**Nikhil Machave (Data Science Enthusiast)**

**ABSTRACT:**

These bike sharing systems have become widely popular across different cities in the world. The elements of affordability and accessibility have helped in the promotion and expansion of the concept of a bike rental system. And it definitely comes across as a win-win for just about anyone who is willing to skip the expensive cab services for a bike. Discussions around the positive effects of bike sharing often focus on environmental sustainability and the personal health benefits experienced by users. When new bike share schemes are introduced, there’s often a reference to the “greening” of the city; ideally, more bikes can equal fewer cars, which leads to a reduction in greenhouse gasses and an improvement in local air quality**.**

**PROBLEM STATEMENT:**

Maximize: The availability of bikes to the customer.

Minimize: Minimise the time of waiting to get a bike on rent.

**The main goal of the project is to:**

Finding factors and cause those influence shortage of bike and time delay of availing bike on rent. Using the data provided, this paper aims to analyse the data to determine what variables are correlated with customer churn, if any. Hourly count of bike for rent will also be predicted.

**Introduction :**

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis.

The goal of this project is to combine the historical bike usage patterns with the weather data to forecast bike rental demand. The data set consists of hourly rental data spanning two years.

**Factors Affecting :**

Following are the factors affecting to the number of bike rentals:

1. **Weather** : We observe higher bike rentals when the weather (ie humidity, windspeed solar radiations) is more clear and sunny. We also notice that there is a single instance where there were rentals under heavy rain/snow condition this maybe happen because of outliers in the dataset.
2. **Seasons** : Bike rental counts across the 4 seasons ie. Fall spring summer and winter. Bike reservations are highest during the Summer season and least during the Spring season.

**Working Day** : Bike rental counts on working and non-working days and we observed that the outliers are present in working day.

1. **Holiday** : Bike rental counts on holidays and non-holidays. Holidays correspond to non-working days. Also outliers are present in non holidays.
2. **Temperature** : We observed that there is increase in the bikes rented counts with temperature with a small decrease at the highest temperature. Temperature between 32 and 36 degrees Celsius seems to be the ideal temperature.
3. **Hours** : We observed that there is a peak in the bike rentals counts at around 8am morning and at around 5pm evening.

**Data-set description**

|  |  |
| --- | --- |
| **Feature Name**  Date : year-month-day  Rented Bike Count  Hour  Temperature(**°C)**  Humidity (%)  Wind speed (m/s)  Visibility (10m)  Dew Point temperature (**°C)**  Solar Radiation (MJ/m2)  Rainfall (mm)  Snowfall(cm)  Seasons  Holiday  Functioning day | **Type**  Date  Int64  Int64  Float64  Int64  Float64  Int64  Float64  Float64  Float64  Float64  Object  Object  Object |

**Steps involved :**

The following steps are involved in the project

1. **Exploratory Data Analysis** :

After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is bike rentals counts with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables and also we observed the distribution of variables. It gave us a better idea that how feature behaves with the target variable.

1. **Null values Treatment and Outliers :**

Dataset contains a no null values to disturb the accuracy but outliers are present which can disturb the accuracy. So Again, we use z-score to remove outliers.

1. **Numerical and categorical Features :** With the help of exploratory data analysis we analyzed the categorical as well as numerical features in the dataset.
2. **One hot encoding :**

In this dataset some categorical variables like seasons, holiday and function day, we change it with numerical database.

1. **Correlation Analysis :**

We plot the heatmap to find the correlation between both dependent variable and independent variables.

1. **Train test Split :**

In train test split we take x as dependent variables and y take as independent variable then train the model.

**ALGORITHMS:**

**LINEAR REGRESSION:**

Linear regression is a supervised machine learning model majorly used in forecasting. Supervised machine learning models are those where we use the training data to build the model and then test the accuracy of the model using the loss function.

Linear regression is one of the most widely known time series forecasting techniques which is used for predictive modelling. As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable of interest).

We’re going to fit a line

y = β0 + β1x

to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable. Before we talk about how to do the fit, let’s take a closer look at the important quantities from the fit:

• β1 is the slope of the line: this is one of the most important quantities in any linear regression analysis

• β0 is the intercept of the line.



**RIDGE REGRESSION:**

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.

we have concluded that we would like to decrease the model complexity, that is the number of predictors. We could use the forward or backward selection for this, but that way we would not be able to tell anything about the removed variables' effect on the response. Removing predictors from the model can be seen as settings their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way. This way, we decrease model complexity while keeping all variables in the model. This, basically, is what Ridge Regression does.



**LASSO REGRESSION:**

Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression.

The only difference in ridge and lasso loss functions is in the penalty terms. Under lasso, the loss is defined as:



**DECISION TREE:**

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be *“learned”* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called*recursive partitioning*. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.



**GRADIENT BOOSTING:**

The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term ‘gradient’ is related here.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc

**Conclusion:**

EDA:

1. Demand for bikes got higher when the temperature and hour values were more.
2. Demand was high for low values of Humidity and solar radiation.
3. Demand was high during springs and summer and autumn and very low during winters.
4. Maximum bikes were rented in the year 2018
5. Count of rented bikes is high during no holiday and functioning day especially during office time.

**Model Fitting Conclusion**

1. From above its clear that Gradient Boosting regressor (CV) model is the best model for this dataset.
2. Decision tree Regressor and Gradient Boosting GridSearchCv gives the highest R2 score of 88% and 90% respectively
3. The most important features who had a major impact on the model predictions were; hour, temperature, Humidity, solar-radiation, and Winter.

**The model performed well in this case but as the data is time dependent, values of temperature, wind-speed, 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**