Capstone 1 - Machine Learning Applications

For this section, I wanted to use what I had learned in the machine learning unit on my dataset. At this point I'm very familiar with the features and I know what I can infer/predict based on them. I had to tweak my initial hypotheses and predictive goals that I set out when I proposed the project in the beginning. This is because I learned that I couldn't make those leaps/assumptions and get the prediction I originally wanted out of it.

As a context reminder, these data are about rides that riders take on Lyft Baywheels bikes. To recap, I had the following features available to me:

- Start lat lon of ride
- End lat lon of ride
- Start time
- End time
- Duration of ride (created from the above two)
- User type
- Gender
- Age

Those are the core features. Some were too correlated to be useful and I didn't include those up in the list above. For example the duration of the ride is almost perfectly correlated with the cost of the ride to the rider, since I created the pricing model based on how long they rode.

My focus was thus on doing linear regression to predict the duration of the ride based on three dependent variables: the user type, age, and gender. From these I wanted to find how long they rode. I ran a regular linear regression model from sklearn and found a strong performance with a low RMSE.

I also tried out 5-Fold Cross validation on the linear regression model and X and y variables and got the following result:

```
[-0.00067336  0.01795258 -0.00019745  0.0021323   0.00336142]
Average 5-Fold CV Score: 0.004515098954770735
```

I also wanted to do another algorithm but after EDA and all that I realized there isn't much I can do with these data in the end. As such, I used a standard logistic regression model to see if I might predict what the user type is (either customer or subscriber) based on duration of the ride, gender, and age. It turns out I can quite well. The model was looking at the binary user type class, where 0 = ‘customer’ (rode at least once on the fly) and 1 = ‘subscriber’ (rides and pays monthly). The results of the confusion matrix and classification report show the good performance that was achieved for true positive classifications. This means that the model correctly predicting 92% of the time that the user type was subscriber when it was in fact subscriber. It’s worth noting here however that the data are skewed quite significantly in favor of the Subscriber user type. As such, in the future I would need to figure out a way to undersample that class, or resample and oversample the underrepresented class.

```
[[   18  1075]
 [   19 13238]]
              precision    recall  f1-score   support

           0       0.49      0.02      0.03      1093
           1       0.92      1.00      0.96     13257

    accuracy                           0.92     14350
   macro avg       0.71      0.51      0.50     14350
weighted avg       0.89      0.92      0.89     14350
```

With these results in mind and in light of the EDA I’ve done throughout, I think it’s safe to say that these data really do tell something about the kinds of rides and kinds of people who perform these rides. And it looks like they can be used to figure out future examples of ride durations or user types based on the selected X variables above.