Bike Share Analysis and Prediction

Mahammad Parvez Salim
Department of Computer Information System
University of Houston - Clear Lake
Houston, Texas
Salimm3027@uhcl.edu

***Abstract*— A rising number of bike-share business causes more competition among each other. Several bike-sharing companies fail to sustain in this competitive market, due to lack of predictive analysis of their business. This paper analyses data of the Los Angeles Metro Bike Share and successfully builds the predictive model to forecast future demand for bicycles. Each station of the company, located in a different region of the city; therefore, the demand varies for each station. Building a model to predict demand based on every station for a daily and hourly basis can effectively help the operational issue. Based on the usages of the stations, clustering can also help to reduce the number of stations. Los Angeles Metro Bike Share company has three types of pass holders. The study also finds the relationship between subscriber types with the distance and duration they cover. In this paper, the experiments try to solve the bike-sharing business problems using Data Mining approaches.**

Keywords— bike-share; predictive model; Los Angeles Metro Bike Share; clustering; Data Mining

I. INTRODUCTION

Rising health consciousness, global warming issue, and traveling flexibility dramatically increase the bike-sharing demand in recent years. Bikes are easy to use, fast, and easy to commute through congested places. Metro Bike Share is one of the public transportation options in Los Angeles. In the Metro Bike Share system, people can use the bike for short or long distances within Los Angeles, and it has 246 stations spread around the different regions. The study analyzes the data from the company's website and tries to solve underlying business problems with the approach of Data Mining. The research successfully forecasts the future demand, using Seasonal Autoregressive Intended Moving Average. Using Random Forest Regressor, the research also solves the operational issue for stocking the bikes efficiently in each station. Based on the trip count of each station, the k-means clustering reduces the number of stations. Metro bike share has different types of subscribers. The study also tries to find out the relation between the subscriber types with their trip duration and distance using Decision Tree Classifier, Gaussian Naïve Bayes, and Logistic Regression.

II. DATASET DESCRIPTION

The dataset we use in this paper about Los Angeles Metro Bike Share has trip details from July 7, 2016, to September 30, 2019. This dataset has 761689 unique trip details with 13 attributes; start and end station IDs, start and end time, start and end latitude, longitude along with pass-holder type and one way or round-trip information. The website provides one more dataset for stations, which has 246 unique rows and 5 attributes. This dataset has information about station IDs with their name, location, go-live date, and status (active/ not active).
Another complementary dataset has the weather information which has Date, TMIN (Minimum temperature of the day), TMAX (Maximum temperature of the day), TAVG (Average temperature of the day).

The source of the datasets [1] [2]

III. RELATED WORK

Intensive number of researches appear in this domain. Johan Holmgren [3] provides the day of the week, time of the year, and weather (temperature and precipitation), which influence the amount of bicycle traffic at a point in the traffic network. One research tries to predict potential trip destination and duration [5]. Longbiao Chen [6] provides the number of trips for each station based on the Area function[7], Human activity[8], and Demographics [9] [10].

IV. PROBLEM STATEMENT

A. The Seasonal ARIMA algorithm predicts electricity demand in china [4]. Electricity demand has a seasonal pattern. In this research, the provided data has a seasonal pattern. Bike demand goes high in summer and goes down in winter. The research builds a Seasonal ARIMA model to predict bike demands.

B. Los Angeles Metro Bike share company has 246 stations scattered through different regions. Every station has a different demand, stocking the bikes in those stations in an efficient way can be a problem. Building a model to predict how many bikes should be there can help the operational issue. This research predicts the demand for each station on an hourly and daily basis, taking consideration of weather and holiday impact.

C. The research tries to reduce the station numbers by clustering the stations based on their total daily trips. This model helps reduce the number of stations.

D. Los Angeles Bike share company has monthly-pass, flex-pass, and walk-up pass. The experiment classifies the different passholders based on the duration of their trip

time, the distance they cover, and finds out the relationship between pass-holder types. Classifying the customers based on their demands helps to modify their subscription policy. Also, it adds enormous value to the company, as it can be used to understand the future revenue, which gets generated from a specific pass holder type.

V. DATA PREPROCESSING

A. Data Cleaning

- 9% of the data set has null values. These values get removed.
- The latitude and longitude with (0.0000,0.000) values also gets removed.
- Several trips start from the virtual stations, and these virtual stations are used for testing purposes. For the research, trips starting from the virtual stations get removed.

B. Feature Engineering.

- To forecast the bike demand, the trips gets aggregated on daily basis. Table 1 shows the fragment of final time-series data for the Down Town Los Angeles region.
- The distance of each trip gets calculated from the starting and ending latitude-longitude.
- To predict the number of incoming and outgoing bikes in every station, the experiment counts ingoing and outgoing bikes on an hourly and daily basis.

start_time	trips
2019-03-31	412
2019-04-01	561
2019-04-02	701
2019-04-03	698
2019-04-04	579
2019-04-05	618
2019-04-06	417
2019-04-07	395

Figure 1: Number of trips over time.

VI. DATA ANALYSIS

A. The trend of trips:

The trend of the trips can be observed by the number of trips each day over the period. Figure 2 shows that the number of trips goes down in December, January, and again, it gradually goes up in July and August. The sudden spikes arise on the day of CicLAvia festival. Also, it shows how the experiment divides the data for training and testing. The orange part of the trip implies testing, and the blue part implies the training.

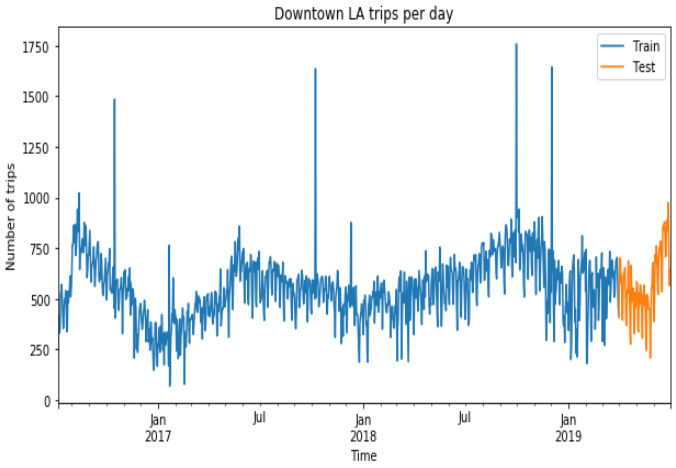

Figure 2: Number of trips over time.

B. The trend of trips on each day:

From Figure 3, it appears that Sundays seem to be busier than other days of the week.

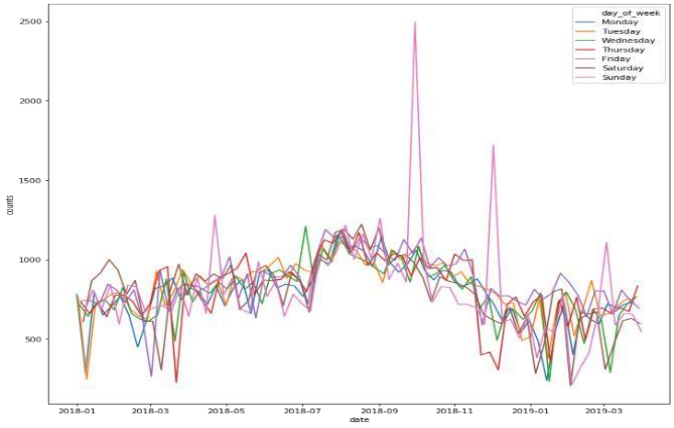

Figure 3: Trip count for day of week

C. Passholder type and duration:

Figure 4 shows that the Walk-up pass holders cover more duration than the Flex Pass and Monthly Pass.

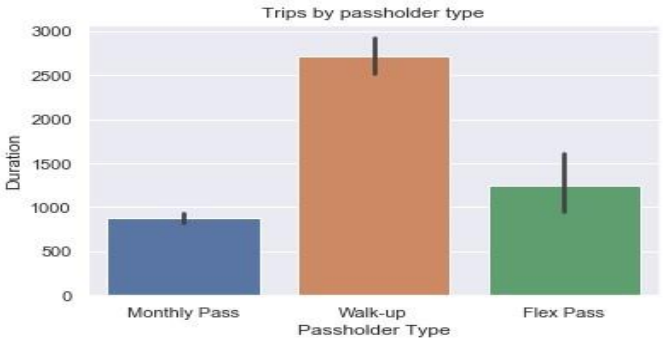

Figure 4: Trip duration by subscription type

D. Top five stations:

Figure 5 shows the top five stations with the highest number of start trips. Station number 4214 has the greatest number of starting trips.

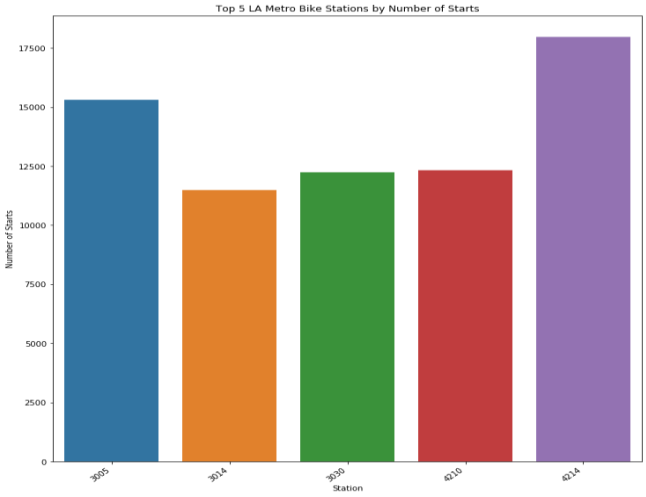

Figure 5: Top five popular stations

VII. EXPERIMENTS

Experiment I: *Bicycle Demand Prediction:*

In this experiment, the research takes the trip count from July 7, 2016, to September 30, 2019, on a daily basis, and divides the data into training and test sets; 2016-07-01 to 201903-31, as training set and 2019-03-31 to 2019-9-30, as a test set, considering the trip counts for the Down Town Los Angeles region..

A. Method:

The dataset does not have a constant mean and variance over time and has a significant amount of seasonality. In that case, ARIMA model works even with non-stationary data that has a trend of element, for instance. However, a limitation occurs this time. The time series displays seasonality (recurring patterns),;in this case the ARIMA model will not work. To make it on ARIMA model, the data should convert into a stationary time series. This implies on considering SARIMA model that can deal with seasonality. This model, generally termed as the SARIMA (p,d,q) × (P, D, Q) m, where p: Trend autoregression order. d: Trend difference order. q: Trend moving average order. And P: Seasonal autoregressive order. D: Seasonal difference order, Q: Seasonal moving average order, m: The number of time steps for a single seasonal period. Using grid search with the combination of the parameters, the experiment finds that the lowest AIC value obtains using SARIMA (1,0,1) × (1,2,2)28. Fitting this parameter in the model gives the best result.

B. Result:

The model gives MAPE (Mean Absolute Percentage Error) 27.028 when tested on the test dataset. Figure 6 shows how the model performs on the test data set. The orange part of the data shows the predicted value, and the blue part indicates the original data. The grey part shows 95% confidence level.

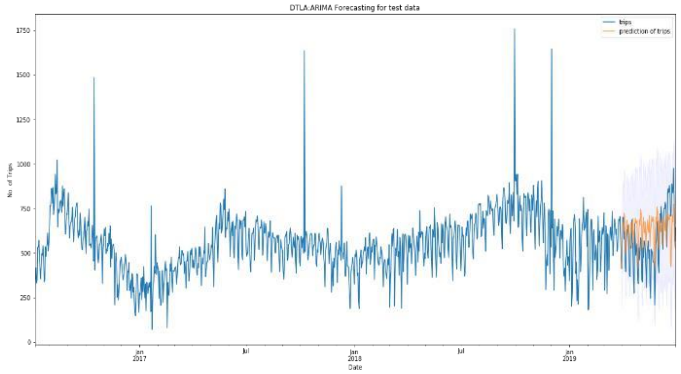

Figure 6: Forecast on test dataset

Experiment II: *Demand Management for every station:*

This experiment seeks to build a model to predict user demand for each station on an hourly and daily basis. It helps to stock the bikes in each station adequately. To understand the usage pattern, the machine learning approach appears significant.

A. Method:

For this model, the experiment considers the impact of temperature and holidays. Merging the trip data from July 7, 2016, to September 30, 2019, with weather data on zip code gives a new modified dataset. Two more columns are added for ingoing and outgoing bikes. Now the dataset has information about the incoming and outgoing number of bikes for each station, along with the temperature and holidays. Saturdays and Sundays are also considered as a holiday for this experiment. To predict the number of incoming and outgoing bikes, this research uses, Random forest regressor and Decision tree regressor models. The ingoing and outgoing bikes get categorized into dependent attributes, and temperature, holiday, time, and date into independent attributes. This experiment divides the data into training (75%), and testing (25%).

B. Results:

While evaluating on the test data, the Decision Tree Regressor gives the following results: Mean Absolute Error 4.2906, Mean Squared Error : 82.4047, R-square value : 0.1336 for predicting the outgoing bike from each station, and the Random Forest Regressor gives Mean Absolute Error : 3.2107, Mean Squared Error : 30.7215, R-square value : 0.6770. Random Forest regressor gives more accuracy for predicting the bike usage based on each station. For predicting incoming bike, Random Forrest Regressor gives, Mean Absolute Error: 0.3256, Mean Squared Error : 0.4035, and R-square value : 0.5964.

Experiment III: *Reducing the number of stations:*

This experiment seeks to reduce the number of stations based on the trip counts along with taking the consideration of the daily average temperature.

A. Method:

Three groups of the trip count have been created, high (trips >100), medium (trips between 50-100), and low(trip < 50), based on the number of trip counts of each station. For clustering, the research uses k-means clustering, with the value of k being 3 here. k-means clustering uses unsupervised learning.

B. Result :

The model can cluster 34% of the stations correctly based on their trip counts.

Experiment IV*: Classifying the passholders:*

In this experiment, the study tries to find out the relationship between the different pass-holder types (Flex Pass, Monthly Pass, Walk-up) with distance and the duration. This experiment calculates the distance and stores them in a column 'Distance'. Distance and Duration are independent attributes. Passholder Type is dependent attribute.

A. Method:

Decision Tree Classifier, Gaussian Naïve Bayes, and Logistic Regression classify the pass holder types. The data gets divided into training (75%), and testing (25%). After evaluating each model on the test data, it shows that every model gives almost the same accuracy.

B. Results:

Figure 6 shows the accuracy level of different classifiers when tested on the test data set. The study observes that there seems to have no significant difference between accuracy for Decision Tree Classifier, Gaussian Naïve Bayes, and Logistic Regression.

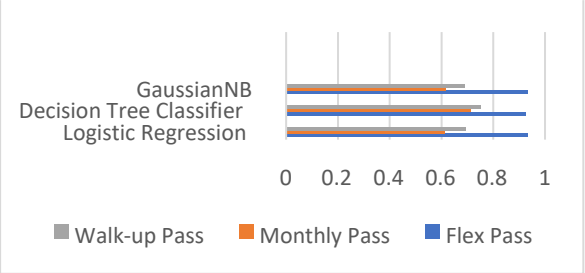

Figure 6: Accuracy comparison of different Classifiers

VIII. CONCLUSION

The research shows that weather and holidays have a significant impact on the number of trips in a period. The study conducts the experiment to solve the problems arising in the domain of bicycle-sharing companies. The data did not have any personal information about the customers, such as age and gender. To make the models more robust, these data can be useful.

The research uses the data from the Los Angeles Metro bikes hare, which explains the effectiveness of the predictive models on real-world datasets.

ACKNOLEDGEMENT

A special token of appreciation for Professor, Dr. Gary D. Boetticher. He helped me throughout the research.

I will remain thankful to him for giving me the opportunity to work on this research.

REFERENCES

[1] https://www.ncdc.noaa.gov/
[2] https://bikeshare.metro.net/about/data/
[3] Holmgren, Johan, Sebastian Aspegren, and Jonas Dahlströma. "Prediction of bicycle counter data using regression." *Procedia computer science* 113 (2017): 502-507.
[4] *Wang, Yuanyuan, et al. "Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: A case study of China." Energy Policy 48 (2012): 284-294*
[5] Zhang, Jiawei, et al. "Bicycle-sharing system analysis and trip prediction." *2016 17th IEEE international conference on mobile data management (MDM)*. Vol. 1. IEEE, 2016.
[6] Chen, Longbiao, et al. "Bike sharing station placement leveraging heterogeneous urban open data." *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015.
[7] García-Palomares, Juan Carlos, Javier Gutiérrez, and Marta Latorre. "Optimizing the location of stations in bike-sharing programs: A GIS approach." *Applied Geography* 35.1-2 (2012): 235-246.
[8] Bikeshare, Capital. "Capital Bikeshare member survey report." Washington, DC (2013).
[9] El-Assi, Wafic, Mohamed Salah Mahmoud, and Khandker Nurul Habib. "Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto." *Transportation* 44.3 (2017): 589-613.