# MACHINE LEARNING PROJECT *TBD*

A PREPRINT

**Hyun Suk Ryoo (Max)**
hr2ee@virginia.edu

**Johnny Wong**
jw6qs@virginia.edu

**Sujin Park**
sjp7yf@virginia.edu

March 22, 2019

## 1 Abstract

The number of people using electronic bikes instead of riding the bus or car has been increasing due to the fact that they are eco-friendly and efficient. In order to maximize the profit of the electronic bike rental companies, our group wanted to predict the busiest and least busy hours of people riding bikes in order for the companies to determine when to put out more bikes and when to take them back to be recharged. Additionally, we wanted to use the prediction model for other bike users for them to see the hours with the lowest bike traffic so that they can enjoy their bike rides in a more efficient manner.

In order to determine the busiest and least busy hours, we took the data for number of bike users in a given hour in the NOVA/DC area and classified them into three classes based on the count percentile: 0 being the least busy, 1 being the middle, and 2 being the busiest. 0 would be the best time for users to ride the electronic bikes. 1 would be regular with normal amount of traffic. 2 would be the time when there is heavy traffic and many electronic bike users in the area. Using this data, we tested several different models to determine the best performing model for our data set. In order to compare the performance, we calculated the accuracy, precision, recall, and F-1 values for Softmax, Random Forest, and Decision Tree models. With our prediction model, both the users and the electronic bike rental companies can benefit and maximize the efficiency of bike usage in NOVA.

## 2 Motivation

Using the data set that contains the number of bike users during a give hour of the day, we wanted to predict the busiest and least busy hours of the day. With this prediction, we can collaborate with electronic bike rental companies that are becoming a big trend right now and apply the information to determine the hours to put out more bikes as well as the hours to take them back to recharge them. During the hours with the greatest number of bike users, there will be more people who are likely to use electronic bikes in order to get to places or just for leisure. Similarly, during the hours with the least number of bike users, there will be less people wanting to use electronic bikes, which means that bike companies can take most of them back to be recharged. Additionally, from the prediction, other regular bike users will be able to know whether to ride the bike or not. Regular bike users will most likely want to avoid riding bikes during the busiest hours since it might take them longer to get to places due to traffic, especially if they do not have much time. However, during the less busy hours, it might be faster for them to ride bikes than to take the bus or drive for short distances since they would get slowed down by the traffic of other bike users. The data will be classified into 3 types: 0 for not busy, 1 for average, and 2 for busy. Therefore, the electronic bike rental companies will put out the most number of bikes during the hours that are classified as 2, less bikes during the hours that are classified as 1, and very few number of bikes during the hours that are classified as 0. For the users, they should ride the bike during the hours that are classified as 0, ride the bike during the hours that are classified as 1 if they have time, and not ride the bike during the hours that are classified as 2. This will result in maximizing the efficiency of people using bikes as well as the profit for the electronic bike rental companies. From the user's perspective, it can decrease bike traffic, and with more electronic bikes available, more people will choose to ride electronic bikes over regular scooters, cars, or buses, decreasing air pollution as well. Furthermore, this can be applied to locations other than Virginia, though Charlottesville's recent boom of electronic scooters was our inspiration.

## 3 Method

As of right now, our main focus was training and testing the Softmax model in order to see if the Softmax model is a good prediction model for our data set. Before training and testing the Softmax model, we first checked the data to see if data cleaning was necessary by checking for missing values, etc. Then we determined the mean, standard deviation, 25 percentile, and 75 percentile of the bike rider count in order to classify them into three classes: 0 within the 25 percentile, 1 being the middle, and 2 above the 75 percentile. This classification was necessary for the users to determine whether they should ride the bike and for the electronic bike rental companies to determine whether they should put out more bikes or not. After classifying the data, we split the data into training and testing sets in order to prepare for training and testing the Softmax model. To see the performance, we calculated the accuracy, precision, recall, and F-1 score from the Softmax model in order to see how the Softmax model performs with our data set. Lastly, we fine tuned the model using grid search in order to improve its performance by trying different values for the hyperparameter C to find the optimal value that will improve the model.

## 4 Preliminary Experiments

After training and testing the Softmax model, we calculated the accuracy, precision, recall, and F-1 score to determine its performance. Accuracy was 0.6700230149597238, precision was [0.71734694 0.64963119 0.65719064], recall was [0.78372352 0.73045024 0.44107744], and F-1 was [0.74906766 0.68767429 0.52787105]. After looking at these accuracy, precision, recall, and F-1 score values, we decided that Softmax model might not be the best preforming model for our data set since the values for all four were relatively low. We then decided to do a grid search to fine tune the hyper parameter C, which we found 3 to be the optimal parameter. The accuracy with the new hyper-parameter became 0.6703107019562716. The precision became [0.71763507 0.64997364 0.65719064]. The recall became [0.78483835 0.73045024 0.44107744]. The F-1 score became [0.74973376 0.68786611 0.52787105]. We could see that our model was able to increase by a small factor for some performance measures when fine tuning the parameters. Also, our F-1 score increased for the first two classes. Therefore, we can say that our model after fine tuning was a better predicting model for our purpose.

## 5 Next Steps

Since the accuracy, precision, recall, and F-1 score from the Softmax model were relatively low, we want to see if there are other models that we can test to find a model that performs better with our data set. Therefore, for the next steps, since we already tested the data using the Softmax model, we plan on using Random Forest, as well as Decision Tree, to further test the data and determine the best model for predicting the busiest and least busy hours for riding bikes. For each of the models, Random Forest and Decision Tree, we plan on getting their accuracy, precision, recall, and F-1 score, in order to compare all three models accordingly. By using and comparing different models, we want to maximize the performance of the prediction model in order to provide the best prediction model that we can provide for both the bike riders and the electronic bike rental companies. During this process, we will perform some feature selection and search for the best hyperparameter and weights for each model. As we get further along the project, we would like to combine all our models into an ensemble that is better than the individual models by themselves.

## 6 Member Contribution

We all worked on the proposal and checkpoint sections together, having meetings to talk about what we are going to do and checking over each others work to make edits and additions. We also all worked to perform the necessary preliminary experiments for this checkpoint, checking with the TAs as well.