

Predicting Bike Usage in Washington, DC using Machine Learning

MTH 496 Final Project
Katie Pocock

Introduction

The ability to predict how weather may affect bike usage can be critical to companies planning on developing more rentable bike services. Information found from this study can factor into what cities more money should be put into the rent-a-bike funds; such as spending more on cities that have less snow and rain rather than cities with regularly uncomfortable weather conditions. Also, when planning or designing road infrastructure in cities such as Washington, DC decision making should consider the number of bike commuters in order to promote traffic safety..

Choosing a bike for everyday commutes can have a great affect on an individual's carbon footprint along with having considerable health benefits. According to research from the University of Oxford, choosing a bike over a car just once a day can reduce the average person's transportation-related emissions by 67%. While routine exercise can have both mental health and physical health benefits.

I utilized an SVM, Random Forest, and ANN classifiers to find predictive factors for bike usage. Using classification techniques in Machine Learning may help us identify factors that have substantial effect on bike usage; learning where rentable bike services will be the most beneficial and impactful.




Dataset Overview

The given data is a combination of how many bikes were used from Capital Bikeshare and the weather on such dates of collection. Capital bikeshare publishes downloadable data regarding over 500 thousand bicycles around Washington, DC including factors such as start and end destinations, Bike Number, Count, and more. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues; my dataset tackles some of the effects of the environment on bike usage.

The given data is split into training and testing subsets, of which the training holds 5,000 individual points at which data was collected. Each date of collection is divided into 10 separate features of various statistics. This ranges from the weather conditions of that day accumulated in different forms on measurement, to the time of day collected, month, and whether it is a holiday.

My first step in making the data usable for my research making it a dataframe and labeling the categorical type column values with numbers. I did so by changing the 'Yes' and 'No' values to be 1 and 0 respectively for both the testing and training subsets. Similarly, I changed the seasons to range from 0 to 3 and weather columns to be from 0 to 2. The company was analyzing this data in the period of 2011 to 2012.

The data is fed into my chosen algorithms in 10 separate features and goes through its respective calculations before outputting a prediction of the number of bikes rented.



Machine learning algorithms:

1. SVM - Support Vector Machines is machine learning algorithm that analyzes data for classification and regression analysis. SVMs are used in categorization, image classification, handwriting recognition and in the sciences, though it is best suited for classification. SVM algorithms aim to find a hyperplane in an N-dimensional space that classifies the data points distinctly. My algorithm minimizes the error in a 10 dimensional hyperplane, according to the dimensions of my dataset, in order to classify the count of bikes used.
2. Random Forest - Random forest consists of a large number of individual decision trees that operate as an ensemble for regression, classification, and other tasks. . Each individual tree in the random forest spits out a class prediction and the prediction that is returned the most by all of the trees becomes our model's prediction. My algorithm will train itself to best predict the status of the weather with the weather related inputs in my dataset.
3. ANN - An Artificial Neural Network is an information processing technique that includes a large number of connected processing units that work together to process information. The collection of connected units or nodes, which loosely model the neurons in a biological brain, can transmit a signal to other nodes. After receiving signals they process it and can signal nodes connected to it. For each node, a number is used as a weight of the strength between connections of nodes to make calculations. Meaningful results are generated by doing classification, regression of continuous target attributes, pattern recognition, and more from it. My ANN is used with 4 layers of 200 neurons each for this specific project.



Evaluation Metrics

We want to have a model with both good precision and recall; which is found using the F1 score type of evaluation metric. The F1 score sort of maintains a balance between the precision and recall for your classifier. If your precision is low, the F1 is low and if the recall is low again your F1 score is low.

Here we can weed out accuracy because accuracy is only a valid choice of evaluation for classification problems that are well balanced and don't have any skewed points which I do in my dataset. Precision can also be ruled out since it is implemented in the F1 and doesn't include recall. Precision is a good choice when you want to be very sure about your prediction only; it tends to ignore important features since it focuses on being very precise with a small amount of data. Recall alone is also not the best choice since it can be measured with F1. This alone is best when you want to capture as many positives as possible. Not suitable for our problem since accuracy can be skewed and it would not be a good accuracy measurement.

$$(10.1) \text{ Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$(10.2) \text{ Precision} = \frac{T_p}{T_p + F_p}$$

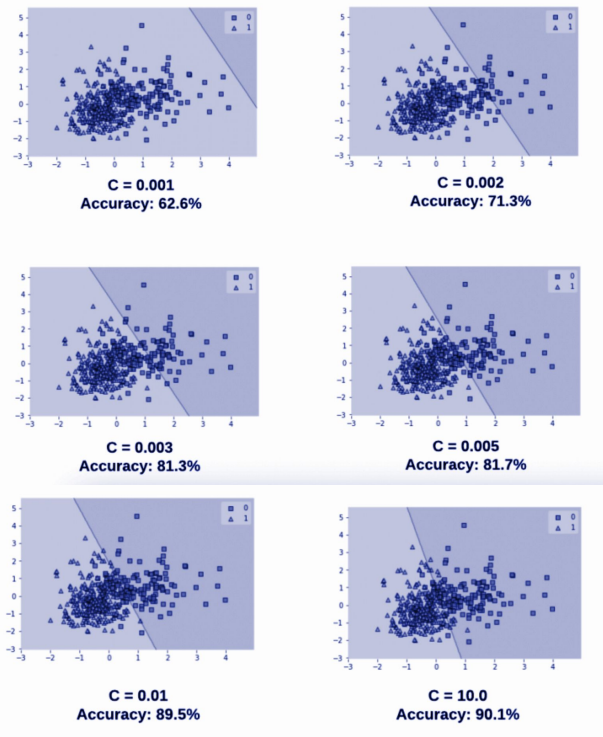
$$(10.3) \text{ Recall} = \frac{T_p}{T_p + T_n}$$

$$(10.4) F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Methodology

For my SVM, I used a radial basis function. Since the number of observations in the dataset far outweighs the number of features this is the best usage of the function.

For the Random Forest method and the ANN method, both methods can intrinsically select features through the training process; therefore I did not need to choose the features for each of these methods. Feature selection Random Forest method is an “embedded method”. Pulling the important factors and focusing on them. Similarly, the ANN uses its neurons that have less impact to not be referred to as often while the more important features that have a bigger impact are magnified therefore the features with more impact are chosen.



Results

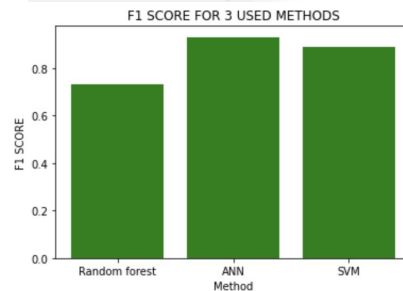
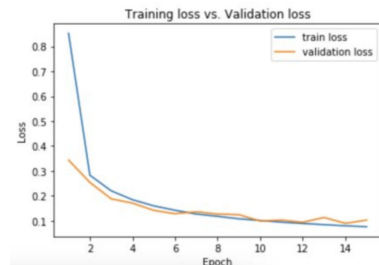
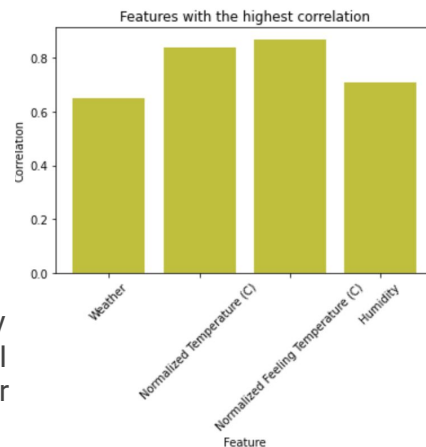
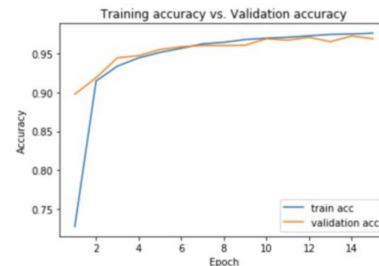
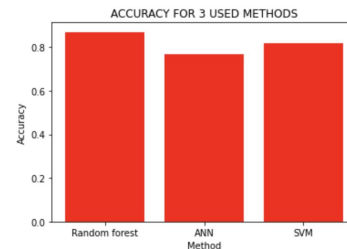
I decided to format the results in boxplots in order to compare the three types of methods and what results I got from each.

Looking at the accuracy score results each of my methods had a good score; not reaching the 90% but all at about the same range of high 70% to low 80%. The random forest generator did the best in this comparison.

Now, looking at the epoch F1 scores there are similar results. The random forest did notably worse than my other methods; ANN did great in this case; which makes sense!

Finally, I plotted the features that had the highest correlation for my dataset. There were not many features therefore I chose the top 4; I could've chosen the top three since there is a drop in correlation for the weather feature.

I also graphed the accuracy and loss for the training set to see both validation loss and train loss and validation accuracy and train accuracy.



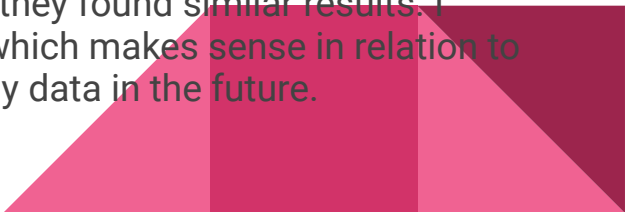
Discussion

Clearly from my project there is a strong correlation between the weather and bike usage from Capital bikeshare's dataset on rentable bikes. In other similar reports a similar relation is found; with air temperature being the outstanding feature with the highest correlation.

This result makes a lot of sense; since Capital bikeshare users may have to rent and leave the bicycle at a station, which can be located away from their destination, they may need to spend more time in the rain or colder temperature therefore discouraging them from using a bike that day.

This is an interesting result; now knowing the effect of the heavy correlated features changes could be made such as making sheltered bike renting areas or choosing to build similar bike rental businesses in areas that are getting less uncomfortable weather conditions. Also, making road infrastructure decisions based on how common it would be to see a biker in the winter months or typically rainier months.

Pazdan, Sylwia found similar results by using a logistic regression model; they found similar results. I believe the random forest classifier was strong in finding the predictions which makes sense in relation to using logistic regression. I would be interesting in trying this method on my data in the future.




Future Work

With more time, it would be interesting to see how different layers of my ANN would impact the results I got. I used 4 layers of 200 neurons in order to produce my predictions; but from our Homework 3 there is a clear difference in the amount of layers and neurons used in an ANN model. Such changes in accuracy and run time may have an interesting output that may change my findings.

Another shortcoming was my given dataset only having certain features from the Capital Bikeshare dataset. There are certain factors disregarded such as the distance or duration of bike usage that could be interesting when comparing to the weather conditions. For example, if shorter duration rides are more common on non-bikeable days research could be done about those trips. If I were to attempt a similar project in the future I would use the Capital Bikeshare dataset to its full extent and add the weather data so that I can still get similar results when looking at certain features but not weed out those bike trip characteristics.

Finally, I would love to look at how the results I found compare to the current findings of Capital Bikeshare. Factors such as climate change were not calculated in my findings but may have a significant impact on bike usage today.



Conclusion

Briefly summarize dataset, important results and conclusion presented in the report

With the motive of finding and predicting bike usage depending on weather conditions, my method of using the Random Forest for classification was the best. This was due to its accuracy being very high especially compared to the other values and also a strong F1 score. My findings of accuracy scores was quite high; I believe this was due to my dataset having an imbalance of it being common for a high bike usage count number. The given dataset was a bit skewed toward a majority of high bike usage since it is so common in Washington, DC. The more predictive features I found were Weather, Normalized temperature C, Normalized Feeling Temperature(C), and Humidity. Such features are great predictors in guessing bike usage in order to make rational decisions based on how likely there is to be a high bike usage in a city.

In conclusion, I believe my project was a great taste of using Machine Learning and Deep learning methods to answer a real world applicable problem. I got to choose a topic I am very interested in and not only learned about Machine and Deep Learning but how to manipulate a dataset I am unfamiliar with in order to produce appropriate findings.



Bibliography

Yiu, Tony. "Understanding Random Forest." Medium, Towards Data Science, 29 Sept. 2021, <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.

"What Is a Support Vector Machine (SVM)? - Definition from Techopedia." *Techopedia.com*, <https://www.techopedia.com/definition/30364/support-vector-machine-svm>.

"How Much Carbon - and Money - You Can Save by Biking." *Future*, <https://www.future.green/futureblog/save-carbon-biking#:~:text=Reducing%20Carbon%20Emissions,of%20CO2%20per%20mile%20traveled>.

Acharya, Shwetha. "How to Improve the Accuracy of a Regression Model." *Medium*, Towards Data Science, 22 June 2021, <https://towardsdatascience.com/how-to-improve-the-accuracy-of-a-regression-model-3517accf8604>.

Get on your bike: Active transport makes a significant impact on carbon emissions - University of Oxford 2021

"Regression Artificial Neural Network." *Regression Artificial Neural Network · UC Business Analytics R Programming Guide*, http://uc-r.github.io/ann_regression#:~:text=Regression%20ANNs%20predict%20an%20output,require%20a%20numeric%20dependent%20variable.

Kumar, Aditya. "SVM (Support Vector Machine) for Classification." *Medium*, Towards Data Science, 8 July 2020, <https://towardsdatascience.com/svm-support-vector-machine-for-classification-710a009f6873>.

DataTechNotes. "SelectKBest Feature Selection Example in Python." *SelectKBest Feature Selection Example in Python*, 11 Feb. 2021, <https://www.datatechnotes.com/2021/02/selectionbest-feature-selection-example-in-python.html>.

"Seven Most Popular SVM Kernels." *Dataaspirant*, 17 Dec. 2020, <https://dataaspirant.com/svm-kernels/#t-1608054630726>.

Pazdan, Sylwiab, "Impact of Environment on Bicycle Travel Demand-Assessment Using Bikeshare System Data." *Sustainable Cities and Society*, Elsevier, 23 Jan. 2021, <https://www.sciencedirect.com/science/article/pii/S2210670721000196>.

