BUAN 6356.003 Business Analytics with R

Group IV

Project Report

Predicting the Demand of Bike Share Users, Based on Natural factors

Group Members:

Kushagra Jain - KXJ230017

Harshita Kala - HXK230047

Ananya Canakapalli - AXC230004

Ramya Padmini Jandhyala - RXJ230032

Sai Tarun Angadipeta - SXA230199

ABSTRACT

This project explores the correlation between natural factors such as time, date and weather to forecast the demand of bike share users. Our project studies the relationship between natural factors and how they impact the number of bike share users on a particular day.

Bike share services provide short-term bicycle rentals, usually via automated kiosks or smartphone apps. In order to improve bike distribution and guarantee that bikes are accessible when and where needed, businesses forecast demand. In order to predict variations in demand, predictive models make use of variables such as past usage data, meteorological conditions, events, and urban patterns. By reducing wait times, this forecasting improves the user experience and guarantees the service provider allocates resources profitably.

OBJECTIVE

We aim to determine the relationship between the natural factors such as date, weekday, holiday, temperature, wind speed etc. and the demand for bike share users. We also aim to determine the trends in demand of bike share users based on the weather situations. Through our project, we aim to create a meaningful model for bike share service operators, to plan their resources accordingly.

DATA SOURCES AND DESCRIPTION

The data that has been used for this project is of the city of Ahmedabad. Ahmedabad is an Indian urban city, and is one of the most developed cities in India. This city is one of the first cities to adopt the concept of bike share services, and hence fits the description for a good dataset for our project. The recorded data runs over a span of 2 years. The following table shows the various metrics that affect demand for bike share users.

S.No.	Metrics	Description
1	**dteday**	The date and day of the data collected. The data collected is for the year 2018-2019
2	**season**	Specifies the season. Coded as follows: 1: spring, 2: summer, 3: fall, 4: winter
3	**yr**	The year of the data recorded. Year is coded as: 1: Year 2018, 2: Year 2019.
4	**mnth**	Specifies the month of the data recorded. The month is coded as 1,2,3….11,12
5	**holiday**	Specifies whether the particular day is a holiday or not. The holiday is coded as: 1: the day is a holiday, 0: the day is not a holiday
6	**weekday**	Specifies the weekday of the data recorded. The weekday is coded as 0: Sunday, 1: Monday ….. 6: Saturday
7	**workingday**	Specifies whether the particular day is a working day or not. The workingday is coded as: 0: the day is not a working day, 1: the day is a working day
8	**weathersit**	Specifies the weather situation for the data recorded. The weathersit is coded as: **1**: (Clear, few clouds, partly cloudy) **2**: (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) **3**: Light Snow, Light Rain + Thunderstorm + Scattered clouds,

		Light Rain + Scattered clouds **4**: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
9	**temp**	Specifies the temperature recorded for the data.
10	**atemp**	Specifies the average temperature of the particular day
11	**hum**	Specifies the average humidity of the particular day
12	**windspeed**	Specifies the average windspeed of the particular day
13	**casual**	Specifies the number of casual (unregistered) users who used the bike share services on that particular day.
14	**registered**	Specifies the number of registered users who used the bike share services on that particular day.
15	**cnt**	Specifies the total count of actual number of users who used the bike share services on that particular day.

The source of the data is Kaggle.

PROCESS:

DATA CLEANING

We commence the project by cleaning the data that we have. For this:

- We first checked for missing values and NA values.
- Secondly we checked the appropriate formatting and data type resourcing.
- We then checked for missing or NULL values in the dataset.
- We checked to remove any duplicate columns in the dataset

DATA PRE-PROCESSING

We then moved to pre-processing the data. This was essential as our variables were not standard and scaled. So, to ensure a good model, we preprocessed the data:

- The 'yr' and 'dteday' fields were removed from calculations as they are irrelevant to the project. We also removed the 'instant' column as it represents a record serial number. We created a dummy dataset without the said variables.
- We tested the collinearity and distribution of features by creating pairplots for numerical variables and boxplots against target variables for categorical variables.
- We scaled numerical variables for better model interpretability and to avoid numeric instability.

- The data was split into 80% training and 20% testing datasets to evaluate the regression model.

REGRESSION MODEL

We created three linear regression models in R with the selected features as our X variables and had our target variables: Total Number of Users, Casual Users and Registered Users. We created 3 models so as to assist companies in accurately predicting the number of registered users, casual users and ultimately the total number of users.

The regression equation coefficients for the models are as follows:

Coefficients	Casual Users	Registered Users	Total No. of Users
Intercept	1406.63225	2458.53585	3865.168101
Season	66.72063	408.46119	475.181818
Month	-20.75978	-14.38490	-35.144684
Holiday	-255.94203	-203.18939	-459.131419
Weekday	34.29869	60.40001	94.698703
Working Day	-864.57093	869.47665	4.905724
Weather Situation	-70.00986	-351.53371	-421.543571
Temperature	-306.87353	-1070.94247	-1377.815999
Average Temperature	702.64120	1721.38724	2424.028449
Humidity	-70.61166	-197.11072	-267.722373
Windspeed	-44.38772	-117.74777	-162.135491

INFERENCE:

- We can conclude that Average Temperature has the highest positive impact in all three models. This implies that the number of users, whether casual, registered or total, are most positively affected by the average temperature on a given day.
- In the casual users model, we can observe that Working Day has the highest negative impact on the target variable. This implies that the number of casual users are less on a working day, and they keep going down as we move further along the week.
- In the registered and total users model, we can observe that Temperature has the highest negative impact on the target variable. This implies that the higher the temperature on a given day, the lower the number of registered and total users a company can expect.

EVALUATION:

We tested our regression models on the testing dataset, and evaluated our regression models using Adjusted R squared and mean squared error. The following were our findings:

Parameter	Casual Users	Registered Users	Total Users
R Squared	0.664277	0.4900311	0.5148073
Adjusted R Squared	0.6584383	0.481162	0.5063691
Mean Squared Error	214051.1	1364052	2024931

Through the R Squared, Adjusted R Squared and the Mean Squared Error, we came to conclude that while the model is decently accurate to fit all sorts of datasets, it still does not guarantee a very high accuracy. This lack of high accuracy is due to the absence of a few key factors from the dataset. Further, we figured that the average Adjusted R Squared in predicting use of public transport mediums is around 40% - 60% only.

FUTURE SCOPE:

We can further improve the accuracy of our model by factoring for a few more variables such as population density, availability of public transport, awareness about bike share services, traffic density etc. With the added variables and improved scope of better feature selection, we can improve the implied accuracy of our model.