Aim- Perform Regression Analysis for the Uber ride.

Objectives- Predict the price of the Uber ride from a given pickup point to the agreed drop-off location. Perform following tasks:

1. Pre-process the dataset.

2. Identify outliers.

3. Check the correlation.

4. Implement linear regression and ridge, Lasso regression models.

5. Evaluate the models and compare their respective scores like R2, RMSE, etc

Theory- Regression is a tool that allows you to estimate how the dependent variable changes as the independent variable(s) change. Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line.

Regression models can be used for many purposes:

- Evaluating the effect of an independent variable on a dependent variable.
- Forecasting future values of the dependent variable based on prior observations of both variables

Terminologies Related to the Regression Analysis:

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.

- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Types of Regression

There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables

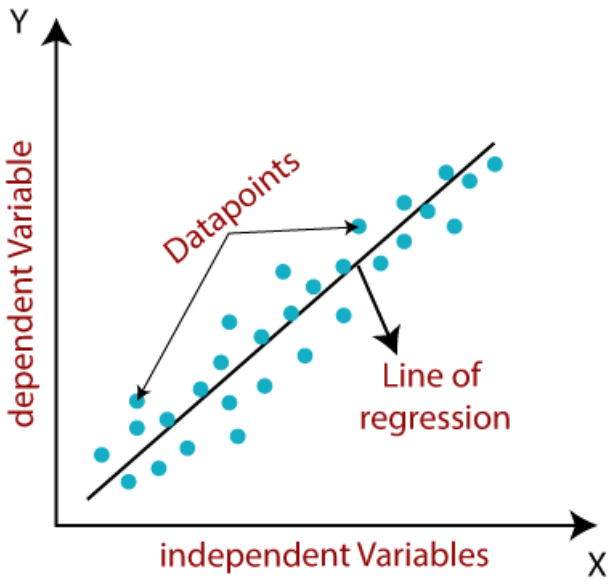
1. **Linear Regression**
2. **Logistic Regression**
3. **Polynomial Regression**
4. **Support Vector Regression**
5. **Decision Tree Regression**
6. **Random Forest Regression**
7. **Ridge Regression**
8. **Lasso Regression**

A) Linear Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

$y= a_0+a_1x+ \varepsilon$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (Predictor Variable)

a_0= Intercept of the line

a_1= Linear regression coefficient

ε= Random Error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression-

1) Simple Linear Regression- If a single independent variable is used to predict the value of a numerical dependent variable, and then such a Linear Regression algorithm is called Simple Linear Regression.
2) Multiple Linear Regression- If more than one independent variable is used to predict the value of a numerical dependent variable, and then such a Linear Regression algorithm is called Multiple Linear Regression.

Applications of linear regression:

- Analyzing trends and sales estimates
- Salary forecasting
- Real estate prediction
- Arriving at ETAs in traffic.

B) Lasso regression

Lasso regression, also known as L1 regularization, is a linear regression technique that helps prevent overfitting and performs feature selection by adding a penalty term to the linear regression cost function. This penalty term is based on the absolute values of the regression coefficients. Regression consists of following methods:

- Data Preparation
- Data Preprocessing
- Model Selection
- Feature Selection
- Model Evaluation
- Hyperparameter Tuning
- Interpretation

C) Ridge regression

Ridge regression is a variant of linear regression that adds a regularization term to the traditional linear regression cost function. This regularization term, represented by the L2 norm of the regression coefficients, helps prevent overfitting by penalizing large coefficient values.

1. Regularization: Ridge regression adds a penalty term based on the sum of the squared values of the regression coefficients (L2 regularization) to the linear regression cost function.

2. Overfitting Prevention: The regularization term discourages the model from assigning excessive importance to any single predictor variable, which helps prevent overfitting, especially when dealing with high-dimensional data.

3. Shrinking Coefficients: Ridge regression tends to shrink the coefficients towards zero but rarely forces them to be exactly zero, meaning it doesn't perform feature selection as aggressively as Lasso regression.

4. Hyperparameter: The strength of regularization (alpha) in ridge regression is a hyperparameter that needs to be tuned. Smaller values of alpha lead to results closer to standard linear regression, while larger values increase the regularization effect.

5. Mathematical Formulation: The ridge regression cost function seeks to minimize the sum of squared errors (like linear regression) while also minimizing the sum of squared coefficients times the regularization parameter alpha.

Conclusion- We implements linear regression, Ridge, and Lasso regression to find out the price of the Uber ride from a given pickup point to the agreed drop-off location.