To Eric Garcetti,

My name is Emily Tran and I am researching about bike sharing programs in my Machine Learning course at Occidental College. Through applying machine learning techniques, I believe that the city of Los Angeles will greatly benefit from installing its own a bike sharing system. Reasons include LA's tendency to have good weather year round being put to good use, LA's large demographic of low-income residences benefiting from cheaper mean of transport, and LA's battle against the high production of smog becoming easier. I will now proceed to explain how I have arrived to this claim.

Gathering Data

Our class was given a dataset full of bike sharing information used by Laboratory of Artificial Intelligence and Decision Support (LIAAD) in University of Porto[1]. Essentially, the dataset is a record containing the number of bikes being registered per day in Washington D.C. Each set has metadata on 11 features: *season, year (2011 or 2012), month, holiday, weekday, workday, weather situation, temperature (measured temperature), aTemperature (actual temperature), humidity, wind speed.*

About the Model

For this project, the multiple regression will be used as the prediction model. Multiple Regression is a version of Linear Regression that uses more than one feature to predict a value of outcome. This model aims to create a system that produces the lowest loss function, which is the Residual Sum of Squares (RSS) in this case.

RSS: the sum of the square of difference between the prediction outcome and the true outcome.

Finding the Best Feature Set

Although having 11 feature sets seems more than ideal, sometimes using too many feature sets can hinder a prediction model. This is due to the chance that some features are too related and will then produce inaccurate models. I used three methods to choose the best feature set:

> Adjusted R-Squared ($AdjR^2$): $AdjR^2$ is the percentage of true labels that is explained by the model that has also been adjusted to the number of features. In other words, $AdjR^2$ will be able to catch redundant features. Luckily, there are only 11 features to work with. Therefore there was ample computational power and space to apply linear regression to every possible set of features (see see Table 1). As the set gets the smaller, the more irrelevant the feature sets are to each other.
>
> *Note: Notice how having 11 and 10 feature sets produces the same $AdjR^2$, which is due to the correlation between *temperature* and *aTemperature.* This means that it is useless to include both features since they have similar trends to each other.

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.
[2] Holland, Gale, and Doug Smith. "L.A. County homelessness jumps a 'staggering' 23% as need far outpaces housing, new count shows." Los Angeles Times. May 31, 2017. Accessed October 6, 2017. http://www.latimes.com/local/lanow/la-me-ln-homeless-count-20170530-story.html.

Confidence Levels (CL): CL is the interval below and above the mean that has a 95% chance of containing the true value. In other words, the smaller the interval, the more confident your model is to finding the correct output. Based on the best features of each number of sets, the confidence levels for each coefficient were calculated and recorded (see Table 2).

P-value: P-value of a variable is the probability that the outcome will be true under the condition that the variable is null. In other words, the smaller the p-value of a feature, the better that feature is to include in the final regression model since it will have a higher probability of generating correct outcomes (see Table 3).

Although a feature set of 11 or 10 has the highest AdjR^2, the number of features that produces the smallest confidence level is 7. Regardless, the feature set of 7 has a high AdjR^2 value that is remarkably close to the max value. In addition, the p-values of that feature subset are incredibly low, where the largest p-value is 0.00007%. Therefore I will choose to work with the 7 feature subset (highlighted in the tables).

Table 1: Best Feature Sets

# of feats	Best Features	AdjR^2
2	Year, ATemperature	0.628
3	Year, ATemperature,Working Day	0.701
4	Year, ATemperature,Working Day, Season	0.762
5	Year, ATemperature,Working Day, Season, Weather Situation	0.805
6	Year, ATemperature,Working Day, Season, Weather Situation, Windspeed	0.808
7	Year, ATemperature,Working Day, Season, Weather Situation, Windspeed, Weekday	0.812
8	Year, ATemperature,Working Day, Season, Weather Situation, Windspeed, Weekday, Humidity	0.813
9	Year, ATemperature,Working Day, Season, Weather Situation, Windspeed, Weekday, Humidity, Month	0.814
10	Year, ATemperature,Working Day, Season, Weather Situation, Windspeed, Weekday, Humidity, Month, Holiday	0.815
11	Year, ATemperature,Working Day, Season, Weather Situation, Windspeed, Weekday, Humidity, Month, Holiday, Temperature	0.815

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.
[2] Holland, Gale, and Doug Smith. "L.A. County homelessness jumps a 'staggering' 23% as need far outpaces housing, new count shows." Los Angeles Times. May 31, 2017. Accessed October 6, 2017. http://www.latimes.com/local/lanow/la-me-ln-homeless-count-20170530-story.html.

Table 2: Confidence Levels for each # of Best Feature Sets

	2	3	4	5	6	7	8	9	10	11
Season			1.93	1.783	1.795	1.779	1.782	3.066	3.965	3.066
Year	5.17	4.61	4.1	3.72	3.68	3.647	3.658	3.652	3.649	3.601
Month								0.956	0.956	0.956
Holiday									11.25	11.26
Weekday						0.909	0.909	0.907	0.911	0.913
Workday		4.96	4.41	4.00	3.965	3.931	3.915	3.91	4.030	4.032
Weather Situation				3.45	3.418	3.389	4.393	4.389	4.384	4.389
Temp.										78.59
aTemp	15.86	14.17	13.1	12.27	12.22	12.11	12.26	12.34	12.33	88.99
Humidity							17.51	17.57	17.59	17.58
Windspeed					24.54	24.31	25.26	25.23	25.2	25.54

Table 3: p-values of the best feature set of 7

Feature	p-value
Season	4.0786e-44
Year	4.2436e-145
Weekday	7.3349e-05
Workday	9.0101e-56
Weather Situation	2.6689e-31
aTemperature	3.9282e-70
Windspeed	3.9282e-70

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.
[2] Holland, Gale, and Doug Smith. "L.A. County homelessness jumps a 'staggering' 23% as need far outpaces housing, new count shows." Los Angeles Times. May 31, 2017. Accessed October 6, 2017. http://www.latimes.com/local/lanow/la-me-ln-homeless-count-20170530-story.html.

Predicting Using Model

With the best features subset, I created a multiple regression model. Then, I created fake data points that are based on LA's weather (information taken from usclimatedata.com and weather-and-climate.com) depending on the season and the days of the week (see Table 4). You can read Lowest Prediction as the lowest demand and the Highest Prediction as the highest demand.

Table 4: Original Prediction Table (10,000 trials for each category)

Season	Weekday or Weekend	Average Prediction	Lowest Prediction	Highest Prediction
Spring	Weekend	2163	560	3786
Spring	Weekday	2987	1297	4669
Summer	Weekend	3090	1785	4400
Summer	Weekday	3908	2521	5276
Winter/Fall	Weekend	1933	370	3480
Winter/Fall	Weekday	2726	1107	4353

Analysis of Data

Indeed, these numbers seem low in comparison to the nearly 4 million people who resides in LA. However, since the model was using training data from Washington D.C. that has a population size of almost 700,000, the predictions will not scale to LA's dense population. That being said, we can multiply each prediction by 5 since LA's population is around 5 times larger than that of Washington D.C (see Table 5).

Table 5: Scaled Prediction Table (10,000 trials for each category)

Season	Weekday or Weekend	Average Prediction	Lowest Prediction	Highest Prediction
Spring	Weekend	10815	2800	18930
Spring	Weekday	14935	6485	23345
Summer	Weekend	15450	8925	22000
Summer	Weekday	19540	12605	26380
Winter/Fall	Weekend	9665	1850	17400
Winter/Fall	Weekday	13630	5535	21765

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

[2] Holland, Gale, and Doug Smith. "L.A. County homelessness jumps a 'staggering' 23% as need far outpaces housing, new count shows." Los Angeles Times. May 31, 2017. Accessed October 6, 2017. http://www.latimes.com/local/lanow/la-me-ln-homeless-count-20170530-story.html.

The numbers might still seem low in comparison to LA's population, but I want to mention several key ideas:

Help workers navigate: Notice that the predicted number of registered bikes is higher during weekdays for all seasons than during weekends. This implies that residents will mostly use bikes as a means of traveling to and from work. I understand that LA has the metro system as well as busses, but offering bikes will reduce the chances of dense crowds underground or in vehicles.

Healthier and eco-friendly option of travel: At the worst case scenario, only 2800 people will register for a bike. However, that is 2800 people who will be partaking in a healthier lifestyle and an ecofriendly approach to navigating around — concepts that every city in the US and the world should promote.

Jumpstart lives of the homeless and the low income: According to LA times, there are 57794 homeless people living in LA, which is a 23% increase from the previous year[2]. This means that nearly 60 thousand people most likely cannot afford a car. A bike sharing program will be able to help those who cannot afford basic essential needs, those who are struggling as low income residents, and those who do not feel comfortable taking public transportation with these labels.

Think about the tourists: LA is a very popular destination for travelers. Offering a bike sharing program will attract tourists as taking Uber, Lyft, or taxi gets expensive. Therefore the actual number of registered or casual bikes will be higher than the predicted values.

Other Methods Used to Improve Prediction Model

Using the subset of best features, a multiple regression model was created. To reiterate, the multiple regression model aims to find the smallest Residual Sum of Squares (RSS). While RSS is powerful, sometimes it can produce inaccurate models for large sets of features that creates an illusion of a good model without accounting for redundant features. Since RSS is related to $AdjR^2$, this explains why the full feature set of 11 produces the highest $AdjR^2$ but had poor confidence levels. Therefore two other loss functions were created that allow data scientists to — in a way — penalize the weight of feature sets: Ridge and Lasso. Unfortunately, these methods performed less adequate compared to the original multiple regression model.

Conclusion

Based on the predictions using multiple regression, it seems that the city of Los Angeles will greatly benefit from installing a bike sharing program. Residents will be able to use bikes as an additional way of traveling that will be healthier for them and for our planet. The bike sharing program will reduce the crowds in busses and metros, and it will also help those who cannot afford cars. Those who can afford to travel to LA can use bikes as a cheaper and more fun alternative to taxi services. According to the prediction model, the worst case scenario is that only 2,800 people will register for a bike on that particular day; best case scenario is 26,380

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.
[2] Holland, Gale, and Doug Smith. "L.A. County homelessness jumps a 'staggering' 23% as need far outpaces housing, new count shows." Los Angeles Times. May 31, 2017. Accessed October 6, 2017. http://www.latimes.com/local/lanow/la-me-ln-homeless-count-20170530-story.html.

bikes. No matter the case, the predictions estimates that thousands of people will benefit from registering for a bike daily.

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.
[2] Holland, Gale, and Doug Smith. "L.A. County homelessness jumps a 'staggering' 23% as need far outpaces housing, new count shows." Los Angeles Times. May 31, 2017. Accessed October 6, 2017. http://www.latimes.com/local/lanow/la-me-ln-homeless-count-20170530-story.html.