Data Science and Big Data Analytics Student Guide

Hubway visualization Project

Name of student	ID
1. Manar Abdelkarim	434200161
2. Nada alhajaj	434200053
3. Ohoud Alotaibi	434011415

Section: 8I7

Table of content

Abstract:

Women's participation in cycling for transport and recreation is approximately half that of men. Research is required to investigate the individual, social and environmental determinants of women's participation in cycling for transport and recreation. Discussion: Few studies have systematically investigated women's perceptions and experiences of cycling and little is known about what motivates and sustains their involvement. Preliminary indications are that, for women, there may be an interest in and capacity to participate in cycling that is not being translated into practice. Safety concerns appear to be a significant deterrent to women cycling. Safety factors have a differential impact on women as they are generally more risk averse than men. Quantitative risk assessments suggest that the risk of injury associated with cycling is small and that the health benefits outweigh the health costs. Cycling promotion campaigns may achieve greater success with women if they enable women to experience cycling in an environment that both is, and is perceived to be, safe and supportive.

Introduction:

Hubway is a bicycle sharing system in the Boston area. Launched at the end of July in 2011, Hubway expanded to the neighbouring cities: Cambridge, and Somerville. In this research, in seek to choose trips that are purposed to ensure that the duration is in the minimum rate. We had chosen the data set, which the terminal stations at universities in two cities to prove that women spend time more than men in bike ride, because of the five Cs, which are:

1. Comfort: distracted driving, speed of cars and no separated lanes on streets are the main factors that makes the bike riding uncomfortable and unsafe for women.

2. Convenience: lack of time, inability to carry children and other passengers, and inability to carry more stuff are the reason that women are not feel convenience.

Women who do not ride on a daily basis cited higher levels of fashion and equipment concerns, including: [1]

- It is difficult to bring spare clothes (44%).
- Clothing/grooming are a problem (36%).
- I hate arriving somewhere all red and sweaty (34%).
- Helmets mess up my hair (31%).

3. Confidence: most of women feel no secure in their skills, because they can`t fix any problem that could face such as fixing a flat tire which makes them conference.
4. Consumer Products: women are a powerful consumer force, but too often, they do not feel welcome in bike shops or do not feel products address their desires and needs.
5. Community: in some cultures, riding bike is a shame for women and it could refer to the social level as rich women don`t drive bikes or it`s only for poor people.[2]

Related Work:

They found a visualization about Hubway trips taken by male riders vs. female riders for the top three origin stations, late night trips and early morning trips. This visualization Gender is only provided by registered riders. They were intent on not allowing other designs to fixate their focus. However, there is a possibility they were slightly influenced by some-since they were able to see the narrow screenshot as shown below.

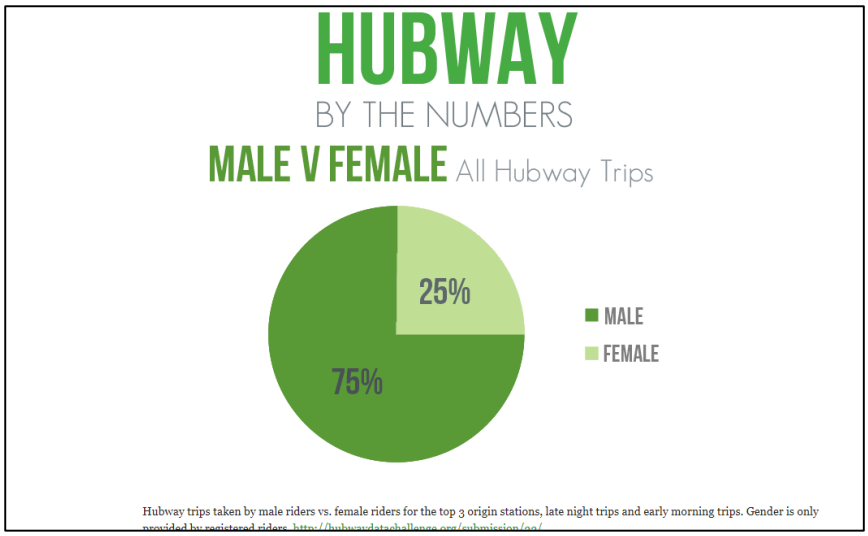


Figure 1- all Hubway trips

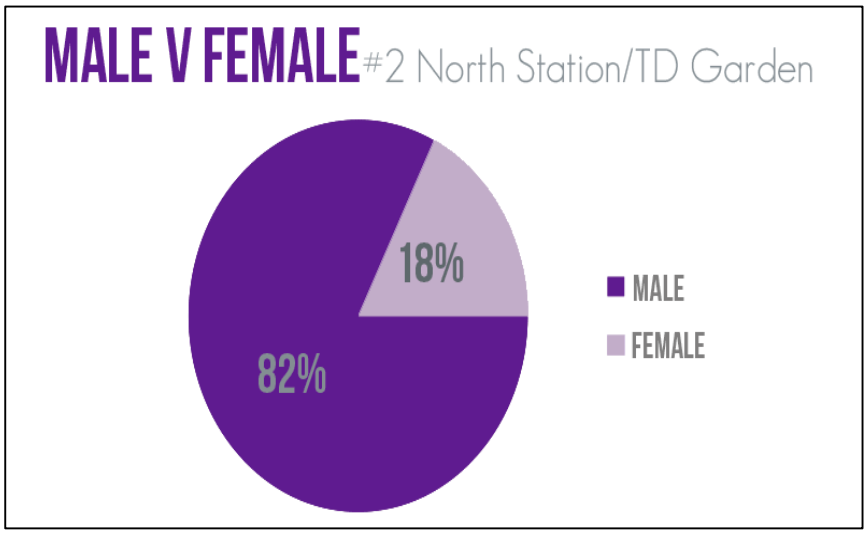


Figure 2- partial Hubway trips "location based"

Up to our knowledge, no one discuss the duration for female. Most of the competitors used Python language but we are the only team who used Rapid Miner. [3]

Methodology:

1. **Discovery phase:**

We installed our dataset from hubway challenge website.[4] it contains two datasets : stations dataset that has attributes (station ID, Start and End station, status) and trips dataset that has attribute(seq_id,hubway_id,status,duration,start_date,strt_statn,end_date
, end_statn, bike_nr, subsc_type, zip_code, birth_date, gender)

Hypothesis:

H1: female take longer time than male?

H2: distribution for male is more than female. From city wise, from location wise

We took a small portion of data, almost 30raws. From these small amount, we discovered that male and female could be our label for this problem.

Tool used for analysis:

- RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. [5]

In RapidMiner we choose the female and male to be the label that to be predicted, though auto model, which is new feature from RapidMiner that will help in machine learning problems , Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

2. **Data preparation:**

Hub way Trips data:

- Filtered male \female from blank variables(empty)
- Birth date and zip code removed due to the small amount and huge amount missing.
- All rest data no blank variables in it.
- Balance the male and female amount through filter. Male was (314722), Female was (103974)
- Remove start date and end date since we would like to predict duration suitable for male and female not months and seasons.
- Convert seconds to min to make it more reasonable.
- Subscription type & Status are not required since they are unbalanced and noisy data will add only nothing since its one class.

- From station number we merge both datasets to add station name and station location.
- Remove two end stations because their status are “Removed”.

1. Duration :The length of the trip, in seconds. We wanted to use the duration to look at what sort of trips riders were taking. We wanted to define each trip as either a leisure ride or a purposeful ride. For our purposes, we converted this to minutes, eliminating any trip that was less than 60 seconds. (Often, those trips are "trials" by people(
2. Start + End Stations :Each station has a unique ID number as well as the station name. The start station is where the bicycle is checked out from and the end station is where the bicycle is checked back into. We can use the end station information to ensure the riders are students.
3. Location name: the street the contains Hubway stations. We need to know which station is near to a university or a central library domain.

1. Preprocessing contains many sub-process:
   - Define target (we chose gender to be our label)
   - Should discretize? (we desecrated single: Age > type of data)
   - Map values
   - Remove column? (We removed selected column that we do not need it).
2. Replace missing value: in 4 steps we replace nominal, positive infinity, negative infinity and numerical missing values by "missing"
3. Recorder attributes (we ordered columns alphabetically "ascending")
4. Filter examples

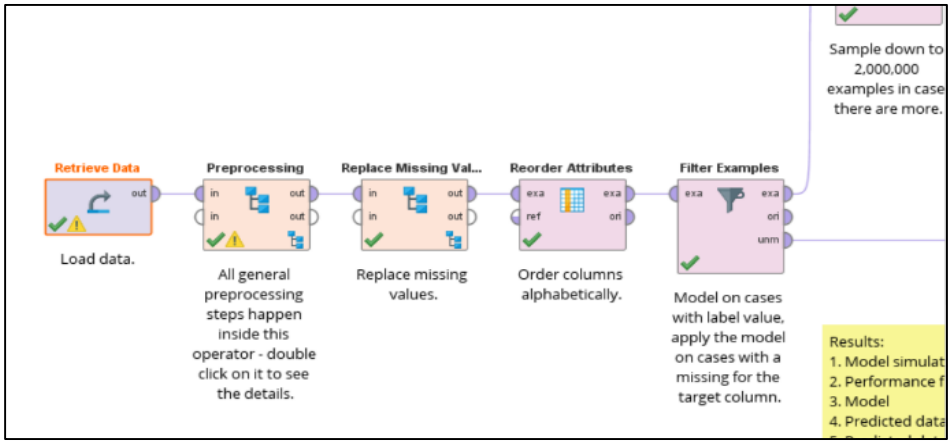


Figure: preprocessing

3. Modeling phase:

Training / Testing:

We spilt the dataset to two datasets (training and testing) we toke 80% from original dataset for training and 20% for testing. Then we used Naïve Bayes training model. After that, we duplicated the dataset to train our dataset more to give more accuracy.

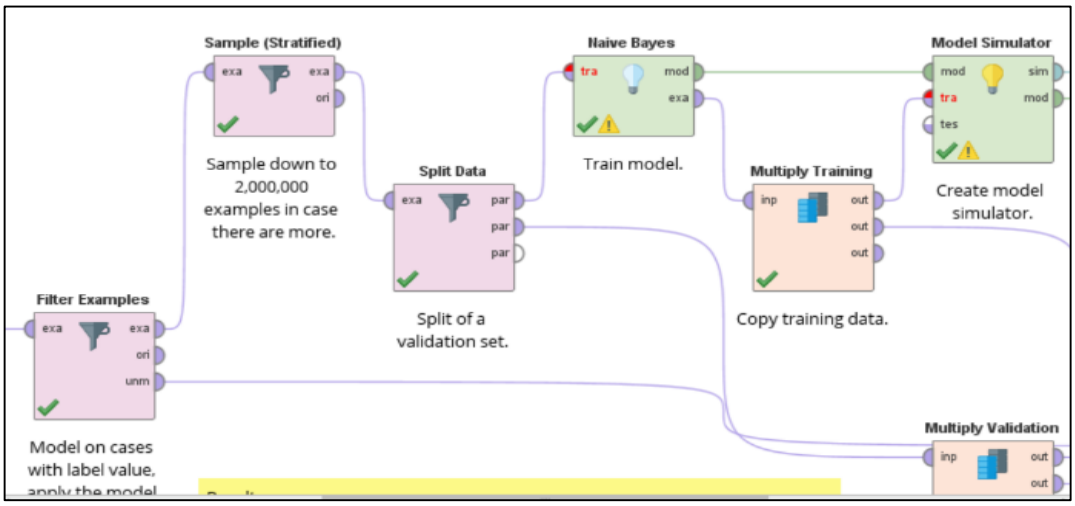


Figure 4- training and testing

4. building phase

- **Performance measure in unbalance dataset:**

	Accuracy	Precision	Recall
Naïve Bayes	69.6%	72.7%	92.5%
Logistic Regression	71.8%	71.9%	99.9%
Deep Learning	72%	77.1%	99.8%

- **Performance measure in balance dataset:**

	Accuracy	Precision	Recall
Naïve Bayes	63.5%	64.5%	59.9%
logistic Regression	63.1%	64.1%	59.3%
Deep Learning	64.3%	66.8%	81.9%

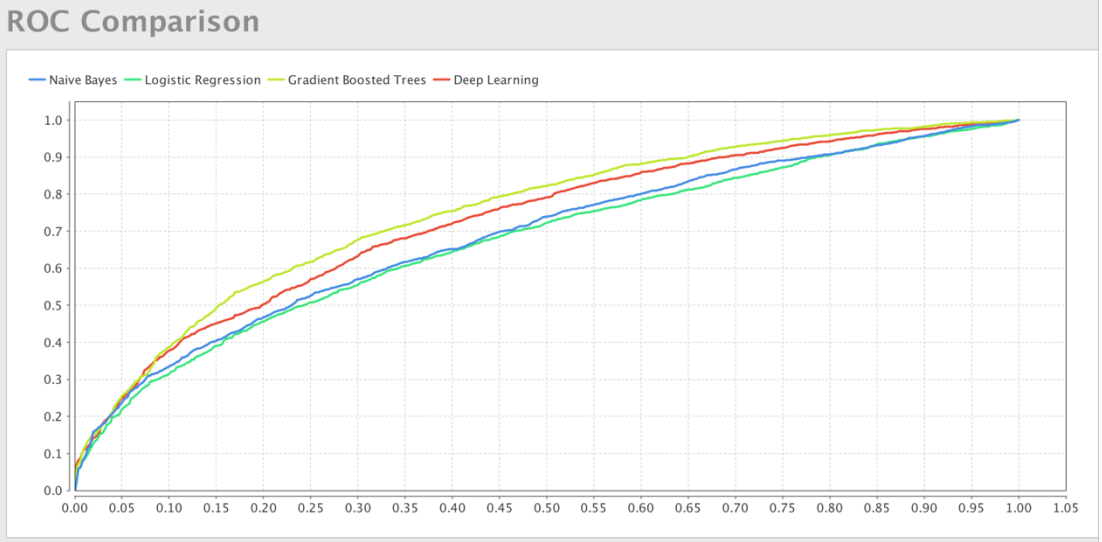


Fig :ROC comparison

5. Communication result:

H1: female take longer time than male?
We had proved that women in average take more time than men.
H2: distribution for male is more than female "From city wise, from location wise"?
The distribution for male is more than female in all cities and location except for the city of Boston.

Results:

- First, we tried to use an unbalanced dataset that contains 43695 rows “43696 with label” but we faced errors because the training model using predict had bias to males due to their large number.

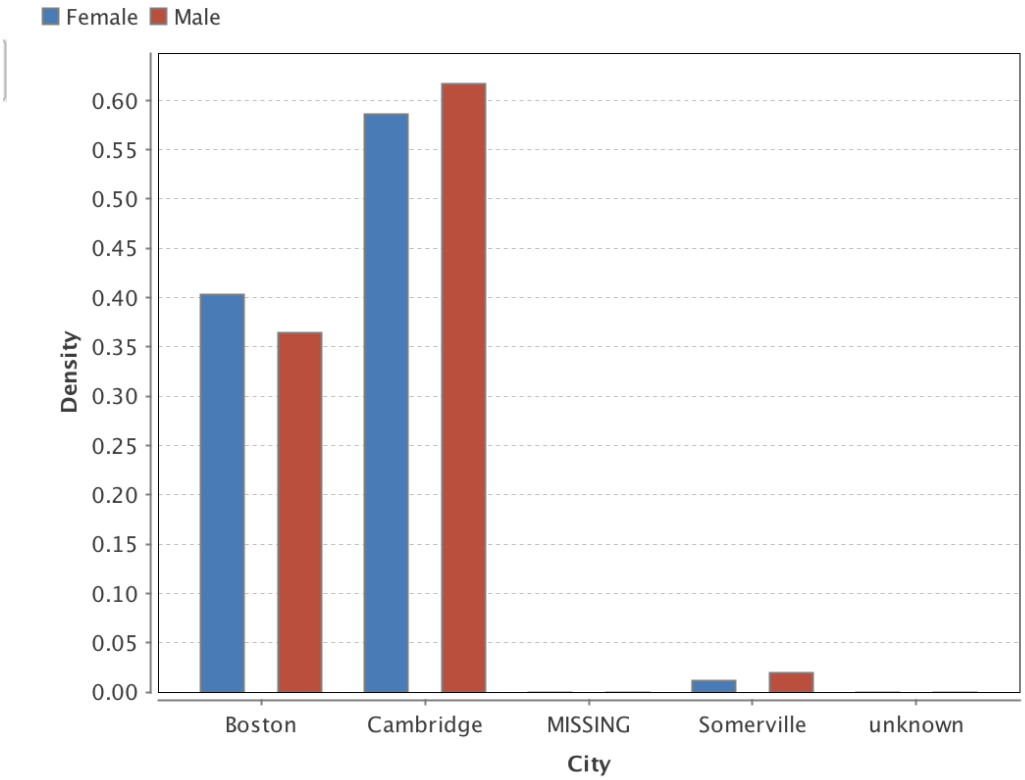


Figure: naïve Bayes

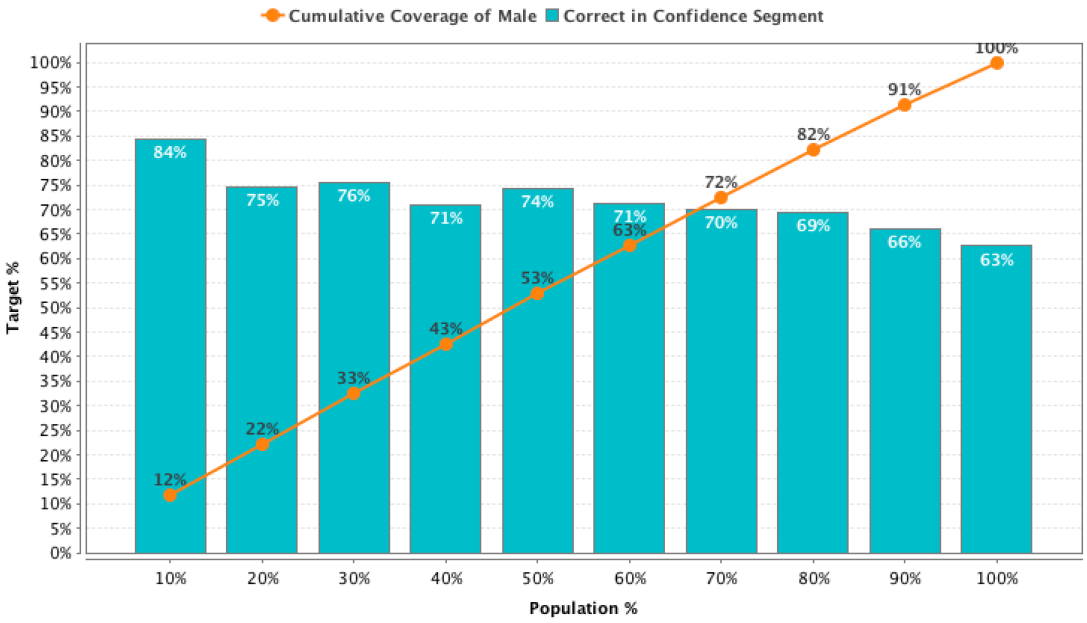


Figure: logistic Regression

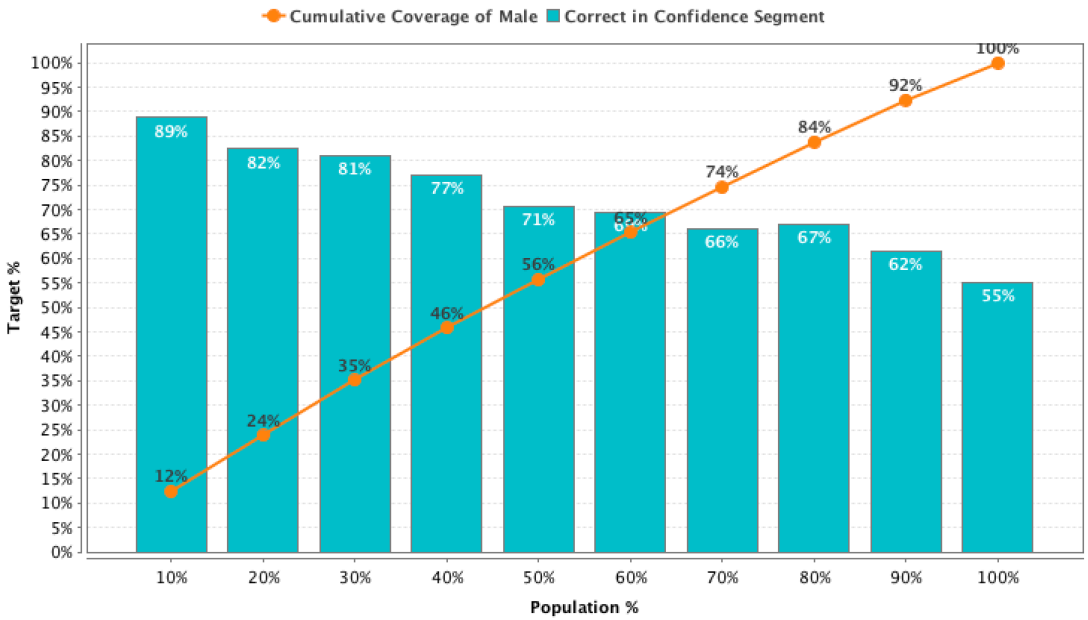


Figure: Deep learning

Second in seek to prevent the bias we decided to convert the dataset to balance dataset "female = male" that contains 24540 rows, we used it with predict model.

The best model was Naïve Bayes due to the high level of accuracy

- By using Naïve Bayes model, we had proved that female takes more time "duration" than male as in the figure shown bellow

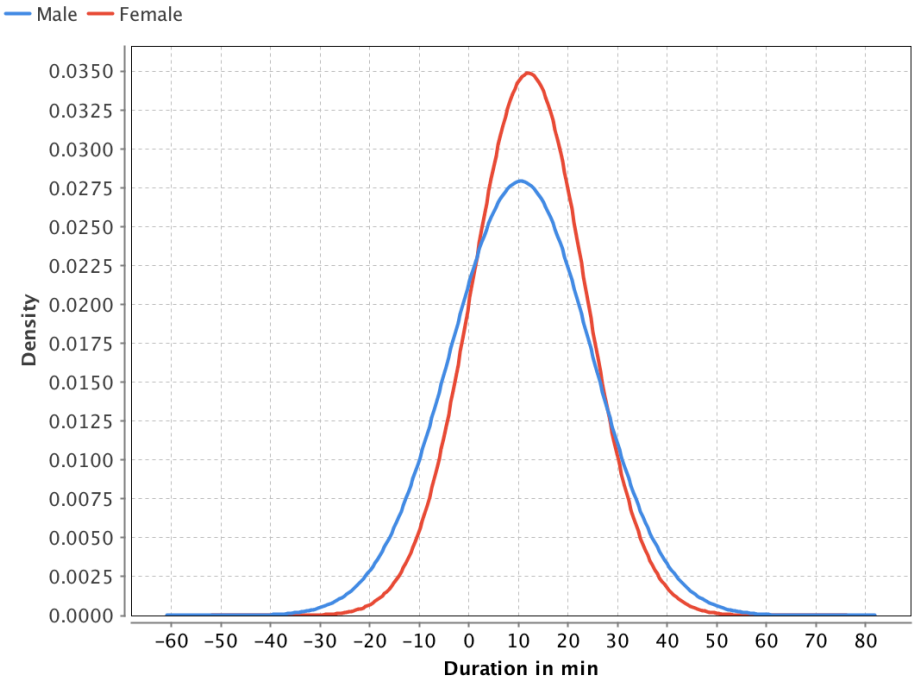


Figure:comparing female and male by duration - density chart using Naive Bayes

- By using Naïve Bayes model, we had noticed that in general the male in average use Hubway more than female as in the figure shown bellow

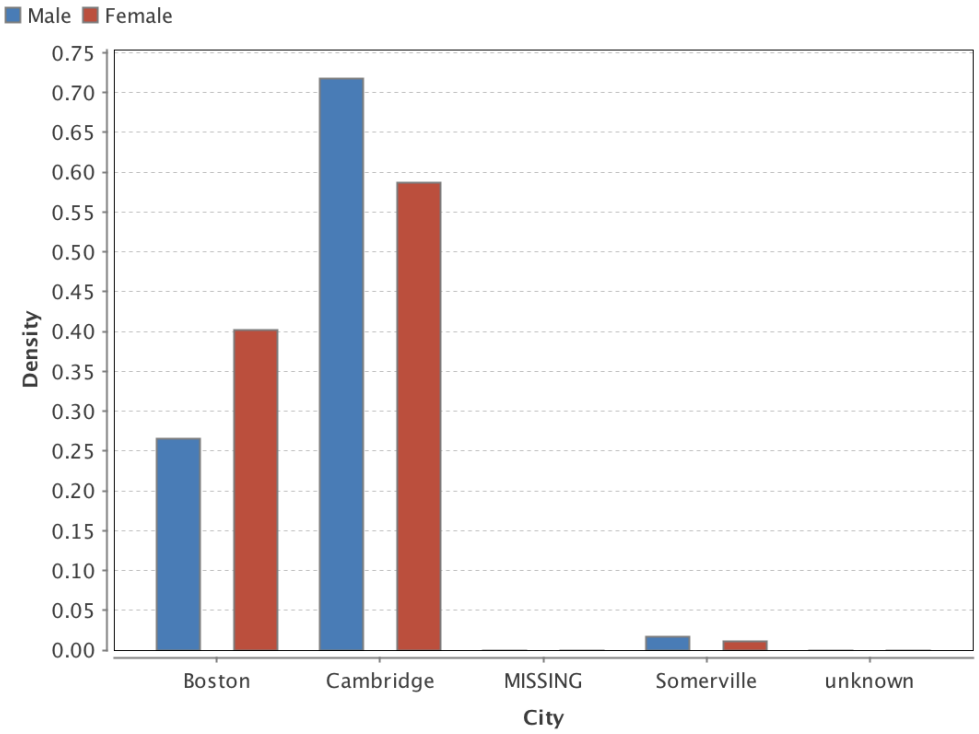


Figure: comparing female and male by city- density chart using Naive Bayes

- By using Naïve Bayes model, we had proved that female takes more time "duration" than male as in the figure shown bellow

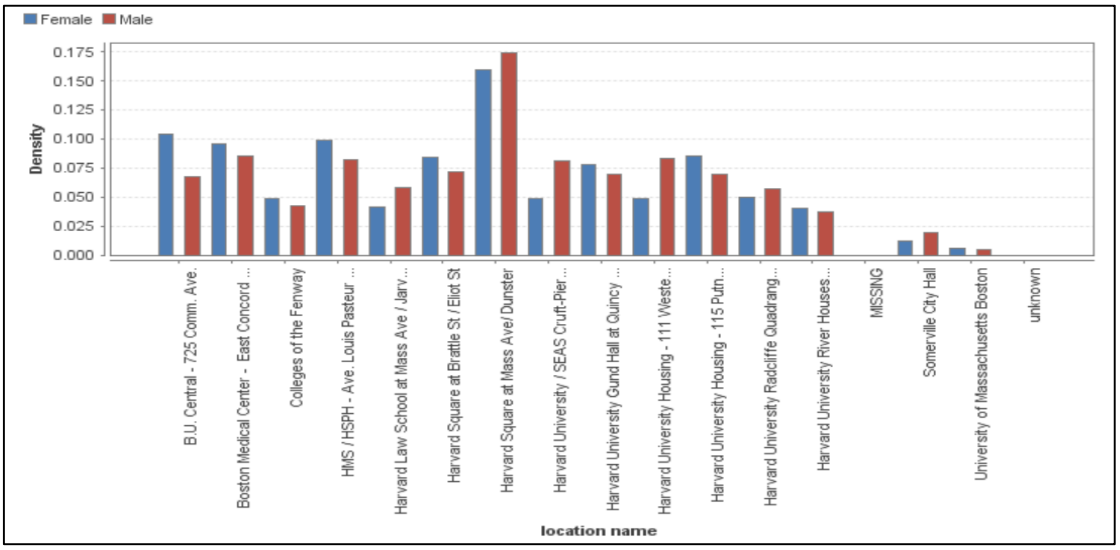


Figure: comparing female and male by city- density chart using Naive Bayes

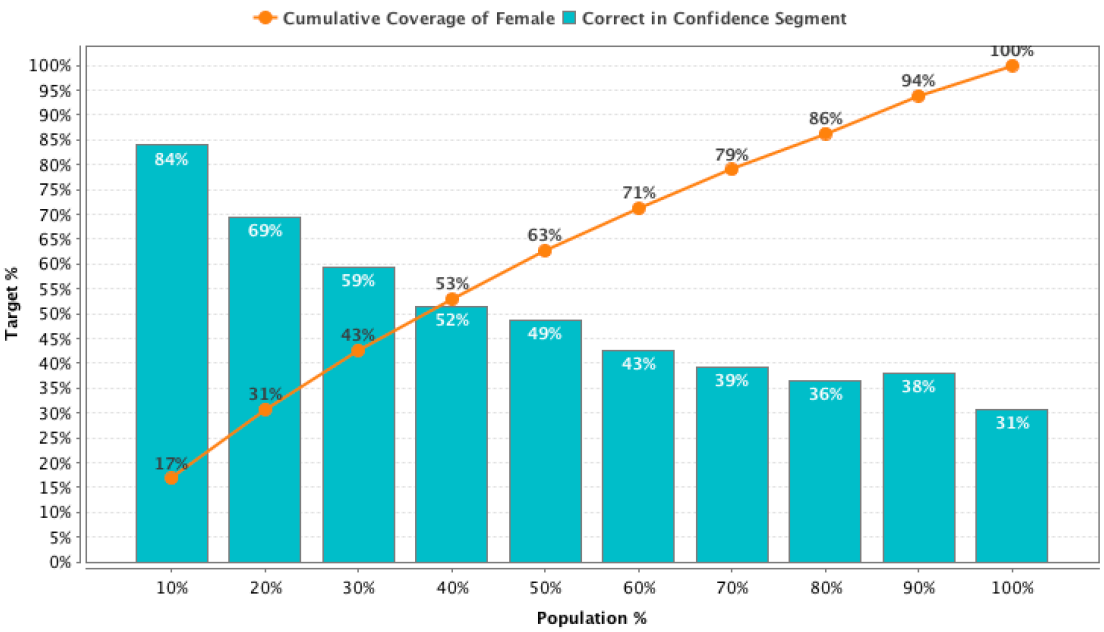


Figure: *logistic Regression*

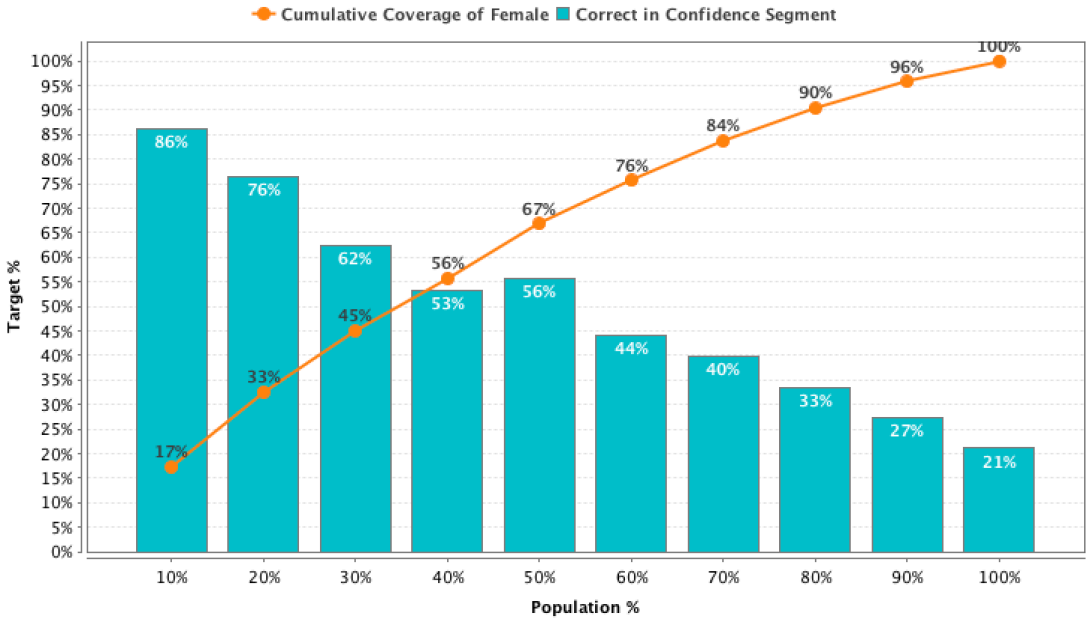


Figure: Deep learning

Conclusion:

At the end, we have some recommendation that will increase the percentage of women who ride bikes; we think that if more bike lanes and bike paths have been provided would increase the safety for women to ride. Hubway can provide women-specific courses, like Women on Wheels series from the Marin County Bicycle Coalition that will increase the women confidence. In addition, plentiful, secure bike parking would increase women riding.

Reference

[1] Alita J. Cousins, & Steven W. Gangestad , Christine E. Garver-Apgar & Jeffry A. Simpson. (2007). Changes in Women's Mate Preferences Across the Ovulatory Cycle.

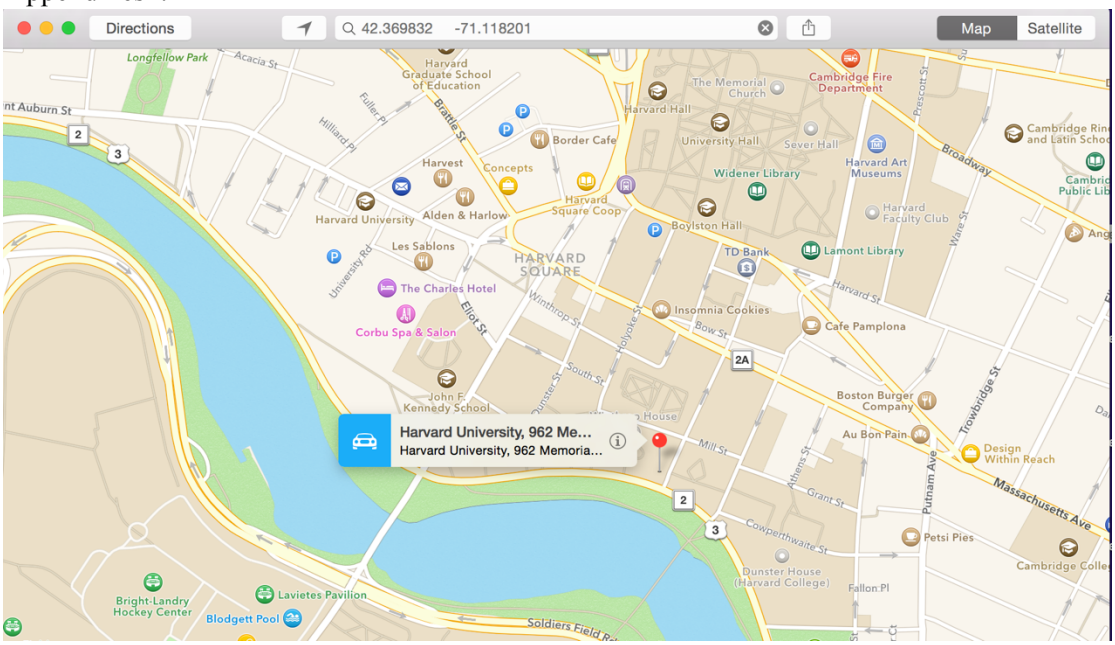
[2] Broache, Anne. Perspectives on Seattle Women's Decisions to Bike for Transportation. 2012

[3] Hubway.tumblr.com. (2018). HUBWAY. [online] Available at: http://hubway.tumblr.com/ [Accessed 12 Apr. 2018]..

[4]"Hubway: Metro-Boston's Bike share Program | The Hubway", *The Hubway*, 2018. [Online]. Available: https://www.thehubway.com. [Accessed: 11- Apr- 2018].

[5] RapidMiner. (2018). Lightning Fast Data Science Platform | RapidMiner. [online] Available at: https://rapidminer.com/ [Accessed 12 Apr. 2018]

Appendixes :


Directions
42.369832 -71.118201
Map
Satellite
Harvard University, 962 Me...
Harvard University, 962 Memoria...
Harvard Graduate School of Education
Harvard Hall
The Memorial Church
Cambridge Fire Department
University Hall
Sever Hall
Harvard Art Museums
Widener Library
Border Cafe
Harvest
Concepts
Harvard Square Coop
Harvard University
Alden & Harlow
Boylston Hall
Harvard Faculty Club
TD Bank
Lamont Library
Les Sablons
HARVARD SQUARE
The Charles Hotel
Corbu Spa & Salon
Insomnia Cookies
Cafe Pamplona
John F. Kennedy School
Boston Burger Company
Au Bon Pain
Design Within Reach
Petsi Pies
Dunster House (Harvard College)
Bright-Landry Hockey Center
Lavietes Pavilion
Blodgett Pool
Soldiers Field
Longfellow Park
Acacia St
Auburn St
Fuller Pl
Hilliard Pl
Brattle St
University Rd
Eliot St
Winthrop St
Holyoke St
Bow St
South St
Mill St
Harvard St
Grant St
Cowperthwaite St
Fallon Pl
Athens St
Putnam Ave
Trowbridge St
Massachusetts Ave
Broadway
Prescott St
Ware St
2
3
2A