BikeIndia Project on Linear Regression

Background Information

You are a data scientist at BikeIndia, a bike-sharing system which provides individuals with access to bikes for short-term use, either for a fee or at no cost. Users can borrow bikes from designated docks, typically computer-controlled, where they enter payment information to unlock the bike. Bikes can be returned to any dock within the same system.

BikeIndia, a bike-sharing provider in the US, had experienced significant revenue declines during the COVID-19 pandemic. Facing challenges in the current market environment, the company aims to develop a strategic plan to boost revenue once lockdown measures ease and the economy improves.

To achieve this goal, BikeIndia intends to assess post-quarantine demand for shared bikes nationwide. By understanding customer needs post-pandemic, the company aims to differentiate itself from competitors and achieve substantial profits.

To analyze demand factors for shared bikes, BikeIndia has engaged a consulting firm. Specifically, they seek to identify significant variables influencing bike demand in the American market. Key questions include: which variables are critical in **predicting bike demand**, **and how effectively do these variables describe demand patterns**?

Through extensive data collection, including meteorological surveys and user preferences, BikeIndia has compiled a comprehensive dataset on daily bike demands across the American market, focusing on various influencing factors.

Problem Statement

We need to create a model that predicts the demand for shared bikes using the available independent variables. This model will help management comprehend how demand changes based on various features, enabling them to adjust the business strategy to match demand levels and fulfill customer expectations. Additionally, the model will offer insights into the demand dynamics of a new market, providing valuable information for management decision-making.

Dataset

Download your dataset from here - Dataset

Dataset Name - BikeIndia

Dataset Schema -

RangeIndex: 730 entries, 0 to 729
Data columns (total 16 columns):

S. no.	Column	Non-Null Count	Dtype
1	instant	730 non-null	int64
2	dteday	730 non-null	object
3	season	730 non-null	int64
4	yr	730 non-null	int64
5	mnth	730 non-null	int64
6	holiday	730 non-null	int64
7	weekday	730 non-null	int64
8	workingday	730 non-null	int64
9	weathersit	730 non-null	int64
10	temp	730 non-null	float
11	atemp	730 non-null	float
12	hum	730 non-null	float
13	windspeed	730 non-null	float
14	casual	730 non-null	int64
15	registered	730 non-null	int64
16	cnt	730 non-null	int64

dtypes: float64(4), int64(11), object(1)

Dataset Description -

1. Instant - serial number
2. Dteday - the date on which the bike was rented
3. Season - the four seasons of the year represented by integral values - 1,2,3, and 4
4. Yr - represents the year. 0 stands for the year 2018, and 1 stands for the year 2019.
5. Mnth - represents all the months of the year by integral values from 1-12 representing months from January to December, respectively.
6. Holiday - Binary values - 0 for non-holiday and 1 for holiday.
7. Weekday - consists of integral values 0-6 representing all the days of a week
8. Workingday - binary values - 0 for non-working days and 1 for working days.
9. Weathersit - situation of the weather represented by two integers - 1 and 2.
10. Temp - the temperature at which the bike was rented
11. Atemp - the average temperature of the day
12. Hum - humidity represented by a float value
13. Windspeed - speed of the wind on that day represented by a float value
14. Casual - number of casual bike riders
15. Registered - number of registered bike riders
16. Cnt - represents the count of the total number of bike riders in that day

Solution Deliverables

Your solution must cover the following checkpoints -

1. Dataset information - View the dataset's overall structure that shows all the columns and their data types, number of columns that are categorical, etc.
2. Data quality check - Check for null/duplicate/missing values
3. Data Cleaning - Pre-process the data so that it becomes suitable for training a model on it.
4. Identifying and removing redundant/unwanted features solely by observing the features and whether they are required for the prediction of the dependent variable.
5. Exploratory Data Analysis - involves visualizing numerical and categorical values and correlation matrix
6. Feature Rescaling, if required
7. Model Implementation - Based on EDA, first implement a simple linear regression model using the feature that has maximum correlation with the dependent variable and then implement a multiple linear regression model.
8. Evaluation Metrics - evaluate the results of both the models and compare them.
9. Final Results - Write down the key findings from the EDA and results.