Bike Sharing Demand
Forecast use of a city bikeshare system

Dong Zhu

Deakin University

(None)

Overview

Problem Definition

Bike Sharing Demand Prediction

Data exploration

Check for missing vaules

Check for outliers

Data visualization

Time characteristic analysis

Weather characteristics analysis

Feature selection

Correlation analysis

Modelling and Forecasting

Conclusion

Problem Definition

Definition

Bike-sharing systems are a means of renting bikes, through which people can rent a bike from any place and return it when they arrive at their destination. The bike-sharing system clearly records the time of travel, the place of departure, the place of arrival and the time. Therefore, it can be used to study mobility in cities. In this project, historical usage patterns were combined with weather data to predict bike rental demand in Washington, D.C.

- Researchers can use bike sharing systems as a sensor network, which can be used for studying mobility in a city.
- This is a Supervised regression machine learning task.
- 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.

Data exploration

Check for missing vaules


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  int64
 3   workingday  10886 non-null  int64
 4   weather     10886 non-null  int64
 5   temp        10886 non-null  float64
 6   atemp       10886 non-null  float64
 7   humidity    10886 non-null  int64
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64
 10  registered  10886 non-null  int64
 11  count       10886 non-null  int64
dtypes: float64(3), int64(8), object(1)
memory usage: 1020.7+ KB
None
```

Figure 1: Training data information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6493 entries, 0 to 6492
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    6493 non-null   object
 1   season      6493 non-null   int64
 2   holiday     6493 non-null   int64
 3   workingday  6493 non-null   int64
 4   weather     6493 non-null   int64
 5   temp        6493 non-null   float64
 6   atemp       6493 non-null   float64
 7   humidity    6493 non-null   int64
 8   windspeed   6493 non-null   float64
dtypes: float64(3), int64(5), object(1)
memory usage: 456.7+ KB
None
```

Figure 2: Test data information

Check for outliers

- Statistical description

	count	mean	std	min	25%	50%	75%	max
season	10886.0	2.506614	1.116174	1.00	2.0000	3.000	4.0000	4.0000
holiday	10886.0	0.028569	0.166599	0.00	0.0000	0.000	0.0000	1.0000
workingday	10886.0	0.680875	0.466159	0.00	0.0000	1.000	1.0000	1.0000
weather	10886.0	1.418427	0.633839	1.00	1.0000	1.000	2.0000	4.0000
temp	10886.0	20.230860	7.791590	0.82	13.9400	20.500	26.2400	41.0000
atemp	10886.0	23.655084	8.474601	0.76	16.6650	24.240	31.0600	45.4550
humidity	10886.0	61.886460	19.245033	0.00	47.0000	62.000	77.0000	100.0000
windspeed	10886.0	12.799395	8.164537	0.00	7.0015	12.998	16.9979	56.9969
casual	10886.0	36.021955	49.960477	0.00	4.0000	17.000	49.0000	367.0000
registered	10886.0	155.552177	151.039033	0.00	36.0000	118.000	222.0000	886.0000
count	10886.0	191.574132	181.144454	1.00	42.0000	145.000	284.0000	977.0000

Figure 3: Data description

Check for outliers

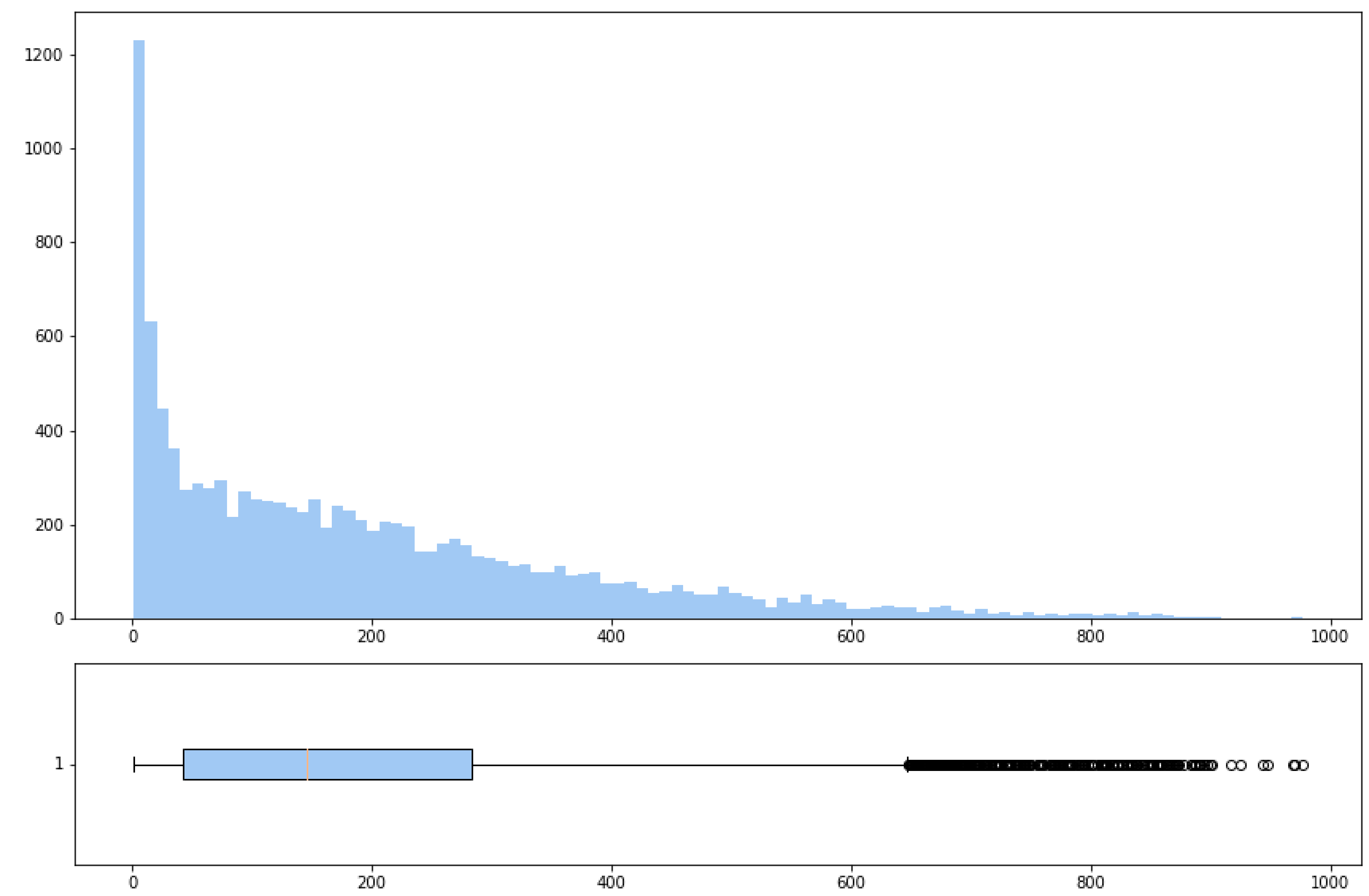


Figure 4: The distribution of the label ”count”

Check for outliers

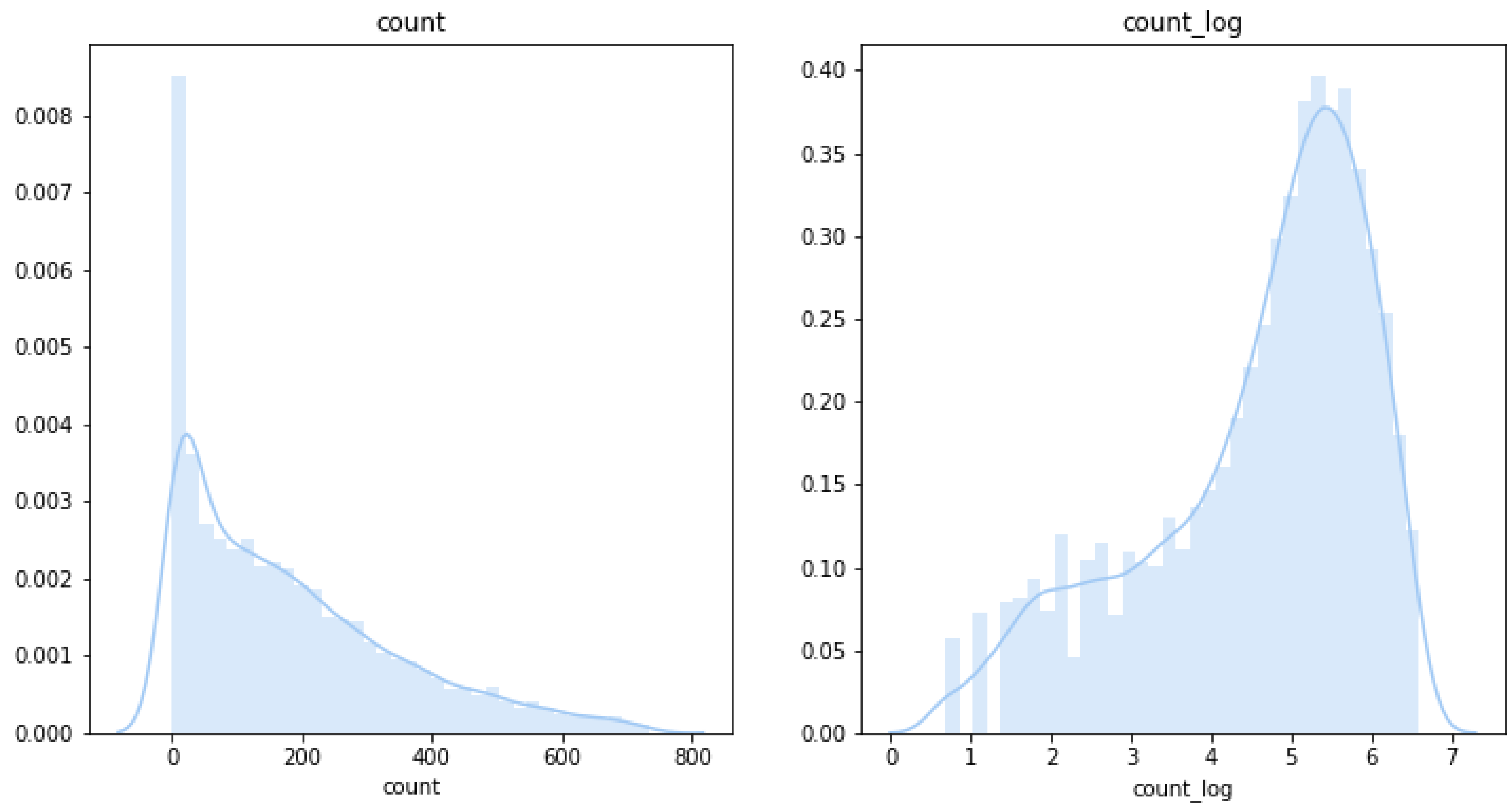


Figure 5: Count distribution compare

Check for outliers

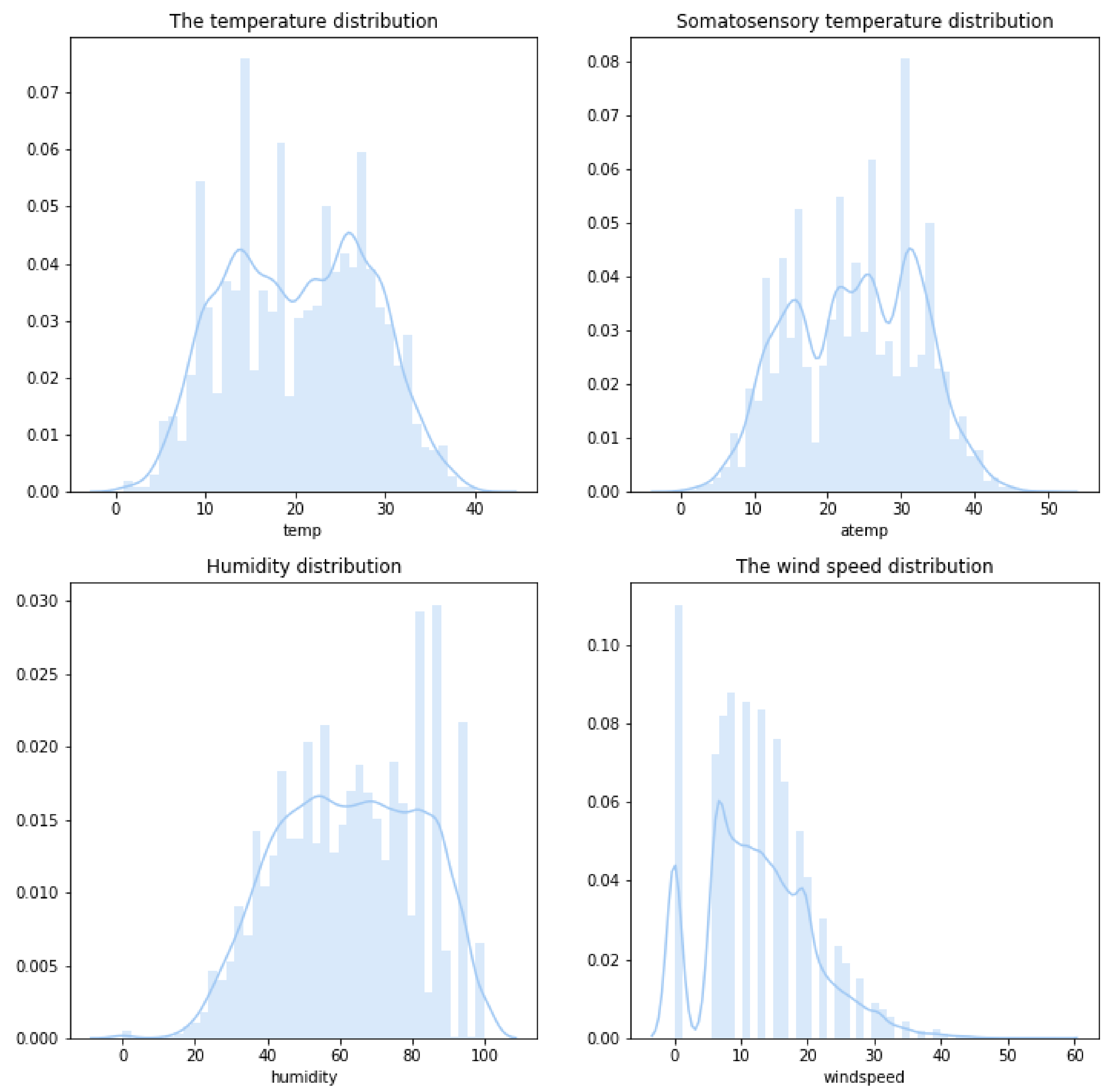


Figure 6: Main features distribution

Check for outliers

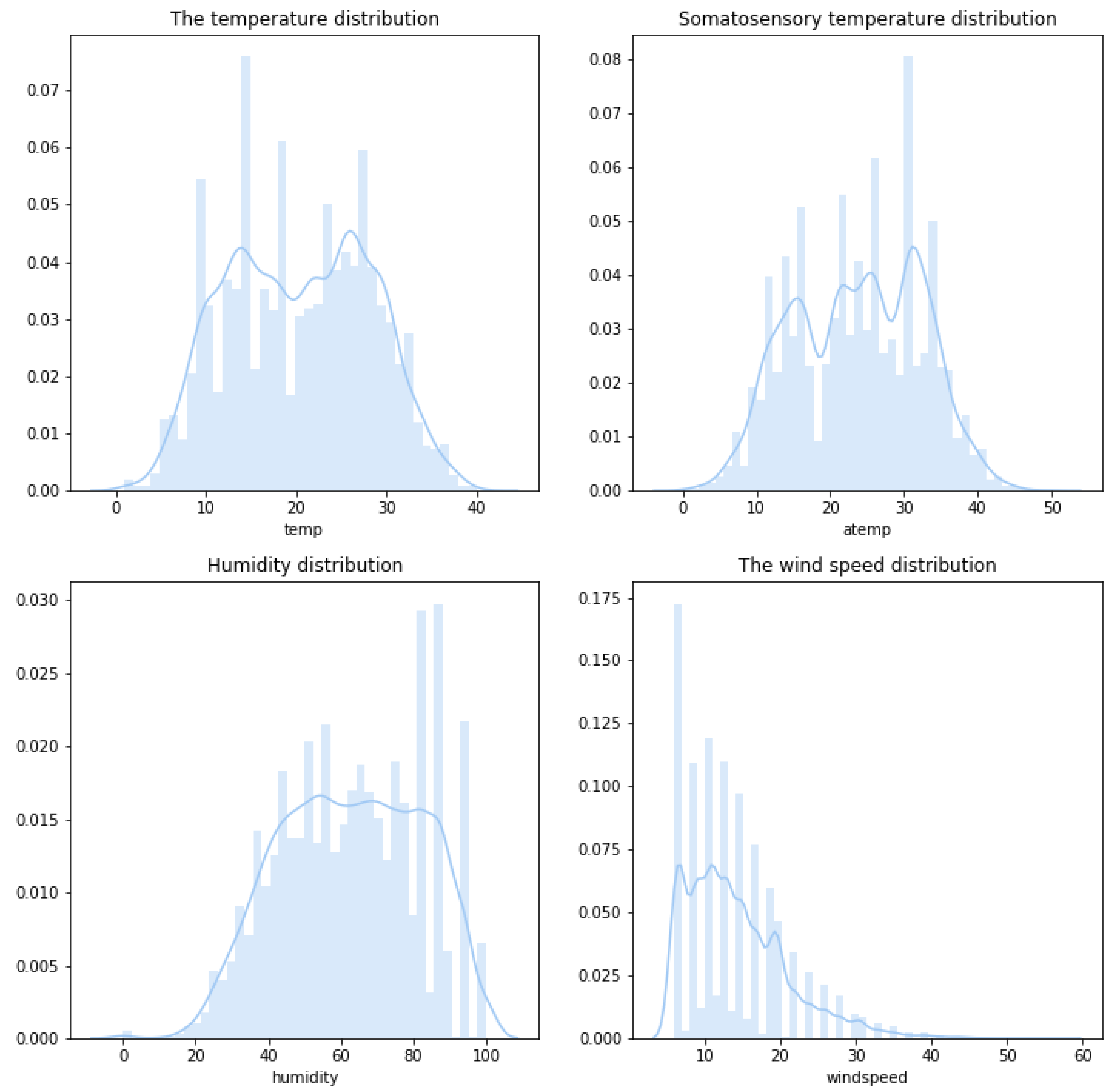


Figure 7: Main features distribution

Data visualization

Time characteristic analysis

- there are two peaks in the graph, one is from 7-8 in the morning, the other is from 5-6 in the afternoon, which is the morning peak and the evening peak respectively, which is in line with the actual situation.

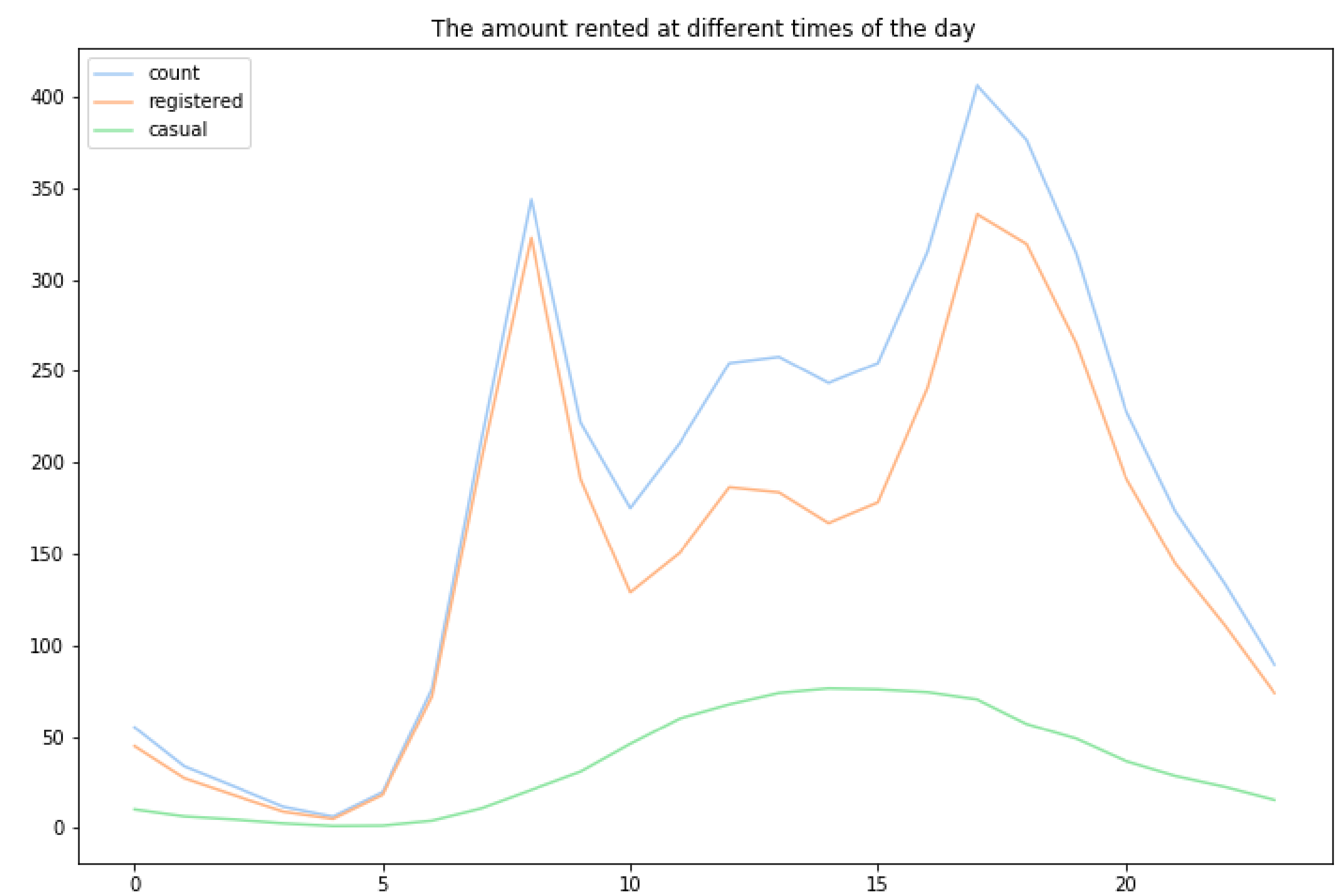


Figure 8: The amount rented at different times of the day

Time characteristic analysis

- from Monday to Friday, 8 in the morning of the day - 9 am and 5 to 7 PM, usage is more, may be caused by time going to work in the morning and evening after work time, include the reason of eating out at the same time, for the weekend, time is more focused, basic usage around 11 PM to 5 PM, This time is supposed to be everyone's leisure time.

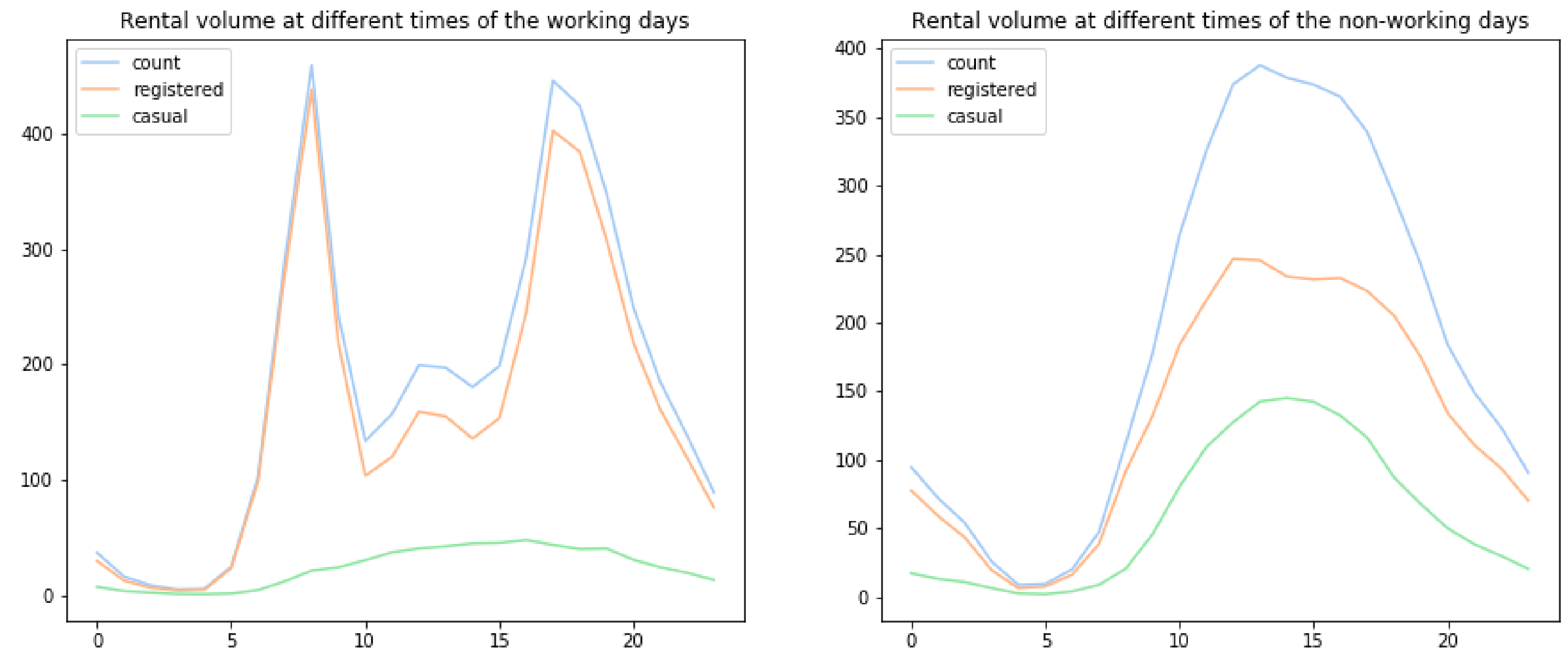


Figure 9: Rental amount at different times of the non-working days and the non-working days

- The usage is obviously lower in spring, probably due to the lower temperature.

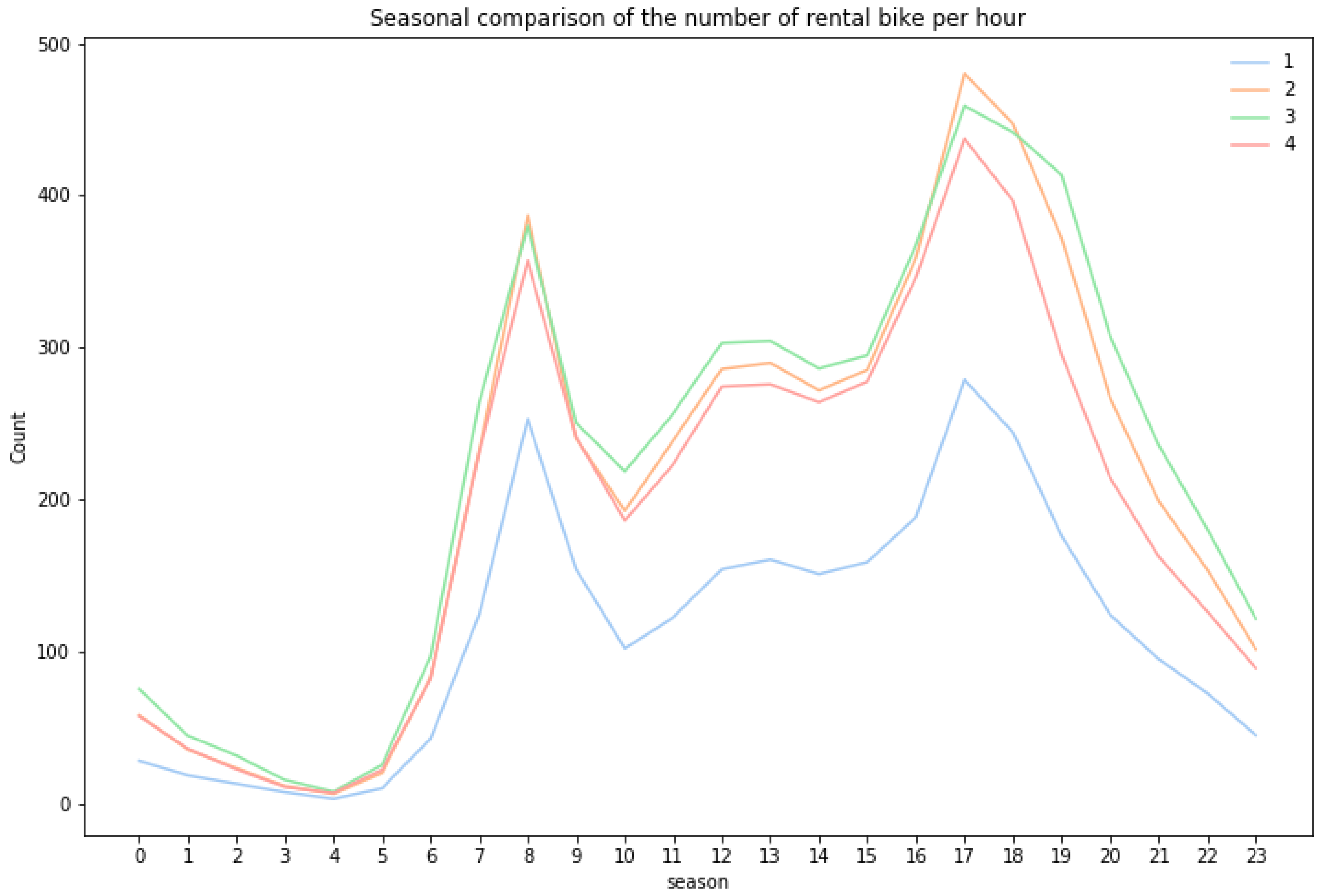


Figure 10: Seasonal comparison of the number of rental bike per hour

Weather characteristics analysis

- Temperatures below 10 degrees, above 30 degrees, and fewer bike rentals – too cold or too hot will damper rental demand.
- The higher the wind, the fewer bike renters - high winds dampen rental demand.
- The higher the humidity in the air, the fewer people who hire bikes - it's more comfortable to ride on dry days.

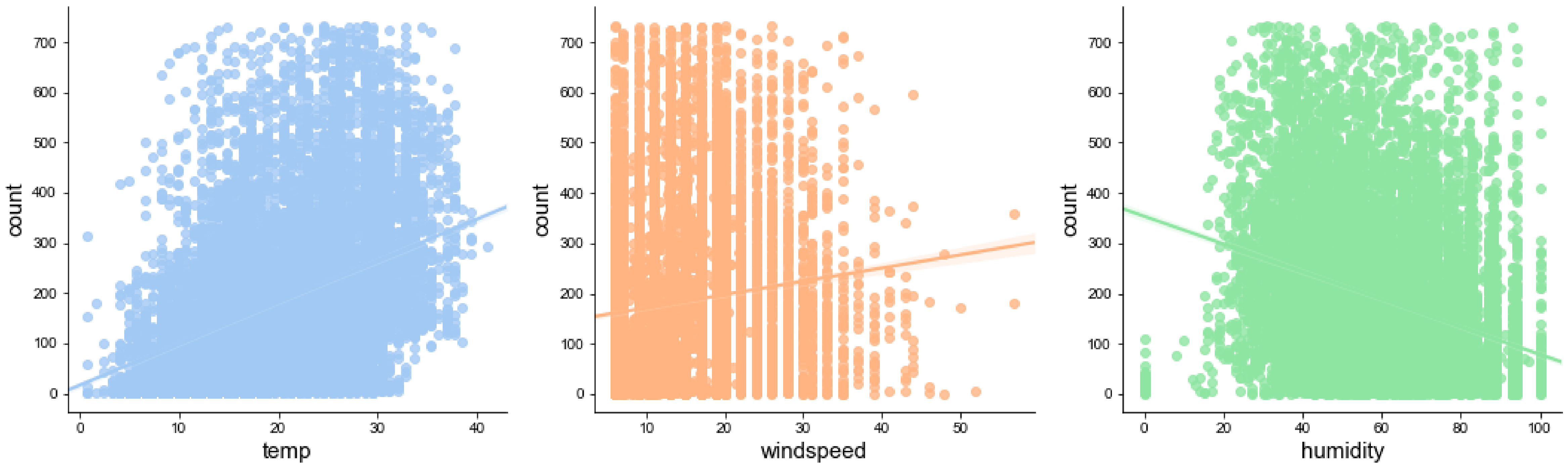


Figure 11: The effect of weather on rental amount

Feature selection

Correlation analysis

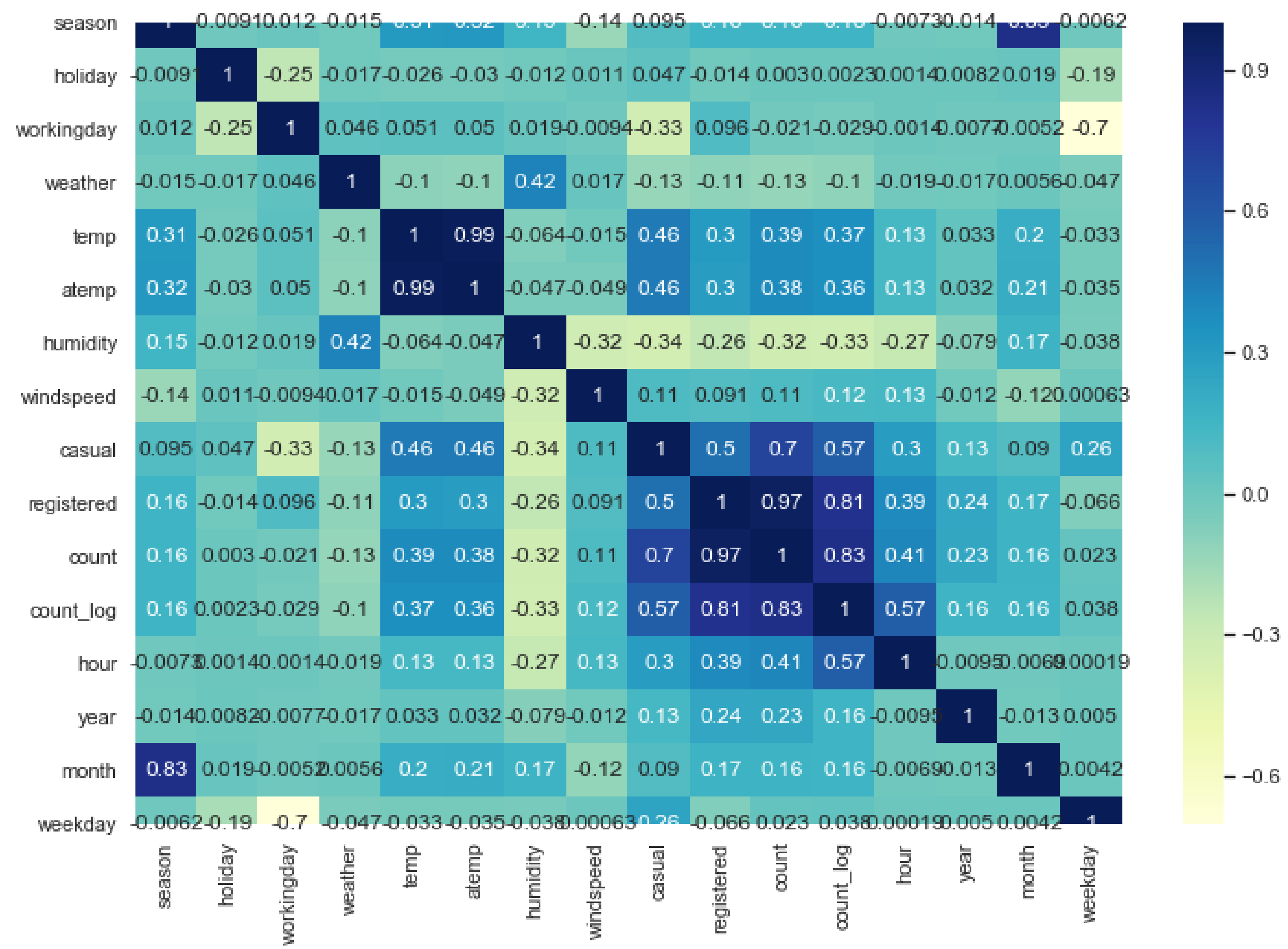


Figure 12: Correlation analysis

Correlation analysis

- The influence of characteristics on count is as follows: hour>temp>atemp>humidity>month>season>year>weather>windspeed >workingday>weekday>day>holiday

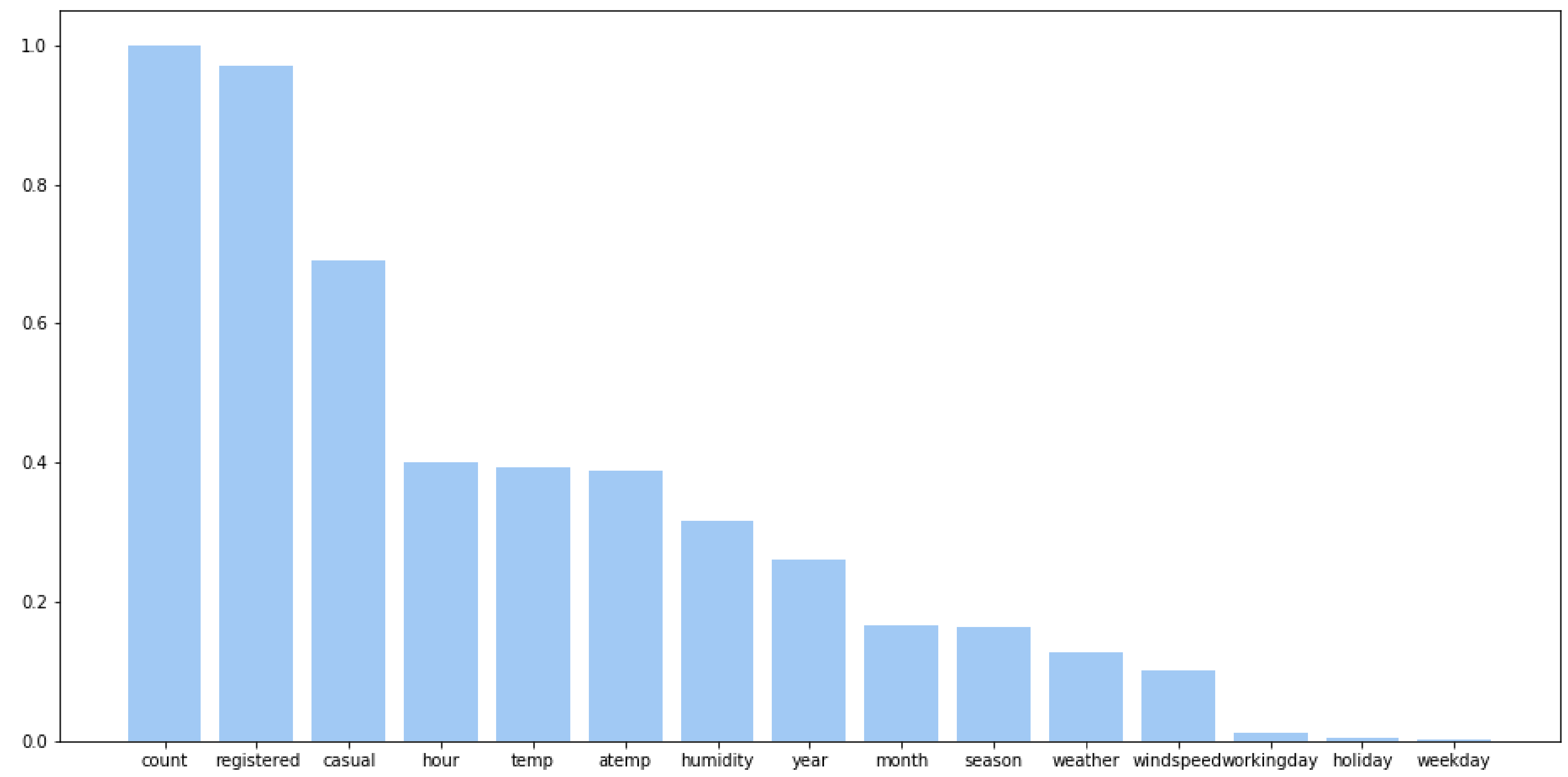


Figure 13: Correlation rank

Modelling and Forecasting

- Select random forest model and cross validation using grid search

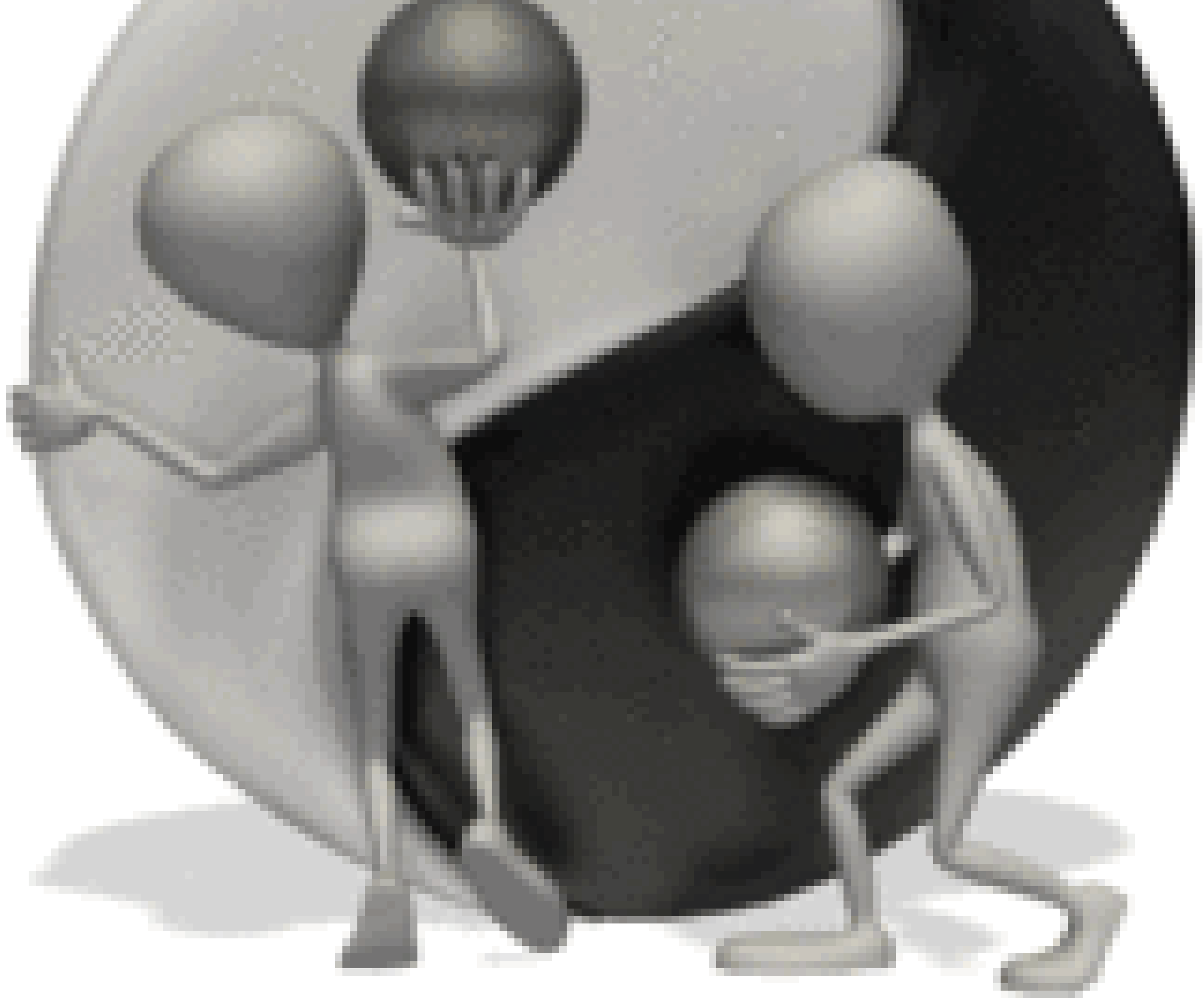
Result	The initial model	After cross validation
Accuracy	0.9338	0.9249
MSLE	0.0152	0.0159

Conclusion

Conclusion

- Cross-validation by grid search did not decrease the RMSE of the model and did not improve the accuracy of the model. The effect did not come up to expectations.
- The limitation of this study is that it does not consider whether there is overfitting of the model, and further experiments can be carried out in future studies.

Questions?



Contact Information

Student Dong Zhu
School of Information Technology
Faculty of Science, Engineering Built Environment (SEBE)
Deakin University, Australia

gangli@tulip.org.au

Team for Universal Learning and Intelligent Processing