THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Prediction of Bike Sharing Demand

By

Chun Yim

School of Information Technology and Electrical Engineering

University of Queensland

Submitted for the degree of Master of I.T

11/2017

Chun Yim

Chun.yim@uq.edu.au

06/11/2017

Prof Michael Brünig
Head of School
School of Information Technology and Electrical Engineering
The University of Queensland
St Lucia QLD 4072

Dear Professor Brünig

In accordance with the requirements of the Degree of {Bachelor of Engineering (Honours)} in the School of Information Technology and Electrical Engineering, I submit the following thesis entitled

"Prediction of Bike Sharing Demand"

The thesis was performed under the supervision of Professor Neil Bergmann. I declare that the work submitted in the thesis is my own, except as acknowledged in the text and footnotes, and that it has not previously been submitted for a degree at the University of Queensland or any other institution.

Yours sincerely

Chun Yim

Chun Yim

ACKNOWLEDGEMENT

The accomplishment of this thesis would not have been possible without the support of many people. I would first like to thank my supervisor Professor Neil Bergmann of the School of Information Technology and Electrical Engineering. His office was always open whenever I had a question about my master thesis. He is willingly to give some useful suggestions and steer me in a right direction if needed. I would also like to thank my tutor Dr. Larissa Meinicke for her contribution. Without her comments of my thesis, the thesis could not be presented in a professional look. Finally, I would like to express my gratitude to my parents for the support and encouragement throughout my years of study. Thank you.

Chun Yim

Abstract

With a significantly increase in the usage of bike sharing service in global, mechanisms that are able to predict the bike rental demand become gradually important. In this paper, I present different linear models and tree based models that work in the bike usage data in Washington, D.C. The dataset is integrated by features selection and features engineering to ensure the quality of data. The results show that weather and time factors are strongly correlated to the demand of bike rental. The prediction model is then evaluated with Root Mean Square Log Error (RMSLE). Parameters tuning, and blending are applied in the model to improve the performance of classifiers. The outcome is that random forest has the best performance against other algorithms. Random forest has successfully control the overfitting and gives a lowest RMSLE. Finally, this prediction model is able to predict the number of bike rental on hourly basis, this information is valuable for bike sharing company to make decision about the supply of bike and bike station in Washington D.C.

List of Figures

List of Tables

Table of Contents

Chapter 1 Introduction

1.1 Significance of prediction of bike sharing demand

Bike sharing system is a service which provides bike rentals to the public for shared use. People can borrow a bike for short distance transit, it allows individuals to borrow a bike from a bike station and return it at another bike station. The service is usually free or inexpensive for a short-term basis to encourage the environmentally friendly mode of transportation. This system began in Europe in 1965 [1]. With the automation of membership, bike rental and return, the bike sharing programs are now available in more than 50 countries, over 800,000 of bicycles are operated at 37,500 stations in 712 cities [2].

With the rapid advancements in information technology, the demand of bike sharing service has increased mostly in recent years. Most of the bike sharing systems have their own smartphone mapping app so that user can check the nearest bike station with available bike and do the payment easily [3]. With the heavy traffic in global cities and the desire of a green transportation with the least environmental impact, bike is now become an acceptable transportation option for public. As a result, the popularity of bike sharing services has grown quickly and the prediction of bike sharing demand has become an important problem to address.

1.2 Objectives of this thesis

As the bike sharing systems are used by millions of users everyday around the world, a large volume of data is generated, such as the duration of travel, departure location, arrival location, and time elapsed. These historical usage data are useful when study the mobility of a city and used to determine the location of bike stations. In addition, demographic and weather are important factors when determine the bikes renting demand in a city. And thus, machine learning will be a right tool to understand and identify the data patterns [4]. By estimating the bike rental usage and flow, we can build up a prediction model to forecast the bike rental demand. This model could be used as a planning tool when bike sharing company decide to expand the city's bike sharing system.

1.3 Problem Statement

Over the past few years, several studies have analyzed the factors which will affect the bike rental usage and flow. Alexander R. et al [5], investigates the demographic information and the characteristics of the build environment of different cities, and finds a strong correlation between population density and the bike usage. Etienne and Latifa et al [6], investigates the influence of population and the bicycle lanes, and used clustering to explore the usage of bike sharing systems. Imani et al. [7], uses meteorological data to analyzes the Montreal's Bicycle Sharing System, and concludes the temporal characteristics and built environment attributes that will be affect the bike rental usage and flow. All these studies give a good idea in predicting the bike rental demand. In this task, the topic of Bike Sharing Demand is a competition on Kaggle [8]. The aim of this project is to forecast the use of bike sharing system in Washington, D.C. The historical bike usage patterns are used to study the mobility of the city. The bike usage and flow are then combined with the weather data of Washington, D.C to build up a prediction model. Therefore, several machine learning algorithms are implemented to build a decent prediction model. The model should be able to predict the number of bike rental on hourly basis which is beneficial for various civic and logistical applications. Also, an accurate prediction of bike rental demand is helpful for traffic management and logistical support for the bike sharing services in the future.

Chapter 2 Literature Review

In past, many people have studied the bike rental demand for many cities around the world. Some of the relevant researches are mentioned below.

D. Singhvi et al. (2015) [19] used the statistics form 332 bike stations in New York City to predict the bike demand for the bike sharing system. The study focuses on the bike usage on the weekdays morning which between 7 AM to 11 AM, and use taxi usage, weather and, demographic variables as covariates to predict the bike demand. They use regressing analysis to build the prediction model and use log-log regression models to examine the results based on different weather and temporal characteristics. The average Root Mean Square Log Error (RMSLE) was 0.42 which shows a good level of accuracy. The authors concluded that the prediction of analyzing pairwise bike demand at neighborhood level are much better than individual station in bike sharing systems. This is useful for strategic decisions when decide to expand the bike sharing system in a city.

J. Malani et al. (2013) [20] applied serval machine learning algorithms to predict the hourly bike demand in Washington DC area. The authors collected the quarterly usage of bike from 2011 to 2012. The dataset consists 3,288,891 of rides across 204 bike stations. The authors inserted the population and housing data to the dataset to strengthen the prediction model. Linear Regression, Support Vector Machine and Random Forest are used to build the model. The authors used Root Mean Square Error (RMSE) to evaluate the result which is 0.570283. The authors state that the developed model can forecast the bike demand in urban environment. With adding more covariates and try different analysis models, the prediction can be improved and able to predict the bike demand on hourly basis. Thus, this can assist bike sharing company making a better decision with the number of bike placed at different stations.

R.A. Rixey (2013) [21] investigated serval bike sharing ridership in three operational U.S. systems. The author considered the demographic and built environment characteristics of the bike sharing stations and developed a regression model for predicting station ridership. The regression analysis identified the population density, retail job density, median income levels and the share of alternative commuters are significantly correlated to the prediction of bike sharing ridership. This study suggested that bike sharing system planners should consider those significant factors when determine the extent and spatial distribution of bike stations. The author also integrated the prediction model with the potential destinations of biking network. The model is more widely applicable in different communities than the earlier models. It is useful for bike sharing systems to predict the potential levels of ridership and identify the popular station locations.

Y.Li et al. (2015) [21] investigate the bike sharing system in Washington D.C and predict the bike rental demand for the city. The prediction model is based on hourly basis which improve the performance of the predictor. The authors argue that the covariate of weather variables is less dominant in actual prediction. Root Mean Square Error (RMSLE) is used to evaluate the performance of prediction model. The results show that the performance of CTree and Random Forest are better than the other models as the tree based models could capture the nonlinear interaction effects between variables. The RMSLE of CTree is 0.460 and 0.503 for the Random Forest. The authors have also performed regression model on registered and non-registered users separately, but the regression model works better on combined users with less noise and variance. The authors concluded that CTree and Random Forest have the best prediction accuracy as they could capture the relationship between the data.

Chapter 3 Background

3.1 Introduction of Machine Learning

Machine learning is an emerging technique that gives computer ability to learn without being explicitly programmed [4]. Machine learning focuses on the study and construction of algorithms to expose the pattern of data. By learning from historical relationships and trends on the data, computers can build analytic models and make prediction of data. The analytic models will be significant for bike sharing system to examine the key factors that influence the bike usage and flow. In this thesis, machine learning will be used to perform analysis and applied to data of bike sharing services to find out the correlation between bike rental usage and the weather and time. And thus, a predictive model is formulated to forecast the demand of bike sharing by using serval machine learning algorithms, as will be discussed below.

3.2 Categories of Machine Learning

In machine learning, supervised learning and unsupervised learning are the two main categories for machine learning techniques.

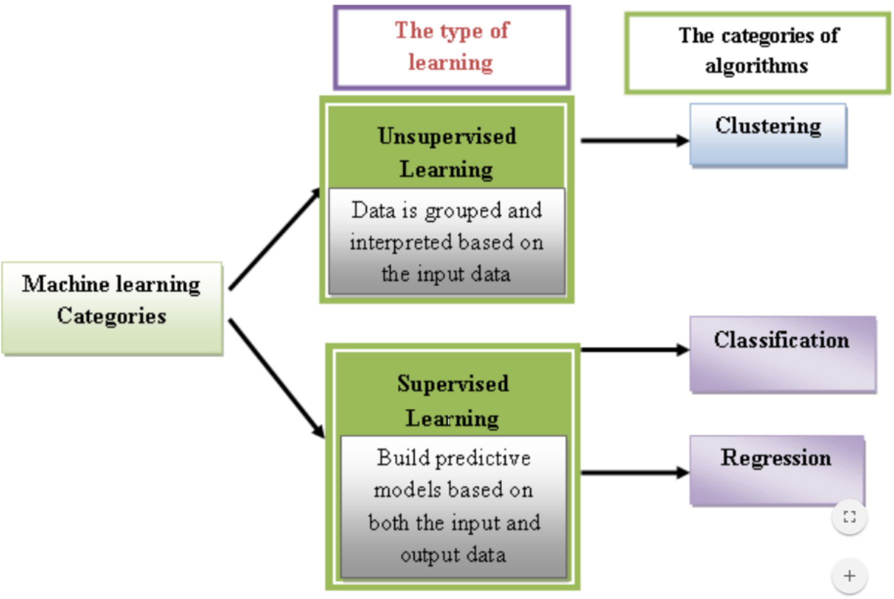

The type of learning
The categories of algorithms
Unsupervised Learning
Data is grouped and interpreted based on the input data
Clustering
Machine learning Categories
Classification
Supervised Learning
Build predictive models based on both the input and output data
Regression

Figure 1 Categories of Machine Learning [9]

3.2.1 Supervised Learning

The aim of supervised learning is to build a prediction model by using a given training dataset. The training dataset consists of both input data (independent variable) and output data (dependent variable). Recent studies [4, 9] indicated that using a larger training dataset will always give a better predictive result. Thus, supervised learning algorithms will identify the patterns of data and generate a function to predict corresponding data of a new dataset. Also, the testing set will be used to verify the quality of prediction of the new dataset.

3.2.2 Unsupervised Learning

The aim of unsupervised learning is to deduce a function which can describe the structure of an unclassified dataset. Unlike supervised learning, there is no corresponding output variable or correct result, therefore, the evaluation of the data structure and the accuracy of prediction results are subjective. Unsupervised learning algorithms are mainly used for association and data clustering; the data structure can be derived by the relationships among the variables in the dataset [16]. Examples of unsupervised learning algorithms include k-means and Apriori algorithm.

3.3 Classification Techniques

Classification techniques are used for predicting categorical-response values. It is a method to assign data into specific classes according to their similarities. The following sections explain some common classification algorithms including Support Vector Machines, Decision Trees, Random Forest and Boosting.

3.3.1 Support Vector Machine (SVM)

Support vector machine is a supervising learning algorithm which used for classification and regression analysis. SVM training algorithm builds a model which assigns new examples to one category or the others after given a set of training examples [10]. The SVM model represents the mapping of samples as points in space, and then the samples of different categories are divided by the distinct gap as wide as possible. The category of new samples is predicted and mapped into the space which they fall. Figure 2 shows the possible decision boundaries when split the data into two class. As two types of data points are completely separated in the figure, both decision boundaries give a perfect training result. However, the margin of decision boundary B1 is larger than B2 which shows a better generalization error. On the other hand, a small margin results overfitting of model. [10]

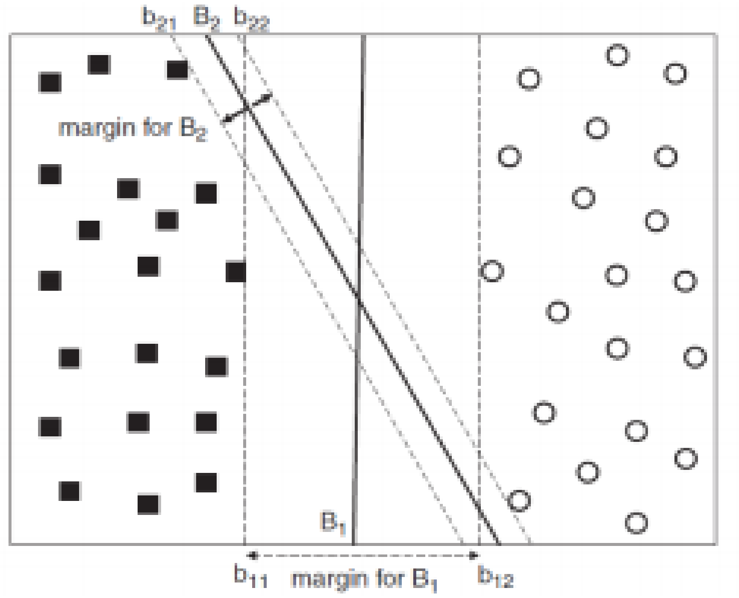


Figure 2 Possible Decision Boundaries [10]

3.3.2 Decision Trees

Decision tree uses a tree-like structure diagram as shown in figure 3 to visualize the classification rules and present the relationships between inputs and outputs. The internal node in decision tree represents a “decision rule” on an attribute. By splitting the rules, the data will be split into a hierarchy of branches, each branch represents the result of the rule. The bottom nodes in decision tree are called leaves and each leaf node represents a class value, impurity measure is commonly used to measure the purity of data in leaf node [11]. Generally, the size of tree is proportional to prediction accuracy. However, overfitting may be occurred when the tree size is too big [11], thus, it is hard to determine the best size of tree. The size of training set, the performance on training set and the complexity of classifiers are also the key factors of good test performance [11]. There are two kinds of tree models in decision trees; classification trees and regression trees. The leaf node in classification trees represents categorical-response values and the lead node in regression trees represents continuous-response values.

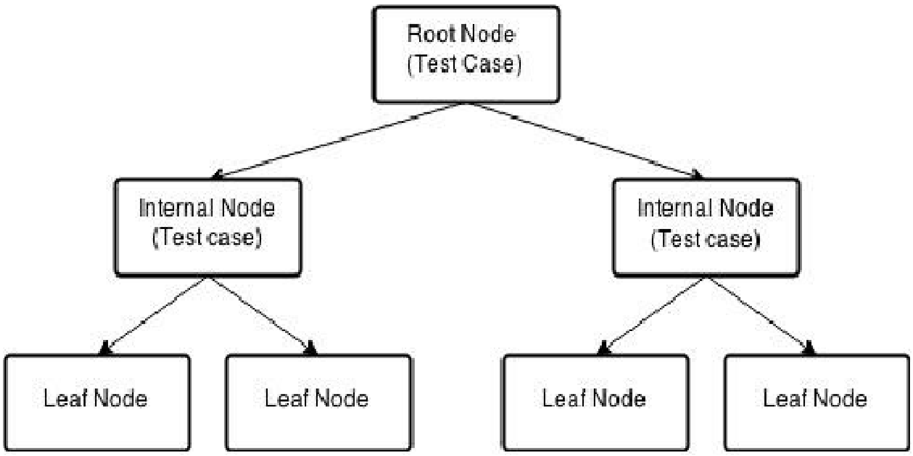


Figure 3 Structure of Decision Tree [11]

3.3.3 Random Forest

Random forest is an ensemble approach for classification problems. It takes a set of variables to construct multiple decision trees. The multiple weak decision trees for different subsets of the training set are combined to form a strong decision tree [12]. By randomly selected the input features to split each node of the decision tree, it can determine the splitting decision of a node. The integrity of the tree grows with no pruning as not all the features are needed to examine. This can reduce bias in the resulting tree. The fundamental idea of random forest is to train a decision tree repeatedly and form a better decision tree. By using multiple such tress, it can reduce the overall error of prediction model. These decision trees will then join to operate the training set and use averaging to improve the predictive accuracy and control over-fitting [12]. The size of the features is correlated to the strength of random forest, decision trees become more correlated with a bigger size of features.

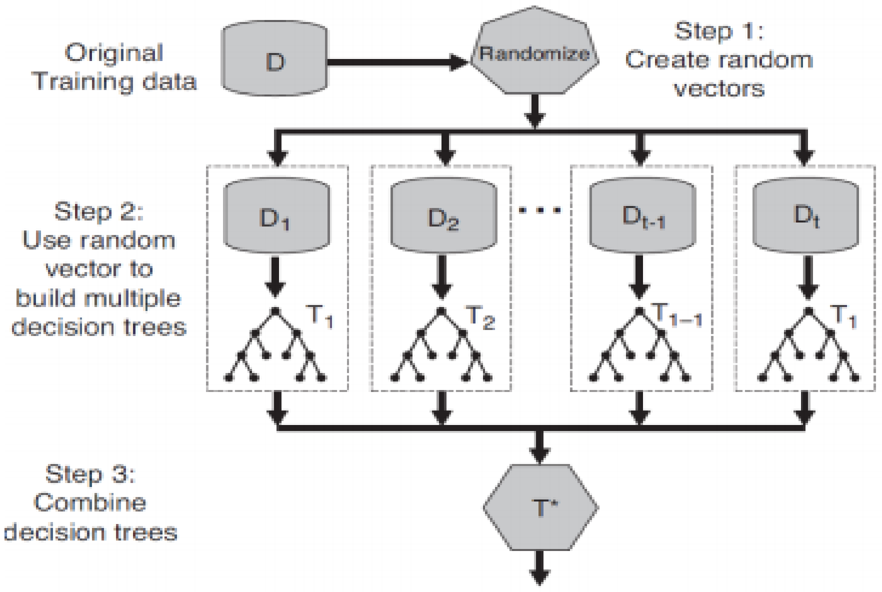


Figure 4 Random Forest [12]

3.3.4 Boosting

Boosting is an ensemble learning method which converts several weak classifiers to a strong classifier in classification problem. Boosting method repeats to creates a new classifier in each stage; the classifier in early stage will be used to train on the error residuals and produce a new classifier which is stronger than the previous stage. With a large amount of data, boosting algorithm will iteratively work in more stages. By combining the outputs from weak classifier, a strong classifier will be created. The performance of using strongly correlated classifier is significantly better than normal classifier [13], and this greatly improves the prediction power of model. Thus, boosting is a common machine learning algorithm for reducing variance in supervised learning [13]. It can also be combined with other algorithms and result a highly accurate prediction.

3.4 Regression Techniques

Regression is a statistical technique which used for predicting continuous-response values with one or more independent variables. It is a process for identify the relationship between a response input and output. The probability distribution is used to describe the variation of the output data (dependent variable) in regression analysis [9].

3.4.1 Linear Regression

Linear regression is a method for predicting continuous-response values in classification problems. This linear approach is used to model the relationship between the scale dependent variable Y and one or more independent variables expressed as X. For one independent variable, the process is called simple linear regression. The case of more than one independent variable is called multivariate linear regression [14]. In addition, the probability distribution is used to describe the variation of the output data (dependent variable) in regression analysis. Ridge and Lasso regression are two common extensions for linear regression. They are closely related, Ridge regression is proposed to solve multicollinearity, but it does not have ability to select predictors as it cannot shrink the parameters to zero. Lasso regression fixed the problem and able to do variable selection. Both are attempts to minimize residual sum of squares of predictors in a given model.

3.4.2 Time Series Techniques

In the case of time series techniques, the prediction of dependent variable is based on the past behavior rather than the independent variables. Function of time is used to set up the predicting relationship. This technique is mainly used to find out the pattern of past dataset if the behavior of data is hard to understand. As it only predicts the future results, not the reason of results, therefore, it is easier to implement time series techniques instead of regressing analysis [15].

3.5 Evaluation model & Model validation

The purpose of modeling is to infer the logical relationship to understand the current behavior and predict the future behavior of a system. The response variable is the predicted variable and the inputs are the predictor variables in a model. The evaluation model and model validation are presented in the following sections.

3.5.1 Root Mean Square Log Error (RMSLE)

Root Mean Square Error (RMLSE) is a common technique to measure the different between the predicted value and the actual observed values. The deviation of data samples is called prediction error when computed outliers in the dataset. Therefore, RMSLE is a measure of predictive power of model as it is used to aggregate the prediction errors for various time

[17]. A low RMSLE value means that the chosen predictor variables are well fit for the data, but it may lead overfitting of data and high complexity of model fitting. On the other hand, a high RMSLE value means a bad fit of data, but it results a simple model fitting.

3.5.2 K-fold Cross Validation

K-fold cross validation is a common method for model validation. Firstly, the dataset will be split into k equaled size of subsets. K-1 number of subsets are used for the training set and the remain subset is used for validation [18]. The validation of the dataset is done for k times until all the subsets have been used for the validation. At a result, the score functions are averaged and estimate the prediction error. Figure 5 illustrated the 4-fold cross-validation method. The grey part of the examples is the testing set for the validation process and the remain dataset represents the training set.

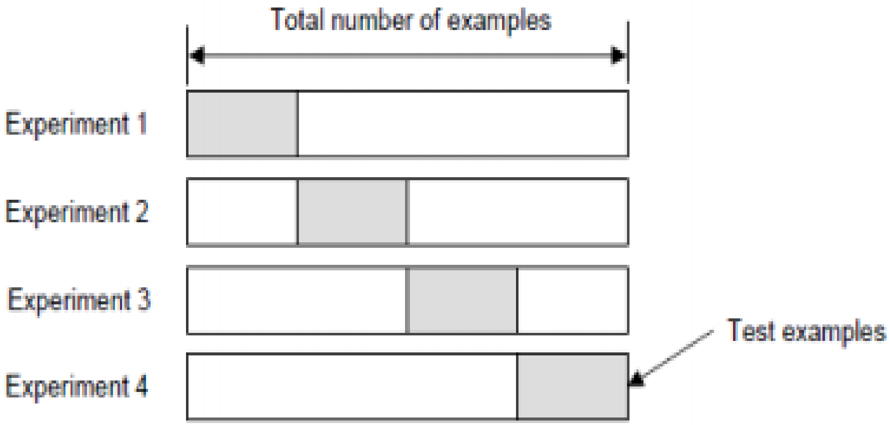


Figure 5 4-Fold Cross-Validation [18]

Chapter 4 Data Analysis

This section describes the dataset and the steps of data preprocessing in detail. Data preprocessing includes missing value processing and data transformation. The data analysis is based on the statistics and data distribution.

4.1 Introduction to dataset

In this task, the bike sharing usage data are collected from Capital Bikeshare program et al [8] to forecast bike rental demand in Washington DC. The dataset consists the hourly bike rental data with weather and date information from 2011 to 2012. The training set represented the first 19 days of each month from 2011 to 2012 and the testing set represents the rest of day of each month in the same 2 years period. Totally, there are 10,886 ride records in the training set and 6,493 ride records for the testing set. The training set consists of 12 variables including date time, season, holiday, working-day, weather, temp, atemp, humidity, windspeed, casual, registered and count.

	Training Set	Testing Set
Number of Entries	10,886	6,493
Number of Variables	12	9

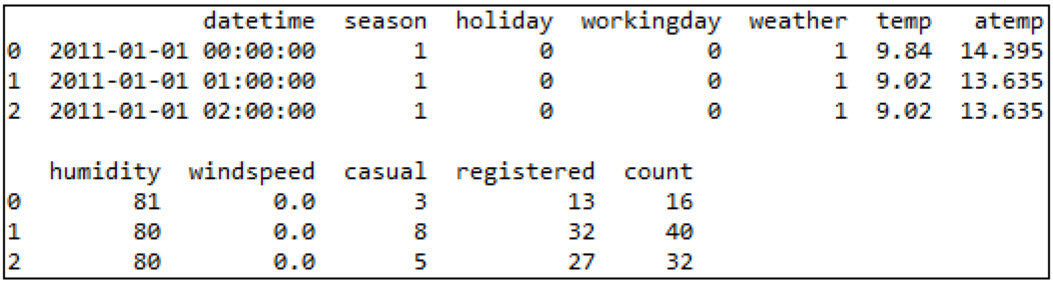
Table 1 Shape of the dataset

The detailed description of variables is shown in table 2. Besides the numerical variables, there are also categorical variables and time variable. Therefore, we need different treatments for different type of variables. Date time is a string of timestamp, 4 categorical variables are used to represent different weather conditions and season, holiday and working-day are binary indicator variable to represent whether a day was a holiday/working-day. In addition, casual and registered represented the number of bike rental by non-registered and registered users, and count is the sum of the non-registered and registered users. Count, registered and non-registered variables are not included in testing set, as the goal of prediction model is to predict the total number of bikes rented on hourly basis. Therefore, only the first 9 variables of training set are included in testing set. First three rows of training set are shown in figure 6.

Variable	**Datatype**	**Description**
Date time	Object	Rental Date and time information.
Season	Category	Four seasons: 1 = Spring, 2 = Summer, 3 = Fall, 4 = Winter
Holiday	Boolean	Whether the day was a holiday: 0 = Not holiday, 1 = Holiday
Working day	Boolean	Whether the day was a working day: 0 = weekends or holiday, 1 = weekdays
Weather	Category	Weather of bike rental day: 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
Temperature	Float	Temperature in Celsius
Apparent temperature	Float	Feels like temperature in Celsius

Humidity	Integer	Relative humidity percentage
Windspeed	Float	Wind speed
Casual	Integer	Number of bike rented by non-registered user
Registered	Integer	Number of bike rented by registered user
Count	Integer	The sum of registered users and non-registered

Table 2 Dataset variables and types



```
              datetime  season  holiday  workingday  weather  temp   atemp
0  2011-01-01 00:00:00       1        0           0        1  9.84  14.395
1  2011-01-01 01:00:00       1        0           0        1  9.02  13.635
2  2011-01-01 02:00:00       1        0           0        1  9.02  13.635

   humidity  windspeed  casual  registered  count
0        81        0.0       3          13     16
1        80        0.0       8          32     40
2        80        0.0       5          27     32
```

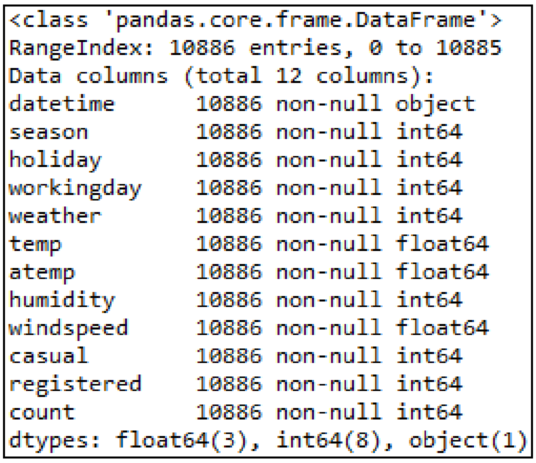
Figure 6 First 3 row of training set

4.2 Data Pre-processing & Visualization of dataset

Data pre-processing is an important step in machine learning. Generally, real world dataset often consists of missing values, outliers and errors. These dirty data will greatly affect the analysis of data and lead to misleading results [23]. Therefore, data pre-processing is foremost to ensure the quality of data. The noisy of data should be removed before making any analysis of the dataset to perform an accurate prediction. This section will describe the data pre-processing of the bike rental dataset and improve the generalization error.

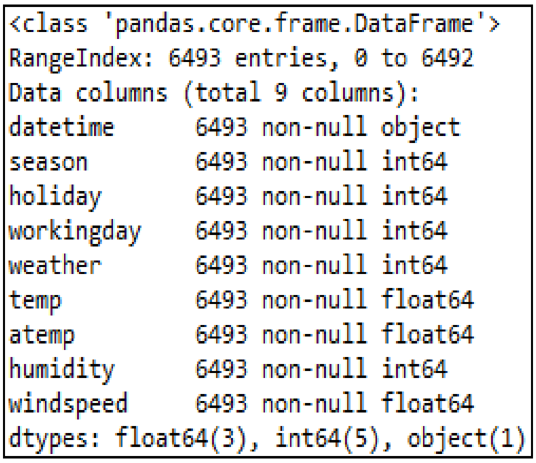
4.2.1 Missing Value Processing

Checking for missing value is always the first step of data pre-processing. When no value is stored in the data entry, missing data occurred. An incomplete or erroneous dataset will affect the sensitivity and specificity of classifier performance [23]. Also, an improper handling of missing values will end up an inaccurate inference about the dataset. Generally, deletion and replacement of data are two common techniques to deal with the missing values. If the number of missing data is less than 5% of the dataset, those data can be dropped from the dataset or replaced with median or mean value [23]. If there is a number of missing values, data imputation is necessary to replace them. Therefore, this section will explore the training and testing set to check for missing values, the DataFrame.info() function from pandas library is used to display the summary data. As shown in figure 7 and 8, both the training and testing do not contain any missing value, missing value processing is not needed in this task.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
datetime      10886 non-null object
season        10886 non-null int64
holiday       10886 non-null int64
workingday    10886 non-null int64
weather       10886 non-null int64
temp          10886 non-null float64
atemp         10886 non-null float64
humidity      10886 non-null int64
windspeed     10886 non-null float64
casual        10886 non-null int64
registered    10886 non-null int64
count         10886 non-null int64
dtypes: float64(3), int64(8), object(1)
```

Figure 7 Summary of Training Set



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6493 entries, 0 to 6492
Data columns (total 9 columns):
datetime      6493 non-null object
season        6493 non-null int64
holiday       6493 non-null int64
workingday    6493 non-null int64
weather       6493 non-null int64
temp          6493 non-null float64
atemp         6493 non-null float64
humidity      6493 non-null int64
windspeed     6493 non-null float64
dtypes: float64(3), int64(5), object(1)
```

Figure 8 Summary of Testing Set

4.2.2 Detecting the outliers

An outlier is an observation that is out of range or erroneous from the normal distribution of data. As like as the missing values in dataset, including outliers in training process will mislead the model and poorer the prediction results. However, sometimes the outliers may contain valuable information about the dataset [23], thus, we should try to understand them before the eliminate these values from the data. The box plot is used as a graphical technique to visualize the data and examine the overall shape of the graphed data. The box plot uses median, the lower and upper quartiles to represent the range of data. By graphically display the distribution of data, any point that is far away from the mass of the data will be identified as outlier. Figure 4.8 shows the box plot of the count variable which is the total number of bike rental for both registered and non-registered customers. The range of data is from 1 to 977 and the median is 145. As shown in the figure 9, there are many data points beyond the Outer Quartile Limit which skews the distribution to the right. Those data points are considered as outliers and will be removed from the dataset. The results from table 3 indicate that 147 observations are removed from the dataset.

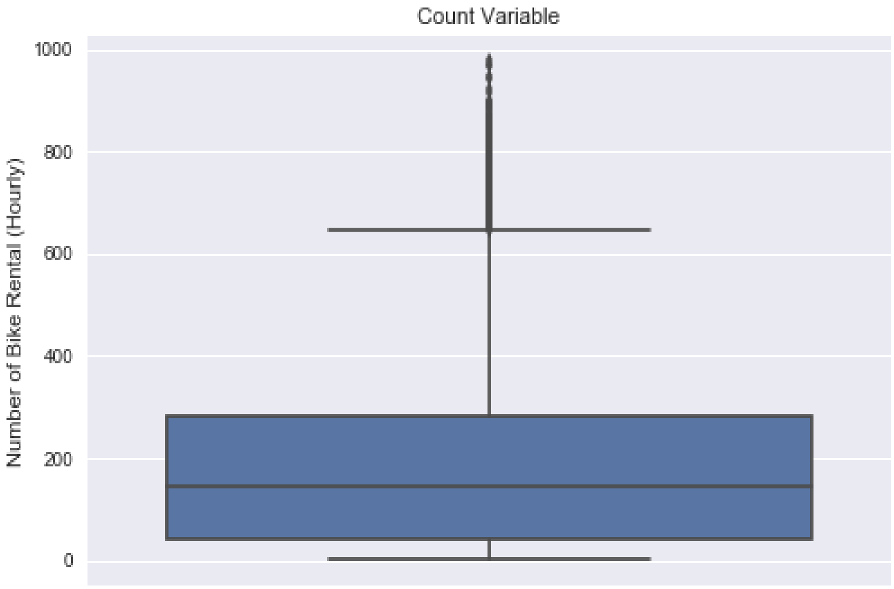


Figure 9 Data Distribution of Count Variable

Shape before remove outliers	(10,886, 12)
Shape after remove outliers	(10,739, 12)

Table 3 Shape of the Training Set

4.3 Pre-processing and Visualization of Categorical Type Variables

The categorical variables including datetime, season, holiday, working day, weather are visualized and then determine the pre-processing method based on the data distribution.

4.3.1 Decomposing the date time into hour, weekday and month

Datetime	Month	Weekday	Hour
2011/1/1 12:00:00 AM	January	Saturday	0
2011/2/2 01:00:00 AM	February	Tuesday	1
2011/3/3 02:00:00 PM	March	Wednesday	14

Table 4 Example of Date Time Decomposition

It is hard to interpret the knowledge of the date time variable as it combined the date and time of ride as one feature [24]. Therefore, it was decomposed into three sub-variables which are month, weekday and hour. As the dataset only consider the bike usage from 2011 to 2012, year variable is ignored in this task. Generally, month, weekday and hour are more correlated to the bike rental demand on real life and easier to corporate and integrate into the model. The column of date time was dropped from the dataset after extracted useful features. The time variable only considers the hour of bike rental as the column of date time was already rounded up to nearest hour.

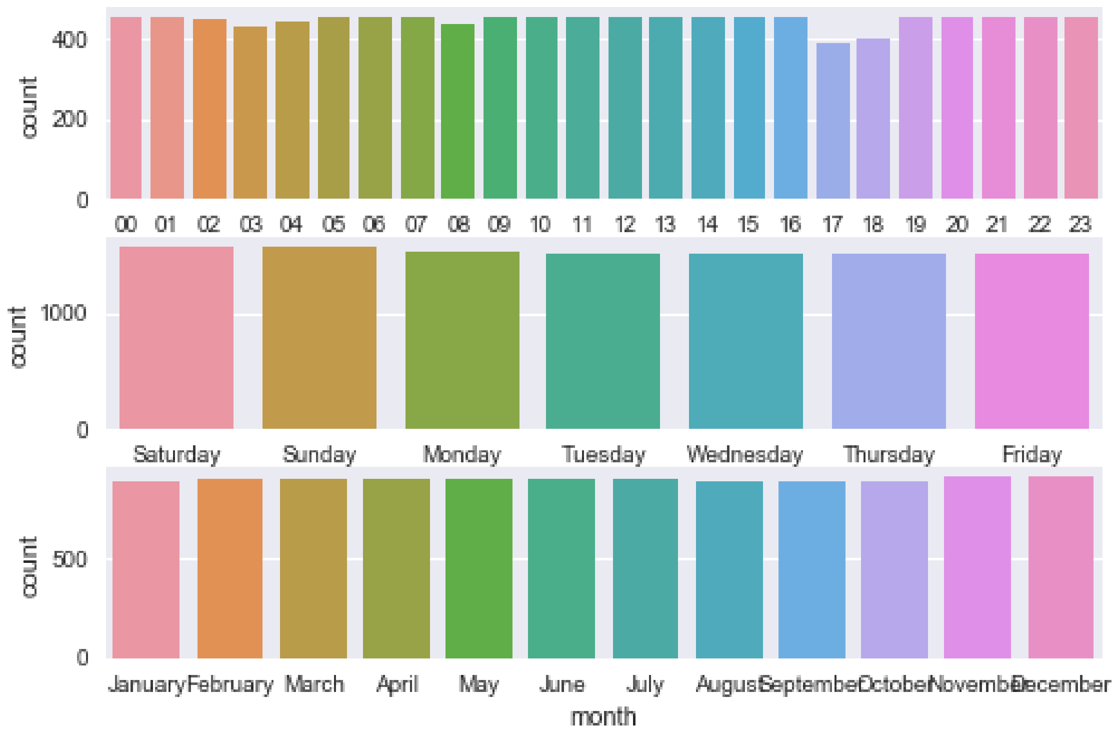


Figure 10 Count of Hour, Weekday and Month Variables

As shown in figure 10, the data distribution in these features is quite balanced, there is no certain period with too much or too less observations. Therefore, we do not need to use data oversampling or undersampling to balance the distribution of data.

4.3.1.1 Hour

From figure 11, we can see the number of bike rental is highly related to the hour of the day as expected. The mean values of bike rental are higher on the rush hour which is at 7am to 8am and 5pm to 6pm. In these part of the day, many students and office users are commuting to school or work, this results traffic congestion on roads and crowding on public transport. Therefore, more people choose bike as a mode of transportation to avoid the heavy traffic and reduce the commute time.

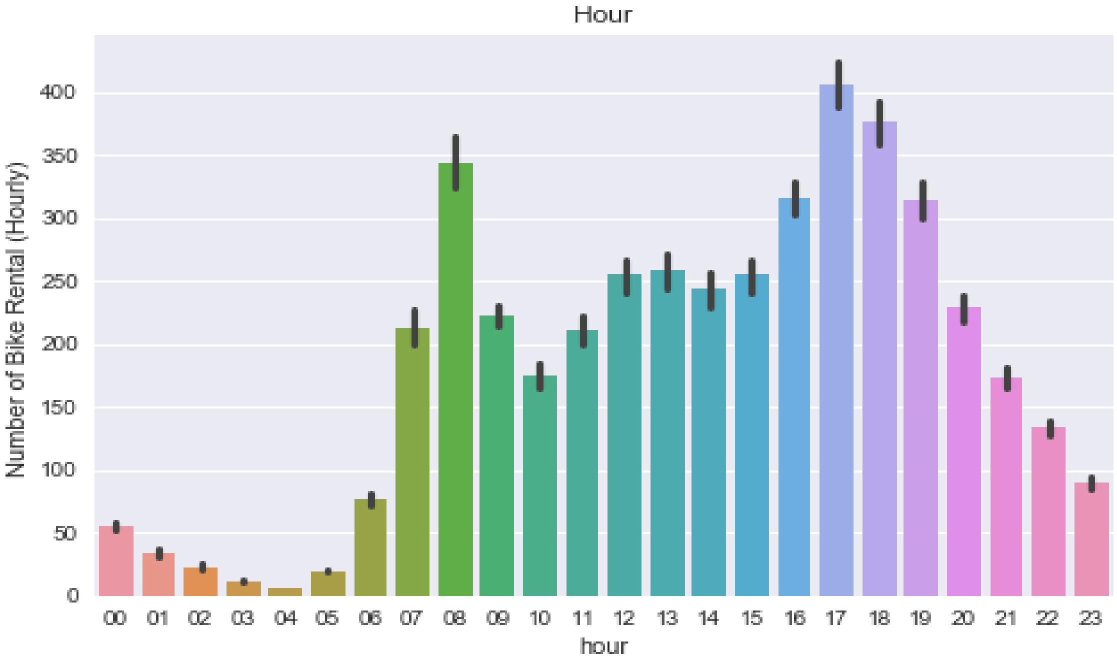


Figure 11 Number of Bike Rental by Hour

4.3.1.2 Weekday

The result from figure 12 shows that the number of bike rental is similar on different weekday. Around 175 bikes are rented in hourly average each day. However, the mean value of bike rental on Saturday is relatively higher. This is quite surprise and a bit contrary from experience. The demand of bike rental become higher with the decrease of commuters. Overall, the number of bike rental is similar between weekday and weekend, this may because of the improvement of public environmental awareness. More people are preferring to take a bike for either relaxing or commuting.

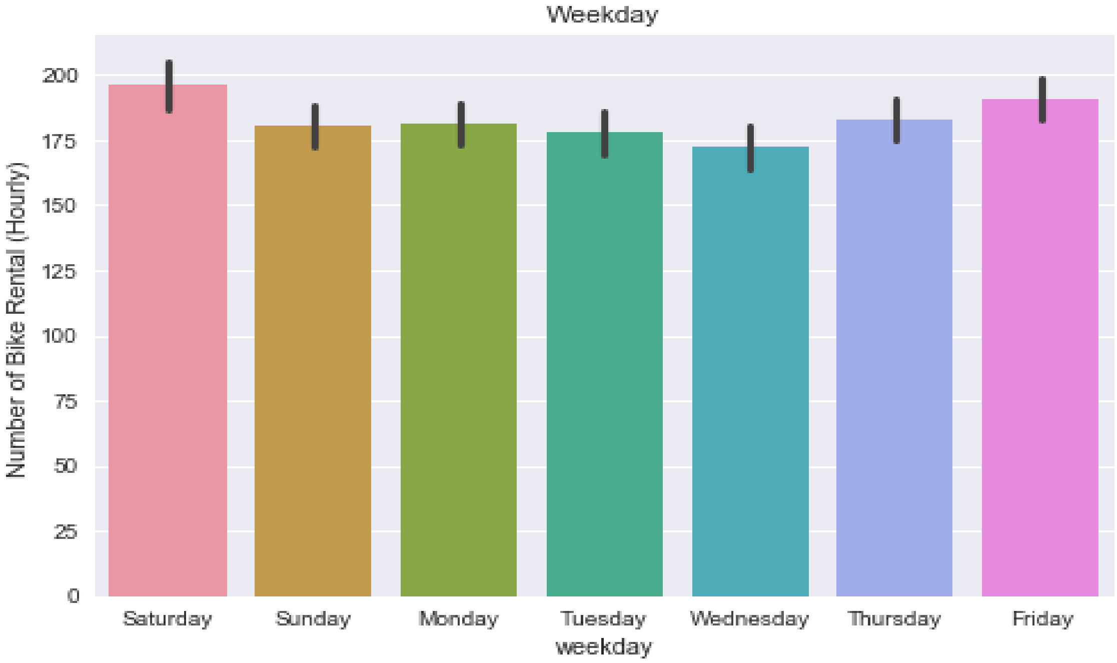


Figure 12 Number of Bike Rental by Weekday

4.3.1.3 Month

As shown in figure 13, most bikes are rented from June to August, the median values show that the hourly averages of bike rental are around 220 bikes in these months, and January and February have the least number of bike rental, around 100 bikes are rented hourly in average. This may correlate to the weather condition and the temperature of the month, the summer months in Washington D.C start from May to August, summer is traditionally associated with warm and good weather condition. Thus, more people are preferring to go out and take a ride either on commuting or cycling. On the other hand, winter may not suitable for outdoor activities with a cold weather and snow. Thus, less people are willing to take a ride in January and February.

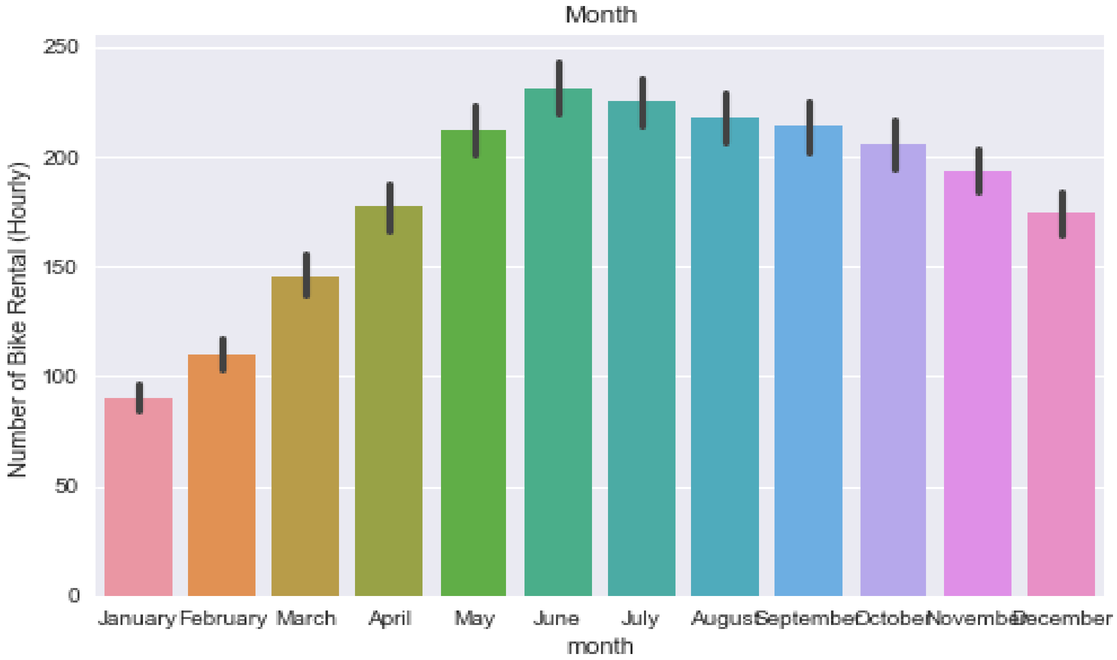


Figure 13 Number of Bike Rental by Month

From the above analysis, it is visible that hour and month are important features in this task. The weekday variable may need to explore further to find out the correlation between the demand of bike rental.

4.3.2 Replacing the season variable with month variable

As each categorical variables of season represents three individual months of a year, therefore using season variable may not capture the difference between individual months of the same season. The column of season is removed and replaced with the month variable.

4.3.3 Weather

As shown in figure 14, the number of bike rental is correlated to the weather condition. Most bike are rented when the weather is clear or partly cloudy, the good weather prompt the level of outdoor activities and people are more willing to take a bike for transportation. The number of bike rental of cloudy and mist day is slightly less than clear weather, bike is still an acceptable transportation option for the public. The number of bike rental drop significantly when it is a light raining day, the inter quartile range and median are half when compared to good weather or a cloudy day. On a raining day, the floor become slippery and it is risky to ride on a wet floor. Therefore, some of people have chosen other transportation instead of bike. However, many of outliers are observed in light rain. As expected, people are rarely riding a bike on a heavy raining day. For over two years from 2011 to 2012, only one record in dataset shows that bikes are rented on such bad weather condition.

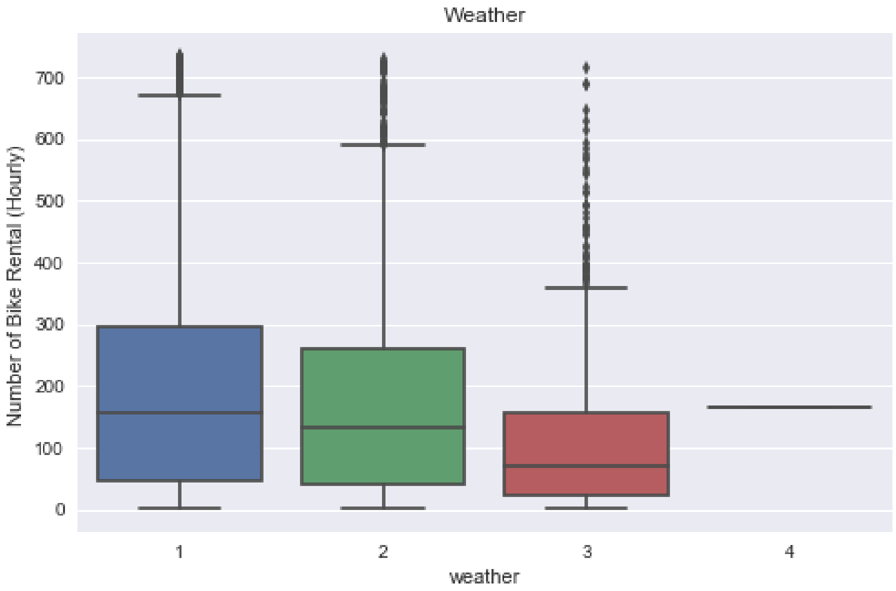


Figure 14 Number of Bike Rental by Weather

4.3.4 Removing Holiday Variable

As mentioned before, the result from weekday variable show that there is no big different between working day and weekend. This is a bit contrary to expectation. As the dataset also includes the working day and holiday variables, we will then analysis these two variables and compare the data distribution.

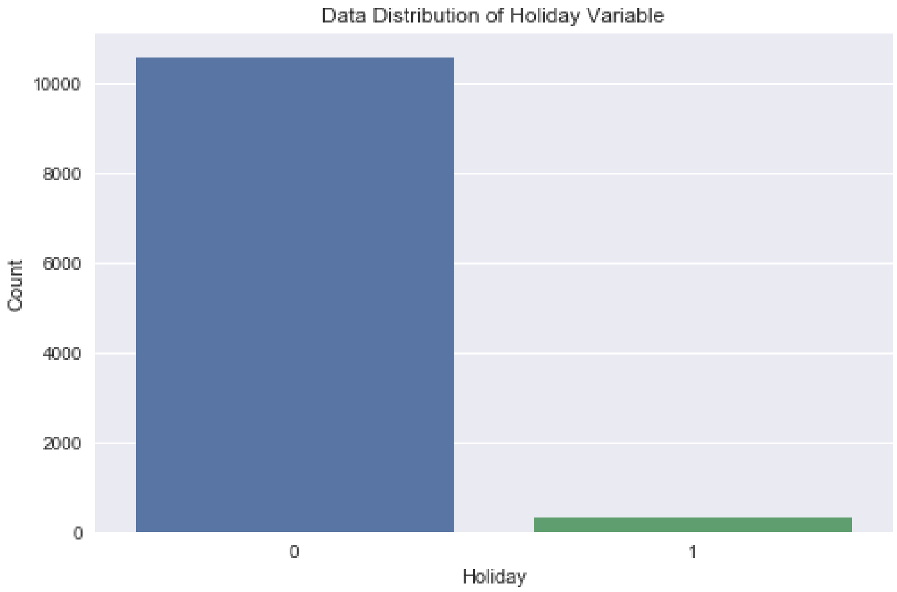


Figure 15 Count of Holiday Variable

As shown in figure 15, the data distribution of holiday variable is almost one-sided. 10,575 out of 10,886 rides are taken on a working day, only 311 bikes are rented on a holiday. A small data will affect the predictability of classifiers as it is hard to generalize universally, and overfitting may occur. As there are too few data of public holidays, it is difficult to notice the relationship between bike rental and holiday. Therefore, it seems that holiday variable is not useful in this task, the column of holiday is removed from the dataset.

4.3.5 Working Day

In addition, the dataset has a working day feature, if the day is a working day, it implies that the day is not a holiday. As shown in figure 16, the number of working day data is double more than the data of non-working day in training set. This gives a better representation of the bike rental difference on working day and weekend/holiday.

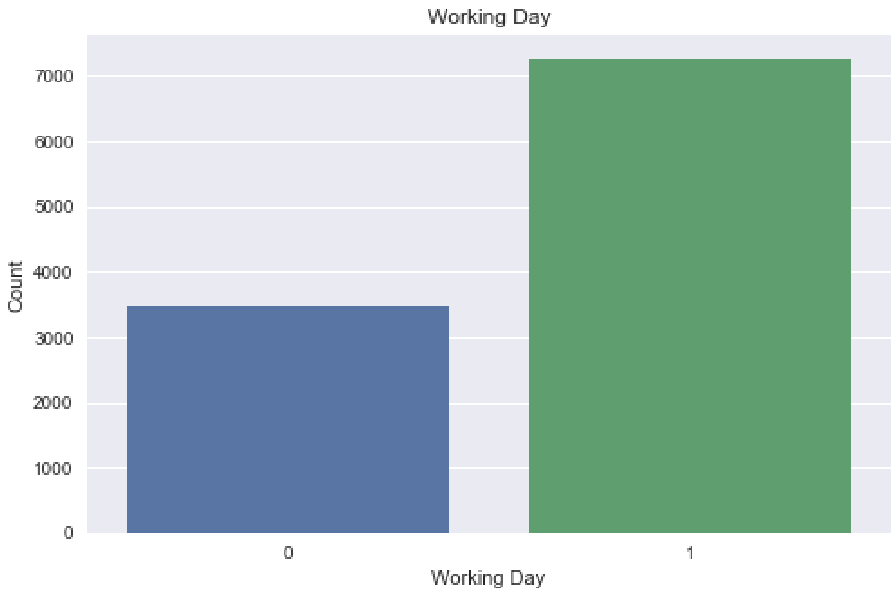


Figure 16 Count of Working Day Variable

As shown in figure 17, around 180 bikes are rented for both working day or non-working day. The result is quite similar, so working day variable may need to integrate with other features to find out the relationship between working day and bike rental demand.

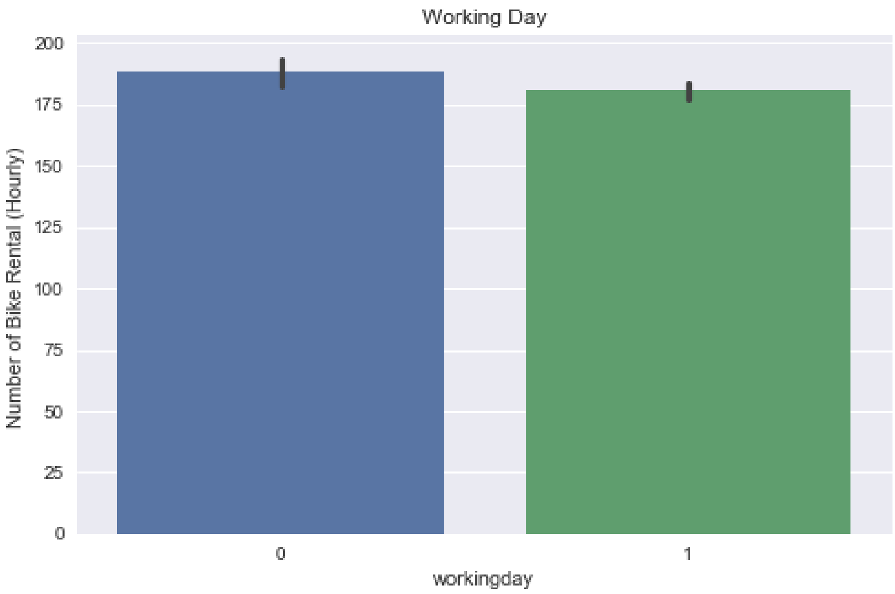
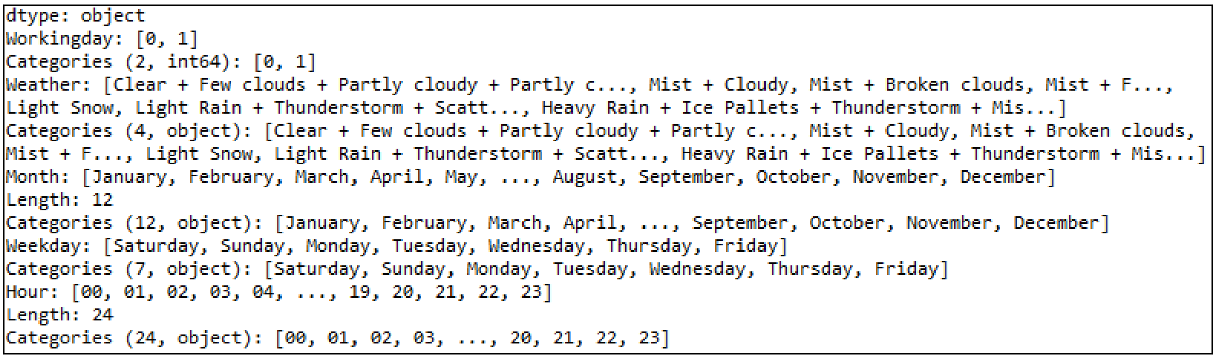


Figure 17 Number of Bike Rental by Working Day

4.4 Converting the data type of categorical variables

Categorical variables always have valuable information of dataset [23], including those knowledge in modeling can improve the prediction accuracy and the understanding of dataset. In addition, there are many machine learning libraries available to handle categorical features. In this task, astype() function from Numpy library is used to convert the columns of season, holiday, working day, weather, month, weekday and hour from integer to categorical data type. This can convenience the exploratory data analysis and visualization of the data. After the detection of missing values, unique() function is used to check for erroneous data of categorical variables. As shown in figure 18, all the columns of categorical variables do not have any mistaken value.



```
dtype: object
Workingday: [0, 1]
Categories (2, int64): [0, 1]
Weather: [Clear + Few clouds + Partly cloudy + Partly c..., Mist + Cloudy, Mist + Broken clouds, Mist + F...,
Light Snow, Light Rain + Thunderstorm + Scatt..., Heavy Rain + Ice Pallets + Thunderstorm + Mis...]
Categories (4, object): [Clear + Few clouds + Partly cloudy + Partly c..., Mist + Cloudy, Mist + Broken clouds,
Mist + F..., Light Snow, Light Rain + Thunderstorm + Scatt..., Heavy Rain + Ice Pallets + Thunderstorm + Mis...]
Month: [January, February, March, April, May, ..., August, September, October, November, December]
Length: 12
Categories (12, object): [January, February, March, April, ..., September, October, November, December]
Weekday: [Saturday, Sunday, Monday, Tuesday, Wednesday, Thursday, Friday]
Categories (7, object): [Saturday, Sunday, Monday, Tuesday, Wednesday, Thursday, Friday]
Hour: [00, 01, 02, 03, 04, ..., 19, 20, 21, 22, 23]
Length: 24
Categories (24, object): [00, 01, 02, 03, ..., 20, 21, 22, 23]
```

Figure 18 Class of Categorical Variables

4.5 Pre-processing & Visualization of Numerical Variables

In this part, the numerical variables including atemp, temp, humidity, windspeed are visualized and then determine the pre-processing method based on the data distribution.

4.5.1 temp variable

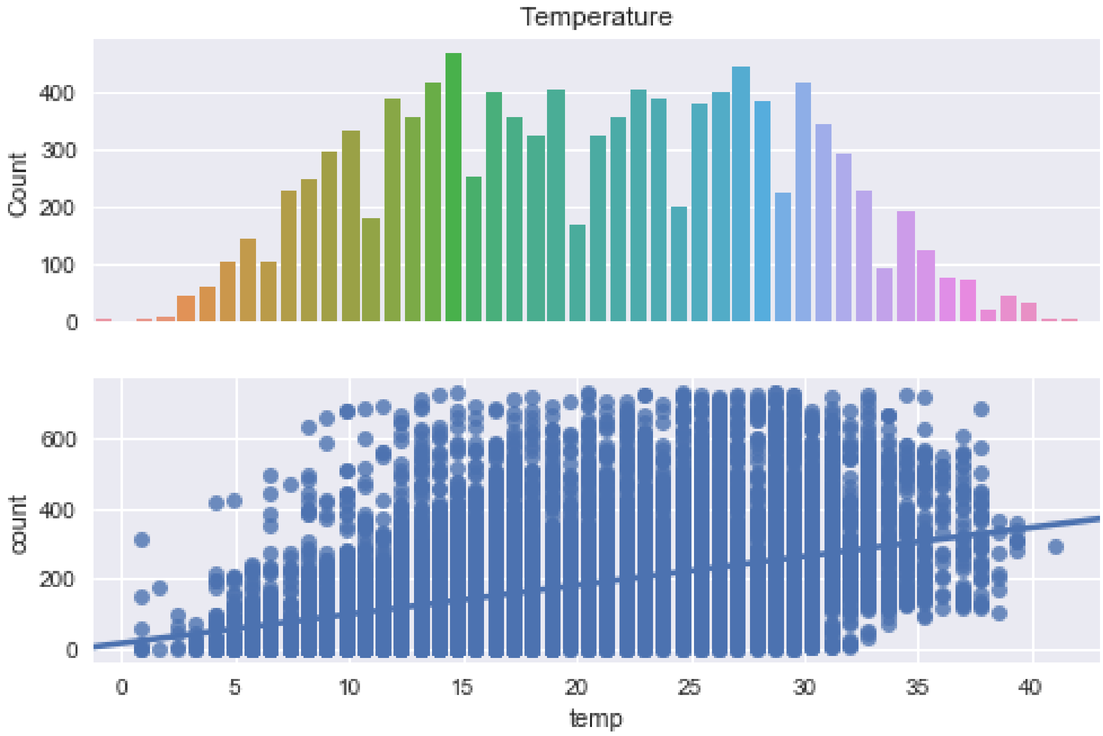


Figure 19 Number of Bike Rental by temp

As shown in figure 19, the distribution of temp variable is close to normal distribution, most of the data are distributed in the middle range of temperature, only a few observations with low or high temperature. The result from the regression plot shows that the bike rental demand is directly proportional to temperature. With the increase in temperature, the number of bike rental also increase. However, the rising trend is not match with the normal expectation, but there is a positive correlation between bike rental demand and the temperature, so that the temperature gradually increases with the number of bicycle rental increased within a certain range. As shown in figure 19, the optimal temperature for biking is around 15-30 Celsius, only a few observations occurred when the temperature is lower than 5 or higher than 35 Celsius. These indicate that people prefer go cycling under a comfort temperature.

4.5.2 Removing "atemp" variable

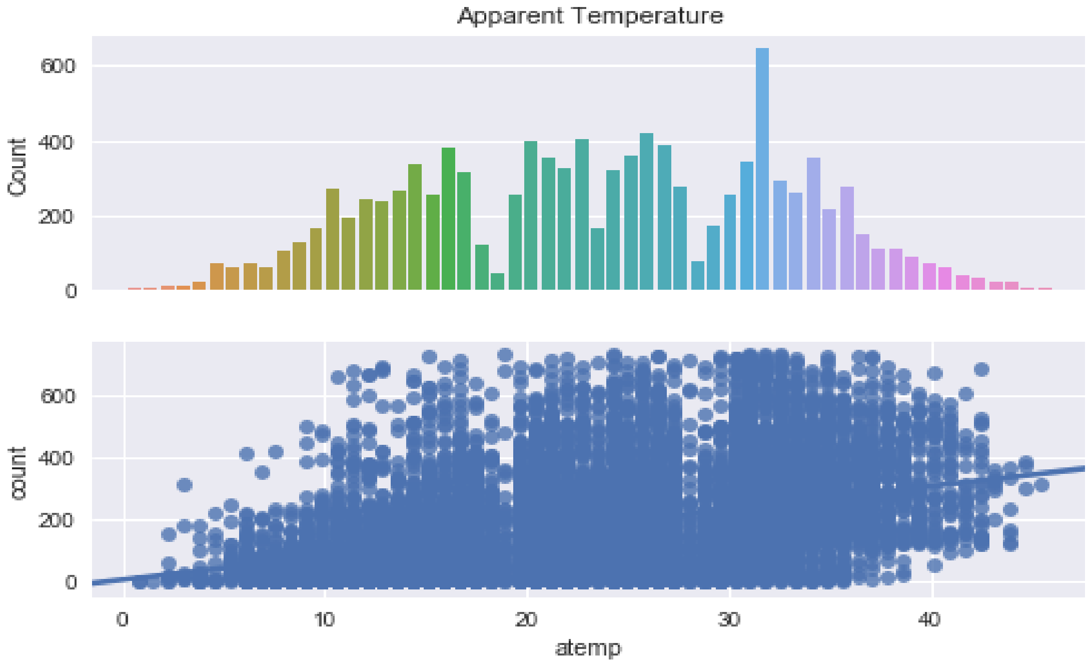


Figure 20 Number of Bike Rental by atemp

The apparent temperature has visualized in figure 20 to study the correlation with bike sharing demand. The result is similar to the temp variable, but these are fewer observations when the apparent temperature is around 20 and 30 Celsius and the number of bike rental increase significantly in 33 Celsius. The regression plot shows that atemp have a directly proportional relationship to the bike rental demand.

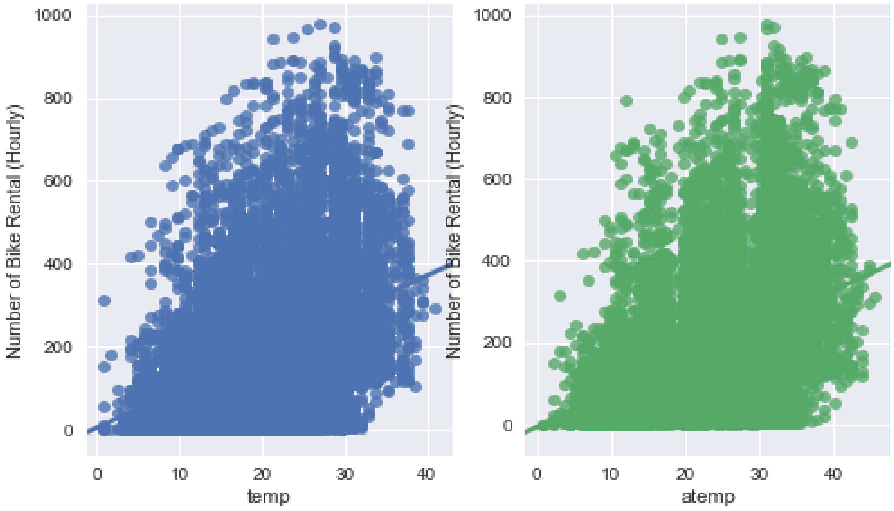


Figure 21 Regression plot of temp and atemp

Generally, the higher temperature of the day, the higher feels like temperature. Therefore, temp and atemp variables should have a strong correlation. Regression plot in seaborn library is a good method to display the relationship between two variables. Figure 21 illustrates the distribution and data range of temp and atemp variables, the results is similar as expected. Thus, one of them should be removed to avoid collinearity issues. Temp should give a more objective information of the actual temperature than atemp. The temperature data is more relevant reason for people to make decision of renting a bike. Also, the data in temp variable is more uniformly distributed than atemp. Then, the column of atemp is removed from the dataset.

4.5.2 Windspeed

As shown in figure 22, the number of bike rental decrease with a higher windspeed rate. The optimal windspeed for biking is between 0-15. However, the data of windspeed is quite sparse, further evaluation is needed to identify the correlation with the demand of bike rental. Also, outliers are observed when the windspeed is higher than 50, those data points are needed to remove in training process.

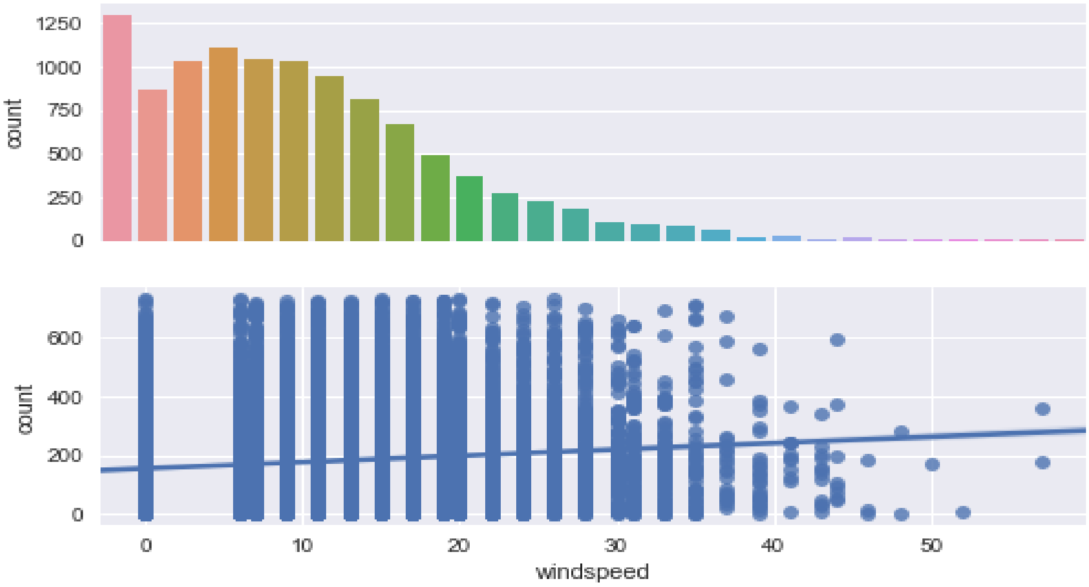


Figure 22 Number of Bike Rental by Windspeed

4.5.3 Humidity

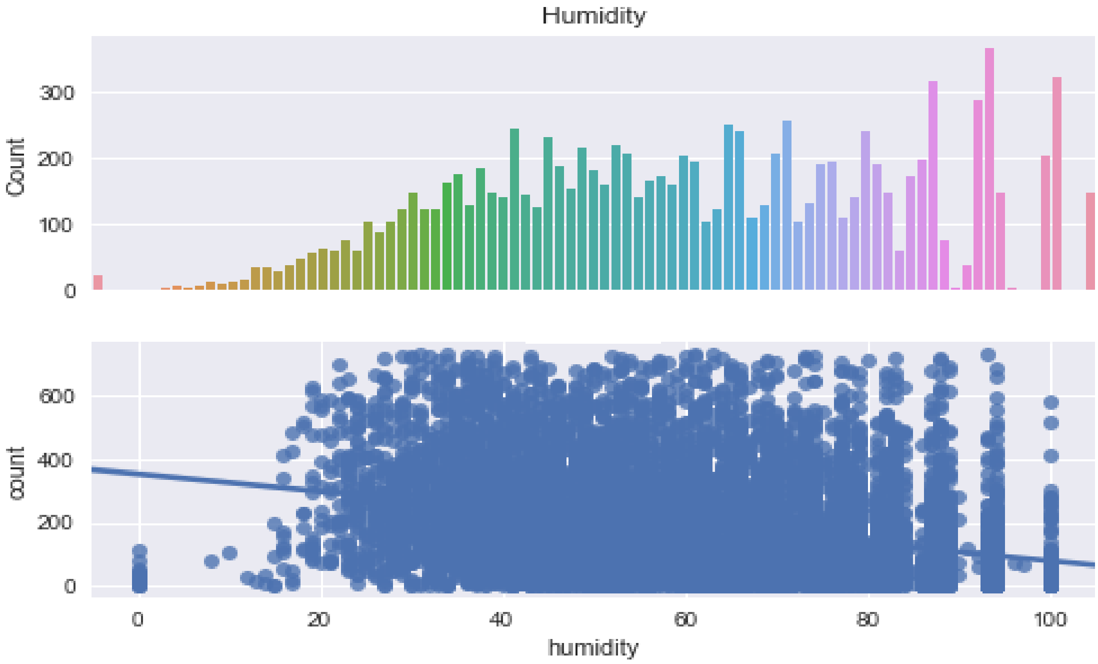


Figure 23 Number of Bike Rental by Humidity

The bike rental demand is relatively discrete on the distribution of humidity, as many peak values are shown in figure 23. Overall, a low or high humidity will decrease the demand of bike rental. People tends to go cycling with a human comfort humidity which is between 20% to 80%. The regression plot shows a negative proportional between the number of bike rental and humidity which implied a correlation exists between these two variables.

4.6 Correlation Analysis

To understand the correlation between the dependent variable and numerical variables, correlation matrix in Seaborn library is a good technique to investigate the dependence between count, temp, atemp, humidity and windspeed. The correlation coefficients between each variable and the others is shown in figure 24. The following conclusions are drawn based on the correlation matrix results.

Temp feature has a positive correlation with count which is 0.39. Even though the value is not that significant, it still represents that count is slightly depend on temp feature. On the other hand, humidity gives a negative correlation with count which -0.32, as like as atemp, count has got a little dependency on humidity. The correlation of windspeed with count is 0.11, the coefficient is much smaller when compare to others and implies a poor correlation with count. Therefore, windspeed is not a useful feature in this task and will not consider in training process. As the casual and registered are the leakage variables in nature, they are not taken into account when building the model.

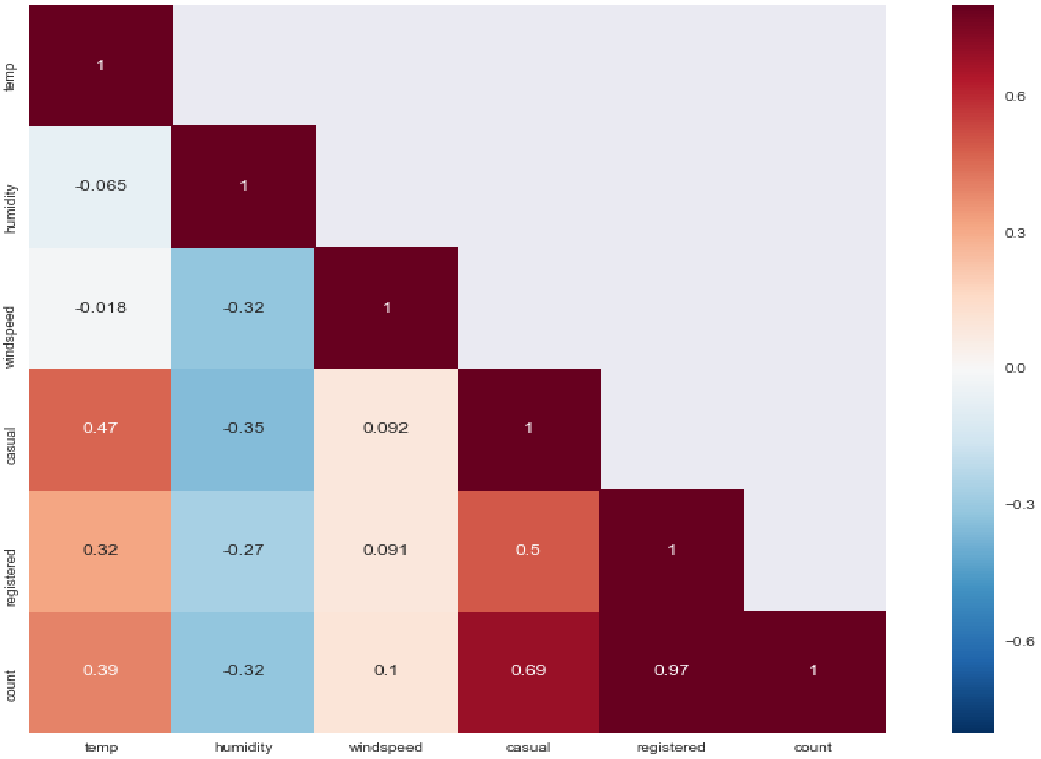


Figure 24 Correlation Matrix

4.7 Data Distribution of Independent Variable

As shown in figure 25, the data distribution of count variable is skewed towards right. The dependent variable is desirable to have normal distribution for the modeling techniques. Thus, log transformation is used to deal with the skewed data to make the relationship clearer. Log transformation is a process to reduce the variability of data and make the data more normally in distributed [25]. The patterns in the data will be more interpretable and visible. Distplot() function in Seaborn library is used to visualize the distribution of count variable and the log transformation result. The log transformation is taken on count variable after removing outliers. The data looks better after the log transformation but still not fitting the ideal normal distribution.

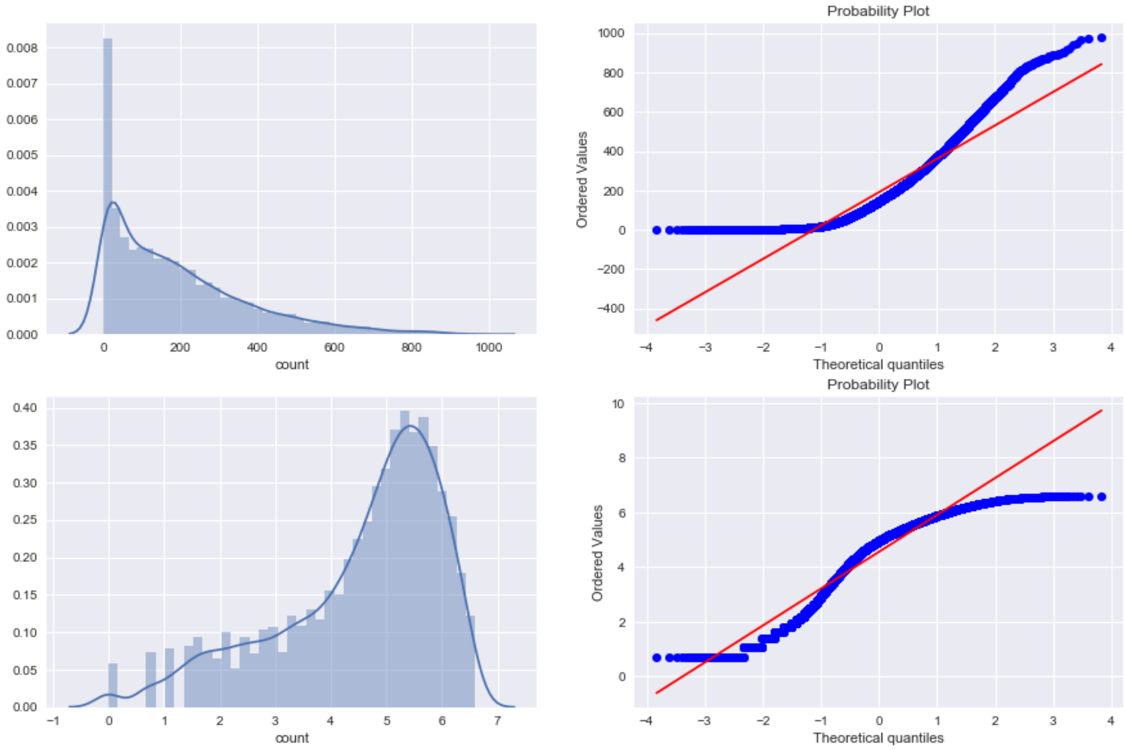


Figure 25 Distribution of Independent Variable

4.8 Data Integration

Data integration is to combine the information from various data and discover useful pattern and knowledge from the dataset. In this task, month, weekday and user type are combined with the hour of the day to examine the bike usage pattern in greater extent. In addition, some inferences are made from the graphs given below.

4.8.1 Hourly Bike Rental by Months

As shown in Figure 4.6.1, the pattern of bike rental by hour is consistent across months. Th demand of bike rental is higher around 7am to 8am and 5pm to 6pm. As mentioned before, school and office commuter have chosen bike as a mode of transportation to avoid the traffic congestion on roads and crowding on public transport. Also, January and February have a relatively lower demand for bike due to the temperature and weather condition.

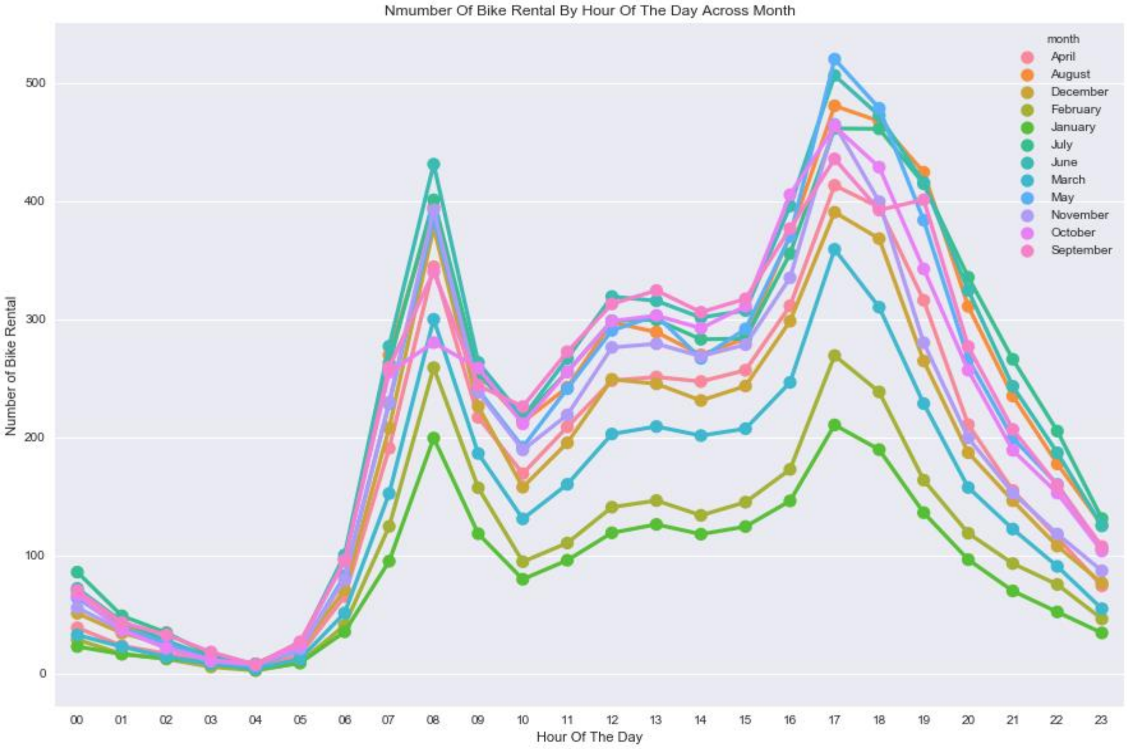


Figure 26 Hourly Bike Rental by Months

4.8.2 Hourly Bike Rental by Weekdays

Figure 27 shows the hourly bike rental across weekdays. The bike usage pattern is visible and understandable. The peak hour for bike rental from Monday to Friday is around 7am to 8am and 5pm to 6pm because of the school and office commuters, we can infer that most people rent bike on working day for commuting. However, the peak hour of Saturday and Sunday is around 12pm to 4pm. The possible reason for this is that people do not need to work or school on weekend, they tend to rent bike for entertaining instead of commuting. Therefore, we can deduce that most people rent a bike on weekend for relaxing cycling.

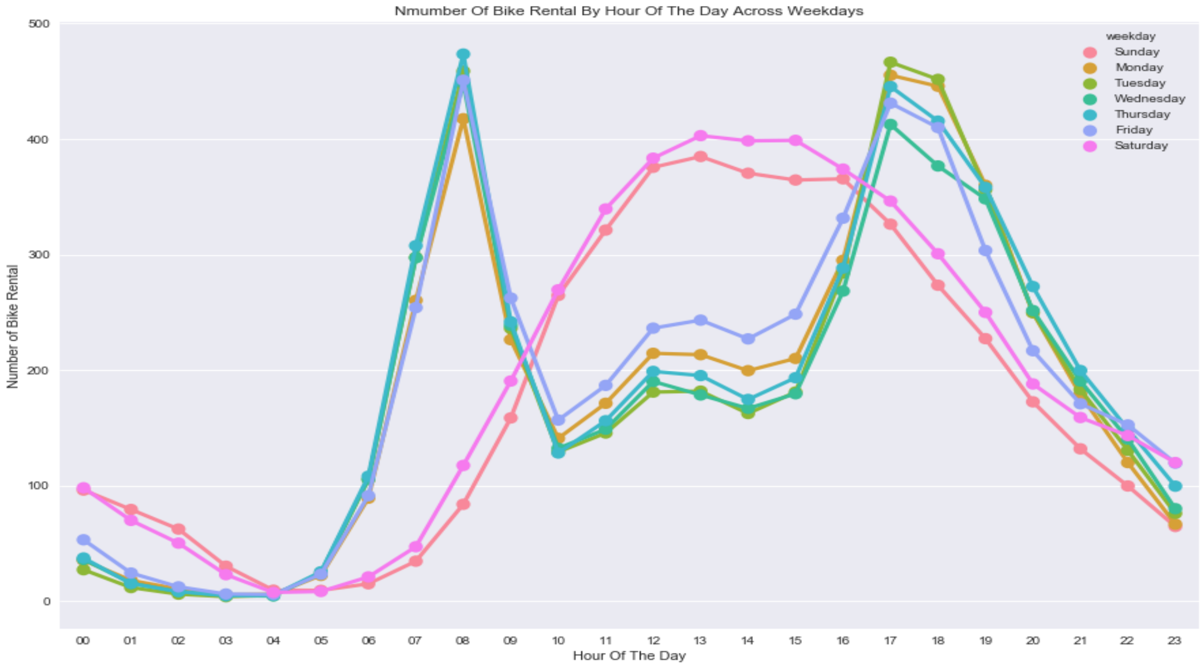


Figure 27 Hourly Bike Rental by Weekdays

4.8.3 Hourly Bike Rental by User Types

The hourly bike rental demand of casual and registered users is visualized in figure 28. It is obvious that registered users contribute most on the peak hour which is around 7am to 8am and 5pm to 6pm. The pattern is as like as the bike usage in weekdays which is shown in figure 27. This means that most of the commuters are registered for the bike sharing system. Usually, the members of bike sharing services can enjoy a lower bike rental rate and various benefits. Therefore, for those who need a bike regularly for commuting, they are recommended to apply for a membership. On the other hand, the pattern of bike rental by casual user is similar to bike usage in weekend. The peak hour of bike rental is around 12pm to 5pm. Thus, it can be deduced that most of the casual users rent bike for entertaining purpose. As they are not riding frequently, they may not want to apply for a membership.

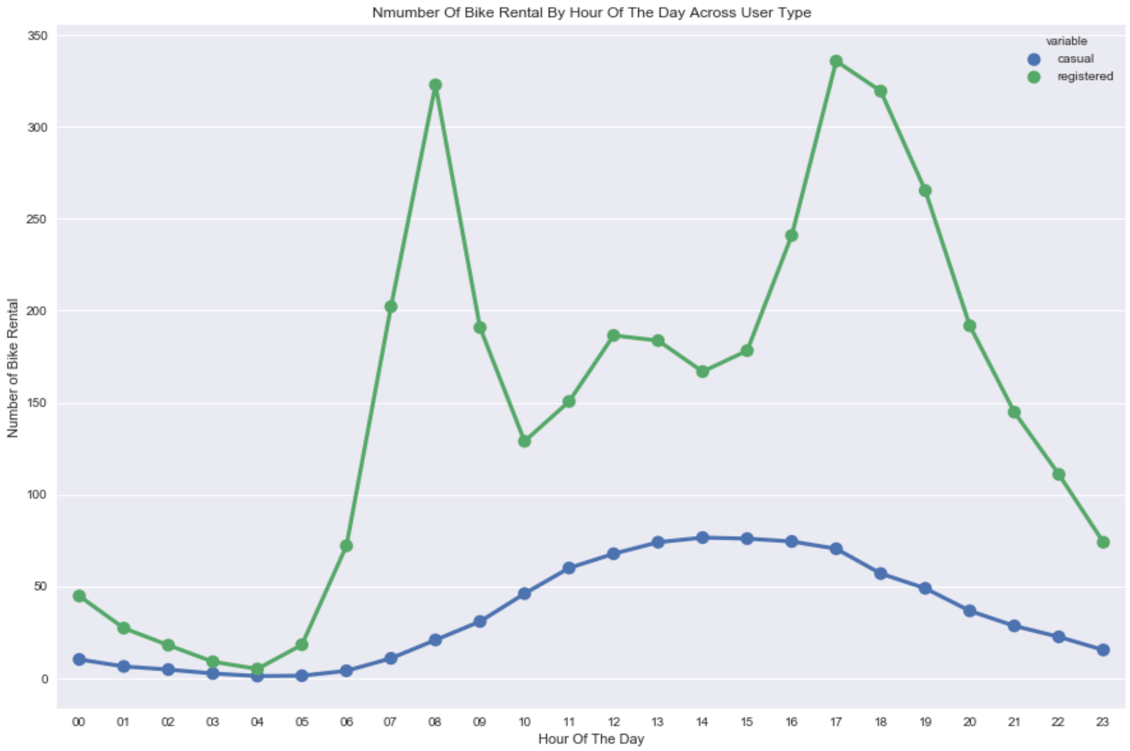


Figure 28 Hourly Bike Rental by User types

Chapter 5　Methodology and Results

This section analysis the data obtained in the previous chapter, after the data pre-processing, the dataset become more complete and consistent without missing values and noisy.

5.1 Feature Engineering

Feature engineering occupies an important position in machine learning, practically speaking, a good feature engineering is a key of success in machine learning. Throughout the Kaggle, KDD and other different competitions which related to machine learning, some of the champions did not use some advanced algorithms but did a good work in feature engineering, then use some common algorithms, such as linear regression, and result an excellent performance. Unfortunately, not many academic materials have mention the feature engineering but the feature selection. Basically, feature engineering is a process of transforming raw data into useful features [26]. These new features can describe the dataset better and used to build a model with a decent performance. From the view of mathematical, feature engineering is to design the input variable manually.

In this task, the feature engineering is relatively simple, the analysis of variables is given below. The original features of the project are described as follows:

	Features	**Description**
1	Datetime	Timestamp of bike rental: in year/month/day format
2	Season	Season of bike rental: 1: Spring 2: Summer 3: Fall 4: Winter
3	Holiday	Whether the day was a holiday 0: Not holiday 1: Holiday

4	Working day	Whether the day was a working day 0: Not working day 1: Working day
5	Weather	Weather condition: 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
6	Temp	Temperature on Celsius
7	Atemp	Apparent temperature (Feel like temperature)
8	Humidity	Relative humidity percentage
9	Windspeed	Wind speed
10	Casual	Number of bike rented by non-registered user
11	Registered	Number of bike rented by registered user
12	Count	The sum of registered users and non-registered

Table 5 Dataset variables and types

As mentioned in chapter 4 Data Analysis, Datetime, Season, Holiday, Atemp, Windspeed features are removed from the training set, and the Casual, Registered and Count are not features in this task. The remain features are classified as follow.

Features	**Data type**	**Coding**
Datetime	Discrete value	-
Working day	Category	One-hot encoding
Weather	Category	One-hot encoding
Temp	Continuous	One-hot encoding
Humidity	Continuous	One-hot encoding

Table 6 Variables and data types of new training set

The general method of feature processing is to transform the category type into one-hot coding. One-Hot coding is mainly used to register the status, each state has its own independent bits, and only one state is valid at any time. In the actual application of machine learning, sometimes the variables are not continuous value, there may be categorical variables. Such as gender, it can be divided into "male" and "female". For such categorical features, we usually need to encode it, as in the following example:

Consider the following three characteristic attributes:

Gender: ["male", "female"]

Region: ["Europe", "US", "Asia"]

Browser: ["Firefox", "Chrome", "Safari", "Internet Explorer"]

The gender attribute is two-dimensional, the area is three-dimensional, and the browser is four-dimensional, so that we can use One-Hot coding on the sample "[" "male" US "," Internet Explorer "]", "male" corresponds to [1,0], the "US" corresponds to [0,1,0], "Internet Explorer" corresponds to [0,0,0 ,1]. The result of the complete encoding is: [1,0,0,1,0,0,0,0,1]. Therefore, if we have a number of dimensional, the data will become very sparse.

In this example, we will encode the hour, weekday and month variable with another encoding methods. These multi-dimensional features are treated as a discrete sequence.

Features	Type	Coding
Hour	Category	Discrete sequence
Weekday	Category	Discrete sequence
Month	Category	Discrete sequence

Table 7 Time dummy variable coding

5.2 Model building

Linear models and tree based models are common models for regression task. In this task, linear regression is used as a linear model and optimized by using Ridge and Lasso regression. These three regression methods are mostly applicable to the linear distribution. In this task, the data does not have a linear distribution but more likely a normal distribution. So, the prediction result of linear models will have a greater error than tree based models. The tree base model including gradient boost decision tree (GBDT), random forest and XGBoost.

5.3 Baseline

In this task, we selected multiple linear regressions as the baseline of the prediction. Univariate linear regression only be able to explain the change of an independent variable with the dependent variable. In the real-life problem, the change of the dependent variable is often influenced by several important factors. Therefore, it is necessary to use two or more features as independent variables to explain the changes of the dependent variable. This also known as multiple regression. When multiple independent variables are linearly related to the dependent variable, the regression analysis is a multiple regression.

The simple statement is as follows:

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

5.4 Model Evaluation

In this task, Root Mean Log Square Error (RMLSE) is used to evaluate the result of model. It is used to measure the difference between the predicted value and actual value. RMSLE is suitable in this task because it can compare the accuracy between different prediction models. It can evaluate the measure of fit, the smaller value of RMSLE, the more accurate of the prediction model.

The formula is written as:

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

Where p is the predicted value, a is the actual value and n is the total number of samples. The value of RMSLE is related to under perdition more than over prediction so it is appropriate method to evaluate the prediction. The results can make sure that the demand for bike rental is always greater than the number of bike provided which is valuable for bike sharing company.

5.5 Result of Linear Models

In this task, we selected Linear Regression, Ridge and Lasso as linear models to build up the prediction model. 20% data of the training set is used as validation partc The results of linear models are shown in next section.

In this task, Grid Search from scikit-learn library is used to tune the alpha values in Ridge and Lasso. Alpha represents the regularization strength, a higher alpha value, stronger regularization. It is used to improves the performance of estimators and reduce the variance in the problem.

5.5.1 Ridge Regression

Mean	Standard Deviation	Alpha value
-1.02396	0.04924	0.1
-1.02397	0.04925	1
-1.02397	0.04925	2
-1.02397	0.04925	3
-1.02397	0.04925	4
-1.02399	0.04926	10
-1.02405	0.04930	30
-1.02443	0.04944	100
RMSLE: 1.0158500003240123		

Table 8 Alpha Tuning for Ridge

The selected parameter for Ridge is 0.1 for the alpha as it results a lowest standard deviation.

5.5.2 Lasso Regression

Mean	Standard Deviation	Alpha value
-1.09593	0.05742	0.1
-1.11427	0.10265	1
-1.15999	0.12526	2
-1.22742	0.13492	3
-1.30491	0.11899	4
-1.44749	0.08882	10
-1.44749	0.08882	30
-1.44749	0.08882	100
RMSLE: 1.0375138554059375		

Table 9 Alpha Tuning for Lasso

The selected parameter for Lasso is 0.1 for the alpha as it results a lowest standard deviation.

Models	**RMSLE**
Linear Regression	1.0158500002746595
Ridge	1.0158500003240123
Lasso	1.0375138554059375

Table 10 Results of Linear Models

Multiple linear regression is used as a baseline in this task, the RMSE of linear regression is 1.0158500002746595 which is consistent with the analysis.

5.6 Results of Tree Based Models

In this task, we selected gradient boost decision tree (GBDT), random forest and XGBoost as tree based model to build up the prediction model. 20% data of the training set is used as validation part with 5 iterations of stacking. The performance of these models is listed below.

Models	**RMSLE**
XGBoost	0.43
Random Forest	0.11100668489749002
Gradient Boosting Decision Tree (GBDT)	0.20820457537732653

Table 11Results of Tree Based Models

5.6.1 XGBoost

Number of iterations	**RMSE (Train)**	**RMSE (Valid)**
[195]	0.105365	0.424261
[196]	0.105288	0.424227
[197]	0.105052	0.424328
[198]	0.104972	0.424389
[199]	0.104754	0.424495
[200]	0.103931	0.424791
[201]	0.103712	0.424673
[202]	0.103571	0.42458
[203]	0.103292	0.425369
[204]	0.103207	0.425269
[205]	0.102782	0.425161

Table 12 Xgboost iterative training loss

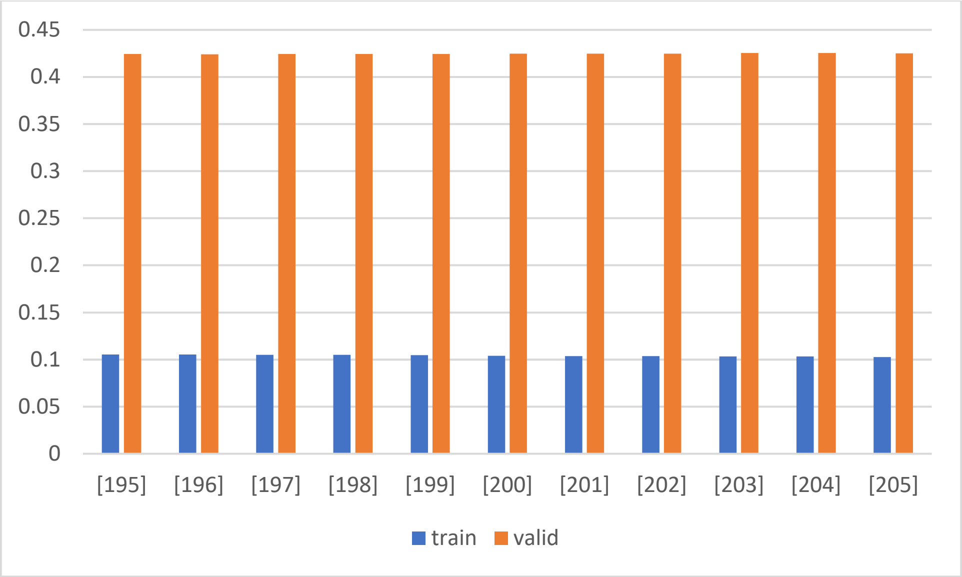


Figure 29 Visualization of Training Loss

As shown in figure 29, the value of RMSE of XGBoost is lower than the linear models. However, over-fitting occurs, the training RMSE decrease with the increase of validation RMSE. Overfitting means the hypothesis of model is too closely to the training set data. [12] The model tends to memorize the existing data points instead of predicting the new data point. Therefore, the model only performs better on the training set rather than the testing set or new data. As a result, avoid overfitting is a critical task in machine learning.

5.6.2 Random Forest

Random Forest shows an excellent performance in this task with 0.111~, the RMSLE is much lower than XGBoost as overfitting has been solved in modeling. By training the decision tree repeatedly, random forest will form a better decision tree. By using multiple such tress, it can reduce the overall error of prediction model. These decision trees will then join to operate the training set and use averaging to improve the predictive accuracy and control over-fitting. When the number of trees tend to be infinite in random forest, theoretically, the training error and validation error can be converged and reduce the RMSLE to a minimum. But it is impossible to build infinite trees in actual practice. In addition, the parameter settings of model will also affect the degree of overfitting. By tuning the parameters, random forest can effectively avoid overfitting.

We selected the best single model and used 5-fold cross-validation to explore the effect of the n_estimators and max_features parameters on the result. The results are recorded in the table below.

Number of n_estimator	**RMSLE**
50	0.43693
60	0.43599
70	0.43596
80	0.43646
90	0.43621
100	0.43617
110	0.43538
120	0.43540
130	0.43459
140	0.43529
150	0.3538

Table 13 N_estimator Tuning Results

As we can see, the result of RMSLE is similar between 50 – 140 trees in random forest. However, the RMSLE significantly drop when the number of trees is greater than 150 which mean overfitting occurs. Generally, using more estimators will increase the generalization of the tree model. But after a certain point, overfitting will occur, and result an accuracy loss.

Max_features is also another parameter in random forest, it represents the maximum number of features to consider for the best split. In this task, we consider some common value for max_features, those values are listed as follow. The result shows that using Sqrt value for max_features give a lowest RMSLE with 0.111.

max_feature	**RMSLE**
Sqrt (sqrt(n_features))	0.111
Log2 (log2(n_features))	0.1430591768451
Auto (Default)	0.1430
0.99 (percentage)	0.112

Table 14 Max_features Tuning Results

Models	**RMSE**
LinearRegression	1.0158500002746595
Ridge	1.037513855405938
Lasso	1.0158500003240178
xgboost	0.43
RandomForest	0.11100668489749002
GBDT	0.20820457537732653

Table 15 Final Result of Models

The result of models is listed as above, generally, the performance of tree based models is better than linear models as the data do not have a linear distribution. Among the models, decision tree gives the best performance. Finally, we select the single model to conduct the tuning process, and get the optimal parameters with n_estimators = 130, max_feature = sqrt in random forest. The model is then submitted on Kaggle and results a 0.47 test score which is a fair perdition result.

5.7 Ensemble learning

Stacking and blending are two similar approaches which used to improve the predictive performance by combining classifiers. These two approaches are applied in the modeling.

5.7.1 Blending

This ensemble approach is to combine the prediction result of several classifiers and then outperform any single classifier. Even in the worst-case scenario, the model result after blending is still better than the worst classifier. The average value is the most used rule of the combination of classifiers. By using disjoint data to train different base models, and then take an average value of the outputs. This approach is easy to implement but reduce the utilization of training set.

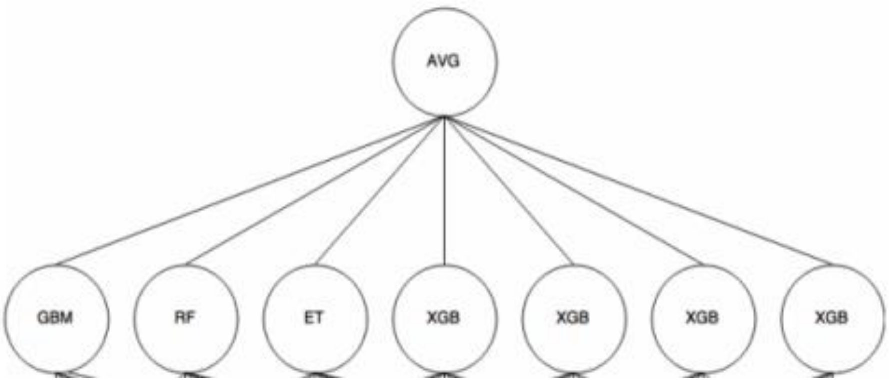


Figure 30 Concept diagram of Blending [27]

In this task, random forest and GBDT are outperform other classifiers. Therefore, these two classifiers are combined in blending and averaged the output values.

5.7.2 Stacking

Stacking is also an ensemble approach of combining multiple models, the mechanism combined the output from the level 0 classifiers as shown in figure 31. Those results are used as training data for the advanced classifier (level 1 classifier) to figure out the combining mechanism.

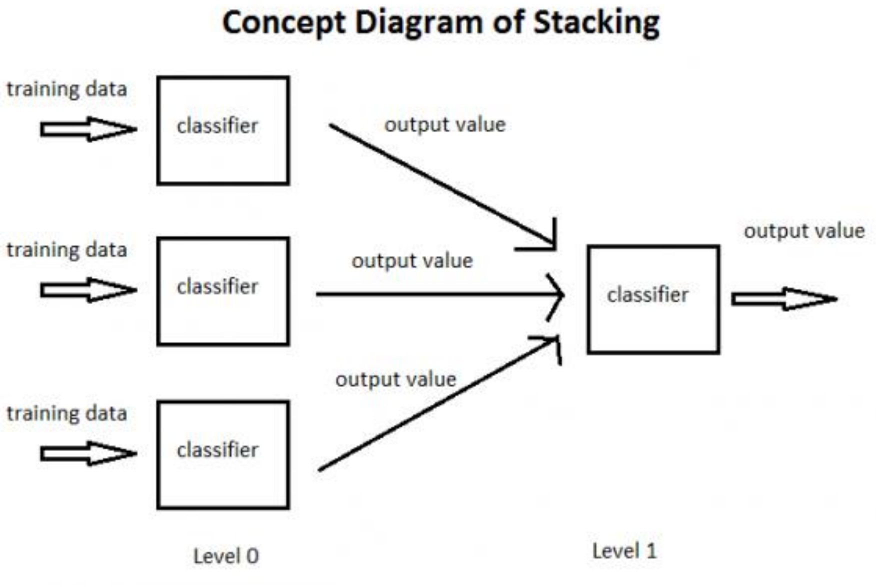


Figure 31 Concept diagram of Stacking [27]

In this task, the process of stacking is similar to the cross validation. The training data is first divided into 5 parts and trained with 5 iterations. At each iteration, 4 parts of data are used as training set to train the base model, the base model is then tested by the remaining hold-out set. The prediction result is recorded at each iteration. A new base model is generated and used to predict one part of the training data in each iteration. After 5 iterations have been completed, we obtain a matrix of *Number of training data rows x Number of Base Model*, which is then used as training set for another classifier in the next layer. When the training of the advanced classifier is completed, the previously saved prediction results of base model are used to test the data. Finally, the advance model is taken out to make a prediction and get the final result.

Chapter 6 Conclusion

Microscopic aspect

In this Kaggle competition, different linear models and tree based models are used to build a prediction model of bike rental demand in Washington D.C. We found that time variables such as hour and month and weather are the most predictive features for this problem. Also, data pre-processing and data engineering are useful in contributing to accurate predictions. By eliminating the noisy and data transformation, the performance of classifiers is improved. In addition, random forest is able to best capture the relationships within the data and avoid overfitting, which led to the best performance among the classifiers with 0.111 RMSLE.

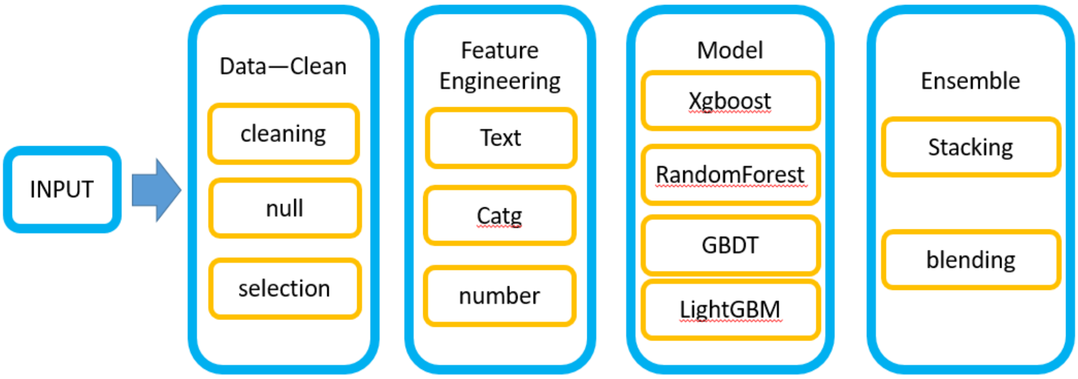


Figure 32 General Process of this task

Macroscopic aspect

After this Kaggle competition, I am more familiar with the general process of handling machine learning application, especially some skills are highlighted in the model integration. The process can be summarized as follows:

1. Improve the quality and understand of dataset

Enhancing the completeness and consistency of dataset by handle the missing values and remove the outliers. Then, check the data type of variables and create some univariate/bivariate plots to understand the nature of the variables.

2. Optimization

Each problem has a unique evaluation method and understand how it changes with the features. Select the features with a strong correlation to the prediction object in training set. Feature engineering is necessary to transform the raw data into useful features, such as transforming categorical variables and creating advanced features.

3. Measurement

Select the suitable algorithms for the task and compare the performance of the models. Understand the different of the results and determine the cross-validation strategy to avoid overfitting. Try to tune the parameters to improve the performance of classifiers.

4. Evaluation

Select the classifiers which are relevant to best result for integration. Ensemble approach is suitable to improve the predictive performance by combining classifiers and generate the final prediction model.

Chapter 7 Future Work

7.1 Adding more relevant covariates

As there are only 7 features useful in training set, I would like to add more covariates in future work and improve the predictions. The possible covariates that can be integrated in future are listed below.

1. Population density data

There could be a positive correlation between the population density and the demand of bike. A higher population density usually results a higher demand of bike rental. For example, the demand for bike in central business district and residential area generally higher demand as people may need a bike for commuting.

2. Public transportation data

There could be a positive correlation between demand for short transportations and bike within an area. If the area is lack of public transportation, the demand of bike may increase and vice versa.

3. The bikeway system

A good bikeway system is always attractive to cyclists. People feel safer and comfort with great bike routes. The demand of bike may increase with a mature of bikeway system.

These specific demographic features can represent the demand of bike rental in a city in a greater extent. By considering these covariates, the result of prediction will be more accurate and representative.

7.2 Using different classifiers and analysis model

In this task, all the classifiers I used are supervised learning algorithms. Therefore, I may apply some unsupervised methods in the future, such as clustering techniques. Clustering is a good technique to identify some of the meaningful relationships that we may missed in the data [28]. Also, time-series analysis is useful in predicting the demand of bike rental in real world and improve the generalization of the result.

Reference

[1] S. Shaheen, S. Guzman and H. Zhang, "Bikesharing in Europe, the Americas, and Asia", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2143, pp. 159-167, 2010.

[2] S. Parkes, G. Marsden, S. Shaheen and A. Cohen, "Understanding the diffusion of public bikesharing systems: evidence from Europe and North America", *Journal of Transport Geography*, vol. 31, pp. 94-103, 2013.

[3] K. Tu, "From Bike Messengers to App Stores: Regulating the New Cashless World", *SSRN Electronic Journal*, 2013.

[4] M. Narasimha Murty and D. Susheela Devi, *Introduction to pattern recognition and machine learning*. Singapore [u.a.]: World Scientific, IISc Press, 2015.

[5] R. Alexander Rixey, "Station-Level Forecasting of Bike Sharing Ridership: Station Network Effects in Three U.S. Systems", *TRB 2013 Annual Meeting*, 2012.

[6] C. Etienne and O. Latifa, "Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Vélib' System of Paris", *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 1-21, 2014.

[7] A. Faghih-Imani, N. Eluru, A. El-Geneidy, M. Rabbat and U. Haq, "How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal", *Journal of Transport Geography*, vol. 41, pp. 306-314, 2014.

[8] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge", *Progress in Artificial Intelligence*, vol. 2, no. 2-3, pp. 113-127, 2013.

[9] "Supervised Learning Workflow and Algorithms - MATLAB & Simulink - MathWorks United Kingdom", Au.mathworks.com, 2017. [Online]. Available: https://au.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html#. [Accessed: 19- Mar- 2017].

[10] B. Wang, Y. Yao, X. Wang and X. Chen, "PB-SVM Ensemble: A SVM Ensemble Algorithm Based on SVM", *Applied Mechanics and Materials*, vol. 701-702, pp. 58-62, 2014.

[11] R. Schapire, "Machine Learning Algorithms for Classification", 2015. [Online]. Available: http://www.cs.princeton.edu/schapire/picasso-minicourse.pdf. [Accessed: 24- Mar- 2017].

[12] D. Sharma, "De-Biased Random Forest Variable Selection", *SSRN Electronic Journal*, 2011.

[13] Y. Freund, "An introduction to boosting based classification", AT&T Labs, 2013. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.3483&rep=rep1&type=pdf.

[Accessed: 24- Mar- 2017].

[14] W. Liu, Y. Han, F. Wan, F. Bretz and A. Hayter, "Simultaneous Confidence Tubes in Multivariate Linear Regression", *Scandinavian Journal of Statistics*, vol. 43, no. 3, pp. 879-885, 2016.

[15] G. Janacek, "Time Series Analysis Forecasting and Control", *Journal of Time Series Analysis*, 2009.

[16] Chalodhorn, R. *Machine Learning II: Unsupervised Learning*, 2016 [online] Available at: https://courses.cs.washington.edu/courses/csep573/10wi/06-unsup-learning-2up.pdf [Accessed 24 Mar. 2017].

[17] L. Mentaschi, G. Besio, F. Cassola and A. Mazzino, "Problems in RMSE-based wave model validations", *Ocean Modelling*, vol. 72, pp. 53-58, 2013.

[18] J. Fan, S. Guo and N. Hao, "Variance estimation using refitted cross-validation in ultrahigh dimensional regression", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 1, pp. 37-65, 2011.

[19] D. Singhvi, S. Singhvi, P. Frazier, S. Henderson, E. Mahony, D. Shmoys, and D. Woodard. Predicting bike usage for new york city's bike sharing system. In *AAAI Workshops*, 2015.

[20] J.Malani, N.Shinha, N.Prasad, V. Loksh, "Forecasting Bike Sharing Demand", 2013 [online] Available at: https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a050.pdf [Accessed 24 Sep. 2017].

[21] R. Rixey, "Station-Level Forecasting of Bikesharing Ridership", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2387, pp. 46-55, 2013.

[22] Y.Li, Y. Zheng, H. Zhang, L. Chen, "Traffic prediction in a bike-sharing system", *SIGSPATIAL,* 2015.

[23] X. Sun, "Preprocessing of Examination Analysis System Data Based on Data Mining", *Applied Mechanics and Materials*, vol. 608-609, pp. 300-303, 2014.

[24] Q. Feng, J. Hannig and J. Marron, "A note on automatic data transformation", *Stat*, vol. 5, no. 1, pp. 82-87, 2016.

[25] P. Sedgwick, "Log transformation of data", *BMJ*, vol. 345, no. 081, pp. e6727-e6727, 2012.

[26] Y. Sheikh, "Effective Feature Selection for Feature Possessing Group Structure", *International Journal Of Engineering And Computer Science*, 2017.

[27] L. Efanova, "Blending. Part 2. Main varieties of blending", *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya*, no. 139, pp. 5-14, 2016.

[28] H. Kriegel, P. Kröger and A. Zimek, "Detecting clusters in moderate-to-high dimensional data", *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1528-1529, 2008.