Capstone Project- II

Bike Share Demand Prediction

INTRODUCTION TO SUPERVISED MACHINE LEARNING

Supervised learning is the types of machine learning in which machines are trained using well "labeled" training data, and on basis of that data, machines predict the output. The labeled data means some input data is already tagged with the correct output. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

Insights of project:-

We are provided with a data set containing relevant data required to predict the demand for bike. Utilising the concepts of EDA(Exploratory Data Analysis),supervised learning algorithms ,Data Visualisation we will be using the data available to analyze the factors affecting the demand .

Algorithm followed:

1.Importing the necessary packages and libraries.

2.Mounting the drive for importing the data.

3.Checking for missing, Nan values, Null values.

4.Observing the data types .

5.Observing the correlation among independent variables.

6. Exploring the data set.

7.Exploring the categorical values, numerical features from data set.

8.Exploring different target variable.

9.Splitting the data and training the data.

10.Observing the results.

1.Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

.

2.Feature Engineering

Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy**.** Feature engineering is required when working with machine learning models. Regardless of the data or architecture, a terrible feature will have a direct impact on our mode.

3.Variance Inflation Factor

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

4.Multicollinearity

Multicollinearity, or collinearity, is the existence of near-linear relationships among the independent variables. For example, suppose that the three ingredients of a mixture are studied by including their percentages of the total. These variables will have the (perfect) linear relationship: P1 + P2 + P3 = 100. During regression calculations, this relationship causes a division by zero which in turn causes the calculations to be aborted. When the relationship is not exact, the division by zero does not occur and the calculations are not aborted. However, the division by a very small quantity still distorts the results. Hence, one of the first steps in a regression analysis is to determine if multicollinearity is a problem.

Effects of Multicollinearity

Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression

coefficients, give false, nonsignificant, pvalues, and degrade the predictability of the model (and that's just for starters).

5. Homoscedasticity

In regression analysis , homoscedasticity means a situation in which the variance of the dependent variable is the same for all the data. Homoscedasticity is facilitates analysis because most methods are based on the assumption of equal variance.
The best way for checking homoscedasticity is to make a scatterplot with the residuals against the dependent variable.

Now we will switch to building different types of models. Firstly we will train it by feeding and then will find the accuracy.Supervised machine learning Regression models are:-

1.Linear Regression

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:

2.Lasso Regression

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean..Lasso Regression uses L1 regularization technique . It is used when we have more features because it automatically performs feature selection.The word "LASSO" stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularisation of data models and feature selection.

Limitation of Lasso Regression:

- Lasso sometimes struggles with some types of data. If the number of predictors *(p)* is greater than the number of observations *(n)*, Lasso will pick at most n predictors as non-zero, even if all predictors are relevant (or may be used in the test set).
- If there are two or more highly collinear variables then LASSO regression select one of them randomly which is not good for the interpretation of data.

3.Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable. Another biased regression technique, principal components regression, is also available in NCSS. Ridge regression is the more popular of the two methods.

Limitation of Ridge Regression:

Ridge regression decreases the complexity of a model but does not reduce the number of variables since it never leads to a coefficient been zero rather only minimizes it. Hence, this model is not goods for feature reduction.

4.Elastic Net Regression

- The elastic net method performs variable selection and regularization simultaneously.The elastic net technique is most appropriate where the dimensional data is greater than the number of samples used.Groupings and variables selection are the key roles of the elastic net technique.

5.Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.

In this project the maximum accuracy we have obtained through this model.

6.Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

Conclusion

For our project we observe_we observe the models are performing well with the test data with good results.• Random forest Regressor, Random forest Regressor with gridsearchCV and XGB Regressor gives the highest R2 score of 92%, 89% and 86% recpectively for test dataset.• Feature Importance value for Random Forest and Gradient Boost are different.

However, this is not the ultimate end. As this data is time dependent, the values for variables like temperature, windspeed, solar radiation etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time. Therefore, having a quality knowledge and keeping pace with the ever evolving ML field would surely help one to stay a step ahead in future.

References:

1.https://ncss-wpengine.netdna-ssl.com/

2.https://www.kdnuggets.com/

3.https://corporatefinanceinstitute.com/