BIKE SHARING DEMAND

YAO YANG

Abstract. The goal of this project is to forecast bike rental demand given the input feature like weather, temperature, humidity, and windspeed.

Sixteen basic machine learning regression prediction family models were used. The final prediction is composed of the predictions of the best two models Stacking and LightGBM with a weight of 0.6 and 0.4.

Contents

Date: July 22, 2023.
Key words and phrases. Machine Learning.
Thanks to the seniors who guided Flip00 learning.

GLi:

A good paper introduction is fairly formulaic. If you follow a simple set of rules, you can write a very good introduction. The following outline can be varied. For example, you can use two paragraphs instead of one, or you can place more emphasis on one aspect of the intro than another. But in all cases, all of the points below need to be covered in an introduction, and in most papers, you don't need to cover anything more in an introduction.

1. 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. ~~Currently, there are over 500 bike-sharing programs around the world.~~

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. This dataset was provided by Hadi Fanaee Tork using data from Capital Bikeshare. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

The raw datasets contain eight data fields, which attributes are shown below.

- 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, 2: Mist + Cloudy, 3: Light Snow, 4: Heavy Rain
- temp - temperature in Celsius
- atemp - ~~"feels like" temperature in Celsius~~ body sensed temperature
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

GLi:

A few general tips: Don't spend a lot of time into the introduction telling the reader about what you don't do in the paper. Be clear about what you do do. Does each paragraph have a theme sentence that sets the stage for the entire paragraph? Are the sentences and topics in the paragraph all related to each other?

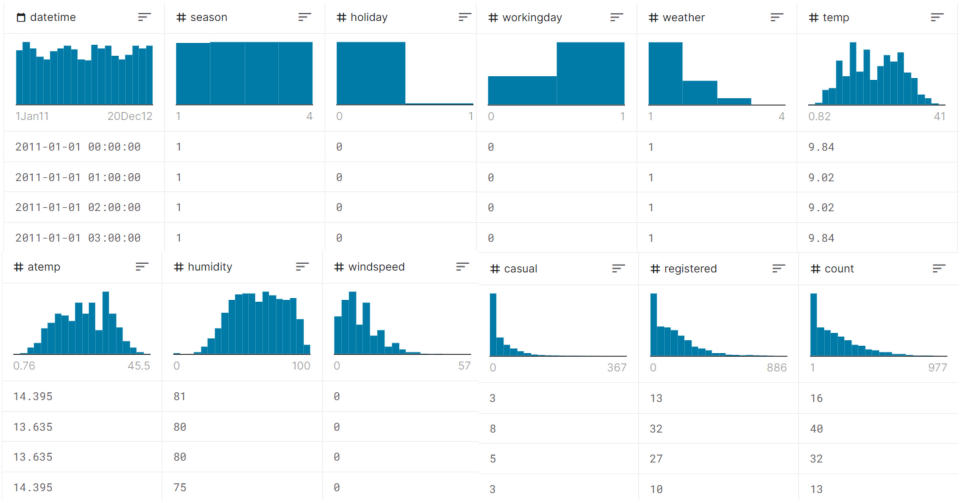

Figure 1. Data Describe

2. Problem Defination

The goal of this project is to forecast bike rental demand given the input feature like the duration of travel, departure location, arrival location, and time elapsed.

Evaluation metrics: RMSLE (Root Mean Squard Logarithmic Error) is required to evaluate the model.

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

n is the number of test set samples, pi is the test value, and ai is the actual value. When the root mean square error is smaller, it means that the fitting effect of the data is better and the test value is closer to the actual value.

3. Data Clean

You are provided hourly rental data spanning two years. For this competition, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each ~~hour111covered~~ hour covered by the test set, using only information available prior to the rental period.

- train.csv It contains a training set of target variables.
- test.csv It does not contain a training set of target variables.
- sampleSubmission.csv It is a properly formatted sample submission file.

4. Data Preprocessing and Feature EnginPeering

- Process date data (datetime module)
- Transform categorical features (calendar module)
- Analyze missing value and handle outlier
- Analyze target variable (logarithmic transformation)
- Fill in zero values in the windspeed feature (random forest model)

5. Model Solution

When modeling, we mainly consider the three numerical features "temp", "humidity" and "windspeed".

Sixteen basic machine learning regression prediction ~~family~~ models were used.

The final prediction is composed of the predictions of the best two models Stacking and LightGBM with a weight of 0.6 and 0.4.

6. Modeling and Result

- Model: Stacking*0.6+LightGBM*0.4
- Score: 0.50413
- Rank: 1741/3243

~~Summary~~ Here are summary of RMSLE scores for the 16 models:

	Model	RMSLE
15	LightGBM	0.316161
11	RandomForestRegressor	0.375379
10	BaggingRegressor	0.394187
14	XGBoost	0.422559
13	GBRT	0.435759
8	DecisionTreeRegressor	0.523695
9	ExtraTreeRegressor	0.554145
12	AdaBoostRegressor	0.697286
4	KernelRidge Regression	0.813210
7	KNN	0.864965
6	SVR	1.045943
5	ElasticNet Regression	1.053736
3	Ridge Regression	1.053749
2	Lasso Regression	1.054156
0	Linear Regression	1.054414
1	Logistic Regression	1.127804

FIGURE 2. RMSLE Scores

SCHOOL OF CYBER SECURITY, CHONGQING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS, CHONGQING 400065, CHINA
Email address: yangyao23432@foxmail.com