

Data Mining & Visualization

by Yang Lu

Submission date: 01-May-2023 10:49AM (UTC+0100)

Submission ID: 204090691

File name: 40920_Yang_Lu_Data_Mining__Visualization_493224_1232565848.pdf (1.4M)

Word count: 3937

Character count: 21264

Data Mining & Visualization

Abstract

Finding meaningful patterns in data is the process of data mining. Finding trends and patterns in the data is the goal of a data mining session. Data mining places an emphasis on analyzing a lot of observable databases. Database administration, neural networks, deep learning, recognizing patterns, and data visualization are all used in this specialized field. It may be thought of from a statistical perspective as the automated computer-assisted exploratory investigation of several complicated data sets. A data set is a collection of data, sometimes referred to as an information set, data collection, or data gathering. A data set is a grouping of data, frequently shown as a table. Every row corresponds to a different variable. Each one of the rows relates to a specific question from the dataset. Each variable's value is listed, including the height and weight of an item or an arbitrary sequence of numbers. A data profile is what we name each value. The number of rows in the dataset's data may be represented by one or more members. The dataset cited in this paper is the UCI's Bike Sharings dataset, and WEKA and Jupyter Notebook were applied to process and analyze the data and present the results.

1 Introduction of Bike Share Dataset

1.1 Background

Bike sharing systems, a modern take on traditional bike rentals, automate each stage of the process, from signing up for membership to borrowing and returning bikes. Customers may rent a bike from one location and return it at another thanks to these technologies. Currently, there are in excess of 500,000 bikes utilized in 500 bike-sharing programs worldwide. Due to their crucial role in conveyance, preservation of the environment, and health issues, these systems are of utmost importance today. The data generated by bike-sharing systems is of great interest to many academics since it is essential to the exciting real-world applications of these networks. The duration of bike-sharing use, departure timings, and arrival locations are all explicitly documented in the system, unlike other modes of transportation like buses or subways. With this function, the bike-sharing service becomes a fake sensor network that is capable of tracking urban movements. Therefore, it will be expected that the surrounding and seasonal conditions will strongly influence the rental process for bike sharing. For instance, the weather, precipitation, day of the week, season, hour of the day, etc. can all have an impact on a renter's behavior.

The two-year historical log for the years 2011 and 2012 from the Capital Bike-share system in Washington, D.C., USA, is the basis of the main data collection. The most significant occurrences in the city may be discovered by watching this data. There are two csv files in this dataset: 17379 records in hour.csv, which aggregates bike sharing numbers on an hourly basis. day.csv, 731 entries, daily aggregated bike sharing counts.

1.2 Associated tasks

Regression: Hourly or daily estimation of the amount of bicycles rented depending on the local climate as well as the time of year.

Event and Anomaly Detection: The amount of leased bikes is also tied to various local events that may be easily located online. For instance, a Google search for "2012-10-30 washington d.c." yields information about Hurricane Sandy. As a result, the data may also be utilized to verify algorithms for anomaly or event detection.

1.3 Dataset characteristics

This data set contains 17389 data, each of which contains 16 attributes, respectively. Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv.

Name	Intro
instant	record index
dteday	date
season	season (1:springer, 2:summer, 3:fall, 4:winter)
yr	year (0: 2011, 1:2012)
mnth	month (1 to 12)
hr	hour (0 to 23)
holiday	weather day is holiday or not
weekday	day of the week
workingday	if day is neither weekend nor holiday is 1, otherwise is 0
weathersit	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	Normalized temperature in Celsius. The values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius. The values are divided to 50 (max)
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	count of casual users
registered	count of registered users
cnt	count of total rental bikes including both casual and registered

2 Related work

2.1 Weka

Weka, which is also an acronym of a bird in New Zealand, was developed at the

University of Waikato as well as first used in its current version in 1997. The full acronym is Waikato Intelligent Analysis Environment. The GNU General Public License (GPL) governs its licensing. The program is designed in the Java™ programming language and has a GUI for working with data files and producing visual output (such tables and curves). Additionally, it has a standard API, allowing users to incorporate WEKA into their own programs just like they would an additional library to perform functions like automated server-side data mining and visualization on a fresh interactive interface. Figure 2-1 shows the WEKA's original user interface:

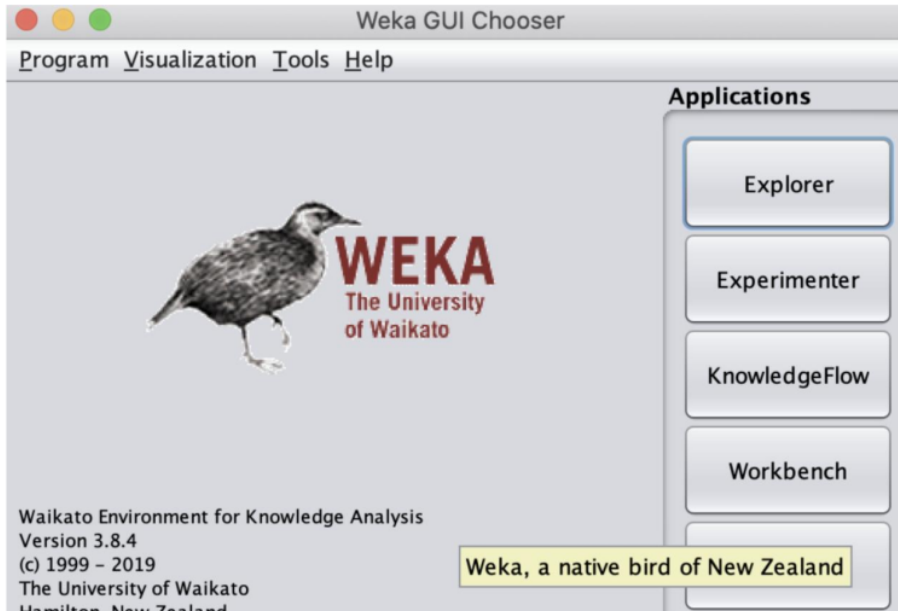


Figure 2-1. WEKA

Additionally, WEKA is equipped with generic API that enables users to incorporate WEKA into their own applications just like any other libraries in order to perform tasks like server-side programmed data mining. In order to perform data mining using jpython, WEKA is incorporated into Python in this article. The embedding concept of WEKA is displayed in Figure 2-2:

Python

Weka's functionality can be accessed from Python using the [Python Weka Wrapper](#). Conversely, Python toolkits such as scikit-learn [can be used from Weka](#).

Figure 2-2. Embedding principle of WEKA.

2.2 Data mining

Data mining is the process of discovering hidden patterns, patterns, and knowledge from big datasets through the use of various algorithms and technologies. The related work of data mining includes data preprocessing, feature selection, data transformation, algorithm selection and application, as well as result evaluation and interpretation. The development process of data mining can be traced back to the 1960s and 1970s, when it mainly focused on data analysis in commercial databases and social sciences. With the continuous development of computer and network technology, the number of data is increasing exponentially, and the diversity and complexity of data are also increasing day by day. In this context, data mining has gradually become one of the main methods for processing big data. In recent years, there have been some new technologies and trends in the field of data mining, such as deep learning, machine learning, big data analysis, cloud computing, distributed computing, etc. The emergence of these new technologies has brought infinite possibilities and new challenges to data mining. At present, data mining has become a very important branch in the fields of data science and artificial intelligence. The development history of data mining can be roughly divided into the following stages:

- Exploratory data analysis stage: In the 1960s, statistician Tukey proposed the concept of exploratory data analysis, and he emphasized the understanding of the intrinsic structure and laws of data by means of graphics and visualization.
- Data management and processing phase: From the 1970s to the 1980s, with the development of computer technology and the popularity of database systems, data processing gradually became an important task, and data mining began to turn to database based techniques.
- Machine learning and data mining stage: From the 1990s to the present, machine learning and data mining gradually became an independent discipline with a wide range of application areas, such as finance, medical care, e-commerce, etc.
- Big data era stage: Since the 21st century, with the development of the Internet, Internet of Things and other technologies, the amount of data is exploding, and data mining is facing new challenges and opportunities, such as large-scale data storage, distributed computing, etc.

The process of data mining are:

1. Data cleaning and pre-processing: This includes handling missing values, outliers, and duplicates, and performing feature selection and transformation.
2. Data integration: Combining data from several different sources into a single set.
3. Data Transformation: Normalization, discretization and normalization of data, as well as dimensionality reduction and denoising of data for better application of data mining algorithms.
4. Data mining algorithm selection and application: Different data mining algorithms are suitable for different data types and mining tasks, such as classification, clustering, association rule mining, and time-series data mining.

5. Model evaluation and optimization: Evaluate the performance of different algorithms using methods such as cross-validation, and optimize the mining models by tuning parameters.
6. Result visualization and interpretation: Display mining results to users in the form of graphs, reports, etc., and provide relevant interpretations and recommendations.

3 Data processing

3.1 Data description

First import the dataset using Pandas and Numpy and use the .info() function to obtain the dataset details as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 17 columns):
#   Column          Non-Null Count  Dtype
---  -
0   instant         17379 non-null  int64
1   dteday          17379 non-null  object
2   season          17379 non-null  int64
3   yr              17379 non-null  int64
4   mnth           17379 non-null  int64
5   hr              17379 non-null  int64
6   holiday         17379 non-null  int64
7   weekday         17379 non-null  int64
8   workingday      17379 non-null  int64
9   weathersit       17379 non-null  int64
10  temp            17379 non-null  float64
11  atemp           17379 non-null  float64
12  hum             17379 non-null  float64
13  windspeed       17379 non-null  float64
14  casual          17379 non-null  int64
15  registered      17379 non-null  int64
16  cnt             17379 non-null  int64
dtypes: float64(4), int64(12), object(1)
memory usage: 2.3+ MB
None
```

Figure 3-1. Overview of Bike Share Dataset

At the same time, perform missing value analysis:

```

instant      False
dteday       False
season       False
yr           False
mnth         False
hr           False
holiday      False
weekday      False
workingday   False
weathersit    False
temp         False
atemp        False
hum          False
windspeed    False
casual       False
registered   False
cnt          False
dtype: bool

```

Figure 3-2. Missing value of Bike Share Dataset

Then, using the head() function to print the first five examples of the dataset to explore the data:

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	0	1	1

Figure 3-3. First five examples of Bike Share Dataset

Subsequently, the describe() function is used to observe the statistical characteristics of each attribute and preliminarily estimate the distribution of each feature:

	instant	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp
count	17379.00000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	8690.0000	2.501640	0.502561	6.537775	11.546752	0.028770	3.003683	0.682721	1.425283	0.496987	0.47571
std	5017.0295	1.106918	0.500008	3.438776	6.914405	0.167165	2.005771	0.465431	0.639357	0.192556	0.17181
min	1.0000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.020000	0.00000
25%	4345.5000	2.000000	0.000000	4.000000	6.000000	0.000000	1.000000	0.000000	1.000000	0.340000	0.33333
50%	8690.0000	3.000000	1.000000	7.000000	12.000000	0.000000	3.000000	1.000000	1.000000	0.500000	0.48485
75%	13034.5000	3.000000	1.000000	10.000000	18.000000	0.000000	5.000000	1.000000	2.000000	0.660000	0.62121
max	17379.0000	4.000000	1.000000	12.000000	23.000000	1.000000	6.000000	1.000000	4.000000	1.000000	1.00000

Figure 3-4. Statistical characteristics of each attribute

The following conclusions can be drawn: most of the characteristic mean values

are between 1/4-1/2 quantile, and the noise value is small and basically non-existent;
A few noise values exist between 1/2-3/4 quantile.

3.2 Data Analysis

Plotting cnt against other attributes in a box plot:

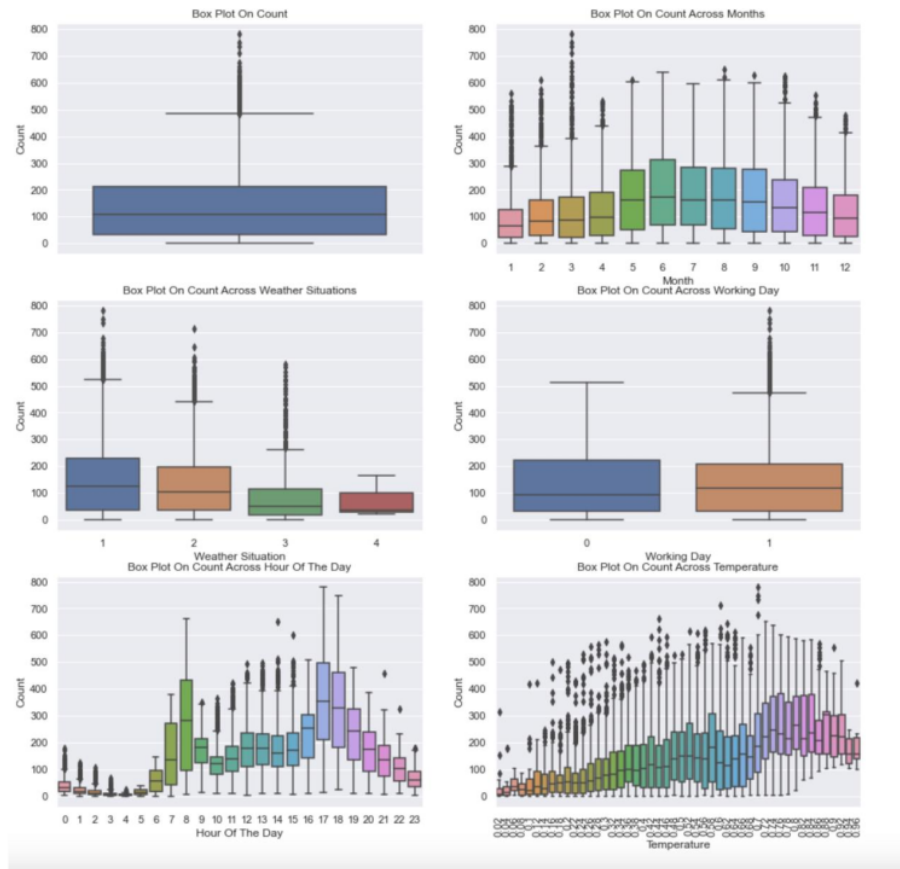


Figure 3-5. cnt against other attributes in a box plot

The box plots lead to the subsequent conclusions: Bicycles are hired with greater frequency on regular weekdays compared on weekends or holidays, according to the weekday and holiday box plots. The hourly box plots reveal local peaks around 8:00 a.m. and 5:00 p.m., showing that the majority of customers utilize the bike renting service to commute to work or school. Higher temperatures seem to boost the overall amount of bicycle rentals, whereas chilly temperatures not only decrease the median amount of rentals but also increase the amount of outliers in the data.

Histogram/distribution of the total number of car rentals cnt:

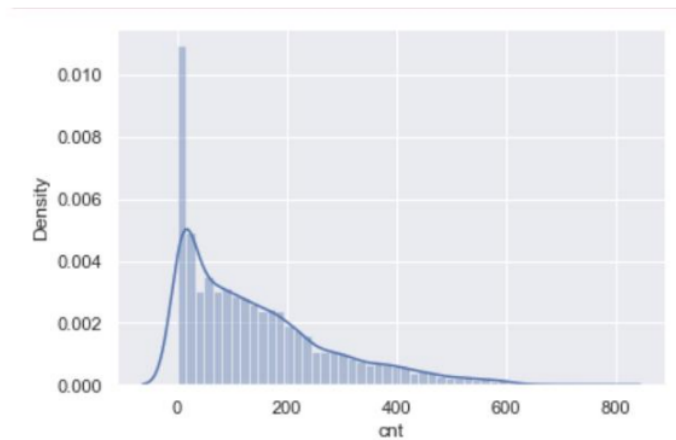


Figure 3-6. Histogram-distribution of cnt

The probability distribution plot of cnt demonstrates that a normal distribution is not followed by the count values. To locate and eliminate outliers from the data, we will apply the median as well as the interquartile range (IQR). The desired values can also be transformed into a normal distribution and used together with the mean as well as the standard deviation.

The distribution map after processing is:

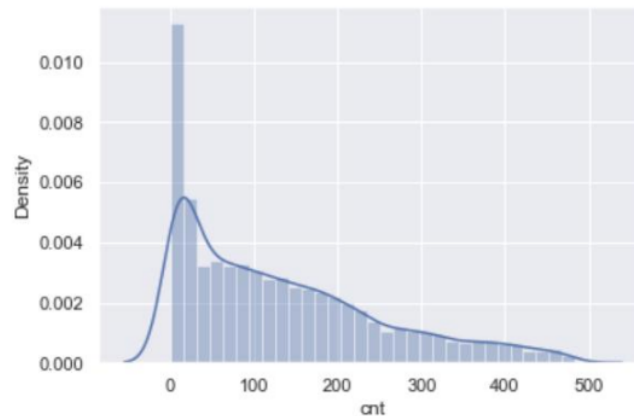


Figure 3-7. Histogram/distribution of cnt (after processing)

Plotting unique heat maps for correlation analysis of features:

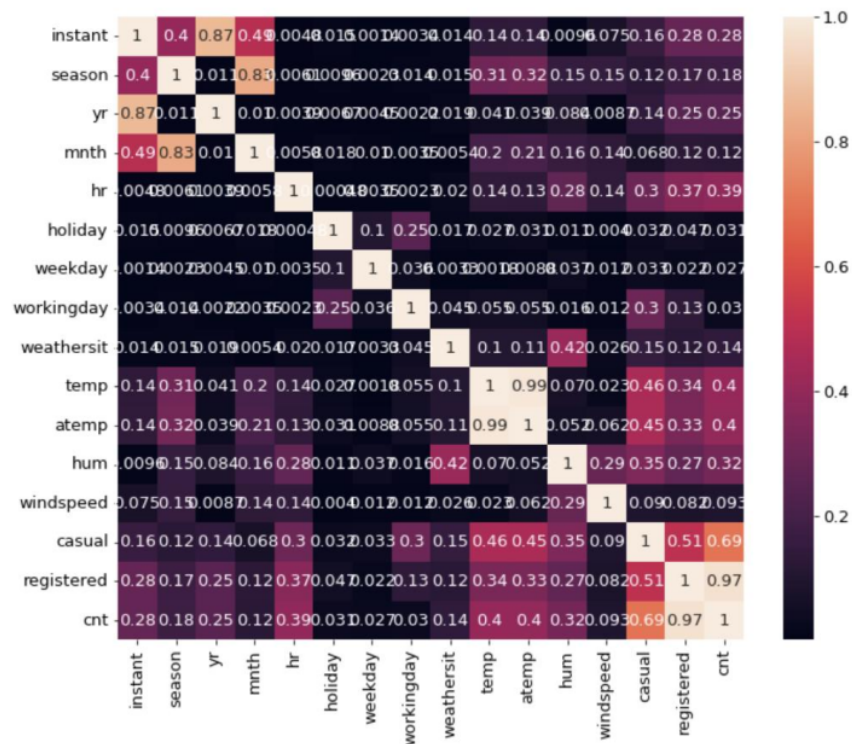


Figure 3-7. Unique heat maps of features

Setting the correlation threshold at 0.5, the correlation is considered highly relevant when it is greater than 0.5, from which it follows that:

```
weathersit and temp = 0.99
casual and registered = 0.97
instant and season = 0.87
dteday and yr = 0.83
windspeed and registered = 0.69
windspeed and casual = 0.51
```

Figure 3-8. Highly correlated features

From this, we can conclude the following:

1. the variables "temp" and "atemp", weather temperature and human sensory temperature are directly related. In order to reduce the dimensionality of the prediction model, we can delete the feature "atemp".
2. the variables "atemp" and "cnt" are related to the number of rental cars, and there will be more rental cars if the temperature is suitable.
3. variables "weathersit" and "hum" weather conditions will directly affect the moderate, so the correlation is high.
4. The variables "hr" and "temp" seem to be the characteristics that contribute more to predicting the number of bikeshare.

5. The variables "Casual" and "registered" contain direct information about bike share counts, which would lead to data leakage if used for prediction. Therefore, they are not considered in the feature set.

4 Data Analysis with Machine Learning

4.1 Model Selection

The qualities of the issue that is being raised are:

Regression: The target variable has a value of the continuous kind.

A small sample size of fewer than 100K.

Few characteristics should be important since, according to the correlation matrix, only a small number of them may be used to forecast the target variable.

These characteristics provide an excellent chance for techniques like ridge regression, support vector regression, integrated regression, and random forest regression to significantly improve the results. Therefore, we will evaluate the performance of these models as follows:

Model	Mean Squared Error	R ² score
SGDRegressor	52202.88	-0.12
Lasso	43103.36	0.07
ElasticNet	54155.92	-0.17
Ridge	42963.88	0.07
SVR	50794.91	-0.09
SVR	41659.68	0.10
BaggingRegressor	19283.78	0.58
BaggingRegressor	38698.05	0.17
NuSVR	41517.67	0.11
RandomForestRegressor	18949.93	0.59

Figure 4-1. The performance of models

Therefore we finally chose the random forest method as our method to predict cnt.

4.2 Random Forest Algorithm

Random forest is an integrated learning method that combines multiple decision trees together for prediction tasks such as classification and regression. The basic process of the random forest algorithm is as follows:

The random forest algorithm will randomly select N samples from the training set for training each decision tree, where N can be less than or equal to the training set size.

For each decision tree, it randomly selects m features from all features ($m \ll$ total number of features) and performs node splitting based on these m features.

The decision tree is predicted by voting when the classification algorithm is used, and by averaging when the regression algorithm is used.

The above steps are repeated several times to construct multiple decision trees.

The final decision result is obtained by voting or averaging the predictions of each decision tree according to certain weights.

The random forest algorithm has the following advantages:

The random forest algorithm has better scalability and robustness for high-dimensional data and large data sets.

The random forest algorithm is more capable of handling missing data.

Random Forest algorithm can detect important features in the sample, but it is achieved by random selection, and there may be some errors when the data volume is small.

The training process of random forest algorithm is not prone to overfitting problem.

There are more application scenarios for the random forest algorithm, such as recommendation systems, credit evaluation, classification, regression, etc. The algorithm performs well in practical applications and is widely used.

4.3 Result of Random Forest

The result of the random forest is:

Model	Dataset	MSE	MAE	RMSLE	R ² score
RandomForestRegressor	training	297.61	10.90	0.21	0.98
RandomForestRegressor	validation	19006.60	96.51	0.47	0.59

Figure 4-2. Evaluation indicators of Random Forest

Plotting the importance of features in a random forest:

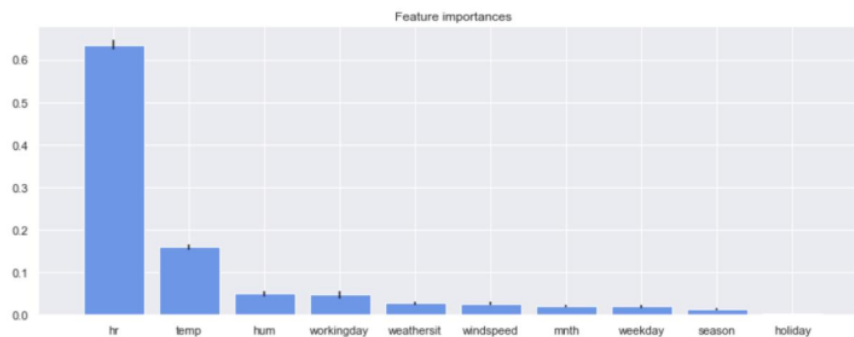


Figure 4-3. Importance of features of Random Forest

From the above analysis, we can see that time and temperature are the main characteristics that affect the number of people using shared bikes.

5 Data Analysis with Deep Learning

We now use this dataset to predict bike-sharing usage.

5.1 Secondary processing of data

Process Dummies

Dummy Variable, or dummy variable (the name always feels weird), refers to a

variable that reflects an attribute. The value of a dummy variable usually takes 0 or 1, for example, 0 means non-bachelor degree and 1 means bachelor degree. In a lot of data processing, we need to treat the data with dummy variables. For example, the month in this data is represented by 1-12, but the value of the month itself has no numerical meaning, for example, there are more months in February than in January, which is obviously not reasonable.

From the explanation of the dataset properties, we know that season, mnth, weekday, hr, weathersit, yr, holiday are all dummy variables. Among them, yr, holiday have only two values, which can be expressed by 0 or 1 season, mnth, weekday, hr, weathersit are data to be processed, use `get_dummies()` to process the data, y use `concat()` to splice with the original data, and finally use `drop()` to delete the dummy variables in the original data, the processing results The result is as follows:

yr	holiday	workingday	temp	hum	windspeed	casual	registered	cnt	season_1	...	weekday_1	weekday_2	weekday_3	weekday_4	weekday_5	weekda
0	0	0	0	0.24	0.81	0.0	3	13	16	1	...	0	0	0	0	0
1	0	0	0	0.22	0.80	0.0	8	32	40	1	...	0	0	0	0	0
2	0	0	0	0.22	0.80	0.0	5	27	32	1	...	0	0	0	0	0
3	0	0	0	0.24	0.75	0.0	3	10	13	1	...	0	0	0	0	0
4	0	0	0	0.24	0.75	0.0	0	1	1	1	...	0	0	0	0	0

5 rows × 60 columns

Figure 5-1. Process Dummies

Standardizing

The results of standardizing the dataset are as follows:

yr	holiday	workingday	temp	hum	windspeed	casual	registered	cnt	season_1	...	weekday_1	weekday_2	weekday_3	weekday_4	we
0	0	0	0	-1.334609	0.947345	-1.553844	-0.662736	-0.930162	-0.956312	1	...	0	0	0	0
1	0	0	0	-1.438475	0.895513	-1.553844	-0.561326	-0.804632	-0.823998	1	...	0	0	0	0
2	0	0	0	-1.438475	0.895513	-1.553844	-0.622172	-0.837666	-0.868103	1	...	0	0	0	0
3	0	0	0	-1.334609	0.636351	-1.553844	-0.662736	-0.949983	-0.972851	1	...	0	0	0	0
4	0	0	0	-1.334609	0.636351	-1.553844	-0.723582	-1.009445	-1.039008	1	...	0	0	0	0

5 rows × 60 columns

Figure 5-2. Standardizing

5.2 Deep learn network

We just employ one hidden layer, and the sigmoid function is used as the activation function connecting input and hidden. The total amount of iterations is 2000, the learning rate is 0.8, and the total amount of hidden nodes is 12, the prediction accuracy is as follows:

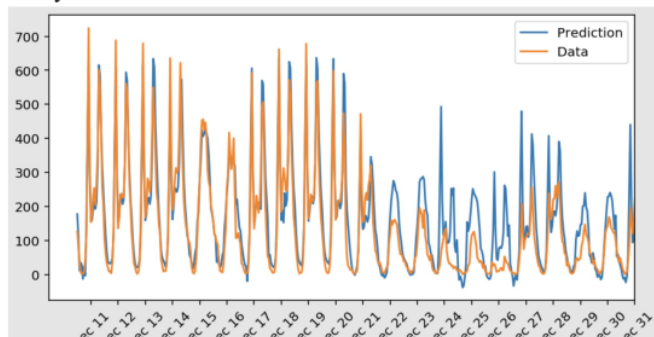


Figure 5-3. Prediction accuracy

6 Reason discussion

6.1 Characteristics of Random forest

- Principal benefits of Random forest:

- (1) Excellent performance and a significant edge over competing algorithms. Random forest doesn't require to perform feature selection since it can handle very high dimensional data (data with numerous characteristics).
- (2) Random forest may reveal which qualities are more crucial after training.
- (3) Because the trees are trained independently of one another, the training process is quick and simple to parallelize.
- (4) During the training process, it may identify the effect between FEATURES.
- (5) Random forest can balance the error in imbalanced datasets. Random forest may be a useful tool for balancing the dataset's inaccuracy when there is an imbalance in categorization.
- (6) Accuracy can still be upheld even if a significant portion of the characteristics are absent.
- (7) There is minimal interference when using the random forest technique. So it is also a good idea to use RF when there is a lot of missing data.
- (8) Random Forest has an exceptional capacity to withstand overfitting (even though mathematically the random forest doesn't generate the overfitting phenomena, in reality the effect of noise is not insignificant, even though expanding the tree may decrease overfitting, there isn't any way to entirely eradicate overfitting, irrespective of how you boost the total amount of trees, and once more the number of trees can't be raised enormously).
- (9) The generalization capability of the random forest generation error (generalization error) employing an unbiased estimate model is high.

- Principal drawbacks of Random forest:

- (1) Since it does not provide a continuous output, random forest analysis does not do as well when handling regression issues as it does in classification. Regression algorithms like Random Forest are limited in their ability to predict outcomes outside the boundaries of the data in the training set, which might cause overfitting when analyzing certain kinds of noisy data. (PS: Research on several noisy regression, classification, and other issues has revealed that random forests overfit).
- (2) For some statistical designers, random forests are like a mysterious black box over which they have no control. Only experimenting with various parameters and random seeds is possible.
- (3) There could be a lot of decision trees that look alike, hiding the actual outcomes.
- (4) Small amounts of data as well as data that is low-dimensional (data with few characteristics) may not generate appropriate categorization. The advantages of random forests are that they can handle high-dimensional data, feature-missing data, and imbalanced data.
- (5) Processing data is quicker than boosting etc. but significantly slower than a single decision tree. The random forest is bagging.

6.2 Characteristics of Full connected neural network

- Principal benefits of fully connected neural network:

(1) No requirement for fixed size input and integration of all information. For most neural network layers, fixed input size is required, but for the Dense fully connected layer or the whole fully connected neural network, no input size is required, as long as all are connected, thus ensuring the integration of all information.

- Principal drawbacks of fully connected neural network:

(1) Large number of parameters and poor expressiveness of spatial structure (tiling down the dimensional information). The number of parameters in a single Dense fully-connected layer is multiplied to the next layer, resulting in a large number of parameters, because the fully-connected processing is equivalent to linear combination of features, truncation, for example, three-dimensional information will only be represented as two-dimensional information through the fully-connected layer.

6 Conclusion

This project uses the Bike Sharings dataset project dataset from the UCI website, which contains the number of hourly bicyclists per day from January 1, 2011 to December 31, 2012. The bicyclists are divided into casual and registered users. This project uses Python and WEKA, through machine learning and other data analysis and data mining techniques, to firstly perform descriptive statistics on the data and complete the pre-processing of the data; secondly use the dataset to visualize and analyze the correlation of each attribute of the dataset and the distribution of each attribute; finally build a prediction model of bike-sharing users based on the dataset. Based on the evaluation indexes of machine learning, this paper finds that the random forest model can effectively predict the number of bicycle users, in which the two features "hr" and "temp" are more important, and the number of users can be predicted from the perspective of these two features.

In the bike-sharing model, the ride length, departure, and arrival locations are explicitly documented in the system, unlike other modes of transportation like buses or subways. This feature transforms the bike-sharing program into a fictitious sensor network that monitors a city's movement. Thus, it is hoped that by tracking this data, numerous significant events taking place in a city can be tracked. New smart hotels have appeared with the introduction of the web and the era of big data. One of the components that is unavoidably difficult to be absent is data mining, and this architecture has excellent practicality.

Through the study, design and implementation of this design, it makes me have a deeper understanding of Python programming and WEKA application. Through data mining and data analysis, many conclusions are drawn, which are also of great significance in reality, and I hope to use the data processing methods more flexibly in future practice. Through the use of a variety of feature models, a deeper understanding of feature engineering, as well as the ability to use simple machine learning algorithms, is important for future learning.

References

- [1] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', *Progress in Artificial Intelligence* (2013): pp. 1-15, Springer Berlin Heidelberg.
- [2] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. "The WEKA Data Mining Software: An Update." *ACM SIGKDD Explorations Newsletter* 11 (1): 10–18. doi:10.1145/1656274.1656278.
- [3] Tan, P.-N., Steinbach, M., & Kumar, V. (2013). *Introduction to Data Mining*.
- [4] Xindong Wu, et al. "Data Mining with Big Data." *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, June 2013, pp. 97 – 107, <https://doi.org/10.1109/tkde.2013.109>.
- [5] Liaw, Andy, and Matthew C. Wiener. *Classification and Regression by randomForest*. Jan. 2007.
- [6] Lawrence, Rick L., et al. "Mapping Invasive Plants Using Hyperspectral Imagery and Breiman Cutler Classifications (RandomForest)." *Remote Sensing of Environment*, vol. 100, no. 3, Dec. 2005, pp. 356 – 62, <https://doi.org/10.1016/j.rse.2005.10.014>.
- [7] Au, Timothy C. "Random Forests, Decision Trees, and Categorical Predictors: The 'Absent Levels' Problem." *Journal of Machine Learning Research*, Jan. 2018.
- [8] Asuncion, A. *UCI Machine Learning Repository*. Jan. 2007.
- [9] Pedregosa, Fabian, et al. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research*, Feb. 2011.
- [10] LeCun, Yann, et al. "Deep Learning." *Nature*, May 2015, pp. 436 – 44, <https://doi.org/10.1038/nature14539>. Hans Weytjens, Enrico Lohmann, & Martin Kleinstaubert (2021). Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet *Electronic Commerce Research*.
- [11] Schmidhuber, Jürgen. "Deep Learning in Neural Networks." *Neural Networks*, Oct. 2014, pp. 85 – 117, <https://doi.org/10.1016/j.neunet.2014.09.003>.
- [12] LGoodfellow, Ian, et al. "Deep Learning." MIT Press eBooks, Nov. 2016.
- [13] Paszke, Adam, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *Neural Information Processing Systems*, Jan. 2019.
- [14] van der Walt, Stéfan, et al. "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science & Engineering*, vol. 13, no. 2, Mar. 2011, pp. 22–30, <https://doi.org/10.1109/mcse.2011.37>.

Data Mining & Visualization

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

Instructor

63/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

YANG LU

63

KNOWLEDGE AND UNDERSTANDING

Dataset description (15 marks)



Case dataset taken from UCI's core repository appraised and core rationale for its examination addressed in your submission.

12

COGNITIVE/ INTELLECTUAL SKILLS

Primary and Secondary data analysis, reflection (35 marks)



Stages of analyses for interpreting bike sharing dataset explained sufficiently well for replication of actions to be undertaken.

23

PRACTICAL APPLICATION SKILLS

Primary and Secondary data processing (20 marks)



Preprocessing using Jupyter undertaken sufficiently.

14

TRANSFERABLE SKILLS

Data reporting and report writing (30 marks)



Figures included are expounded on and implications of results addressed well.