BIKE SHARING DEMAND PREDICTION

Muhammet Emin Turan
Nesibe Betül Döner
Emre Bağbakan
Ahmet Serhat Strazimiri

June 2022

1 Introduction: Problem Statement and Goal

The bicycle sharing system could be a benefit where bikes and diverse sorts of bicycles are arranged for common utilize within the brief term for a charge. The bicycle sharing framework permits individuals to choose up a bicycle from one parking lot and return it to the other parking lot within the city, as long as the two parking lots have a place to the same system. So, numerous bicycle share systems permit individuals to borrow a bicycle from a "dock" which is more often than computer-controlled wherein the client enters the installment data, and the system opens it. This bicycle can at that point be returned to another dock having a place to the same framework. Also, in recent years many cities added electrical bikes as well as mechanical bikes in their bike sharing system. As of now, rental bicycles are presented in numerous urban cities for the improvement of mobilty consolation. The main reason why bike sharing systems are widespread in big cities is traffic. As an example of other reasons are transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals. It is critical to form the rental bicycle available and accesible to the public at the proper time because it reduces the holding up time. Inevitably, giving the city with a steady supply of rental bicycles gets to be a major concern. The vital portion is the prediction of bicycle count at each hour for the steady supply of rental bicycles.

The aim of this project is predict bike sharing demand under various circumstances (temperature, humidty, whether it is holiday or not etc.) by using Machine Learning. It is important to predict the demand especially in big cities because customers may refuse to rent if they wait too long which may cause both economical problems and customer dissatisfaction to bike sharing company and they may not amortized their bikes in time and lost the trust of their customer. Opposite to that factor, if demand prediction went wrong and company increase the amount of bikes in their fleet, they may waste their money with such unnecessary expenses. That is why our machine learning application is a life saver

for such systems.

2 Literature Search

In a span of few decade, the sharing of bicycle system has seen enamours growth (Fishman, 2016). Shortly, this system has changed people's daily transportation with ensuring them bicycle for common use. Amsterdam in Netherlands was the one where initial bicycle sharing system has started in 1965 (Shaheen, Guzman, and Zhang, 2010). Their main motivation was providing environment a green future, social welfare, and meet the needs of Dutch citizens because The Dutch are the people who use bicycle the most. Nowadays, the system spread all of the world especailly after 00s. Thanks to development of mobile applications, people are allow to know nearby bike station and can rent their bike from these type of applications. Till today there are more than 50 countries having 712 which implemented bicycle sharing method (Shaheen, Martin, Cohen, Chan, and Pogodzinski, 2014).

Ddareungi is a bike sharing system in South Korea, which started in the year 2015, known as Seoul bike in English. It was started to overcome issues like greater oil prices, congestion in traffic and pollution in the environment and to develop a healthy environment for citizen of Seoul to live. Han River is the initial place where Ddareungi was first started on October 2015 in Seoul, few months later, total number of bike sharing station touched 150 with as much as 1500 were there. In order to cover the entire people in Seoul, in 2016 there is a gradual incline in number of docking station. As large as 20,000 bikes were made available which was confirmed by Seoul Mayor Park won-soon. With the help of growing technologies, Seoul city is now equipped with 1500 bike renting station which are operational round the clock (Sathishkumar V E and Yongyun Cho 2020).

Studies are done to enhance the background information about the bike sharing systems, explaining from the starting era of bike sharing to the latest generation approaches (DeMaio, 2009). Fishman mentioned the research carried out on the categories of bicycle sharing system, consisting of the history of the documentation process, usage analysis, user relevance, mentioning the fact that some researchers used automated computerised devices for gathering data related to bicycle sharing system (Fishman, 2016). Investigates moreover proposed an dynamic open bicycle sharing issue based on day by day strategies comprising of frameworks, while an inventive procedure for overseeing rental bike stations and number of models to fathom the issue is additionally created (Raviv and Kolka, 2013). Prescient models for assessing the plausible run of total bicycle ask in a given geographic zone subordinate on data from drive to-work ponders are made (Barnes and Krizek, 2005). A factual demonstrate for foreseeing the amount of bicycles enlisted each hour, which included a number of variables considering the amount of supporters, the time data amid the week, the occasion of strikes or occasions, and the climate (temperature, measure of rain, humidty etc.) (Borgnat, Abry, Flandrin, Rouquier, 2009). Analysts too

inspected human versatility data subordinate on the amounts of bicycles open within the stations, utilizing this to recognize worldly and topographical designs of versatility interior the city and anticipate the amount of bicycles available in any given station minutes or hours in progress (Kaltenbrunner, Meza, Grivolla, Codina, and Banchs, 2010).

Future accessibility of bikes in the stations was predicted by examining the moment of a continuous time Markov-chain population model with time-dependent rates (Feng, Hillston, and Reijsbergen, 2017). The investigation was carried out to think about different climate conditions and worldly qualities, in station-level and system level examination traits. Within the viable station-level examination, analysts utilized a clustering technique to recognize get-togethers of stations with indistinguishable properties, considering the affect of temperature and mugginess by showing a temperature–humidity file and a warm wave marker variable. They moreover completed a framework level examination, demonstrating that particular factors had basic impacts at different events of day. Particularly, temperature variable, precipitation, and whether it was a workday impact influenced the rental bicycle request at particular events (Kim, 2018).

With Artificial neural systems, Back Proliferation Systems were proposed to broaden the scope of issues that can be dealt with and have since been utilized for figure in a wide assortment of locales. Counterfeit Insights and information mining methods have moreover been connected to the counterfeit neural arrange models to upgrade the precision (Chen, 2007). The main downside of utilizing manufactured neural organize is its complicated structure and computational fetched. Profound learning procedures moreover utilized in bicycle sharing request forecast. Even though various examinations have inquired about activity stream and open rental bicycle request forecast, as it were some have been focussed around the moment-based request in open bicycle sharing frameworks (Gao and Lee, 2019). Moreover, the ceaseless variety in bicycle sharing system is outstandingly eccentric and furthermore impacted by various outside components.

3 Dataset Description

Our dataset include weather information such as temperature, humidty, windspeed, visibility, dewpoint, solar radiation, snowfall, rainfall. In addition to weather information, dataset also include the number of bikes rented per hour and date information of city of Seoul. Attribute informations are Date as year-month-date, Ranted Bike Count as ranted bike each hour, Hour as time of the day, Temperature as Celsius , Humidty as

4 Methods

4.1 Exploratory Data Analysis

Before we start modeling, first getting an idea on how the number of bike rentals depend on the various features provided to us.

4.1.1 Seasons

Below are the plots that show the average bike rental (count) for seasons. Bike reservations are highest during the Summer (April to June) and Fall (July to September) season and least during the Spring season (January to March).

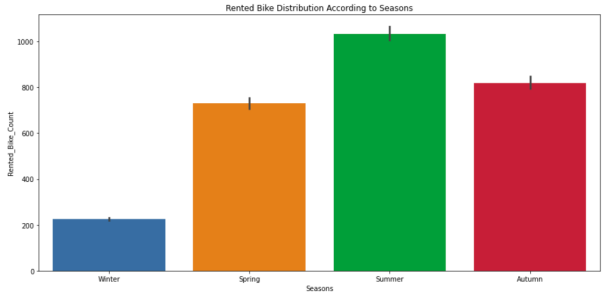

Figure 1: average bike rental across different seasons

4.1.2 Holiday

Below are bar plots and box plots of average bike rental counts on holidays and no holidays.

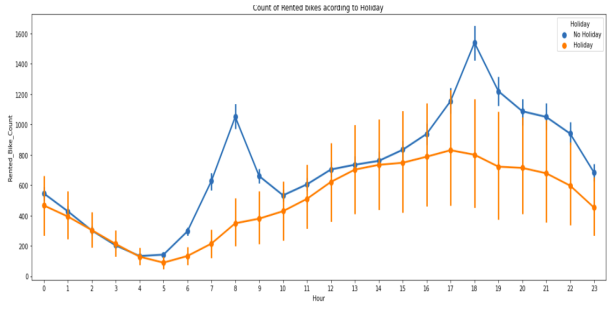

Figure 2: average bike rentals on holidays and non-holidays

We can see more outliers on no holidays.

4.1.3 Month

Below plot contains the average bike count over each month of a calendar year. The above figure is highly correlated with the seasons bar plot since seasons

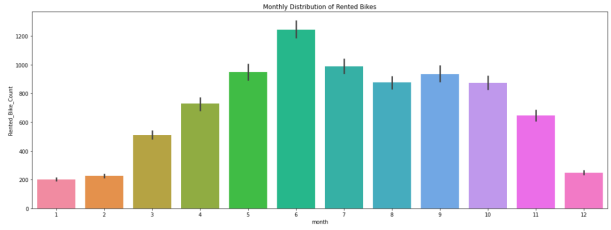

Figure 3: Average Monthly bike rental count

plot effectively is the average count for 3 of these months. We can see that the most rentals are in the months of June and July while least are on January and February.

4.1.4 Weekdays

Below are bar plots that show weekdays and weekends.

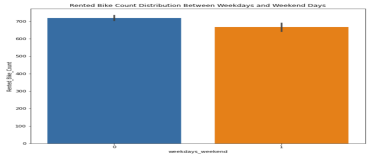

Figure 4: average bike rentals on Weekdays and Weekends

4.1.5 Hour

Now let us examine how the average bike count varies across the day (vs. hour) for the below four categories respectively; weekdays and weekends, functioning day, seasons, holidays etc.

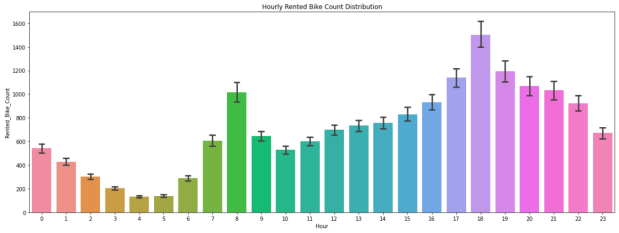

Figure 5: count of rented bikes according to weekdays and weekends

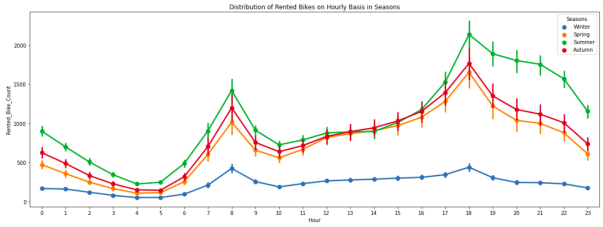

Figure 6: count of rented bikes according to functioning day

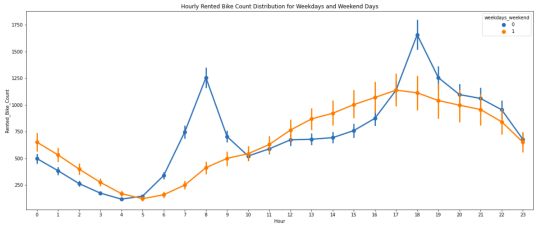

Figure 7: count of rented bikes according to seasons

4.2 Correlation Analysis

With correlation analysis, we can find the ratio of each variable to the rented bike count. We are doing this by using pandas profiling report. We can understand how much it affects the number of bike rentals whether it is dependent with the parameters(temperature, humidity, hours etc) or not. If it is correlated, what is its degree of correlation?

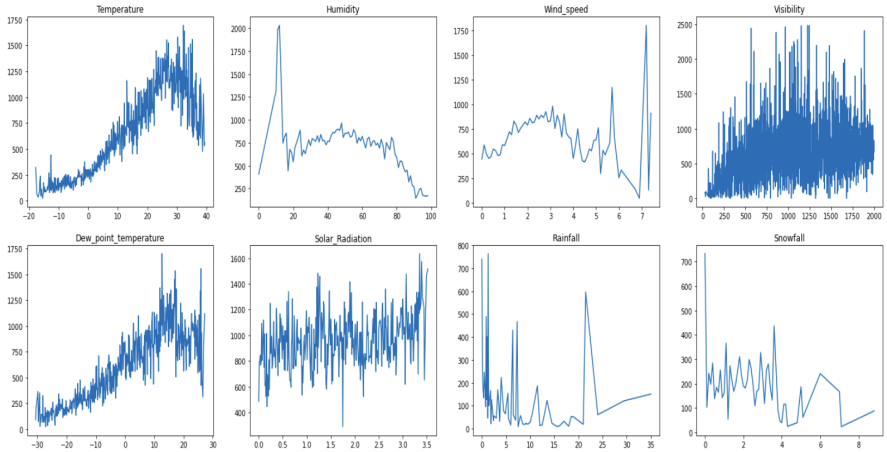

Figure 8: rented bikes distribution of numerical distributiion

4.3 Feature Engineering

The provided data in its raw form wasn't directly used as an input to the model. Several feature engineering was carried out where few features were modified, few were dropped, and few were added. Below is a summary of the feature engineering carried out with the provided data set.

1. The datetime column which contained the date-time stamp in 'yyyy-mm-dd hh:mm:ss' format was split into individual ['month', 'date', 'day', 'hour'] categorical columns.
2. OneHotEncoding of categorical feature set
3. drop date column: Intuitively, there is should be no dependency on date. Hence drop this column.
4. Drop windspeed column: Very poorly correlated with count and has several missing/erroneous data. Hence drop this column.
5. Drop atemp column: temp and atemp are very highly correlated and essentially indicate the same thing. Hence retain only the temp column
6. Drop holiday and day columns: The workingday column had information about holiday embedded in it. workingday = weekday and not a holiday. Since we noticed that there were two kinds of bike rental behaviors - during working days and not a working day, we will retain only the workingday column and drop 'day' and 'holiday' column.
7. Drop season column: This is because month column has a direct mapping with season (Winter: January to March, Summer: April to June, Fall: July to September and Spring: October to December). Hence, we retain month column due to its higher cardinality.

4.4 Train/Test Split

1. Training set
   (a) This contains data of from the 1 st to 17th of every month
   (b) This set is used to train various model and obtain the best set of hyperparameters for these models. We use GridSearchCV to tune the hyperparameters using this training set
2. Test set
   (a) This contains data of from the 18th to 19th of every month
   (b) This is used to evaluate all our models. The model with the best test score is finally chosen for submission

4.5 Modelling

Regression algorithms used. They are linear algorithms(linear regression, decision tree regressor), ensemble algorithms(random forest regression) and stacking algorithms(gradient boost, grid search CV) where predictions from linear and ensemble methods were used to make final predictions.

5 Experiment Results

	Model	MAE	MSE	RMSE	R2
0	Linear regression	4.407855	33.788208	5.812763	0.785453
1	Decision tree regressor	5.662666	63.739822	7.983722	0.595267
2	Random Forest regressor	2.228264	13.005861	3.606364	0.917416
3	Gradient Boosting	3.451114	21.116961	4.595319	0.865912
4	Grid Search CV	2.328538	11.940064	3.455440	0.924183

Figure 9: metrics

6 Discussion

We see that the r2 score is high in all of the models we tested. This shows us that the models capture the majority of the data variance. Thanks to this, we can store our findings and compare them. The results we obtained, especially in random forest and gradient boosting models, show us that the most important factors in people's rental bike preferences are weather events such as humidity and temperature. We combined these with the name of functioning day by generalizing. Another part is the hours of the day. The dominance of the preferences between 3-7 p.m. can be seen. Seasons and months have little effect. It seems that people are looking for more rental bikes during the hours when the humidity is high and the temperatures are suitable for sightseeing and after work hours. If we make a sociological evaluation of the work, the correct prediction rate of our models is quite high. In addition, although the predictions seem effective in general, especially in the linear regression model, it cannot be said that the problem we are dealing with is suitable for linear regression, and the reason for this can be shown as both the large number of outliers that appear and the fact that the predicted parts have no closeness with the actuals at some points.

7 Conclusion

To summarize our study in general, we first used the EDA method to understand the data in detail. Using EDA made our job easier and we interpreted the obscure relationships and trends well. We began by analyzing and transforming our dependent variable, 'Rented Bike Count.' Then, we looked at categorical data and eliminated those that had a predominance of one class. We also looked at quantitative variables, determining their correlation, distribution, and link to the dependent variable. We additionally hot encoded the categorical variables and deleted certain numerical features with primarily 0 values. We implemented algorithms that are linear regression, decision tree, random forest and gradient boost. It was remarkable that no overfitting was observed as a result of the applied algorithms. Besides, feature importance values in random forest and gradient boost were different. In addition, the r2 score in random forest regression was very high compared to the others (99). Because this data is time-dependent, it might not always be accurate. As a result, there will be period when the model does not work adequately. Because machine learning is a discipline that is rapidly growing, we must be prepared for any eventuality and periodically evaluate our model. It is known that to be successful in machine learning, it is not only to create a model, but also to update the model in the changing conditions in the real world and by obtaining new information. This project showed us this fact the most.

8 References

-DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. Journal of Public Transportation, - Fishman, E. (2016). Bikeshare: A review of recent literature. Transport Reviews, 36(1), 92–113. doi:10.1080/01441647.2015.1033036 - Gao, X., Lee, G.M. (2019). Moment-based rental prediction for bicycle-sharing transportation systems using a hybrid genetic algorithm and machine learning. Computers Industrial Engineering, 128, 60–69. - Kim, K. (2018). Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations. Journal of Transport Geography, 66, 309–320. doi:10.1016/j.jtrangeo.2018.01.001 - Shaheen, S.A., Guzman, S., Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: Past, present, and future. Transportation Research Record, 2143(1), 159–167. doi:10.3141/2143-20 - Shaheen, S.A., Martin, E.W., Cohen, A.P., Chan, N.D., Pogodzinski, M. (2014). Public Bikesharing in North America during a period of rapid expansion: Understanding business models. Industry Trends User Impacts, MTI Report, San Jose State University, 12–29. - Feng, C., Hillston, J., Reijsbergen, D. (2017). Moment-based availability prediction for bike-sharing systems. Performance Evaluation, 117, 58–74. doi:10.1016/j.peva.2017.09.004 - Raviv, T., Kolka, O. (2013). Optimal inventory management of a bike-sharing station. Iie Transactions, 45(10), 1077–1093. doi:10.1080/0740817X.2013.770186