Abstract. This research developed a model of linear regression for forecasting the demand for shared bikes in a bike sharing system. By analyzing a dataset sourced from Kaggle, the study focuses on identifying the factors that have the most impact on bike demand and building a model based on these factors. The methodology involves data cleaning, creating dummy variables for categorical variables, and conducting exploratory data analysis. The features are rescaled, and the model building process includes recursive feature elimination and analysis of VIF and p-values. The outcome indicated that linear regression model accurately predicts bike demand based on various factors. This model can assist employers in adapting their business strategies, understanding customer expectations, and effectively managing bike-sharing systems. The findings contribute to the optimization and success of sustainable urban transportation, emphasizing the potential of bike sharing as an eco-friendly transportation option.

1 Introduction

Zhang and Mi utilize big data analysis to assess the environmental benefits of bike sharing [1]. The study demonstrates that bike sharing can lead to significant energy savings and emissions reduction, particularly when replacing short car trips. This highlights the potential of bike sharing as a sustainable transportation option. Ricci conducts a comprehensive review on impacts and processes of bike sharing implementation and operation [2]. It consolidates evidence from various studies to assess the impacts and challenges associated with bike sharing. The review reveals positive impacts such as reduced traffic congestion, improved air quality, and increased physical activity, while also addressing infrastructure, funding, and user behavior as important considerations in bike-sharing implementation. Vogel, Greiser, and Mattfeld employs data mining techniques to analyze user activity patterns within bike-sharing systems. The authors identify different user groups and emphasize the importance of understanding user behaviors for system design and resource allocation, ultimately improving the user experience and system performance [3]. Tran, Ovtracht, and d’Arcier focus on the role of built environment factors in modeling bike-sharing systems. Their study highlights the influence of factors like land use, population density, and transportation infrastructure on bike-sharing usage. Considering these factors in system design and planning can enhance user satisfaction and overall system performance [4].

Overall, these articles collectively underscore the potential of bike sharing as a sustainable transportation solution. They highlight the environmental benefits, positive impacts on urban mobility, the significance of understanding user behaviors, and the influence of built environment factors on bike-sharing system performance. The findings contribute to the design, operation, and optimization of bike-sharing systems, aiding in the promotion and success of sustainable urban transportation. The objective of this study is to create a model of the bike-sharing system that will help business owners determine the variables that have the biggest effects on predicting demand for shared bikes. The model based on the demand for shared bikes will be constructed using the independent variables that are readily available. The management will use it to determine precisely how the needs change with changing features. They can adjust their business approach in accordance with demand levels and client expectations. The model will also help management better grasp the dynamics of demand in a new market.

2 Data

2.1 Variables

2.1.1 Variables and definitions

The dataset used in this project is sourced from Kaggle and contains information on daily demands across American markets for a two-year period [5]. The dataset consists of 16 columns, including variables such as “instant,” “date,” “season,” “year,” “month,” “holiday,” “weekday,” “working day,” “weather situation,” “lowest temperature,” “highest temperature,” “humidity,” “wind speed,” “casual user,” “registered user,” and “total demand.”

In the dataset, the “season” variable is represented by the numbers 0, 1, 2, 3, and 4, indicating the four seasons as follows: season one (January to March), season two (April to June), season three (July to September), and season four (October to December). The “year” variable is represented by the numbers 0 and 1, corresponding to the years 2018 and 2019. The “month” variable is recorded using the
numbers 1 to 12, representing the months from January to December. The “holiday” variable is represented by 0 and 1, where 0 signifies “No” and 1 signifies “Yes.” The “weekday” variable is recorded using the numbers 0 to 6, corresponding to Monday to Sunday. The “working day” variable is represented by 0 and 1, with 0 indicating “No” and 1 indicating “Yes.” The “weather situation” variable is recorded using the numbers 1, 2, and 3, representing three different weather situations.

2.1.2 Creating dummy variables

In statistics and econometrics, a dummy variable, also known as an indicator variable, is a binary (0 or 1) variable created to represent categorical variables in a quantitative analysis [6]. It is used to incorporate qualitative information into a regression model, where numerical values are required.

Dummy variables are created by assigning a value of 1 to represent the presence of a particular category and a value of 0 to represent the absence of that category.

Dummy variables are primarily used in regression analysis, especially when dealing with categorical variables as predictors. By including dummy variables in regression models, the effects of categorical variables can be estimated quantitatively. For example, in a multiple regression model, the coefficients of the dummy variables indicate the difference in the outcome variable’s mean between the reference category (represented by 0 in all dummy variables) and the corresponding category (represented by 1 in the dummy variable).

Dummy variables allow for the inclusion of categorical information in regression models while maintaining compatibility with numerical analysis techniques. They help capture the influence of qualitative factors that cannot be directly represented by numerical values.

In this study, create the DUMMY variables for 4 categorical variables ‘mth’, ‘weekday’, ‘season’ and ‘weathersit’.

2.2 Data preprocessing

2.2.1 Check for null/missing values

Checking for null values in Python is an essential step in data analysis and preprocessing. Null values, also known as missing values, refer to the absence of data in a specific variable or column. Understanding and handling null values is crucial because they can impact the accuracy and reliability of data analysis and modeling. There are no missing or Null values in either the columns or rows of our data collection, according to the check.

2.2.2 Removing unnecessary columns

The following variables can be excluded from further analysis based on a high-level examination of the data and the data dictionary:

- instant: Its only an index value

2.3 Data analysis

2.3.1 EDA

Exploratory Data Analysis (EDA) refers to the process of visually and statistically exploring a dataset to understand its main characteristics, patterns, and relationships [7]. It involves using various techniques to summarize, visualize, and interpret the data before diving into more formal modeling or analysis. EDA is a crucial step in the data analysis process as it helps analysts gain insights, detect anomalies, validate assumptions, and inform subsequent modeling or analysis decisions. By thoroughly exploring the data, analysts can make better-informed choices and develop a solid foundation for further analysis or modeling tasks.

2.3.2 Visualising numeric variables

All the numerical variables in the training dataset should be plotted as a pairplot. According to the Pair-Plot result, there is a LINEAR RELATION between the variables "temp," "atemp," and "cnt."

2.3.3 Visualising categorical variables

Before producing dummies, create a boxplot of all category variables against the goal variable "cnt" to observe how each predictor variable compares to the target variable.

Six category variables were present in the dataset. Use a box plot to examine their effect on the dependent variable ('cnt'). The conclusions that could be drawn were:

- season: With a median of over 5000 bookings (over the course of two years), about 32% of bike reservations took place in season 3. Seasons 2 and 4 came in second and third, respectively, with 27% and 25% of all reservations. This suggests that the dependent variable's season may be predictably related.
- mth: With a median of over 4000 bookings per month, about 10% of the bike reservations took place in the months 5, 6, 7, and 8. This suggests that mth has a clear trend and can serve as a reliable predictor of the dependent variable.
- weathersit: A median of about 5000 bike bookings occurred during "weathersit" (during the span of two years), accounting for nearly 67% of all bike bookings. With 30% of all bookings, weathersit2 came in second. This shows that WeatherSit does exhibit a trend toward bike reservations, which may be a useful indicator of the dependent variable.

2.3 Data preprocessing

2.3.1 EDA

Exploratory Data Analysis (EDA) refers to the process of visually and statistically exploring a dataset to understand its main characteristics, patterns, and relationships [7]. It involves using various techniques to summarize, visualize, and interpret the data before diving into more formal modeling or analysis. EDA is a crucial step in the data analysis process as it helps analysts gain insights, detect anomalies, validate assumptions, and inform subsequent modeling or analysis decisions. By thoroughly exploring the data, analysts can make better-informed choices and develop a solid foundation for further analysis or modeling tasks.

2.3.2 Visualising numeric variables

All the numerical variables in the training dataset should be plotted as a pairplot. According to the Pair-Plot result, there is a LINEAR RELATION between the variables "temp," "atemp," and "cnt."

2.3.3 Visualising categorical variables

Before producing dummies, create a boxplot of all category variables against the goal variable "cnt" to observe how each predictor variable compares to the target variable.

Six category variables were present in the dataset. Use a box plot to examine their effect on the dependent variable ('cnt'). The conclusions that could be drawn were:

- season: With a median of over 5000 bookings (over the course of two years), about 32% of bike reservations took place in season 3. Seasons 2 and 4 came in second and third, respectively, with 27% and 25% of all reservations. This suggests that the dependent variable's season may be predictably related.
- mth: With a median of over 4000 bookings per month, about 10% of the bike reservations took place in the months 5, 6, 7, and 8. This suggests that mth has a clear trend and can serve as a reliable predictor of the dependent variable.
- weathersit: A median of about 5000 bike bookings occurred during "weathersit" (during the span of two years), accounting for nearly 67% of all bike bookings. With 30% of all bookings, weathersit2 came in second. This shows that WeatherSit does exhibit a trend toward bike reservations, which may be a useful indicator of the dependent variable.
• holiday: The fact that nearly 97.6% of bike reservations took place during non-holiday times shows that this data is skewed. This suggests that the dependent variable CANNOT be well predicted by the holiday.
• weekday: Weekday variables have separate medians between 4000 and 5000 reservations and exhibit a fairly close trend (between 13.5 percent and 14.8 percent of total booking on all days of the week). This variable may or may not have any bearing on the predictor. If this needs to be added or not, I’ll leave that up to the model.
• workingday: Nearly 69% of bike bookings occurred on “workingdays” with a median of nearly 5000 bookings (during a two-year period). The working day may therefore be a useful predictor of the dependent variable, according to this.

2.3.4 Correlation matrix
To find out which variables are strongly associated, look at the correlation coefficients.
• The heatmap clearly shows which variables are multicollinear in nature and which are highly correlated with the target variable.
• While the linear model is being constructed, this map will be utilized periodically to validate various correlated values, together with VIF and p-value, for selecting the proper variable to include in or exclude from the model.

2.4 Rescaling the features
Use the MinMaxScaler from sklearn.preprocessing to rescaling the feature. Rescaling the features can enhance the speed and convergence of machine learning algorithms, improve the data's readability, and lessen the likelihood that some characteristics would dominate the study or be biased.

3 Methodology: Linear regression
A dependent variable and one or more independent variables are modeled statistically using the linear regression technique [8]. It is a linear approach to modeling the relationship between variables, assuming a linear relationship between the dependent variable and the independent variables.

In linear regression, the goal is to fit a linear equation to the data points in such a way that the sum of the squared differences between the observed and predicted values is minimized. This is known as the method of least squares.

The equation for a simple linear regression model with one independent variable is typically represented as:

\[ y = b_0 + b_1 x \]  (1)

Where \( y \) represents the dependent variable (also known as the response variable or target variable); \( x \) represents the independent variable (also known as the predictor variable); \( b_0 \) represents the y-intercept (the value of \( y \) when \( x \) is 0); \( b_1 \) represents the slope of the line (the change in \( y \) for a unit change in \( x \)). The linear regression model estimates the values of \( b_0 \) and \( b_1 \) that best fit the data by minimizing the sum of the squared differences between the observed \( y \)-values and the predicted \( y \)-values based on the given \( x \)-values.

Linear regression can also handle multiple independent variables, resulting in multiple linear regression. The equation for multiple linear regression can be represented as:

\[ y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \]  (2)

The dependent variable is represented by \( y \), the independent variables are represented by \( x_1, x_2, \ldots, \) and \( x_n \), and the coefficients or weights associated with each independent variable are represented by \( b_0, b_1, b_2, \ldots, \) and \( b_n \). Linear regression is commonly used for various purposes, including prediction, forecasting, and understanding the relationship between variables. The type and strength of the link between the independent and dependent variables are clarified. Additionally, linear regression can be extended to handle more complex relationships by incorporating polynomial terms, interaction terms, or applying other techniques such as regularization.

4 Empirical analysis

4.1 Splitting the data
Divide the data into TRAIN and TEST (in a ratio of 70:30) using the train_test_split method from the sklearn package.

4.2 Building the model

4.2.1 Recursive feature elimination
Recursive Feature Elimination (RFE) is a feature selection technique used in machine learning and data analysis [9]. It aims to identify the most important features or variables from a given dataset by recursively eliminating less significant features. The Linear Regression function from SciKit Learn will be utilized due to its compatibility with RFE (a utility from sklear) for our purposes.

4.2.2 Use ‘ststsmodels’ to building the model

Every time a model has been built, the VIF value and P-value need to be checked, and one variable should be removed to enhance the model. Repeat this process until the model reaches a satisfactory level.

Variance Inflation Factor is a statistical measure used to assess multicollinearity in regression analysis [10]. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can lead to unstable and unreliable coefficient estimates.

The VIF quantifies the degree of multicollinearity by measuring how much the variance of the estimated regression coefficient is inflated compared to what it would be if the variables were not correlated. A high VIF indicates high multicollinearity, suggesting that the corresponding independent variable may not provide unique or independent information in the regression model.

The formula for calculating the VIF of a variable \( i \) is as follows:
\[ VIF(i) = 1/(1 - R(i)^2) \]  

where \( R(i) \) square is the coefficient of determination (R-squared) obtained when regressing the variable \( i \) against the other independent variables in the model. The VIF value ranges from 1 upwards, with a value of 1 indicating no multicollinearity (perfectly independent variable) and higher values indicating increasing levels of multicollinearity.

VIF analysis is particularly relevant in regression analysis, where it helps identify variables that may be causing multicollinearity issues and affecting the stability and interpretability of the regression model.

In a linear regression model, the p-value is a statistical measure that quantifies the strength of evidence against the null hypothesis for each independent variable’s coefficient [11]. It aids in establishing the relationship between the independent variables and the dependent variable's statistical significance.

The null hypothesis in linear regression states that there is no relationship between the independent variables and the dependent variable, meaning the coefficient of the independent variable is zero. The coefficient is not zero, which is consistent with the alternative hypothesis that there is a meaningful link.

The p-value represents the probability of observing a coefficient as extreme as the one estimated in the regression model if the null hypothesis is true. In other words, it indicates the probability of obtaining the observed relationship or a stronger relationship purely due to random chance.

Typically, a smaller p-value indicates stronger evidence against the null hypothesis and suggests that the independent variable has a significant impact on the dependent variable. Conventionally, a common threshold for significance is a p-value less than 0.05 (5 percent). If the p-value is below this threshold, it is considered statistically significant, and the null hypothesis is rejected in favor of the alternative hypothesis.

4.2.3 Build the final model

Make that the final model's VIF value and P-value are good, that there is very little multicollinearity between the predictors, and that all of the predictors' p-values appear to be significant.

Here is the equation for the surface that fits the final model the best: 
\[
\text{cnt} = (0.084143 \times \text{const}) + (0.230846 \times \text{yr}) + (0.043203 \times \text{workingday}) + (0.563615 \times \text{temp}) + (-0.155191 \times \text{windspeed}) + (0.082706 \times \text{season}_2) + (0.128744 \times \text{season}_4) + (0.094743 \times \text{mnhth}_9) + (0.056909 \times \text{weekday}_6) + (-0.074807 \times \text{weathersit}_3) + (-0.306992 \times \text{weathersit}_3)
\]

4.3 Making prediction

In linear regression, the “predict” function is used to generate predictions or estimates of the dependent variable based on the values of the independent variables [12]. It allows you to apply the learned relationship between the independent and dependent variables to new or unseen data.

Using the predict function in Python with the scikit_learn library to predict in the final model.

4.4 Find R² value

The R-squared value, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables in the model [13]. It provides an assessment of how well the linear regression model fits the observed data. The R-squared value for testset is 0.8203, this seems to be a really good model that can very well ‘Generalize’ various datasets.

4.5 Results

According to the final model, the following are the top 3 predictor variables that affect the booking of a bike:

- **Temp:** A coefficient value of 0.563615 meant that for every unit increase in the temp variable, the number of bikes available for rent increased by 0.563615 units.
- **Weathersit-3:** The coefficient value of "-0.306992" showed that the number of bike rentals falls by 0.306992 units for every unit rise in the Weathersit3 variable.
- **Yr:** The bike rental numbers increased by 0.230846 units for every unit that the yr variable increased, according to a coefficient value of '0.230846'.

Therefore, it is advised to give these factors top priority when planning in order to achieve maximum booking.

The following qualities are also worthy of consideration:

- **Season-4:** According to a coefficient value of "0.128744," the number of bikes rented increases by 0.128744 units relative to season-1 for every unit increase in the season-4 variable.
- **Windspeed:** The bike rental numbers decreased by 0.155191 units for every unit rise in the windspeed variable, according to a coefficient value of "-0.155191".

**NOTE:** Weathersit-3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

5 Conclusion

In summary, this study developed a model for forecasting the demand for shared bikes in a bike-sharing system using linear regression analysis. The study identified significant factors that impact bike demand, such as temperature, season, month, weekday, and weather situation. The model-building process involved data cleaning, feature selection, and rescaling of features, resulting in an accurate and reliable predictive model.

The findings of this research have practical implications for bike-sharing system operators and urban transportation planners. By understanding the key factors that influence bike demand, operators can optimize their resources, improve service quality, and enhance user satisfaction. Additionally, this research contributes to the promotion of sustainable transportation solutions by...
encouraging the use of bike-sharing systems as an eco-friendly alternative to traditional modes of transportation.

It's crucial to recognize the limits of this study, though. The model's performance may vary in different geographical locations and cultural contexts. Further research and validation are necessary to ensure the generalizability of the findings. Nevertheless, this research provides a solid framework for future studies on the subject of predicting bike-sharing demand and provides useful information for both practitioners and regulators.

References

5. kaggle datasets. Available at https://www.kaggle.com/datasets/gauravduttakiit/bike-sharing
13. T. J. Smith, C. M. McKenna. Multiple Linear Regression Viewpoints, 39, 2 (2013)