Analysis and visualization of bike rental data - Prediction of class of user

A project poster submitted to Dr. PATTABIRAMAN V - School of Computer Science and Engineering
in partial fulfilment of the requirements for the course of CSE3020 – DATA VISUALIZATION

Abstract

We have a dataset of year 2014 of a company named "Capital Bikeshare" which contains information about different bikes along with their distinct numbers that are being used by either a member or a casual user for their journey from one station to another provided with the start time and end time.
Through this project we will be basically analyzing the previous trends of the bike rental company and hence predicting the class of user (A member or a casual user).
We will analyze the trip history of the company and hence arrive at the conclusion. We made visually interactive output for better consumption of data.

Introduction

Through this project we will be basically analysing the previous trends of the bike rental company and hence predicting the class of user (A member or a casual user).
We will analyse the trip history of the company and hence arrive at the conclusion.
Data set Description: It contains the following columns:

- Duration
- Start date
- Start time
- End date
- End time
- Start station number
- Member type
- Bike number
- End Station name
- End station number
- Start station name

➢ This dataset comes from a bike sharing company in US
➢ There are a total of 4,01,122 entries in the tabulation
➢ The entries are dated from 1/1/2014 to 31/3/2014 and belong to Quarter-1 of the year 2014

The coding is done in R language and we used fisher's linear discriminant analysis to predict the class of users.

Methods and Materials

We will be using linear discriminant analysis.
Linear Discriminant Analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in Statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.
This method projects a dataset onto a lower-dimensional space with good class-separability to avoid overfitting ("curse of dimensionality"), and to reduce computational costs.
The we plotted the graph and visualized the output to conclude and able to optimize the business.

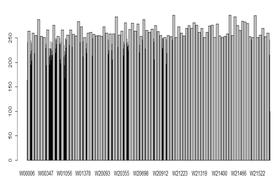


Figure 1. bike number vs frequency

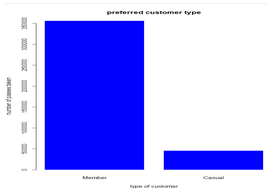


Figure 2. Class of user M or C

Results

This type of project can be helpful for the companies who run trade, car/bike rental or any other commercial business to optimize the supply and maximise their profits in the market.
This can also help them in keeping track of the commodities that are in demand region wise so that they can increase/decrease the supply according to the demand.
This can also help in keeping track of the quality of product. For example-Like in this project, if a particular bike is been used more than 200 times, it can alert the company about its servicing and maintenance.

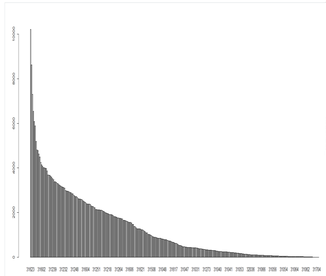


Chart 1. station number used

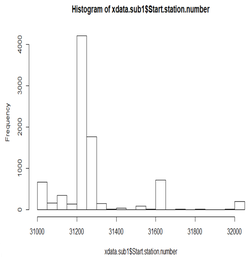
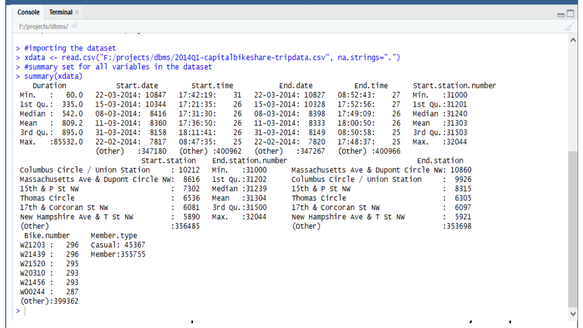


Chart 2. station number vs frequency

We could predict from the sources that the probability that the next person visiting for rental services for more than 1 hour duration would be a member is 0.1139939 and for casual is 0.8860061.
The accuracy of our prediction is 89.25804%.
So by using all these data, we can maximise the profit of the company.
Based on the graph we can decide weather we need to service a bike or not based on its current condition and duration of its service.

Dataset



	Duration	Member.type	Start.station	Start.station.number
440	6666	Casual	15th & Crystal Dr	31003
441	6640	Casual	15th & Crystal Dr	31003
520	5582	Casual	Florida Ave & R St NW	31503
527	5517	Casual	Florida Ave & R St NW	31503
528	5534	Casual	Florida Ave & R St NW	31503
541	8791	Casual	24th & N St NW	31255
542	8769	Casual	24th & N St NW	31255
553	11641	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
554	11624	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
576	3638	Casual	17th & Corcoran St NW	31214
583	6924	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
592	6600	Casual	Smithsonian-National Mall / Jefferson Dr & 12th St SW	31248
705	4272	Casual	Convention Center / 7th & M St NW	31223
707	4317	Casual	Convention Center / 7th & M St NW	31223

Showing 1 to 15 of 8,518 entries

Conclusions

By referring to the histogram between frequency and time duration, we infer that most number of bikes (around 4100) are hired for a time duration of range 5000-10000 seconds.
After studying the histogram between frequency and station number, we infer that most number of bikes (790) are hired from station number 31258. So we learn that availability of bikes at station number 31258 should be kept high.
After studying the histogram we would be able to infer that which bikes are being used more than 250 times. So the bikes that are being used more than 250 times should be sent for servicing.

Contact

Utkarsh Brajnil (18BCE1158)
Himanshu Lohar (18BCE1158)
Email: utkarsh.brajnil2018@vitstudent.ac.in
Website: utkarshbrajnil.github.io
Phone: 9182455113

References

1. Internet, YouTube, Google
2. Stack Overflow
3. Keggal, towards datascience, medium

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com