DBA5106 Term project

Meta City Bicycle Sharing Demand Prediction

Ke Ma, A0212524U
Mingzhe Xu, A0232022A
Nanhai Zhong, A0231953E
Xiao Liang, A0232007X

1 Introduction

1.1 Background

Meta city is the first city built in Meta Universe, we are hired by a bike sharing company to use data about bicycle usage and spot characteristics and generate valuable business prediction model from it to support new market operation strategy making.

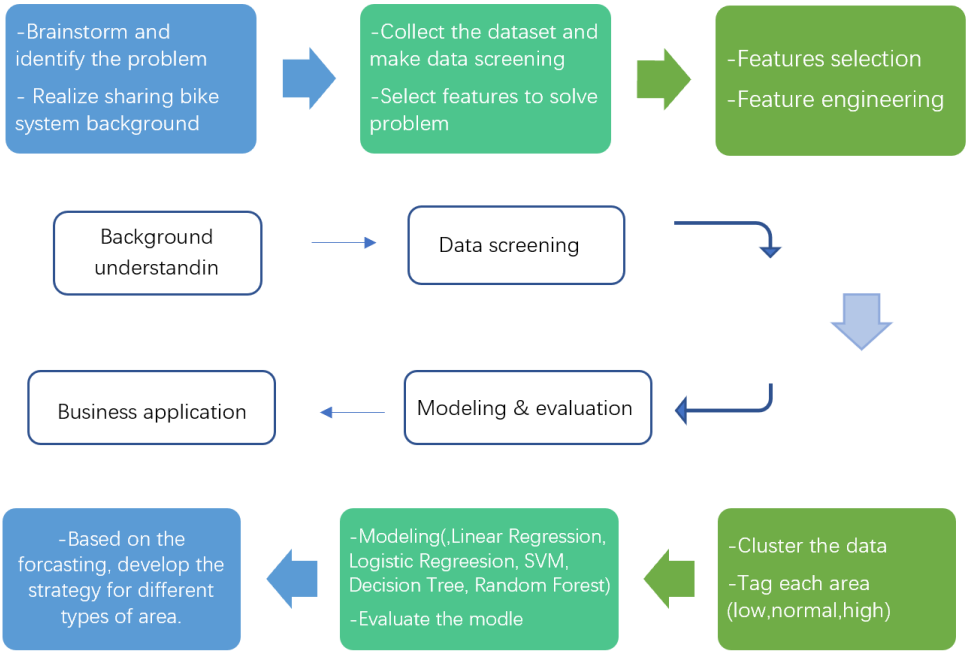
In our Meta City, people would be able to spend more time outdoors. Therefore, we plan to utilize our dock-less bicycle to improve our public transportation system, especially for the last 1 mile. People could use this new transportation system affordably and conveniently. On the other hand, this innovative approach would reduce carbon emission if people would like to use the sharing bicycle.

1.2 Problem statement and objectives

In this project, we are required to solve some problems such as the bike deployment because we do not provide exact places to lock the bike. In order to fit the demand, we use linear regression to forecast the number of the demand. Moreover, we combine with some practical problem to apply in business field. Therefore, we utilize the classification method to solve these problems. Finally, we separate into 3 types which include low, normal and high for the whole city and customize different strategies for each type.

1.3 Methodology

For the whole project, we follow the classic business analytics workflow.



2 Data Exploration and Preprocessing

2.1 Abstract

The original data set is a nine-day data set of shared bicycles in a certain city from Apr.26th 2017 to May.4th 2017, including the number of bicycles used and the total number of bicycles, as well as other related factors that may affect the volume of use. There are 460300 records in the original data set without any missing values. Each record represents the information for a specific area in a specific hour. The dataset contains 28 features, which show information in five dimensions: number of shared bicycles, time, geographic location, geographic features and weather condition.

Feature	Definition	Time sequenced	Spot features
ID	key	N	N
CELL_ID	region ID	N	N
DAYS	date	Y	N
HOURS	hour	Y	N
SEQ	the i-th hour	Y	N
MOVE	volume of use per hour	Target	Target
TOTAL	total volume per hour	N	N
HDB	public residential building floor area ratio	N	Y
PRIVATE	private residential building floor area ratio	N	Y
COMM	commercial building floor area ratio	N	Y
INDU	industrial building floor area ratio	N	Y
CYCLPATH	total length of bicycle lane (m)	N	Y
BUS_STOP	total number of bus stops	N	Y
ROADINT	total number of intersections	N	Y
ROAD_LIN	total number of sections	N	Y
ENTROPY	entropy	N	Y
MRTDIST	distance to the nearest subway station (m)	N	Y
DISTCEN	distance to the city center (m)	N	Y
RAIN	rainfall in one hour (cm)	Y	Y
TEMPO	temperature in one hour (Celsius)	N	Y
LABORFREE	Free Ride Activity on May Day	Y	N
WENDFREE	Free Ride Activity on Weekends	Y	N
Xmin	the west border in grid map	N	Y
Xmax	the east border in grid map	N	Y
Ymin	the south border in grid map	N	Y
Ymax	the north border in grid map	N	Y
Longitude	longitude	Y	Y
Latitude	latitude	Y	Y

2.2 Exploratory data analysis (EDA)

2.2.1 The Volume in Different Days

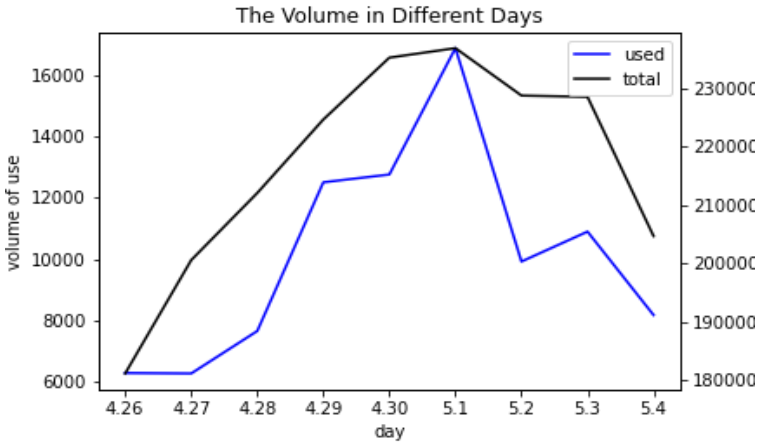


Figure 1 The volume in different days

The ***Figure 1*** shows that the fluctuation of demand during the 9 days are quite large. The single-day usage on the 26th and 27th is relatively small, while the usage on the 29th, 30th and 1st is so big that it is more than twice the usage on the 26th and 27th. The doubling of usage on weekends and May Day can be preliminarily inferred to be related to the free riding activities on May Day and on weekends. In addition, it can also be confirmed from the figure above that not all Wednesdays and Thursday have such low usage as 26th and 27th. The daily usage during the working day may be approximately 6,000 to 11,000. In summary, we could suppose that under normal circumstances, the daily usage on weekdays is lower than it on weekends.

Moreover, the figure shows that the volume of use has the same trend with the total volume. Hence, we hypothesize that there is a potential relationship between the total volume and the volume of use. Then, we use Pearson to test the hypothesis. The result shows that the correlation is 0.311905 and the p-value is 0.00 (<0.05), which means they do have significant correlation.

2.2.2 The Volume of Use of Different Time Slots

In order to compare the demand at different time slots, we use heatmap to show the volume of use in different hours on different days.

Comparing the data horizontally from ***Figure 2,*** we can find that the morning peak is from about 7:00 to 8:00 every weekday and appears a bit later on weekends. The value of the morning peak is from about 300 to 500. The evening peak in usage is from 18:00 to 19:00. The value is larger than the morning peak and spans a large range from about 500 to 1,000. In addition, from 1:00 to 5:00 on weekdays and from 2:00 to 6:00 on weekends and May 1st are low demand periods.

Comparing the data vertically, we could find that the colors of weekends and May 1st, which are much darker than weekdays, which shows that they are in line with the analysis of demand change over week: there are more demands of shared bike on weekends and holidays.

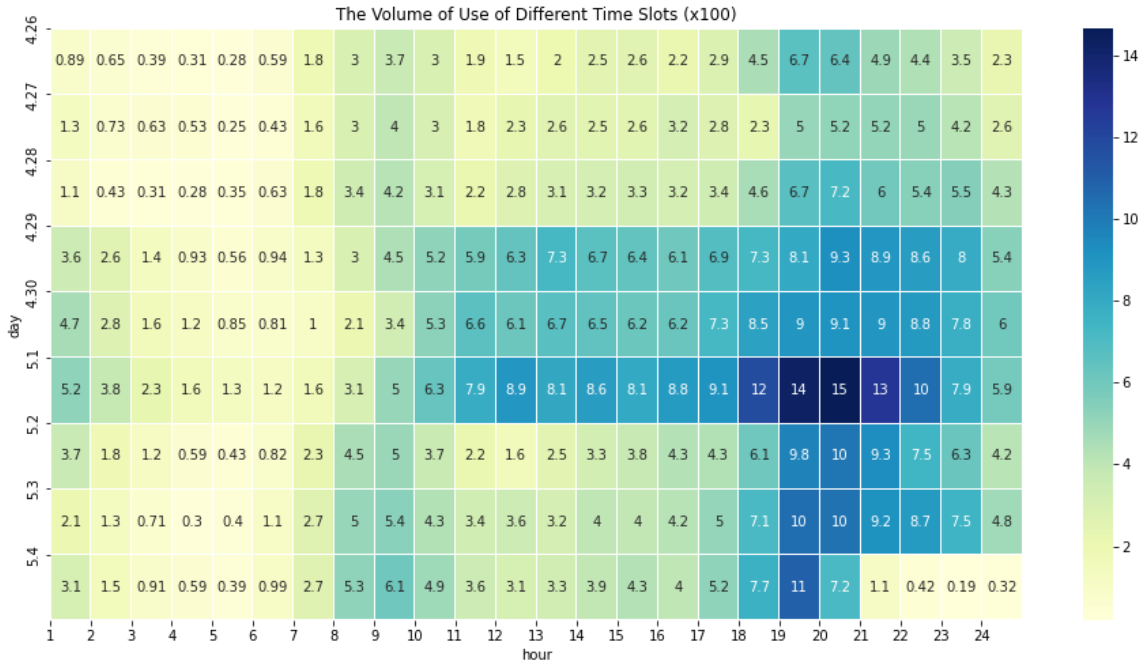


Figure 2 The volume of use of different time slots (x100)

2.2.3 The Volume of Use in Different Regions

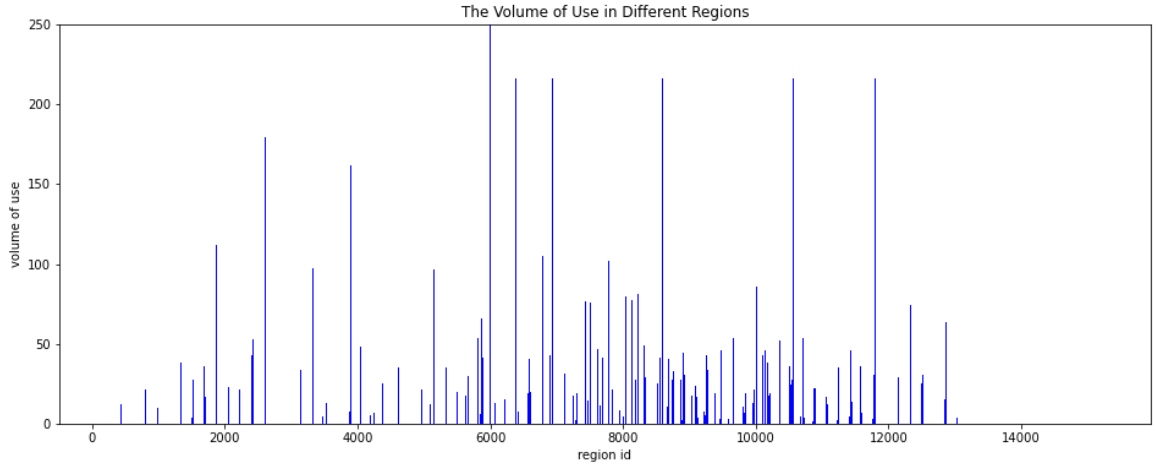


Figure 3 The volume of use in different regions

From the ***Figure 3***, we could know that the total number of used shared bikes among different regions has high variance, which will influence the prediction accuracy. The reason of the high randomness is that the area of each region is relatively small. Therefore, in the following process, the randomness can be decrease by clustering each region into a bigger region. In addition, it is more practical for bike-sharing companies to implement regional management rather than focused on each small area.

3 Clustering

3.1 Clustering

Bike-sharing companies develop different strategies for different management areas to carry out commercial activities such as advertising, bike delivery and so on. Therefore, it is important to do

clustering in order to help them divide the city into different management areas and to decline randomness in data set.

The purpose of clustering is to combine the adjacent and similar regions together. Therefore, we use K-medians with Manhattan distance, which is more in line with the actual distance calculation rules. Moreover, in order to satisfy the similarity requirement, we not only take the location features (*xmin, xmax, ymin* and *ymax*) into account, but also consider those features related to location, such as *MRTDIST*, *DISTCEN* and so on. Because the difference in scaling, the location features have higher weight when calculating distance. At the same time, other features will play a role in ensuring similarity.

Firstly, we try to identify the ideal K for K-medians. The curve between WSS and K ***Figure 4-a*** shows that when K is larger than 20, the WSS fluctuates in a small range between 0.5 and 1.5. Considering the difficulty of regional management and the value of WSS, 33 is chosen to be the value of K.

Then, we get 33 new combined regions shown ***Figure 4-b***. Compared with the clustering only

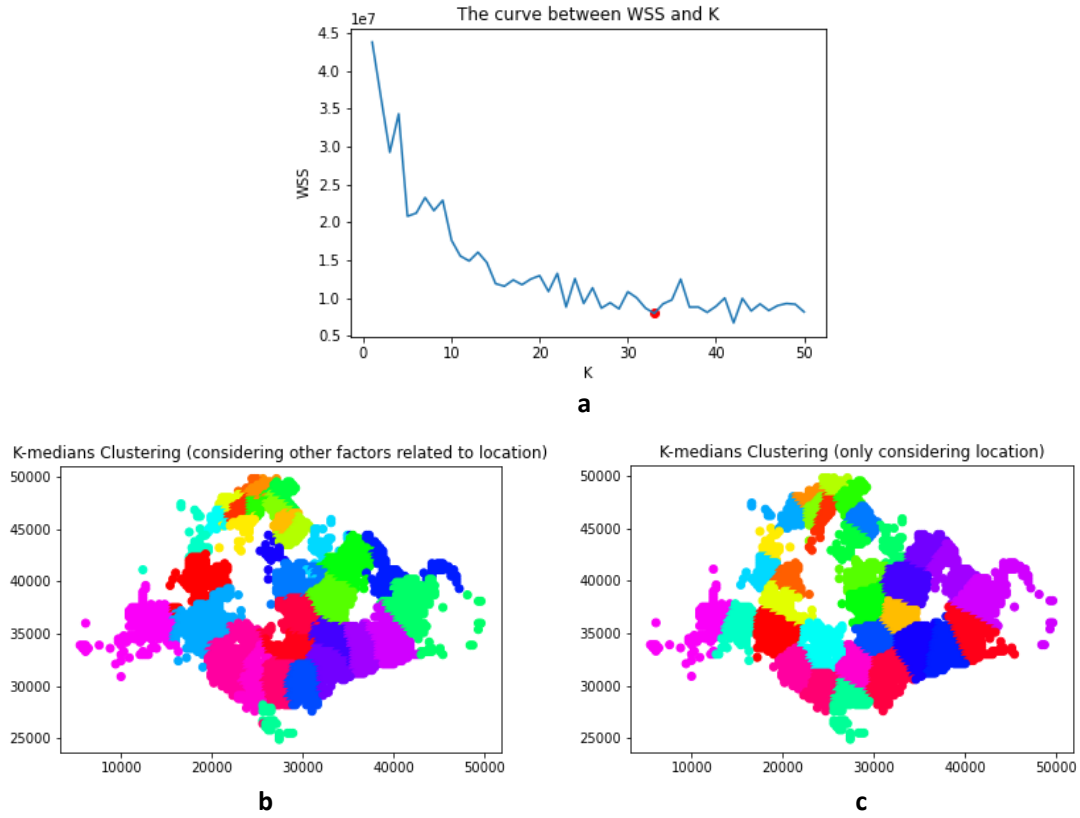


Figure 4 Clustering and parameter tuning

considering location ***Figure 4-c,*** although the ways of dividing the area are different, there is more similarity within each new region. For example, by adding *MRTDIST* into the model, the clustering will combine those closer regions with similar distance to the MRT as a cluster rather than only combine the closest regions together.

Finally, integrate data based on clusters and time slots. The original features could be divided into three groups: related to time, integrated by summing, integrated by averaging. The group of features related to time includes *SEQ*, *DAYS*, *HOURS*, *LABORFREE* and *WENDFREE*. The group of features needed to be integrated by summing includes *MOVE*, *TOTAL*, *CYCLPATH*, *BUS_STOP*, *ROADINT*, *ROAD_LIN* and *RAIN*. The group of features needed to be integrated by averaging includes *HDB*, *PRIVATE*, *COMM*, *INDU*,

ENTROPY, *MRTDIST*, *DISTCEN*, *TEMPO*, *x* and *y*. Then, we get the new data set with 7122 records.

3.2 Level Evaluation

By evaluate the demand level of each cluster, the companies can differentiate the potential of creating value of different regions so that they can set up targeted operation approaches to different regions based on their level.

Table 3-1

Indicator	Value
count	33
mean	2766.424
std	2687.119
min	54
25%	880
50%	1945
75%	3359
max	9951

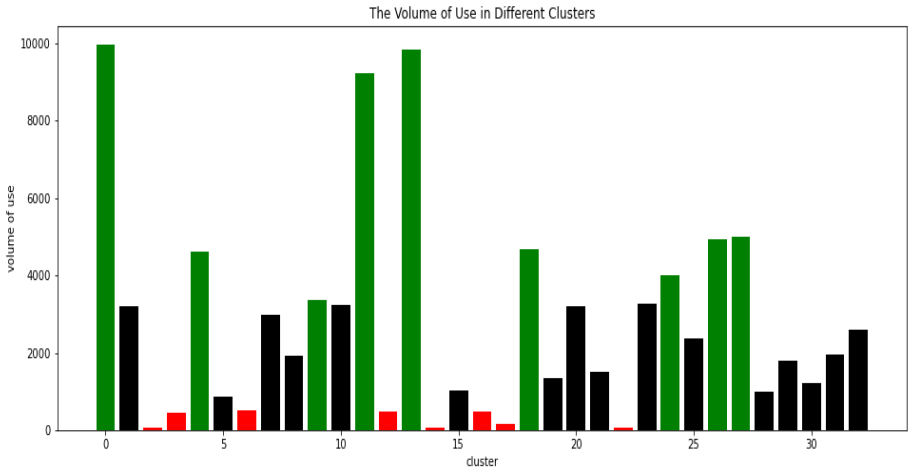


Figure 5 The volume of use of different clusters

The *Table 3-1* above shows that the 1st quartile and 3rd quartile of the volume of use in different clusters are 880 and 3359. We use these two values as two thresholds to divide the clusters into 3 level ***Figure 5***: low (red), normal (black) and high (green). The aim is to distinguish the high demand level clusters and low demand level clusters from the normal clusters. Hence, the companies can pay more emphasis on the high demand level to create more value and on the low demand level to reduce cost.

4 Demand Prediction model

After clustering, we get the demand of shared bike at each new combined region. For bike-sharing companies, predicting the demand at each management area will give them a better view about how to deliver and recycle the bicycles. Therefore, in this section, we use LASSO to figure out the relationship between demand of shared bike and other factors. In addition, cross validation is used to select the ideal models.

4.1 LASSO Model

In LASSO, the penalty term is in L-1 norm, which means it will force some coefficient estimates to be exactly zero. The sparse model is more useful in reality, because sometimes when we face a new case, we cannot get all the features we have now.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{14}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{14}|\beta_j|$$

In the first model, considering the feasibility of reusing this predictive model in future, only features that can be obtained before the actual demand occurs are included in the model.

$$\text{demand} = \beta_0 + \beta_1 HOURS + \beta_2 HDB + \beta_3 PRIVATE + \beta_4 COMM + \beta_5 INDU + \beta_6 CYCLPATH + \beta_7 BUS_{STOP} + \beta_8 ROADINT + \beta_9 ROAD_{LIN} + \beta_{10} ENTROPY + \beta_{11} MRTDIST + \beta_{12} DISTCEN + \beta_{13} LABORFREE + \beta_{14} WENDFREE$$

In the second model, in order to figure out how the supply will influence the demand, we add the feature *TOTAL* into the model.

$$\text{demand} = \beta_0 + \beta_1 HOURS + \beta_2 HDB + \beta_3 PRIVATE + \beta_4 COMM + \beta_5 INDU + \beta_6 CYCLPATH + \beta_7 BUS_{STOP} + \beta_8 ROADINT + \beta_9 ROAD_{LIN} + \beta_{10} ENTROPY + \beta_{11} MRTDIST + \beta_{12} DISTCEN + \beta_{13} LABORFREE + \beta_{14} WENDFREE + \beta_{15} TOTAL$$

4.2 Cross Validation and Model Selection

There are two main traits to consider when choosing a model: parsimony and accuracy. The ideal model should have less features but still has a good performance.

Therefore, we firstly use 10-fold cross validation to check the accuracy of each λ. From the results ***Figure 6***,the LASSO model performs best when $\lambda = 0.01$ for the first model and performs best when $\lambda = 0.02$ for the second model. The differences of performance are small in both model for $\lambda \in$ {0.00005, 0.0001, 0.0005, 0.001, 0.0015, 0.002, 0.01, 0.02, 0.05, 0.1, 0.2}.

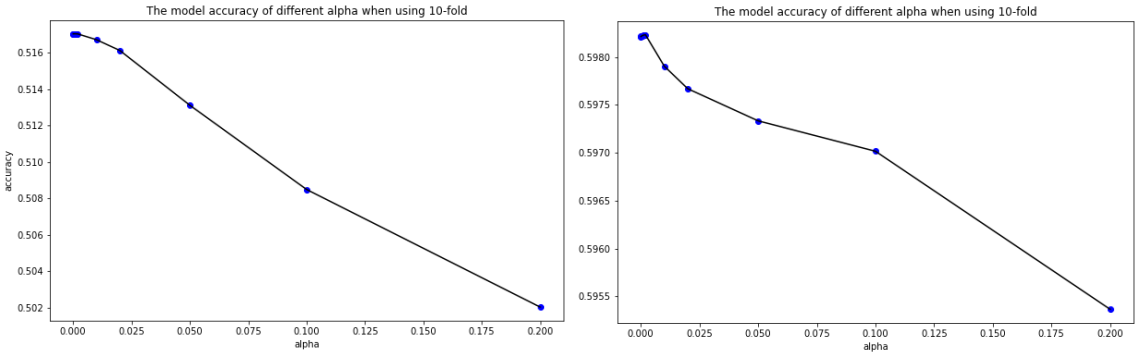


Figure 6 The Accuracy of the Model with (left) and without (right) TOTAL

However, for the first model when $\lambda = 0.01$, the model still contains all the features. When $\lambda = 0.2$, the model only contains 10 features and has similar performance ($R^2 = 0.502$) as $\lambda = 0.01$. Therefore, we finally choose $\lambda = 0.2$ for the first model. For the second model, considering the consistency and parsimony, we also choose $\lambda = 0.2$ where $R^2 = 0.613$. Then, the corresponding coefficients of each model are shown in ***Table.4-1.***

Table.4-1 The Coefficients of Two Predictive Model

Coefficient	1st Model	2st Model
intersection	-12.551	-9.7745
HOURS	0.8454	0.8434
INDU	-0.8344	-
CYCLPATH	0.0007	0.0002
BUS_STOP	0.149	-0.0914
ROADINT	-0.0047	0.012
ROAD_LIN	0.0009	-0.0042

MRTDIST	-0.0026	-0.0012
DISTCEN	0.0004	0.0001
LABORFREE	8.4356	7.106
WENDFREE	3.5575	2.3894
TOTAL	-	0.0581

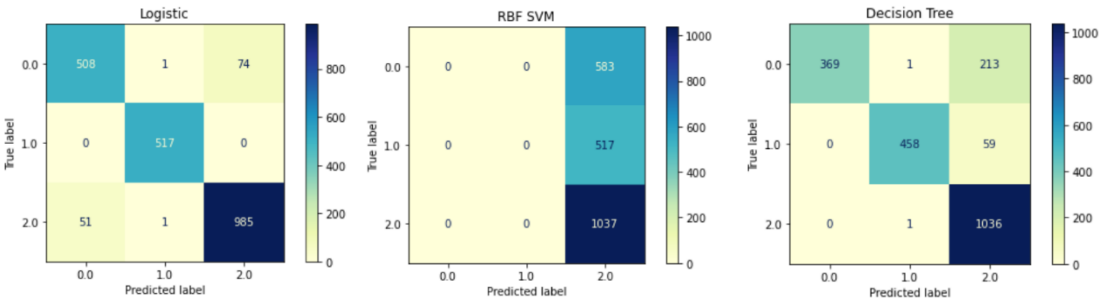
The results of the two models show that *WENDFREE* and *LABORFREE* have the biggest influence on the demand, which is in line with our previous analysis. Also, by taking *TOTAL* into consideration, the accuracy of predicting the demand is increase. The coefficient of *TOTAL* is positive and significant, which means that the more we deliver the shared bikes the more bikes are used. Moreover, except for *ROADINT* and *ROAD_LIN*, the coefficients of other features are still quite similar. Therefore, we could infer that by introducing the feature *TOTAL* we have further explain the model in a certain dimension which used to be reflected in the error term.

5 Demand level prediction model

5.1 Basic Models

We divide the demand for bicycles into three categories: high, normal, and low. Therefore, the bicycle demand forecasting problem is transformed into a classification problem. We select 14 variables that may affect the demand, namely *HOURS, HDB, PRIVATE, COMM, INDU, CYCLPATH, BUS_STOP, ROADINT, ROAD_LIN, ENTROPY, MRTDIST, DISTCEN, LABORFREE, WENDFREE,* and use *LEVEL* to represent the level of bicycle demand. According to the previous data preprocessing, we have 7122 pieces of data, 70% of which are used as the training set and 30% as the test set. We select several basic models to check the accuracy of the classification and then select two of these models with the best performance for parameter tuning.

First, we establish the Logistic, RBF SVM, Decision Tree, Random Forest, AdaBoost, Naive Bayes models. For each model, we return a classification report and confusion matrix as follows. According to the results, we can conclude that the performance of the Decision Tree and the Random Forest is the best, so we choose these two models for parameter tuning.



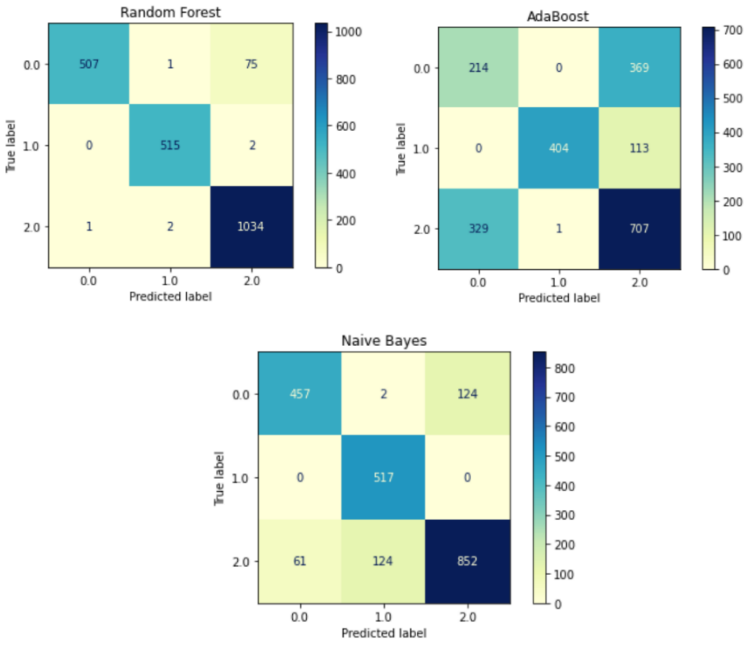


Figure 7 Confusion matrixes

5.2 Model Tuning

For the Decision Tree, we adjust the parameters of *criterion, max_depth, min_samples_leaf* and *min_impurity_decrease*. In the end, we choose the Decision Tree whose *criterion* is entropy, *max_depth* is 30, *min_impurity_decrease* is 0.1, *min_samples_leaf* is 2 as the optimal model.

For Random Forest, we adjust the parameters of *n_estimators, criterion, max_features, max_depth* and *min_samples_split*. In the end, we choose the Random Forest whose *criterion* is gini , *max_depth* is 25, *max_features* is 4, *min_samples_split* is 5, *n_estimators* is 11 as the optimal model. The followings are the confusion matrix for these two models.

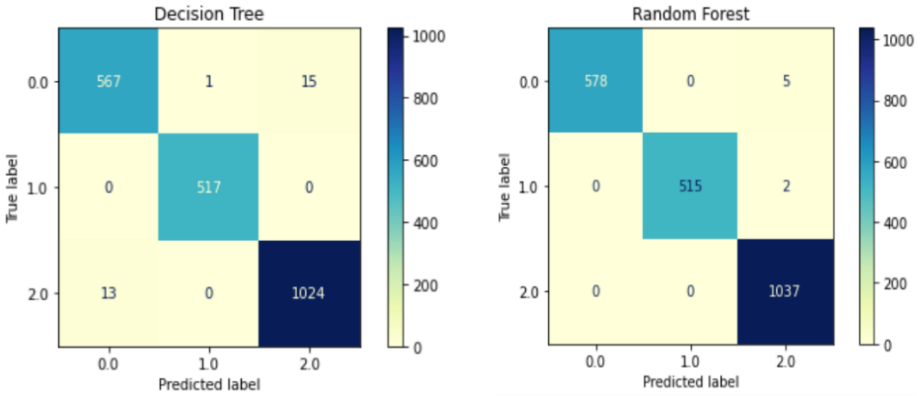


Figure 8 Tunning model results

Comparing the above two models, we can discover that **Random Forest** has a higher precision, recall, f1-score than Decision Tree. Random forest can handle high-latitude data, is insensitive to noise, and is not easy to overfit. These advantages make it perform better. Below is the classification report for our Random Forest model.

	precision	recall	f1-score	support
high	1.00	0.99	1.00	583
low	1.00	1.00	1.00	517
normal	0.99	1.00	1.00	1037
accuracy			1.00	2137
macro avg	1.00	1.00	1.00	2137
weighted avg	1.00	1.00	1.00	2137

5.3 Variable importance

According to our Random Forest, we also return the feature importance. We can find that *ROADINT, BUS_STOP, DISTCEN, HDB, ROAD_LIN* have a big impact on demand level. These five features represent the basic population and number of commuters. It indicates that the greater the population base and the greater the number of commuters, the greater the demand for bicycles. We can also find that *HOURS* has a low impact on demand level, for each level contains *HOURS* range from 0 to 24. *WENDFREE* and *LABORFREE* have even lower impact of the model. This is because although the promotion activity "free ride" stimulates the demand, this effect is equivalent in areas regardless of their basic demand level. The stimulation affect is also diluted in 9 days.

From the insights gathered above, we preliminarily suggested the bike sharing company when determine area demand level, relies more on ***geographic characteristics*** that indicates high transportation flow (e.g., number of bus stop). And be careful to ***time-dependent*** disturbed factors when dealing with time-series data, for these factors have no contribution for demand prediction.

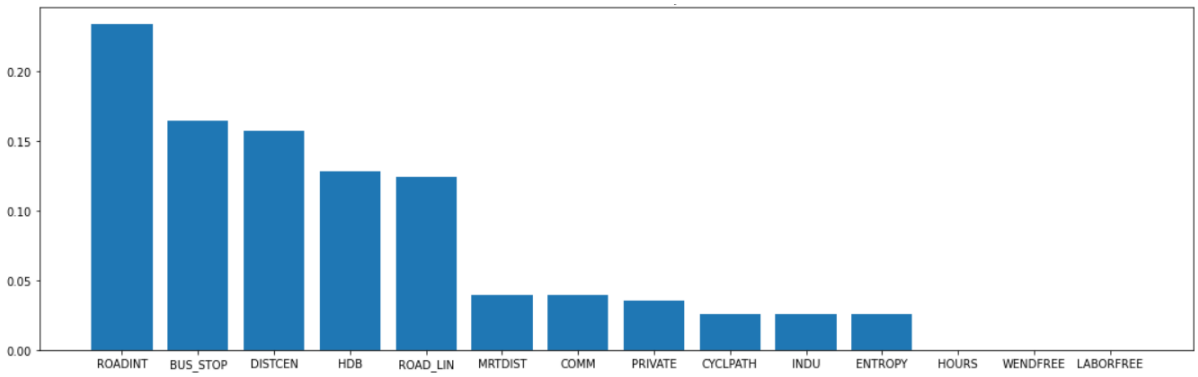


Figure 9 Feature importance plot

5.4 Model evaluation: Expected value

After tuning hyperparameters of top two performance models, their confusion matrixes are generated. In each model, number of total test instances is 2137. To furtherly compare the two models, we introduce domain knowledge to create a benefit-cost matrix, so called *profit matrix* to calculate expected value. Expected value framework is especially useful in business decision making, it decomposes data-analytics thinking into elements that extracted from dataset and that need to be acquired from outsource. The expected value calculation framework is as follows,

$$EV = \Pr(o_1) \times v(o_1) + \Pr(o_2) \times v(o_2) + \Pr(o_3) \times v(o_3) + \cdots$$

5.4.1 Benefit-cost matrix: with domain knowledge

Later on, we established a very simple business model for benefit-cost matrix. Bikes are shared among people with short rides for each time used and paid one-time or as club membership fee. We assume that a certain number of bikes are decided to be transported and placed in each area based on different demand estimation at the beginning of a week. And bikes are barely ridden out of the area where they are placed initially. The value chain of a sharing bike is illustrated as follows,

Net profit = gross profit per use × used time – placing cost per unit × placing amount (supply)

– operation cost per use × used time – promotion spending

Among them, $used\ time = MIN\{supply, demand\}$. *Placing amount*, *promotion spending* is determined ahead of the week based on estimation results (i.e., yellow columns in ***table 5-1***). In contrast, gross profit and operation cost are dependent on actual number of uses (i.e., blue columns in ***table 5-2***).

We assume unit cost/profit as follows, and generate the benefit-cost matrix in ***table 5-2***

Gross profit per use	$ 1.2
Placing cost per unit	$ 0.15
Operation cost per use	$ 0.3

Table 5-1 profit model parameter

Demand level	Average demand	Supply	Gross Profit	placing cost	operation cost	promotion spending
Low	276	300	331.2	45	82.8	20
Normal	2093	2500	2511.6	375	627.9	120
High	6177	7000	7412.4	1050	1853.1	300

Table 5-2 profit matrix

		Actual		
	Demand Level	Low	Normal	High
Est.	Low	$ 183.4	$ -542.9	$ -1,768.1
	Normal	$ -439.8	$ 1,388.7	$ -1,098.1
	High	$ -1,294.8	$ -931.4	$ 4,209.3

5.4.2 Calculate expected value

Based on benefit matrix and confusion matrix we generated in above models, we are able to calculate expected value for each model. Specifically, we calculate estimation rate for each column something like true positive rate with total actual number as denominator. Then aggregate product of two matrix on the

weight of prior class proportion. Expected value are displayed under two tables.

Table 5-3 Decision Tree confusion matrix and expected value

Decision Tree		Actual		
		Low	Normal	High
Est.	Low	97.76%	0.19%	1.44%
	Normal	0.00%	99.81%	0.00%
	High	2.24%	0.00%	98.56%
Prior Class proportion		27.14%	24.24%	48.62%
Expected value		**$ 2381.08**		

Table 5-4 Random Forest confusion matrix and expected value

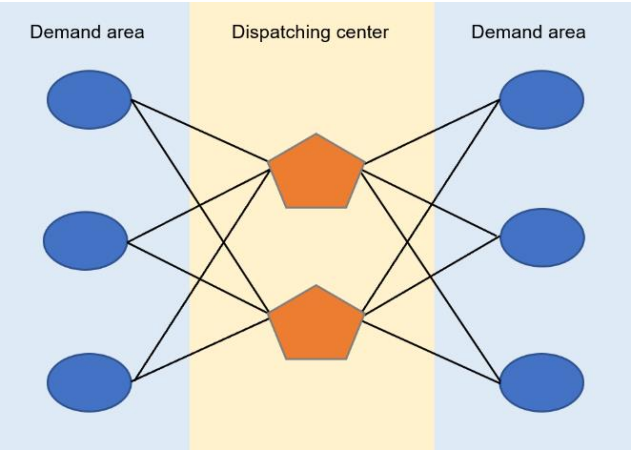
Random Forest		Actual		
		Low	Normal	High
Est.	Low	100.00%	0.00%	0.48%
	Normal	0.00%	100.00%	0.19%
	High	0.00%	0.00%	99.33%
Prior Class proportion		27.05%	24.10%	48.85%
Expected value		**$ 2421.71**		

As results implies, random forest model is prior to decision mode by expected value $40.63 in average for each cluster per week. For ***33*** business areas in meta city, decisions based on ***Random Forest*** model are expected to generate profit **$ 1340.71** more than the other model.

6 Business Application

6.1 Dispatching system

Above, we have built two models to predict demand or demand level in areas. Lasso linear regression is able to predict numeric demand quantity given geographical characteristics and hour. Though this is still a rough prediction model, it can be deployed in company's dispatching system. As shown in the plot, demand distributed sparsely in meta city. The company only allows to decide the initial number of bikes in each area. However, the demand is fluctuating between day to day. So, we designed a dispatching system to transport bikes from areas to areas after peak hour to meet supply shortage. The dispatching system, theoretically, integrated the data we predicted for next day with number of bikes remained in each area after peak hour and decide dispatch strategy. To furtherly solve the problem, it



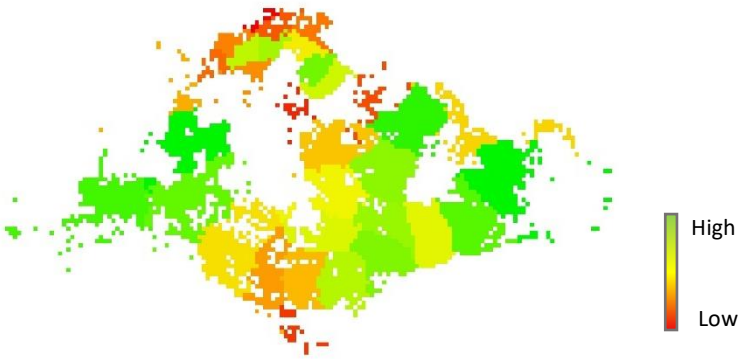


Figure 10. Predicted Demand distribution over meta city*. The plot displays predicted demand in test set. Red represents low demand area and green represents high demand area. Dispatch system transport spare bikes in red areas to green areas and also transport bikes between high demand areas in different hour.*

might need to introduce linear programming methods, e.g., ***Multi-Location Newsvendor Problem.*** Solving this problem is not the major goal for this project, but the demand prediction model surely provides crucial information.

6.2 New market entry decision

According to linear regression model, the interpretability is not satisfied and features like *HOUR* is not exactly linearly related to demand, since as the ***Figure 2*** shown, peak hour is around 6-8 pm. From another aspect, for most business decisions, accuracy of the prediction outcome is weighted more than precision. So, in the second model, we rank each area according to their total bike move over a week and label the data into l*ow, normal and high* three categories. We convert a numerical prediction problem into a classification problem to extend its business application.

By applying the second model, we help bike sharing company to predict the demand level of any potential area that might be the new market way earlier than they begin to invest into the market. Model features except *HOUR* are all geographical features which are relevantly fixed on the map and are easy to be collected. Moreover, according to variable importance ***Figure 9***, *HOUR* has little effect to the Random Forest model.

Demand level forecasting contributes a lot in business decision making in topics of market entry strategy, operation strategy, marketing activities, etc.

Est. demand level	Action
LOW	Low placing number, or set as forbidden areas. Design *Bike Hunting* campaign to encourage people move bikes out of such areas. Reward “hunters” with coupons or cash, which helps reduce transport cost. Low maintenance intensity for bicycles in this area.
Normal	Normal placing intensity, Normal promotion activity. This kind of areas takes a major proportion.

High	Place more bikes to meet high demand in this area. Allocate more dispatch capacity in such area to increase efficiency and reduce average bike idle time in other lower demand areas. Design and apply promotion activities to release potential demand.

Furtherly, the model can be used in a broad range of business operation area and decision making. Combine with domain knowledge or even profit and cost information of the company, the model can be used to evaluate market value and business performance. In ***5.4 Model Evaluation: Expected value*** section, we give a very simple demonstration predicting average profit per area generated by using Random Forest model. Though the benefit-cost matrix is based on heavy assumptions, it has a lot potential value for any bike sharing company over the world.