BIKE SHARING DEMAND PREDICTION

Chouyin Zhang

Hunan University, China

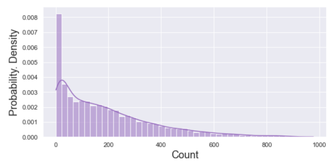
Problem Introduction

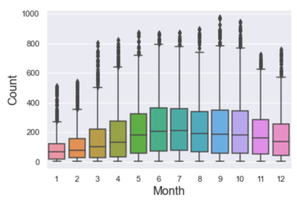
Background: Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

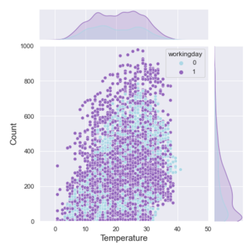
Problem: The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. 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.

Exploratory Data Analysis

- Distribution of dependent variables *count*, which represents bike usage.
- Distribution of *count* with *season*,*weather*,*month*, and *working day*.
- *A correlation heatmap* showing the *2D correlation matrix* for the features in our dataset.
- *Four jointplots* showing how daily usage varies with the continuous variables.







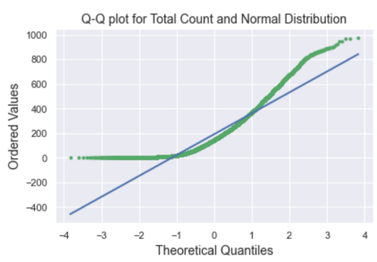
Data Preprocessing

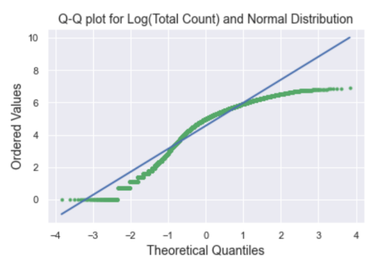
Drop some variables from our dataframe

1.Drop casual and registered as the test set does not contain data for the casual count or the registered count.
2. Drop atemp since it has almost perfect correlation with temp.
3.Drop all day of the month data from the training set before training our models.

Replace the target variable count with the natural logarithm of count

As the natural logarithm of count is more normally distributed, Let us replace the target variable count with y=ln(count+1).





Data Preprocessing

Replace the categorical data with binary dummy variables.

The categorical data is expressed as an arbitrary numerical value. Therefore we can use dummy coding, replacing the categorical data with binary dummy variables using the pandas function.

Split the labelled training set into two sets

One for training our models, and a validation set for determining the best set of hyperparameters.

Define a prediction function

Finally, let's define a function predict that will report prediction scores for a given model.

Train and Apply Models

Build and apply some chosen models

Model	RMSLE For Each Model
LinearRegression	0.136
Ridge	0.137
Lasso	0.148
RandomForestRegressor	0.084
GradientBoostingRegressor	0.080
XGBRegressor	0.085
LGBMRegressor	0.078

It can be observed that the LGBMRegressor model can get the best score. So, we apply this best model to test data.

The output of the prediction is as below

```
# Predict test_ohe with LGBMRegressor (selected model)

pred = lgbm_reg.predict(test_ohe)
pred[0:10]
```

```
array([2.42233246, 1.53272438, 1.30873403, 1.32273984, 0.98087189,
       1.99283686, 3.72056736, 4.60189455, 5.35332948, 4.74563852])
```

```
# back to count's original values
predicted_count = np.expm1(pred)
predicted_count[0:10]
```

```
array([ 10.27212045,   3.63077566,   2.70148479,   2.75369181,
         1.66678036,   6.33631636,  40.28781259,  98.6729723 ,
       210.31068191, 114.08126425])
```

```
# predict count

predicted_count = np.round(predicted_count, 0)
predicted_count[0:10]
```

```
array([ 10.,   4.,   3.,   3.,   2.,   6.,  40.,  99., 210., 114.])
```

Submission

	Unnamed: 0	datetime	count
0	0	2011-01-20 00:00:00	10.0
1	1	2011-01-20 01:00:00	4.0
2	2	2011-01-20 02:00:00	3.0
3	3	2011-01-20 03:00:00	3.0
4	4	2011-01-20 04:00:00	2.0

Conclusion

In this data mining task, I use the dataset from Kaggle. According to the problem background given, I apply EDA, data preprocessing to the dataset, and finally apply the best model to test data, successfully finish this task.