Supervised ML(regression) - Bike sharing demand prediction

Technical documentation

—

Mohammed Arifuddin Atif
arifuddinatif63@gmail.com

Introduction

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

Problem statement

We are tasked with predicting the number of bikes rented each hour so as to make an 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 the performance of the model.

    C. Converting the data into appropriate data types to create a regression model.

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

    E. Creating more columns in our dataset which would be helpful for creating 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 remove correlated and multicollinear variables.

II. Drawing conclusions from the data

Plotting necessary graphs which provides relevant information on our data like :

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

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

C. autumn and spring seasons have almost equal amounts of bike rent count.

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

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

F. Very few bikes have been rented in december which is winter season.

G. Most bikes have been rented in december in the year 2017 as we don't have data before that.

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

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

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

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

L. The rentals were more in the morning and evening times.This is because people not having personal vehicles , commuting to offices and 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.

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 different models used.

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

   b. RMSE - Root Mean Square Error (RMSE) is the standard deviation of 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 dependent variable that's explained by an independent variable or variables in a regression model.

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

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

Models used

Linear regression:

Linear regression is a linear approach for modelling the relationship between a scalar response 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 linear 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 values and 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 change in the independent variable (X).
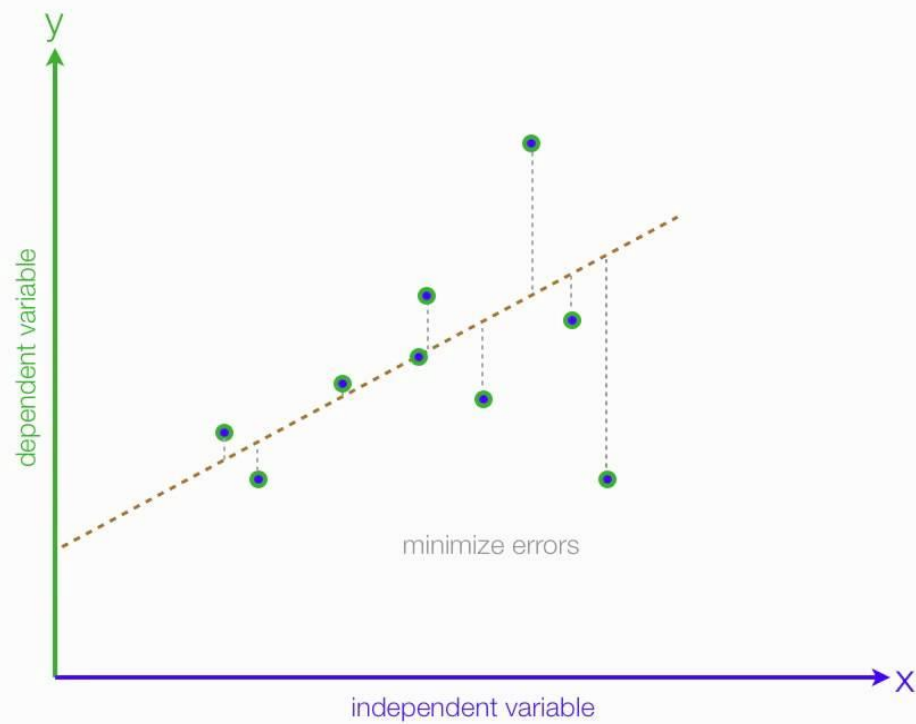
Lasso regression model:

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 goal of 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 the issue 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:

$$Min(||Y - X(theta)||\text{^}2 + \lambda||theta||\text{^}2)$$

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 magnitude of 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 linear regression 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 average prediction of the individual trees is returned.[1][2] Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform 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 the predictions 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 the ensemble.

It is also easy to use given that it has few key hyperparameters and sensible heuristics 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 lasso and 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 inclusion of "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., lasso and ridge regression.

Challenges faced

1. Pre-processing the data was one of the challenges we faced which includes removing 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 was challenging because it might decrease the models performance.
3. Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

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 from the given data and then we trained our model on linear regression and other models .

Out of all models used , with extra-trees regression model we were able to get the r2-score of 0.85.The model which performed poorly was elastic net regularization with r2-score of 0.42.

Given the size of data and the amount of irrelevance in the data , the above score is good.