Detail Project Report
Bike Share Prediction

Revision Number : 1.2
Last Date of Revision : 10/06/2023

Document Version Control

Date	Version	Description	Author
02/06/2023	1.0	Abstract Introduction General Description	SAHIL JOSAN
04/06/2023	1.1	Technical Requirements Data Requirements Data Pre-Processing Design Flow	SAHIL JOSAN
10/06/2023	1.2	Data from user and its validation Rendering the results Deployment Conclusion	SAHIL JOSAN

Contents

Abstract

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city.

I am considering variables as season, yr,mnth, holiday, weekday, workingday, weathersit, Temp, atemp, hum, windspeed, casual, registered to predict the count of total rental bikes That will be required.

1. Introduction

1.1 Why this DPR Document?

The main purpose of this DPR documentation is to add the necessary details of the project and provide the description of the machine learning model and the written code. This also provides the detailed description on how the entire project has been designed end to end.

Keypoints:

- Describes the design flow
- Implementations
- Software requirements
- Architecture of the project
- Non-functional attributes like
- Reusability
- Portability
- Resource Utilization

2. General Description

2.1 Problem Perspective

The Bike Share Prediction is a machine learning model that helps companies to predict how many number of bikes will be required on that particular day based on some input data.

2.2 Problem Statement

The main goal of this model is to predict Bike Share counts based on the input features

2.3 Proposed Solution

To solve the problem, we have created a user interface for taking the input from the user To predict Count of Bikes Shared using our trained ML Model after processing the input and at last the predicted value from the model is communicated to the user.

3. Technical Requirements

As technical requirements, we don't need any specialized hardware for visualization of the application. The user should have a device that has the access to the web and the fundamental understanding of providing the input. And for the backend, we need a server to run all the required packages to process the input and predict the desired output

3.1 Tools Used

Python programming language and frameworks such as numpy, pandas, scikit-learn, Vs Code and streamlit are used to build the whole model.

- VS Code is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn, and Plotly are used.
- Streamlit is used for deployment of the model
- Front end development is done using streamlit
- Python is used for backend development
- Github is used as version control system

4. Data Requirements

The Data requirements totally supportted the matter statement and also the data is available on the UC Irvine Machine Learning Repository in (.zip) file format.

4.1 Data Collection

The data for this project is collected from the UC Irvine Machine Learning Repository, the URL for the same is https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

4.2 Data Description

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. I have used daily count data set which has 731 records and 14 attributes, which shows information like season, yr, month, weekday, weather the day was working day or not elc.

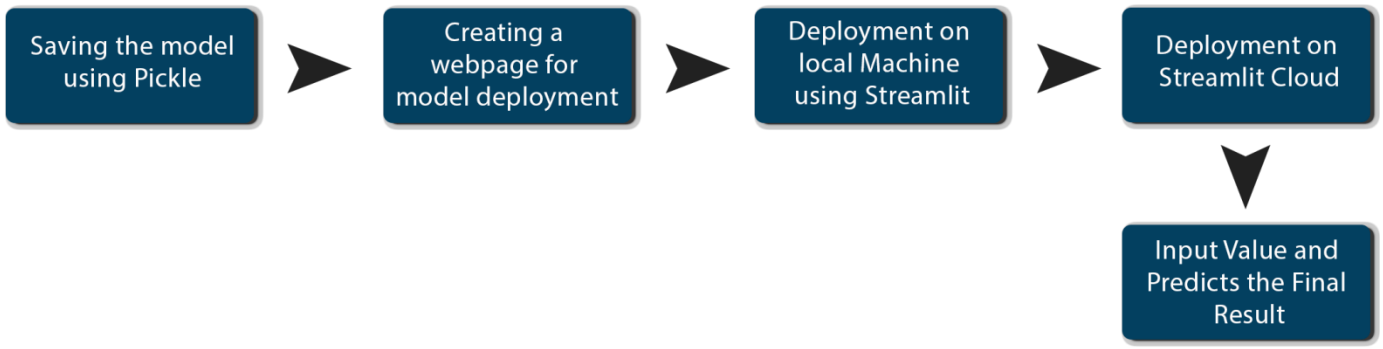
4.3 Data Pre-processsing

- Checked the datatype of features in dataset using df.info()
- Checked for Null values, because the null values can affect the accuracy of the model.
- Perform Label Encoding for the features that have categorical data.
- Checked the distribution of the features to interpret its importance.

  Now, the data is prepared to train a machine learning model.

5. Design Flow

5.1 Deployment Process



5.2 Data Description

In Logging, at each time an error or an exception occurs, the event is logged into the system log file with reason and timestamp. This helps the developer to debug the system bugs and rectify the error.

5.3 Data from user

The data from the user is retrieved from the created streamlit web app.

5.4 Data Validation

- The data provided by the user is then being processed by app.py file and validated.
- The validated data is then sent to the prepared model for the prediction.

5.5 Rendering the results

The data sent for the prediction is then rendered to the web page.

6. Deployment

The tested model is then deployed to streamlit. So, users can access the project from any internet device.

7. Conclusion

The Bike Share Prediction syatem will predict the count of rental bikes with the trained knowledge with set of rules. The company can use this system to predict how many bike will be rented on that particular day at that particular time.

8. Frequently Asked Questions (FAQs)

Q1) What's the source of data?

The data for training is accessible to all at UC Irvine Machine Learning Repository, the URL for the same is https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

Q2) What are the type of the data ?

The data was the combination of numerical and categorical values.

Q3) What's the complete flow you followed in this Project ?

Refer Page no 6 for better understanding

Q4) After the file validation what you do with incompatible file or files which didn't pass the validation ?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q5) How logs are managed ?

We are using different logs as per the steps that we follow in validation and modelling like Data validation log, Data Insertion, Model Training log, Prediction log etc.

Q6) What techniques were you using for data pre-processing?

- Removing unwanted attributes.
- Visualizing relation of independent variables with each other and output variable
- Checking and changing distribution of continuous values.
- Removing Outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values using label_encoder

Q7) How training was done or what models were used ?

- Before dividing the data in training and validation set, we performed pre-processing over the data set and made the final dataset.
- As per the dataset training and validation data were divided.
- Algorithms like Linear Regression, SVM, Decision Tree, Random Forest, XGBoost were used based on the recall, final model was used on the dataset and we saved that model.

Q8) How prediction was done?

The testing files are shared by the client. We performed the same life cycle on the provided dataset. Then, on the basis of dataset, model is loaded and prediction is performed . In the end we get the accumulated data of predictions.

Q9) What are the different stages of deployment ?

- First, the scripts are stored on Github as a storage interface.
- The model is first tested in the local environment.
- After successful testing, it is deployed on streamlit cloud.