**BIKE SHARING DEMAND**

**- DATA ANALYSIS**

**DSCI 4780/6780 Final Project Report**

**BY**

Adedamola Adebamiro

Adjoa Myra Light Dagba

Jesse Annan

Prince Osei Bonsu

April 22, 2023

**PROBLEM DESCRIPTION**

Africa is the world's second largest and second-most-populous continent, after Asia in both aspects, with over 1.2 billion people, most of whom would need some means of transportation to go about their daily lives. Unfortunately, the continent is plagued with major challenges of transportation, largely attributed to the massive population, in addition to insufficient and sometimes unreliable transport systems (SSATP, 2015).

**Problem**

Living in a developed country is often taken for granted, the luxury of having good roads, affordable transportation, and quick and easy access to vital services. Imagine having to travel many miles to work, school, and the market on foot every day, instead of being able to hop in your car and go where you need to when you need to. It would be exhausting, and you would waste many hours each day just commuting. You would also miss out on countless opportunities for education, economic growth, and healthcare. Unfortunately, for some West African countries like Sierra Leone and Ghana, this is not a hypothetical scenario - it's their daily reality. Lack of transportation is one of the biggest challenges facing rural residents in these West African nations. The distances they need to travel are often long, the roads are mostly unpaved and bumpy, and public transportation is either prohibitively expensive or nonexistent.

These are complex issues that can seem impossible to solve, however, something as simple as a bike can make a huge difference. With a bike, inhabitants of these rural areas can cover greater distances faster and easily, providing much-needed access to new opportunities for education, work, and healthcare. It's a small solution that can have a big impact on improving the quality of life for those who are struggling. Growing up in Africa and experiencing some of these problems first-hand has been a major motivation for picking this project. Another motivation is the potential positive impact of this innovation on the environment and public health. Biking is a sustainable and eco-friendly mode of transportation that can help reduce carbon emissions and air pollution. By assisting biking companies in optimizing their services, more people may be encouraged to choose biking as a means of getting around, leading to improved public health and a cleaner environment.

**Dataset**

The dataset used in this project was sourced from [Kaggle.com](https://www.kaggle.com/competitions/bike-sharing-demand/overview). Below is a list of all the features in the dataset, as well as a brief description of each feature:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| datetime | Hourly date + Timestamp |
| season | 1 = spring, 2 = summer, 3 = fall, 4 = winter |
| holiday | Whether the day is considered a holiday |
| workingday | Whether the day is neither a weekend nor a holiday, i.e., a working day of the week |
| weather | 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 |
| temp | Temperature in Celsius |
| atemp | ‘feels like’ temperature in Celsius |
| humidity | Relative Humidity |
| windspeed | Wind Speed |
| casual | Number of Non-registered user rentals initiated |
| registered | Number of Registered user rentals initiated |
| count | Number of total rentals |

The target features are *casual, registered,* and *count* (a summation of *casual* and *registered* features).

**Proposed Analytics Solution**

The objective of this project is to build a machine learning model that will assist bike companies predict the demand for their services. Through data analysis, this project will aid business owners in identifying when to provide bike rentals to consumers to maximize profitability. The findings of this project will also enable bike sharing systems to determine the optimal times to increase their bike inventory for rental, to meet the demands of consumers during peak seasons. By utilizing data-driven insights, bike companies can maximize their profits while providing their customers with the convenience and accessibility they need.

**DATA EXPLORATION AND PREPROCESSING**

The dataset, as seen above, consists of 7 descriptive features and 3 target. To begin the data exploration process, the *datetime* column was first transformed into 5 components namely: *‘year’, ‘month’, ‘day’, ‘dayofweek’, and ‘hour’*. To begin our exploratory data analysis, we first observe the linear relationship between descriptive features and our target features.

Graphical user interface, application

Description automatically generated

***Figure 1:*** *A heatmap of the continuous features and the ‘count’ target feature*

The correlation matrix of the continuous features and targets shows that as *humidity* increases the demand of bikes by both *casual* and *registered* members reduced, it also shows *temp and atemp’s* increase, increased the demand of bikes for both *casuals* and *registered* users. However, both features are highly correlated. This makes sense because the actual temperature should not differ significantly from the ‘feels like’ temperature.

A picture containing text, electronics

Description automatically generated

***Figure 2:*** *A heatmap of the categorical features and the ‘count’ target feature*

Observing the linear relationship between categorical features and targets. We observed a general increase in features like *year, month, hour, season,* and *weather* had a positive impact on the demand of bikes. It is also very clear to observe that *month* and *season* are highly correlated thus using both in any predictive analysis could negatively affect the predictive power of any model. Same could be said about *workingday* and *dayofweek.* In view of that, we chose to select *season* over *month.*

The *windspeed* feature although it has very weak linear relationships with the target features. We carried out some descriptive analysis. Average and maximum *windspeeds* for the *year 2011* were both greater than *year 2012*. Riding bikes in high *windspeed* is generally dangerous and therefore we deduce it could be the reason for the few bikes demanded during the *year 2011.*

Bar charts were plotted to examine the distribution and trend of bike rentals by hour, month, year, season, weather, and holiday. It is observed that the number of rentals in 2012 is generally greater than the number of rentals in 2011 for both casual and registered bikes. The most casual bikes were rented in June and the most registered bikes were rented in October.

For casual bikes, the rentals begin to pick up daily around 9 am, peak at around 2-3 pm, and begin to slow down at around 6-7 pm. For registered bikes, we observe two peaks occurring at around 8 am and 5-6 pm, before rentals begin to slow down for the day.

Based on seasons, more bikes are rented in the Fall and Summer seasons than in the other two seasons for casual users. For registered users, there are a lot of rentals in the winter season as well, in addition to the fall and summer seasons.

Based on the weather, more bikes are rented in clear and cloudy weather conditions than in rain or snowy weather for casual users. For registered users, there is a significant percentage of users renting in the rain and in snowy weather too.

Based on holidays, there are more casual rentals when it is a holiday, compared to registered rentals.

On the days of the week when it is a holiday, both casual and registered bike rentals are highest on Mondays and Wednesdays, while on days of the week that are working days, bike rentals are highest on Tuesdays, Thursdays, and Fridays.

**Data Quality Reports**

The data quality reports for both the categorical and continuous features are seen below.

Table

Description automatically generated

***Table 1:*** Data Quality Report for Continuous Features

Table

Description automatically generated

***Table 2:*** Data Quality Report for Categorical Features

**Handling Missing Values and Outliers**

As seen in the data quality report tables above, there are no missing values in the dataset. Hence, there was no need to handle missing values. Since treating high demand of bikes regardless of the season or weather condition as an outlier would be expensive, we decided to trim our dataset. The top and bottom 2% of the dataset was dropped. We also observed our *windspeed* feature, shows very dangerous values when bikes were used. It will be very difficult for even pro athletes to use bikes in *windspeeds* above *30mph* therefore we removed instances where *windspeed* were above 29mph. After these methods had been applied, the initial 10,886 instances of bike demands were reduced to 10,155 instances, implying that approximately 7% of our original dataset were considered outliers.

**Normalization**

The continuous features, *temp, atemp, windspeed,* and *humidity* were normalized into the range of [0,1]. This is important because huge differences in ranges of values in respective features may affect the way our machine learning algorithm(s) think a feature is worth.

**Feature Selection and Transformations**

Since implementing the wrapper method was deemed expensive, the filter method was used for selecting relevant features for our machine learning algorithm. After the exploratory data analysis, the features deemed fit for selection from the correlation heatmap are *temp, humidity,* *year, hour, season, weather,* and *workingday*. Although *workingday* has a bad correlation with all our target features, we observed a general increase in model accuracy when the feature was added. Machine Learning algorithms, inspired by some mathematical concepts usually need numerical values to make sense of the data, therefore, our categorical features were transformed using the one-hot-encoding method (with the first value dropped). After this, *Season* was selected over *month* because month increased the dimensionality of our data without causing any significant increase in our model performances. Transforming regression problem to a classification problem, we took advantage of the symmetric nature of our target features and implemented two methods. In the first method, the targets were grouped into four groups using equal width binning and in the second method, the targets were grouped based on their interquartile positions.