Executive Summary

We will identify and interpret factors that relate to higher or lower levels of pedestrian activity. Insights about drivers of pedestrian activity are expected to be useful when considering new locations across the U.S. for stores and determining the optimal hours for having stores open. To assist you with this problem, we have built a model to predict pedestrian traffic levels that we believe provides such insights.

The data studied is from NYC Open Data and were collected in 2017-2019. The 11,373 records contain information about the number of pedestrians observed during each hour, the time of day (in one-hour segments), the day of the week, and various weather information. Prior to our work, the data had already been cleaned. Because the data was collected in New York City, insights gained from the data may not translate well to other places where climate patterns, public transportation, and pedestrian-friendly city planning may differ.

Preparing for Modeling

To begin, we explored whether or not the variables were likely to predict the number of pedestrians based on data exploration and domain knowledge. We determined that the time of day variable was most likely to be important, and the temperature variable was least likely to be important because it duplicated information contained in the forecasted daily average temperature.

Next, we transformed several variables to make them more suitable for modeling. Additionally, many of the variables in the dataset were related to each other in ways that would make it difficult for the models to separate their impacts. After analyzing the data, we made the following changes:

- Rather than considering each of the seven days of the week on their own, we grouped Monday through Friday since they are weekdays, and considered those separately from each of Saturday and Sunday.
- We regrouped the weather variable into 3 categories: “nice” (clear or partly cloudy), “mild” (cloudy, windy, or fog), and “inclement” (rain, snow, or sleet).
- We changed the time of day variable to measure the number of hours away from 2 p.m.
- We decided to use the forecasted daily average temperature instead of the hourly temperature measure.

Model Selection

Three types of models were considered for this project, decision trees, generalized linear models (GLMs), and random forests. To make a prediction, decision trees use a series of if-else statements, GLMs use a mathematical formula, and random forests combine the predictions of many decision trees. We concluded that random forests were not a good fit for

the project, as they were not likely to yield useful insights about the drivers of pedestrian traffic due to being difficult to interpret. Several GLMs and decision trees were trained and compared.

Our final model is a decision tree. To select this model, we considered (1) the ability to gain useful insights from the model, (2) the model performance as measured by root mean squared error (RMSE) – a commonly used metric for measuring model accuracy, and (3) the ability to apply those insights to new localities. The tree model is easy to interpret, so we will be able to extract meaningful insights. It performed significantly better than the other models as its predictions had smaller errors on average. The decision tree was not as good as the GLM for applying the useful insights to new localities, but the performance of the GLM was worse enough that we did not think the insights from it would be as useful in any localities, so the decision tree was selected.

The Model

The following 5 factors were important for determining the number of pedestrians (listed in order of importance):

- The number of hours difference between the current time and 2 p.m.
- The average temperature forecast
- The hourly rate of precipitation in inches
- Whether it was Saturday or not
- Weather conditions as categorized in the second bullet under “Preparing for Modeling” above

Key insights from the model are:

- The time of day is important. In fact, whether it was between 9 a.m. and 7 p.m. was the most important driver of pedestrian traffic, with more between those times.
- Colder average temperature forecasts were associated with fewer pedestrians, but the temperature forecasts leading to more or less pedestrians is dependent on the time of day.
- The weather conditions are only important at some combinations of time of day and average temperature forecast.
- The two situations that happen with meaningful frequency that predict a large number of pedestrians are
    - Between 9 a.m. and 7 p.m., temperature forecast above 51 degrees, not Saturday, and precipitation less than 0.003. 1825 pedestrians are predicted and this setting occurred 22% of the time.
    - Between 11 a.m. and 1 p.m., temperature below 51 degrees, and precipitation less than 0.006. 1518 pedestrians are predicted and this setting occurred 12% of the time.

- Conversely, the fewest predicted number of pedestrians (131) are before 8 a.m. or after 8 p.m. along with a temperature below 62 degrees. This scenario occurred 22% of the time.

These results make intuitive sense. Our decision tree found that there is less pedestrian activity when in early morning, late evening, or when most are sleeping at night. Additionally, more pedestrians are present in weather conditions that are more comfortable for walking.

Recommendation

A decision tree model was constructed provides insights and predictions regarding pedestrian activity. As expected, time of day and weather conditions were the primary drivers. Because our model was trained using data collected from just New York City, caution should be used when applying the existing model to new localities. We recommend seeking additional data from other localities to improve the model.