Bike sharing demand

Abstract

The goal of this project was to use a multiple linear regression model for the prediction of demand for shared bikes.

Business Goal : Model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

Design

The project originates from the Data Science Bootcamp T5 the data is provided by UCI, We would be interested in prediction the rentals on various factors including season , temperature , weather and building a model that can successfully predict the number of rentals on relevant factors .

Data

This dataset contains the seasonal and weekly count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding temperature and humidity information. Bike sharing systems are a new way of traditional bike rentals. The whole process to from membership to rental and return back has become automatic. Given below is the description of the data which is a (17379,17) shaped data, The variables are

The table represent the features used in the training and analysis:

Features	Description
rec_id	Daily customer index
datetime	The date index for both years
season	Season type (1-winter, 2-spring, 3- summer, 4-fall)
year	The year (0-2011, 1-2012)
month	The months (1-12)
Is_holiday	0 – not holiday, 1-holiday
weekday	Weekdays 0(Monday) – 6(Sunday)
Is_workingday	0- not a working day, 1- workingday
weather	Weather type(1-Clear, 2- Cloudy, 3- Rian, 4-Storm
temp	Normalized value of temperatures at every rec_id
atemp	Normalized value of the absolute temperature
humidity	Contains the normalized value for the humidity
windspeed	Contains the normalized value for the windspeed
casual	Has the number of unregistered users at a given day
registered	Has the number of registered users
Total_count	Total rentals with both casual and registered users

Algorithms

- **Feature Engineering**

The provided data in Its raw form wasn't directly use as an input to the model serval future engineering was carried out where a few features we modified few were dropped

1-Encode categorical column (Hour, season) drop first column

2-Scalar numerical column

3- Fill zero value in windspeed

4- cross validation

- **Models**
  linear regression , polynomial , Ridge regression , lasso regression
  Model Evaluation and selection
  The entire training dataset of 17379 was split 25/75 train vs. holdout
  And all scores reported below were calculated with 5-fold cross validation on the training portion only. Prediction on the 25%holdout were limited to the very end so this split was only used and scores seen just
  The official metric for Bike Sharing Data was regression rate;
  Final regression 5-fold CV scores: 33features

- **linear regression , ACCURCY(0.69)**

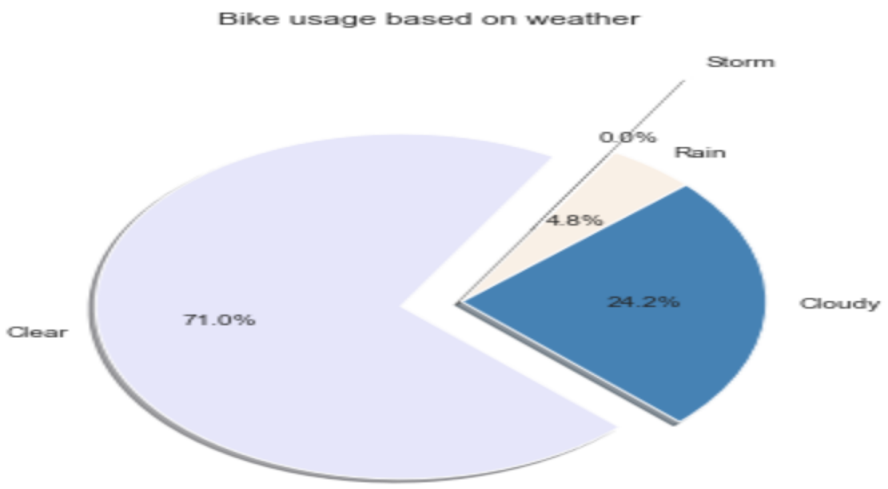
Holdout

- **linear regression , ACCURCY(0.69)**
- **Polynomial degree(2) ,ACCURCY(0.90)**
- **Ridge regression , ACCURCY(0.69)**
- **lasso regression, ACCURCY(0.69)**

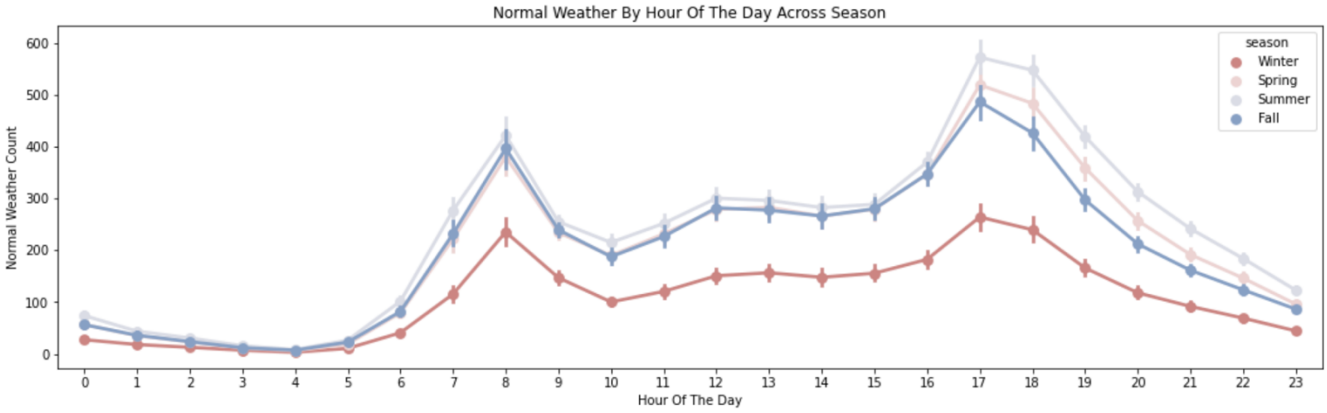
Tools

- Numpay and pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and seaborn for plotting

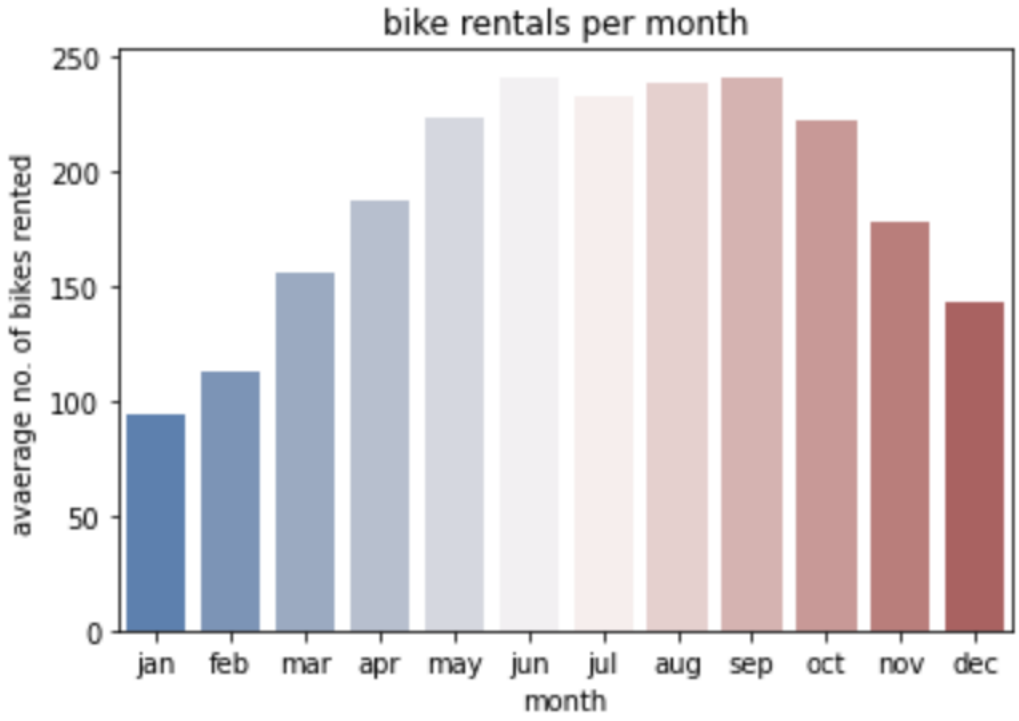
Communication



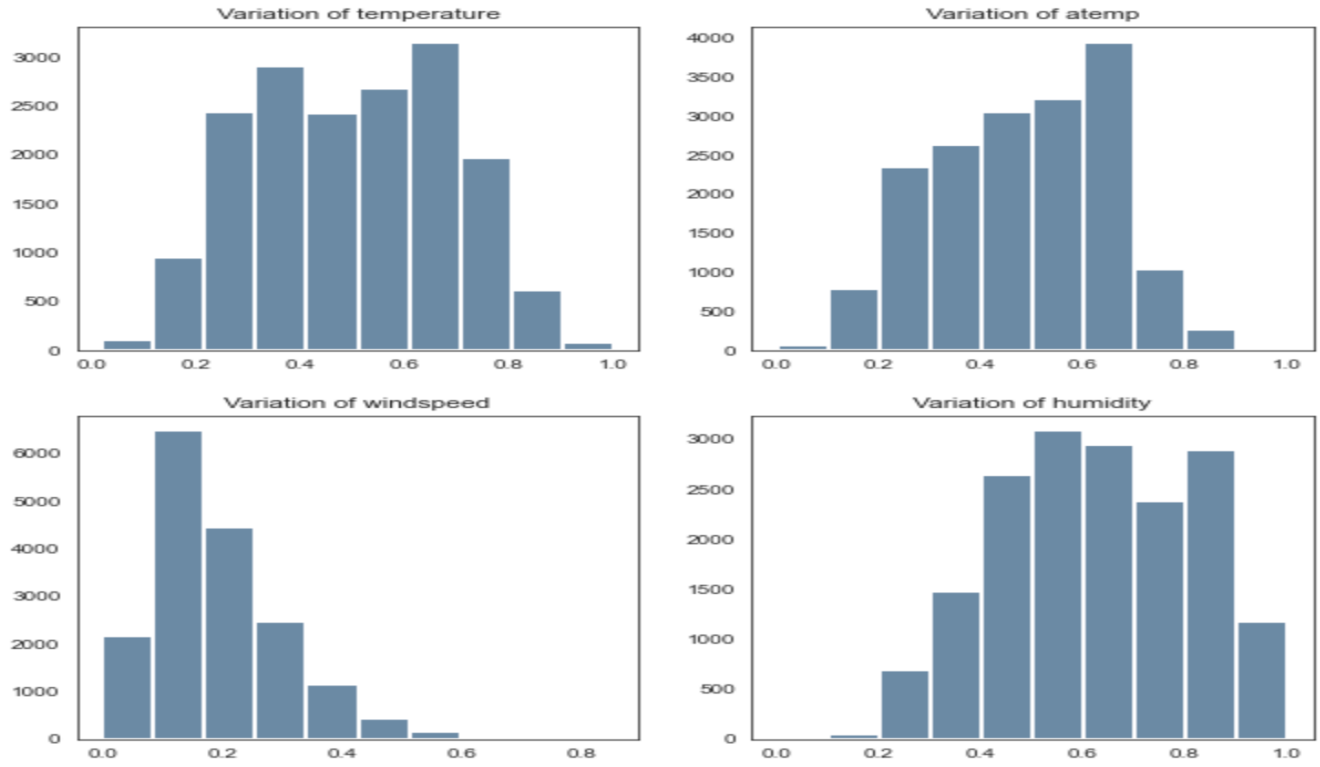
Clear skies have most occured with the 2 year period whereas Storm occurence are zero.

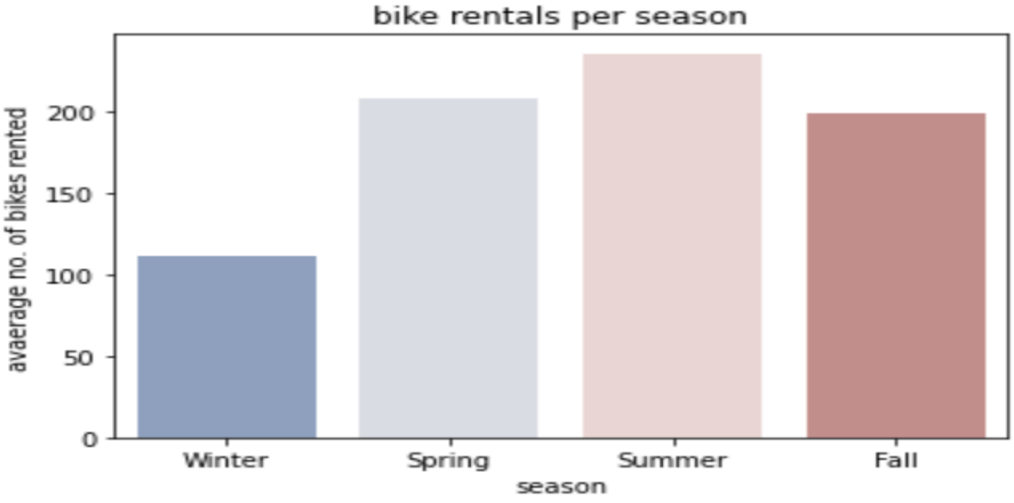


We see a significant rise in the number of the day across season in the hour 7-16

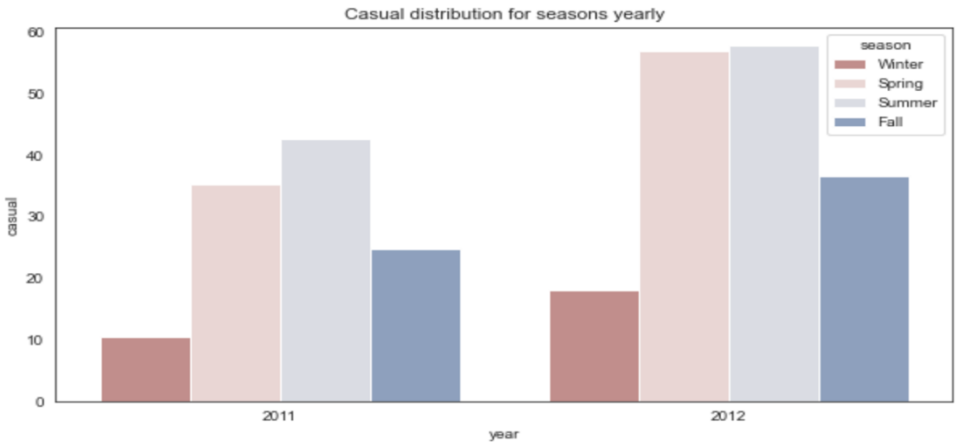


Bike rental is more between May to October month

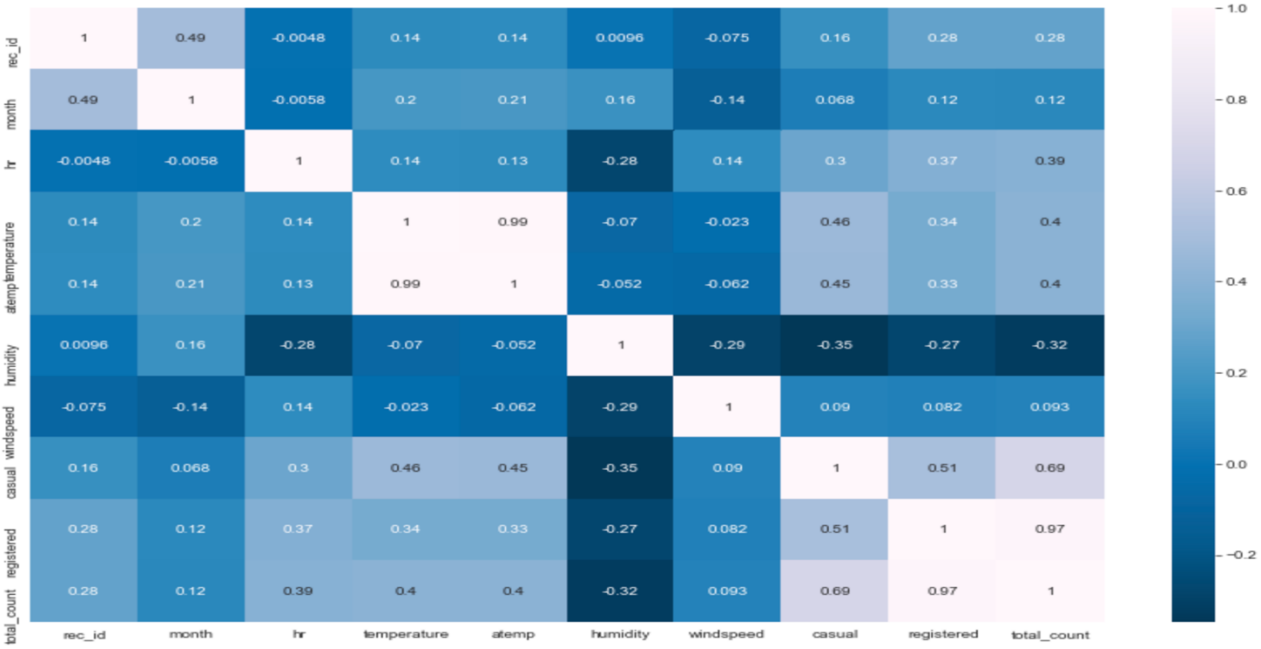




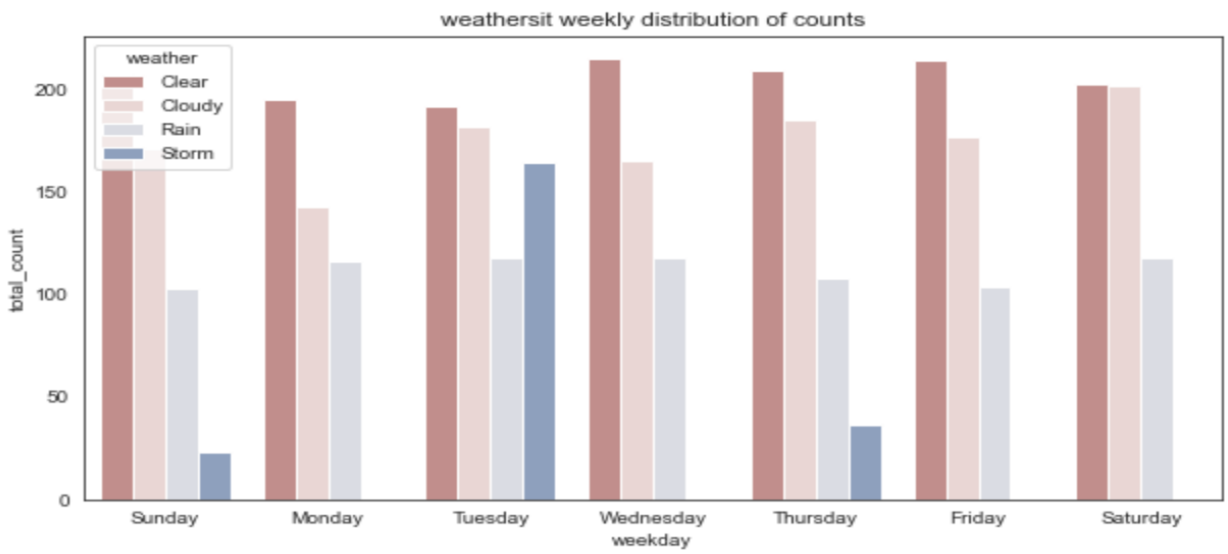
Bike rental is more when season is summer then spring and then fall.



is more in 2012 compared 2011



Highest correlation between temp & atemp so we drop atemp



The median of total_count remains the same on all the days of week