TITLE OF THIS PAPER

AUTHOR 1, GANG LI, AND AUTHOR 3

Abstract. Bike sharing demand aims to forecast the use of the bikeshare system throughout the city. This is an automated system of renting bicycles. The process of obtaining membership, rental and bike return is automated via a network of kiosk locations through out a city. The target of this project is to predict the total count of bikes rented each hour covered by test set using only information available prior to the rental period.

Contents

Date: (None).

2020 *Mathematics Subject Classification.* Artificial Intelligence.

Key words and phrases. Bike sharing, Machine Learning, Data Mining.

1. Introduction

In the shared bicycle system network covering the whole city, users can rent and return bicycles by themselves. Currently, there are more than 500 bike-sharing systems around the world. The data generated by these systems clearly records information such as user rental time, departure and end locations, and acts as a sensor network that can be used to study urban traffic behavior.
In this competition, you are asked to use historical data, including weather conditions, to predict the rental demand of Washington's shared bike system.
The final goal is to use information available before the rental period to predict hourly bike usage for the test set.

The remainder of this paper is structured as follows.Exploratory data analysis in section two gives the description about the data and the visual analysis of those data. Section 3 defines some of the predictive modelling techniques that are used in this project. The evaluation results are displyed in section4. The last section is conclusion.

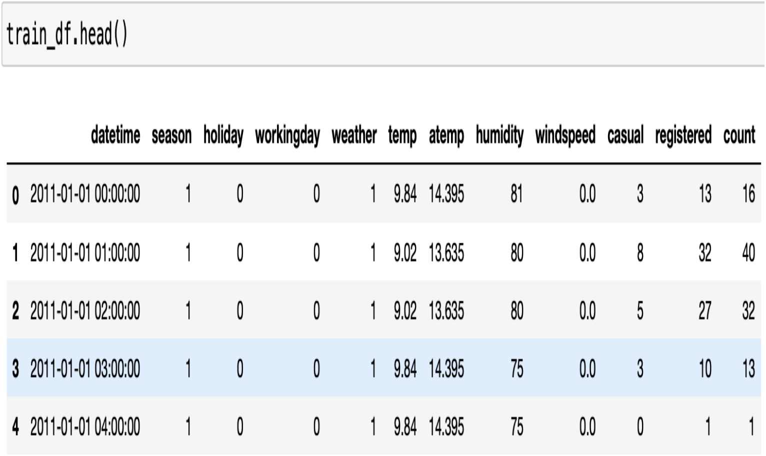
2. Exploratory Data Analysis

The official website provides an hourly car rental data spanning two years, in which training set provides the data and usage of the first 19 days of each month,* test set* provided after the 20th to the end of the month data.
There are two dataset provided for analysis. One dataset is titled 'train.csv' and the other is titled 'test.csv'.
The dataset comprises of various categorical and numerical features.

- Category Features : Hour, weekday, month, Season, holiday, Workingday, weather
- Numerical Features: temp, atemp, humidity, windspeed, registered and causal

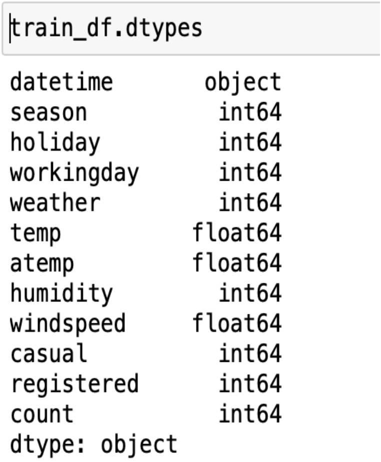
Dataset Description

- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather - 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 - temperature in Celsius
- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- egistered - number of registered user rentals initiated
- count - number of total rentals



```
train_df.head()
```

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0	3	13	16
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	8	32	40
2	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0	5	27	32
3	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0	3	10	13
4	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0	0	1	1



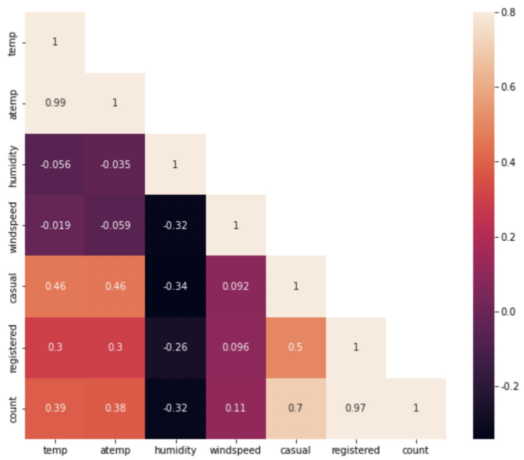
```
train_df.dtypes
```

```
datetime       object
season          int64
holiday         int64
workingday      int64
weather         int64
temp          float64
atemp         float64
humidity        int64
windspeed     float64
casual          int64
registered      int64
count           int64
dtype: object
```

3. Correlation Matrix

A common way to understand how target variables are affected by numerical features is to find the correlation matrix between them, plotting a heatmap of the correlation matrix between count and [temp, atemp, humidity, windspeed, casual, registered].
Based on the above heatmap, we can see that some of the features have no relation with the response variable. we can drop those columns.

- humidity, temp are negatively correlated with count
- There is a strong correlation between temp and atemp, if both are included in the model, it will cause multicollinearity problems, so one of the features must be deleted. We remove the atemp feature because it has a weaker correlation with count than temp.
- Casual, Registered are not considered and removed during model building
- humidity, temp and windspeed features are considered during future modelling



- fill in the zero values in the windspeed feature: usage is high when the wind speed is 0, which may be caused by null filling. Therefore, a random forest model is used here to fill in the zero values.

4. Predictive Modelling

For the purpose of predictive modelling, the given dataset is divided into train set and test set.

- Split the dataset into train set and test set
- Train Data size : 0.7
- Test Data size : 0.3

Prediction with Linear Model

- Linear Regression
- Ridge Regression
- Lasso Regression
- Logistic Regression
- ElasticNet

Prediction with Ensemble Learning Model

- Bagging Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- AdaBoost Regressor

5. Evaluation Results

- Evaluation Indicators: root mean square error is required (Root Mean Squared Logarithmic Error, RMSLE) to evaluate the quality of the model.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\log(p_i+1) - \log(a_i+1)]^2}$$

Among them, n is the number of samples in the test set, p_i is the test value, and a_i is the actual value. The smaller the root mean square error,

TABLE 1. Model Evaluation Results

Model	Accuracy
Random Forest Regression	0.376319
Bagging Regression	0.395248
GBRT	0.430378
AdaBoost Regression	0.703528
Ridge Regression	1.045335
Lasso Regression	1.045453
ElasticNet Regression	1.045489
Linear Regression	1.046341
Logistic Regression	1.131105

the better the fitting effect of the data, and the closer the test value is to the actual value.

6. CONCLUSIONS

This is only the basic modeling of the data. The results can be further enhanced.

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: `xxx@tulip.academy`

(A. 2) School of Information Technology, Deakin University, Geelong, VIC 3216, Australia
Email address, A. 2: `gang.li@deakin.edu.au`

(A. 3) School of Information Technology, Deakin University, Vic 3125, Australia
Email address, A. 3: `xxx@deakin.edu.au`