INFO6105 : Data Science Engineering Methods

Team No : 2 (Rutuja More, Nitant Jatale)

Topic : BlueBikes Demand Forecasting & EDA

Introduction

BlueBikes is a service that allows people to rent bicycles for a fee for a limited time, usually by the hour or day. In many cities around the world, BlueBikes have become a popular mode of transportation. As people become more aware of the environmental impact of transportation, BlueBikes has emerged as a viable and cost-effective alternative to traditional modes of transportation. However, external factors such as weather, air quality, seasons, and Covid-19 can have a significant impact on BlueBikes systems. As the popularity of BlueBikes services grows, it is critical that these companies accurately forecast demand for their bikes in order to satisfy a rapidly growing customer base. The ML forecasting models allow BlueBikes companies to predict demand for bikes over different time periods and see if any factors, such as seasonal, air quality, and weather conditions, affect demand. This will benefit the consumer by ensuring greater accuracy with bike availability for use, reducing the risk of a customer wanting to use a bike but not having access to one.

The motivation for this project is to explore the relationship between BlueBikes demand and external factors such as weather, air quality, Covid-19, and seasons. BlueBikes’ station-based loaning system creates a need for forecasting demand. With a station-based loaning system, a user picks up a bike from a docking station and returns it to any station which means Bicycles must be available at a docking station for the user to rent the bicycle.

The goal of this project is to develop a machine learning model that will forecast the number of bikes rented daily based on historical rider data and other variables such as weather, Air quality, Covid-19 and Seasons. Predicting daily bike ridership can help identify fluctuations and trends that could support the bike rebalancing system as well as business decisions such as the need for additional stations. When weather or air quality is bad, people might be reluctant to rent a shared bike. Demand for bikes might be higher in summer than harsh winters. The impact of Covid-19 on bike renting demand will also be analyzed, with confirmed cases and deaths analyzed with a one-day time lag. We aimed to pinpoint the most important features that can accurately predict bike demand.

Methodology

EDA :

In this project, EDA would involve analyzing the available data related to bike ridership, including historical demand data and external factors such as weather, air quality, and seasonal patterns. EDA would involve analyzing historical demand data to identify any patterns or trends in demand over time, and examining external factors such as weather patterns and air quality to see if they have any impact on bike demand. This information can then be used to select appropriate features for the machine learning model and preprocess the data for optimal performance.

Linear Regression :

Linear regression is a commonly used statistical technique for predicting a continuous variable based on one or more predictor variables. In the case of predicting demand for BlueBikes, linear regression can be used to model the relationship between the number of bikes rented and various factors that might influence demand, such as weather conditions, time of day, day of the week, and holidays.

Decision Trees :

Decision trees work by dividing the dataset into smaller subsets, using a series of binary decisions to determine which subset each observation belongs to. Each decision in the tree is based on a specific feature, and the algorithm continues to split the data until a stopping criterion is met, such as reaching a certain depth or number of observations. Decision trees can handle both categorical and continuous variables, which is useful when predicting demand for BluBikes systems, where variables such as weather conditions, season, and time of day may be important predictors.

Random Forest :

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. Random Forest can identify which variables are the most important predictors of BluBikes demand. This information can be used to identify patterns in the data and to make informed decisions about what factors to consider when forecasting demand. It can also capture nonlinear relationships between variables. For example, the impact of temperature on bike demand may not be linear, and Random Forest can detect these types of relationships. Random Forest is robust to outliers and noise in the data, which is important for predicting bike sharing demand, as there may be unexpected events or anomalies that can affect demand.

By comparing the accuracy of demand forecasting using these three machine learning models, we can determine which model provides the most accurate results for predicting BluBikes demand.

Description

The historical data for BlueBikes rides in Boston was accessed from Bluebikes's website, which contained a zip file of ride information for each day from the year 2019 to 2022. The file contains 14-15 attributes, depending on the year of the data, including trip duration, start date time, stop date time, start station id, start station name, start station latitude, start station longitude, end station id, end station name, etc.

In addition to the above Bluebikes data, we have incorporated weather data. This Boston weather data was accessed from the National Oceanic and Atmospheric Administration (NOAA) website which keeps daily summaries of historical weather for the city. This file contains multiple attributes for the data, the ones which we have included for the ML model are date, average wind speed, precipitation amount, snowfall amount, average temperature, minimum temperature, maximum temperature, etc.

The third dataset is the COVID-19 data. That was accessed from the USA Facts website, which keeps official records on COVID-19 related information and data. The dataset contains multiple records and attributes, the ones we will be using for forecasting in this project include the date, the number of new cases, and the 7-day moving average.

The fourth dataset is the Air Quality Index data, which was accessed from the United States Environmental Protection Agency Website. This dataset contains attributes like Date, Air quality index, carbon monoxide levels, etc.

Dataset

1. **https://www.bluebikes.com/system-data**
   (https://www.kaggle.com/datasets/jackdaoud/bluebikes-in-boston?select=bluebikes_tripdata_2019.csv)
2. **https://www.weather.gov/wrh/Climate?wfo=box**
3. **https://www.epa.gov/outdoor-air-quality-data/download-daily-data**
4. **https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/massachusetts/county/suffolk-count**