ORIE 4741

Progress Report

We used the SF Bay Area Bike Share dataset. Each data sample will contain the name of a station, a specific hour, whether it is rush hour, day of week, whether it is a holiday, zip code for the station, and also mean information about the weather for that day. We believe that these features are relevant to predicting the number of bikes rented from a station. When choosing these features, we looked for factors that we thought would be related to whether or not people would want to rent a bike. Our prediction will be the number of bikes per station that are rented for the specified station at a specific hour. Since the original dataset contains per-ride information, but we plan to use rides-per-hour information, we needed to do a bit of preprocessing to aggregate all of the data to fit what we needed for our prediction. Each row corresponds to a specific day, for a specific starting station and starting hour. It also contains information about the environment (ie. weather, holidays). For a given row, we will predict the number of bike rides that will be taken at a station, for the given hour under the circumstances.

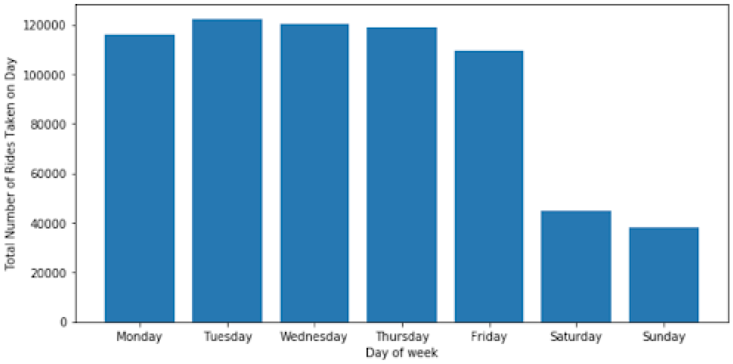
Missing/Corrupted Data

When initially going through our data, we quickly realized that there were rows where the zip code associated with a single bike ride belonged to Oregon, which clearly was incorrect since the data set was supposed to be representative of Californian bike rides. This was an instance of corrupted data in our data set. Since we discovered that there were multiple rows where the zip codes in station.csv did not match up with the actual name of the station, we used a Google Maps API to generate a correct zip code based on the latitude and longitude of the station. To add on, we also had some missing data in the weather data table. For instance, for the feature called precipitation_inches had a floating number for almost every row, except in some, it simply contained a letter T. There were only a handful of sparse rows which had missing precipitation_inches data, so we simply dropped those rows, as we still had more than enough data to work with even after those rows were dropped.

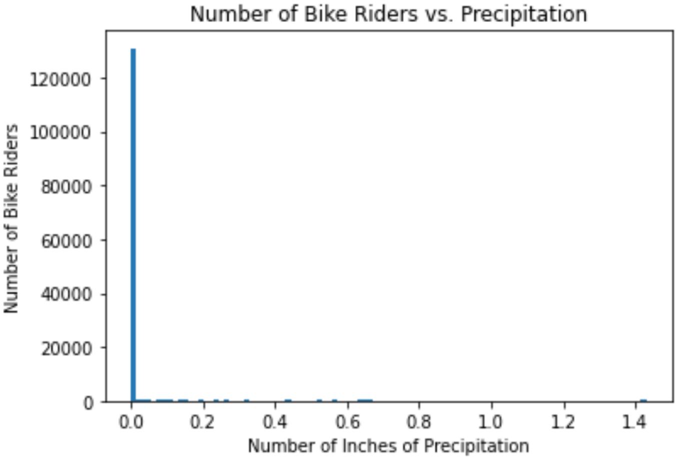
Graphs + Analysis (Initial Visualizations)

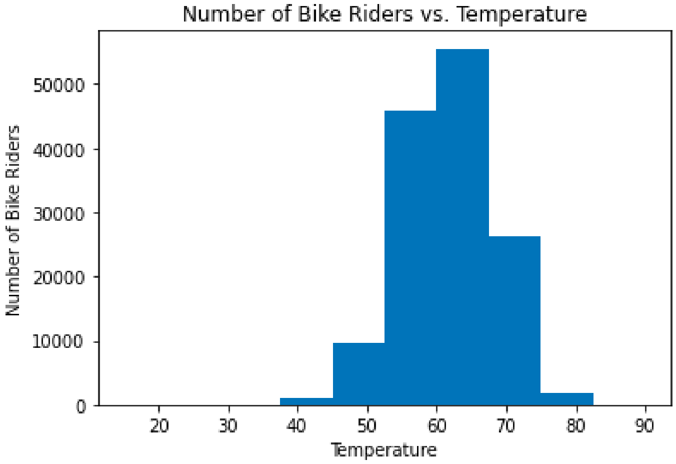
When selecting which features to incorporate into our final data set, there were obvious choices: mean_precipitation, mean_visibility, isRushHour, etc... Clearly, the average inches of precipitation on a particular day would impact the number of bike riders that day, as it would make it much more difficult to go biking. However, in order to ensure the impact of these features on the total number of bike riders, we made some plots that display this correlation.

The following plot demonstrates the correlation between the day of the week and the total number of rides that took place that day. Clearly, the most number of bike rides take place around the middle of the week (Tuesday, Wednesday, Thursday), closely followed by Monday and Friday. Then, there's an evident drop in the number of bike rides during the weekend. Thus, the day of week clearly plays an important factor in determining the number of bike rides.



Similarly, the following two plots also demonstrate that there is a clear relationship between the number of inches of precipitation and the temperature on the number of bike rides.





Results from Regression Modeling

For this milestone, we chose to build various regression models that incorporated different sets of features from our input space. For each of these regression models, we split the dataset into training and testing datasets using Scikit-learn's train_test_split interface. We decided to go for a 67% / 33% train/test split. Then, using sklearn's linear regression models, we fit our dataset, and subsequently logged the error metrics derived from the regression model's predictions. These error metrics included the MSE, MAE and RMSE. As an example, for the first model, we simply included 3 weather-based features - the mean humidity, the mean visibility and the precipitation inches. Our regression model had the following coefficients: [0.02155454 0.00419423 -0.52917649]. These coefficients can be interpreted intuitively as well - the first 2 coefficients have low values, and this makes sense because California's humidity doesn't vary significantly throughout the year. This means that the humidity won't play a role in whether someone chooses to ride a bike. Similarly, visibility doesn't really affect bike-riding on a 'street-level'. Lastly, the coefficient for the 'precipitation_inches' feature is negative because a higher value for that should mean that the number of bike riders decreases (we lose approximately 1 rider for every 2 additional inches of rain).

For our last linear regression model, we chose to include every pertinent feature, with features like the day of the week and the starting station ID being represented by 1-hot vectors. Some important things to note include that the following features have negative coefficients: mean_wind_speed_mph, precipitation_inches, isHoliday, Hour 0, Hour 2, Hour 3, Hour 4, Hour 5, Hour 19, Hour 16, Hour 18, Hour 19, Hour 20, Hour 21, Hour 22, Hour 23, Station 4, Station 5, STation 7, Station 8, Station 27, Station 30, Station 32, Station 61, Station 64, Station 83, Saturday, Sunday, 94041, 94063, while the following features had positive coefficients: isRushHour, Hour 1, Hour 6, Hour 7, Hour 8, Hour 9, Hour 11, Hour 12, Hour 13, Hour 14, Hour 15, Hour 17, Station 3, Station 14, Station 22, Station 28, Station 69, Station 70, Friday, Monday, Thursday, Tuesday, Wednesday, 94107, 95113.

Future Plan

Our future plan is that after we run some regressions and if we come to realize that there are some features that are negligibly impacting the number of bike rides, then we may rerun our regression model on a dataset without those features and see how our predicted values change. Additionally, we plan to take two approaches: First, we will run regression to predict a 16-vector output. We will also run regression for each station in isolation, and we will ultimately compare our results between the two models.