# Cab Fare Prediction

Shriniwas Raju Jagadabhi
*Dept. of Computer Science and Engineering*
*IIIT Naya Raipur*
Chattisgarh, India
shriniwas21100@iiitnr.edu.in

Sontu Akshath Rishi
*Dept. of Computer Science and Engineering*
*IIIT Naya Raipur*
Chattisgarh, India
sontu21100@iiitnr.edu.in

***Abstract*—This project aims to predict cab prices based on various features such as distance, time, and location. A publicly available data-set containing information about cab rides is used, and a detailed analysis of the data-set is performed, including data cleaning, preprocessing, and exploratory data analysis. Different regression models such as linear, multiple linear, and polynomial regression are compared, and various performance metrics are used to evaluate the models. This project provides insights into the factors that affect cab fares and analysis of different regression model while comparing them.This project mainly include analysis of data and model.**

***Index Terms*—Simple Linear Regression, Multiple Regression, Polynomial Regression, Data Visualization, Cab Fare Prediction**

## I. Introduction

Regression analysis is a statistical technique used to analyze the relationship between a dependent variable and one or more independent variables. Regression analysis is widely used in various fields, including economics, finance, healthcare, and marketing, to understand the relationships between variables and make predictions based on the data.

Regression analysis can be classified into two types: simple regression and multiple regression. Simple regression involves analyzing the relationship between two variables, whereas multiple regression involves analyzing the relationship between a dependent variable and multiple independent variables.

Linear regression is a commonly used regression analysis technique that assumes a linear relationship between the dependent variable and independent variables. It is used to predict the value of a dependent variable based on the values of one or more independent variables.

In addition to linear regression, there are other regression techniques such as polynomial regression, logistic regression, and time series regression. Polynomial regression is used to model nonlinear relationships between the dependent variable and independent variables.

Regression analysis is an essential tool for data analysts and data scientists, as it helps them to understand the relationships between variables and make predictions based on the data. Regression analysis is also used in machine learning and artificial intelligence to develop predictive models.

## II. Dataset Description

Understanding of data is the very first and important step in the process of finding solution of any business problem. Here in our case our company has provided a data set with following features, we need to go through each and every variable of it to understand and for better functioning. Size of Dataset Provided: - 16067 rows, 7 Columns (including dependent variable) Below mentioned is a list of all the variable names with their meanings:

| Variables | Description |
|---|---|
| **fare_amount** | Fare amount |
| **pickup_datetime** | Cab pickup date with time |
| **pickup_longitude** | Pickup location longitude |
| **pickup_latitude** | Pickup location latitude |
| **dropoff_longitude** | Drop location longitude |
| **dropoff_latitude** | Drop location latitude |
| **passenger_count** | Number of passengers sitting in the cab |

Fig. 1. Dataset Attributes

## III. Data Preprocessing

### A. Data exploration and Cleaning

The very first step which comes with any data science project is data exploration and cleaning which includes following points as per this project:
a. Separate the combined variables.
b. As we know we have some negative values in fare amount so we have to remove those values.
c. Passenger count would be max 6 if it is a SUV vehicle not more than that. We have to remove the rows having passengers counts more than 6 and less than 1.
d. There are some outlier figures in the fare (like top 3 values) so we need to remove those.
e. Latitudes range from -90 to 90. Longitudes range from -180 to 180. We need to remove the rows if any latitude and longitude lies beyond the ranges.

### B. Creating some new variables from the given variables

Here in our data set our variable name pickup_datetime contains date and time for pickup. So we tried to extract some important variables from pickup_datetime:

- Year

- Month
- Date
- Day of Week
- Hour
- Minute

Also, we tried to find out the distance using the haversine formula which says: The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes. Important in navigation, it is a special case of a more general formula in spherical trigonometry, the law of haversines, that relates the sides and angles of spherical triangles.

### C. Selection of Variables

Now as we know that all above variables are of now use so we will drop the redundant variables:

- pickup_datetime
- pickup_longitude
- pickup_latitude
- dropoff_longitude
- dropoff_latitude
- Minute

Now only following variables we will use for further steps:

| VariableNames | Variable DataTypes |
|---|---|
| fare_amount | float64 |
| passenger_count | object |
| year | object |
| Month | object |
| Date | object |
| Day of Week | object |
| Hour | object |
| distance | float64 |

Fig. 2. Dataset Attributes after Preprocessing

### D. Feature Scaling

Here we have only two continuous variables : fare_amount and distance remaining are categorical varaibles. Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution. Here we tried to show the skewness of our variables and we find that our target variable absenteeism in hours having is one sided skewed so by using log transform technique we tried to reduce the skewness of the same. Below mentioned graphs shows the probability distribution plot to check distribution before log transformation:

Below mentioned graphs shows the probability distribution

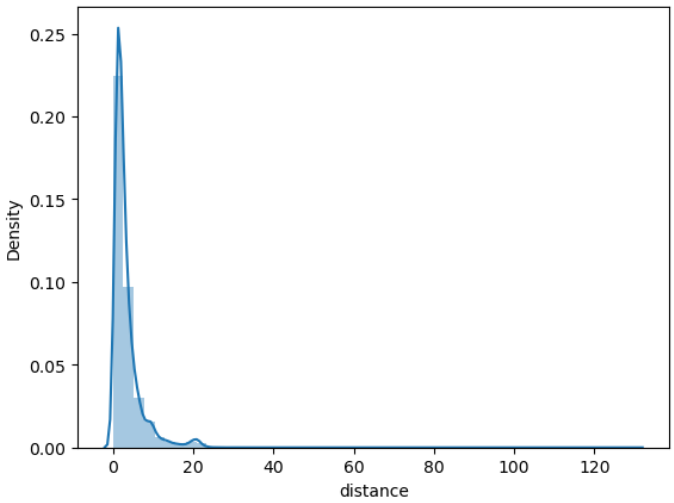


Fig. 3. Distribution of distance before log transformation

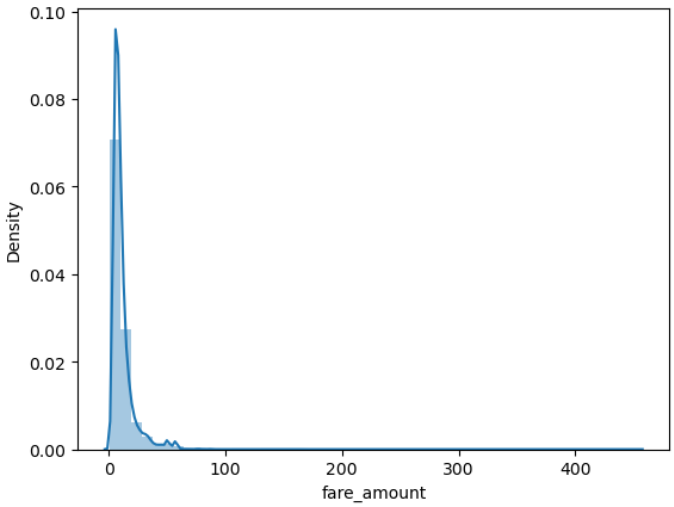


Fig. 4. Distribution of fare before log transformation

plot to check distribution after log transformation:

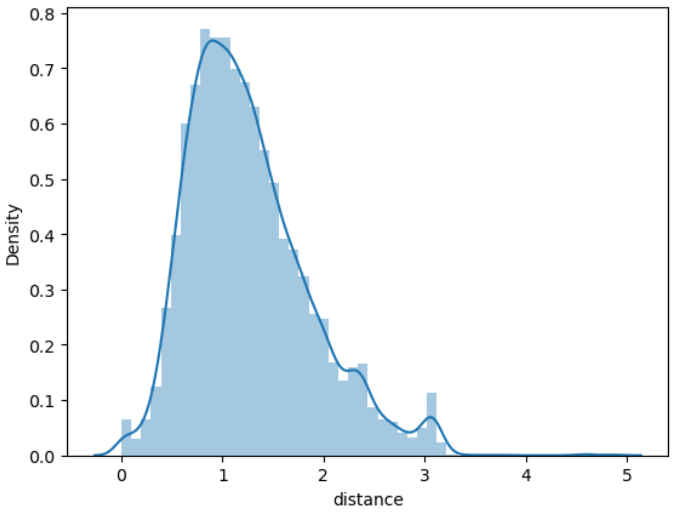


Fig. 5. Distribution of distance after log transformation

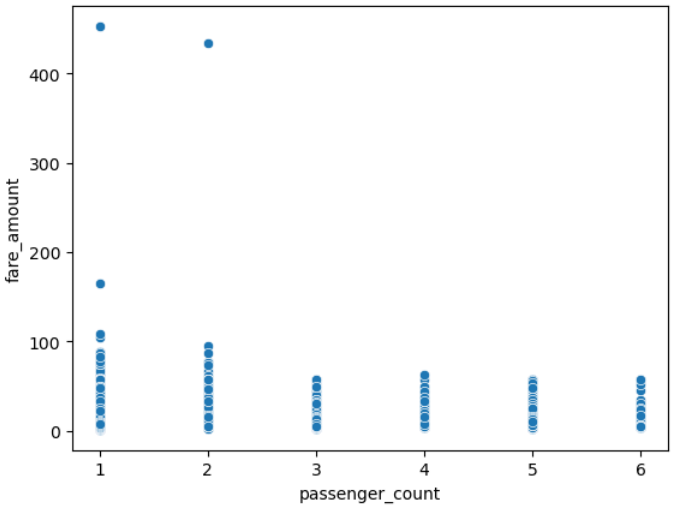


Fig. 7. No. of Passengers vc Fare amount

From the above plot we can conclude that highest Fare are coming from single and double travelling passengers.

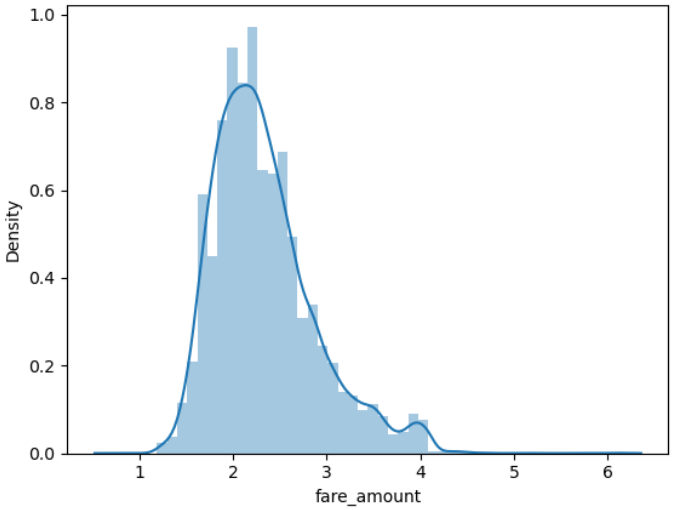


Fig. 6. Distribution of fare after log transformation

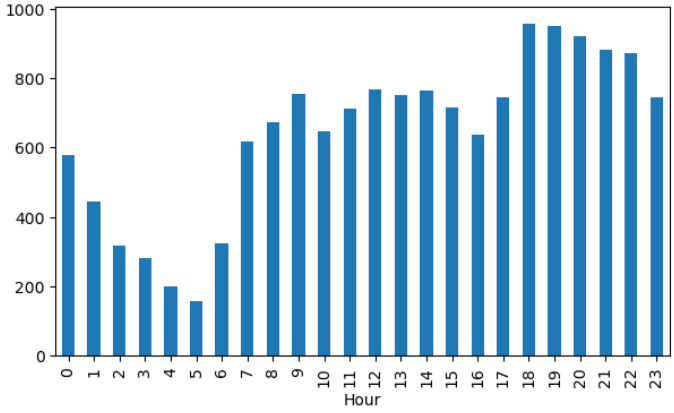


Fig. 8. Hour vs No. of Cabs

Lowest cabs at 5 AM and highest at and around 7 PM i.e the office rush hours

## IV. Data Visualization

In this project, we utilized data visualization techniques to explore and communicate the findings from our analysis. By creating various charts, graphs, and visualizations, we were able to effectively communicate complex information to our audience in a clear and concise manner. Our use of data visualization allowed us to identify patterns and trends within the data that might have been missed otherwise, and to gain insights that helped inform our decision-making process.

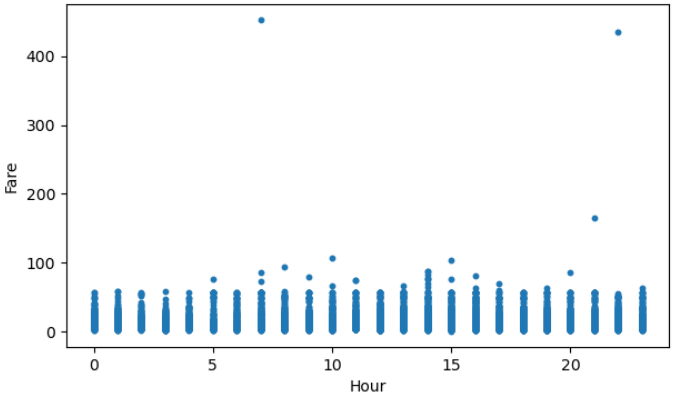


Fig. 9. Scatterplot of Hour vs Fare

From the above plot We can observe that the cabs taken at 7 am and 23 Pm are the costliest. Hence we can assume that cabs taken early in morning and late at night are costliest

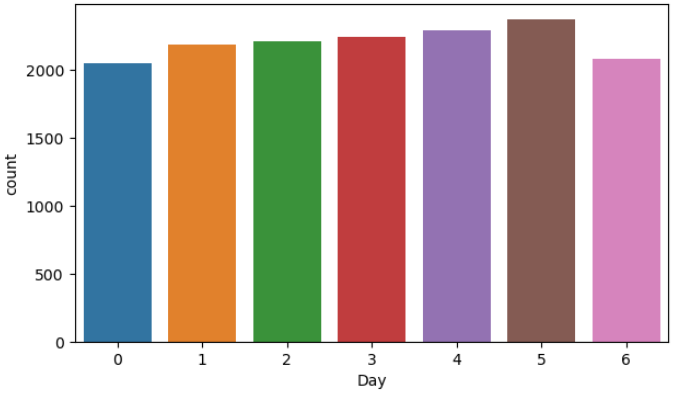


Fig. 10. Day vs No. of Cabs

The day of the week does not seem to have much influence on the number of cabs ride.

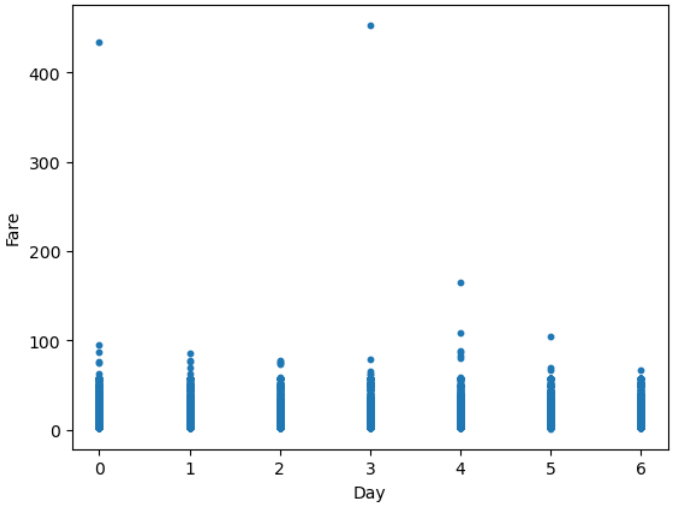


Fig. 11. Scatterplot of Day vs Fare

The highest fares seem to be on a Sunday and Thursday, and the low on Wednesday and Saturday. May be due to low demand of the cabs on saturdays the cab fare is low and high demand of cabs on sunday and monday shows the high fare prices.

### *A. Calculating the Distance*

We have calculated distance using Havrsine formula . Haversine distance is a formula used to calculate the distance between two points on a sphere, such as the Earth. It is commonly used in geographic applications and can help in determining the shortest distance between two locations on a map.

The formula takes into account the latitude and longitude of the two points and calculates the great-circle distance between them. The Haversine formula is based on the law of haversines, which states that the haversine of a central angle in a sphere is equal to the sum of the haversines of its two complementary angles.

The Haversine formula is given as follows:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{lat2-lat1}{2}\right) + \cos(lat1)\cos(lat2)\sin^2\left(\frac{lon2-lon1}{2}\right)}\right)$$

- d is the distance between the two points in kilometers
- r is the radius of the Earth (mean radius = 6,371km)
- lat1 and lat2 are the latitudes of the two points in radians
- lon1 and lon2 are the longitudes of the two points in radians

The formula calculates the distance between the two points as an arc along a sphere, which is why it is particularly useful for geographic applications and since our data set contains such values like latitude and logitude we have used this formula.

This visualization of the new parameter distance -

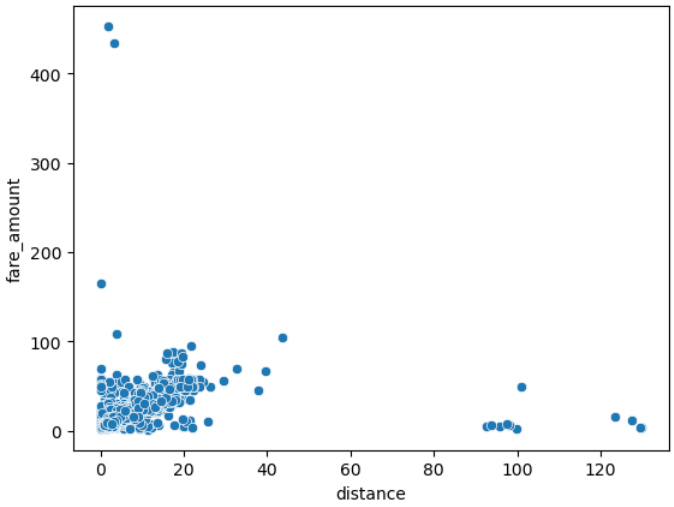


Fig. 12. Scatterplot between Distance and Fare

### B. Correlation Matrix

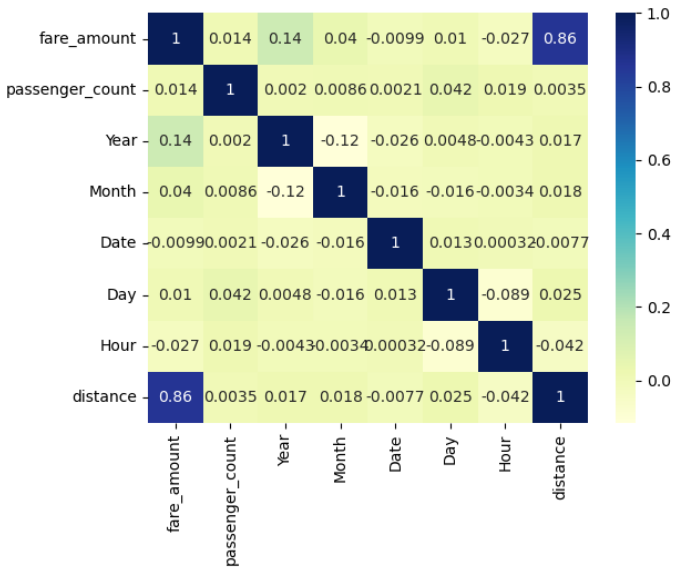


Fig. 13. Correlation Matrix

Here we can observe distance and fare_amount are highly correlated as compared to other attributes.

## V. Methodology

### A. Simple Linear Regression

Simple linear regression is a commonly used statistical method to model the relationship between a dependent variable and a single independent variable. In other words, it is used to establish a linear relationship between two variables, where the aim is to predict the value of the dependent variable based on the value of the independent variable.

From the correlation matrix it is clear that distance is highly correlated with our target variable than any other attribute, so we applied simple linear regression by taking distance as input variable and fare_amount as target variable by splitting the training and testing data in 80:20 ratio.

We can see the regression line plotted between training and testing data as follows:


Fig. 14. Fare vs Distance (Training Data)


Fig. 15. Fare vs Distance (Test Data)

For this model Root Mean Squared Error (RMSE) value is 0.2558318472148484 and $R^2$ Score is 0.7638474817524837.

### B. Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. Multiple regression is an extension of simple linear regression. It is used as a predictive analysis, when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).

Here again we splitted the training and testing data into 80:20 ratio respectively and for this model the Root Mean Square Error (RMSE) value is 0.24540661786977663 and $R^2$ score is 0.7827019104296612.

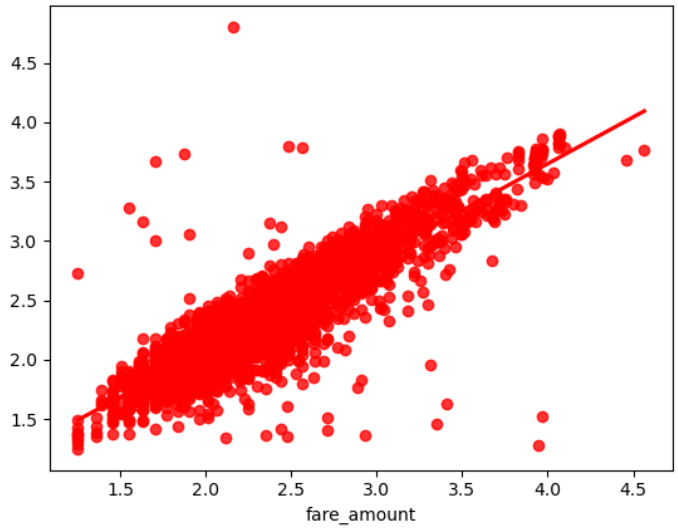


Fig. 16. Actual vs Predicted using Multiple Linear Regression

### C. Polynomial Regression

Polynomial regression is a type of regression analysis used to model the relationship between the independent variable and dependent variable as an nth degree polynomial function. Unlike simple linear regression, which assumes a linear relationship between the variables, polynomial regression can capture non-linear relationships.

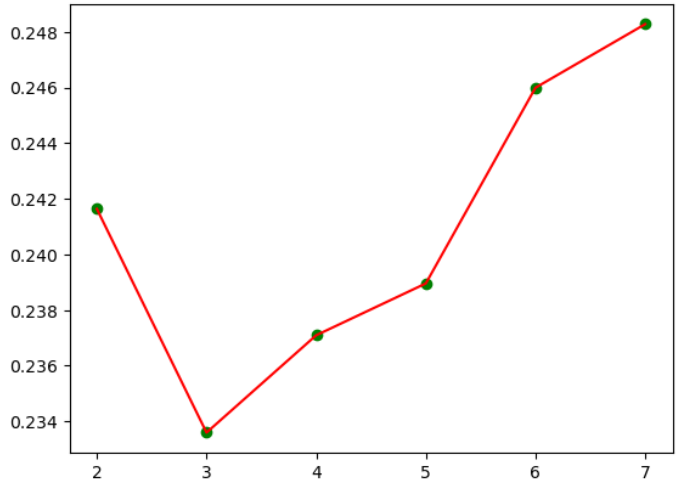


Fig. 17. Degree of Polynomial vs Error

Here we can observe that the error is minimum for the polynomial regression with degree 3 while compared to all the other degrees. So we choose degree 3 for our polynomial regression and apply it on our data to predict values and compare it with actual values. This is graph containing the predicted vs actual fare price values using the polynomial regression of degree 3.

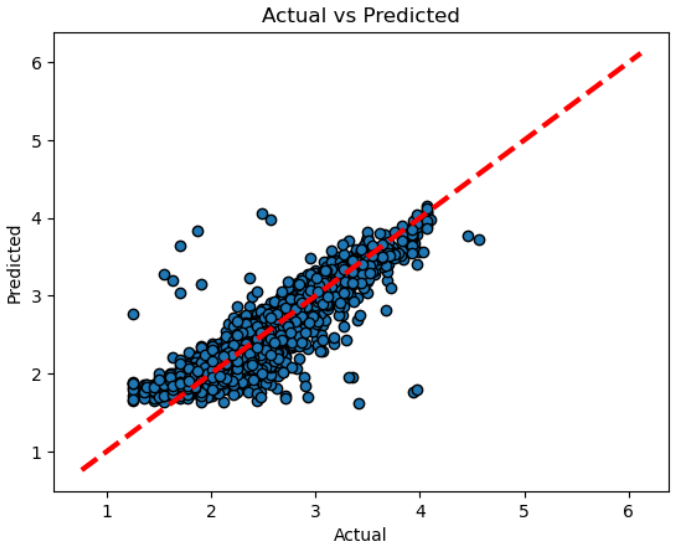


Fig. 18. Actual vs Predicted

For this model Root Mean Squared Error (RMSE) value is 0.2335959094064254 and $R^2$ Score is 0.803114453876693.

## VI. Observation

After evaluating the machine learning models of cab fare prediction using simple linear, multiple linear, and polynomial regression, it was observed that the polynomial regression model outperformed the other models. This was based on the mean values of R-squared score and Root Mean Squared Error (RMSE). The R-squared score of the polynomial regression model was 0.803, which indicates that 85 percentage of the variance in the dependent variable (cab fare) is explained by the independent variables. The R-squared score for simple linear regression was 0.763 and for multiple linear regression, it was 0.782. Additionally, the RMSE for polynomial regression was lower than that of the other models, indicating that the polynomial regression model had less error in its predictions.

## VII. Conclusion

In conclusion, we can say that the the polynomial regression is the better model for cab fare prediction among the three models evaluated. We can say that polynomial is better to use for data since it can effectively capture non-linear relationships between the input variables and the target variable .

## Acknowledgment

We would like to thank Dr. Mallikharjuna Rao Sir for his guidance and support throughout this project. His expertise and insights were instrumental in helping us to produce meaningful results.

## References

[1] https://www.kaggle.com/datasets/pankajkumar90/cab-fare-dataset?resource=download