Predicting the impact of weather on transportation ridership in New York City

Problem Description:

The transportation system in New York City plays a vital role in the daily life of millions of people. Public transportation provides an essential means of travel for residents, workers, and visitors alike. The city's transportation system consists of several modes, including buses, subways, and trains, operated by various agencies.

Weather conditions can have a significant impact on transportation ridership in New York City. Extreme weather events like snowstorms, hurricanes, and heatwaves can disrupt the normal operation of the transportation system, causing delays, cancellations, and closures. Understanding the impact of weather on transportation ridership can help transportation agencies prepare for these events and take appropriate measures to minimize disruptions.

To this end, the City of New York has provided a dataset on Kaggle that contains information on transportation ridership and weather conditions in the city. The dataset covers the period from January 2014 to December 2018 and includes data on subway, bus, and train ridership, as well as weather measurements such as temperature, precipitation, and wind speed.

Dataset Link:

The dataset can be found on Kaggle: https://www.kaggle.com/cityofny/transportation-weather-impact

Objectives:

The primary objective of this project is to build a predictive model that can accurately forecast transportation ridership in New York City based on weather conditions. The model should be

able to predict the number of riders for each mode of transportation (subway, bus, and train) and each day of the year, given the weather conditions on that day.

The project's secondary objectives include:

1. Exploratory data analysis of the transportation ridership and weather data to identify trends, patterns, and relationships.

2. Feature engineering to create new features that capture the impact of weather on transportation ridership.

3. Building and evaluating multiple machine learning models to identify the best-performing model.

4. Developing a web application that allows users to input weather conditions and obtain a prediction of transportation ridership for the selected mode of transportation and date.

Deliverables:

The deliverables for this project include:

1. A Jupyter notebook that documents the data exploration, feature engineering, and machine learning modeling processes.

2. A web application that allows users to input weather conditions and obtain a prediction of transportation ridership for the selected mode of transportation and date.

3. A report summarizing the project's findings, including insights gained from the data exploration, the best-performing machine learning model, and its predictive performance.

Suggested Framework to Solve this problem :

Framework or Approach:

To achieve the objectives outlined in the problem description, the following framework or approach can be used:

1. **Data Collection:** Download and import the transportation ridership and weather dataset provided by the City of New York from Kaggle.

2. **Data Cleaning:** Check the dataset for missing, invalid or inconsistent data, and perform necessary data cleaning techniques such as imputation or deletion of data.

3. **Exploratory Data Analysis (EDA):** Conduct EDA to gain insights into the data, such as identifying trends, patterns, and relationships between the features. EDA can include summary statistics, visualizations, and hypothesis testing.

4. **Feature Engineering:** Create new features that capture the impact of weather on transportation ridership. Examples include combining weather variables to create indices or adding weather information from the previous days to capture trends.

5. **Data Preparation:** Split the data into training and testing sets. Preprocess the data by scaling, encoding, and transforming the features as necessary.

6. **Machine Learning Modeling:** Develop multiple machine learning models such as linear regression, random forest, and gradient boosting to predict transportation ridership based on weather conditions.

7. **Model Evaluation:** Evaluate the performance of each model using appropriate metrics such as mean absolute error, mean squared error, or R-squared. Select the best-performing model.

8. **Web Application Development:** Develop a web application that allows users to input weather conditions and obtain a prediction of transportation ridership for the selected mode of transportation and date.

9. **Reporting and Documentation:** Summarize the project's findings, including insights gained from the data exploration, the best-performing machine learning model, and its predictive performance. Create a Jupyter notebook that documents the data exploration, feature engineering, and machine learning modeling processes.

Code Explanation :

Here is the simple explanation for the code you can find at code.py file.

Section 1: Importing Libraries and Loading Data

This section of the code starts by importing the required Python libraries – pandas, numpy, matplotlib, seaborn, and sklearn. It then loads the dataset into a Pandas DataFrame using the read_csv function.

Section 2: Data Preprocessing

In this section, we preprocess the data to prepare it for machine learning. This includes dropping unnecessary columns, checking for missing values, converting datatypes, and encoding categorical variables using one-hot encoding. We also split the data into training and testing sets using a 70:30 ratio.

Section 3: Building Random Forest Models

This section builds Random Forest models to predict subway and bus ridership based on weather data. We use the RandomForestRegressor class from the sklearn.ensemble module to create the models. We also tune the hyperparameters of the models using grid search cross-validation.

Section 4: Model Training and Hyperparameter Tuning

In this section, we train the Random Forest models using the training data and the hyperparameters selected in the previous section. We also calculate the feature importances of the models to see which weather variables are most important for predicting ridership.

Section 5: Model Evaluation

This section evaluates the performance of the Random Forest models for subway and bus ridership prediction using three different metrics – mean absolute error (MAE), mean squared error (MSE), and R-squared. It calculates the predictions for the test set using the predict method of the trained models, and then compares these predictions with the actual test labels to calculate the metrics. Finally, it prints out the performance metrics for both the subway and bus models.

Section 6: Conclusion and Future Work

This section provides a brief conclusion to the project and suggests potential future work that could be done to improve the models or explore the data further.

We used Random Forest regression algorithm to model the relationship between weather variables and subway/bus ridership in NYC. This algorithm is suitable for this task because it can handle non-linear relationships between the variables and can handle large datasets with many features.

To run the code, you will need to have Python 3 installed, along with the libraries mentioned in Section 1. You can run the code by executing the Python script from the command line or running the code cell by cell in a Jupyter Notebook. You will also need to download the dataset from the Kaggle link provided in the problem description and place it in the same directory as the Python script.

Future Work :

Section 1: Introduction

In this section, we will provide a brief overview of the future work for this project and explain the steps involved in implementing it.

Section 2: Feature Engineering

One potential area for future work is to do more feature engineering to extract more useful information from the data. For example, we could create new features based on time-related variables such as day of the week, hour of the day, and month of the year. We could also incorporate information about major events or holidays that might affect ridership. To implement this, we would need to modify the preprocessing code in Section 2 to create the new features.

Section 3: Model Selection and Evaluation

Another potential area for future work is to experiment with different machine learning algorithms and model configurations to see if we can improve the prediction performance. For example, we could try using other regression algorithms such as Gradient Boosting, Support Vector Regression, or Neural Networks. We could also try different hyperparameter tuning methods or different evaluation metrics. To implement this, we would need to modify the code in Sections 3 and 5 to incorporate the new models and evaluation metrics.

Section 4: Spatial Analysis

A third potential area for future work is to explore the spatial patterns of ridership and weather in NYC. We could use geographic information system (GIS) tools to create spatial visualizations of ridership and weather data, and explore whether there are any spatial relationships between weather and ridership. For example, we could look at whether ridership is higher in certain neighborhoods or regions of the city during certain weather conditions. To implement this, we

would need to acquire and preprocess spatial data for NYC, and modify the code in Section 2 to merge the spatial data with the existing dataset.

Section 5: Time Series Analysis

A fourth potential area for future work is to incorporate time series analysis into the models to account for trends and seasonality in the data. We could use time series forecasting models such as ARIMA or SARIMA to model the time series patterns in the data and incorporate them into the machine learning models. To implement this, we would need to modify the code in Sections 2 and 3 to include time series features and use time series models to make predictions.

Section 6: Conclusion

In this section, we have outlined several potential areas for future work for this project, including feature engineering, model selection and evaluation, spatial analysis, and time series analysis. By exploring these areas, we can gain a deeper understanding of the relationship between weather and transportation ridership in NYC and improve the accuracy of our predictive models.

To implement any of these future work ideas, we would need to modify the existing code in the relevant sections and run the modified code. We would also need to acquire any additional data or software tools required for the analysis. We should also evaluate the results of any changes we make to the models to ensure that we are improving their predictive accuracy.

Exercise Questions :

1.What feature(s) have the most significant impact on the ridership of subway and bus services in New York City?

Answer: The feature importance plots generated by the random forest models indicate that temperature, precipitation, and snowfall are the top three features that have the most significant impact on subway and bus ridership in New York City.

2.How does the performance of the models vary when using different hyperparameters for the random forest algorithm?

Answer: The performance of the models can be evaluated using different hyperparameters such as the number of trees, the maximum depth of the trees, and the minimum number of samples required to split a node. Tuning these hyperparameters using techniques such as grid search or randomized search can improve the model performance.

3.Can the same approach be applied to predict the ridership of other modes of transportation such as taxis or bikes?

Answer: Yes, the same approach can be applied to predict the ridership of other modes of transportation. However, the feature engineering and selection process may differ based on the characteristics of the transportation mode being studied.

4. can the model be improved to account for the impact of external events such as strikes or major events in the city?

Answer: Additional features can be added to the model to account for the impact of external events such as holidays, major events, or strikes. These features can be obtained from publicly available datasets or through web scraping techniques. The impact of these features can then be evaluated using techniques such as correlation analysis or feature importance plots.

5.How can the model be used to optimize the allocation of resources for transportation services in New York City?

Answer: The model can be used to predict the ridership for each mode of transportation on a given day. This information can be used to optimize the allocation of resources such as the number of buses or subway cars needed to meet the demand. The model can also be used to forecast the ridership for future periods, which can aid in capacity planning and resource allocation.

Concept Explanation :

The algorithm we used in this project is called the random forest algorithm. Imagine a forest with lots of trees, and each tree is a decision maker. In our case, each decision maker is a model that predicts the ridership of transportation services based on various weather-related factors.

Let's say we have to predict whether a person will eat pizza or not. We have several factors like age, gender, location, weather, and so on, and each factor has some impact on the decision. Now, let's say we have a bunch of people who have already eaten pizza, and we know whether they liked it or not based on the same factors.

To build our random forest model, we take a random subset of people who have eaten pizza, and we build a decision tree based on the factors that influenced their decision. We repeat this process multiple times with different subsets of people and different factors, and we build multiple decision trees.

Once we have built all the decision trees, we use them to predict whether a new person will eat pizza or not based on the same factors. Each decision tree gives its own prediction, and we take the majority vote of all the decision trees to make our final prediction. This way, we get a more accurate prediction than just relying on a single decision tree.

In our project, we used the random forest algorithm to predict the ridership of transportation services in New York City based on various weather-related factors. We built multiple decision trees using a random subset of data and different factors, and we took the majority vote of all the decision trees to make our final prediction.

I hope this funny and friendly explanation helped you understand the random forest algorithm!