Pune Institute of Computer Technology
Department of Computer Engineering
Academic Year 2021-22

DA
MINI PROJECT REPORT ON

Prediction of bike rental count hourly or daily based on the environmental and seasonal settings

Performed by
41415 Sanket Bhatlawande
41414 Samarth Bhadane
41402 Abhishek Sawalkar

Under the guidance of
Prof. B. Masram

Problem Statement

In this mini project, we intend to make a model that predicts the count of casual and registered users given the environmental and seasonal settings.

We will be performing the following tasks:

Regression:

Prediction of bike rental count hourly or daily based on the environmental and seasonal settings.

- Event and Anomaly Detection:

Count of rented bikes are also correlated to some events in the town which easily are traceable via search engines.

For instance, query like "2012-10-30 WashingtonD.C." in Google returns related results to Hurricane Sandy. Some of the important events areidentified in [1]. Therefore, the data can be used for validation of anomaly or event detection algorithms as well.

Abstract:

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and returnback has become automatic. Through these systems, user is able to easily rent a bike from a particular position and returnback at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed ofover 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic,environmental and health issues.

In this project, we intend to make a model that predicts the count of casual and registered users given the environmental and seasonal settings. After a short EDA and normalization, we'll be using Gradient Boosting, Random Forest, and a simple DNN.

Software and Hardware Requirements:

Software Requirements: Linux OS, Anaconda/Jupyter notebook or Google Colab, python3 3.2

Hardware Requirements: Laptop with 8GB RAM, Intel processor

Introduction:

Bike sharing count prediction

Outcomes:

1. We visualize the data using cross correlation matrix and visualize the time series.
2. We use MultiOutputRegressor object with GradientBoostingRegressor estimator.
3. We use Random Forest and DNN model to predict and then compare the results.

Objectives:

1. Learn how to apply preprocessing steps on a labelled dataset.
2. Learn to build various data prediction models.
3. Learn to split dataset into train and test set and apply normalizing data.
4. Learn visualization with different methods like cross correlation matrix.

Scope:

We have analyzed the daily data of bike share, we can also do the same for monthly data to analyze the trends that occur in the different seasons that can help us plot the next steps.

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated bythese systems make them attractive for the research. Opposed to other transport services such as bus or subway, the durationof travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system intoa virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of importantevents in the city could be detected via monitoring these data.

Dataset details:

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions,precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related tothe two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which ispublicly available in http://capitalbikeshare.com/system-data. The final dataset is taken from Kaggle.

Dataset characteristics:

Bike sharing counts aggregated on daily basis. Records: 731 days

instant: record index

- dteday : date

- season : season (1:springer, 2:summer, 3:fall, 4:winter)

- yr : year (0: 2011, 1:2012)

- mnth : month (1 to 12)

- holiday : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)

- weekday : day of the week

- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

+ weathersit :

  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy

  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- temp : Normalized temperature in Celsius. The values are divided to 41 (max)

- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)

- hum: Normalized humidity. The values are divided to 100 (max)

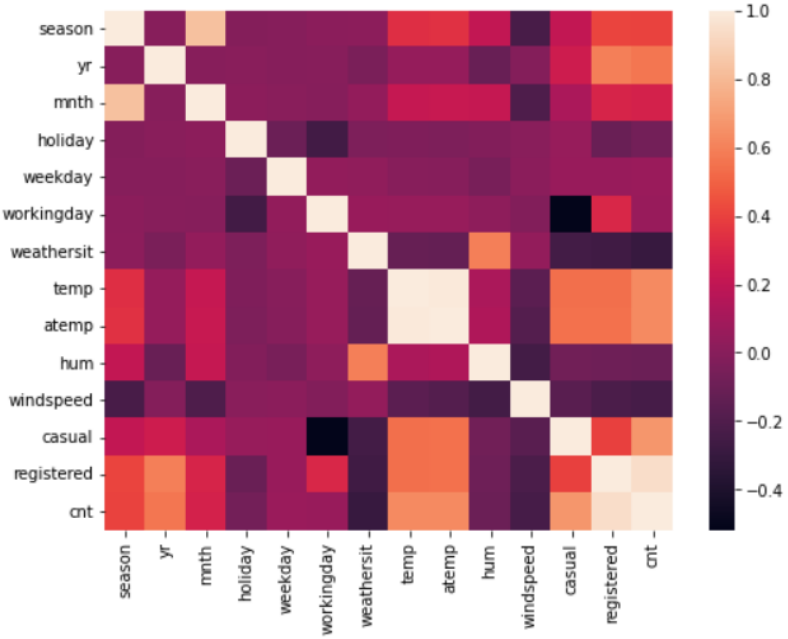
- windspeed: Normalized wind speed. The values are divided to 67 (max)

- casual: count of casual users

- registered: count of registered users

- cnt: count of total rental bikes including both casual and registered

Data visualization using cross correlation matrix:



Feature selection:

There are two types of input data:

- Seasonal data
- Weather data

and the output of the model is the bike user count which can be found in three columns of:

- casual

- registered
- cnt (sum of the causual and registered)

We can drop 'cnt' as it's just the total count of users. We experimented with seasonal data only or weather data only as the input to the model. But, the performance was better when all input features are used.

Theory Concepts:

Models analysed:

1. Gradient Boosting Regressor:
   Gradient boosting is one of the most powerful techniques for building predictive models.
   Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.
   Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

2. Random Forest:
   A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

   A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like scikit-learn).

3. DNN:
   Deep Neural Networks (DNNs) are typically Feed Forward Networks (FFNNs) in which data flows from the input layer to the output layer without going backward[3] and the links between the layers are one way which is in the forward direction and they never touch a node again.The outputs are obtained by supervised learning with datasets of some information based on 'what we want' through back propagation. Like you go to a restaurant and the chef gives you an idea about the ingredients of your meal. FFNNs work in the same way as you will have the flavor of those specific ingredients while eating but just after finishing your meal you will forget what you have eaten. If the chef gives you the meal of same ingredients again you can't recognize the ingredients, you have to start from scratch as you don't have any memory of that. But the human brain doesn't work like that.

Applications:

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated bythese systems make them attractive for the research. Opposed to other transport services such as bus or subway, the durationof travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system intoa virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of importantevents in the city could be detected via monitoring these data.

Implementation details:

1. Loading the data:

According to the description of the dataset, the enviromental features (4 columns of temp, atemp, hum and windspeed) are normalized. I think the temp and atemp (feels like temperature) are too correlated, so I discard atemp here. I need the data in its original scale, so I'll reverse the normalization according to the description below from the dataset owner:

- temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)

2. Data visualisation:

We visualize the time series and also using cross correlational matrix.

3. Used Gradient Boosting Regressor, Random Forest and DNN models with following results:
   Model: Train MAE - Test MAE
   Gradient Boosting: 165.58 - 515.7
   Random Forest: 97.21 - 548.42
   DNN: 394.99 - 630.69

Model chosen:

From the results after analysis, best results based on test MAE is driven from Gradient Boosting method. Perhaps the other models specially DNN can be tuned to achieve at least the same level of performance.

Conclusion:

We have successfully created a prediction model for analysis of bike rental count based on 731 entries of daily data.