Moringa School

Phase 5 project

MODELING AND PREDICTING TRAFFIC CONGESTION LEVELS AND PEDESTRIAN CROSSINGS AT DIFFERENT TIMES OF THE DAY.

Group Members

Benson Kamau

Kevin Muchori

Sally Kinyanjui

Breden Mugambi

Modeling and predicting traffic congestion levels and pedestrian crossings at different times of the day

Overview

The urban mobility and transportation sector is vital for the functioning of modern cities, enabling the movement of people and goods efficiently. Within this industry, traffic management and pedestrian safety are crucial components that directly impact the quality of life in urban areas. Effective traffic pattern analysis and prediction can help mitigate congestion, enhance safety, and improve overall urban mobility. Well managed traffic leads to minimized economic losses, improved quality of life especially on the side of pedestrians.

Challenges

There are so many problems that are encountered especially in most urban towns whose vehicle and pedestrian population continues to grow everyday. One of the problems is the traffic congestion which leads to higher traffic volumes which in turn brings about economic losses due to wasted time and fuel, increased pollution. Another key challenge is the pedestrian safety where High pedestrian traffic in urban areas increases the risk of accidents. A challenge to also note is collecting accurate and real-time data from various sources is challenging which would make accurate traffic and pedestrian predictions challenging.

Proposed Solution

To solve some of these challenges would include measures such as advocating for sustainable urban mobility policies and invest in supportive infrastructure. Use of Use machine learning models to analyze and predict traffic patterns and pedestrian crossings at different times of the day. In order to gather real-time data on traffic and pedestrian movement would require use of high technology like IoT devices.

Conclusion

The analysis and prediction of traffic congestion levels and pedestrian crossings are essential for enhancing urban mobility and safety. Successful implementation of these solutions can lead to reduced congestion, fewer accidents, and an overall improvement in the quality of urban life

Problem Statement

Urban areas continue to face significant challenges in managing their traffic congestion and ensuring pedestrian safety. The changing nature of these areas together with the increasing volume of both vehicle and pedestrian traffic, makes it hard for one to predict traffic patterns affectively.

Objective

Our primary objective is to create an accurate time series model(s) and machine learning model(s) that can model, analyze and predict traffic congestion levels and pedestrian crossings at different times of the day

Specific Objectives

1. To identify key factors that influence traffic and pedestrian movement
2. To develop predictive models for forecasting future traffic congestion and pedestrian crossing patterns.
3. To provide recommendations for urban planners and traffic management authorities to improve traffic flow and pedestrian safety.

Data Understanding

The data to use in this study is sourced from the UC Irvine Machine Learning Repository. It has 4760 rows and 14 data features.

- 'oid': This column represents a unique identifier for each object record in the dataset.
- 'timestamp': This column stores the exact time of each record
- 'date': This column extracts the date portion from the 'timestamp' column, providing the day without the time information.
- 'hour': This column extracts the hour of the day (0-23) from the 'timestamp' column.
- 'x': This column represents the X-coordinate of each object in our data.
- 'y': This column represents the Y-coordinate of each object in our data.
- 'vehicle_count': This column indicates the number of vehicles observed in the vicinity of each object record.
- 'pedestrian_count': This column reflects the number of pedestrians observed in the vicinity of each object record.
- 'congestion_level': This column categorizes the traffic congestion level at the time of each record.
- Environmental Columns:
- 'weather_condition': This column represents the weather condition at the time of each record
- 'temperature': This column holds the temperature recorded at the time of each data point.
- 'location': This column specifies the specific region or intersection where the data was collected.
- 'body_roll': This column is intended to hold data related to the body roll angle (in degrees) of an object.
- 'body_pitch': Similar to body_roll hold data related to the body pitch angle (in degrees) of an object.
- 'body_yaw': Likewise, could be used for data on the body yaw angle (in degrees) of an object.
- 'head_roll': could be used for data on the head roll angle (in degrees) of an object.
- 'head_pitch': Similar to head_roll, could be used for data on the head pitch angle (in degrees) of an object.
- 'head_yaw': Similar to head_roll and head_pitch, this column holds NaN values and could be used for data on the head yaw angle (in degrees) of an object.

Data Cleaning

Data cleaning is a critical step in preparing the dataset for analysis. The goal is to remove any irrelevant or problematic data that could skew the results or lead to inaccurate conclusions. The following steps outline the cleaning process for this dataset:

Step 1: Dropping the Empty Columns

The dataset contains several columns that do not have any data: body_roll, body_pitch, body_yaw, head_roll, head_pitch, and head_yaw. Since these columns are empty and will not contribute to the analysis, they need to be removed.

Step 2: Convert the 'Timestamp' Column to Datetime Format

The timestamp column is essential for time series analysis. However, to perform accurate time-based operations, it must be in the correct datetime format. This step ensures that

Step 3: Check for and Remove Duplicate Rows

Duplicate rows can occur due to data collection errors or redundancy in the dataset. It is important to identify and remove these duplicates to avoid bias in the analysis.

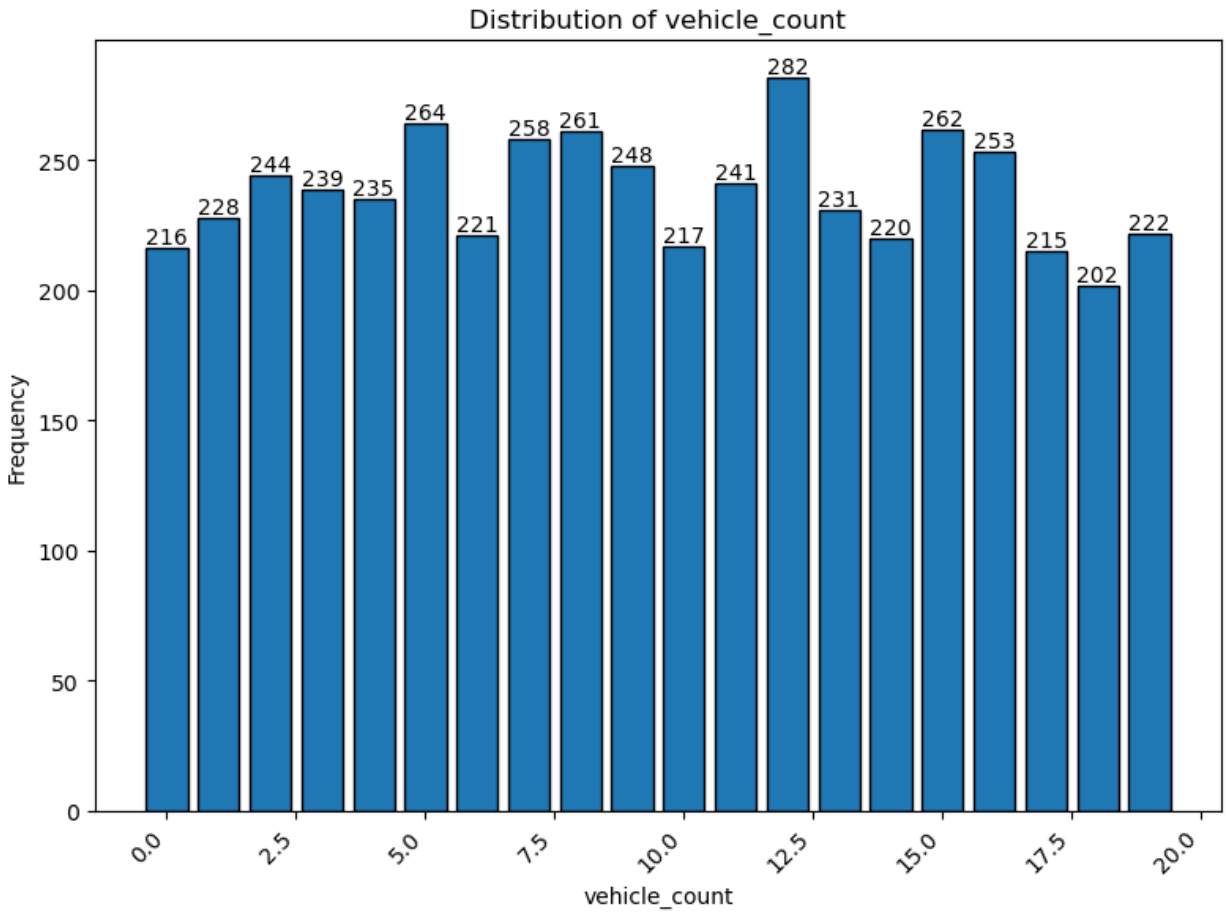
Summary

Through the data understanding process, we identified key components of the dataset, particularly the importance of the timestamp column and the presence of empty positioning columns. The data cleaning steps ensured that the dataset was free of irrelevant columns, correctly formatted for time series analysis, and devoid of duplicate entries. This preparation sets the stage for more accurate and insightful analysis of the traffic data.

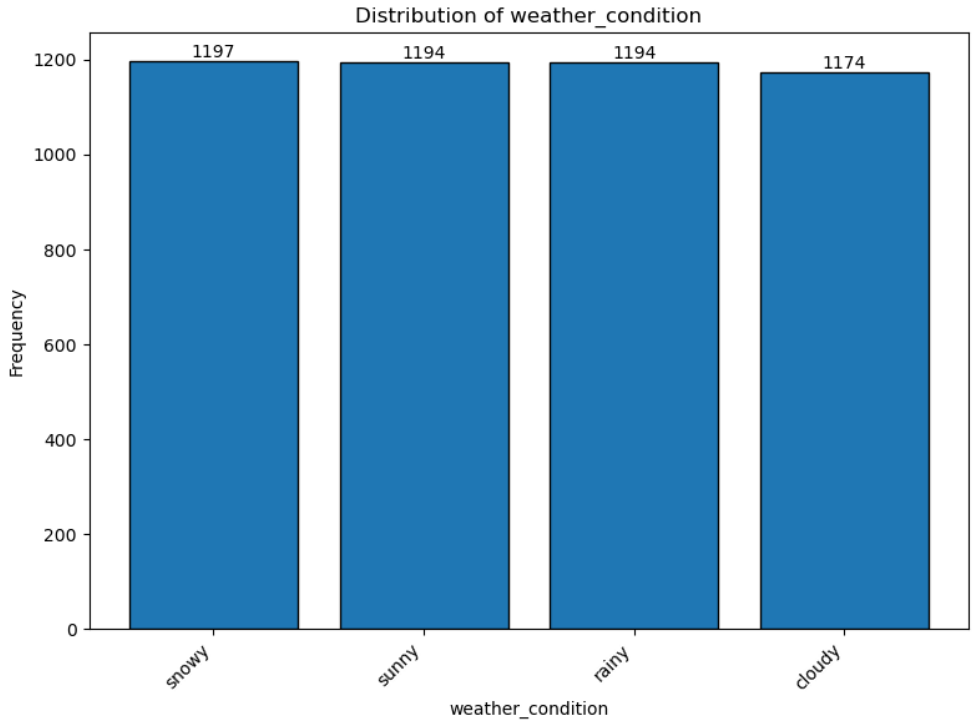
DATA ANALYSIS

We will be using different analytical and visualization techniques to visually describe relationships, either between variables or description of those variables

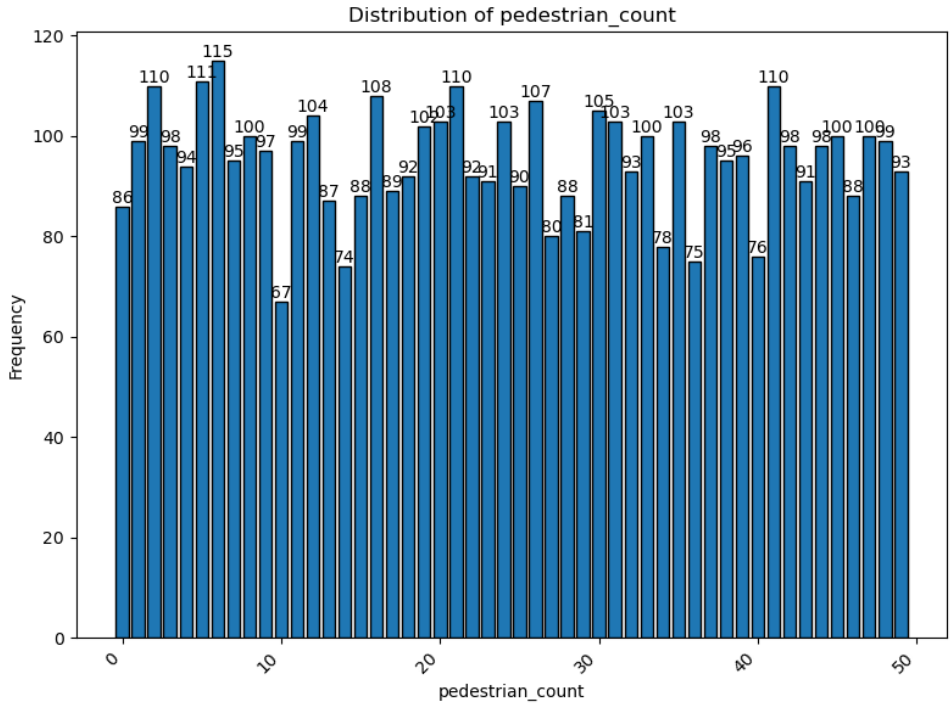
Univariate analysis



In this visualization, we can see the frequency distribution of vehicles accounted for in the dataset. Here, we can see that all vehicles have at least a minimum of at least 200 vehicles, with the hisgest being 282 vehicles

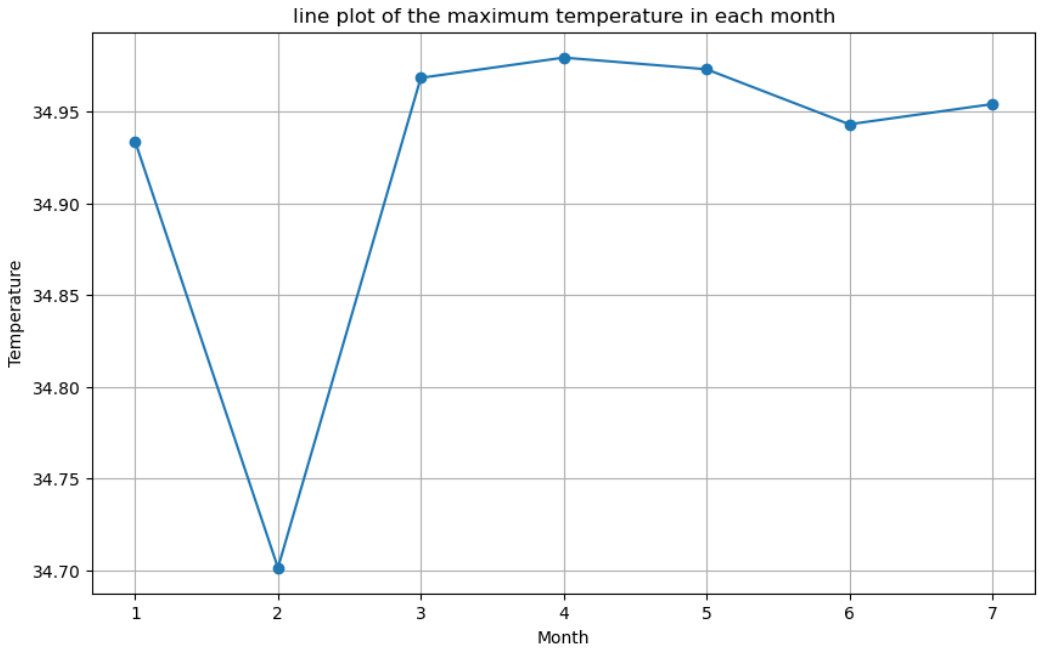


In this visualization, we can see that the weather was snowy at most of the time, but all weather conditions seem to have been present throughout the dataset

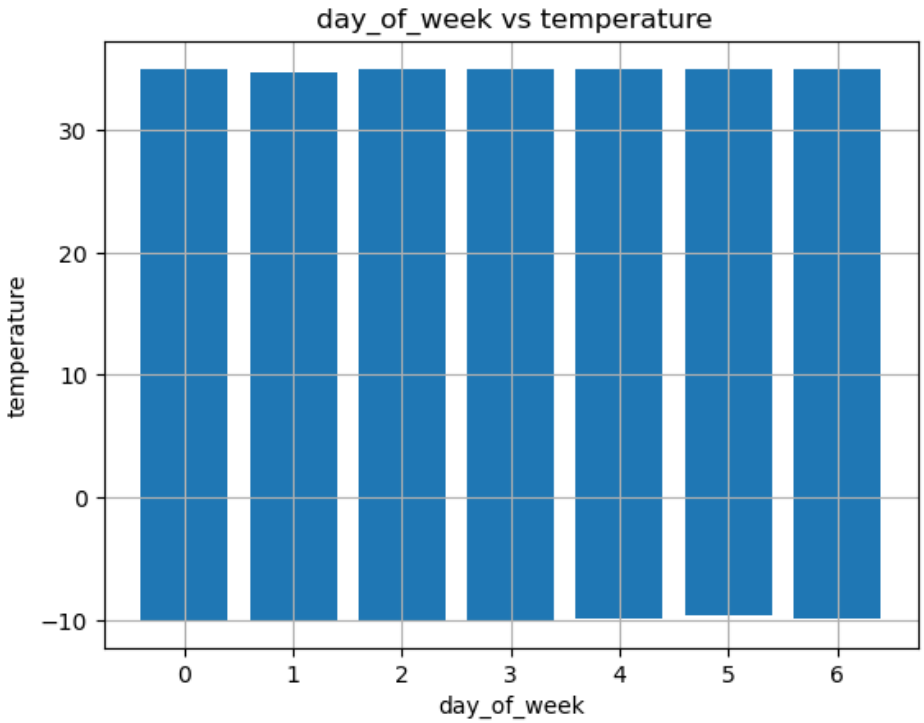


Here, we can see that there are high pedestrian counts in the column, with the highest pedestrian count being 115

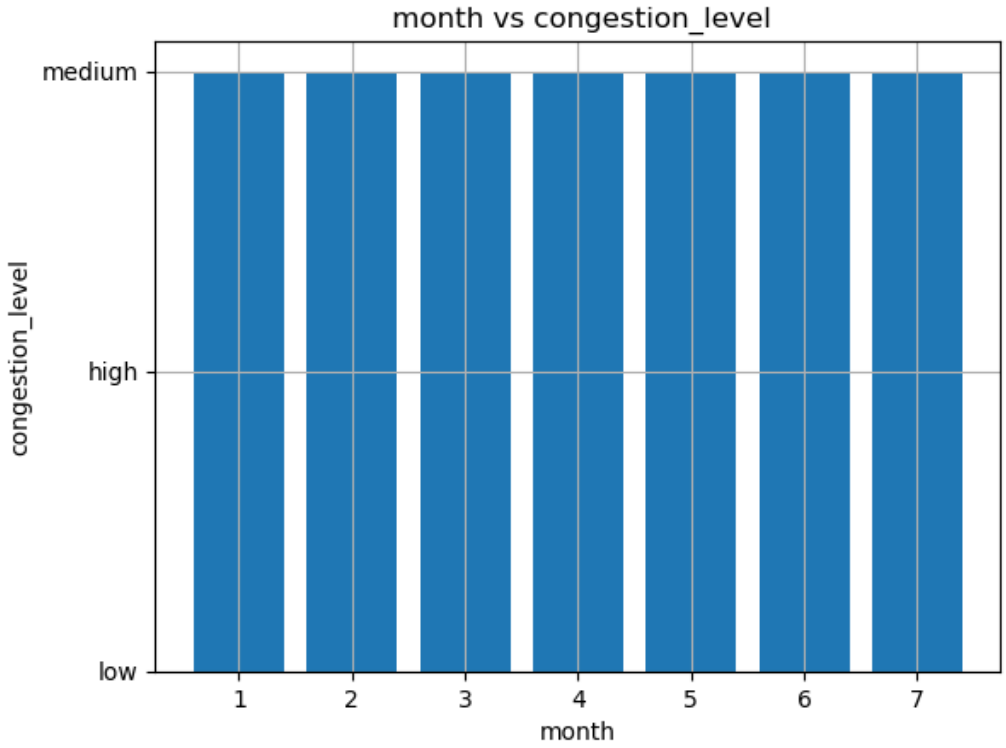
Bivariate Analysis



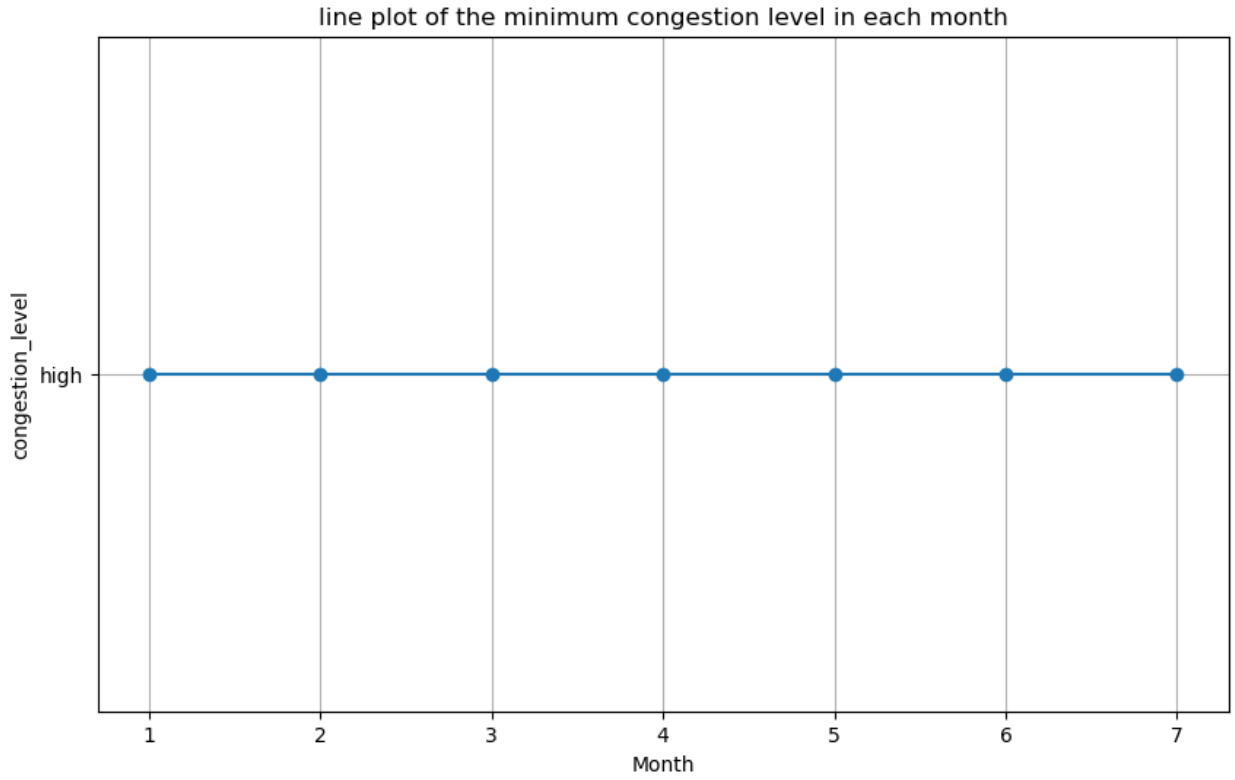
In this, we plotted the month against the temperature, specifically the maximum temperature. The minimum temperature has also been plotted in the notebook. The months range from month 1 to month 7, and we can see that in February, we had the lowest maximum temperature, a substantial dip from January



We did the same thing but with days of the week instead of months. This time, it encompasses the entire temperature range



Here, we aimed to see the congestion levels with each month, and as shown in the next visualization , the minimum congestion level was a medium


line plot of the minimum congestion level in each month
congestion_level
high
1
2
3
4
5
6
7
Month

MODELING

We will be using different models that is both time series models and classification models to model and predict congestion levels and pedestrian crossing at different times of the day

DATA PREPROCESSING

The vehicle count time series has an upward and downward trends at different intervals

From the seasonal plot the data has spikes and it is therefore seasonal

The people count time series has an upward and downward trends at different intervals

From the seasonal plot the data has less spikes compared to our vehicle count time series.It is seasonal

Before the actual modeling I'll first preprocess the data, focusing on the time series features. This includes:

- Handling Categorical Variables: Converting categorical features like congestion_level, weather_condition, and location into numeric values.
- Resampling: Aggregating the data based on a suitable time interval if needed (e.g., hourly or daily).
- Stationarity Check: Checking if the time series is stationary, as ARIMA requires stationarity. If it's not stationary, we'll apply transformations like differencing or moving average.

Our data was stationary.

Baseline model :Naive Forecast

Baseline model is a simple model that anyone can come up with. In time series, this is a model that proves that the current value is a true reflection of previous value.I go ahead to create a simple naive forecast model.
Baseline MAE (Naive Forecast): 6.514
The forecast reflects true values of the vehicle count

1.VEHICLE COUNT

We used both ARIMA and SARIMA models.ARIMA model had an AIC value of 30004.157

Observations

From the forecasted vehicle count we noted the following

From 8:00am the vehicle count is relatively low and gradually increases at 9:00am and continues to rise up to 1200 hrs where it reaches an optimum count at around 1300hrs and remains almost the same all the way

From 8:00am the vehicle count is relatively decreasing all the way to few minutes to 10:00am where the count shoots up and at around 12:00pm the count start to reduce significantly and then rises from 2:00pm to 3:00pm the vehicle count drops drastically and at around 1500hrs the count then increases drastically

2.PEDESTRIAN COUNT

We used both ARIMA and SARIMA models.ARIMA model had an AIC value of 38975.561

Observations

The forecasted pedestrian count forms a curve where the number of pedestrian count at 8pm is the highest and then reduces over time forming rectangular hyperbola.

From the forecasted pedestrian count with SARIMA model we noted the following

From 8:00am the pedestrian count is relatively decreasing all the way to few minutes to 10:00am where the count shoots up and at around 11:00pm the count start to reduce significantly and then rises from 12:00pm to 3:00pm the pedestrian count drops drastically and at around 1500hrs the count then increases drastically

3.CONGESTION LEVELS

We used classification models since our congestion levels were categorical variables

Low -1

Medium - 2

High -0

a)Model 1.Decision Tree

High congestion - 0

Had the highest recall score of 0.55, this shows that the model was able to identify 55% cases of high congestion

low congestion - 1

Recall score of 0.34, this shows that the model was able to identify 34% cases of low congestion

medium congestion - 2

Recall score of 0.08, this shows the model was able to identify 8% cases of medium congestion

The overall model accuracy was 32.52%

b) Model 2 .Random forest

High congestion - 0

Had a recall score of 0.33, this shows that the model was able to identify 33% cases of high congestion

low congestion - 1

Recall score of 0.37 ,this shows that the model was able to identify 37% cases of low congestion

medium congestion - 2

Recall score of 0.31, this shows the model was able to identify 31% cases of medium congestion

The overall model accuracy was 33.53%

c)Model 3 .XGBoost

High congestion - 0

Had a recall score of 0.32, this shows that the model was able to identify 32% cases of high congestion

low congestion - 1

Recall score of 0.32 ,this shows that the model was able to identify 32% cases of low congestion

Medium congestion - 2

Recall score of 0.35, this shows the model was able to identify 35% cases of medium congestion.
The overall model accuracy was 32.86%

Noted the accuracy is low in the models , we went ahead to try and solve the problem and did hyperparameter tuning

d) Model 4 - Decision Trees with pipelines and Grid Search

High congestion - 0

Had a recall score of 0.48, this shows that the model was able to identify 48% cases of high congestion

low congestion - 1

Recall score of 0.34, this shows that the model was able to identify 34% cases of low congestion

Medium congestion - 2

Recall score of 0.12, this shows the model was able to identify 12% cases of medium congestion

The overall model accuracy was 31.34%

e) Model 5 - Random Forest with pipelines and Grid Search

High congestion - 0

Had a recall score of 0.28, this shows that the model was able to identify 28% cases of high congestion

low congestion - 1

Recall score of 0.30, this shows that the model was able to identify 30% cases of low congestion

Medium congestion - 2

Recall score of 0.35, this shows the model was able to identify 35% cases of medium congestion

The overall model accuracy was 30.92%

EVALUATION

We will use both accuracy and recall to evaluate our model performance

Recall is a metric that measures the model's ability to identify true positive cases out of all actual positive cases and Accuracy is a general measure of model performance.

From the accuracy summary and recall summary we noted the following:

Best model - Random forest : Accuracy - 34%

:Recall - 33.5%

second model - XGBoost : Accuracy - 32.85%

:Recall - 32.9%

third model - Decision Trees : Accuracy - 32.5%

:Recall - 32.6%

FEATURE IMPORTANCE

1) Random forest

For Random forest model temperature, location and pedestrian count are the features that heavily affected congestion levels while weather condition as a feature came last

2) XGBoost

For XGBoost model pedestrian count, temperature and location in that order are the features that heavily affected congestion levels while weather condition as a feature came last

CONCLUSIONS AND RECOMMENDATIONS

a) CONCLUSIONS

The project yielded the desired results. The main objective and specific objectives were all satisfied.

The best models that modeled our data were ARIMA both for the pedestrian count and vehicle count and Random Forest for the congestion levels

Temperature, location and pedestrian count are the features that heavily affected congestion levels while weather condition as a feature came last

We also noted from 8:00am the vehicle count is relatively decreasing all the way to few minutes to 10:00am where the count shoots up and at around 12:00pm the count start to reduce significantly and then rises from 2:00pm to 3:00pm the vehicle count drops drastically and ataround 1500hrs the count then increases drastically

Additionally, from 8:00am the pedestrian count is relatively decreasing all the way to few minutes to 10:00am where the count shoots up and at around 11:00pm the count start to reduce significantly and then rises from 12:00pm to 3:00pm the pedestrian count drops drastically and at around 1500hrs the count then increases drastically

b) RECOMMENDATIONS

We would recommend the following to those who formulate policies in the urban mobility and transportation sector:

1. Encourage businesses and schools to adopt staggered start times, particularly around the critical periods of 8:00 AM and 2:00 PM. This can help distribute traffic more evenly throughout the day, reducing peak congestion

2. During identified peak pedestrian and vehicle congestion times, increase and encourage the frequency use of trains. This can help reduce the number of vehicles on the road, reducing congestion.

3. Given that pedestrian counts drop significantly after 3:00 PM, encourage walking and cycling during these times by improving infrastructure, such as creating safer, well-lit walkways and cycling paths. This could reduce vehicle congestion and promote healthier lifestyles.

4. Run public awareness campaigns to educate commuters about the best times to travel and the benefits of using alternative transportation modes during peak congestion times. This could include promoting the use of public transport or cycling.

5. Consider implementing congestion charges during peak hours to discourage unnecessary trips and reduce the number of vehicles on the road

NEXT STEPS

1. Analyze how extreme weather (e.g., heavy rain, snow) impacts traffic patterns. This could inform future infrastructure investments, such as better drainage systems or snow removal plans.

2. Since traffic patterns can change over time due to various factors, we should Continue to collect and analyze traffic data from our country Kenya and try and refine our models.