Insert Your Title Here*

Insert Subtitle Here

FirstName Surname†
Department Name
Institution/University Name
City State Country
email@email.com

FirstName Surname
Department Name
Institution/University Name
City State Country
email@email.com

FirstName Surname
Department Name
Institution/University Name
City State Country
email@email.com

ABSTRACT

Bike sharing is a method of renting bicycles. With the popularity of smart phones, shared bicycles, as an important part of the urban transportation system, are convenient, efficient, economical, and environmentally friendly. Data such as rental time, rental, return location, and time consumption generated by this rental system have attracted people's attention. The rental system has also become a perception network. The aim of our project is to predict the rental needs of the rental system with historical data from Washington.

KEYWORDS

Data Analysis, Data Visualisation, Feature Engineering，

Feature Selection，Random Forest，Grid Search

1 Introduction

The report is aimed to predict bike sharing demand by studying from different variables. After some preparation on this problem and searching other authors' methodology that try to solve it before, the first step was to download data from Kaggle. Then we used Python to visualize the data and random forest was applied to fill the missing values. Then it goes to feature extraction and training to find the best and most accurate model. RMSLE is to measure our performance. The smaller the value is, the better the model fits the data. After taking the logarithm of the data after the algorithm predicts and taking the natural constant exponent, the RMSLE criterion can be met as much as possible.

2 Dataset

Our project aims to predict the number of rental bikes under specific conditions through the historical data given, including the characteristics of hourly weather, temperature, time, season, holiday or workingday, windspeed, number of registered and non-registered users for two years. The training set is the first 19 days of each month, and the test set is from the 20th to the end of each month. We will only use the information available before the rental period to predict the total number of bicycles rented per hour covered by the test set.

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
registered	number of registered user rentals initiated
count	number of total rentals

Figure 2: Data Field

3 Data preprocessing

3.1 Import data

We can observe the train data set and test data set through train.shape, test.shape, train.info() and test.info(). As shown in Figure 1 and 2, there are 12 fields in the train data set and 9 fields in the test data set. It can be found that, compared with the train data set, the test data set does not have the three fields of 'casual', 'registered' and 'count'. In this article, we will predict their values through the models.

```
<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
```

A

```
<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
```

B

Figure 3.1: Information of data set. A: Train data set; B: Test data set

3.2 Check for missing values

From the visual image in Figure 3.2, we can find that there is no missing data in these two data sets. However, no missing does not mean there are no outliers.

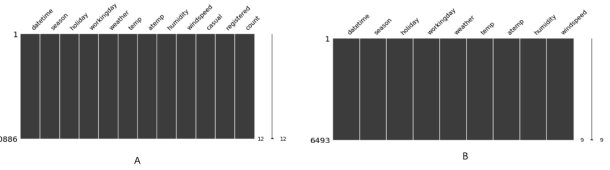

Figure 3.2: Check for missing values. A: Train data set; B: Test data set

3.3 Check for duplicate values

In order to facilitate the simultaneous data cleaning of the training data set and the test data set, the two data sets were merged. From the Figure 3.3, we can see that there are no duplicate values in either dataset.

```
Duplication:   (17379, 12)
Deduplication:   (17379, 12)
```

Figure 3.3: Check for duplicate values

3.4 Take the logarithmic of the count

As shown in Figure 3.4.1, The count variable contains a number of outliers that are skewed to the right (data points that exceed the upper quartile limit). Therefore, we hope to deal with the long tail of this column of data.

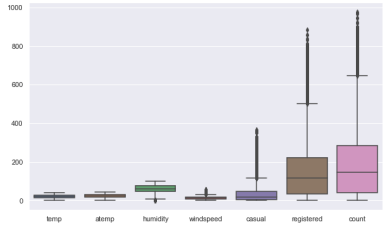

Figure 3.4.1: Visualize outliers

In this step, we take the logarithmic the Y-axis data (count). It can also be found in Figure 3.4.2(A) that the deviation of data density distribution is quite serious and there is a very long tail. Therefore, we hope to deal with the long tail of this column of data and exclude data beyond 3 standard deviations.

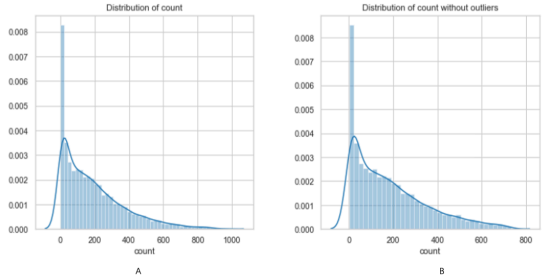

Figure 3.4.2: Logarithm the count. A: Distribution of count; B: Distribution of count without outliers

The comparison Figure 3.4.2(B) shows that almost all the long tail after about 800 has been removed, but the data still fluctuates a lot, which is prone to over-fitting.

Therefore, we want to choose logarithmic changes to make the data relatively stable. As shown in Figure 3.4.3, we can see that, after the logarithmic transformation [2], it can modify the right skew of data to a certain extent, so that it is closer to normal distribution. Taking logarithm can transform the nonlinear variable relation into linear relation, which is more convenient for parameter estimation.

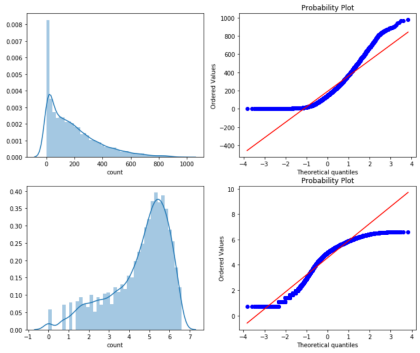

Figure 3.4.3: Logarithmic transformation

3.5 Convert datetime format

Since the datetime format is string, it is not convenient for data analysis later. The datetime is divided into date, hour, year, month and weekday in Table 3.5

	atemp	casual	count	holiday	humidity	registered	season	temp	weather	windspeed	workingday	date	hour	year	month	weekday
0	14.395	3.0	16.0	0	81	13.0	1	9.84	1	0.0	0	2011-01-01	0	2011	1	6
1	13.635	8.0	40.0	0	80	32.0	1	9.02	1	0.0	0	2011-01-01	1	2011	1	6
2	13.635	5.0	32.0	0	80	27.0	1	9.02	1	0.0	0	2011-01-01	2	2011	1	6
3	14.395	3.0	13.0	0	75	10.0	1	9.84	1	0.0	0	2011-01-01	3	2011	1	6
4	14.395	0.0	1.0	0	75	1.0	1	9.84	1	0.0	0	2011-01-01	4	2011	1	6

Table 3.5: Datetime conversion

3.6 Fill the wind speed

We can be found some problems through this distribution in Figure 3.6(A). For example, there is a lot of windspeed data with zero values and the values (1-6) are of the vacancy. We seem to infer from this that the data itself may have missing values, but it is filled with zeros. However, these values are zero for windspeed data can interfere with the prediction.

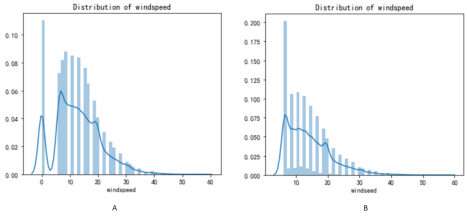

Figure 3.6: Distribution of windspeed. A: Distribution before using the random forest to fill the missing value of windspeed; B: Distribution after using the random forest to fill the missing value of windspeed.

Therefore, we want to use the random forest [3] to fill the missing value of windspeed according to the same year, month, season, temperature, humidity and so on several characteristics. The result as given in Figure 3.6(B).

4 Influencing factors data visualization analysis

4.1 Influence of hour

During working days, there are two peak hours for registered user to commute, and there is also a small peak at lunch time. As for hours, the number of bikes used at 8am and 5pm is the largest. For casual users, the ups and downs are relatively smooth, the peak is around 5pm, while the number of registered users far exceeds the number of casual users; For non-working days, the number of users presents a normal distribution over time, with the peak at about 2pm and the valley at about 4am, and the distribution is relatively even.

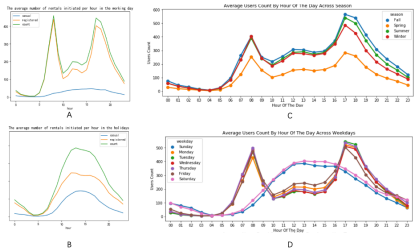

Figure 4.1: The change of bike rental quantity under the influence of time period; A: line diagram of casual, registered, and total number of users in working day within 24 hours; B: line chart of casual, registered and the total number of users in non-working day; C: line chart of the number of users on weekday; D: line chart showing the number of users in each season.

4.2 Influence of temperature

Both atemp (sensible temperature) and temperature change periodically. Since temperature is highly correlated with atemp (positive), we only focus on temperature. The average temperature increases from January to July, but the temperature begins to drop after July and reaches the lowest temperature in January. Atemp has the same trend, which can be seen in Figure 4.2.

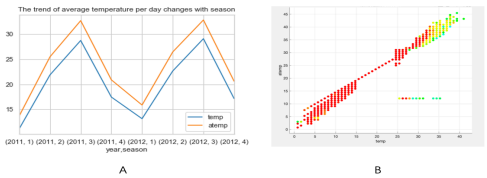

Figure 4.2.1: Temperature and atemp. A: the change of temperature and atemp within two years. B: the correlation analysis between temperature and atemp. Temperature and atemp are positively correlated.

Both the number of casual and registered users show a trend of increase as the temperature rising, but it starts to drop when the temperature exceeds 35 degrees and reaches its lowest point at 4 degrees.

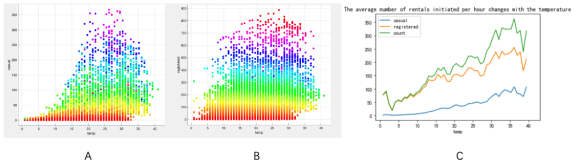

Figure 4.2.2: Variation of the number of users under the influence of temperature; A: scatter diagram of casual users with temperature; B: scatter diagram of registered users with temperature; C: line chart of casual, registered users and total users with temperature.

4.3 Influence of humidity

The change of humidity is not very big. The whole data fluctuated around 60, with a peak of 80. And the number of users quickly peaks around humidity 20 and then slowly declines.

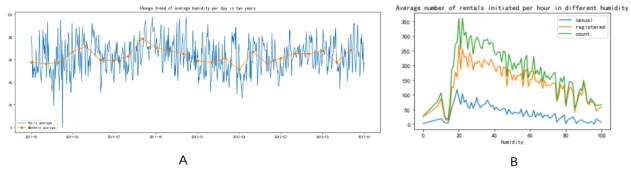

Figure 4.3: Variation of the number of users under the influence of humidity. A: line chart of humidity from January 2011 to December 2012; B: line chart of the number of casual, registered, and total users with humidity.

4.4 Influence of year and month.

There are a lot of local troughs. The rental situation of Shared bikes increased in 2012 compared with 2011.The data fluctuates significantly with months; Data from September to December in 2011, March to September in 2012 fluctuated wildly.

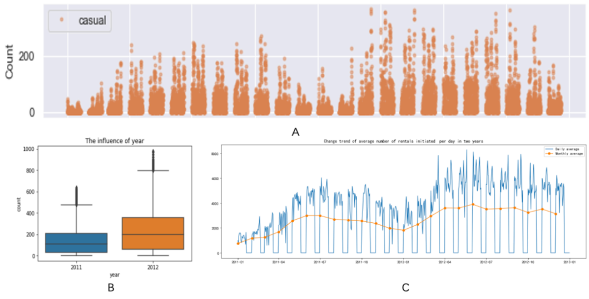

Figure 4.4:Changes in bike rental in 2011 and 2012. A: change of the number of casual users, which can be found consistent with periodicity. B: boxplot of the total number of leases in 2011 and 2012. C: line chart of the total users over time.

4. 5 Influence of season

The number of both casual and registered users peaks in fall, while the number is lowest in the spring. The overall number of users in 2012 was higher than in 2011.By analyzing season, there is no significant difference in the number of bikes, indicating that season has no great influence on the use of bikes.

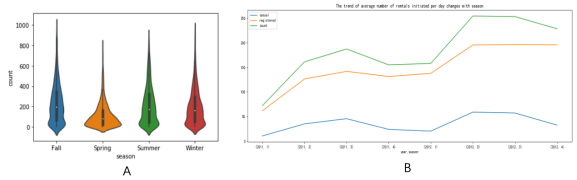

Figure 4.5: Variation of users under influence of season. A: organ diagram of the number of users per season. B: line chart of the number of casual, registered and total users with seasonal variation.

4. 6 Influence of weather

The total number of users for each type of weather and the weather is shown in Figure 4.6.

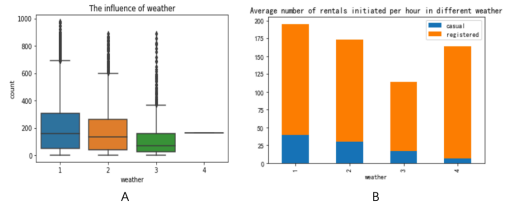

Figure 4.6: The number of users changes under the influence of the weather. A: boxplot of users for each weather. B: the average rental quantity for each weather.

There is some reasonable data: when the weather is 4 but people travel a lot, especially the number is even higher than type 2 weather. According to figure 4.6(A), people travel in type 4 weather should be at least. Data for weather 4 is shown in Table 4.6, we can see the data is generated during the week, which is not typical.

datetime	season	holiday	orkingda	weather	temp	atemp	humidity	indspee	casual	registered	count	windspeed_rfr	date	hour	year	month	weekday
2012/1/9 18:00	1	0	1	4	8.2	11.365	86	6.0032	6	158	164	6.0032	2012/1/9	18	2012	1	1
2011/1/26 16:00	1	0	1	4	9.02	9.85	93	22.003	NaN	NaN	NaN	22.0028	2011/1/26	16	2011	1	3
2012/1/21 01:00	1	0	0	4	5.74	6.82	86	12.998	NaN	NaN	NaN	12.998	2012/1/21	1	2012	1	6

Table 4.6: Data for weather type 4

4.7 Data for weather type 4

The wind speed fluctuated and increased between September 2011 and December 2011 to March 2012. Considering that high wind speed is very rare, so we just take max data according to wind speed. There is no obvious periodic pattern in wind speed. The number of leases decreases as the wind speed increases and decreases significantly when the wind speed exceeds 30.

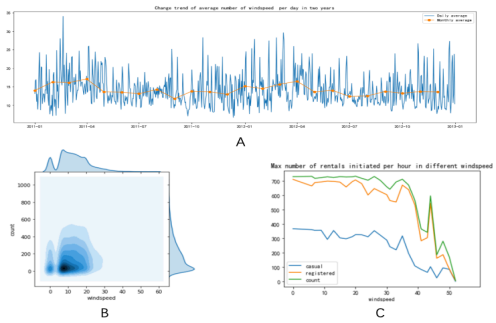

Figure 4.7: Variation of users under wind speed. A: line chart of wind speed changes from January 2011 to December 2012; B: density diagram of wind speed to users; C: line chart of the change in the wind speed of the number of casual, registered, and total users.

However, there was a rebound when the wind speed was around 40. This data is shown in Table 4.7. This data was generated during the rush hour, so it is also an outlier which is not representative.

datetime	season	holiday	orkingda	weather	temp	atemp	humidity	indspee	casual	registered	count	windspeed_rfr	date	hour	year	month	weekday
2012/3/8 17:00	1	0	1	1	25.42	31.06	38	43.999	52	545	597	43.9989	2012/3/8	17	2012	3	4

Table 4.7: Abnormal wind speed data

4. 8 Influence of date

The data are summed by day and averaged by other dates data. Due to the difference in the number of days of working days and other days, the average number of users on working days and non-working days was taken, and the sum of the number of users on each day of the week was calculated. Among them, the number of registered users in weekday is more than casual users. The number of bikes used by registered users in weekend decreased, while the number for casual users increased.

Since there are few holidays in a year, we just take a look at the annual number, which 10 days in 2011, and 11 days in 2012. Registered or casual users in holiday use shared bikes more than in non-holiday.

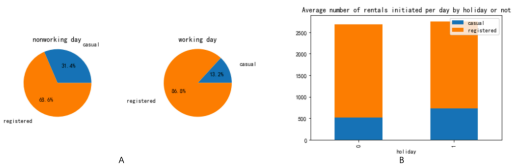

Figure 4.8: The number of users changes under the influence of date. A: pie chart comparing the quantity of casual and registered users in working days and non-working days; B: bar chart comparing the number of casual and registered users on holidays and non-holidays.

5 Feature Engineering

In the second part of data preprocessing, we did some characterization. To sum up, there are three operations. Firstly, like the example of the figure 3.6 Take the logarithmic of the count, the outliers of 'y' have been processed. We have dealt with the long tail of this column of data and exclude data beyond 3 standard deviations. The result is shown in Figure 3.6.1. Then logarithmize the 'y' value. After taking the logarithm, we can correct the right-biased shape of the data to a certain extent to make it closer to normal. The use of log can transform the nonlinear variable relationship into a linear relationship, which is more convenient to parameter estimation.

The second operation is to convert the training set time datetime. Convert to (data, hour, year, weekday, month). Because random forest prediction is needed when the model is finally built, the specific operation is shown in 3.7.

Step 3, we use Random Forest algorithm to fill the windspeed, the specific operation steps are shown in 3.8.

6 Feature selection

Feature selection and extraction is also an important step in this project. Good feature selection can improve the performance of the model. Help us understanding the data characteristics better and get a better underlying structure. In addition to that, it also play a important role in improving models and algorithms. As shown in 6.1 and 6.2

We can know from the two figures that atemp and temp have a strong correlation with each other, so the atemp variable is not considered. Furthermore, features like 'casual' and 'registered' features are missing in the test set, so they are not considered; The last feature we left is the higher correlation coefficient with count, such as: humidity, temp, windspeed, hour, season, month,and weather. In the following algorithm, since the CART decision tree part uses binary classification, we use one-hot encoding to convert multi-category data (season, weather) into multiple dichotomous categories.

7 Select and Train Models

In this project, our main purpose is prediction. Here we have selected common algorithms suitable for prediction such as Linear Regression, Ridge Regularization Model and Lasso Regularization Model algorithms, as well as Random Forest and Gradient Boost algorithms in Ensemble model.
First, we use the feature dataset selected in step 6 to train the model, and then predict the count in the train dataset. Next, substitute the predicted count number and the actual train data set count into rmsle (which was already explained in the second step) to measure the model performance. The smaller the rmsle here, the better the model's fitting ability. After our training, the specific results of training are shown as Table 1.

Model	Rmsle
Linear Regression Model	1.0038124283371648
Grandient Boost	0.2045955069924137
Regularization Model - Ridge	1.0046345184801875
Random Forest	0.11085148914451422
Regularization Model - Lasso	1.0039934908275858

8 Fine-Tune the Model

Through the analysis of the previous steps, we can see from the table 1 that the performance of the Random Forest model is significantly better than several other models. So in this step, we chose the random forest model to adjust its parameters, hoping to get better results.
In the random forest algorithm, the main parameters of the random forest are n_estimators (the number of subtrees), max_depth (the maximum growth depth of the tree), min_samples_leaf (the minimum sample number of leaves), min_samples_split (the minimum sample number of branch nodes), max_features (Maximum number of selected features). Here we choose n_estimators and max_depth that have a greater impact on the performance of the random forest model for adjustment.

First of all, adjust the parameters of n_estimators. We build a random forest for each n_estimators from 0 to 200, get the rmsle scores of different n_estimators and draw them into an image, as shown in the figure 8.1.1.

Figure 8.1.1: The learning curve Figure 8.1.2: Final prediction

Here we find that the effect when using n_estimators = 129 is better than the effect of = 100 in the previous model.

Next, we use n_estimators for 129 to build a random model to adjust the parameter of max_depth #, and use a grid search to adjust max_depth. The result is shown in the figure. It can be seen that when max is 19, the effect is better. So, in the end our model is a random forest model with n_estimators = 129 and max_depth = 19.

Most importantly, we use this model to predict the count data of the test training set, and the prediction results are shown in several figures. We submitted the wanted file and got a score on kaggle. The specific results are shown in Figure 8.1.3.

Figure 8.1.3: The score of Kaggle

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] Muhlestein, Whitney E., et al. "Machine learning ensemble models predict total charges and drivers of cost for transsphenoidal surgery for pituitary tumor." Journal of neurosurgery 131.2 (2018): 507-516. DOI: https://doi.org/10.3171/2018.4.JNS18306

[2] Bartlett, Maurice S., and D. G. Kendall. "The statistical analysis of variance-heterogeneity and the logarithmic transformation." Supplement to the Journal of the Royal Statistical Society 8.1 (1946): 128-138. DOI: https://www.jstor.org/stable/2983618

[3] Pal, Mahesh. "Random forest classifier for remote sensing classification." International Journal of Remote Sensing 26.1 (2005): 217-222. DOI: https://doi.org/10.1080/01431160412331269698