PERFORMANCE COMPARISON OF PUBLIC BIKE DEMAND PREDICTIONS: THE IMPACT OF WEATHER AND AIR POLLUTION

by

Min Namgung

A thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Computer Science

Fort Wayne, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Jin Soung Yoo, Chair

Department of Computer Science

Dr. Adolfo S. Coronado

Department of Computer Science

Dr. Peter A. Ng

Department of Computer Science

Approved by:

Dr. Jin Soung Yoo

Dedicated to my beloved family

ACKNOWLEDGMENTS

I would like to acknowledge all of those who have believed in me: My family and friends. Thank you to my family for your unconditional love and support. I appreciate my parents the most to support and encourage me through everytime. I appreciate my grandmother, who always make me strong and cheer me up. I couldn't have reached this achievement without you. I would also like to show my appreciation to Dr. Jin Soung Yoo for her remarkable guidance on my thesis with her insight and knowledge, Dr. Adolfo Coronado for his outstanding mentorship and invaluable support, and Dr. Peter Ng for his generosity and comments on my research. I appreciate all of their advice and counsel on my thesis. These professors have helped me gain motivation and passion for my research and have encouraged me in my pursuit of a master's degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

NYC New York City

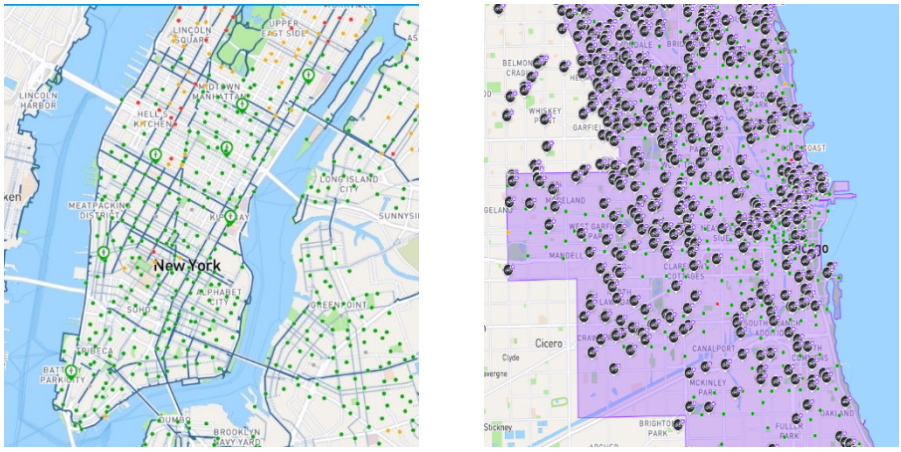
ABSTRACT

Many metropolitan cities motivate people to exploit public bike-sharing programs as alternative transportation for many reasons. Due to its' popularity, multiple types of research on optimizing public bike-sharing systems is conducted on city-level, neighborhood-level, station-level, or user-level to predict the public bike demand. Previously, the research on the public bike demand prediction primarily focused on discovering a relationship with weather as an external factor that possibly impacted the bike usage or analyzing the bike user trend in one aspect. This work hypothesizes two external factors that are likely to affect public bike demand: weather and air pollution. This study uses a public bike data set, daily temperature, precipitation data, and air condition data to discover the trend of bike usage using multiple machine learning techniques such as Decision Tree, Naïve Bayes, and Random Forest. After conducting the research, each algorithm's output is evaluated with performance comparisons such as accuracy, precision, or sensitivity. As a result, Random Forest is an efficient classifier for the bike demand prediction by weather and precipitation, and Decision Tree performs best for the bike demand prediction by air pollutants. Also, the three class labelings in the daily bike demand has high specificity, and is easy to trace the trend of the public bike system.

1. INTRODUCTION

With increasing prevalence in urban environments, bike-sharing systems have become a prominent feature across city space worldwide. They provide a low-cost, environmentally-friendly transportation alternative for cities that eases traffic congestion, improves public health, and cuts carbon emissions. Due to their tremendous eco-friendly characteristics, many cities establish their bike-sharing system as a transportation method. Representative bike-sharing systems in the US are the Citi bike (NYC bike-sharing system) [citibikenyc.com] and Divvy bike (Chicago bike-sharing system) [divvybikes.com]. When the Citi bike first launched the service in May 2013, there were 332 stations and 6,000 bikes. Similarly in Chicago, when the Divvy bike first started the service in June 2013, there were 75 stations and 750 bikes. Currently in 2020, the Citi bike has 1000 stations with 15,000 bikes in Manhattan, Brooklyn, and Queens [citibikenyc.com]. Based on the Nyc.gov report, released in May 2019, nearly 1.6 million people ride a bike at least once a year in NYC, and of those riders, nearly 800 thousand rides the bike-sharing system regularly. Similarly, due to the increment of the Divvy bike-sharing users, based on Chicago.gov, Chicago's Divvy bike had around 600 stations with 6000 bikes [divvybikes.com] in 2020. As the public bike-sharing system becomes popular, and people commonly use public bikes, better bike management is required. To better management, a prediction of bike demand is one of the core factors to manipulate the stations and the public bikes for the users.

The more the public bike user count increases, the more demand for the public bike system accumulates. Due to the popularity of the public bike, effective public bike management is required. This research aims to propound a trend of public bike demand by comparing two metropolitan cities that have managed the public bike system by external factors that could affect bike users. There are possibly many factors that impact the usage of the bike-sharing system. In this research, we assume weather data such as temperature, precipitation, and air pollution are the most important factors to influence the public bike system. To fulfill this goal, the following questions are discussed: How does weather precipitation and temperature affect the daily bike demand? And How much do air pollutants influence the bike usage? This work firstly explores how public bike demand varies by weather conditions, such as



(a) NYC Citi bike station in Manhattan area, 2020

(b) Chicago Divvy bike station, 2020

Figure 1.1. Public Bike Stations in NYC (a) and Chicago (b)

maximum/minimum daily temperature and rain/snow measurement. Secondly, this research shows how much the air quality and the public bike usage correlate or how much the daily air condition affects public bike users' decisions. Once both objectivities are conducted in each city, the result of NYC's and Chicago's trend of bike usage is compared to predict the bike demand model.

However, developing the public bike demand prediction is challenging, based on the bike trip observations and weather conditions. The reason is that the analysis of the bike demand models from NYC and Chicago are different, and the external weather conditions vary between the two cities. Each city has a different distributed population, and the income or tax rate is also noticeably different. Because of those reasons, citizens who live in Chicago may prefer to ride with the public bike system due to the high state tax rate, and citizens who live in NYC might not consider taking other transportation when they commute, or vice versa. To analyze more accurate bike demand prediction models, the total percentage of bike riders based on the city's population should be considered. Also, the scale of city or the length of the available bike path is dissimilar in both cities. Due to peripheral reasons, the public bike system's popularity is challenging to clarify and predict the bike demand model from two metropolitan cities.

Even if this research has a limitation of different scale of population based off of two large cities, the analyzed output data will contribute to several organizations: City Halls which already manipulate the public bike system and New Cities which have not yet organized a public bike system. In academia, this research will be one of the first steps for forecasting the alternative transportation's relationship between external factors, especially in air conditions and air pollutants.

To analyze the Citi bike and Divvy bike data, this work uses a Decision Tree, Naïve Bayes, and Random Forest. Decision Tree in this work aims to predict the range of public bike usage as target data with daily maximum/minimum temperatures and each air pollutants' daily maximum measurement as the prediction's features. This work also adopts Naïve Bayes to generate a probability of daily bike usage and find a relationship between bike demand and weather conditions, and between the range of bike demand and air pollution. Lastly, this work applies Random Forest to discover a more precise correlation and the range of bike usage between daily demand and each air pollutants' measurement or daily average temperature and precipitation. This study will focus on 2019 NYC Citi bike data and 2019 Chicago Divvy bike data for measuring daily bike demand. External affecting factors to the public bike demand include 2019 daily weather data and 2019 daily air pollution measurements categorized in carbon monoxide (CO), nitrogen dioxide (NO_2), particulate matter less than 2.5μm ($PM2.5$), ozone(O_3), and oxygen saturation (SO_2) for both NYC and Chicago.

The case study shows that daily average temperature and maximum precipitation influence the range of public bike demand. When the air pollutants measure high, the bike usage is likely to be low but not have as much of an affect as the weather data. In the prediction performance, Random Forest and Decision Tree have a higher accuracy rate among the three classifiers. Random Forest increases each task's accuracy and probability by bagging individual trees. Decision Tree finds each task's trend of public bike system with features and labels.

To illustrate, the paper will be laid out in the following sections: Section 2 Problem Description; Section 3 Related Work; Section 4 Data Description and Data Exploration; Section 5 Methodology; Section 6 Task One: Bike Demand Prediction by Weather Condition;

Section 7 Task Two: Bike Demand Prediction by Air Pollution; Section 8 Conclusion and Future Work.

2. PROBLEM DESCRIPTION

This work aims to propound the public bike demand prediction based on weather and air condition by comparing two metropolitan cities that manage a public bike system. Among many other external factors which impact the public bike demand, this research focuses on daily temperature, precipitation, and air pollutants' measurement.

To produce a more accurate bike demand prediction models, the public bike analysis applies multiple machine learning algorithms, such as Decision Tree, Naïve Bayes, and Random Forest, then compares each learning's algorithm and explores which algorithms perform the best at each task using performance comparison results. Before exploring the public bike demand prediction models with those external factors, this study conducts data exploration with the bike demand observation data set.

This work conducts how weather conditions affect public bike demand. The first hypothesis in this research is daily weather temperature and precipitation impact on the daily bike demand. Due to daily weather being one of the most impacting external factors for transportation or traffic, the public bike demand is likely to change based on this first external impacting factor. In addition, this research shows how much the air quality and the public bike users are correlated, or how much the daily air condition affects the public bike demand. The second hypothesis is air condition and is one of the affecting factors on daily bike usage. Some countries, such as China, India, or Korea, have been negatively affected by severe air conditions. The air pollution dilemma encourages the public bike system as alternative transportation. Still, at the same time, the severe air pollution discourages outdoor activity for the public. For these reasons, predicting the public bike demand, dependent on the air conditions, is going to be necessary for future research. However, although NYC and Chicago's air environment are not as severe as in other countries, the air quality might be one of the affecting factors for public bike usage. This hypothesis and the previous external factor extend into a second hypothesis, allowing the verification of another external factor which impacts daily bike demand.

Once NYC and Chicago's public bike demand is analyzed these two objectivities, the public bike system trends in NYC and Chicago are compared with the performance comparisons' table.

3. RELATED WORKS

Public bike-sharing system analysis becomes popular in many cities as it becomes an increasingly popular alternative transportation method in modern society. The public bike-sharing system is analyzed in many ways: management-level and user-level. To have a deeper layer of understanding, it is necessary to approach multi different levels in the public bike-sharing system.

3.1 Public bike-sharing demand prediction on management-level

The public bike-sharing system has been expanding for the past decade. As the public bike systems increase in scale, an extensive amount of attention is paid to public bike-sharing demand prediction models. There are three clusters of demand prediction models in the literature: city-level, station-level, and neighborhood-level. For the city-level prediction model, the model aims to predict the bike usage for an entire city. Y. Li et al. (2020) suggested a hierarchical consistency prediction model to predict the citywide bike-sharing system. They proposed an Adaptive transition constraint clustering algorithm, a similarity-based efficient Gaussian Process regressor, and a General least square formulation to predict the causality between rent and return of real-time public bike-sharing usage. R. Giot et al. (2014) proposed a prediction of bike-sharing system usage for the next 24 hours on the city-level side by using two years of bike-sharing models in Washington D.C. area. Their research was tested by multiple regression models such as Ridge regression, Adaboost regression, Support vector regression, Random forecast tree and Gradient boosting regression tree, and evaluated their outputs. However, this research argued that there are over-fitting issues when the delay is greater than one hour, which requires a further utilization by using multiple pieces of information, such as origin, destination, trip duration, check-in/out time, or user information. The bike-sharing prediction model on station-level is one of the most popular and challenging levels of predicting bike demand. P. Hulot et al. (2018) proposed a bike-sharing prediction model to simplified behaviors using the external context data, then predicted the bike-sharing system usage on station-level. They used multiple machine learning techniques such as Linear regression, Multi-layer perception, Gradient boosted tree, and

Random Forest. Y. Li et al. (2015) proposed a hierarchical prediction model by tracking each station cluster. Their research primarily used bipartite clustering to sort out stations into two levels. Then, the total number of bikes were predicted by a Gradient boosting regression tree for evaluating Washington D.C. and NYC. However, the bike-sharing prediction models on the neighborhood-level were quite challenging. Many researchers have paid attention to neighborhood-level in multiple ways. D. Singhvi et al. (2015) predicted New York Citi bike-sharing system usage on neighborhood-level by analyzing regression models with several external data contexts such as NYC weather and taxi data, demographic and housing factors as covariates in predicting pairwise trips. R. Rixey (2013) found a contiguity to other bike share stations and to densely populated neighborhoods, and higher levels of income and education are positively correlated with bike-sharing stations usage. Also, non-white populations and precipitated weather negatively affected the demand of bike-sharing system. The author concluded with those correlation factors by analyzing Census block data to border on neighborhood-level.

3.2 User prediction of Public bike-sharing demand on user-level

Not only the increased demand of bike-sharing trend, but also user predictions on bike-sharing system have been popular. Along with the demand on bike-sharing prediction models, a prediction of bike riders has also been given from multiple researches. There are some valuable findings in the bike-sharing riders' predictions. R. Beecham et al. (2014) analyzed cycling trips by riders of London's bike-sharing system. They analyzed the Spatio-temporal context under each cycling journey. They found that women's trips tend to be structured, such that women use public bikes at weekends and within London's parks, while men tend to use public bikes for commuting. J. Zhao et al. (2015) explored bike-sharing travel time and trip cycling patterns by gender and day of the week by analyzing z-score values. They found that women were more likely to make multiple-circle trip chains than men on weekdays. H. I. Ashqar et al. (2017) modeled the number of available bikes in San Francisco Bay by using Random Forest and Least-squared boosting, and Partial least-squared regression. Their research found that station neighbors, the prediction horizon time, and weather variables were significant factors in modeling, to predict the number of available bikes. D Freund et

al. (2017) provided new approaches to rebalance overnight and new optimization for other non-motorized rebalancing efforts during the day in NYC. Their goal was customer satisfaction, and they suggested a new integer program considering pick up/drop off bikes or moving to an adjacent station. J Yoon et al. (2012) suggested the navigating advisor application for the usage of public bike-sharing system. The navigating advisor provided the best pair stations and found the shortest paths for bike users by using machine learning algorithms. Y Zhou. (2019) proposed bike placement based on bike demand for existing and new stations. They set different distance thresholds by considering place embedding or station geography, then evaluated the model's RMLSE and error rate with the real bike data sets. Froehlich et al.(2009) implemented four predictive models to forecast the total available bikes at each bike station: last value, historical mean, historical trend, and a Bayesian network. They used two methods to analyze with time series methods: Auto-Regressive Moving Average (ARMA) and Auto-Regressive Integrated Moving Average (ARIMA). These methods were used to predict the number of available bikes/docks for each station. Gallop et al. (2011) used a seasonal Autoregressive integrated moving average time series analysis to account for the complex serial correlation patterns and tested the model against actual bicycle traffic counts in Vancouver, Canada. The results demonstrated that the weather had a significant impact on the bike demand, particularly temperature, rain, humidity, and clearness were significant. Ashqar et al. (2019) demonstrated the bike availability model in San Francisco Bay Area. They used Poisson regression model and Negative Binomial Regression model for the bike count model, Random Forest algorithms for the bike availability prediction model, and Bayesian information criterion for the comparison between models following a forward step-wise regression guided by the results of Random Forest. They found that the time-of-the-day, temperature, and humidity levels were significant count predictors in the bike system.

4. DATA AND DATA EXPLORATION

This section explains the 2019 public bike-sharing data set in NYC and Chicago. This bike data set is commonly used for each task and shows data exploration in each city.

4.1 Data

There are two different types of bike trip data used in this research. In this section, the first data set is the NYC Citi bike trip data provided by the Citi Bike official website [citibikenyc.com]. This work used the data from January 1, 2019, to December 31, 2019. The second data set is the Chicago Divvy bike trip data collected from the Divvy official page [divvybikes.com]. To match with the NYC Citi bike data set, the Divvy bike trip data set includes bike trip records from January 1, 2019, to December 31, 2019. Unlike the Divvy bike, the Citi bike contains the station latitude/longitude. However, both NYC's Citi bike and Chicago's Divvy bike comprise similar trip features.

NYC Citi Bike Original Data Set:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station ID
- Start Station Name
- Start Station Latitude/Longitude
- End Station ID
- End Station Name
- End Station Latitude/Longitude
- Bike ID
- User Type (Customer=24-hour/3-day pass user; Subscriber=Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- User's Birth Year

Chicago's Divvy bike trip data is less tidy than NYC Citi bike data. There are some missing values in the gender and customer's birth year, and the headers are inconsistent. Below is a list of 2019 Chicago's Divvy bike trip data indicators.

Chicago Divvy Bike Original Data Set:

- Trip ID
- Start Time
- End Time
- Bike ID
- Trip Duration
- From Station ID
- From Station Name
- To Station ID
- To Station Name
- User Type
- Gender
- Birth Year

4.2 Data Cleaning

The final bike data trip is required for data preparation due to missing data for customer observation in Divvy trip data, and inconsistency between the bike trip data sets from NYC and Chicago. For the NYC bike trip data, start time and stop time are formatted with 2019-01-01 00:01:47 to categorize each observation into a particular time, including the observation at 7 or 10 am. Along with the time, a new column "wday" is added to separate weekdays and weekends. To calculate the correct "wday", each date is labeled with numbers 1 to 7 and adequately replaced with 1: Sunday, 2: Monday, ... , until and 7: Saturday. The "trip duration" is also converted to a minute value by dividing 60 from the original trip duration (sec) in addition to the bike trip date. The "age" and "age group" are also evaluated from the trip observation's birth year to categorize each age group's trip data. Similarly, the Chicago trip data also adds "wday" to represent weekdays, and "age" and "age group" are

also labeled in the same way as NYC bike trip data preparation. The missing values are omitted because the research aims to classify the bike user class for each situation.

As shown in Table 4.1, the final bike trip data in NYC and Chicago is formatted with these ten indicators. The final bike data is the fundamental data set in this research. The total number of data in NYC is 20,551,697, and the number of data in Chicago is 3,279,253. Among these tremendous numbers of data, we only focus on morning rush hour, 7 am to 9 am, daily time frames. After sorting the morning rush hour data, the remaining number of data in NYC is 3,003,190, and the total number of bike trips during Chicago during rush hour is 631,268.

Table 4.1. Tided Bike Trip Data Example

tripduration	**usertype**	**gender**	**wday**	**hour**	**tripMin**	**age**	**age group**	**tripDate**	**bikeid**
834	Subscriber	Female	Tuesday	7	13.9	41	40s	2019-01-01	3244
562	Subscriber	Female	Tuesday	7	9.4	43	40s	2019-01-01	33186
1332	Subscriber	Male	Tuesday	7	25.4	32	30s	2019-01-01	3544
844	Subscriber	Male	Tuesday	7	14.1	25	20s	2019-01-01	3228

4.3 NYC Data Exploration

To get more familiar with the experiment data, we first explore the public bike system data from NYC and Chicago using data visualization. The Citi Bike trip data in 2019 contains over 20,551,694 trip records. Among those trip data, male users take almost three-quarters of all trips, while female users take only a quarter.

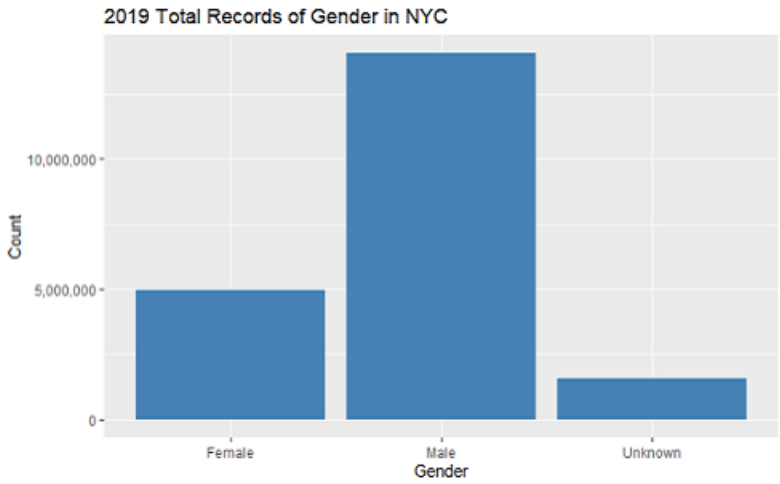
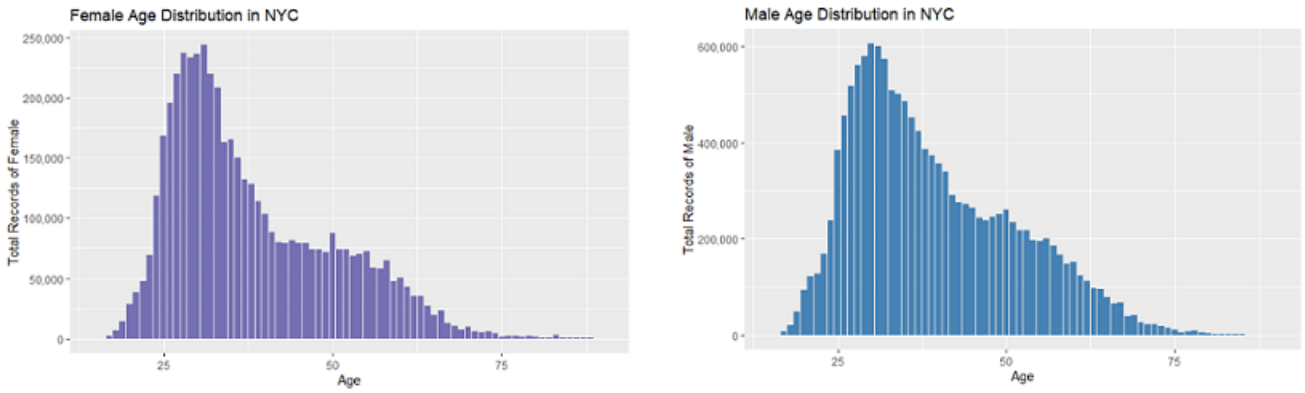


Figure 4.1. Gender Distribution in NYC



(a) Female Age Distribution (b) Male Age Distribution

Figure 4.2. Bike User Distribution in NYC

As shown in Figures 4.2 and 4.3, the later 20s and early 30s age group have the most public bike system usage. An interesting finding from the results is that the 40s generation's bike usage is less than the 50s generation. Figure 4.2 (a, b) shows that the age distribution

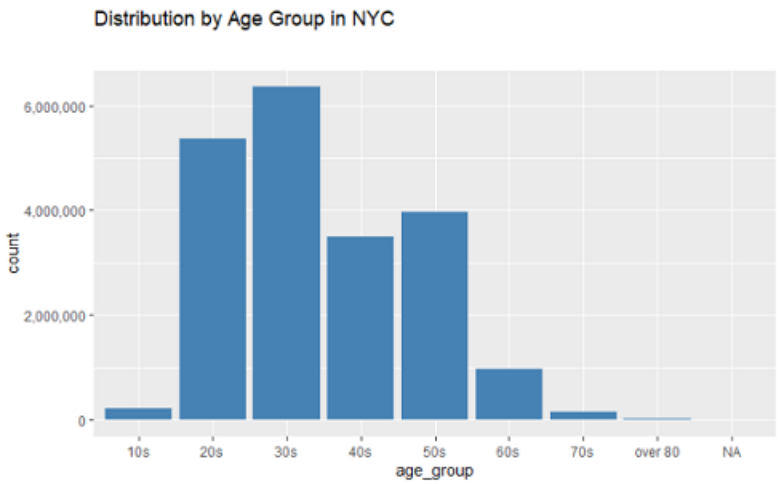


Figure 4.3. Age Group Distribution in NYC

in the 50s is relatively smaller than in the 40s; however, the overall age group distribution in the 50s is larger than the 40s, as shown in Figure 4.3. The age distribution in females and males draws a similar distribution. As we can expect so far, many young bike users who travel to work daily have used the public bike system.

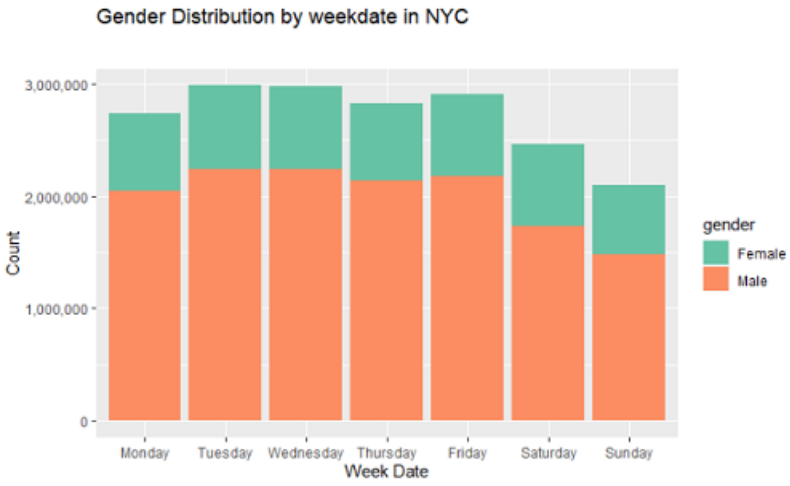


Figure 4.4. Bike Demand Distribution per gender by Week Date in NYC

Figure 4.4 displays that the public bike demand is relatively greater during weekdays than weekends. Also, the trends of male bike users are three times larger than the trends of male bike users on a daily basis. Monday through Friday has the most public bike usage, compared to the weekends.

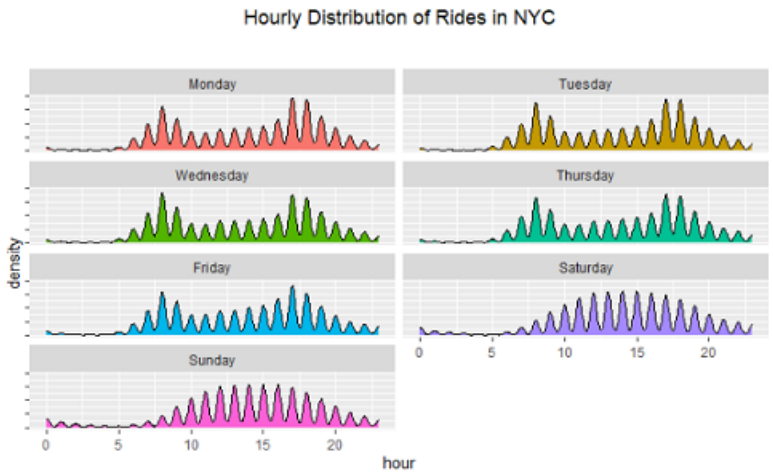


Figure 4.5. Hourly Demand Distribution by Date in NYC

Figure 4.5 shows that the peak hours are 7 am to 9 am and 4 pm to 6 pm on weekdays during the usual commute to work. During weekends, public bike usage steadily increases and decreases as the day goes on. For this reason, this research focuses on the 7 am to 9 am data, for the 2019 NYC Citi Bike data set, in consideration of the higher bike demand during the weekdays.

4.4 Chicago Data Exploration

The Divvy Bike trip data in 2019 contains over 3,279,253 trip records. Among those trip data, male users take almost three-quarters of all trips as similar to the NYC Citi bike data set. In contrast, female users take only a quarter.

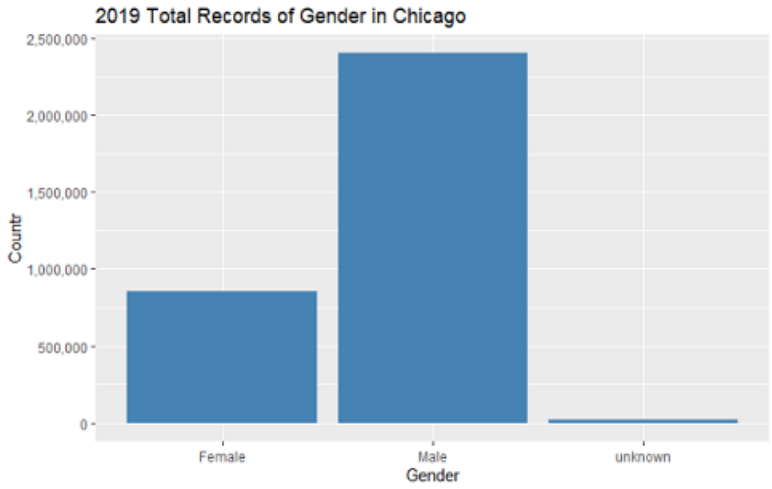
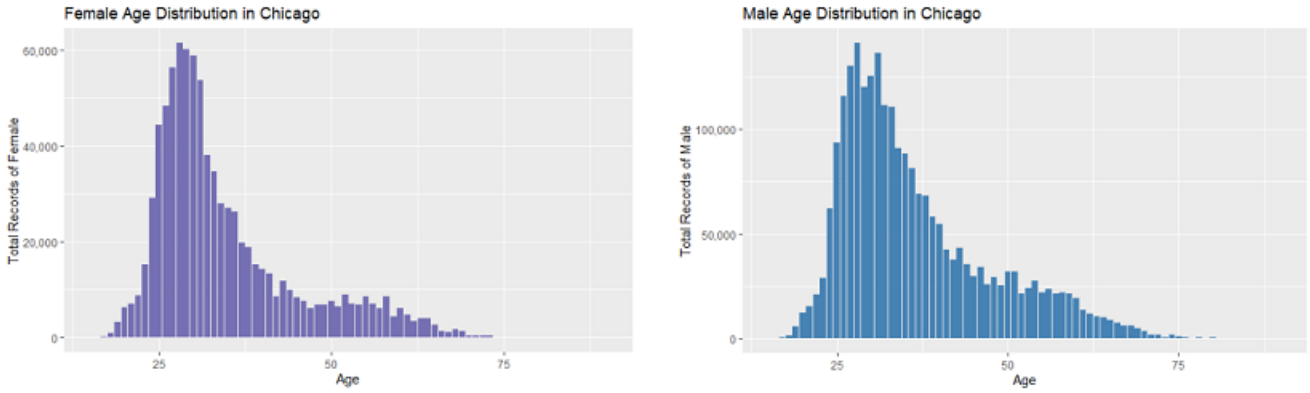


Figure 4.6. Gender Distribution in Chicago



(a) Female Age Distribution (b) Male Age Distribution

Figure 4.7. Bike User Distribution in Chicago

As shown in Figures 4.7 and 4.8, the later 20s and early 30s age group have the most public bike usage, similar to the NYC Citi bike. Compared to the NYC public bike system in Figure 4.3 and 4.8, the 10s' generation is relatively small in Chicago, and the 20s and 30s groups use the bike the most in both cities. In Figure 4.8, the 20s age group has more users

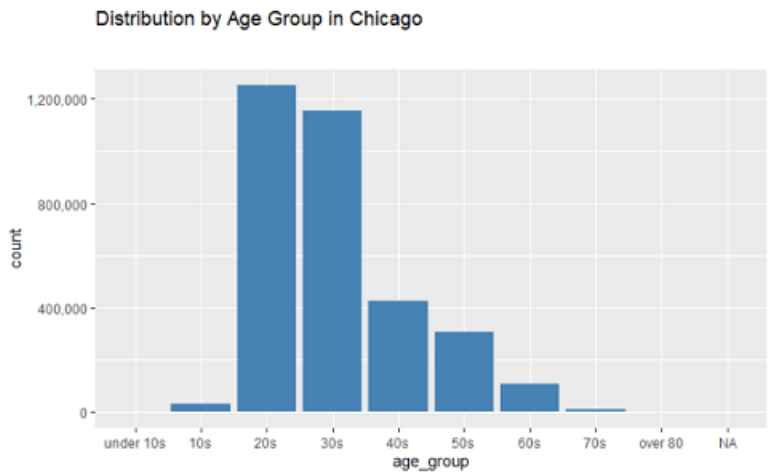


Figure 4.8. Age Group Distribution in Chicago

than the 30s age group. Oppositely the 30s age group is larger than the 20s in NYC, based on Figure 4.3. Also, the 40s and 50s bike users count is relatively small, compared to the bike users count from the same age group in NYC.

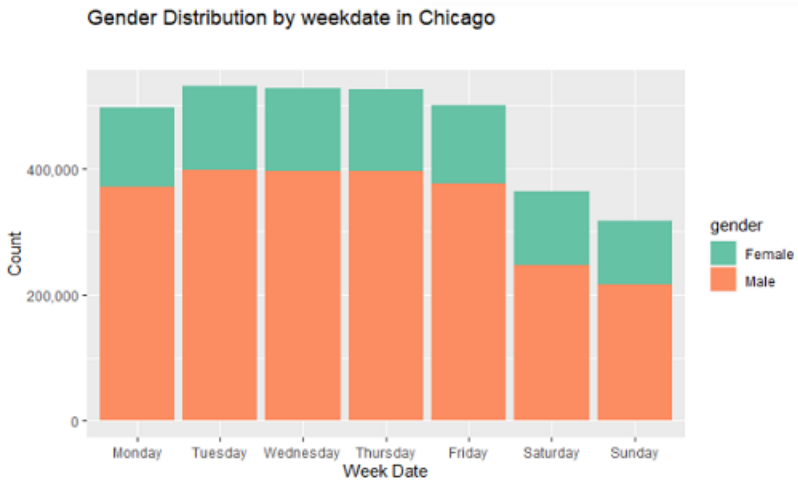


Figure 4.9. Bike Demand Distribution per gender by Week Date in Chicago

Figure 4.9 shows that the public bike demand is relatively greater during weekdays than weekends. Also, similar to NYC, the male bike users count are three times larger than the female users count daily. Monday through Friday has the most public bike usage, compared to the weekends.

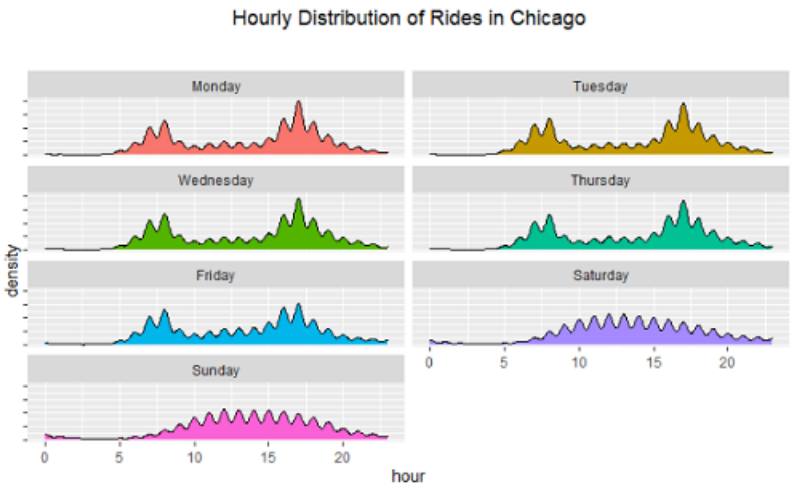


Figure 4.10. Hourly Demand Distribution by Date in Chicago

As shown in Figure 4.10, the busiest hours are 7 am to 9 am and 4 pm to 6 pm on weekdays. Compared to the NYC Citi bike, Figure 4.10 represents a much more clear bike usage record during rush hour. Therefore, this research focuses on 7 am to 9 am for the 2019 Chicago Divvy Bike data set, considering the higher bike demand during weekdays at that particular time frame. In summary, the research aims to process morning rush hour data to predict the public bike analysis from both cities.

5. METHODOLOGY

This section will discuss multiple machine learning algorithms, Decision Tree, Naïve Bayes, and Random Forest, applied in the prediction of daily bike demand by weather and air pollution.

5.1 Decision Tree Induction

Decision Tree technique is a popular tree classification because both numeric and categorical data are well-suited. Decision Tree consists of internal nodes, each associated with a logical training test set and possible consequences. Therefore, the final tree form makes it easy to find the trend of popularity. In Decision Tree learning, predictive values have built Decision Tree from observations to conclusions about the item's target value. Also, Decision Tree is very flexible in choosing the number of training features.

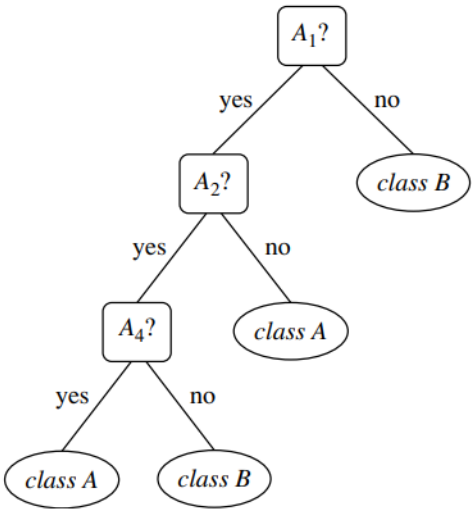


Figure 5.1. Decision Tree example, Han, J. et al. (2012) Figure 8.6

As shown in Figure 5.1, Decision Tree constitutes the root node, internal nodes, and leaf nodes. A rectangle denotes internal nodes, and ovals denote leaf nodes. Decision Tree follows a top-down approach, which begins with a training set of tuples and their associated class labels. The training set is recursive partitioning into smaller subsets as the tree becomes formulated. The root node is a-top-most decision node in a Decision Tree, the internal nodes

are tree nodes or parent nodes that split into one or more child nodes, and the leaf nodes are the bottom nodes that do not split further. In Figure 5.1.1, the root node asks whether a data record satisfies the condition; if it is true, the next decision node will be on the left (A_2 condition), and it runs until reaching the bottom leaf node. If the target data is not satisfied with the condition, the next node goes to the right (False), and the target data runs until it reaches class B, and the data condition is defined as class B.

Accordingly, each class in Decision Tree classifiers will be evaluated with the Gini index in this research. Gini index calculates the probability of specific features, which is classified incorrectly when selected randomly. The Gini index measures the impurity of a data partition or set of training tuples.

$$Gini = \sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{i=1}^{n}(p_i)^2 \quad \text{where } p_i \text{ is } p_1, p_2, ..., p_i \tag{5.1}$$

p_i denotes the probability of an element being classified for every distinct class. $Gini$ will be zero if some $p_i = 1$ and it since each $p_i < 1$, it will be maximized if all p_i are equal. When the classification is correctly analyzed, the Gini index approaches zero. Thus, the lower the Gini index, the better fitting classification.

The Gini index considers a binary split for each attribute. Let's assume a case where A is a discrete value attribute having x distinct values, a_1, a_2, … , a_x. To determine the best binary split, all the possible subsets can be formed using known values of A. Each subset, S_A, can be considered a binary test for the attribute A of the form "$A \in SA$?". Given a tuple, this test is satisfied if the value of A for the tuple is among the values listed in SA. (Han, J. et al. (2012) Ch 8.7)

In this research, each possible split point is considered for continuous-valued attributes. The point giving the minimum Gini index for a continuous-valued attribute is taken as the split-point of the attribute. For a possible split point of A, D_1 is the set of tuples in D satisfying $A \leq$ *split point*, and D_2 is the set of tuples in D satisfying $A >$ *split point*. The attribute which has the minimum Gini index is selected as the splitting attribute. Decision Tree classifier represents decisions and helps decision-making visually and explicitly. By

researching Decision Tree learning, classifying nodes, help decide the final class, and conclude the class.

5.2 Naïve Bayes

Naïve Bayes is a classifier in the probabilistic machine learning model that's used for classification. Naïve Bayes assumes a bottom-line probability model that found the probability that an instance belongs in multiple classes, rather than a straightforward classification from a class. This algorithm assumes all related attributes are analytically independent and equivalently crucial in the probability.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A)\prod_{\mathrm{i}=1}^{n} P(B_{\mathrm{i}}|A)}{P(B)} \tag{5.2}$$

These expressions are:

- P(A|B) is a posterior probability of target class given predictor attributes
- P(A) is a prior probability of target class
- P(B|A) is a likelihood which is the probability of predictor attribute
- P(B) is a prior probability of predictor attribute

Equation 5.2 is a simple Naïve Bayes formula. The algorithm calculates the prior probabilities first at each class attribute and computes the posterior probabilities' values. Naïve Bayes approach works adequately when all the features/predictor attributes and the dependent attribute are categorical. The second reason for selecting Naïve Bayes is that training test data set is very quick because a single pass of the data is required to take account of discrete variables' frequencies or calculate a normal probability for continuous variables.

In this research, the first task is the bike demand prediction, dependent on daily maximum precipitation and daily average temperature uses multi nominal Naïve Bayes model. The multi nominal Naïve Bayes distributes data multi-nominally, and it conducts Naïve Bayes variants used in each separated classification, such as an application used for word count.

$$\hat{\theta}_{y_{\mathrm{i}}} = \frac{N_{y_{\mathrm{i}}} + \alpha}{N_y + \alpha n} = \frac{\sum_{x \in T} x_{\mathrm{i}} + \alpha}{\sum_{\mathrm{i}=1}^{n} N_{y_{\mathrm{i}}} + \alpha n} \quad \text{where } \theta_y \text{ is } \theta_{y_1}, \theta_{y_2}, ..., \theta_{y_{\mathrm{i}}} \tag{5.3}$$

The $\theta_{y_{\mathrm{i}}}$ is the probability $\mathrm{P}(x_{\mathrm{i}}|\mathrm{y})$ of feature I appearing in a sample belonging to class Y. The $N_{y_{\mathrm{i}}}$ is the number of times feature I appears in a sample of class y in the training set T, and N_y is the total count of all features for class Y.

In addition, the second task which is the bike demand prediction dependent on daily air condition uses Gaussian Naïve Bayes model.

$$P(x_{\mathrm{i}}|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_{\mathrm{i}} - \mu_y)^2}{2\sigma_y^2}) \tag{5.4}$$

Gaussian Naïve Bayes is also known as a normal distribution that supports continuous-valued features and models (Equation 5.4). An approach to creating a simple Naïve Bayes model assumes that a Gaussian distribution describes the data with no co-variance (independent dimensions) between dimensions. This model can be fit by merely finding the mean and standard deviation of the points within each label, which is needed to define such a distribution for measuring daily air pollutants.

5.3 Random Forest

Random Forest is ensemble with multiple Decision Trees. The individual Decision Tree in Random Forest splits out a class prediction, and the class with the most votes becomes the final model's prediction (Han, J. et al. (2012)). More specifically, each Decision Tree depends on a random vector's values sampled independently and with the same distribution for all trees in the forest. While the tree is classified, each tree votes, and the most popular class is returned and selected as the final decision. Random Forest is built by bagging, which stands for bootstrap aggregation. Each training set is a bootstrap sample [Bagging Method]. Because of sampling with replacement, some of X's original tuples may not be included in Xi, whereas others may occur more than once. The tuples' frequency is generally known as voting; thus, when the particular tuple has the most considerable frequency, the bagging

assigns the class with the most votes to X.

Algorithm 1 Bagging Method

1: **procedure** BAGGING(X)
2: **for** i $= 1, 2, 3, ..., k$ **do** ▷ Create k models:
3: create bootstrap sample, X, by sampling D with replacement
4: use X_{i} and the learning scheme to derive a model, M_{i}

Given a set, X, of X tuples, bagging works as follows. For iteration i (i $= 1, 2, ..., k$), a training set, X_{i}, of x tuples is sampled with replacement from the original set of tuples, X.

To construct a Decision Tree classifier, M_{i} randomly selects each node's number of attributes as candidates for the split at the node. The CART (Classification And Regression Trees, Machine Learning Terminology) methodology is used to grow the trees. The trees are grown to maximum size and are not pruned. Random Forest formed this way, with random input selection, are called Forest-RI (Han, J. et al. (2012)). The bagged classifier has relatively greater accuracy than a single classifier derived from the original training data, such as Decision Tree. The Random Forest is more robust to the effects of noisy data and over-fitting because Random Forest is insensitive to the number of attributes selected for consideration at each split. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers. From the bagging method that Random Forest adopted, the accuracy of Random Forest depends on the individual classifiers' strength and a measure of the dependence between individual classifiers. The ideal scenario is to maintain the strength of individual classifiers without increasing their correlation.

5.4 Evaluation Metrics

For evaluating the prediction performance of three different classification algorithms, Decision Tree, Naïve Bayes, and Random Forest, this research uses each model's accuracy, overall weighted average precision, recall, and measure, and each class's sensitivity and specificity.

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Figure 5.2. General Confusion Matrix Example, Han, J. et al. (2012) Figure 8.14

Figure 5.2 shows a general confusion matrix, which is 2x2. The accuracy, sensitivity, specificity, precision, recall, and F measure are found from this confusion matrix. Firstly, the classifier's accuracy on a given test set is the percentage of test set tuples correctly classified by the classifier. The accuracy computes by (TP + TN) / (P + N). Secondly, sensitivity measures the proportion of positives that are correctly identified, which is true positive, and specificity measures the proportion of correctly identified negatives, which is true negative. The sensitivity computes by TP / P, and specificity is by TN/ N from the general confusion matrix. Thirdly, precision can be thought of as a measure of exactness (i.e. What proportion of positive identifications was actually correct). In contrast, recall is a completeness measure (i.e. What proportion of actual positives was identified correctly). The precision calculates by TP/ (TP+FP), and recall calculates by TP/ (TP + FN). Lastly, the F-1 score (F measure) is the harmonic mean of precision and recall, and it gives equal weight to precision and recall. F measure is also found by (2 * precision * recall) / (precision+ recall) from these computations.

6. BIKE DEMAND PREDICTION BY WEATHER CONDITION

This section aims to reveal a relationship between the weather, daily weather temperature, precipitation, and daily bike demand. The record of climatological observations (NOAA data) and 2019 public bike trip data from NYC and Chicago are analyzed.

6.1 Data Preparation

We select trip records from the public bike trip data during the morning rush hour from 7 am to 9 am based on what we have learned in data exploration. Table 6.1 shows that the total number of trip count between 7 am and 9 am per each date is accumulated and displays the total count with the trip date.

Table 6.1. Pre-processing Bike Trip Records Data

#	**Trip Count**	**Trip Date**
1	5031	2019-01-01
2	4400	2019-01-02
3	7504	2019-01-03

The second data set we used in this task is the NOAA 2019 weather data from the National Oceanic and Atmospheric Administration (NOAA) [noaa.gov]. The NOAA provides access to a Global Historical Climatology Network (GHCN) in a daily weather record database, which offers historical daily temperature, precipitation, and snow records over global land areas. Among the NOAA's extensive database records, we focus on 2019 daily temperature and precipitation in NYC and Chicago land area. Each month's data set includes over 40 meteorological elements, including temperature daily maximum/minimum, the temperature at observation time, precipitation, snowfall, snow depth, evaporation, wind movement, wind maximums, soil temperature, and cloudiness. However, we opt for the daily maximum and minimum temperature and daily precipitation (rain/snow) in NYC and Chicago land area. The weather data in both NYC and Chicago have the same format after the collected one-year data is tidied.

Table 6.2. NOAA Tidied Data set in NYC

#	Date	Month	Day	MaxTemp	MinTemp	Rain	Snow
1	1/1/2019	1	1	58	39	0.06	0
2	1/1/2019	1	2	40	35	0	2
3	1/1/2019	1	3	44	37	0	0

The weather data shown in Table 6.2 is processed to summarize the average daily temperature and sum of daily precipitation. In Table 6.3, the "AvgTemp" column represents the daily average temperature by dividing two from the sum of "MaxTemp" and "MinTemp" columns, and the "Total Precipitation" column means the sum of "Rain" and "Snow" columns.

Table 6.3. Pre-processed Weather Data from Table 6.2

#	Date	MaxTemp	MinTemp	Rain	Snow	AvgTemp	Total Precipitation
1	1/1/2019	58	39	0.06	0	48.5	0.06
2	1/1/2019	40	35	0	2	37.5	2
3	1/1/2019	44	37	0	0	40.5	0

The bike trip in Table 6.1 and weather data in Table 6.4 are merged by the "Trip Date" column. The "Trip Count" indicators are from Table 6.1, and the "AvgTemp" and "Total Precipitation" columns are from Table 6.4.

Table 6.4. Summarized Data

#	Trip Date	Trip Count	AvgTemp	Total Precipitation
1	1/1/2019	5031	48.5	0.06
2	1/1/2019	4400	37.5	2
3	1/1/2019	7504	40.5	0

To label daily trip observation data, the precipitation is divided into five different ranges to represent the categorical data sets. Based on each range, this precipitation rate attribute helps to establish how precipitation affects the daily bike demand. Along with the assumption between precipitation range and public bike system demand, five categorical values on the daily average temperature are prepared in Table 6.5.

Table 6.5. Precipitation Label

Precipitation Rate	Range (inches)
None	None
Little	0.01 <= Precipitation < 0.1
Weak	0.1 <= Precipitation < 0.5
Moderate	0.5 <= Precipitation < 1.0
Heavy	1.0 <= Precipitation

Table 6.6. Temperature Label

Average Temp	Range (F)
Cold	Avg temp < 35
Chilly	35 <= Avg temp < 50
Moderate	50 <= Avg temp < 65
Warm	65 <= Avg temp < 805
Hot	80 <= Avg temp

Initially, the entire daily ride is numeric data by counting the number of observations by date. Since daily bike trip records are a continuous value, it is not easy to find the trend of daily bike demand. To track the daily bike demand trend, we set the range of the number of public bike usage. The average demand for the entire ride data is divided into three labels to resolve the continuous values' labeling problem. Finding the average number of rides consists of taking the sum of trip observations and diving it by the number of total observations. Accordingly, the average number of rides for the 2019 NYC Citi bike-sharing system was 6802, and for the 2019 Chicago Divvy bike-sharing system, was 1435 per day. The number of rides column is separated into three categories using this average value, "More than Average," "In Average," and "Less than Average."

Table 6.7. Bike Demand Label in NYC

Demand Of Daily Ride	Range (counts)
More than Average	0 <= DemandOfDailyRide <= 3500
In Average	3501 <= DemandOfDailyRide <= 7500
Less than Average	7501 <= DemandOfDailyRide

Table 6.7 and Table 6.8 depict the demand for bike labels by each range. To be more specific, the range in NYC is divided into 0-3500, 3501-7500, and more than 7501, and

Table 6.8. Bike Demand Label in Chicago

Demand Of Daily Ride	Range (counts)
More than Average	0 <= DemandOfDailyRide <= 400
In Average	401 <= DemandOfDailyRide <= 1435
Less than Average	1436 <= DemandOfDailyRide

the range in Chicago is divided into 0-400, 401-1435, and more than 1436. These divisions reflect to prevent imbalance labeling problems when the test data is trained, and these ranges produce a more reliable output for each task. The final data set consists of "weatherlabel," "templabel," and "DemandOfDailyRide" columns as shown in Table 6.9. Both NYC and Chicago data sets have the same columns and the final data set is ordered by the bike trip data. Therefore Table 6.9 has a total of 365 observations for both in NYC and Chicago.

Table 6.9. Final Data Set for Task 1

#	weather label	templabel	DemandOfDailyRide
1	Chilly	Little	Less than Average
2	Chilly	Heavy	Less than Average
3	Chilly	Little	Less than Average

6.2 Prediction Model

In this section, the bike demand prediction model by temperature and precipitation will be shown.

6.2.1 NYC

Figure 6.1 displays a NYC Bike Demand Prediction by Weather Condition as a tree classification. The Decision Tree folds max depth=7 and test size=0.30 to produce a more accurate result.

The root node starts with: IF templabel <= 0.5 (daily avg temp < 35, "cold") THEN Bike Demand = "Less Than Average"
IF templabel <= 0.5 ("cold") AND weatherlabel <= 1.5 (0.01 <= daily total precipitation

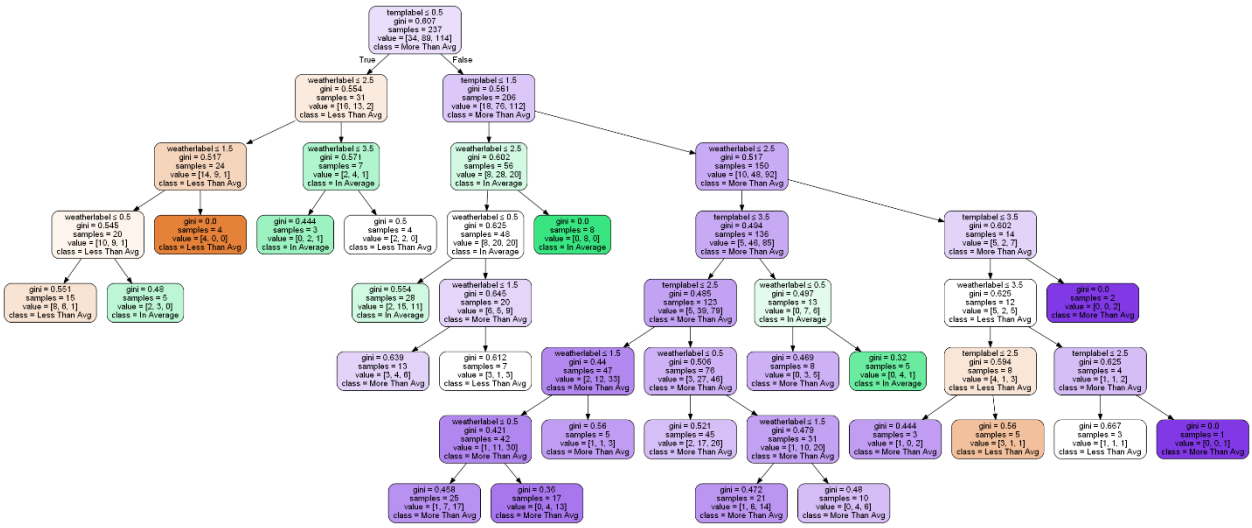


Figure 6.1. Decision Tree in NYC

< 0.1, "little") THEN Bike Demand = "Less Than Average" (1)

IF templabel <= 0.5 ("cold") AND weatherlabel <= 2.5 (0.1 <= daily total precipitation < 0.5, "weak") and THEN Bike Demand = "Less Than Average" (2)

IF templabel <= 0.5 ("cold") AND weatherlabel > 2.5 (daily total precipitation > 0.5, "moderate" or "heavy") and THEN Bike Demand = "In Average" (3)

IF templabel <= 0.5 ("cold") AND weatherlabel > 0.5 (daily total precipitation > 0.01, "little" or "precipitation existed") THEN Bike Demand = "In Average" (4)

Among the left-side child branch classes on the NYC Decision Tree, most of the results show "Less Than Average" when the daily temperature is cold, even though the maximum precipitation is little or none (1). However, the bike usage increases when the total precipitation is more than 0.5, while the bike usage decrease when the daily total precipitation is between 0.1 and 0.5, as shown in (2) and (3). Also, the overall bike usage is "In Average" when the precipitation exists, rather than the bike usage is "Less Than Average" when precipitation does not exist (4).

Among the right-side branch classes on the NYC Decision Tree, bike users tend to use the public bike "More Than Average" when the overall daily temperature is chilly (35 <

daily average temperature < 50). However, there are multiple interesting findings among the chilly weather condition.

IF templabel <= 1.5 ("chilly") AND weatherlabel <= 0.5 (daily total precipitation < 0.01, "none") THEN Bike Demand = "In Average" (5)
IF templabel <= 1.5 ("chilly") AND weatherlabel > 0.5 (daily total precipitation > 0.01, "little") THEN Bike Demand = "More Than Average" (6)
IF templabel <= 1.5 ("chilly") AND weatherlabel <= 1.5 (0.1 <= daily total precipitation < 0.5, "weak") THEN Bike Demand = "More Than Average" (7)

Among the first right-side child branch classes on the NYC Decision Tree, the daily bike demand tends to be as "In Average" when the daily average temperature is chilly, and the precipitation is none. Another interesting finding is that the bike demand is "More Than Average" when the precipitation exists, rather than when the precipitation does not exist. This interesting trend becomes much clearer in the next class (7) as shown in (5), (6) and (7). In another case, when the daily average temperature is between 65 and 80 ("warm"), bike users tend to use the public bike system "More Than Average." However, the bike demand decreases to "Less Than Average" when the precipitation is "moderate" or "heavy" (precipitation > 2.5).

IF templabel > 2.5 (daily avg temp > 65, "warm," or "hot") AND weatherlabel > 3.5 ("heavy") THEN Bike Demand = "More Than Average" (8)
IF templabel <= 3.5 ("warm") AND weatherlabel <= 3.5 (0.5 < daily total precipitation < 1.0, "moderate") THEN Bike Demand = "Less Than Average" (9)
IF templabel <= 3.5 ("warm") AND weatherlabel > 3.5 (daily total precipitation > 1.0, "heavy") THEN Bike Demand = "In Average" (10)
IF templabel > 3.5 (daily avg temp > 80, "hot") AND weatherlabel > 1.5 (daily total precipitation > 0.1, "weak", "moderate" or "heavy") THEN Bike Demand = "More Than Average" (11)
IF templabel > 3.5 (daily avg temp > 80, "hot") AND weatherlabel <=1.5 (0.01 <= daily

total precipitation < 0.1, “little”) THEN Bike Demand = “In Average” (12)

Among the second right-side child branch classes on the NYC Decision Tree, there are interesting findings in “warm” or “hot” weather conditions transition. Between (9) and (10), the bike demand increases to “In Average” from “Less Than Average” when the precipitation increases. Also, when the weather label is hot, the bike demand increases to “More Than Average” as the precipitation increase, as shown in (11) and (12). When the temperature is hot with moderate precipitation, the bike demand is “In Average”; while it is “Less Than Average” when the weather is warm with moderate precipitation. We usually assume that there would be more bike demand when the precipitation is moderate during a warm day, rather than during hot temperatures; however, some unexpected results show the opposite.

6.2.2 Chicago

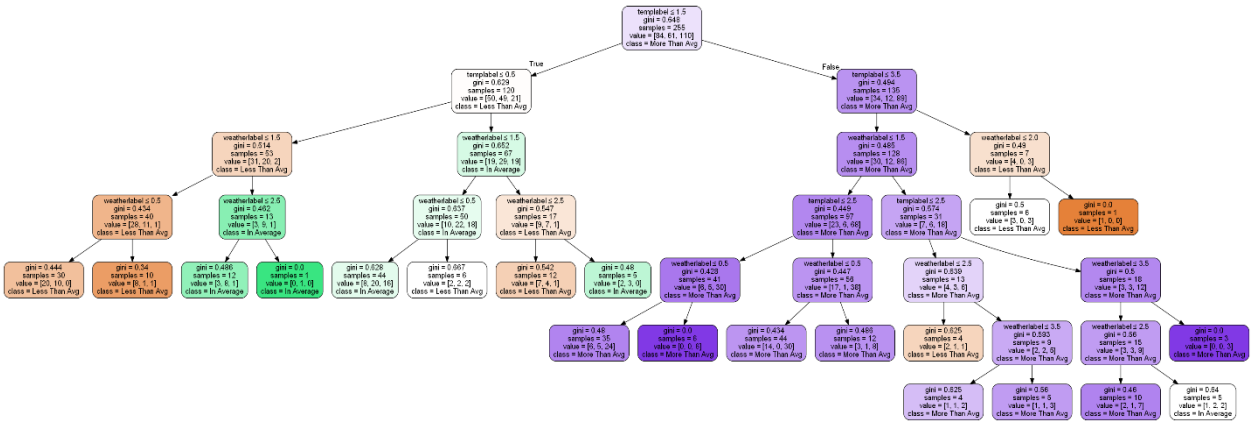


Figure 6.2. Decision Tree in Chicago

Figure 6.2 displays a Chicago Bike Demand Prediction by Weather Condition as a tree classification. The Decision Tree folds max depth = 7 and test size = 0.30 to produce a more accurate result and match the same environment as NYC.

The root node starts with: IF templabel <= 1.5 (35 <= daily avg temp < 50, “chilly”) THEN Bike Demand = “Less Than Average”

Among the first left-side child branch classes on the Chicago Decision Tree, the majority of results show “Less Than Average” or “In Average” as the daily temperature is cold.

IF templabel > 0.5 (daily avg temp > 35, “chilly”) AND weatherlabel <= 1.5 (0.01 <= daily total precipitation < 0.1, “little”) THEN Bike Demand = “In Average”
IF templabel > 0.5 (“chilly”) AND weatherlabel <= 0.5 (daily total precipitation < 0.01, “none”) THEN Bike Demand = “In Average”
IF templabel <= 0.5 (daily avg temp < 35, “cold”) AND weatherlabel > 1.5 (daily total precipitation > 0.1, “weak” or “moderate”) THEN Bike Demand = “In Average” (1)
IF templabel <= 0.5 (daily avg temp < 35, “cold”) AND weatherlabel <= 1.5 (0.01 <= daily total precipitation < 0.1, “little”) THEN Bike Demand = “Less Than Average” (2)
IF templabel > 0.5 (daily avg temp > 35, “chilly”) AND weatherlabel <= 2.5 (0.1 <= daily total precipitation < 0.5, “weak”) THEN Bike Demand = “Less Than Average” (3)
IF templabel > 0.5 (daily avg temp > 35) AND weatherlabel > 2.5 (daily total precipitation > 0.5, “moderate” or “heavy”) THEN Bike Demand = “In Average” (4)

We usually assume there is more bike demand when the precipitation is little or none during similar temperatures; however, based on (1) and (2), the bike usage increases to “In Average” when the precipitation is more than 0.1. Also, the bike demand increases when the weather precipitation has little, rather than zero, precipitation in the days referenced in (3) and (4).

IF templabel <= 3.5 (65 <= daily avg temp < 80, “warm”) AND weatherlabel <= 1.5 (0.01 <= daily total precipitation < 0.1, “little”) THEN Bike Demand = “More Than Average” (5)
IF templabel <= 2.5 (50 <= daily avg temp < 65, “moderate”) AND weatherlabel <= 2.5 (0.1 <= daily total precipitation < 0.5, “weak”) THEN Bike Demand = “Less Than Average” (6)
IF templabel <= 2.5 (“moderate”) AND weatherlabel > 2.5 (daily total precipitation > 0.5, “moderate” or “heavy”) THEN Bike Demand = “More Than Average” (7)

IF templabel > 3.5 (daily avg temp > 80, "hot") AND weatherlabel <= 2.0 (0.1 < daily total precipitation < 0.5, "weak") THEN Bike Demand = "Less Than Average" (8)

Among the second right-side child branch class in the Chicago Decision Tree, almost all class represents "More Than Average" when the daily average temperature is chilly, moderate, and warm (between 35 and 80), as shown in (5). When precipitation is heavier during a moderate temperature, the bike demand is "More Than Average" rather than "Less Than Average" when there is weak precipitation, as shown in (6) and (7). When the daily temperature is hot and the precipitation rate is weak, the bike demand decreases to "Less Than Average," as shown in (8).

6.3 Feature Evaluation

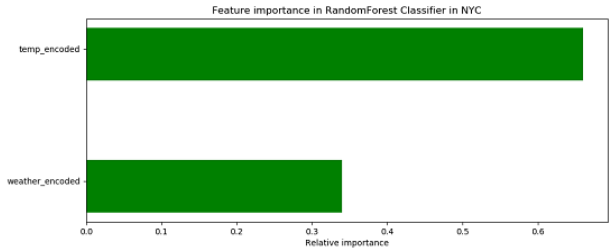


Figure 6.3. Random Forest Feature Importance Graph in NYC

Figure 6.3 shows that the "temp encoded" has more effect on Random Forest decision in this task. The "temp encoded" label represents the temperature labels by encoding each category in Table 6.6, and the importance rate is about 0.65 out of 1. The "weather encoded" label represents the weather labels by encoding each category in Table 6.5, and the importance rate of the "weather encoded" is 0.35 out of 1.

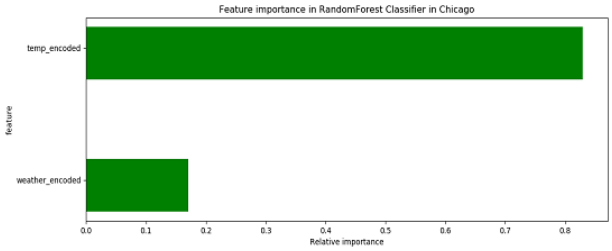


Figure 6.4. Random Forest Feature Importance Graph in Chicago

Based on Figure 6.4, the "temp encoded" has more effect on the Random Forest decision in this task. The "temp encoded" label represents the temperature labels by encoding each category in Table 6.6, and the importance rate is about 0.85 out of 1. The "weather encoded" label represents the weather labels by encoding each weather category in Table 6.5, and the importance rate of the "weather encoded" is 0.15 out of 1. Figure 6.3 and 6.4 show that the "weather encoded" feature is much more dominant in both cities.

6.4 Performance Comparison

This section compares the prediction performance of three different classification algorithms, Decision Tree, Naïve Bayes, and Random Forest, using six metrics, accuracy, sensitivity, specificity, weighted average precision, weighted average recall, and weighted average F-measure.

6.4.1 NYC

There are three performance comparison tables below. Table 6.10 displays the three classifiers' performance metrics with three labels, Table 6.11 represents sensitivity and specificity of each class (label), and Table 6.12 shows another performance comparison results sorting by two labels.

Table 6.10. Performance Comparison in NYC

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**
Decision Tree	53.64	0.53	0.54	0.53
Naïve Bayes	42.73	0.38	0.43	0.37
Random Forest	55.45	0.55	0.55	0.54

Table 6.11. Sensitivity and Specificity by class in NYC

Classifier	**Decision Tree**		**Naïve Bayes**		**Random Forest**	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
In Average	0.52	0.84	0.38	0.77	0.65	0.84
Less Than Average	0.38	0.74	0.19	0.69	0.37	0.74
More Than Average	0.62	0.69	0.49	0.59	0.62	0.72

Table 6.12. Performance Comparison with two labels in NYC

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Sensitivity**	**Specificity**
Decision Tree	62.73	0.63	0.63	0.63	0.59	0.66
Naïve Bayes	56.36	0.56	0.56	0.55	0.53	0.58
Random Forest	62.73	0.63	0.63	0.63	0.59	0.66

Table 6.10 shows that Random Forest classifier has the highest accuracy, highest weighted average precision, recall, and f-measure among the three classifiers. The highest precision,

which is referred to as completed labeling, is from Random Forest as 0.55. Moreover, the lowest recall, which presents incorrect labeling, is from Naïve Bayes as 0.43. Among classifiers in Table 6.11, the highest sensitivity in the "In Average" class is Random Forest, in the "Less Than Average" class is Decision Tree, and in the "More Than Average" class is both Decision Tree and Random Forest.

Overall, the "More Than Average" class has the highest sensitivity rate as 0.62, and the "Less Than Average" class has the lowest sensitivity rate as 0.38 in Decision Tree, 0.37 in Random Forest, and 0.19 in Naïve Bayes. Thus, we note that as the Random Forest classifier has high accuracy, the Random Forest's ability to label the positive class correctly is also remarkable given its' high sensitivity from three classes. In general, multi-class has high specificity, meaning that it can accurately recognize negative tuples. Especially in the "Less Than Average" class, although the class has the lowest sensitivity rate, the specificity is high enough.

Table 6.12 represents the overall performance comparison when only two labels exist. The labels, which consists of "More Than Average" and "Less Than Average," lead the models to higher accuracy comparatively, and higher precision and recall results. However, one of the shortages in the two class labeling, "More Than Average" and "Less Than Average," is the tree classification does not depict the range difference as much as three classes. We can find out that the three class labels model's trend of bike demand is easier to trace, compared to the two labels' bike prediction model.

6.4.2 Chicago

There are three performance comparison tables below, similar to NYC. Table 6.13 displays the three classifiers' performance metrics with three labels, Table 6.14 represents sensitivity and specificity of each class (label), and Table 6.15 shows another performance comparison results sorted by two labels.

As shown in Table 6.13, Random Forest classifier has dominantly the highest accuracy, highest weighted average precision, recall, and f-measure among three classifiers. The highest weighted precision is from the Random Forest as 0.56, and the lowest recall is from Naïve Bayes as 0.44.

Table 6.13. Performance Comparison in Chicago

Classifier	Accuracy	Precision	Recall	F-Measure
Decision Tree	47.27	0.53	0.47	0.49
Naïve Bayes	43.64	0.36	0.44	0.33
Random Forest	54.54	0.56	0.55	0.54

Table 6.14. Sensitivity and Specificity by class in Chicago

Title	Decision Tree		Naïve Bayes		Random Forest	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
In Average	0.56	0.77	0.25	0.66	0.63	0.76
Less Than Average	0.24	0.82	0.31	0.78	0.40	0.83
More Than Average	0.65	0.63	0.47	0.75	0.58	0.72

Table 6.15. Performance Comparison with two labels in Chicago

Classifier	Accuracy	Precision	Recall	F-Measure	Sensitivity	Specificity
Decision Tree	63.64	0.64	0.64	0.62	0.63	0.65
Naïve Bayes	66.36	0.66	0.66	0.65	0.67	0.66
Random Forest	65.45	0.65	0.65	0.65	0.70	0.60

Among the classifiers in Table 6.14, the highest sensitivity in the "In Average" class is Random Forest, in the "Less Than Average" class is Random Forest, and in the "More Than Average" class is in Decision Tree. Overall, the "More Than Average" class has the highest sensitivity rate of 0.65, and the "Less Than Average" class has the lowest sensitivity measure of 0.24 in Decision Tree, 0.31 in Naïve Bayes, and 0.4 in Random Forest. Thus, we note that as the Random Forest classifier has high accuracy, the Random Forest's ability to label the positive (rare) class correctly is also remarkable given its high sensitivity from three classes. In general, the multi-class has high specificity, meaning that it can accurately recognize negative tuples. Especially in the "Less Than Average" class, although the class has the lowest sensitivity rate, the specificity is the highest in the Decision Tree and Random Forest.

Table 6.15 represents the overall performance comparison when there are only two labels that exist in Chicago. The labels, which consists of "More Than Average" and "Less Than Average," lead the models to higher accuracy comparatively, higher precision and recall

results. However, one of the shortages in the two classes labeling, "More Than Average" and "Less Than Average," is the tree classification does not depict the bike demand range difference. We can find out that the three labels model's trend of bike demand is easier to trace, compared to the two labels' bike prediction model.

7. BIKE DEMAND PREDICTION BY AIR POLLUTION

This section aims to discover a relationship between daily air pollution rate and daily bike demand. The 2019 daily air quality index from the United States Environmental Protection Agency (EPA data) [epa.gov/outdoor-air-quality-data] and 2019 public bike trip data are analyzed.

7.1 Data Preparation

The first bike data we select in this task is Table 6.1 Bike Trip data, which has to pre-process for the previous task. The bike demand prediction by air pollution relates to the daily bike demand similar to the previous task. Thus the same bike trip recorded data is used, reference on the Table 6.7 and Table 6.8 Bike Demand Label.

The second data set we use in this work is the 2019 daily air quality index collected from the United States Environmental Protection Agency (EPA). EPA air pollution measurements are categorized in carbon monoxide (CO), nitrogen dioxide (NO_2), Particulate Matter Less than 2.5μm ($PM2.5$), Ozone (O_3), oxygen saturation (SO_2) for both NYC and Chicago. The EPA sets National Ambient Air Quality Standards (NAAQS) Table for pollutants considered harmful to public health and the environment. This research focuses on five pollutants with primary standards among a total of seven pollutants, excluding Lead (Pb) and $PM10$.

Table 7.1. National Ambient Air Quality Standards

Pollutant	**Primary/Secondary**	**Averaging Time**	**Level**
Carbon Monoxide (CO)	Primary	8 hours	9 ppm
		1 hour	35 ppm
Nitrogen Dioxide (NO_2)	Primary	1 hour	100 ppb
	Primary and Secondary	1 year	53 ppb
Ozone (O_3)	Primary and Secondary	8 hours	0.070 ppm
Particle Pollution ($PM2.5$)	Primary and Secondary	24 hours	35 $\mu g/m^3$
Sulfur Dioxide (SO_2)	Primary	1 hour	75 ppb

The primary standards provide public health protection, including the health of sensitive populations such as asthmatics, children, and the elderly. Oppositely, the secondary standards provide welfare protection, including protection against decreased visibility and

damage to animals, crops, vegetation, and buildings. Each air pollutant should not exceed the maximum level in the averaging time to be harmful to the primary targeting group.

Table 7.2. Daily Air Quality Data in Chicago

Date	CO		NO_2		$Ozone$		$PM2.5$		SO_2	
	Max	**AQI**	**Max**	**AQI**	**Max**	**AQI**	**Mean**	**AQI**	**Max**	**AQI**
1/1/2019	0.4	5	19.3	18	0.029	27	6.6	28	0	0
1/2/2019	0.4	5	31.8	29	0.02	19	9.1	38	0.8	0
1/3/2019	0.2	2	23.7	22	0.016	15	8.9	37	1.6	1

Among those indicators in Table 7.2, each air pollutant's AQI is excluded because the AQI is stabilized, and the task needs the same continuous variable to produce a comprehensive variety of cases. Therefore, DailyMaxCO, DailyMaxNO2, DailyMaxOzone, DailyMeanPM2.5, DailyMaxSO2 columns are selected as features in this task.

Table 7.3. Final Data Set for Task 2

Date	**MaxCO**	**MaxNO2**	**MaxOzone**	**MeanPM2.5**	**MaxSO2**	**Bike Demand**
1/1/2019	0.4	19.3	0.029	6.6	0	Less than Average
1/2/2019	0.4	31.8	0.02	9.1	0.8	Less than Average
1/3/2019	0.2	23.7	0.016	8.9	1.6	In Average

As shown in Table 7.3, five air pollutants' measurements become features, and the bike demand becomes a target class in this task.

7.2 Prediction Model

This section analyzes the relationship between the daily bike demand and the daily air pollutants' measurement with the Decision Tree, Naïve Bayes, and Random Forest classifiers.

7.2.1 NYC

Figure 7.1 displays an overall bike demand prediction by daily air quality as the tree classification in NYC. The Decision Tree folds max depth=5, test size=0.3, and random state=42 to produce an accurate result.

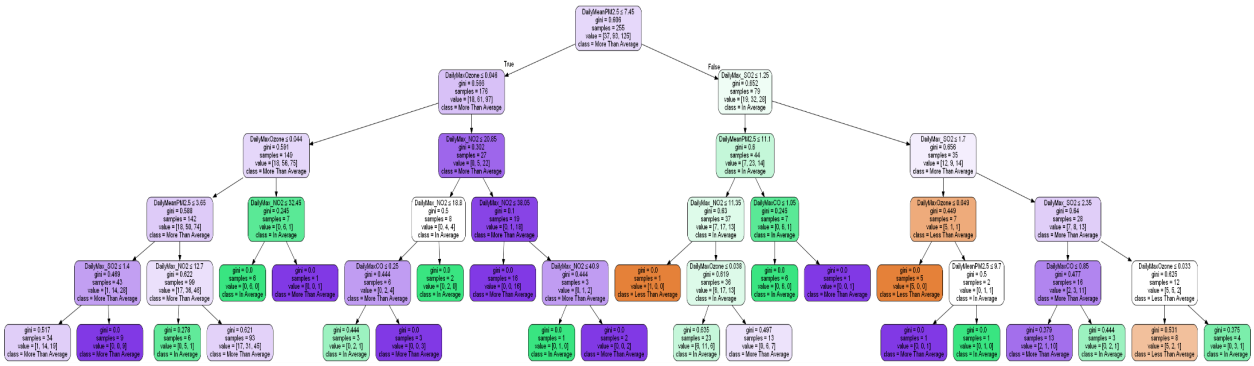


Figure 7.1. Decision Tree in NYC

Bike demand under the condition of DailyMeanPM2.5 <= 7.45 $\mu g/m^3$ tends to be "More Than Average" due to a satisfactory PM 2.5 condition. However, there are multiple cases of "In Average" bike demand affecting different air pollutants. The root node starts with: IF DailyMeanPM2.5 <= 7.45 THEN Bike Demand = "More Than Average"

IF DailyMaxOzone > 0.046 AND DailyMaxNO2 <= 18.8 THEN Bike Demand = "More Than Average" (1)

IF DailyMaxOzone > 0.046 AND DailyMaxNO2 > 18.8 THEN Bike Demand = "In Average" (2)

IF DailyMaxNO2 <= 18.8 AND DailyMaxCO <= 0.25 THEN Bike Demand = "In Average" (3)

IF DailyMaxNO2 <= 18.8 AND DailyMaxCO > 0.25 THEN Bike Demand = "More Than Average" (4)

IF DailyMaxOzone > 0.046 AND DailyMaxNO2 <= 40.9 THEN Bike Demand = "In Average" (5)

IF DailyMaxOzone > 0.046 AND DailyMaxNO2 > 40.9 THEN Bike Demand = "More Than Average" (6)

Based on Table 7.1 NAAQS Table, Ozone should not exceed more than 0.07 ppm, and NO2 should not exceed 100 ppb per day. Under the moderate Ozone condition based on (1) and (2), the daily bike demand increases when the DailyMaxNO2 rate is less than 18.8

ppb. DailyMaxNO2 is less than 18.8 ppb, which is still a satisfactory NO2 condition; the public bike demand increases to "More Than Average" by following DailyMaxCO, which is greater than 0.25 ppm. Based on Table 7.1, DailyMaxCO should not exceed over 9 ppm; otherwise, it will negatively affect people. Thus, it is interesting finding that bike demand increases when DailyMaxCO is higher. The daily bike demand increases when DailyMaxNO2 is greater than 40.9 ppb compared to the condition when DailyMaxNO2 is less than 40.9 ppb. Although the weather condition is not excellent, under moderate Ozone and moderate daily NO2 conditions, the daily bike demand increases to "More Than Average." Oppositely, the bike demand decreases to "In Average" when the NO2 condition is better based on (5) and (6).

IF DailyMeanPM2.5 > 3.65 AND DailyMaxNO2 <= 12.7 THEN Bike Demand = "In Average" (7)

IF DailyMeanPM2.5 > 3.65 AND DailyMaxNO2 > 12.7 THEN Bike Demand = "More Than Average" (8)

IF DailyMaxOzone > 0.044 AND DailyMaxNO2 <= 32.45 THEN Bike Demand = "In Average" (9)

IF DailyMaxOzone > 0.044 AND DailyMaxNO2 > 32.45 THEN Bike Demand = "More Than Average" (10)

To compare (7) with (8), the bike demand increases to "More Than Average" when DailyMaxNO2 is greater than 12.7 ppb, oppositely the bike demand is "In Average" when the NO2 condition is less than 12.7 ppb. Besides, under DailyMaxOzone > 0.044 ppm conditions, which is satisfactory but not good Ozone condition, the daily bike demand is varied based on NO2 measure. When DailyMaxNO2 is less than 32.45 ppb, the bike demand is "In Average." On the other hand, the bike demand is "More Than Average," when DailyMaxNO2 is greater than 32.45 ppb based on (9) and (10). The annual mean of NO2 is 53 ppb, which could negatively impact the elder and young kids.

On the first right-side child branch in the NYC Decision Tree, most of the bike demand is "In Average" under the condition: IF DailyMeanPM2.5 > 7.45 AND DailyMaxSO2 <= 1.25 THEN Bike Demand = "In Average"

One of the interesting findings in the first right-side child branch classes is:
IF DailyMaxNO2 > 11.35 AND DailyMaxOzone <= 0.038 THEN Bike Demand = "In Average" (11)
IF DailyMaxNO2 > 11.35 AND DailyMaxOzone > 0.038 THEN Bike Demand = "More Than Average" (12)
IF DailyMeanPM2.5 > 11.1 AND DailyMaxCO <= 1.05 THEN Bike Demand = "In Average" (13)
IF DailyMeanPM2.5 > 11.1 AND DailyMaxCO > 1.05 THEN Bike Demand = "More Than Average" (14)

The bike demand is likely to increase when the air pollutants' measurements are not high, and the bike demand decreases when the air pollutants increase. However, the bike demand increases to "More Than Average" when DailyMaxOzone is greater than 0.038 ppm. Ozone should not exceed 0.070 ppm in an eight hours period based on Table 7.1, and 0.038 ppm is comparably moderate but not excellent condition. Also, under DailyMeanPM2.5 > 11.1 $\mu g/m^3$ conditions, the bike demand increases when DailyMaxCO is greater than 1.05 ppm. Inversely, the bike demand is "In Average" when DailyMaxCO is less than 1.05 ppm under the same DailyMeanPM2.5 condition. From these results, even though the first right-side child branch has mostly moderate air conditions, unexpected results occur.

On the second right-side child branch classes in the NYC Decision Tree, three different bike demand classes experience different circumstances, "Less Than Average," "In Average," and "More Than Average."
IF DailyMaxSO2 <= 1.7 AND DailyMaxOzone <= 0.049 THEN Bike Demand = "Less Than Average" (15)
IF DailyMaxSO2 <= 1.7 AND DailyMaxOzone > 0.049 THEN Bike Demand = "In Aver-

age” (16)

IF DailyMaxSO2 <= 2.35 AND DailyMaxCO <= 0.85 THEN Bike Demand = “More Than Average” (17)

IF DailyMaxSO2 <= 2.35 AND DailyMaxCO > 0.85 THEN Bike Demand = “In Average” (18)

IF DailyMaxSO2 > 2.35 AND DailyMaxOzone <= 0.033 THEN Bike Demand = “Less Than Average” (19)

IF DailyMaxSO2 > 2.35 AND DailyMaxOzone > 0.033 THEN Bike Demand = “In Average” (20)

To compare (15) and (16), the daily bike demand increases when DailyMaxOzone is greater than a satisfactory threshold, 0.049 ppm, under DailyMaxSO2 <= 1.7 ppb condition. DailyMaxOzone 0.049 ppm is close to 0.07 ppm, which negatively affects the elder and young children. The bike demand decreases to “In Average” from “More Than Average,” as the DailyMaxCO increases based on (17) and (18). The bike demand oppositely increases to “In Average” again when DailyMaxOzone is greater than 0.033 ppm under DailyMaxSO2 > 2.35 ppb condition.

7.2.2 Chicago

Figure 7.2 displays an overall bike demand prediction by daily air quality as the tree classification in Chicago. The Decision Tree folds max depth=5, test size=0.3, and random state=42 to produce a more accurate result and match NYC’s same environment.

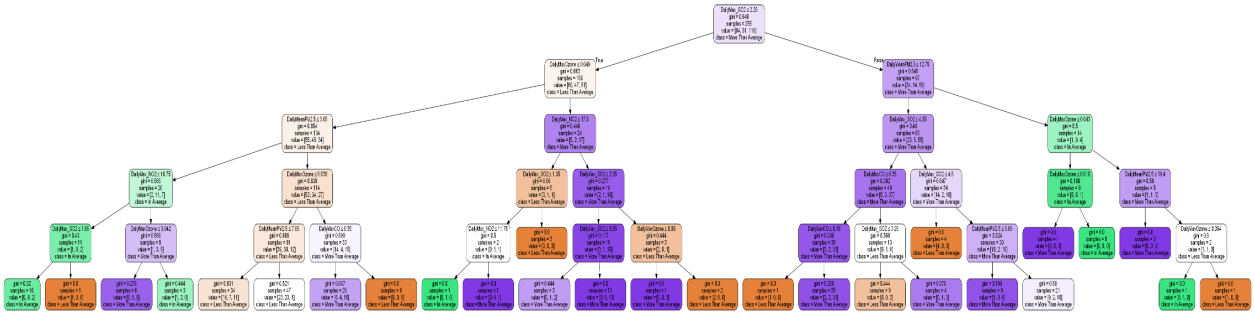


Figure 7.2. Decision Tree in Chicago

The bike demand under the conditions of DailyMaxSO2 is less than 2.25 ppb, and DailyMaxOzone is greater than 0.049 ppm tend to be "More Than Average." Ozone should not exceed a maximum of 0.07 ppm in an eight hour period, based on Table 7.1. Due to the closeness with the threshold, the measurement of 0.049 ppm in Ozone depicts a poor air condition.

IF DailyMaxOzone > 0.049 AND DailyMaxNO2 <= 17.8 THEN Bike Demand = "Less Than Average" (1)
IF DailyMaxOzone > 0.049 AND DailyMaxNO2 > 17.8 THEN Bike Demand = "More Than Average" (2)
IF DailyMaxSO2 < 1.35 AND DailyMaxNO2 <= 11.75 THEN Bike Demand = "In Average" (3)
IF DailyMaxSO2 < 1.35 AND DailyMaxNO2 > 11.75 THEN Bike Demand = "More Than Average" (4)
IF DailyMaxSO2 > 2.05 AND DailyMaxOzone <= 0.06 THEN Bike Demand = "More Than Average" (5)
IF DailyMaxSO2 > 2.05 AND DailyMaxOzone > 0.06 THEN Bike Demand = "Less Than Average" (6)

Based on (1) and (2), the bike demand increases to "More Than Average" when DailyMaxNO2 is greater than 17.8 ppb, while bike demand decreases to "Less Than Average" when DailyMaxNO2 is less than 17.8 ppb. Both conditions are satisfactory since NO2 has not reached the threshold, 100 ppb, but Ozone is close to the daily threshold. However, the bike demand inversely shows an increment when the air condition is negatively affected by the increase in NO2. Similarly, under DailyMaxSO2 < 1.35 ppb condition, which is a healthy air condition. The bike demand increases when DailyMaxNO2 is greater than 11.75 ppb. Oppositely, the bike demand decreases when DailyMaxNO2 is less than 11.75 ppb, as shown in (3) and (4). However, a potentially unhealthy Ozone condition is detected on (5) and (6). Based on Table 7.1, Ozone should not exceed more than 0.07 ppm in an eight hour period. The effect of Ozone measurement is found in (5) and (6). Accordingly, the bike demand decreases due to the harsh Ozone levels, which are greater than 0.06 ppm, and the

bike demand increases again when DailyMaxOzone is less than 0.06 ppm.

IF DailyMeanPM2.5 <= 3.85 AND DailyMaxNO2 <= 16.75 THEN Bike Demand = "In Average" (7)
IF DailyMeanPM2.5 <= 3.85 AND DailyMaxNO2 > 16.75 THEN Bike Demand = "More Than Average" (8)
IF DailyMaxNO2 > 16.75 AND DailyMaxOzone <= 0.042 THEN Bike Demand = "More Than Average" (9)
IF DailyMaxNO2 > 16.75 AND DailyMaxOzone > 0.042 THEN Bike Demand = "In Average" (10)
IF DailyMeanPM2.5 > 3.85 AND DailyMaxOzone <= 0.038 THEN Bike Demand = "Less Than Average" (11)
IF DailyMeanPM2.5 > 3.85 AND DailyMaxOzone > 0.038 THEN Bike Demand = "More Than Average" (12)

An interesting finding on the most left-side child branch in the Chicago Decision Tree (7) and (8) is the daily bike demand increases to "More Than Average" when DailyMaxNO2 is greater than 16.75 ppb, under DailyMeanPM2.5 <= 3.85 $\mu g/m^3$ condition. On the other hand, the bike demand decreases to "In Average" when DailyMaxNO2 is less than 16.75 ppb. To not affect negatively, NO2 should not exceed more than 100 ppm. The bike demand increases to "More Than Average", even though NO2 increases, which is still under the ideal NO2 condition. To compare the pair (9, 10) and (11, 12), there is one interesting finding depending on the O_3, $PM2.5$, AND NO_2. The bike demand naturally decreases to "In Average" from "More Than Average" when DailyMaxOzone is greater than 0.042 ppm under DailyMaxNO2 > 16.75 ppb condition. As mentioned earlier, Ozone measurement 0.042 ppm is close to the healthy Ozone threshold due to it being close to 0.07 ppm. On the other hand, the bike demand increases to "More Than Average" when Ozone is greater than 0.038 ppm under DailyMeanPM2.5 > 3.85 $\mu g/m^3$ condition. Since PM 2.5 does not exceed 35 $\mu g/m^3$, the air condition is still healthy except for Ozone condition.

On Chicago's right-side child branch classes in the Decision Tree, the trend of bike demand is "More Than Average" under DailyMeanPM2.5 <= 12.75, which is not even close to 35 $\mu g/m^3$ per day.

IF DailyMeanPM2.5 <= 12.75 AND DailyMaxSO2 <= 4.5 THEN Bike Demand = "Less Than Average" (13)
IF DailyMeanPM2.5 <= 12.75 AND DailyMaxSO2 > 4.5 THEN Bike Demand = "More Than Average" (14)

There is one interesting finding under DailyMeanPM2.5 <= 12.75 $\mu g/m^3$ condition. As shown in (13) and (14), the bike demand is "More Than Average" when DailyMaxSO2 is greater than 4.5 ppb, which is not a significant enough value to affect the air condition negatively.

IF DailyMeanPM2.5 > 12.75 AND DailyMaxOzone <= 0.043 THEN Bike Demand = "In Average" (15)
IF DailyMeanPM2.5 > 12.75 AND DailyMaxOzone > 0.043 THEN Bike Demand = "More Than Average" (16)
IF DailyMaxOzone > 0.043 AND DailyMeanPM2.5 <= 19.4 THEN Bike Demand = "More Than Average" (17)
IF DailyMaxOzone > 0.043 AND DailyMeanPM2.5 > 19.4 THEN Bike Demand = "Less Than Average" (18)

Most of the bike demand trends under DailyMeanPM2.5 > 12.75 $\mu g/m^3$ condition is comprehensible. To be specific, in (17) and (18), the bike demand decreases to "In Average" or "Less Than Average" from "More Than Average" based on an increase in Ozone and PM2.5 measurement. Among those daily bike trends, there is one unusual case found. The bike demand increases when DailyMaxOzone is greater than 0.043 ppm under DailyMeanPM2.5 > 12.75 $\mu g/m^3$ condition based on (15) and (16). Based on Table 7.1, Ozone should not exceed more than 0.07 ppm, and PM2.5 should not exceed 35 $\mu g/m^3$ per day. Even though

there are high Ozone levels and PM2.5 measurements in (15) and (16), the bike demand represents the opposite, which is considered an interesting finding.

7.3 Feature Evaluation

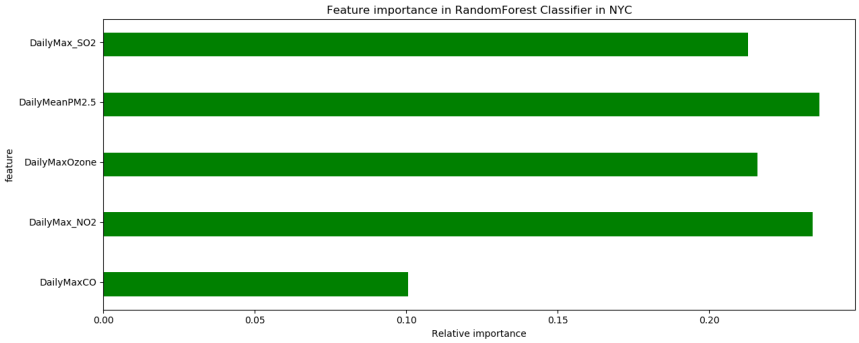


Figure 7.3. Random Forest Feature Importance Graph in NYC

Based on Figure 7.3, among the five features, DailyMaxNO2 has the most effect on the daily bike demand, and DailyMeanPM2.5 and DailyMaxOzone have a comparably high impact on the bike demand. DailyMaxNO2's importance rate is about 0.24 out of 1, DailyMeanPM2.5 is about 0.24, and DailyMaxOzone is about 0.22 out of 1. The least affecting feature of the bike demand is DailyMaxCO, and the importance rate is about 0.1 out of 1.

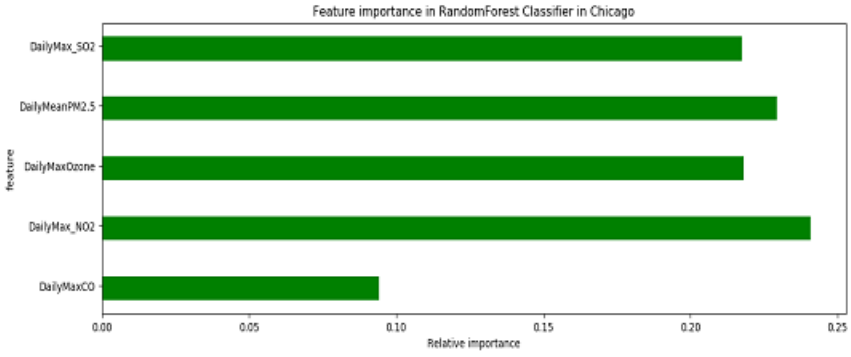


Figure 7.4. Random Forest Feature Importance Graph in Chicago

Based on Figure 7.4, DailyMaxNO2 has the most effect on the daily bike demand among the five features. The importance rate in DailyMaxNO2 is about 0.24 out of 1, Daily-

MeanPM2.5 is 0.23 out of 1, and DailyMaxOzone is about 0.22 out of 1. The least affecting feature is DailyMaxCO, and the importance rate is less than 0.1 out of 1. DailyMaxNO2 and DailyMeanPM2.5 are the most affecting air pollutants on the bike demand from both NYC and Chicago.

7.4 Performance Comparison

This section compares each classifiers' performance of three different algorithms, Decision Tree, Naïve Bayes, and Random Forest, using six metrics, accuracy, sensitivity, specificity, weighted average precision, weighted average recall, and weighted average F-measure.

7.4.1 NYC

There are three performance comparison results. Table 7.4 displays the three classifiers' performance metrics with three labels, Table 7.5 represents sensitivity and specificity of each class (label), and Table 7.6 shows another performance comparison results sorting by two labels.

Table 7.4. Performance Comparison in NYC

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**
Decision Tree	53.64	0.51	0.54	0.51
Naïve Bayes	45.45	0.42	0.45	0.42
Random Forest	46.36	0.45	0.46	0.45

Table 7.5. Sensitivity and Specificity by class in NYC

Title	**Decision Tree**		**Naïve Bayes**		**Random Forest**	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
In Average	0.56	0.82	0.29	0.77	0.33	0.78
Less Than Average	0.35	0.72	0.37	0.73	0.32	0.72
More Than Average	0.59	0.74	0.51	0.61	0.60	0.66

As shown in Table 7.4, Decision Tree classifier has the highest accuracy, highest weighted average recall, and f-measure among the three classifiers. Since a perfect precision and recall score are one, the highest precision, referred to as completed labeling, is Decision Tree

Table 7.6. Performance Comparison with two labels in NYC

Classifier	Accuracy	Precision	Recall	F-Measure	Sensitivity	Specificity
Decision Tree	50.00	0.51	0.50	0.50	0.46	0.55
Naïve Bayes	58.18	0.59	0.58	0.58	0.53	0.64
Random Forest	54.54	0.55	0.55	0.55	0.50	0.59

as 0.51. The lowest recall, which is referred to as incorrectly labeling, is Naïve Bayes as 0.45. Among the three classifiers in Table 7.5, the highest sensitivity in the "In Average" class is Decision Tree, in the "Less Than Average" class is Naïve Bayes, and in the "More Than Average" class Random Forest, but similar in Decision Tree. Thus, the "More Than Average" class has the highest sensitivity rate of 0.6 from Random Forest, and the "Less Than Average" class has the lowest sensitivity rate of around 0.3. Therefore, Decision Tree classifier has high accuracy, and Decision Tree's ability to correctly label the positive class is also remarkable, given its high sensitivity from three classes. In general, multi-class has a high specificity, meaning that it can accurately recognize negative tuples. Especially in the "Less Than Average" class, although the class has the lowest sensitivity rate, the specificity is the highest.

Table 7.6 shows the same model with only two labels: "More Than Average" and "Less Than Average" for the daily bike demand. Overall, the two class labels have higher accuracy and precision, recall, and F-Measure. An interesting finding is Decision Tree in three class labels, which consists of "In Average," "More Than Average," "Less Than Average," has higher accuracy than the two labels, "More Than Average" and "Less Than Average." Since the data set uses the normal distribution to predict the bike demand model, the three labels have more support for distributing the output. However, the two class labeling model represents higher accuracy in Naïve Bayes and Random Forest due to their limited amounts of sorting when training and testing the data set. Since Decision Tree output is also only sorted with two class, it is not easy to perceive where the class actually belongs to due to its' simplicity. Therefore, although three class labels have lower accuracy than the two class labels, the three labels' results contain more interesting results.

7.4.2 Chicago

There are three performance comparison tables in this section, similar to NYC. Table 7.7 displays the three classifiers' performance metrics with three labels, Table 7.8 represents sensitivity and specificity of each class (label), and Table 7.9 shows another performance comparison results sorting by two labels.

Table 7.7. Performance Comparison in Chicago

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**
Decision Tree	47.27	0.47	0.47	0.47
Naïve Bayes	45.45	0.45	0.45	0.44
Random Forest	39.09	0.40	0.39	0.39

Table 7.8. Sensitivity and Specificity by class in Chicago

Classifier	**Decision Tree**		**Naïve Bayes**		**Random Forest**	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
In Average	0.41	0.69	0.37	0.66	0.34	0.64
Less Than Average	0.28	0.82	0.33	0.82	0.27	0.82
More Than Average	0.61	0.67	0.56	0.66	0.50	0.59

Table 7.9. Performance Comparison with two labels in Chicago

Classifier	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Sensitivity**	**Specificity**
Decision Tree	57.27	0.57	0.57	0.56	0.60	0.53
Naïve Bayes	55.45	0.54	0.55	0.54	0.58	0.50
Random Forest	57.27	0.56	0.57	0.56	0.59	0.53

As shown in Table 7.7, Decision Tree classifier has the highest accuracy, highest weighted average precision, recall, and f-measure among the three classifiers. Since a perfect precision and recall score are one, the highest precision, which is referred to as completed labeling, is both Decision Tree as 0.47. The lowest recall, which is referred to as incorrectly labeling, is Random Forest as 0.39.

Among classifiers in Table 7.8, the highest sensitivity in the "In Average" class is Decision Tree, in the "Less Than Average" class is Naïve Bayes, and in the "More Than Average" class is in Decision Tree. Thus, the "More Than Average" class has the highest sensitivity

rate as approximately 0.61 from Decision Tree. The "Less Than Average" class has the lowest sensitivity rate of 0.33 from Naïve Bayes. Therefore, Decision Tree classifier has high accuracy, and Decision Tree's ability to correctly label the positive class is also remarkable, given its' high sensitivity from those three classes. In general, multi-class has high specificity, which means it can accurately recognize negative tuples. Especially in the "Less Than Average" class, although the class has the lowest sensitivity rate, the specificity is high enough.

Table 7.9 shows the same model with only two labelings: "More Than Average" and "Less Than Average" in Chicago. The Chicago's comparison results in two labelings, "More Than Average" and "Less Than Average," also has a higher accuracy than three labeling, "Less Than Average," "In Average," and "More Than Average." The accuracy increases in two labels due to its' simplified training and test data set. However, the sensitivity and specificity in "More Than Average" in two labeling and three labeling do not contain a significant difference. Even though two labeling is simpler when the test data is trained, two labeling is not easy to find an attractive feature in the classification model. Therefore, even though the three labels have lower accuracy than the two class labels, the three class labels' results contain more interesting features and worth discussing.

8. CONCLUSION

Generally, the public bike demand is lower when the daily temperature is low, precipitation exists, or air condition is not satisfactory. However, there are multiple interesting bike demand predictions which inversely increases bike demand under unsatisfactory air condition based on this research.

Decision Tree is one of the best models to visually track all the classification models to discover the unusual bike usage trend. For the first weather and precipitation task, the Random Forest bike demand prediction learning has the highest accuracy for both NYC and Chicago. Even though only two labels exist in the data, the results are the same as the three labeling models.

The second task, air pollution and bike demand, is slightly different from the first task. Since the air pollution data set contains continuous variables, the Decision Tree produces the highest accuracy in NYC and Chicago for the second task. Even though the accuracy and precision in the two labeling system are higher in many spots due to its simplicity in training and test data sets, the specificity, which can sort out non-class groups, is high enough in three class labeling. Therefore, we can conclude that the Decision Tree is the most reliable technique among the three classifiers.

For our model, there is still much work that is worth exploring in the future. Other external factors, such as geographical inference, can be added along with the weather, precipitation, and air pollution data to increase the overall accuracy rate. Secondly, instead of using the maximum or mean value of air pollutants, air pollutants can be replaced with the air quality index and standardize the air pollutants' features. Thus, overall accuracy can increase with more simplified features. The discovery of the relationship between air pollution and bike demand in NYC and Chicago is the first step forward from this research. Even though the result does not seem to produce a critical impact or result in the city, because of their moderate air condition, this result will benefit the city management in the future for possible air pollution condition management.

REFERENCE

[1] Li Y., and Zheng, Y. (2020). Citywide Bike Usage Prediction in a Bike-Sharing System *IEEE Transactions on Knowledge and Data Engineering,* 32(06), 1079-1091.

[2] Giot R., and Cherrier R. (2014). Predicting bikeshare system usage up to one day ahead. In *2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS)* 22-29.

[3] Hulot, P., Aloise, D., and Jena, S. (2018). Towards Station-Level Demand Prediction for Effective Rebalancing in Bike-Sharing Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 378–386.

[4] Li, Y., Zheng, Y., Zhang, H., and Chen, L. (2015). Traffic Prediction in a Bike-Sharing System. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems.* 33, 378–386.

[5] Singhvi, D., Singhvi,S., Frazier,P., Henderson, S., O'Mahony, E., Shmoys,D., and Woodard, D. B. (2015). Predicting Bike Usage for New York City's Bike Sharing System. In *AAAI Workshop: Computational Sustainability.* (pp. 1)

[6] Rixey, R. Alexander (2013). Station-Level Forecasting of Bikesharing Ridership: Station Network Effects in Three U.S. Systems. In *Transportation Research Record* 2387(1), 46–55.

[7] Beecham, R., and Wood, J. (2013). Exploring gendered cycling behaviours within a large-scale behavioural data-set *Transportation Planning and Technology* 37(1), 83-97.

[8] Zhao, J, and Wang, J. and Deng, W. (2015). Exploring bikesharing travel time and trip chain by gender and day of the week. In *Transportation Research Part C: Emerging Technologies* 58(b), 251-264.

[9] Ashqar, H. I., Elhenawy,M., Almannaa, M. H., Ghanem, A., Rakha, H. A., and House, L. (2017). Modeling bike availability in a bike-sharing system using machine learning.

In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* 374-378.

[10] Freund, D., Henderson, S. G., and Shmoys, D. B. (2017). Minimizing multimodular functions and allocating capacity in bike-sharing systems. In *Integer Programming and Combinatorial Optimization Proceedings (Lecture Notes in Computer Science)* 10328, 186–198.

[11] Lin, L., He, Z., Peeta, S. (2017) Predicting Station-level Hourly Demands in a Large-scale Bike-sharing Network: A Graph Convolutional Neural Network Approach. *arXiv preprint arXiv:1712.04997*

[12] Yoon, J. W., Pinelli, F., and Calabrese., F., (2012). Cityride: a predictive bike sharing journey advisor. *Mobile Data Management (MDM), 2012 IEEE 13th International Conference* 306–311.

[13] Zhou, Y. and Huang, Y. (2019) Place Representation Based Bike Demand Prediction. *2019 IEEE International Conference on Big Data (Big Data)* 1577-1586.

[14] Han, J., Kamber, M., and Pei, J. (2012). Data mining: Concepts and techniques, third edition (3rd ed.). Waltham, Mass.: Morgan Kaufmann Publishers.

[15] Froehlich, J., Neumann, J., and Oliver, N. (2009). Sensing and Predicting the Pulse of the City through Shared Bicycling. In *Proceedings of the 21st international jont conference on Artifical intelligence (IJCAI'09)* 1420–1426.

[16] Gallop, C., Tse, C., and Zhao, J. (2011). A seasonal autoregressive model of Vancouver bicycle traffic using weather variables. *i-Manager's Journal on Civil Engineering* 1, 9.

[17] Breiman, L. (2001). Random forests *Machine learning* 45, 5-32.

[18] Loh, W. Y. (2011) Classification and regression trees *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1,14-23.

[19] Ashqara, I., Elhenawyb, M., Rakhac, H. A. (2019) Modeling bike counts in a bike-sharing system considering the effect of weather conditions, Case Studies on Transport Policy. *Case Studies on Transport Policy* 7(2), 261-268.

A. ORIGINAL DATA SCHEMA

This section shows an original data.

A.1 NYC Citi Bike Schema

Citi Bike Company publish downloadable files of Citi Bike trip data. The data includes:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station ID
- Start Station Name
- Start Station Latitude/Longitude
- End Station ID
- End Station Name
- End Station Latitude/Longitude
- Bike ID
- User Type (Customer=24-hour/3-day pass user; Subscriber=Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- User's Birth Year

A.2 Chicago Divvy Bike Schema

Divvy Bike Company publish downloadable files of Divvy Bike trip data. The data includes:

- Trip ID

- Start Time
- End Time
- Bike ID
- Trip Duration
- From Station ID
- From Station Name
- To Station ID
- To Station Name
- User Type (Member, Single Ride, and Day Pass)
- Gender (Zero=unknown; 1=male; 2=female)
- User's Birth Year

A.3 The NOAA record of climatological observations Schema

U.S. Department of Commerce
National Oceanic & Atmospheric Administration
National Environmental Satellite, Data, and Information Service
Current Location: Elev: 140 ft. Lat: 40.7790° N Lon: -73.9693° W
Station: **NY CITY CENTRAL PARK, NY US USW00094728**

Record of Climatological Observations
These data are quality controlled and may not be identical to the original observations.
Generated on 09/09/2020

National Centers for Environmental Information
151 Patton Avenue
Asheville, North Carolina 28801

Observation Time Temperature: Unknown Observation Time Precipitation: 2400

Year	Month	Day	Temperature (F)			Precipitation					Evaporation		Soil Temperature (F)					
			24 Hrs. Ending at Observation Time		At Obs.	24 Hour Amounts Ending at Observation Time				At Obs. Time	24 Hour Wind Movement (mi)	Amount of Evap. (in)	4 in. Depth			8 in. Depth		
			Max.	Min.		Rain, Melted Snow, Etc. (in)	Flag	Snow, Ice Pellets, Hail (in)	Flag	Snow, Ice Pellets, Hail, Ice on Ground (in)			Ground Cover (see *)	Max.	Min.	Ground Cover (see *)	Max.	Min.
2019	01	01	58	39		0.06		0.0		0.0								
2019	01	02	40	35		0.00		0.0		0.0								
2019	01	03	44	37		0.00		0.0		0.0								
2019	01	04	47	35		0.00		0.0		0.0								
2019	01	05	47	41		0.50		0.0		0.0								
2019	01	06	49	31		T		0.0		0.0								
2019	01	07	34	25		0.00		0.0		0.0								
2019	01	08	45	34		0.17		T		0.0								
2019	01	09	45	34		0.06		0.0		0.0								
2019	01	10	34	28		0.00		0.0		0.0								
2019	01	11	30	21		0.00		0.0		0.0								
2019	01	12	34	20		0.00		0.0		0.0								
2019	01	13	33	25		T		T		0.0								
2019	01	14	32	22		0.00		0.0		0.0								
2019	01	15	36	25		0.00		0.0		0.0								
2019	01	16	39	30		0.00		0.0		0.0								
2019	01	17	33	24		0.00		0.0		0.0								
2019	01	18	39	29		0.05		0.5		1.2								
2019	01	19	37	32		0.29		0.2		0.0								
2019	01	20	42	14		0.88		0.0		0.0								
2019	01	21	14	4		0.00		0.0		0.0								
2019	01	22	31	13		0.00		0.0		0.0								
2019	01	23	52	31		T		0.0		0.0								
2019	01	24	59	35		1.33		0.0		0.0								
2019	01	25	40	28		0.00		0.0		0.0								
2019	01	26	35	24		0.00		0.0		0.0								
2019	01	27	49	32		0.00		0.0		0.0								
2019	01	28	38	25		0.00		0.0		0.0								
2019	01	29	43	25		0.23		T		0.0								
2019	01	30	35	6		0.01		0.4		0.0								
2019	01	31	16	2		0.00		0.0		0.0								
		Summary	39	26		3.58		1.1										

Empty, or blank, cells indicate that a data observation was not reported.
*Ground Cover: 1=Grass; 2=Fallow; 3=Bare Ground; 4=Brome grass; 5=Sod; 6=Straw mulch; 7=Grass muck; 8=Bare muck; 0=Unknown
"s" This data value failed one of NCDC's quality control tests. "At Obs." = Temperature at time of observation
"T" values in the Precipitation or Snow category above indicate a "trace" value was recorded.
"A" values in the Precipitation Flag or the Snow Flag column indicate a multiday total, accumulated since last measurement, is being used.
Data value inconsistency may be present due to rounding calculations during the conversion process from SI metric units to standard imperial units.

A.4 The EPA Outdoor Air Quality Data Schema

The United States Environmental Protection Agency(EPA) provide downloadable files of Outdoor Air Quality Data. Total six air pollutants are provided:

- Ozone (O3)
- Particulate matter (PM10 and PM2.5)
- Carbon monoxide (CO)
- Nitrogen dioxide (NO2)
- Sulfur dioxide (SO2)
- Lead (Pb)