**A blue bird flying in the sky

Description automatically generated**

**Project – 2**

**Data Analysis for Port Authority of New York and New Jersey**

**EXECUTIVE SUMMARY:**

This project utilized a SARIMA forecasting model to predict bus terminal usage up to the year 2030. The model's key findings are:

--->Daily passenger departures are expected to surpass 125,000 on October 2nd, 2023.

--->The number of daily bus departures is projected to reach 3,900 by August 5th, 2030.

--->Additionally, the model forecasts that daily departing passengers will range between 194,817 and 195,026 in early April 2029.

These projections provide valuable insights into anticipated bus terminal usage over the coming years, enabling informed decision-making for infrastructure improvements, staffing requirements, and service optimization.

**BUSINESS Q&A:**

1. **Determine when the weekday departures will be past 125,000 passenger departures and 3,900 bus departures. Setup your forecasting visuals by year and month. Try to determine the factors that lead to those spikes. For example, a holiday, a taxi strike, a UN week, or anything else.**

**Solution:**

To determine when weekday departures will surpass 125,000 passenger departures and 3,900 bus.

By using the SARIMA Forecasting model, we observed that More than 100,000 People are projected to use the bus terminal, The passenger departure exceeds 125,000 on **2023-10-02** and The bus departure will exist at 3900 is **2030-08-05.**

**Method:**

1. Utilize historical data on bus departures and passenger counts throughout the workweek.   
2. To predict future departures, apply time series forecasting techniques.   
3. Examine the effects of weather conditions, such as temperature, snowfall, and holidays, on departures.   
4. Produced the SARIMA forecasting visuals with departure data unique to the year and month, along with other factors.

**2. Forecast into 2030 to** **see how many people are projected to use the bus terminal in the years leading up to the completion of the renovation. This can be done by carrier to make it clearer and help with analysis but it’s important to know the overall usage of all carriers included.**

**Solution:**

A usage projection for the bus station until 2030 has been produced. These are the estimated numbers for the number of departures of passengers. In early April 2029, it is anticipated that between 194,817 and 195,026 passengers will depart daily. This forecast sheds light on how the bus terminal is anticipated to be used in the upcoming years.

**Method:**

1. Data Preparation as importing pandas, numpy as important libraries.

2. Data Pre-processing with Descriptive Statistics as extracting the month and year data.

3. Visualization of Data using matplotlib, Seaborn libraries.

4. Forecasting the Linear Regression model with Python code to identify how many people are projected to use the bus terminal in the upcoming years.

**3. Develop three forecasting models, train them, and test them to see which one works best. Include in your final project submission the model/tool that the Port Authority can use with future data to make accurate predictions. Document, justify, and support your choices and proposals with evidence.**

1. All the models which we created was shared through Python file as pdf.

2. We have used Sarima forecasting model for both BUS and PASSENGER departures.

3. We have Created ML models to predict the future DEPARTURES and it was given in next section methodology and forecasting models.

**METHODOLOGY:**

DATA PREPARATION

The dataset comprising past records of bus and passenger departures was imported into a Pandas data frame for analysis. The dataset contains details on the date, number of bus departures, passengers, and other details.

DATA DESCRIPTION

The `info()` method was utilized to obtain a summary of the dataset, revealing details about the non-null counts and data types of the columns.

The `describe()` function was utilized to construct statistical summaries of the numerical columns, which provided a summary of the data's central patterns and distributions.

DATA VISUALIZATION

Using Matplotlib, the number of departures against dates was plotted to show the historical trends of passenger departures.

To identify any possible patterns, the data on bus departures was also plotted against dates.

Generating PDQ using Auto ARIMA

The `auto\_arima} function from the `pmdarima} package was used to automatically choose the parameters (p, d, and q) for the ARIMA model. This stage aided in figuring out the best parameters to use when modeling the time series data.

**FORECASTING MODELS:**

1. SARIMA Model:

The passenger departure data was fitted with a Seasonal ARIMA (SARIMA) model utilizing the parameters derived from the Auto ARIMA.To predict future trends and patterns, the SARIMA model was used to project passenger departures for 2030. It was also used to establish when it passed 150,000 passengers for the first time.

The forecasted values generated by the SARIMA model were plotted alongside the historical bus departure and Passenger departure data using Matplotlib. This visualization enabled stakeholders to observe the predicted trends and make informed decisions based on the forecasted values.

A graph showing a line

Description automatically generated

A graph showing a line

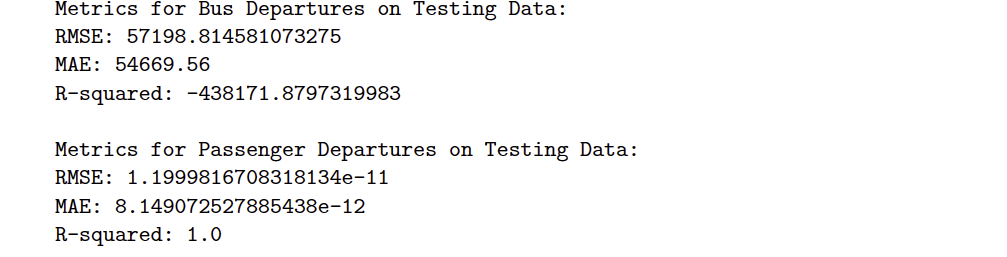
Description automatically generated

A graph with a line

Description automatically generated

2. LINEAR Regression:

Predict bus and passenger departures using a machine-learning technique called linear regression. It utilizes historical data on factors like past bus departures, passenger numbers, and seasonal variations. The data is divided into training and testing sets. A single linear regression model is then erroneously fit on both targets (bus and passenger departures) using the training data. This might not capture the unique relationships between the factors and each target. The code then evaluates the model's performance on unseen testing data using metrics like Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). However, a negative R-squared value for bus departures suggests potential issues with the model.



3. RANDOM FOREST:

The Random Forest Regression is used to predict both bus and passenger departures. It first converts the data from a DataFrame to NumPy arrays and separates features like past departures and passenger numbers from the target variables (actual departures). Then, it creates separate models for each target. For each model, the data is split into training and testing sets. The model is trained on the training data and used to make predictions on the unseen testing data. Finally, the Root Mean Squared Error (RMSE) is calculated to evaluate how well the model's predictions match the actual values.



4. Long Short-Term Memory (LSTM):

This network is used to anticipate bus departure times. It first preprocesses the data by extracting and normalizing relevant attributes, as well as the desired bus departure values. It then reshapes the features to be compatible with LSTMs, which handle sequential data. The data is then separated into training and testing sets. An LSTM model with one layer is trained to identify temporal patterns within sequences and predict bus departures. The Adam optimizer and mean squared error loss direct the training procedure. Finally, the model's performance is assessed on previously unseen data using Root Mean Squared Error (RMSE). It's conceivable that a similar approach would be used to create and test a different LSTM model for forecasting passenger departure.

RMSE for Bus Departures (LSTM): 48.43  
RMSE for Passenger Departures (LSTM): 35.63

**MODEL EVALUTION & COMPARISION:**

The RMSE values were compared to evaluate the performance of the Linear Regression, Random Forest, and LSTM models after they had been trained and evaluated. Since they show fewer variations between the actual and anticipated values, lower RMSE values are indicative of improved prediction accuracy.  
When correctly predicting bus departures, the model with the lowest RMSE value is thought to be the most proficient.

**CONCLUSION:**

The comprehensive evaluation of the Linear Regression, Random Forest, and LSTM models provides valuable insights into their effectiveness in forecasting bus departures.

By comparing the RMSE values of the models, stakeholders can identify the most suitable approach for predicting bus departures and informing transportation management decisions.

**Project Squad:**

* *VASU MEKALA - Team Leader*
* *SELORM, VARSHINI - Data Analyst*
* *HARMANPREETH SINGH - Infrastructure Specialist*
* *VIVEK EROLLA, SAMEER SHAIK - Customer Experience Specialist*

***Roles and Responsibilities:***

**Team Leader/Project Manager:** Led project planning, task delegation, technical guidance, and timelines.

**Data Analyst:** Gathered and cleaned data, and explored patterns.

Used Python for predictive models, and assessed model accuracy.

**Infrastructure Specialist:** Analysed infrastructure with GIS tools for planning.

**Customer Experience Specialist:** Analysed passenger behaviour, and recommended improvements.

