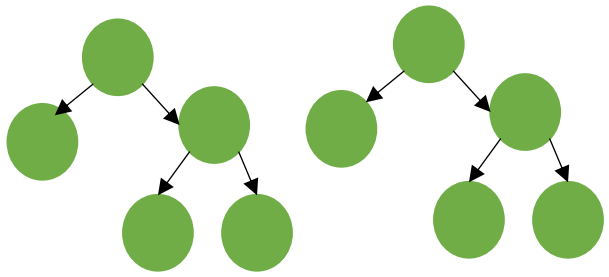


SJSU SAN JOSÉ STATE UNIVERSITY

MODELING

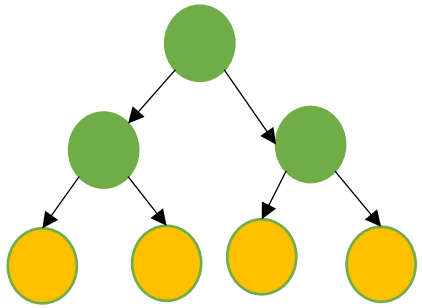
Ensemble Learning Algorithms

RANDOM FOREST



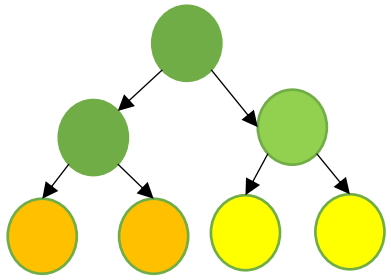
Level wise

CatBoost



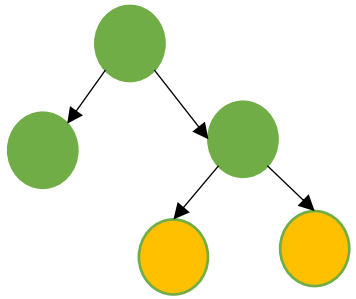
Symmetry

dmlc XGBoost



Level wise

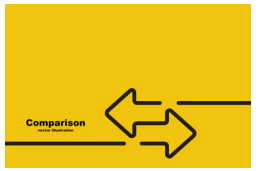
LightGBM



Leaf wise

EXPERIMENTAL RESULTS

SJSU SAN JOSÉ STATE UNIVERSITY



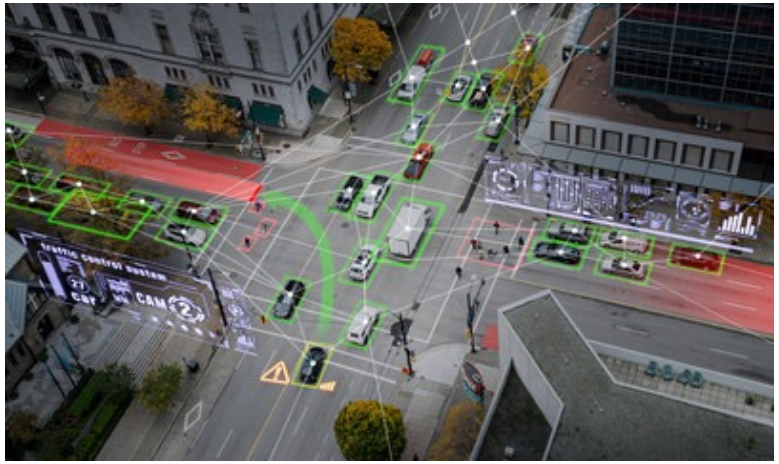
METRICS / MODELS	RANDOM FOREST	CATBOOST	XGBOOST	LIGHTGBM
RMSE	32.63	53.23	54.66	36.94
R2_Score	0.52	0.35	0.34	0.42
Time_taken (mins)	~18 min	~ 46 min	~68 min	~10 min

TEAM MEMBER AND CONTRIBUTION

SJSU SAN JOSÉ STATE UNIVERSITY

Area	Team Member
Brain storming and Topic selection	All
EDA	All
Feature Engineering	All
Random Forest	Divya
XGBoost	Divya
Cat Boost	Nupur
LightGBM	Vani

CONCLUSION AND FUTURE WORK



- Traffic congestion is a major problem in many cities, leading to increased travel times, air pollution, and fuel consumption.
- By accurately predicting traffic congestion, transportation agencies can better plan and manage their infrastructure, reducing congestion and improving the overall transportation experience for commuters.
- City planners could also use these models to plan and design new roads and transportation systems, taking into account predicted traffic patterns.
- One possible direction is to incorporate real-time data sources, such as traffic cameras and GPS data from vehicles, to improve the accuracy of the predictions.
- Another direction is to develop more complex models that take into account factors such as road construction, commercial places, holidays and special events.
- Finally, the models could be integrated with other transportation planning tools, such as route optimization and public transportation scheduling, to create a more comprehensive transportation management system.

Video Source: Adobe Stock