Introduction to Data Analytics & Discovery Informatics
Project Report

Divvy Bike Demand Estimation Based on Weather Conditions in Chicago

Team Members
Naga Venkata V Vinnakota (nvv13)
Pranav Shivkumar (ps1029)
Swapnil Kamate (sk2181)

Table of Contents

Introduction

In today's world, there are various types of transportation options available to commuters daily. Most of these commuters use public transportation, like trains or buses to commute to work or other places, especially in densely populated regions. However, the problem arises when one must commute to a nearby place where a train station or bus stop is not available. In such cases, biking is a viable option. But what do commuters do if they do not own a bike? In such cases, bike sharing becomes a viable option.

In North America, bike sharing [1] tends to be affiliated with the municipal governments, or, in most cases, with colleges or universities to facilitate the travel of college students to the campus. For this project, we are focusing on a particular bike sharing system, the Divvy bike sharing system owned and operated by the Chicago Department of Transportation.

This project aims to understand and analyze the impacts of external factors on the Divvy bike demand. The factors that we are considering for our analysis are the weather conditions, seasonal changes, and holidays. The utilization of these bikes is not only restricted to transportation to work or college campuses; subscribers can also use them for personal purposes, like entertainment. In one way or the other, these bikes tend to have a high demand due to the above-mentioned reasons.

These shared bikes are usually docked at "stations", which are essentially high-tech bike racks consisting of several docks, where the bikes are usually located. Depending on the demand for these bikes, the number of docks will need to increase/decrease. Determining the demand manually is a very tedious process as the bikes that are checked out and returned are highly random and the bikes returned at a station can be from various stations.

To facilitate this, our project also aims to estimate the demand of these shared bikes per hour to meet the demands of the users. Doing so will also help increase the profits for the owners of these shared bike systems since the profits for the organization rely on customer satisfaction. We determine this estimate by using and analyzing various regression models and selecting the model that provides the best fit for our data.

Motivation

The major motivation for this project stemmed from the arrival of the Veo electric scooters in the Rutgers New Brunswick campus. We initially decided to predict the demand of these scooters, however, we were unable to find sufficient data to work with; the data size of the Veo bikes were only available for the summer. We decided to look for other similar options and chanced upon the Divvy bike data in Chicago. This turned out to be an incredible motivation for us, since Chicago is a city with high bike usage and has diversity in climatic conditions, which would help us in our analysis.

Initially, we planned to only use the weather conditions as a factor that would help us analyze the demand of these bikes. We figured that the weather would play a major role in the demand of these bikes, since if the weather conditions are pleasant, it would facilitate people to commute through these bikes. However, if the weather conditions are harsh, like a thunderstorm or a blizzard, it will be very difficult and hazardous to travel using these bikes. As mentioned before, Chicago is also known for its diverse weather conditions, prompting us to proceed with this as the primary factor in our analysis.

However, on further analysis, we found that other factors can impact the bike demand; these factors include the seasons and whether the day is a holiday or a normal working day. The reason behind taking these factors into consideration is that since Chicago is a densely populated city, the holiday factor would increase in many bike checkouts. Also, the bike checkouts would increase if the season is summer or fall, attributing to the pleasant weather conditions as well as the fact that most schools will be on vacation then. Since the holiday factors overlap with weather conditions as well as seasons, we decided to factor these into our analysis.

Literature Review

We mainly referred to these three machine learning papers for our literature review.

1. Short-term prediction for bike-sharing service using machine learning.
2. Regression Model for bike-sharing service by using machine learning
3. Predicting bike sharing demand using recurrent neural networks
4. Use of deep learning to predict daily usage of bike sharing systems.

The first paper was published by Bo Wang and Inhi Kim, who are professors from the Department of Civil Engineering, Monash University. This machine learning paper mainly focused on short-term prediction for a particular station. This was also a motivation behind our analysis as well. Two latest and highly efficient models, LSTM (Long Short-Term

Memory) and GRU (Gated Recurrent Units) were adopted to predict the number of short-term available bikes in a particular station. Random forest was used as a benchmark for the comparison, which we also used in our analysis. We calculated the total number of bikes checked out from a particular station which helps to determine the busiest stations and more information about them. In the paper, the results showed that both the models as well as the random forest algorithm were able to achieve good performance with acceptable errors and comparatively high accurate results. The docking bike-sharing system in Suzhou, China was chosen for this study. The data has 46,000 bikes with 2000 stations across the city. The paper used 95% of all the records as a training set and the remaining 5% for the testing set. Both the algorithms are run using the Keras and TensorFlow libraries with GPU. In the prediction, the Random forest algorithm and both the models show good trends. The difference between the real data and prediction data is less than 1 or 2 bikes. This motivated us to work on getting results for checkouts from each station. The output data in this paper only has one-time step, but multiple time steps need to be done. The paper did not include any of the external factors such as weather, day of the week, holidays etc. which led us to include these factors in our analysis, and hence incorporating different data to get good results.

The second paper, Regression Model for Bike-Sharing Service by Machine Learning, was published on November 4, 2019 by Zhifeng Wang from the Department of Engineering and Computer Science, Australian National University. This paper covers some of the most important points, starting from visualization technology, to visualize data and using the regression models, neural network models, ELM models and DELM models to predict the possible number of bike users. This paper used a dataset from the study case chosen from Capital Bikeshare System, Washington D.C., USA. The data used in the paper utilized data that was like the data that we used; the difference was that, for our project, we merged three different datasets which helped us to know more about the data and bike-sharing systems and understand how their incorporation in the data affects the original dataset. The dataset in the paper utilized important attributes related to weather, seasonal and bike data which led us to use these ideas for our project. To understand the impact on the number of bikes, the author used bar plots, box plots and scatter plots to get an idea of how the data looks. The author also used the correlation between the season and the number of users. Along with other models, the papers also built a multiple linear regression model which modelled the correlation between two or more variables and a response variable by fitting a linear equation to observed data. The idea motivated us to try different regression models in our projects. To reduce the standard error of the variables in the linear regression model, the paper used the AIC method to remove some variables in the linear regression model to get the best model. The author also spoke about improving the accuracy of predicting the total number of bike usage, by using multiple neural networks. The author also suggested to consider the time factor and previous total number of users for future research, which will give better prediction results.

The third paper, Predicting Bike Sharing Demand using Recurrent Neural Networks, was published by students of Renmin University of China in 2018. This paper proposed a real-time method for predicting bike renting and returning in different areas of a city during a future period based on historical data, weather data and time data. The prediction model in this machine learning paper is limited to finding two communities with the most demand for shared bikes, and they used the data of stations in the two communities as the dataset and trained a deep LSTM model with two layers to predict the bike renting and returning. They evaluated the model with root mean square error (RMSE) of the data and showed that the prediction of the proposed model outperforms that of other deep learning models, based on the comparison of their corresponding root mean square error values. The main motivation of this paper is about the use of LSTM sequence learning models for their methodology. It can process sequential data and memorize data of past time steps. Since the RNN's are computationally expensive, the author has used only two layers in the model. They used the Citi bike system data of 2017 as the training set and used data of the first 3 months of 2018 to the test set to conduct the experimental study. This gave us an idea to use a complete year data for our project. Our project had nearly 3.8 million rows of data which gave us good insight into data processing and modelling. They have used deep neural networks to predict the result as well, the root mean square error for the training set was 3.6 and for the test case it was 2.7. Considering the number of docks in each station, the numbers were affordable. The results show that the model fits well with two layers of LSTM which implies that LSTM is better at predictions with sequential data than DNN. The author also suggested that we can get the net demand by calculating the difference between the number of rents and returns. Since we can get proper results about each station, the prediction model can make suggestions for bike companies on how to distribute the bikes specifically to each station to satisfy the customer demand as well as saving unnecessary cost of keeping bikes. This is a good idea for the future scope of our project.

Approach

Data Collection

The first step to get started with our project was to collect the data needed for the analysis and the estimation. For this, we are collecting data from three separate sources:

1. **Divvy Bike Dataset**

   The datasets for the Divvy bikes were readily available as zip files in the official Divvy website: [Index of bucket "divvy-tripdata" (https://divvy-tripdata.s3.amazonaws.com)](https://divvy-tripdata.s3.amazonaws.com). We decided to use the data from the last quarter of 2019, which included the data for the months from October 2019 to December 2019, the first quarter of 2020, which included the data from January 2020 to March 2020 and the monthly data from April 2020 to October 2020. After this, we read each of the datasets into a Pandas data frame and merged all the datasets together.

2. **Weather Dataset**

   Unlike the Divvy dataset, we couldn't find any readily available datasets for the weather data by the hour of every day, so we had to do some web scraping from the Wunderground website: Chicago, IL Weather History | Weather Underground (wunderground.com). We coded up a Python script that uses Selenium web drivers, specifically the Mozilla web driver, to return the monthly data and append it to a csv file. In this way, we obtained the complete weather data.

3. **Holidays Dataset**

   As for the holidays dataset, we found a dataset available on this website: Download Excel File List of Public Holidays in US | Download Excel Files. We got a list of holidays for the 2020 year along with the description of each holiday.

Feature Selection

As can be seen from all the dataset images, each dataset consists of several attributes, both categorical and numerical. We had to select certain attributes that would be useful for our analysis and predictions.

For the divvy data, the features that we selected were:

- **ride_id** - This is the unique ID assigned for each ride on the bike.
- **started_at** - This gives the time the ride began.
- **start_station_name** - This attribute gave the station where the ride began.
- **start_station_id** - This attribute gave the ID of the station where the ride began.
- **ended_at** - This gave us the time when the ride ended.
- **end_station_name** - This gave us the name of the station where the ride ended.
- **end_station_id** - This gave us the ID of the station where the ride ended.
- **member_casual** - This feature defined whether the rider is a subscriber or staff.

For the weather data, the features we selected were:

- **Date**
- **Temperature**
- **Humidity**
- **Wind Speed**
- **Wind Gust**

- **Precipitation**
- **Condition**

For the holiday data, the features that we selected were:

- **Unnamed Column** - The date of the holiday
- **Official Name** - The name of the holiday
- **Remarks** - The description of the holiday

Data Merging

Once we got all the data collected and selected the features to use, the next step was to merge all the datasets into one dataset. We observed all the three datasets and discovered that the started_at attribute in the divvy dataset, the Date attribute in the weather dataset and the Unnamed column in the holiday dataset had the date in common, which could be used for merging.

However, we were posed with the problem that the dates in all the three datasets were different timestamps, which would significantly reduce our data if merged. To bypass this, we converted all the date columns in the three datasets into a datetime format, which allowed us to extract the individual year, month, day, and hour from the timestamp and create each as a separate attribute, which we used to merge the datasets.

Data Preprocessing and Cleaning

As seen in the images of the datasets, there exist missing values, duplicate values, and various types of attributes that needed to be addressed. The first of these, as mentioned above, was the Date attribute in all three datasets which was handled by converting the attribute into a datetime format and splitting it into three numerical attributes. By doing this, we were able to merge the divvy and weather data together to create the partial merged data. However, the major challenge was to preprocess and clean the holiday dataset, and this dataset took more than 75% of our time to process the data.

Starting off, the unnamed column (Date of the holiday) had to be named, so we named that to “Date”. The data included in this dataset only included the federal holidays, like Thanksgiving, Christmas, New Year’s Day, Labor Day, Martin Luther King’s Birthday, etc. Since our aim was to analyze bike checkouts using different types of holidays, we decided to include the school vacations as well, like spring break, summer vacation and winter break. In addition to this, we decided to factor in whether a particular day is a weekend or not. As a result, we had to incorporate this as well into the holiday dataset.

On adding these to the dataset, we ran into another problem, which was to prioritize the holiday. For example, Christmas falls in between the Winter Vacation, or one day in the Winter Vacation is a Weekend. In order to handle this problem, we assigned a priority to the holidays that were initially present in the holidays dataset. This meant that if we inserted any holiday into the dataset, it would check if that date is already present in the dataset and if it is not, it would assign that holiday to that date. Once this was done, we faced another issue. When we merged the current dataset to the divvy and weather merged dataset, there were a huge number of NaN or missing values for most of the dates. We quickly realized that this was due to the other dates not being a holiday. We figured that if we proceeded in this fashion, it would create further issues in our prediction, so we decided to handle this issue.

To handle that issue, we created a data frame with all the dates from November 2019 to October 2020. We then checked if the date was a holiday or not, and if it was not a holiday, we assigned that date with the tag 'Not Holiday'. In this way, the number of missing values was drastically reduced. Finally, we merged this dataset with the already merged data and obtained a final dataset consisting of 3.8 million rows and 21 columns.

Most of the data cleaning had to be done while merging, including handling the missing values in the holidays dataset. However, we noticed that there were several duplicate and missing values in the merged data, which could be attributed to the divvy and weather datasets. To deal with the duplicate values, we decided to keep the first occurrence of the duplicate values and drop the remaining values. To deal with the missing values, we dropped the rows consisting of NaN values. We obtained the cleaned dataset consisting of 3.3 million rows and 20 columns. Finally, we got the number of Divvy bike checkouts by counting the number of rides on a particular hour on a particular day and created this into a separate data frame, which would be used for our exploratory data analysis.

While going ahead with the regression, we realized that datetime attributes and categorical attributes could not be used. As a result, we had to encode these attributes into numerical attributes that could be used by the regression models. To do so, we first converted the date into seconds using the local time tuple. Next, we had to encode the holidays and seasons since they were categorical variables, for which, we used One Hot Encoding for each of these attributes.

Exploratory Data Analysis

To get an idea of the data that we were dealing with, we decided to check how the factors affected the bike checkouts. We wanted to also see how drastically the data analysis would vary for both the merged data that was not cleaned (final_merged_data) and the cleaned dataset (cleaned_data).

Regression Models and Evaluation Metrics

Once we finished with the exploratory data analysis, we had to predict the bike demand and to facilitate this, we decided to use the following regression models:

- Multiple Linear Regression
- Lasso Regression and Ridge Regression
- Ridge Regression
- Decision Tree
- Random Forest

We fitted the above regression models with our dataset and determined the model that best fits with the data. We evaluated the models by using the mean absolute error (MAE), mean squared error (MSE) and R-squared (R2) score. The MAE helps us determine the absolute error between the prediction and the actual data and understand how well the data that we collected fits with the data that we used in the analysis.

The MSE provides the average of the squares of the errors and will be low if the prediction results are close to the actual results. Using these metrics help us determine if the test data fits properly with the generated model. Although both MSE and MAE have the same use i.e., understanding how well the data fits the generated model, as the value of MAE reduces, the fit will be towards the median and might be biased. The other metric we used to evaluate our techniques is R2 score. The R2 score helps us understand how well the generated model fits the data. The higher the value, better the fit. Using these metrics, we were able to decide the model that provided the best fit for our data.

Results

For the exploratory data analysis, we plotted various graphs to see how the below factors listed below affect the number of bike checkouts for both the uncleaned and cleaned data:

- Seasons
- Month
- Day
- Hour
- Holiday
- Weather Conditions

These plots are shown below:

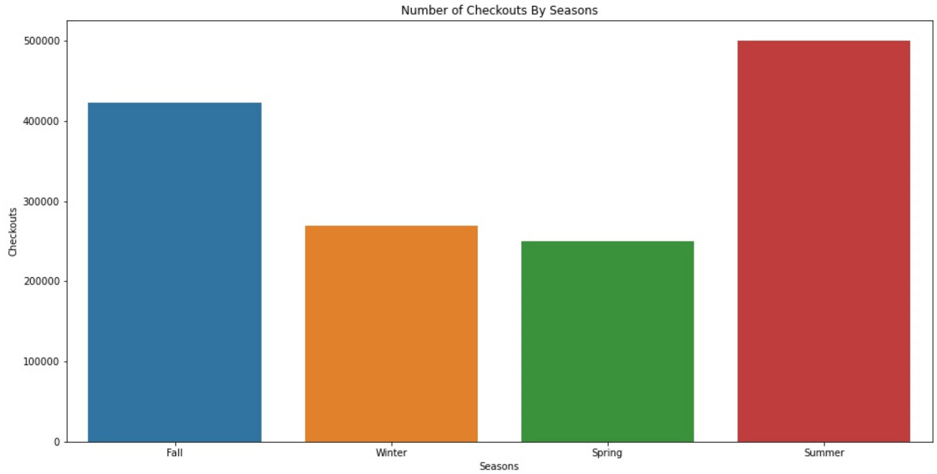


Figure 1: Number of bike checkouts based on Seasons

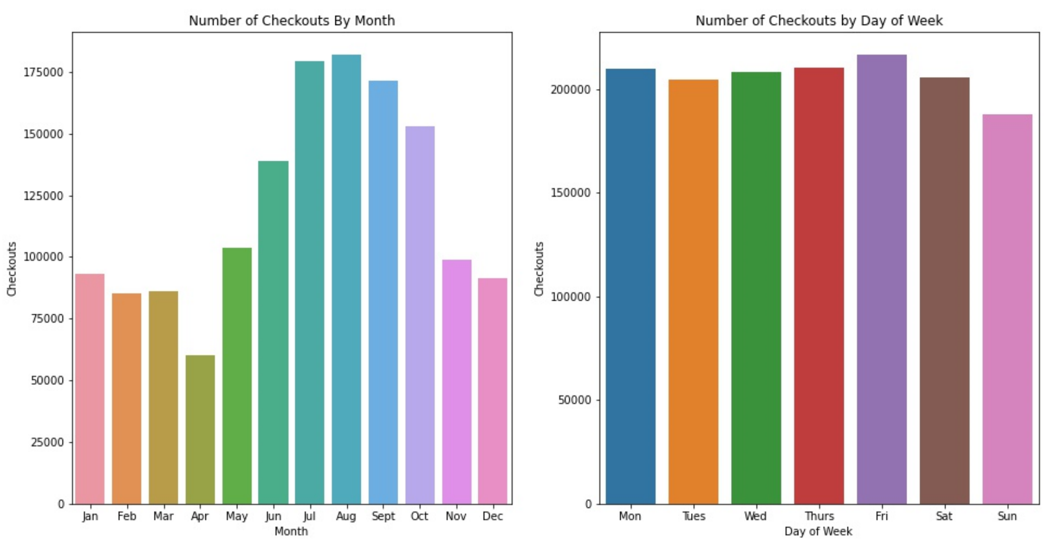


Figure 2: Number of bike checkouts by month (L) and day (R)

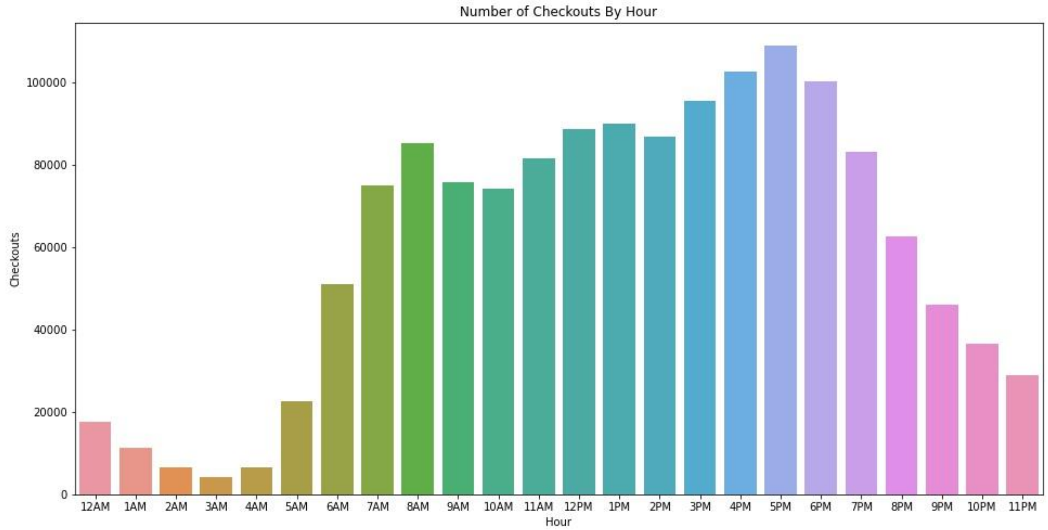


Figure 3: Number of bike checkouts based on hour

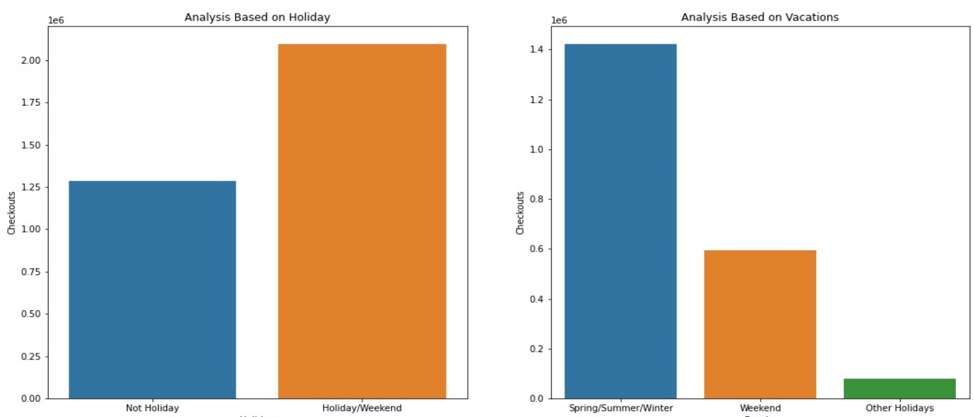


Figure 4: Number of bike checkouts based on holidays (L) and vacations (R)

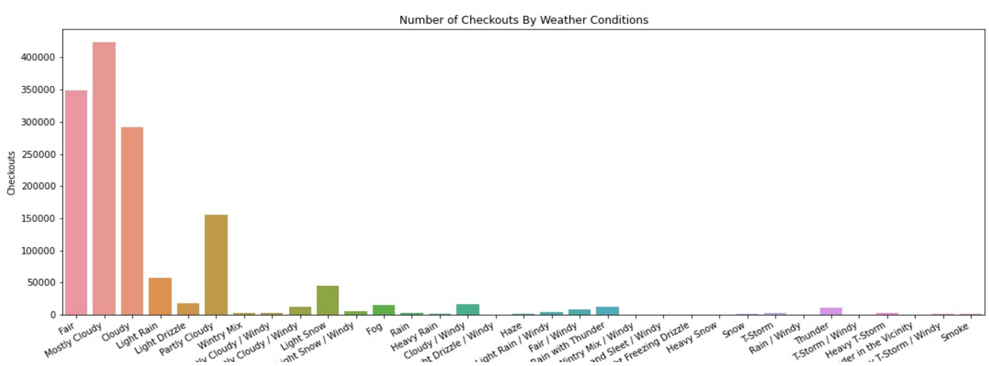
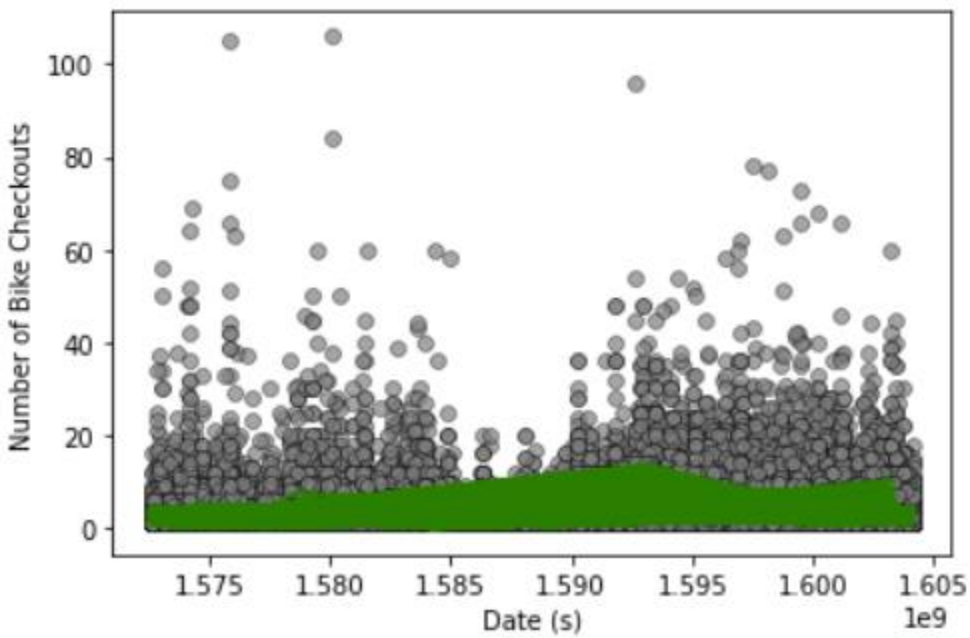


Figure 5: Number of bike checkouts based on weather conditions

The merged dataset without cleaning consisted of 3,807,058 rows and 30 columns, while the cleaned dataset consisted of 3,384,105 rows and 30 columns, implying that there were 422,953 rows with missing data or duplicate values.

For the regression models, we fitted each model with the dataset and used the evaluation metrics to decide which model fits the best with the test data.

For linear regression, the regression line is shown below:



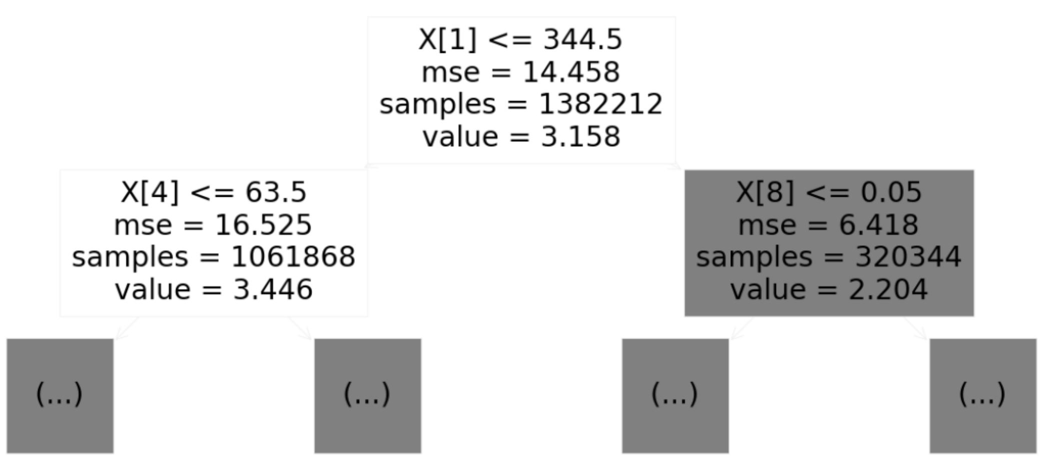


Figure 6: Regression and Decision Tree of depth 1 for unclean data

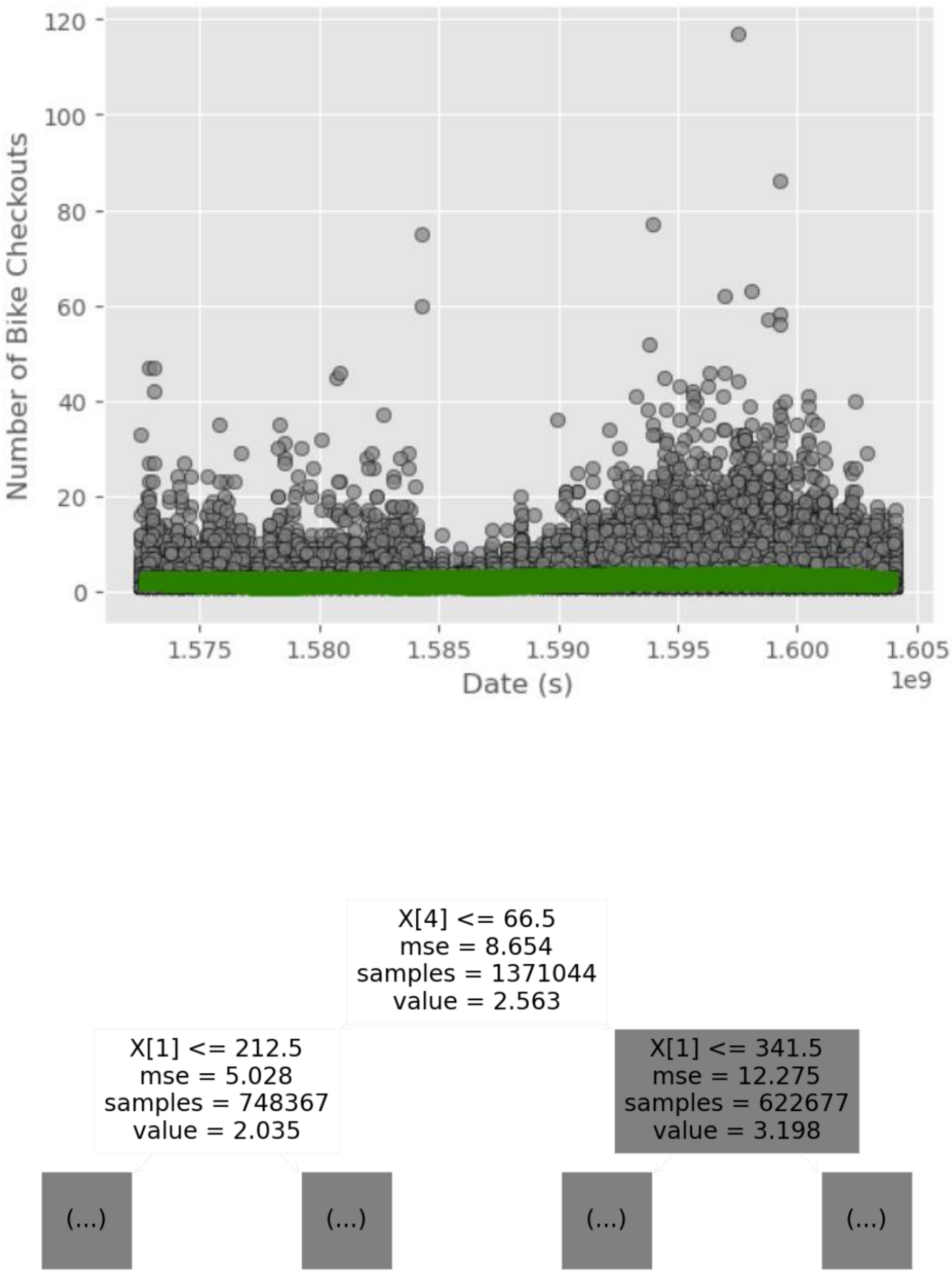


Figure 7: Regression Line and Decision Tree of depth 1 for Cleaned Data

The evaluation metrics for the uncleaned and cleaned data are shown below:

Regression Model	MAE	MSE	R2
Linear Regression	2.105669920368694	13.461859207637687	0.0587081771432435
Lasso Regression	2.1271054955466506	13.649793524921492	0.04556727042561515
Ridge Regression	2.1056718011158204	13.461860052578142	0.05870811806258247
Decision Tree Regression	1.5340421729807006	9.153767113872547	0.3599423378507213
Random Forest Regression	1.386015938582504	5.642355122522269	0.6054703398271429

Table 1: Evaluation Metrics and Results for uncleaned data

Regression Model	MAE	MSE	R2
Linear Regression	1.64111312749547	8.13229914507706	0.0837278807371582
Lasso Regression	1.6428767879709751	8.210707979426987	0.07489349976603654
Ridge Regression	1.6411126271350989	8.132297904458628	0.08372802051854811
Decision Tree Regression	1.379935145022935	6.917659816244232	0.22058218628042836
Random Forest Regression	1.1582206585340582	3.8975152270248037	0.5608640959688438

Table 2: Evaluation metrics and Results for cleaned data

The sample of the output prediction for each model is shown below:

	y_test	y_pred
1330659	1	3.286413
1112994	5	3.870815
562734	2	1.038194
358212	1	1.423497
144939	1	2.159866
...	...	...
850690	1	2.565076
248902	1	1.268136
1253695	16	3.347901
1170611	6	3.288188
447567	1	1.987979

Figure 8: Linear Regression

	y_test	y_lasso_pred
1330659	1	3.225801
1112994	5	3.189319
562734	2	1.829852
358212	1	1.194044
144939	1	1.720799
...	...	...
850690	1	2.363965
248902	1	1.611620
1253695	16	3.284624
1170611	6	3.059778
447567	1	2.329147

Figure 9: Lasso Regression

	y_test	y_ridge_pred
1330659	1	3.286413
1112994	5	3.870800
562734	2	1.038192
358212	1	1.423469
144939	1	2.159853
...	...	...
850690	1	2.565075
248902	1	1.268129
1253695	16	3.347905
1170611	6	3.288179
447567	1	1.987968

Figure 10: Ridge Regression

	y_test	y_dec_pred
1330659	1	4.0
1112994	5	3.0
562734	2	1.0
358212	1	1.0
144939	1	1.0
...	...	...
850690	1	3.0
248902	1	3.0
1253695	16	9.0
1170611	6	6.0
447567	1	1.0

Figure 11: Decision Tree

	y_test	y_rand_pred
1330659	1	2.36
1112994	5	2.84
562734	2	1.52
358212	1	1.37
144939	1	1.42
...	...	...
850690	1	3.09
248902	1	2.08
1253695	16	9.88
1170611	6	3.79
447567	1	1.33

Figure 12: Random Forest

Discussion

On plotting the bar graphs shown in Figures 1 to 5, we were able to analyze the impact of the factors mentioned above on the number of bike checkouts.

In Figure 1, the graph depicts the variation of the number of checkouts vs the season. As expected, the number of checkouts is higher in the summer and fall, owing to the pleasant weather conditions as well as the vacations. Figures 2 and 3 denotes the number of checkouts vs the month, day, and hour. The checkouts in July are the highest since that month falls in the summer vacation time and as for the day of the week, they are the highest on Friday (but the other days are not far behind) and for the hour, towards the evening, many bikes get used. This could be attributed to the officegoers traveling back to their homes, or students roaming around after school/college. As for the holidays, the days where it is a holiday have a greater number of checkouts as expected, and the long vacations like spring, summer and winter break have many checkouts. Finally, for the weather conditions, the days where the weather is 'Mostly Cloudy' have a lot of checkouts.

From the values of the metrics shown above, we can clearly see that the Lasso regression gave the least R2 score and high MAE and MSE values, and hence showed the worst performance from all the models we used. In contrast, the ridge and linear regression gave us the same values for MAE, MSE and R2 score. From this, we can deduce that there is no effect by regularizing the data with a non-linear relationship. As expected, the regression analysis showed abysmal results for the dataset. This could be attributed to the non-linear relationship between the various attributes in the dataset. In addition, the regression line is nearly horizontal and non-linear for the uncleaned data. This is since the regression line was plotted by incorporating the one hot encoded vector for holidays and seasons, increasing the number of attributes. This is seen by the thick line in the regression plot. The line becomes more linear when operating on the cleaned dataset.

The decision tree regression showed a significantly better performance relatively low MSE and MAE values and high R2 score than the regression analysis. This is because decision trees can perform better on the datasets containing categorical variables and non-linear relationships.

The random forest regression gave us the best performance with an R2 score of about 0.5 and the least MSE and MAE values. The reason behind this algorithm providing a better performance than the decision tree algorithm is because the decision trees are pruned at every step to avoid overfitting, while the random forests contain fully grown unpruned trees to have better resolution of the feature space. Also, each tree from the random forest is learned from a random sample of data and a random set of attributes are considered for splitting at each node. In this way, random forests try to perform better than the other models that we worked on.

Now, to understand the effect of running these algorithms on duplicate and missing values and outliers, we ran our models on the unclean data to understand the effect of the duplicates. The regression models show significant improvement in the performance when the data is cleaned compared to the data with no cleaning done. All the metrics are improved i.e., MSE and MAE values are reduced by nearly 40%. The R2 score has been improved significantly for all the regression models. Although data cleaning has improved the performance of the regression models, their overall performance is still the lowest.

While the R2 scores have been improved for the regression models for the cleaned data, the same was not the case for Decision Trees and Random Forests. The R2 scores of these models have reduced compared to the R2 scores for the uncleaned data. However, the MSE and MAE of these models increased significantly, but their increase was not as significant as that of the regression models.

Based on the values of the evaluation metrics for the models for both cleaned and uncleaned data, we can see that cleaning has a significant effect on the performance of the regression model analysis, but the effect is not significant in case of Decision trees and Random Forests. This helps us understand the fact that having outliers, this implies that the presence of outliers and missing values affect performance of the advanced regression techniques like decision trees and Random Forest algorithms, which allowed us to appreciate the purpose of data cleaning.

Conclusion and Future Work

Based on the analysis and regression results, we were able to draw conclusions on the Divvy bike demand. First, the bike demand will be significantly larger during the summer, as seen in Figure 1 and in the left plot of Figure 2, since students have vacations at that time, and fall, since the weather tends to be pleasant in Chicago at that time. The demand tends to be comparatively lesser during the spring and winter, since the weather tends to be significantly harsher during these months, even though winter break and spring break fall in these two seasons.

Second, the bike demand tends to increase towards the evening of the day and is highest on Fridays. As mentioned earlier, this could be due to the officegoers leaving to their homes, or students biking around to enjoy the fresh air after their school/college classes get over. Third, the demand of the Divvy bikes tends to increase when the weather conditions are pleasant and not too harsh, as seen in Figure 5.

Linear regression did not provide us with favorable results, implying that the attributes share a nonlinear relationship, which is further corroborated by the fact that the tree algorithms like decision trees and random forest provided a much better result (low MAE, MSE and higher R2 score). In addition, the cleaned dataset provided a more linear line than the uncleaned dataset, although in both the cases, the regression line is thick, because many attributes were considered. The tree algorithms were less affected by the presence of outliers, missing values, and duplicate values, as seen in the lower R2 score compared to the unclean data compared to their effect on the regression models.

As for the future work of this project, we had a few ideas. We extend the analysis to using more powerful models like neural networks (LSTMs, GRUs) as seen in the literature review. Another idea is that we could incorporate a classification algorithm to categorize the user of the bike as a subscriber or a guest user. Another idea is that we could incorporate a Google Maps API that could compute the fastest route between the start point and the destination. This could be used in conjunction with this project to get the demand of bikes based on the start and end point.

References

1. What Is Bike Sharing? How Bike-Share Programs Work, Pros & Cons (moneycrashers.com)
2. https://www.sciencedirect.com/science/article/pii/S2352146518303181
3. https://www.researchgate.net/publication/326542794_Use_of_Deep_Learning_to_Predict_Daily_Usage_of_Bike_Sharing_Systems
4. https://www.sciencedirect.com/science/article/pii/S1877050919302364
5. https://www.researchgate.net/publication/337062461_Regression_Model_for_Bike-Sharing_Service_by_Using_Machine_Learning/link/5dc307b1299bf1a47b1bf470/download
6. https://divvy-tripdata.s3.amazonaws.com/index.html
7. https://www.wunderground.com/history/daily/us/il/chicago/KMDW
8. https://www.downloadexcelfiles.com/us_en/download-excel-file-list-public-holidays-us#.X9OXEWhKhPZ