HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 2 (2023)

Utilizing Machine Learning to Predict the Number of Bikes in an Area

Tavishi Bansal[1] and Guillermo Goldsztein[#]

[1]Irvington High School
[#]Advisor

ABSTRACT

Machine learning is a type of Artificial Intelligence that uses data to make predictions and improve the accuracy of its outcomes. In this article, the problem discussed is classified as supervised learning and the technique utilized is Logistic Regression. After a description detailing what supervised learning and logistic regression are, using a data set to develop a model which predicts the number of bikes a rental bike company should provide based on certain conditions is discussed. The accuracy of this model is also communicated and the challenges and how the final estimations were reached are covered.

Introduction

Machine learning is a subfield of Artificial Intelligence. Its goal is to create models that can be utilized to make predictions accurately. Bikes are a crucial method of eco-friendly transportation and a way to relieve stress and enjoy oneself. They create no emissions and bring various health benefits. Biking is one of the healthiest forms of exercise as it provides a feeling of independence, accomplishment, and happiness and also improves cardiorespiratory fitness and cardiometabolic health [3]. Indubitably, bikes are an integral part of life for people worldwide. They can improve health and are an effective means of transportation, even more so than walking. Furthermore, they are an economically smart and useful tool for many in the world who can not afford cars. A recent study done by Lund University to identify the societal costs of air pollution, climate change, paths used, road wear, health, and congestion found that bike travel is six times cheaper than car travel [4][5].

Rental bike companies are important as they provide bikes to not only residents that use them for their commute but also to tourists. The success of bike rental integration in cities and companies that are offering this service largely depends on making sure there is an availability of bikes at the right places and times [6]. Such companies look for ways to predict and ensure a balanced supply of bikes for all hours of the day, as demand varies frequently. Since bikes are an integral method of transportation and inventory prediction is crucial for successful businesses, this topic is quite relevant to society and is relatable for people everywhere.

In this article, a model which utilizes machine learning in order to predict the number of bikes a rental bike company should provide is explained. Such numbers are helpful to companies, as they can use this information to store as many bikes as necessary. This brings ease of access to tourists and local people and by planning the inventory throughout the day, such companies can optimize their costs. This model is built using data found on Kaggle [1], a website containing a multitude of public datasets that can be used to create machine learning models. This article is organized as follows: after discussing supervised learning and the dataset at hand, it'll delve into the type of problem being tackled, the process of predicting the number of bikes, and discuss the accuracy of this model.

Supervised Learning and Dataset

The problem addressed is predicting the number of rental bikes required at a given time. The dataset consists of data from two years (2011-2012). The features provided are instant, date, season, year, month, hour, holiday, weekday, workingday, weathersit, temp, atemp, hum, windspeed, count of casual users, count of registered users, and the total number of rental bikes. Since many of the features are overlapping, only some of the features will be utilizied. These features and their meanings are as follows:

1. season (1: winter, 2: spring, 3: summer, 4: fall)
2. yr : year (0: 2011, 1:2012)
3. mnth: month (1 to 12)
4. hr : hour (0 to 23)
5. holiday: whether the day is a holiday or not
6. workingday : if the day is neither weekend nor a holiday
7. weathersit :
    a. 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    b. 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    c. 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    d. 4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
8. cnt: count of total rental bikes including both casual and registered

	season	yr	mnth	hr	holiday	workingday	weathersit		cnt
0	1	0	1	0	0	0	1	**0**	16
1	1	0	1	1	0	0	1	**1**	40
2	1	0	1	2	0	0	1	**2**	32
3	1	0	1	3	0	0	1	**3**	13
4	1	0	1	4	0	0	1	**4**	1
...	...	...	...	...	...	...	...	**...**	...

Figure 1. Depicts the first 5 examples in the dataset [2]

This data set contains a total of 17379 examples. The objective of this article and model is to predict the number of total rental bikes needed at a given time after taking into consideration various crucial features. To make this prediction, a model which will take into account all the important features will be utilized. This set can be analyzed using supervised learning, as it takes in the data from the set and learns from the trends to help predict accurately. In this article the makings of the model and why it is the way it is will be detailed.

Numerical Classification Problem

This problem is a numerical classification problem, as the possibilities lie in the range of various numerical values. One day, there may only be the necessity for 50 bikes, while another day may require a supply of over 100 bikes. Problems such as these require the usage of a model that is a function whose input is the features of

the day and whose output is the predicted number of rental bikes required. This prediction is represented by $\hat{y}$. $\hat{y}$ is a function using the features of an example. In this case, each example has 8 features. The features are assigned to $x_1, x_2, \ldots, x_8$ respectively. Therefore, the function is $\hat{y} = \hat{y}(x_1, x_2, \ldots, x_8)$. After feeding the features' values into the model through their representative variables, the output is a number, such as 100, which would indicate that on that day 100 rental bikes will be used.

Logistic Regression

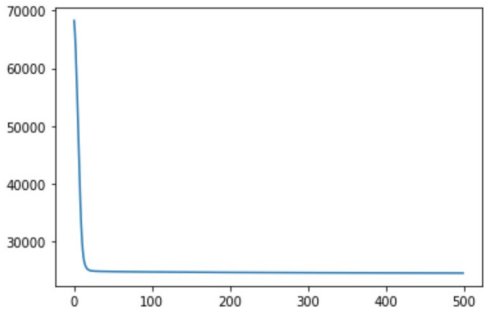

Figure 2. The plot of the function [2]

Throughout this article, one should be under the assumption that each and every example has only 8 features. Logistic regression is a machine learning technique that is used to predict the probability of something occurring. Though often used for binary classification problems, it has been utilized for numerical classification in regards to this dataset.

Training and Validation Set

The data's examples are split into 2 different categories, a training set and a validation set. 80% of the examples are placed in the training set and 20% of the examples are used in the validation set. The reason for this split is to ensure that the model and machine learning are accurate. Furthermore, this prevents the model from overfitting, which is when the model becomes really good at classifying the training set samples but cannot accurately predict data it has not used when testing. The split is performed randomly, ensuring an equal probability for every example to be in either one of the sets. (Ex: example #1 has an 80% chance of being in the training set, as does example #2, #3, and so on).

The validation set is also used to show how accurate a model is. In this case, with a numerical classification problem, visualizing the accuracy in percent form is not a feasible method. Instead, one can examine the trends among the validation set and what the model actually predicted.

```
Y_val[0:10]
```

```
array([[289],
       [487],
       [  5],
       [ 58],
       [177],
       [ 58],
       [219],
       [ 57],
       [119],
       [599]])
```

```
Y_hat[0:10]
```

```
array([[293.44296 ],
       [334.4713  ],
       [100.27577 ],
       [319.50375 ],
       [320.13632 ],
       [ 54.504807],
       [148.30597 ],
       [ 31.399504],
       [221.15968 ],
       [216.48163 ]], dtype=float32)
```

Figure 3. Depicts the validation set alongside the model's predictions [2]

When looking at the results, one can see that both sets follow a similar trend; where the validation set dips, the prediction set shows a significant dip as well. The conclusion can be made that the model is fairly accurate and can be used by companies to predict the supply of bikes that they may need. Since this is a numerical classification problem, there is no specific numerical value that can properly express how accurate the predictions made by the model are. Therefore, the preferred approach is to compare and analyze the training and validation set; after examining the evidence, the logical conclusion made is that the model is pretty accurate because it follows the trends seen in the training set.

Conclusion

In this article, the importance of bike inventory prediction, the type of dataset used, the type of problem at hand, the type of function used in the model, and how the data in the set was used to train the model were discussed. One can observe that the model is fairly accurate, and can be utilized in the real world to help rental bike companies become more efficient and widespread.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Patel, Akash. "Rental Bike Sharing Dataset." *Kaggle*, [circa 2021], https://www.kaggle.com/datasets/imakash3011/rental-bike-sharing. [1]

Bansal, Tavishi. "Tavishi Bansal - Bike Rental." *Kaggle*, [circa August 2022], https://www.kaggle.com/tavishib/tavishi-bansal-bike-rental. [2]

Peruzzi M, Sanasi E, Pingitore A, et al. "An Overview of Cycling as Active Transportation and as Benefit for Health." *Europe PubMed Central*, 01 Apr. 2020, https://europepmc.org/article/med/32429627. [3]

Stefan Gössling, Andy S. Choi. "Transport Transitions in Copenhagen: Comparing the Cost of Cars and Bicycles." *Science Direct*, [circa May 2015], https://doi.org/10.1016/j.ecolecon.2015.03.006. [4]

Stefan Gössling, Andy Choi, Kaely Dekker, Daniel Metzler. "The Social Cost of Automobility, Cycling and Walking in the European Union." *Science Direct*, [circa April 2019], https://doi.org/10.1016/j.ecolecon.2018.12.016. [5]

Yanyong Guo, Jibiao Zhou, Yao Wu, Zhibin Li, Jian-Guo Liu. "Identifying the Factors Affecting Bike-sharing Usage and Degree of Satisfaction in Ningbo, China." *National Center for Biotechnology Information*, 21 Sept. 2017, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5608320/. [6]