DATA1030 Midterm Report

Guanjie Linghu

October 2021

1 Introduction

Bike sharing has been regarded as the most popular environmental-friendly traffic mode in recent years and developed all around the world, especially in developed countries. Therefore, governments or bike sharing companies need to know the number of bikes they should launch, which means they need to estimate the bike usage. It has traits of low usage barrier and easily being disturbed by external factors, such as weather. Therefore, The objective of this project is to develop regression models to make a prediction on the sharing bike usage amount in London given a set of 9 attributes.

My data set contains 17414 data and each data has the following 9 attributes: timestamp field, count of the new bike shares, real temperature, "feels like" temperature, humidity, wind speed, weather, season, and boolean attribute of whether that day is holiday or weekend. A detailed description of the data set can be accessed from Kaggle website given as reference[1].

There are several public projects based on this data set. Most of them are merely exploratory data analysis. In *Bike sharing prediction*[2], the author break down the timestamp into several distinct ordinal features and develop an XGBoost regressor without considering time-series property. It gives a good performance, an R2 score of 0.96. In *London Bike Share - Prophet, XGB, LSTM*[3], the author did a great job on developing models with four different algorithms and finally got the best result from the LSTM model, with an RMSE score of 414.

2 Exploratory Data Analysis

After carefully investigate each feature, I have already had a good grasp on the overall pattern of the data set as a time series and the relation between every two features. Another thing that I need to mention here is that to perform EDA thoroughly, I create four new ordinal categorical features, Month, Day of Month, Day of the week, and Hour. I will discuss it later in the Feature Engineer Section. This section will cover some interesting facts that I found by performing EDA.

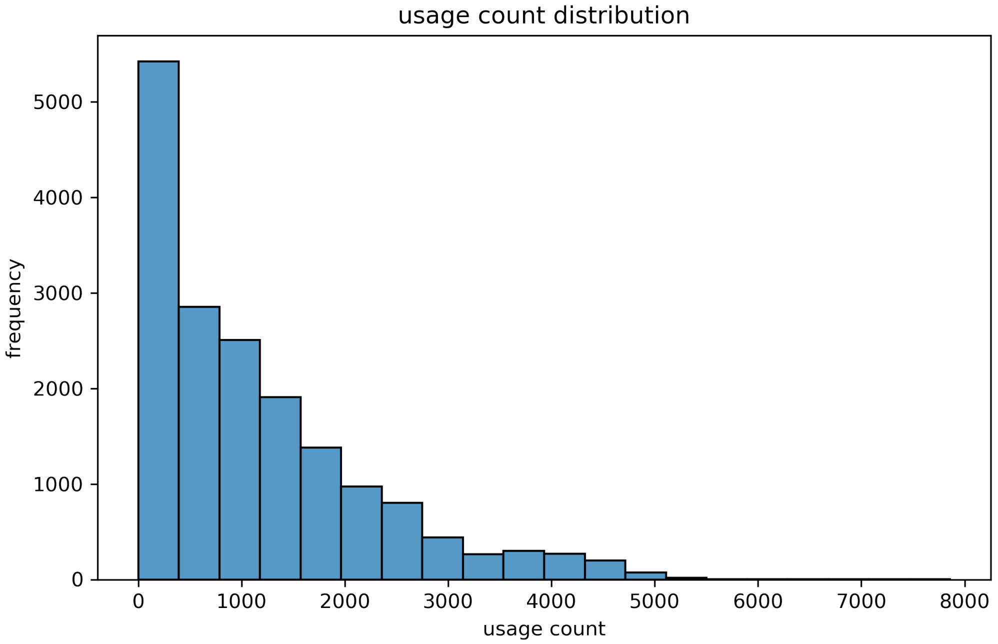
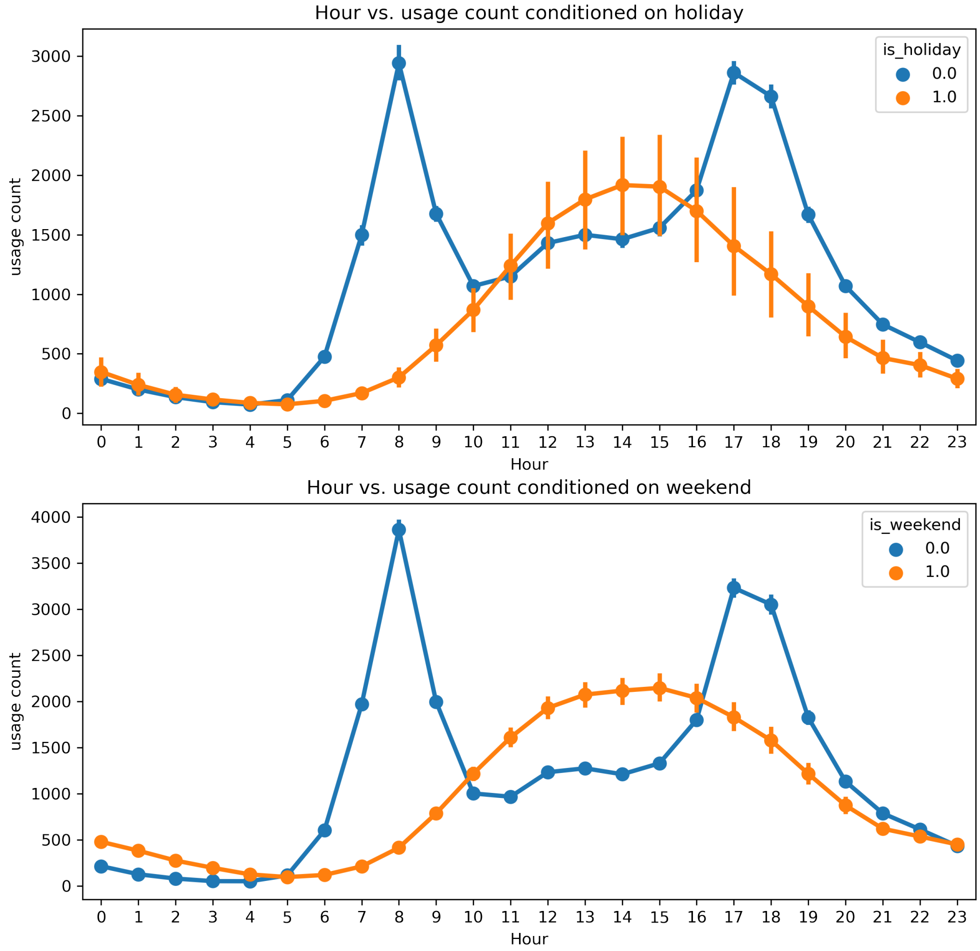


Figure 1: Distribution of target variable

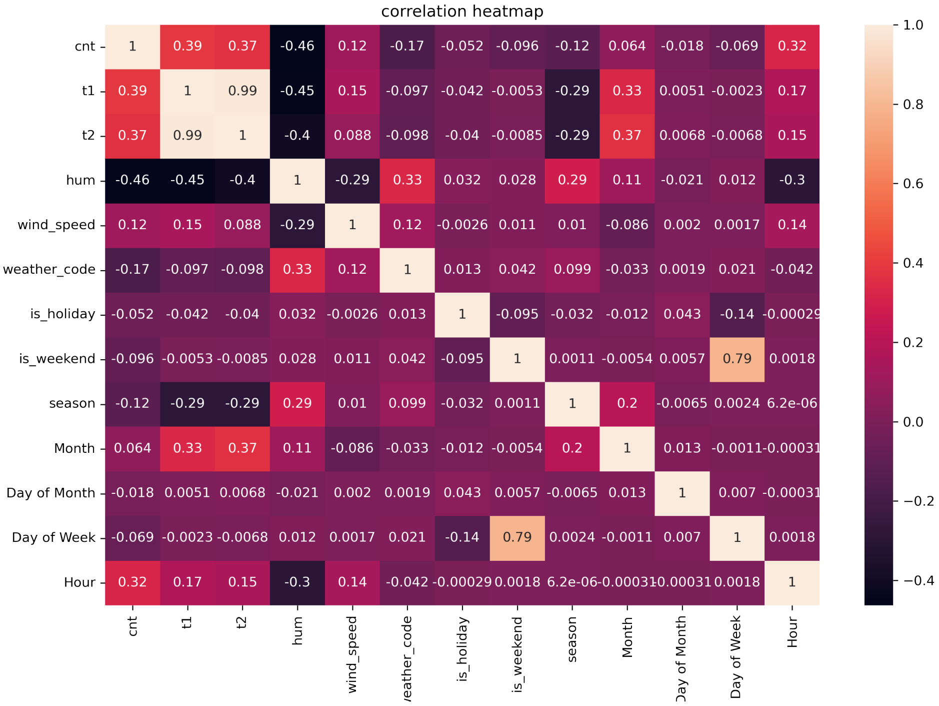
Figure 1 is the histogram that shows the distribution of the target variable ”cnt”. From this graph, we can see that the distribution is highly skewed to the right. Combined with the descriptive result, we have a mean of 1143. The first interval contains more than 25% of values. The maximum value is 7860.


Hour vs. usage count conditioned on holiday

Hour vs. usage count conditioned on weekend

Figure 2: Hour vs. usage count conditioned on holiday and weekend

Figure 2 includes two graphs that show the average amount of shared bike use each hour. Each point with a vertical bar represents an estimate of the central tendency for usage amount of each hour and provides some indication of the uncertainty around that estimate using error bars. The first graph is conditioned on feature is_holiday and the second one is conditioned on feature is_weekend. An interesting and useful fact is that during workdays, people tend to use the shared bike around 8 am and 5-6 pm, which are the rush hour traffic times. While during the break, people usually use the shared bike in the afternoon. It might be useful to extract the pattern as a new feature to improve the performance of machine models.



	cnt	t1	t2	hum	wind_speed	veather_code	is_holiday	is_weekend	season	Month	)ay of Month	Day of Week	Hour
cnt	1	0.39	0.37	-0.46	0.12	-0.17	-0.052	-0.096	-0.12	0.064	-0.018	-0.069	0.32
t1	0.39	1	0.99	-0.45	0.15	-0.097	-0.042	-0.0053	-0.29	0.33	0.0051	-0.0023	0.17
t2	0.37	0.99	1	-0.4	0.088	-0.098	-0.04	-0.0085	-0.29	0.37	0.0068	-0.0068	0.15
hum	-0.46	-0.45	-0.4	1	-0.29	0.33	0.032	0.028	0.29	0.11	-0.021	0.012	-0.3
wind_speed	0.12	0.15	0.088	-0.29	1	0.12	-0.0026	0.011	0.01	-0.086	0.002	0.0017	0.14
weather_code	-0.17	-0.097	-0.098	0.33	0.12	1	0.013	0.042	0.099	-0.033	0.0019	0.021	-0.042
is_holiday	-0.052	-0.042	-0.04	0.032	-0.0026	0.013	1	-0.095	-0.032	-0.012	0.043	-0.14	-0.00029
is_weekend	-0.096	-0.0053	-0.0085	0.028	0.011	0.042	-0.095	1	0.0011	-0.0054	0.0057	0.79	0.0018
season	-0.12	-0.29	-0.29	0.29	0.01	0.099	-0.032	0.0011	1	0.2	-0.0065	0.0024	6.2e-06
Month	0.064	0.33	0.37	0.11	-0.086	-0.033	-0.012	-0.0054	0.2	1	0.013	-0.0011	-0.00031
Day of Month	-0.018	0.0051	0.0068	-0.021	0.002	0.0019	0.043	0.0057	-0.0065	0.013	1	0.007	-0.00031
Day of Week	-0.069	-0.0023	-0.0068	0.012	0.0017	0.021	-0.14	0.79	0.0024	-0.0011	0.007	1	0.0018
Hour	0.32	0.17	0.15	-0.3	0.14	-0.042	-0.00029	0.0018	6.2e-06	-0.00031	0.00031	0.0018	1

Figure 3: Correlation Heatmap

Figure 3 is the correlation heatmap that gives the correlation between each feature. From the heatmap, we can see that feature "t1", "t2", "hum", and "Hour" has the strongest correlation with target variable "cnt".

3 Data splitting and preprocessing

In this section, I will discuss how I preprocess the data. First of all, my data set is time-series data, which is not IID data. This is easy to understand since, for a limited amount of available shared bikes in the city, the usage amount in the previous time period must affect the usage amount in the future period.

I keep the first 20 months of data as my training set, the next 2 months as a validation set, and the last two month of data as a testing set Since Standard Scaler is always a good tool to treat continuous variables, I use Standard Scaler to encode "t1", "t2", "hum", and "wind_speed", which are four continuous features. All the remaining 8 features are categorical and don't have an ordinal pattern, so I encode them by One Hot Encoder. After preprocessing, the data set has 93 features plus 1 target variable.

4 References

[1] Mavrodiev, H. (2019, October 10). London Bike Sharing Dataset. Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset.
[2] niks8411. (2021, March 29). Bike sharing prediction. Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/niks8411/bike-sharing-prediction.
[3] Maartenvandevelde. (2021, February 21). London bike share - prophet, XGB, LSTM. Kaggle.

Retrieved October 12, 2021, from https://www.kaggle.com/maartenvandevelde/london-bike-share-prophet-xgb-lstm.

5 Github repository

https://github.com/lah-dee-dah/DATA1030_project.git