Abstract

The objective of this project was to use classification models to predict the utilization status of New York rail traffic in order to help improve income and plan development for this sweet factory. I worked with the data provided by the MTA, taking advantage of data science engineering along with a random forest model to achieve promising results for this multi-layered problem. After optimizing the model, I built an interactive dashboard to visualize and communicate my results using Python.

Design

This project arose out of the development and utilization of New York City's congested rail system. The data is provided by the MTA, and we want to expand the Halawa factory across the country. Accurately classifying cases via machine learning models will enable the MTA to take action to improve services, allocate resources more quickly to required areas, and ensure that there is no stampede in the subway.

Data

The data set contains 2,669,039 rows of data. Some of the notable features include measurements of the amount of entry and exit, and since it is the Halawa factory, we care about the exit. Approximately one-third of the individual features can be grouped into more general categories, and an 8-h in-depth analysis of the exit and data engineering was performed.

Algorithms

Pandas reset_index() is a method to reset the index of a Data Frame.
sum() is a built-in method in java which returns the sum of its arguments.
set_index() is a method to set a List, Series or Data frame as index of a Data Frame.
They were all helpful to clear the data.

Tools

Pandas for data manipulation
Matplotlib and Seaborn for plotting
Python

Communication