**Linear regression Analysis of Bike Sharing Company**

Department of Computer science

New Jersey Institute of Technology (NJIT)

Newark, New Jersey -07102

**Linear Regression:**

It is on for the supervised learning methods, which use past data for building labels. It has a linear relationship between dependent/target and independent variables. It consists of labels that are continuous in nature. Linear regression models are often fitted using the least squares approach and best fit line. In regression, a best-fit line is a line that fits the given scatter plot in the best way. If there is no association between the proposed target and the independent variables (i.e., the scatterplot does not indicate any increasing or decreasing failure trends), then the linear regression model to the data can’t provide any useful data for interpretation. Linear regression finds the straight line, called the least squares regression, that best represents observations in a bivariate data set.

A linear regression line equation form:

**Y = Β0 + Β1X**

Y -> value of the dependent variable,

X -> value of the independent variable

Β0 -> constant,

Β1 -> regression coefficient.

**Problem Statement:**

In this project, we build a multiple linear regression model for the prediction of demand for shared bikes.

A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short-term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.

A US bike-sharing provider **Boom Bikes** has suffered a dip in its sales due to the pandemic. So, the company wants to understand the dip in sales and wanted to prepare itself for after-pandemic conditions to increase sales. To address this issue, we need to find the factors affecting the demand for bikes and know which variables are useful in predicting the demand for shared bikes.

I am modeling the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

**Dataset Description:**

The data consists of 730 datapoints and 16 variables. The data consist of the following 16 columns: instant, dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, temp, atemp, hum, windspeed, casual, registered, cnt. Out of these variables, cnt is the target variable for the data set. After the analysis, I figured out that there are no missing values in that data set.

**Methodology:**

Here, I used sklearn and statsmodel libraries to perform a linear regression. The data has been converted into object form using dummy variables. The given input data has been split into train and test with a 70:30 split ratio. This means the training data will consist of 70% of the data and the test data will have 30% data. I scaled the numeric columns using the MinMaxScalar function.

The modeling process followed is a top-down approach where we consider all the variables to model and remove the variables which are not important. To filter out the required variables, I used the Recursive feature elimination (RFE) and Variance Inflation Factor (VIF) methods.

We have the following conditions when we consider p-value and VIF.

High p-value, high VIF -> Drop

High p and low VIF (Remove these first) OR

Low p and high VIF (Remove these after once above)

Low p and low VIF

I considered the 2 factors to eliminate the data namely We always prefer VIF<5 & P value <0.05.

The test data here has been evaluated/predicted with the 30% remaining data to check how well the model will work on unseen data.

These are the different variations observed in the categorical data.

Chart, bar chart

Description automatically generated

The accuracy of the model is calculated using the R2 score. The R2 score of the training model observed was 82.3% and for the test model, 81%. This shows that the data is not overfitting and is performing well on the unseen data. The mean square error is also on the lower side i.e., 1.01%. To understand more about the data, the company can wait and get more data to investigate more.

**Conclusion:**

The coefficients of the variables for the respective independent variables are listed below. Here we can observe that variable temperature has the highest co-efficient and the Light Snow and Rain has the lowest co-efficient with respect to the target variable.

Text

Description automatically generated

**Future Works:**

In conclusion, the bike-sharing company must take a decision based on these factors/variables. Here are a few suggestions which can be used by the company to increase the customer market. The company could roll out a few offers in the summer month to attract more people in good temperature seasons. They could provide more good services to the users in rainy seasons. Since the performance of the company is not good in the month of December and July, they can increase their advertising budget in those months so that more people are aware of this and not lose potential customers. They can remove all the offers during the holiday season since no one will be using them during that period.

**Contribution:**

Rohan Pujari(RP992) - Research, Dataset, Methodology, Experiments, Conclusions, and Future works.

**References:**

https://www.kaggle.com/datasets/hmavrodiev/london-bike-sharing-dataset