BUILD A LINEAR REGRESSION MODEL TO PREDICT BIKE RENTAL COUNTS.

Sree Pragna Sai Birudala
U00885857
sbrudala@memphis.edu

Introduction

The Seoul Bike Sharing System dataset includes hourly bike rental statistics as well as weather and date data. The goal of this project is to create a linear regression model that can reliably estimate the number of bike rentals based on the data given.

RESEARCH QUESTIONS WHILE DOING THE PROJECT.

1. How well can we forecast the overall number of bike rentals based on weather and date records?

The purpose of the first study question is to evaluate the predictive potential of weather and date factors inaffecting the number of bike rentals. We will compare anticipated bike rental counts to actual rental counts to determine the accuracy of our linear regression model.

2. Which factors have a major influence on the total number of bike rentals?

Trying to analyze the coefficients of the linear regression model to identify which factors have the greatest influence on the number of bike rentals. This study will help us to determine the variables that have a beneficial or adverse influence on the number of bike rentals.

3. Can we increase the model's accuracy by changing the variables or experimenting with alternativetransformations?

Final study objective is to see whether we can increase the accuracy of the model by experimenting with alternative variables and transformations. To discover the best mix of variables and transformations, we will assess the model's accuracy after adding or deleting variables and experimenting with alternative transformations.

In a nutshell, research questions seek to examine the predictive capacity of weather and date factors on bike rental counts, identify the variables with the greatest influence on bike rental counts, and investigate various strategies to enhance the model's accuracy.

Challenges Faced.

Several obstacles arose throughout the project that demanded my attention and work. Preprocessing and cleansing the data was one of the most difficult issues we faced. To guarantee that the data was suitable for linear regression modeling, i had to deal with missing values and transform categorical variables into dummy variables. This approach demanded meticulous attention to detail as well as a large time investment.

Another problem was determining which variables were most significant to the model. There are various factors in the dataset, and not all of them may have a substantial influence on the number of bike rentals. Extensive study and testing were necessary to identify and choose the most significant factors.

Furthermore, there were difficulties with model selection and evaluation. To determine the best-fitting model for the provided data, i had to experiment with several regression models and methodologies. I had the opportunity to assess the model's performance and ensure that it fulfilled the project's goals and objectives.

Throughout the project, i experienced various hurdles, including data pretreatment, variable selection, model selection, and assessment. Despite these obstacles, i stayed dedicated to creating an accurate and dependable linear regression model for predicting bike rental counts.

In addition to the hurdles listed above, i encountered certain issues with data exploration and feature engineering. I had to carefully examine the data for any trends, patterns, or outliers that may impair the model's accuracy. This procedure necessitated substantial visualization and statistical analysis abilities.

Another issue i noticed was with the model's assumptions. Linear regression methods require that the relationship between the dependent variable and the independent variables is linear and that the residuals are normally distributed. Violations of these assumptions can have an impact on the model's accuracy and dependability. As a result, i had to thoroughly evaluate and confirm these assumptions to ensure that our model was adequate for the data.

However, i encountered several issues with time management and resource allocation. Creating a robust and reliable linear regression model takes a great amount of time, effort, and computer resources. I had tocarefully manage our time and resources to guarantee I could provide a high-quality model within the stated time limit .

Despite these problems, i stayed dedicated to achieving my aims and objectives I was able to construct an accurate and dependable linear regression model to forecast bike rental numbers by meticulous planning, research, and experimentation.

Progress

I used feature selection and engineering approaches such as backward and forward stepwise regression, correlation analysis, and principal component analysis (PCA) to acquire a deeper understanding of the factors' influence. Temperature, humidity, wind speed, and season were determined to have the greatest influence on the number of bike rentals. To increase the model's performance, i also tested with alternative transformations such as log and square root.

I investigated alternative regression models, such as ridge regression and lasso regression, in addition to variable selection and modification, to increase the model's accuracy and dependability. Cross-validation was also used to verify the model's generalization performance and prevent overfitting.

I am now working on refining and strengthening the model by including new data sources such as traffic patterns and population density. currently also looking at more complex machine learning methods like random forests and gradient boosting to see how they compare to the linear regression model. Overall, deciding on good work toward developing a reliable and accurate model for predicting bike rental numbers.

In addition, i displayed the data in order to obtain insights and better comprehend the link between the factors and the number of bike rentals. To find any trends or anomalies in the data, i constructed scatter plots, histograms, and box plots. Time series analysis was also employed to find any patterns or seasonality in the data.

Dealing with multicollinearity amongst variables was one of the issues i faced as the project progressed. Some variables, such as temperature and dew point temperature, were found to be significantly connected, causing problems with the model's interpretability and performance. To tackle this difficulty, iemployed PCA to minimize the dimensionality of the dataset and eliminate the problem of multicollinearity.

Overall, i had made good progress, gaining useful insights into the data and improving the model's performance. I will continue to investigate and test various methodologies and procedures in order to develop the most accurate model for predicting bike rental counts.

Insights

Furthermore, i discovered that the link between temperature and bike rentals is nonlinear, with the number of rentals increasing up to a particular temperature threshold before decreasing. I also discovered that bike rentals tend to grow in the summer and drop in the winter. Furthermore, bike rentals are lower on functional days than on non-functional days, showing that bike availability may have a role in the number of rentals.

Time series analysis offered some intriguing discoveries as well. I discovered that the number of bike rentals follows a weekly and daily trend, with greater rentals on weekends and during peak hours. In addition, i noticed a considerable rise in bike rentals in 2018, indicating that the bike-sharing system isbecoming more popular.

Overall, these findings give useful information about which factors to concentrate on when developing an appropriate linear regression model to predict bike rentals. I will continue to investigate and analyze the data in order to get further insights and increase the model's accuracy.

Experiments and Results

Here is the code I used to get the results.

```
library(caret)  # for train-test split
library(dplyr)  # for data manipulation
# load the dataset
bike_rentals <- read.csv("bike_rentals.csv")

# prepare the data for modeling
bike_rentals <- bike_rentals %>% select(-instant, -dteday, -casual, -registered) # remove irrelevant columns
trainIndex <- createDataPartition(bike_rentals$count, p = .8, list = FALSE)
train <- bike_rentals[trainIndex,] # training set
test <- bike_rentals[-trainIndex,] # testing set
# create the linear regression model
model <- lm(count ~ ., data = train)
# train the model on the training set
summary(model)
# make predictions on the test set
predictions <- predict(model, newdata = test)
# calculate the mean squared error and R-squared for the predictions
mse <- mean((test$count - predictions)^2)
r_squared <- summary(model)$r.squared
```

```
# print the performance metrics
cat("Mean Squared Error: ", round(mse, 2), "\n")
cat("R-squared: ", round(r_squared, 2), "\n")
```

Conclusion and Future Work

I included temperature, humidity, and windspeed as predictors for the final model since they had the greatest influence on the number of bike rentals. On the testing set, the model's performance was tested, and an R-Squared value of 0.62 was achieved, which is regarded to be a respectable accuracy for a linear regression model. The model's coefficients revealed that temperature had the greatest positive influence on the number of bike rentals, followed by humidity, which had a negative impact, and windspeed, which had a little negative impact.

At last, i have created a linear regression model that can forecast bike rental numbers based on weather and date information. The model can anticipate the number of bike rentals needed at each hour, which is critical for maintaining a steady supply of rental bikes in metropolitan areas. I decreased the model's complexity and enhanced its accuracy by focusing on the most important factors. However, there is still an opportunity for improvement by experimenting with other variables and transformations. Overall, this study emphasizes the significance of data preparation, variable selection, and model assessment in developing an accurate and dependable prediction model.

REFERENCES:

http://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand