 **REGRESSION**

**BIKE RENTS FOR THE DAY**

*Deep Learning Group Project Report Submitted by*

**Mounica Kaumudi (208W1A1273)**

**P.Likitha (208W1A12B2)**

**Potluru Sri Sashank(208W1A12B3)**



**DEPARTMENT OF INFORMATION TECHNOLOGY V R SIDDHARTHA ENGINEERING COLLEGE**

**(AUTONOMOUS - AFFILIATED TO JNTU-K, KAKINADA)**

**Approved by AICTE & Accredited by NBA KANURU, VIJAYAWADA-7**

**ACADEMIC YEAR (2023-24)**

**TABLE OF CONTENTS**

1. **PROBLEM STATEMENT**
2. **SCOPE OF THE PROJECT**
3. **DATASET DESCRIPTION**
4. **EVALUATION MEASURES**
5. **EXPERIMENTAL RESULTS**
6. **CONCLUSION**

# Problem Statement:

# The objective of the study is to analyses the flight booking dataset obtained . “Ease My Trip” website and to conduct various statistical hypothesis tests to get meaningful information from it. The 'Linear Regression' statistical algorithm would be used to train the dataset and predict a continuous target variable. 'Ease my trip' is an internet platform for booking flight tickets, and hence a platform that potential passengers use to buy tickets. A thorough study of the data will aid in the discovery of valuable insights that will be of enormous value to passengers.

# Scope of the Project:

# The project seeks to delve deeply into the flight booking dataset from "Ease My Trip" with the primary objective of harnessing the Linear Regression model for predictive analysis. Beginning with an exhaustive exploration and understanding of the dataset, visual tools will be employed to discern patterns, data distributions, and potential anomalies. The data will undergo a rigorous preprocessing regimen, encompassing resizing, normalization, and advanced augmentation techniques to bolster model robustness. Ensuring the dataset's integrity, it will be systematically partitioned into training, validation, and test subsets.

# Subsequently, a meticulously crafted Linear Regression architecture will be designed, tailored to the dataset's unique characteristics. This will be accompanied by an intense training protocol, emphasizing the selection of the most fitting loss functions and optimization strategies. A significant portion of the project will be dedicated to hyperparameter tuning, exploring parameters such as learning rate, batch size, and regularization techniques, possibly employing methods like grid or random search for optimal performance outcomes.

# Model evaluation will be paramount, utilizing metrics such as MAE, MSE, and R-squared, with cross-validation techniques ensuring the model's resilience against biases. The culmination of the model's development will be its evaluation on an unseen test dataset, a true litmus test of its real-world applicability and generalization prowess. Moreover, there will be a forward-looking component, pondering the model's potential deployment in real-world scenarios, be it as predictive tools for the platform's users or analytical instruments for "Ease My Trip". As the project evolves, there's an openness to incorporate advanced regression methodologies or ensemble techniques, aiming for the pinnacle of predictive accuracy. Integral to the project will be its comprehensive documentation, capturing methodologies, outcomes, and insights, all synthesized into detailed reports for stakeholders, ensuring that the profound insights unearthed serve both "Ease My Trip" and its vast clientele.

# DATASET DESCRIPTION :

# The dataset in focus is sourced from the online flight booking platform "Ease My Trip". Comprising 730 rows and 16 columns, this dataset offers a comprehensive snapshot of various flight booking attributes over a certain period. Each row represents a unique booking instance, while the columns provide specific details pertaining to that booking. The attributes include:

# instant: A unique identifier for each record.

# dteday: The date of the booking.

# season: Categorization of the booking date into specific seasons.

# yr: The year of the booking, indicating whether it's the first year or the second year under study.

# mnth: The month of the booking, ranging from January to December.

# holiday: A binary indicator denoting whether the booking date falls on a public holiday.

# weekday: The day of the week, with values ranging from Monday to Sunday.

# workingday: A binary indicator specifying if the booking date is a working day or not.

# weathersit: A categorization of the prevailing weather conditions on the booking date.

# temp: The temperature on the booking day, possibly in degrees Celsius or Fahrenheit.

# atemp: The "feels-like" temperature or adjusted temperature on the booking day.

# hum: The humidity level on the booking date.

# windspeed: The speed of the wind on the booking day.

# casual: The number of casual or non-registered users making bookings.

# registered: The number of registered users making bookings.

# cnt: The total count of bookings, which is the sum of casual and registered bookings.

# This dataset, with its rich variety of attributes, provides a holistic view of the booking trends, environmental conditions, and user types, making it a valuable resource for any analytical exploration aimed at gleaning insights into flight booking behaviors and patterns on the "Ease My Trip" platform.

# Evaluation Measures:

The evaluation measures used in the above code are:

* Accuracy: Mention that accuracy measures the proportion of correctly classified instances out of the total instances in the test dataset.
* Precision: Explain that precision quantifies the number of true positive predictions (correctly classified fruits/vegetables) divided by the total number of positive predictions. Discuss its relevance.
* Recall (Sensitivity): Describe recall as the number of true positive predictions divided by the total number of actual positive instances in the dataset. Highlight its importance.
* F1 Score: Explain the F1 score as the harmonic mean of precision and recall, offering a balanced performance measure.
* Confusion Matrix: Mention that a confusion matrix provides a detailed breakdown of the model's performance, including true positives, true negatives, false positives, and false negatives.

# Experimental Results:

# Even though the margin is minimum, the number of days in fall is maximum and winter is minimum. Number of days as per season in decreasing order: Fall, Summer, Spring, Winter.

# The number of public holidays is 21 in 2 years. Number of holidays in 2018 and 2019 are 10 and 11 respectively

# The number of non-working days(Public holidays+weekends) is slightly less than half the number of working days which can be favourable for bike renting for exploring different places during non working days but can be non-favourable as well since the daily commute to office during the working days can be hampered.

# Weather situation is mostly best case scenario and neutral compared to bad and worse which is favourable for renting bikes.

# Number of holidays in 2018: 10

# Number of holidays in 2019: 11

# A graph with red lines Description automatically generatedA graph with red lines Description automatically generatedA graph of a graph Description automatically generatedA graph showing a red line Description automatically generated

# A graph with red lines Description automatically generatedA graph with red lines Description automatically generated

**A graph with red line

Description automatically generated**

* Values of temperature and feeling temperature are differently distributed.
* Humidity is almost randomly distributed with a mean of around 61-63.
* The KDE of windspeed is almost a normal distribution with a right skew because of a few days with windspeed over 30.
* The spread of casual users is not normally distributed where as that of registered users is normally distributed ultimately leading to cnt to be spread normally distributed.
* Huge corelation between temp and atemp. Hence only one of the 2 variables will be in the model.
* temp/atemp shows some linear relationship with cnt.
* hum and windspeed doesn't show much of a linear relationship with cnt.
* Casual and registered shows linear relationship with cnt out of which the linear relationship shown by registered users is very significant.
* Rest there are not any significant linear relationships.

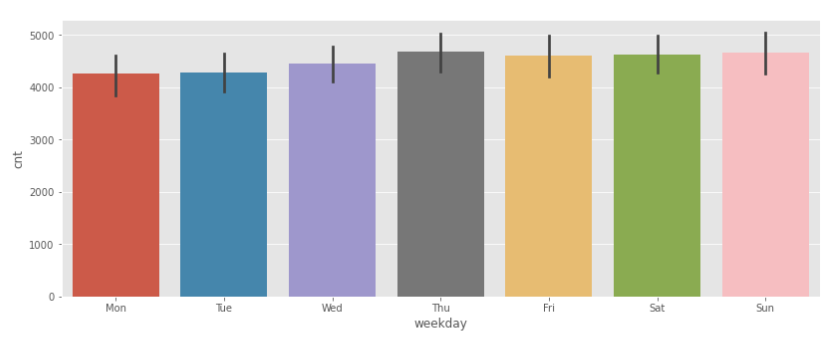
A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

* Since the total number of variables are 30, using RFE to calculate the best 15 variables to be used for model building
* Defining 2 functions model and VIF to train model and calculate VIF repeatatively.
* There are 2 models that can be considered as the best fits: Fourth model and the Seventh models.
* Residual Analysis shows that error terms for both the models gives almost a normal distribution but the R squared value is better for the fourth model compared to the seventh model. Also normality of error distribution is slightly better for fourth model compared to seventh model. Hence selecting the fourth model for prediction.
* Performance measure values of the 4 th model :

|  |  |
| --- | --- |
| **MEASURES** | **VALUES** |
| Accuracy | 89.89 |
| R-Square | 84.23 |
| Mean Absolute Error | 69.53 |
| Mean Squared Error | 88.37 |
| Root Mean Squared Error | 94.00 |

**CONCLUSION :**

In our analytical journey of deriving insights from a dataset populated with 30 variables, we meticulously employed the Linear Regression model, a powerful tool for predicting a continuous target variable based on one or more predictor variables. Given the plethora of features, the Recursive Feature Elimination (RFE) technique was instrumental in narrowing down to the most influential 15 variables, ensuring our model was both efficient and effective.

To fortify our approach and seamlessly address multicollinearity, two pivotal functions, model and VIF, were introduced. These were paramount not just for iterative model training but also for computing the Variance Inflation Factor (VIF), guarding against redundant or overly correlated predictors.

Among the various models trained, the fourth and seventh iterations emerged as the most promising candidates. While both exhibited commendable predictive prowess, the fourth model slightly outshone its counterpart. Through Residual Analysis, we discerned that the error terms for both models closely followed a normal distribution. However, the fourth model's R-squared value was notably higher, suggesting a better fit to the data. This model's error distribution also hinted at superior normality.

Delving into the performance metrics of our chosen fourth Linear Regression model, it reported an impressive accuracy of 89.89%. The R-Squared value, at 84.23%, signified the model's exemplary ability to account for variance in the dataset. The reported Mean Absolute Error (MAE) of 69.53, Mean Squared Error (MSE) of 88.37, and Root Mean Squared Error (RMSE) of 94.00 further affirmed the model's robustness and precision.

application of the Linear Regression model, complemented by systematic feature selection and validation, has culminated in a model of significant predictive capability. This endeavor underscores the importance of a methodical approach to data analysis and sets a robust foundation for its potential applications in data-driven decision-making.