Bike Rental Prediction



Predicting Bike Rental by Hour

Maja Pedigo – Data Scientist at XYZ

# Abstract

This case study is about a bike rental shop. They want to predict the demand of bikes at any given hour of the day, so that they can arrange for sufficient number of bikes for the customers.

Our task is to create a machine learning model which can predict the count of bikes rented at a given hour of the day.

## Data

The data consists of 12 features including the target variable, and 10886 instances. It is contained within one dataset in train.csv format, and another test.csv dataset to be used for testing the model(s).

Data description:

* datetime: Date on which the other values were recorded.
* season: Season (1:winter, 2:spring, 3:summer, 4:fall).
* holiday: whether day is holiday or not
* workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
* weather: Weather conditions (1: Clear, Few clouds, Partly cloudy, Partly cloudy, 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
* temp: The true temperature on that day
* atemp: The feeling temperature on that day
* humidity: Humidity on that day
* windspeed: The Windspeed on that day
* casual: Count of casual users
* registered: Count of registered users
* count: Total users - TARGET

## Method

The goal is to find a close prediction of the number of bikes likely to be rented on hourly basis This requires building **a regression model**. This will be done through data wrangling, exploration, preprocessing and ultimately building of prediction models.

## Data Wrangling

Initial data inspection revealed no missing values in the dataset. This significantly accelerated the phase of data cleaning.

Data frame was transformed by setting datetime feature as index, leaving all the other features in their original numerical data type. Additionally, parts of the datetime were extracted into new features to better obtain insights into year, month, day, day of week and hour – in relation to the target feature.

## Exploratory Data Analysis

During this phase, based on initial statistical analysis it was discovered that:

* There are about twice as much registered compared to casual users.
* There are many more rentals on workdays compared to holidays.
* Most rentals occur when the weather is dry, around 20 degrees C, wind around 12mph

Chart, bar chart

Description automatically generated

* As expected, number of rentals grows over the summer and peaks in the fall.

Chart, line chart

Description automatically generated

* One of the busiest months, July, and hourly trend shows that people rent bikes the most between 11am and 2pm, with a relatively small decline until 5pm, then climbing again until 7pm, going down from there.

A picture containing graphical user interface

Description automatically generated

* A close-up of month July shows slight seasonality every 5 days, possibly means slowing down over the weekend.

Chart, bar chart, histogram

Description automatically generated

* Our data has gaps between every 20th and last day of the month, and the above is a close-up.

Chart, bar chart

Description automatically generated

* Year to year comparison shows significant growing trend over time, and similar distribution over the months in a year.

Chart, scatter chart

Description automatically generated

* Looks like the largest counts are between 20-30 degrees Celsius, and in general higher temperature means more rentals, up until 35 degrees.

Chart

Description automatically generated

* Highest positive correlation is between our target and registered users, which will impact our prediction model in a wrong way, so we will drop them. We are left with temperature and hour with the highest correlation. Humidity affects our target negatively, meaning higher humidity leads to less rentals (probably due to rain).

## Data Preprocessing and Modeling

Firstly, we split the data into previously seen (train data used for machine learning) and unseen (test data used for testing the models).

In the next step, we scaled the numerical data with a high range of values.

With a processed dataset, we started building the first regression model:

**Linear Regression:**

This model gave poor accuracy on the trained and test data:

* Median r2 score – train and test data respectively = 0.40, 0.36
* MAE – train and test data respectively = 103.99, 109.90
* MSE – train and test data respectively = 19282.41, 21827.33

Feature Selection by Linear Regression:

hour 55.906530

atemp 41.595752

year 41.224821

month 33.389449

temp 9.599761

windspeed 4.191408

day 2.768528

holiday -0.921197

weather -2.436207

season -7.439409

humidity -39.544382

**Random Forest:**

This model performed much better, and gave the following results:

* Mean CV Score = 0.94
* STD CV Score = 0.0027
* MAE (mean) = 26.91
* MAE (std) = 0.68
* MAE for test data = 25.60

After we refined the features and fit our data into the grid search, we found that 1000 trees will be best for this model and scaling the data did not have much effect on the results.

Feature Selection with this model ordered Hour, Year and temp as the top 3 features.

Chart

Description automatically generated

Checking if more data would improve results:

Chart, line chart

Description automatically generated

Confusion Matrix

## **Saving the Best Model**

In the final phase, we saved the Random Forest model, and tested it on the ‘test’ data we were provided with.

**Our model predicted 192 rentals per hour as a general demand.**