MGT 6203 Group Final Report

Predicting Bike Rentals from Weather Data

TEAM INFORMATION

Team #: 65

Team Members: Tyler Watt (TWATT6), Mark Wisinger; (MWISINGER3), Joy Watt; (JWATT7), Ramana Vanga; (RVANGA3)

OBJECTIVE/PROBLEM

Background Information:

Bike renting is an activity that has been growing in popularity in recent years due to increased environmental awareness as well as a need for low-contact transportation. It is expected that this trend will continue, and the demand for biking rental options will increase. However, biking is only a viable option in certain areas. High precipitation, extreme weather, and poor city infrastructure can all act as major barriers to biking, while high population density and traffic conditions can be accelerators. Both New York City and Seoul have heavy pressures towards biking and therefore towards the business of rental bikes.

Problem Statement:

One of the most challenging parts of developing a business is predicting the sale of product over time. It is critical to know how market factors will affect demand. For services that revolve around outdoor activities, such as bike rentals, factors such as weather, air quality, and location will also play a role in the number of customers. Businesses planning on opening bike rental locations need to be able to predict sales based on environmental factors to determine business viability and how much inventory they need to keep at a time.

Primary Research Question (RQ):

What effect does the external environment have on the number of bikes rented?

Secondary Research Questions:

1. What effect do the seasons have on the number of bikes rented?
2. What effect does the geographic location (New York versus Seoul) have on the number of bikes rented?
3. What effect does previous business have on the number of bikes rented? (I.e. if a company rented out a lot of bikes last month, does that impact the number of bikes this month?)

Business Justification:

According to some sources, the global bike rental market produced $2.1 billion in 2021 and is anticipated to increase to $11.3 billion in the next 10 years. Factors like the public's desire for environmentally friendly transportation, improved electrical bike technologies, and increased safe biking lanes in recent years may play a huge role in this. Given this increase in demand, companies may want to start setting up bike rental services, but some locations may not be suitable for these types of businesses. For instance, locations with low air quality, dangerously low temperatures, or frequent rainfall may not be able to support bike rental as a service. A business with a model that can accurately predict rental demand can save a lot of money by avoiding unusable areas while focusing on capitalizing on ideal locations before competitors do.

Initial Hypothesis:

We initially hypothesized that the most substantial, statistically significant, effects would come from rainfall and temperature. Factors like geographic location will have a small but significant effect.

Literature Review

Growing concerns about urban sustainability and climate changes have led to an increase in the use of bikes as part of green transportation solutions. In the past decade, bike sharing has grown in as many as 50 countries. There are a lot of articles and algorithms by researchers predicting the bike demand for a certain city or for a certain bike rental node. Authors Pan and Zheng used deep LSTM sequence learning model to predict the demand of bikes renting and returning in different areas of city based on the historical data, weather data and time. According to the authors LSTM sequence learning model could process sequential data, memorize data of past time, learn complex functions and predict sequential data very accurately. Experiments with their model have demonstrated an average Root Mean Squared Error of 2.7069 and have been able to predict demand accurately (Pan and Zheng, 2018).

Authors Sathishkumar and Cho have a different approach for predicting bike demand for Seoul bike sharing data. They think that although artificial neural networks could enhance accuracy, they have complicated structure and high computational cost. Furthermore, because the continuous variation in bike sharing frameworks is exceptionally unpredictable and is influenced by numerous external factors, explaining these connections is challenging. Thus, they used various statistical algorithms CUBIST, Regularized Random Forest, CART, KNN and Conditional Inference Tree. Their results show that CUBIST algorithm improved the R^2, mean squared error, and mean absolute error compared to the other algorithms (Sathishkumar and Cho, 2020).

Authors Wang and Cheng used a regression model with spatially varying coefficients to predict bike sharing demand for Montreal in Canada. They developed a new SVC regression model by regularization. The authors claim that their model outperforms machine learning baseline models for predicting the demand for new stations and the only limitation being that their model does not consider temporal factors, such as weather and time of the day (Wang and Chen, 2021).

All these models only used data set from a single city or only focused on a few rental stations. None of the authors used data from different cities and generated a model based on this combined set of data. However, we feel that combined data might give more insights into the general trend of bike renting.

OVERVIEW OF DATA

We have three different sets of data, one for Seoul city bike share data with weather information for each day of years 2017 and 2018 and have over 2000 data points roughly. This data was found from the link below:
http://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

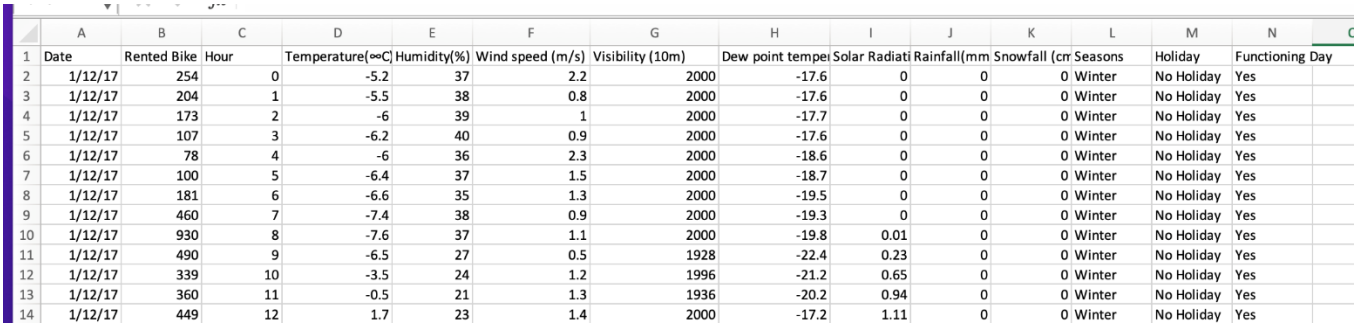
The following link was used to download the Bike share data for New York City for the years of 2017 and 2018. This website has one data file for each month of the year and only contains bike rental data but not the weather data.:
https://s3.amazonaws.com/tripdata/index.html

We used the link below to get weather data for NYC for each day of the years 2017 and 2018:
https://www.visualcrossing.com/weather/weather-data-services

We also found data for Citi Bike customer usage from Kaggle website:

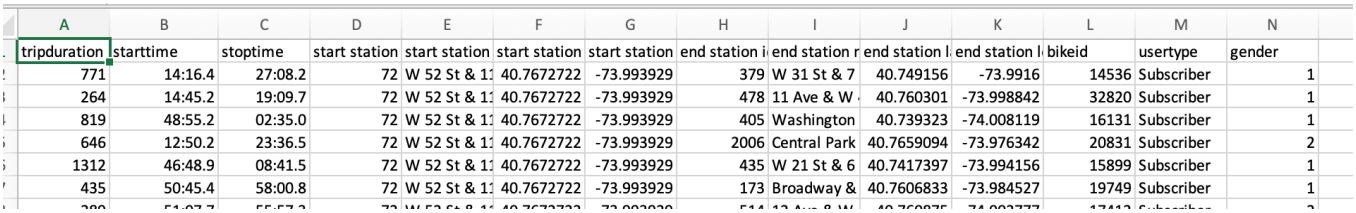
https://www.kaggle.com/datasets/akkithetechie/new-york-city-bike-share-dataset

Our Seoul bike sharing data has a column that shows how many bikes were rented for every hour and we also have the weather details for that hour. Weather details like wind speed, temperature, humidity etc., can be used for algorithm. Additionally, we have an indicator variable indicating it if a holiday.



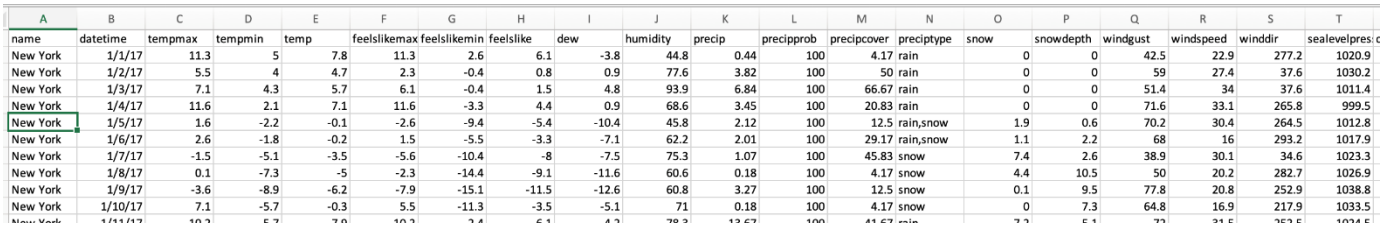
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Date	Rented Bike	Hour	Temperature(∞C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point tempe	Solar Radiati	Rainfall(mm	Snowfall (cm	Seasons	Holiday	Functioning Day	
2	1/12/17	254	0	-5.2	37	2.2	2000	-17.6	0	0	0	Winter	No Holiday	Yes	
3	1/12/17	204	1	-5.5	38	0.8	2000	-17.6	0	0	0	Winter	No Holiday	Yes	
4	1/12/17	173	2	-6	39	1	2000	-17.7	0	0	0	Winter	No Holiday	Yes	
5	1/12/17	107	3	-6.2	40	0.9	2000	-17.6	0	0	0	Winter	No Holiday	Yes	
6	1/12/17	78	4	-6	36	2.3	2000	-18.6	0	0	0	Winter	No Holiday	Yes	
7	1/12/17	100	5	-6.4	37	1.5	2000	-18.7	0	0	0	Winter	No Holiday	Yes	
8	1/12/17	181	6	-6.6	35	1.3	2000	-19.5	0	0	0	Winter	No Holiday	Yes	
9	1/12/17	460	7	-7.4	38	0.9	2000	-19.3	0	0	0	Winter	No Holiday	Yes	
10	1/12/17	930	8	-7.6	37	1.1	2000	-19.8	0.01	0	0	Winter	No Holiday	Yes	
11	1/12/17	490	9	-6.5	27	0.5	1928	-22.4	0.23	0	0	Winter	No Holiday	Yes	
12	1/12/17	339	10	-3.5	24	1.2	1996	-21.2	0.65	0	0	Winter	No Holiday	Yes	
13	1/12/17	360	11	-0.5	21	1.3	1936	-20.2	0.94	0	0	Winter	No Holiday	Yes	
14	1/12/17	449	12	1.7	23	1.4	2000	-17.2	1.11	0	0	Winter	No Holiday	Yes	

Bike share data for NYC only includes the rental details like start station, end station details and the gender of the renter. But we also need weather details for the days the bikes were rented to be able to predict the number of bikes we need per day for rentals.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	tripduration	starttime	stoptime	start station	start station	start station	start station	end station i	end station r	end station l	end station l	bikeid	usertype	gender
	771	14:16.4	27:08.2	72	W 52 St & 11	40.7672722	-73.993929	379	W 31 St & 7	40.749156	-73.9916	14536	Subscriber	1
	264	14:45.2	19:09.7	72	W 52 St & 11	40.7672722	-73.993929	478	11 Ave & W	40.760301	-73.998842	32820	Subscriber	1
	819	48:55.2	02:35.0	72	W 52 St & 11	40.7672722	-73.993929	405	Washington	40.739323	-74.008119	16131	Subscriber	1
	646	12:50.2	23:36.5	72	W 52 St & 11	40.7672722	-73.993929	2006	Central Park	40.7659094	-73.976342	20831	Subscriber	2
	1312	46:48.9	08:41.5	72	W 52 St & 11	40.7672722	-73.993929	435	W 21 St & 6	40.7417397	-73.994156	15899	Subscriber	1
	435	50:45.4	58:00.8	72	W 52 St & 11	40.7672722	-73.993929	173	Broadway &	40.7606833	-73.984527	19749	Subscriber	1

Weather data that we downloaded for New York City contains all the weather-related information that we need to run our algorithm. We have information like wind speed, temperature, humidity just like the data from Seoul city.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	name	datetime	tempmax	tempmin	temp	feelslikemax	feelslikemin	feelslike	dew	humidity	precip	precipprob	precipcover	preciptype	snow	snowdepth	windgust	windspeed	winddir	sealevelpres
	New York	1/1/17	11.3	5	7.8	11.3	2.6	6.1	-3.8	44.8	0.44	100	4.17	rain	0	0	42.5	22.9	277.2	1020.9
	New York	1/2/17	5.5	4	4.7	2.3	-0.4	0.8	0.9	77.6	3.82	100	50	rain	0	0	59	27.4	37.6	1030.2
	New York	1/3/17	7.1	4.3	5.7	6.1	-0.4	1.5	4.8	93.9	6.84	100	66.67	rain	0	0	51.4	34	37.6	1011.4
	New York	1/4/17	11.6	2.1	7.1	11.6	-3.3	4.4	0.9	68.6	3.45	100	20.83	rain	0	0	71.6	33.1	265.8	999.5
	New York	1/5/17	1.6	-2.2	-0.1	-2.6	-9.4	-5.4	-10.4	45.8	2.12	100	12.5	rain,snow	1.9	0.6	70.2	30.4	264.5	1012.8
	New York	1/6/17	2.6	-1.8	-0.2	1.5	-5.5	-3.3	-7.1	62.2	2.01	100	29.17	rain,snow	1.1	2.2	68	16	293.2	1017.9
	New York	1/7/17	-1.5	-5.1	-3.5	-5.6	-10.4	-8	-7.5	75.3	1.07	100	45.83	snow	7.4	2.6	38.9	30.1	34.6	1023.3
	New York	1/8/17	0.1	-7.3	-5	-2.3	-14.4	-9.1	-11.6	60.6	0.18	100	4.17	snow	4.4	10.5	50	20.2	282.7	1026.9
	New York	1/9/17	-3.6	-8.9	-6.2	-7.9	-15.1	-11.5	-12.6	60.8	3.27	100	12.5	snow	0.1	9.5	77.8	20.8	252.9	1038.8
	New York	1/10/17	7.1	-5.7	-0.3	5.5	-11.3	-3.5	-5.1	71	0.18	100	4.17	snow	0	7.3	64.8	16.9	217.9	1033.5

Key Variables:

Below are the variables used in this study and what they mean.

- **Date**: The date for which the bike count was summed.
- **normalized_bike_count**: Bike count is the number of bikes rented on a particular day (Date), but normalized_bike_count is that data normalized to a mean of 0 and an stdev of 1.
- **Location**: The location where the data is from. Either New York or Seoul
- **Temperature**: The temperature on the Date
- **Dew.point.temperature**: For the given Date, the air temperature needed for dew to form
- **Humidity**: The humidity of the Date
- **Rainfall**: The number of inches of rainfall on the Date
- **Snowfall**: The number of inches of snowfall on the Date
- **Wind.speed**: The average windspeed for the Date
- **Solar.Radiation**: The amount of sun exposure during the daytime for the Date

Based on our initial hypothesis, we expected Rainfall and Temperature to play the biggest role in predicting bike rental count.

Data Cleaning

We have data from three different sources. The data format is not consistent, and we must clean the data for us to use the data for our modelling. For example, in the data set for Seoul we have data for each hour of the day and each row includes weather variables and holiday variables.

A1 fx Date

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Date	Rented Bike	Hour	Temperature	Humidity(%)	Wind speed	Visibility (10	Dew point te	Solar Radiati	Rainfall(mm	Snowfall (cm	Seasons	Holiday	Functioning Day
2	1/12/17	254	0	-5.2	37	2.2	2000	-17.6	0	0	0	Winter	No Holiday	Yes
3	1/12/17	204	1	-5.5	38	0.8	2000	-17.6	0	0	0	Winter	No Holiday	Yes
4	1/12/17	173	2	-6	39	1	2000	-17.7	0	0	0	Winter	No Holiday	Yes
5	1/12/17	107	3	-6.2	40	0.9	2000	-17.6	0	0	0	Winter	No Holiday	Yes
6	1/12/17	78	4	-6	36	2.3	2000	-18.6	0	0	0	Winter	No Holiday	Yes
7	1/12/17	100	5	-6.4	37	1.5	2000	-18.7	0	0	0	Winter	No Holiday	Yes
8	1/12/17	181	6	-6.6	35	1.3	2000	-19.5	0	0	0	Winter	No Holiday	Yes
9	1/12/17	460	7	-7.4	38	0.9	2000	-19.3	0	0	0	Winter	No Holiday	Yes

The New York data set is very huge. We have about 24 files, with each file containing one month's data just for New York. The data format for each file is different. For example, as shown below, some of the files have data separated by each day and in some files, we have data separated by hours of the day. We had to aggregate New York's data files like that of Seoul's for us to have consistent data. For 2018 data, Citi Bike changed the format for every entry. Unlike 2017, 2018 data didn't daily data available across the time period. For 2018, the Citi Bike data contained entries for each hour instead of each day. For each month of data file in 2018 we have at least a million rows of data for each month. It was computationally taxing to loop through many rows of data for each month of 2018. We decided to use 2017 NYC Citi Bike data for our analysis, serving as the NYC data counterpart to the 2018 Seoul dataset. Also, the New York data set didn't have weather information and aggregating the data by each day will help us inner join the New York data set with the weather data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	tripduration	starttime	stoptime	start station	start station	start station	start station	end station i	end station r	end station l	end station l	bikeid	usertype	gender
2	970	50:57.4	07:08.2	72	W 52 St & 11	40.7672722	-73.993929	505	6 Ave & W 3	40.7490127	-73.988484	31956	Subscriber	1
3	723	33:30.2	45:33.3	72	W 52 St & 11	40.7672722	-73.993929	3255	8 Ave & W 3	40.7505854	-73.994685	32536	Subscriber	1
4	496	39:18.3	47:35.2	72	W 52 St & 11	40.7672722	-73.993929	525	W 34 St & 11	40.7559416	-74.002116	16069	Subscriber	1
5	306	40:13.4	45:20.2	72	W 52 St & 11	40.7672722	-73.993929	447	8 Ave & W 5	40.7637074	-73.985162	31781	Subscriber	1
6	306	14:51.6	19:57.6	72	W 52 St & 11	40.7672722	-73.993929	3356	Amsterdam .	40.7746671	-73.984706	30319	Subscriber	1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	tripduration	starttime	stoptime	start station	start station	start station	start station	end station i	end station r	end station l	end station l	bikeid	usertype	gender
2	256	12/1/17 0:00	12/1/17 0:04	324	DeKalb Ave &	40.689888	-73.981013	262	Washington	40.6917823	-73.97373	18858	Subscriber	1
3	325	12/1/17 0:00	12/1/17 0:05	470	W 20 St & 8	40.7434534	-74.00004	490	8 Ave & W 3	40.751551	-73.993934	19306	Subscriber	1
4	967	12/1/17 0:00	12/1/17 0:16	347	Greenwich S	40.728846	-74.008591	504	1 Ave & E 16	40.7322185	-73.981656	28250	Subscriber	1
5	125	12/1/17 0:00	12/1/17 0:02	3077	Stagg St & U	40.7087708	-73.950953	3454	Leonard St &	40.7103685	-73.94706	25834	Subscriber	1
6	451	12/1/17 0:00	12/1/17 0:08	368	Carmine St &	40.730386	-74.00215	326	E 11 St & 1 A	40.7295384	-73.984267	14769	Subscriber	1
7	578	12/1/17 0:00	12/1/17 0:10	368	Carmine St &	40.730386	-74.00215	505	6 Ave & W 3	40.7490127	-73.988484	31208	Subscriber	1

Finally, the bike rental data was normalized to have an average of 0 and a standard deviation of 1 (except for the data used for the Poisson Regression).

Models

Linear Regression:

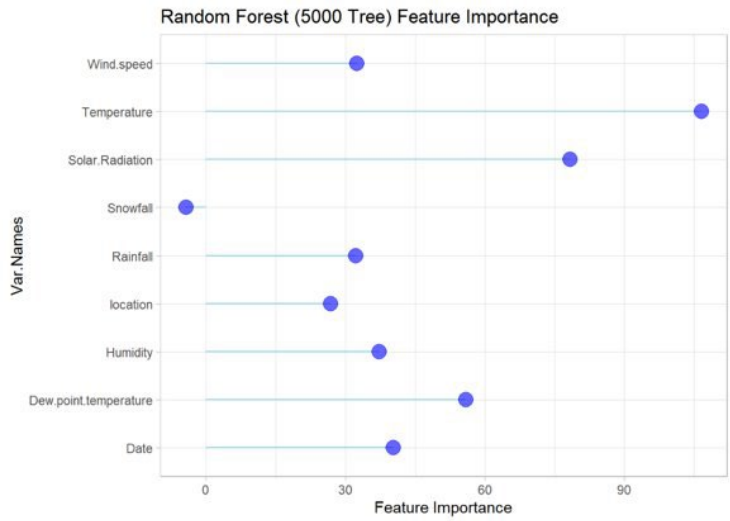
The first model we tried was a simple linear regression using all the factors as predictors of normalized bike count. Based on this model, only temperature, rainfall, and solar radiation were found to be significant. After removing the nonsignificant factors, we saw an Adjusted R^2 value of 0.4919. The model suggests that bike

count has a positive correlation with temperature and solar radiation, but a negative correlation with rainfall. This makes sense, as warm, sunny days are better suited for riding bikes than rainy days.

To compare the full model with the subset model (the model that only used the significant factors), we used Akaike Information Criterion, or AIC. In general, a lower AIC indicates a better model. The full model had an AIC of 480 compared to the subset model's 476, indicating that the subset model is the better of the models. This is likely because AIC favors models with fewer parameters and higher explained variation percentages, so although both models explained similar amounts of variation (R^2 of 0.5196 vs 0.499), the subset model used only 3 predictors vs the full model's 9 predictors.

Random Forest:

Random Forest regression models are excellent methods for prediction, partially because the responses are averaged over many trees, greatly reducing variation. One drawback of random forests is that it is difficult to interpret the model, though it can be done through methods like Feature Importance scoring.



Note how Temperature and Solar Radiation are still considered very important, but rainfall is not. This may be due to solar radiation and temperature being sufficient to describe rainfall (e.g. if it is colder and minimal solar radiation, then it may be raining, but if it's sunny, it is likely not raining).

This process was repeated to create a total of four forests. The hyperparameter ntree was tested by increasing from 5000 to 10000, while the number of features used in the regressions was also changed. In the table below, models with the "Subset" tag only used the features found to significant during simple linear regression.

Number of Trees	Features Used	MSE	R^2
5000	Full	0.472	0.524
10000	Full	0.471	0.525
5000	Subset	0.496	0.499
10000	Subset	0.494	0.502

This table summarizes our results. The option with the highest R^2 value is ntree = 10000, features used = Full. However, all models performed similarly, and therefore more simple models may be more desirable. For

instance, using ntree = 5000, features used = Subset would greatly reduce computation time and data management time while also explaining approximately 50% of the model's variation.

Elastic Net:

Elastic Net is a type of linear regression where parameters are eliminated or weighted based on their importance. We used a cross validation method to tune parameters alpha and lambda. The resulting solution was alpha = 1, lambda = 0.01069142, giving an RMSE of 0.714 and an R^2 of 0.484. The alpha of 1 indicates that the elastic net is strictly lasso regression (as opposed to incorporating Ridge Regression). Consequently, as the elastic net performed best when tuned to be lasso regression, the model minimized certain coefficients to 0 to reduce the feature space/dimensionality which was more effective than Ridge Regression's approach of minimizing certain coefficients but not to absolute 0. However, the losses in R^2 accuracy do not justify the reduction in dimensionality; the accuracy of other models such as linear regression is not high enough to justify trading accuracy for dimension reduction.

The model found that bike rentals had a positive correlation with date, temperature, wind speed, and solar radiation and a negative correlation with humidity, snowfall, and rainfall.

Poisson Regression:

The final model we used for this project was a Poisson regression. The response variable has Poisson distribution [# of bikes per hour]. Contrary to all the other models, Poisson regression found all features to be statistically significant. To confirm that variables have predictive power, we performed the Wald test to confirm that the other variables have predictive power. The P-value of the Wald test was 0, so we reject the null hypothesis and conclude that variables other than Temperature, Rainfall and Solar Radiation also have statistically significant predicting power. Based on this model, bike rentals are positively correlated with location Seoul, humidity, temperature, and solar radiation, and negatively correlated with dew point temperature, snow fall, wind speed, and rainfall. Unfortunately, R^2 is not applicable for this type of model, so directly comparing this model to the other three will present challenges. There is a pseudo-R^2 (Mcfadden's R^2) that calculates to 0.791, but comparing normal R^2 to a pseuodo-R^2 is not a good practice.

Model Comparison:

Model	R^2	Pseudo-R^2
Linear Regression	0.499	NA
Random Forest	0.525	NA
Elastic Net	0.484	NA
Poisson Regression	NA	0.791

R^2 did not vary much from model to model. Across all the models used, even considering the lower performing parameter choices, the range of R2 values goes from 0.474 to 0.525. These results are realistic given the problem reflects large scale human behavior. Each model explains about half of the variation in the data, but there are likely additional factors that drive ridership on a daily basis (such as weekday vs weekend).

Although the Random Forest model performed the best of the R^2 regression models, it is also the hardest to interpret. If its performance is due to random chance, the linear regression model may be the superior choice, as it only uses 3 features, and it is much simpler to make business decisions from it. For instance, if we looked at our linear regression, we could say that moving our bike rental services to a warmer climate would be beneficial since the number of bike rentals is correlated with the temperature. However, for the random

forest model, we could say that temperature is an important factor (based on feature importance), but it's harder to say whether a warmer climate is better than a colder one.

The Poisson regression's Pseudo-R^2 value cannot be directly compared to actual R^2 values, but because it is based on a comparable dataset, it should not be entirely dismissed. Additionally, because Poisson regression models are designed specifically for count data, and its output is easy to interpret, we believe the Poisson regression model is the best choice to solve this problem.

Novelty of the Approach:

Our team leveraged a span of different model types including traditional linear regression (suited for the assumed linearity of the problem), regularization (elastic net), random forest (non-linearity), as well as a Poisson regression (leveraging the Poisson distribution's applicability to arrivals/arrival counts). Attacking the problem from several different angles, we could see certain approaches worked better - random forest indicating non-linearity and Poisson regression reflecting an inherent Poisson distribution in the problem. We could also see regularization struggling, suggesting that this problem was less suited to dimension reduction.

Business Recommendation:

Based on this model, we can draw the following conclusions when trying to set up our bike rental company. Assuming all else equal, we should consider setting up in Seoul instead of New York. We should aim for places with a low dew point, infrequent snowfall, slow wind speeds, and little rainfall. We should also aim for places with high humidity, high temperatures, and lots of sun. Cultural factors (represented in the location feature: NYC/Seoul) had a more minimal impact compared to weather; weather should be the primary driving factor in identifying high-yield bike rental markets in major dense population centers (with less consideration to cultural propensity to ride bikes).

Works Cited:

Sathishkumar, V. E., & Cho, Y. (2020). A rule-based model for Seoul bike sharing demand prediction using weather data. European Journal of Remote Sensing, 53, 166-183. doi:https://doi.org/10.1080/22797254.2020.1725789

Pan, Y., Zheng, R.C., Zhang, J., & Yao, X. (2018). Predicting bike sharing demand using recurrent neural networks. IIKI.

Wang, X., Cheng, Z., Trépanier, M., & Sun, L. (2021). Modeling bike-sharing demand using a regression model with spatially varying coefficients. Journal of Transport Geography, 93, 1. doi:https://doi.org/10.1016/j.jtrangeo.2021.103059