Seoul Bike Sharing Demand Prediction

By: Ayush

Introduction

Currently Rental bikes are introduced in many urban cities for the enhancement ofmobility comfort. It is important to make the rental bike available and accessible tothe public at the right time as it lessens the waiting time. Eventually, providing thecity with a stable supply of rental bikes becomes a major concern. The crucial partis the prediction of bike count required at each hour for the stable supply of rentalbikes.

Problem statement

We are tasked with predicting the number of bikes rented each hour so as to makean approximate estimation of the number of bikes to be made available to the public given a particular hour of the day.

Overview of data

We are given the following columns in our data:

1. **Date: year-month-day**
2. **Rented Bike count - Count of bikes rented at each hour**
3. **Hour - Hour of the day**
4. **Temperature-Temperature in Celsius**
5. **Humidity - %**
6. **Wind Speed - m/s**
7. **Visibility - 10m**
8. **Dew point temperature - Celsius**
9. **Solar radiation - MJ/m2**
10. **Rainfall - mm**

11. **Snowfall - cm**
12. **Seasons - Winter, Spring, Summer, Autumn**
13. **Holiday - Holiday/No holiday**
14. **Functional Day - No(Non Functional Hours), Yes(Functional hours)**

Steps involved

I. Performing EDA (exploratory data analysis)

A. Exploring head and tail of the data to get insights on the given data.

B. Looking for null values and removing them if it affects theperformance of the model.

C. Converting the data into appropriate data types to create a regressionmodel.

D. Creating dataframes which help in drawing insights from the dataset.

E. Creating more columns in our dataset which would be helpful forcreating model .

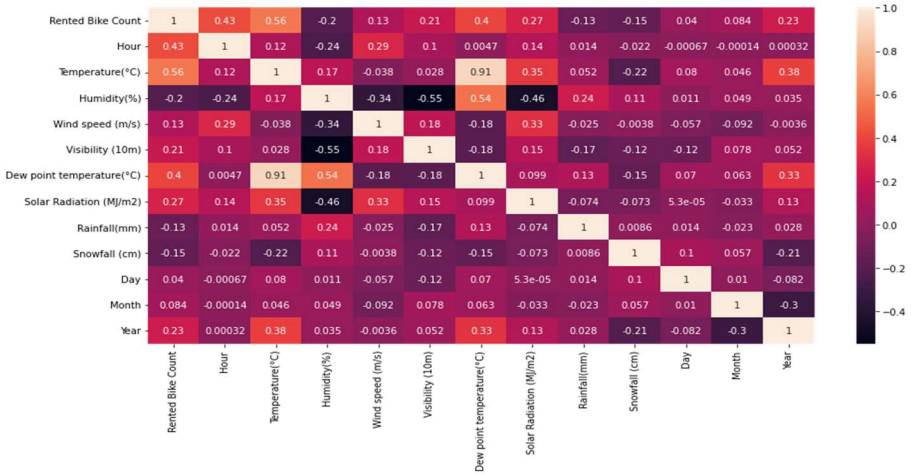
F. Encoding the string type data to better fit our regression model.

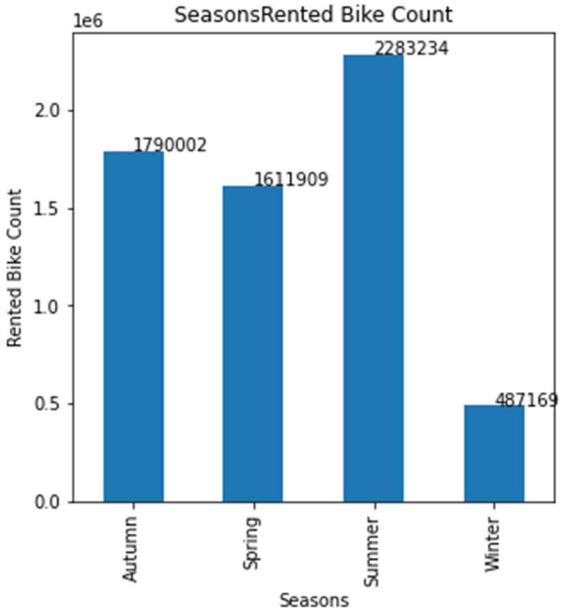
G. Calculating inter-quartile range and filtering our data.

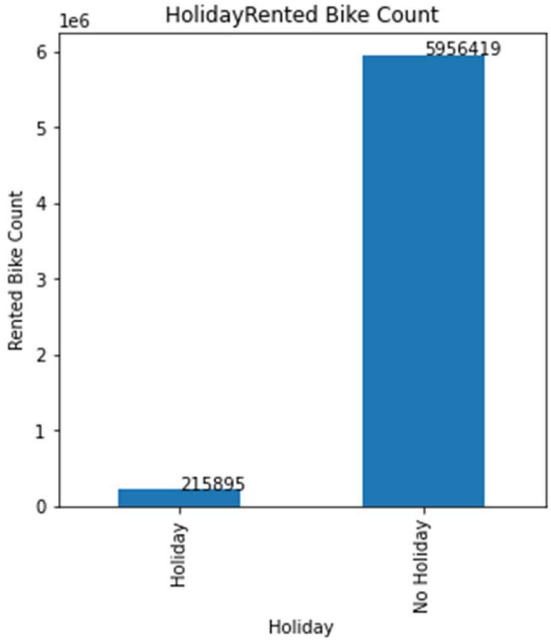
H. Extracting correlation heatmap and calculating VIF to removecorrelated and multicollinear variables.

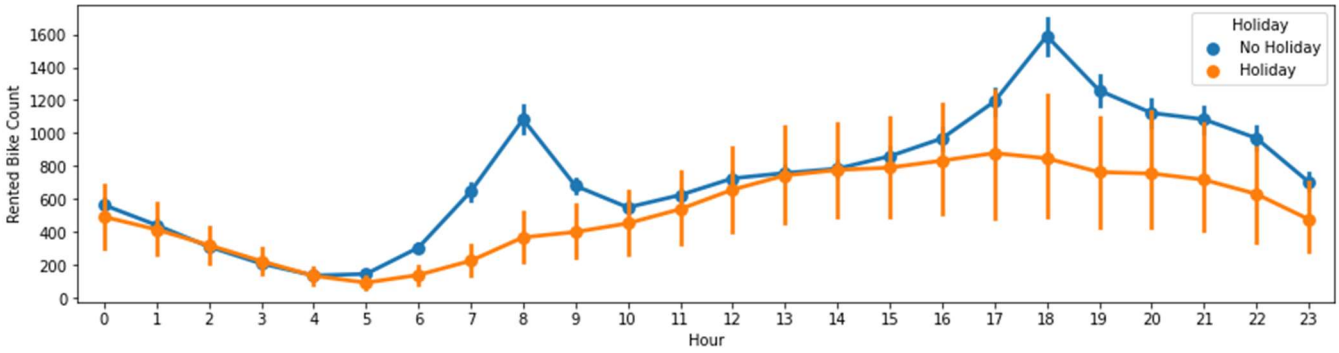
II. Drawing conclusions from the data

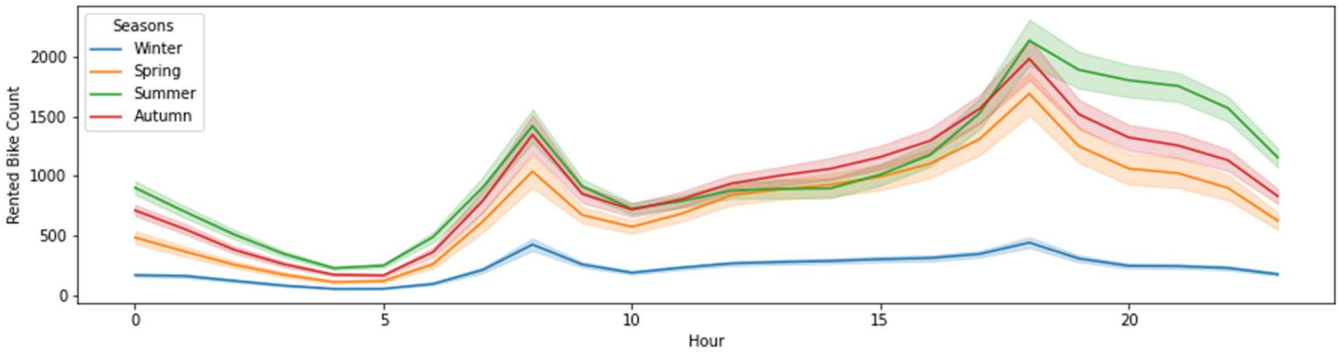
Plotting necessary graphs which provides relevant information on our datalike :




SeasonsRented Bike Count
1e6
Rented Bike Count
1790002
1611909
2283234
487169
Autumn
Spring
Summer
Winter
Seasons


HolidayRented Bike Count
1e6
Rented Bike Count
215895
5956419
Holiday
No Holiday
Holiday


Holiday
No Holiday
Holiday
Rented Bike Count
Hour


Seasons
Winter
Spring
Summer
Autumn
Rented Bike Count
Hour

Plotting necessary graphs which provides relevant information on our datalike :

A. Most bikes have been rented in the summer season.

B. Least bike rent count is in the winter season.

A. autumn and spring seasons have almost equal amounts of bike rentcount.

B. Most of the bikes have been rented in the year 2018.

C. Most of the bikes have been rented on working days.

D. Very few bikes have been rented in December which is winter season.

E. Most bikes have been rented in December in the year 2017 as we don'thave data before that.

F. People tend to rent bikes when there is no or less rainfall.

G. People tend to rent bikes when there is no or less snowfall.

H. People tend to rent bikes when the temperature is between -5 to 25degrees.

I. People tend to rent bikes when the visibility is between 300 to 1700.

J. The rentals were more in the morning and evening times. This is because people not having personal vehicles , commuting to officesand schools tend to rent bikes.

III. Training the model

A. Assigning the dependent and independent variables

B. Splitting the model into train and test sets.

C. Transforming data using minmaxscaler, standardscalar, robustscalar

D. Fitting linear regression on train set.

E. Getting the predicted dependent variable values from the model.

IV. Evaluating metrics of our model

A. Getting MSE , RMSE , R2-SCORE , ADJUSTED-R2 SCORE for differentmodels used.

a. MSE - the mean squared error or mean squared deviation of anestimator measures the average of the squares of the errors.

b. RMSE - Root Mean Square Error (RMSE) is the standard deviationof the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.

c. R2-SCORE - R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependentvariable that's explained by an independent variable or variables in a regression model.

d. ADJUSTED-R2 SCORE - Adjusted R-squared is a modified versionof R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases whenthe new term improves the model more than would be expected by chance. It decreases when a predictor improves themodel by less than expected.

B. Comparing the r2 score of all models used , to get the desiredprediction.

Models used

Linear regression:

Linear regression is a linear approach for modelling the relationship between a scalarresponse and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This

term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis

Logistic regression assumptions:

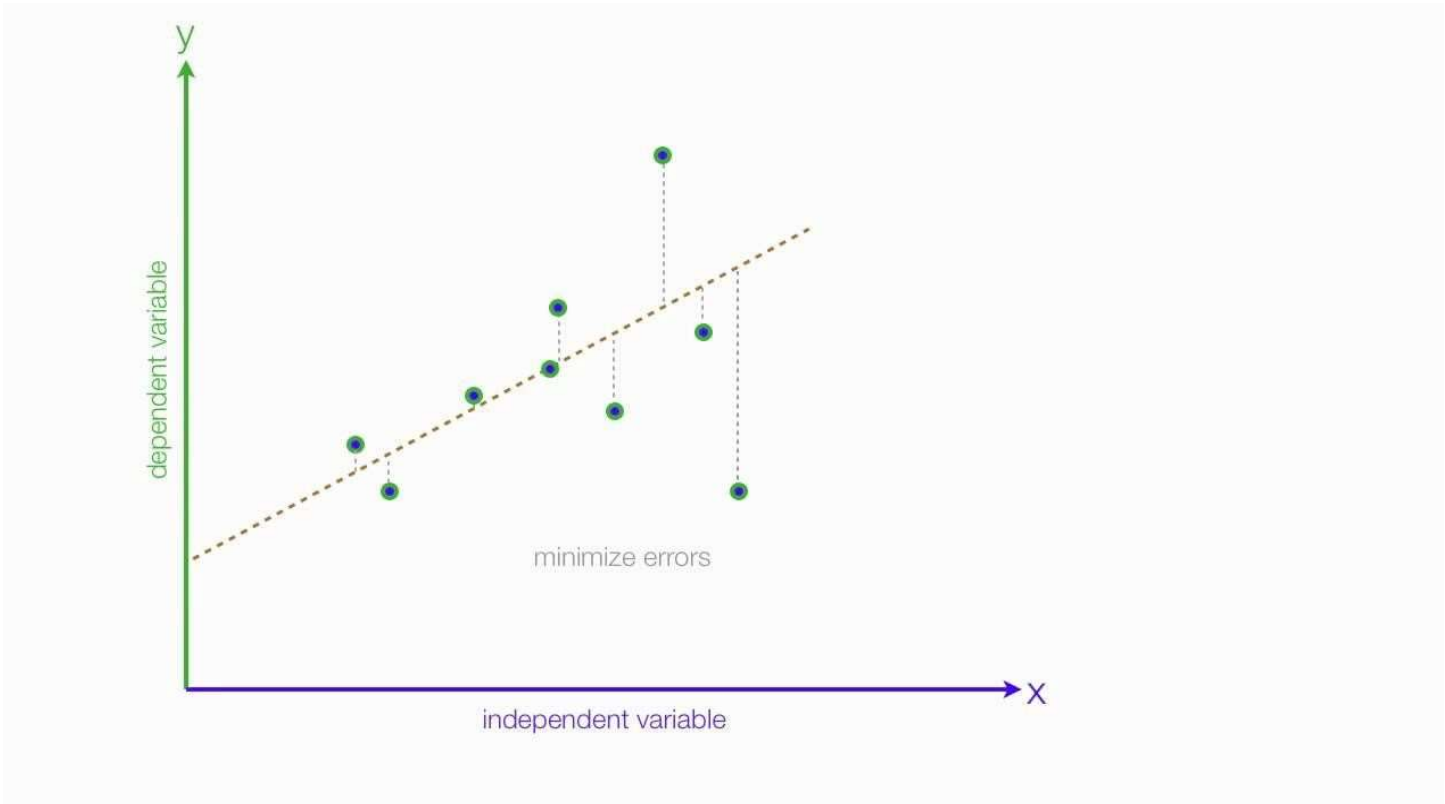
- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.
- There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
- The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

- The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to heteroskedasticity.
- The error terms must be normally distributed.

We have to train our model considering the above assumptions.

Properties of Logistic Regression:

- The line reduces the sum of squared differences between observed valuesand predicted values.
- The regression line passes through the mean of X and Y variable values
- The regression constant (b0) is equal to y-intercept the linear regression
- The regression coefficient (b0) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit changein the independent variable (X).



Lasso regression model:

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso solutions are quadratic programming problems, which are best solved with software (like MATLAB). The goalof the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Ridge regression model:

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When theissue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

The cost function for ridge regression:

Min(||Y – X(theta)||^2 + λ||theta||^2)

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitudeof coefficients is reduced.

- It shrinks the parameters. Therefore, it is used to prevent multicollinearity
- It reduces the model complexity by coefficient shrinkage

Decision tree regression model:

Linear model trees combine linear models and decision trees to create a hybrid model that produces better predictions and leads to better insights than either model alone. A linear model tree is simply a decision tree with linear models at its nodes. This can be seen as a piecewise linear model with knots learned via a decision tree algorithm. LMTs can be used for regression problems (e.g. with linearregression models instead of population means) or classification problems (e.g. with logistic regression instead of population modes).

Random forest regression model:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or averageprediction of the individual trees is returned.[1][2] Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generallyoutperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Extra-trees regression model:

Extra Trees is an ensemble machine learning algorithm that combines thepredictions from many decision trees.

It is related to the widely used random forest algorithm. It can often achieve as- good or better performance than the random forest algorithm, although it uses a

simpler algorithm to construct the decision trees used as members of theensemble.

It is also easy to use given that it has few key hyperparameters and sensibleheuristics for configuring these hyperparameters.

Elastic-net regularization model:

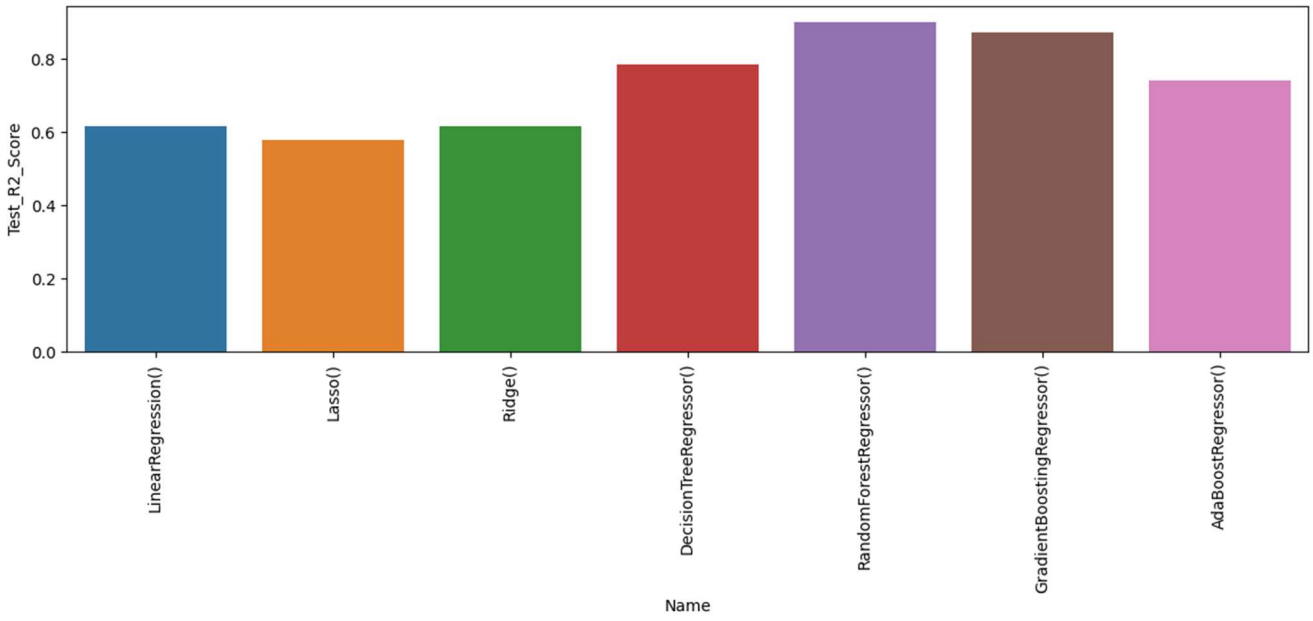
Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lassoand ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

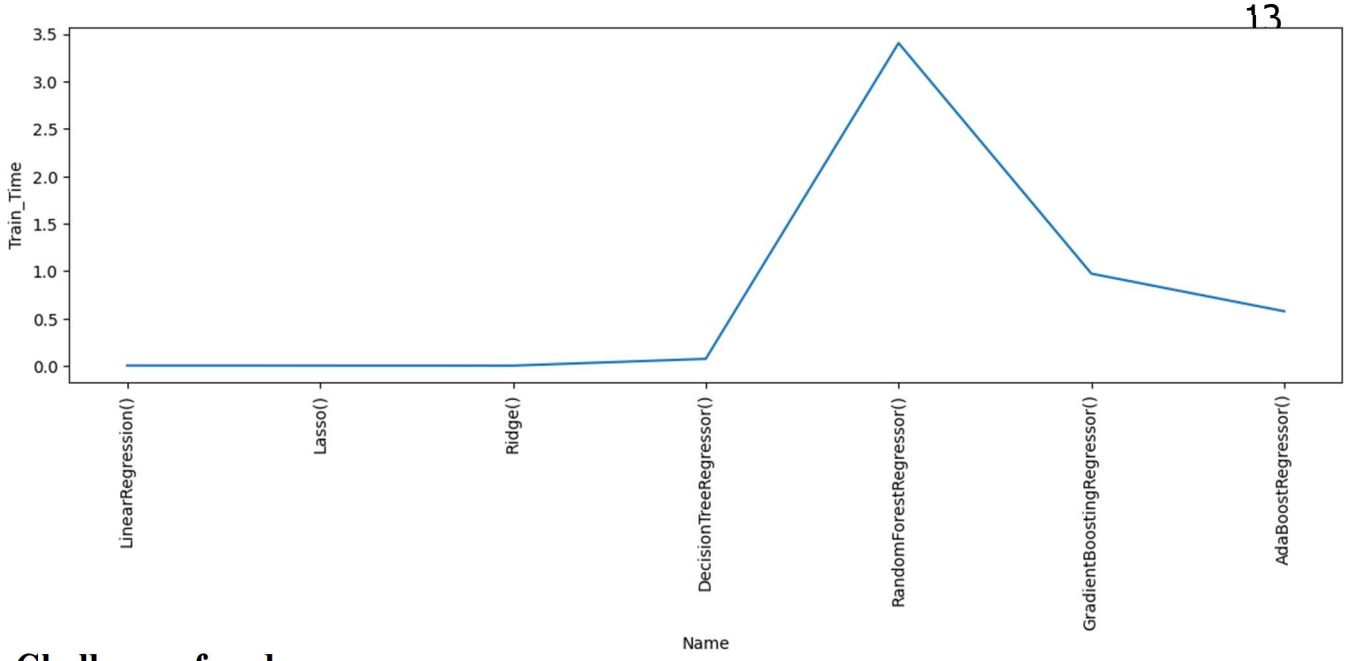
The elastic net method improves lasso's limitations, i.e., where lasso takes a few samples for high dimensional data. The elastic net procedure provides the inclusionof "n" number of variables until saturation. If the variables are highly correlated groups, lasso tends to choose one variable from such groups and ignore the rest entirely.

To eliminate the limitations found in lasso, the elastic net includes a quadratic expression ($||\beta||2$) in the penalty, which, when used in isolation, becomes ridge regression. The quadratic expression in the penalty elevates the loss function toward being convex. The elastic net draws on the best of both worlds – i.e., lassoand ridge regression.

All model scores and training time

	Name	Train_Time	Train_R2_Score	Test_R2_Score	Test_RMSE_Score
0	LinearRegression()	0.003352	0.614822	0.616963	7.499827
1	Lasso()	0.002688	0.577747	0.577491	7.876790
2	Ridge()	0.001956	0.614821	0.616957	7.499891
3	DecisionTreeRegressor()	0.073488	1.000000	0.786056	5.605067
4	RandomForestRegressor()	3.403994	0.984821	0.900992	3.812997
5	GradientBoostingRegressor()	0.972058	0.883240	0.874190	4.298212
6	AdaBoostRegrssor()	0.575346	0.745059	0.742617	6.147814





Challenges faced

1. Pre-processing the data was one of the challenges we faced which includesremoving highly correlated variables from the data so as to not hinder the performance of our regression model.
2. Exploring all the columns and calculating VIF for multicollinearity waschallenging because it might decrease the models performance.
3. Selecting the appropriate models to maximize the accuracy of ourpredictions was one of the challenges faced.

Model Selection

Observation 1: As seen in the Model Evaluation Matrices table, Linear Regression, Ridge and Lasso is not giving great results.

Observation 2: Random forest & GBR have performed equally good in terms of r2 score.

Observation 3: We are getting the best results from lightGBM and RandomForestRegressor and GBR

Observation 4: RandomForestRegressor takes a lot of training time so we use GBR instead as there is very less difference between there r2 score

Conclusion

We are finally at the conclusion of our project!

Coming from the beginning we did EDA on the dataset and also cleaned the data according to our needs. After that we were able to draw relevant conclusions fromthe given data and then we trained our model on linear regression and other models.

Out of all models used, with random forest regression model we were able to get ther2-score of 0.90.The model which performed poorly was ridge with r2-score of 0.57.