An Exploration to Reduction of Chicago Traffic Congestion: Regression, Time Series and Content Analysis on Divvy Bikes System

Yilun Xu

Division of Social Sciences, University of Chicago

yilunx@uchicago.edu

Abstract

With the development of the city, traffic jam in the Chicago area has become increasingly serious. However, a Bike-sharing system is regarded as a possible solution to this situation. In this paper, we used data from Divvy Bikes, a bike-sharing company in Chicago, to analyze how we can promote a bike-sharing system to alleviate traffic congestion. First, we use principal component analysis and Ridge regression to combine variables such as time, road conditions, weather conditions, user's gender, and account properties to regress Divvy Bikes' daily usage and average riding duration. The effects of part of independent variables were analyzed as well. Also, we use the time series model to predict the future usage of Divvy Bikes, including daily usage and average riding duration, and the prediction results are relatively close to the actual situation. Finally, through sentiment analysis and communication networks, we analyze users' attitudes towards Divvy Bikes products and services, and at the same time capture the more representative problems of the products raised by consumers. This study will provide a reference for Divvy Bikes to analyze the usage of Divvy Bikes under different situations and make their business strategy in the future, and how the City of Chicago can promote the bike-sharing market in the future.

Introduction

Traffic congestion describes the situation in which the development of vehicles is deferred by each other in light of restricted street limits (Vencataya et al., 2018). Traffic clog is deteriorating and driving expense is getting higher in each significant metropolitan territory in the United States (Jin & Rafferty, 2018). Gradually, citizens regard traffic congestion as one of the most serious problems in their daily life which bothers millions of commuters on all social levels. Furthermore, in metropolitan cities, traffic congestion is believed to cause more damage to the less developed areas. As a large city in the Midwest, Chicago is also suffering from traffic congestion. Many factors are guessed to cause traffic congestion, including population growth, job opportunities, lack of necessary infrastructure (like new roads), increasing private cars, etc. (Downs, 2000). Also, the approach of neo-liberal urban regionalism and financialized urban governance further exposed the inadequacy of traffic management in the Chicago area (Addie, 2013). Guided by the concept of neoliberal urbanization, Chicago has extensively deployed public transportation in the city center to attract tourists and capital from around the world. This has led to a lack of investment in public transport in other regions, which gave birth to a decline in transport service levels and a lag in transport infrastructure. As a result, neo-urban liberalism has led to the uneven development of public transport in the Chicago area (Farmer, 2011).

Some researchers have given some solutions. To increase the fairness and efficiency of the Chicago bus system, which has been challenged by neo-urban liberalism, a new technique called Bus Rapid Transit (BRT) should be adopted (Sukaryavichute & Prytherch, 2018). Uber has improved the quality of the taxi market in Chicago. However, the impact of the research may be limited to people who use Uber services, not all citizens of the Chicago area (Wallsten, 2015). In recent years, the shared bicycle system in the Chicago area has developed rapidly. Many commuters first use public transportation to travel longer distances and then ride Divvy Bikes to help themselves reach nearby destinations. This shows that the shared bicycle system represented by Divvy Bikes has not only become a choice for people to travel, but also a supplement to public transportation (Zhou, 2015). However, the role of shared bicycles within a small area or a community needs to be further explored (Griffin & Sener, 2016). Furthermore, commercial space transportation is considered to be a future development tendency in Chicago transportation (Wakimoto, 2019), which requires international society to agree on technical standards. This can be an extended solution to Chicago's public transportation.

However, few studies have directly analyzed how current resources like public bicycles can improve the traffic situation in the Chicago area. For example, public bicycles have played in improving the traffic situation in the Chicago area and are precisely one of the most popular transportation for residents in the Chicago area. Also, the reason we chose to study this mode of transportation is that public bicycles are based on the current epidemic-based scenario, and the city of Chicago can maximize the use of existing resources to improve traffic conditions, which is more economical. Bike-sharing systems also have the advantage that they are environmentally friendly. We hope to maximize the effect of public bicycles so that they can improve the transportation experience for the citizens more effectively and with high quality.

There have been studies that predict the flow of bike-sharing systems in sub-regions (Yang et al., 2016). The high accuracy of the models proves that given certain information of a station, the usage of sharing bikes in that specific station can be foreseen. Nevertheless, no such researches have been conducted in Chicago. Therefore, we want to find whether some personal information of the rider, the station, and the time when the consumer rides the bike can contribute to the prediction of Divvy Bikes usage. Also, studies have shown that the usage of shared bicycles is different at different times, for example, in cities like New York City and Washington, D.C (Jia et al., 2020). But we do not know how the usage of shared bicycles in the Chicago area will change over time. Considering that Chicago is located in the central United States and New York City and Washington, D.C. is located in the northeastern United States, we cannot directly apply the changing laws of the usage of shared bicycles in New York City and Washington, D.C. to Chicago. Furthermore, we believe that Divvy Bikes product design will also affect people's use of shared bicycles, so we must understand the user experience of Divvy Bikes. As part of natural language processing technology, content analysis has been widely used in the formulation and analysis of business strategies for a long time (Furrer et al., 2008). We believe that applying content analysis technology can dig out the information behind the user feedback.

In this study, we want to analyze how public bicycles function as transportation means in people’s daily life, where Divvy Bikes, the biggest company focusing on public bicycles in Chicago, will be used as the representation to conduct the research, and see where the company can improve their products using users’ feedbacks as reference. We will explore how different factors influence people’s usage of Divvy bikes, and how the usage change with time going by. Also, we put emphasis on users’ feedbacks with sentiment analysis and communication networks, where we may find the current problems of the products and see how to improve them. We expect this study to provide a reference for the Chicago city government to support the shared bicycle market and Divvy Bikes company to improve their services.

Research Question

The purpose of our research is to find out which factors (including time factors) have a specific impact on the use of Divvy Bikes and use the model to make predictions. At the same time, we want to investigate the user experience of Divvy Bikes through user reviews to better promote Divvy Bikes.

1. Can the independent variables accurately predict Divvy Bikes usage? If not, when we only consider some of these variables, can they better predict the usage of Divvy Bikes? The independent variables include the time of use (including season, whether it is a weekend, month, day, and which day of the week, etc.), user gender, commercial plan (this nature mainly describes the user's account situation, and they may be Subscribers, Customer or Dependent), temperature and weather conditions on the day of riding (including cloudy, rain or snow, clear, thunderstorms or not clear). The dependent variables describe the usage of Divvy Bikes and include Divvy Bikes' daily usage and the average riding span each time.
2. Does the usage of Divvy Bikes show obvious changes over time? Can we predict the usage of Divvy Bikes in a certain period in the future?
3. In user feedback on Divvy Bikes, are there some high-frequency words that can reflect the user's attitude? Which words can highlight whether the user likes or dislikes Divvy Bikes?

Hypothesis

1. The independent variables used in this study can well predict the use of Divvy Bikes. If only some of these factors are considered, we can also get an accurate prediction model.
2. The usage of Divvy Bikes shows a clear change over time. We can summarize the

law of change based on existing data and use this law to predict the future usage of Divvy Bikes.

3. In the users' feedback, there will be some high-frequency words that are closely connected to Divvy Bikes. These high-frequency words will collectively reflect the advantages and current problems of Divvy Bikes.

Data

In the project, we will use off-the-shelf data ("Readymade") to complete data analysis (Salganik, 2018). The Readymade data used in this research come from the Kaggle Open Dataset Platform, the website of Divvy Bikes company, a website named Tripadvisor.

Chicago Divvy Bikes sharing data[1]

This data set includes nearly 950,000 records of consumers using Divvy Bikes from the beginning of 2014 to the end of 2017. Each record is the specific information used by Divvy Bikes once, including the cyclist's personal information, usage time, road condition information, weather conditions at that time, etc. We hope to analyze from this data set how different variables affect the average riding time of consumers and the average daily usage of Divvy Bikes.

Divvy trip history data[2]

These data sets include all usage records from the establishment of Divvy Bikes in 2013 to the present. Unlike Chicago Divvy Bikes' open data, these data sets include all consumption records of Divvy Bikes, but they do not include very detailed consumer information and weather conditions on the day of riding. These data sets include a total of more than 13 million user records. We aim to predict the future usage of Divvy Bikes with these records.

Reviews of Divvy bikes[3]

These data sets contain Divvy Bikes consumers' ratings and specific evaluations of the product. There are nearly 300 English samples. We hope to see from the consumer's analysis of Divvy Bikes what advantages and disadvantages Divvy Bikes

[1] Chicago Divvy Bicycle Sharing Data. (2020). https://www.kaggle.com/yingwurenjian/chicago-divvy-bicycle-sharing-data

[2] Index of bucket "divvy-tripdata". (2020). https://divvy-tripdata.s3.amazonaws.com/index.html

[3] Divvy Bikes Reviews. (2020). https://www.tripadvisor.com/Attraction_Review-g35805-d5074715-Reviews-or5-Divvy_Bikes-Chicago_Illinois.html#REVIEWS

has from the consumer's perspective. The results of the analysis will help Divvy Bikes improve their product quality.

Methods

This article will use different regression algorithms, time series algorithms, multivariate statistical analysis methods, and natural language processing techniques to analyze the data.

Influence of different factors on Divvy Bikes usage

In this section, we wish to analyze the impact of different factors on Divvy Bikes usage, including the impact on Divvy Bikes' daily usage and the average riding span each time, which are the dependent variables in this section. The independent variables we will analyze include: time of use (including season, whether it is a weekend, month, day, and which day of the week, etc.), user gender, commercial plan (this nature mainly describes the user's account situation, and they may be Subscribers, Customer or Dependent), temperature and weather conditions on the day of riding (including cloudy, rain or snow, clear, thunderstorms or not clear). The main methods we will use in this section are Ridge Regression and Principal Component Analysis.

Ridge Regression

Ridge regression is another regression algorithm developed by modifying the cost function based on standard linear regression in machine learning. Ridge regression is a biased estimation regression method dedicated to collinear data analysis. It is essentially an improved least squares estimation method. Ridge regression achieves a more realistic and reliable regression method by giving up the unbiased method of least squares at the expense of partial information and reducing accuracy. Ridge regression fits morbid data better than the least-squares. We can understand Ridge regression as an algorithm that seeks to balance between variance and bias. The objective function is:

$$\mathrm{J} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) \; - \; y_i)^2 + \lambda||w||_2^2$$

The first part is the cost function of the algorithm, and the second part is the length of the fitted coefficients. To write the objective function in this way is to achieve a balance, the error of the first fitting should be small, and the absolute value of the second coefficient cannot be too large. The reason why Ridge regression is used in this case is that many of the variables we are going to analyze are categorical variables. After we transform them into dummy variables, dependent variables will have strong collinearity. At the same time, we will analyze the impact of these variables on the two dependent variables. The purpose of our use of Ridge regression is to find a model that can estimate the daily usage and average riding duration of Divvy Bikes most accurately given these factors.

Principle Component Analysis

Principal Component Analysis, or PCA for short, is a statistical method. Orthogonal transformation is used to convert a set of variables that may correlate to a set of linearly uncorrelated variables. The converted set of variables is called the principal component. The purpose is to reduce the dimensional expression of meaningful sample point data appropriately. If it is reduced to 3 or 2 dimensions, it can be visualized. Among them, for the commonly used sample points, there must be data loss. The problem is how to reduce the loss to a relatively small amount. If some of the variables after dimensionality reduction can explain a high proportion of variance in the original variables, then we think that dimensionality reduction is effective, and the variables after dimensionality reduction can replace the original variables to help us analyze the data (Zou et al., 2006).

Specifically, first, we should normalize the sample attributes, calculate the covariance matrix of the sample vector set, and calculate the eigenvector matrix and eigenvalues of the covariance matrix. Finally, we multiply the proposed K feature vectors with the original sample set to obtain a new sample set after dimensionality reduction. If the fidelity of the new sample set will be high enough, we can use the new sample set after dimensionality reduction to continue the analysis of the data.

The purpose of PCA is to reduce dimensionality, not to return. In addition to compressing data and visualization, PCA can reduce the dimension during supervised learning to reduce the calculation burden and increase the calculation speed. This is the value of using Ridge regression and PCA in this case.

We believe that when all these independent variables are combined, the daily usage and average riding duration of Divvy Bikes cannot be well predicted. Therefore, we want to continue to explore which independent variables can better reflect the two Changes in the dependent variables.

Time series analysis of Divvy Bikes usage

A time series refers to a sequence in which the values of the same statistical indicator are arranged in order of the time when they occur. The main purpose of time series analysis is to predict the future flow of Divvy Bikes based on existing historical data. According to the different observation times, the time in the time series can be a year, quarter, month, or any other time form. Time series requires the use of a realistic and real set of data rather than experimental data, and the data is dynamic. The basic idea of the time series is to establish a mathematical model that can more accurately reflect the dynamic dependencies contained in the sequence based on the system's limited length of operating records and to predict the future of the system.

In this study, we will mainly use the prophet algorithm open-sourced by Facebook. Our research goals include two parts: to simulate the average riding time and daily usage of Divvy Bikes on specific dates from 2013 to the end of 2017; to predict the average riding time and daily usage of Divvy Bikes on specific days throughout 2018, And compare the predicted value with the real value. In summary, the dependent variables in this section are Divvy Bikes' daily usage and the average riding span each time, and the independent variable is the time (each day).

In this algorithm, we borrowed the method of the Decomposition of Time Series. This method divides the predicted value y_t of the time series into several parts, including the seasonal effect S_t, the trend term T_t, the holiday utility H_t and the remaining term R_t. The specific formula is as follows:

$$y_t = g(t) + s(t) + h(t) + r(t)$$

In the formula, $g(t)$ represents the trend term, which represents the changing tendency of the time series in a non-periodic period. $s(t)$ represents the periodic term, or seasonal term, which is generally in units of weeks or years. $h(t)$ represents a holiday item, indicating whether there is a holiday on that day. $r(t)$ indicates an error item or a residual item. The Prophet algorithm is to fit these items, and then finally add them together to get the predicted value of the time series (Guo et al. 2020).

Content Analysis on Divvy Bikes Reviews from Users

We believe that from the objective data, we can extract relevant information about citizens' use of Divvy Bikes and some factors that affect their use of Divvy Bikes. However, these datasets do not allow us to understand the consumer's subjective view of Divvy Bikes, nor does it let us know the specific experience of Divvy Bikes in use. Therefore, we used web crawling technology to obtain all available user reviews about Divvy Bikes in the Chicago area. We will use machine learning algorithms and natural language processing techniques to analyze these reviews. The two main research methods used in this study are: perceptron algorithms and Communication Networks.

Different perceptron algorithms

In this research session, we divided all the data into a training set, validation set, and test set. The ratio of these three data sets is 6: 2: 2. According to the product experience and the specific text content, we think that the text with a score of less than 4 is a negative evaluation, and the text with a score of 4 or 5 is a positive evaluation.

After dividing the data set, we apply the bag-of-words model to each evaluation and extract their word matrix from them. These word matrices will be used as variables to predict the sentiment of the text. In summary, the dependent variable in this section is the sentiment category of each observation, negative or positive. The independent variable is the matrix extracted from the texts using the bag-of-word method.

After extracting the dependent and independent variables for each text, we will use three perceptron algorithms including Perceptron, Average Perceptron, and Pegasos to analyze the text.

A perceptron algorithm uses a feature vector to represent a feed-forward artificial neural network. It is a binary classifier that maps the input (real value vector) on the matrix to the output value (a binary value). $\boldsymbol{w}$ is a vector of real weights, $\boldsymbol{w} * \boldsymbol{x}$ is a dot product. $\boldsymbol{b}$ is a fixed constant. $f(x)$ is used to classify and see if it is affirmative or negative. This is a binary classification problem. If $\boldsymbol{b}$ is negative, then the weighted input must produce a positive value and be greater than $\boldsymbol{b}$, so that the classification result is greater than the threshold $\mathbf{0}$. From a spatial perspective, $\boldsymbol{b}$ changes the position of the decision boundary. The formula $f(x)$ is:

$$f(x) = \begin{cases} 1, & \text{if} \boldsymbol{w} * \boldsymbol{x} + \boldsymbol{b} > 0 \\ 0, & \text{else} \end{cases}$$

The difference between the three perceptron algorithms used in this paper is mainly reflected in how to update the model coefficient vector θ according to the result of misclassification. For data ($\{(x^{(i)}, y^{(i)}), i = 1,2, \dots, n\}$, T), The update methods of these three algorithms are:

Perceptron: if $y^{(i)}(\theta * x^{(i)}) \leq 0, \theta = \theta + y^{(i)}x^{(i)}$

Average Perceptron: $\theta_{final} = \frac{1}{nT}(\theta^{(1)} + \theta^{(2)} + \cdots + \theta^{(nT)})$

Pegasos: $\theta = \begin{cases} (1 - \eta\lambda)\theta + \eta y^{(i)} x^{(i)}, & if\ y^{(i)}(\theta * x^{(i)}) \leq 1 \\ (1 - \eta\lambda)\theta, & \text{else} \end{cases}$, where η and λ are fixed parameters.

We will apply these three algorithms to the data set and tune their hyperparameters. Finally, we can select the algorithm that can obtain the highest accuracy on the test set as a tool to predict the sentiment of Divvy Bikes user feedbacks. This will help Divvy Bikes analyze the new user reviews in the future more quickly. And, with the final selected algorithm, we can also select the words that best reflect the positive or negative evaluations of users to help Divvy Bikes understand the user experience.

Communication Networks

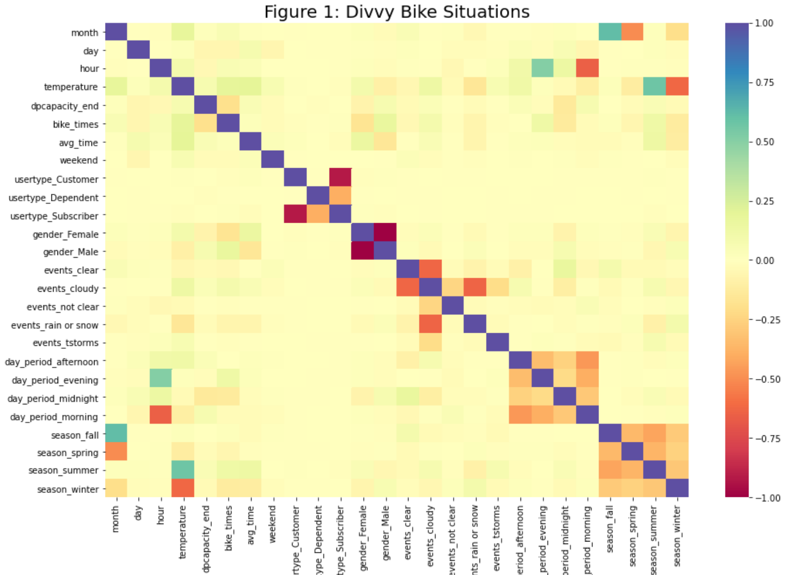
In this part, we will analyze the user's method for Divvy Bikes through the word network according to the semantics. We will construct the relationship between words based on their composition in the sentence. The goal is to understand the structure of how words are connected, and the dynamics of how their meaning flows through the discourse system. For example, we can extract links between subjects, verbs, objects, nouns, and adjectives (or verbs and adverbs that modify them) that modify them. In this part, we want to know which words always appear together, which words are always associated with Divvy Bikes and how strong. After obtaining different word networks, we will use a Graph to visualize these word networks.

Results

Influence of different factors on Divvy Bikes usage

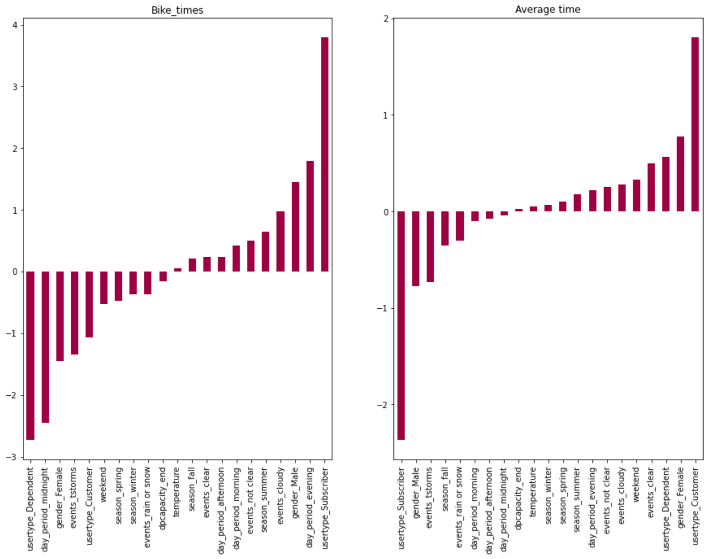
We calculated the covariance between different variables as shown in Figure 1. Most of the variables are related to Divvy Bikes' daily usage (variable name: bike_times) and the average riding duration (variable name: avg_time). Therefore, we infer that not all independent variables can well help us predict two dependent variables. If all variables are used for regression, even if we repeatedly train the algorithm, the accuracy that can be achieved is very low. This prompts us to use principal component analysis to determine which variables may be more important. From the perspective of data covariance, for the daily usage of Divvy Bikes, we see that the three seasons of winter, summer, and autumn have a more obvious impact on it, and on the same day, night, and midnight Obvious impact. Also, consumer gender, the temperature of the day, and the number of total docks at each station (variable name: dpcapacity_end) will also have a more obvious impact. For the average riding time, gender, summer, winter, rain and snow, and the temperature of the day may play a significant role.

Figure 1: Divvy Bike Situations



We use all the independent variables to perform Ridge regression on the two dependent variables, and the resulting coefficients are shown in Figure 2. However, as we predicted, the accuracy of the final model is very low, less than 15%. This validates our previous conjecture: not all factors are very important for predicting two dependent variables.

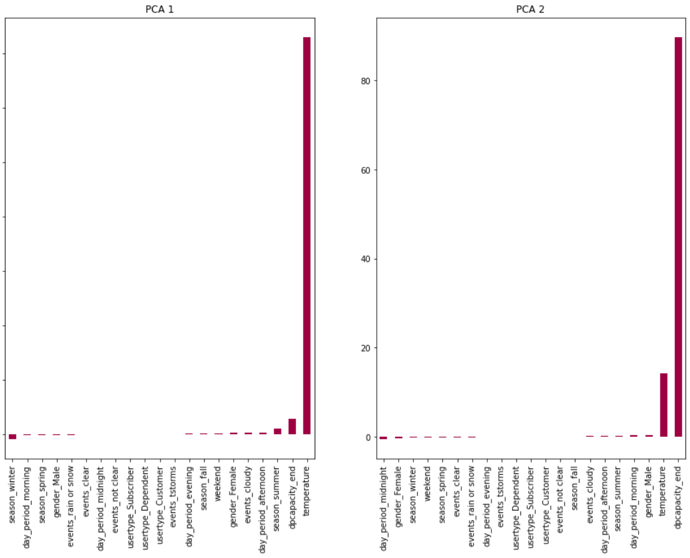
Figure 2: Ridge Regression on All Predictors



In response to this phenomenon, we use principal component analysis to see which variables are more important in the data. We found that the first Principal Component can explain more than 80% of the variance of the data, and the second Principal Component can explain about 11% of the variance of the data. Starting from the third principal component, the variance of the original data that they can explain is very small. Therefore, we believe that the first two principal components can represent the original data well. The coefficients of the different variables in the first two principal components are shown in Figure 3. We can see that in the first principal component, winter, summer, the number of total docks at each station and temperature is the most important feature. In the second principal component, temperature and the number of total docks at each station are the most important features. We suspect that these variables may be more important for predicting the two dependent variables.

Based on this result, we only selected these variables to perform regression on the two dependent variables. However, regardless of whether it is Ridge regression or principal component regression, the final accuracy of the model is still very low. Therefore, we guessed that we may have overlooked very important variables when building the model. For example, people may have different needs and preferences for Divvy Bikes in different regions. And we did not consider the influence of the region in the model.

Figure 3: PCA of All Predictors



Although we haven’t been able to build a model to predict the usage of Divvy Bikes more accurately, from the results of the existing data analysis, we can find that some factors have an important impact on the use of Divvy Bikes, so we decided to separate Explore the effects of different variables.

In Figure 4 and Figure 5, we studied the impact of people on the usage of Divvy Bikes on different days in a week and the usage of people at different hours on the same day. We found that there are a morning peak and evening peak used by Divvy Bikes every day, and the peak time is very close to the time people commute to work on weekdays. In the morning peak of weekdays, people use Divvy Bikes more than on weekends. In the evening, people use Divvy Bikes much less frequently and frequently. At the same time, we found that people have not used Divvy Bikes for a long time. This phenomenon shows that people may prefer Divvy Bikes in short-distance commuting.

Figure 4: Divvy Bike Usage on Different Days

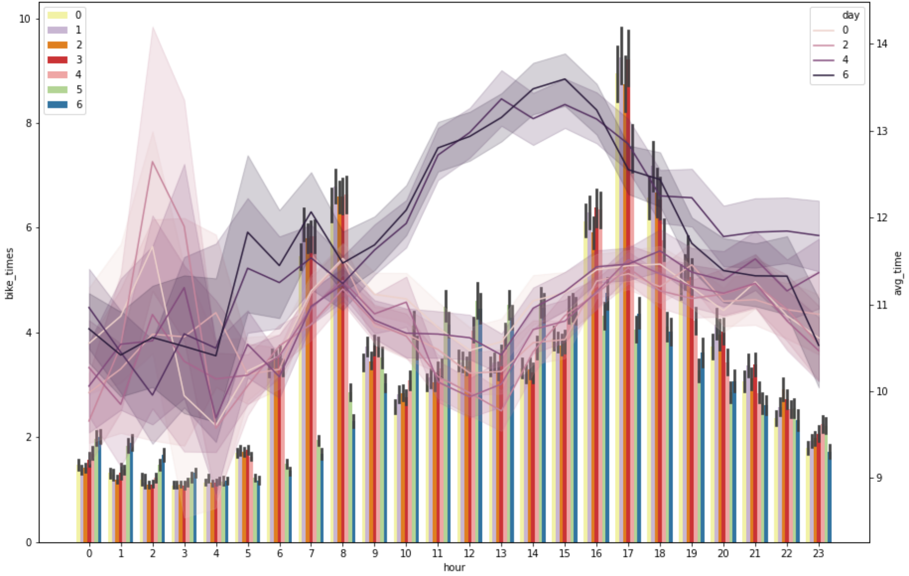
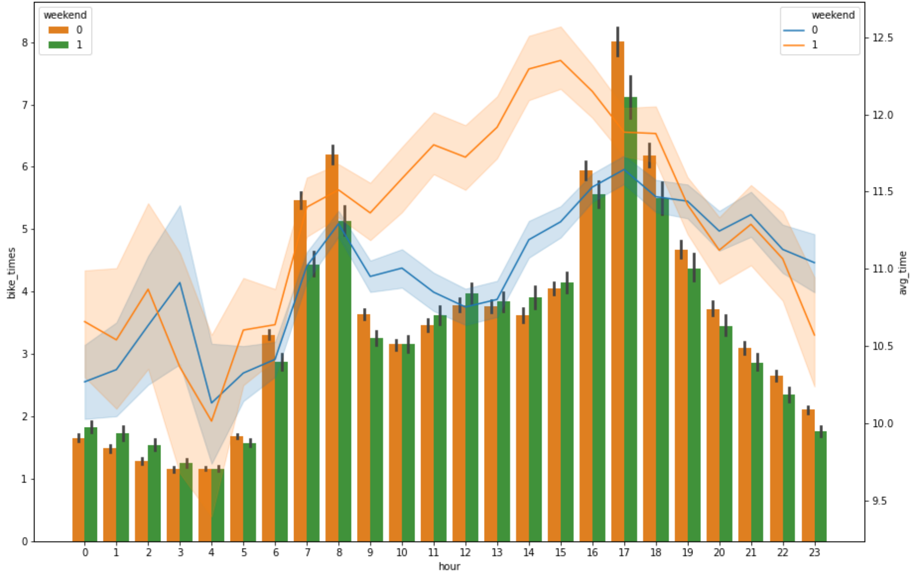
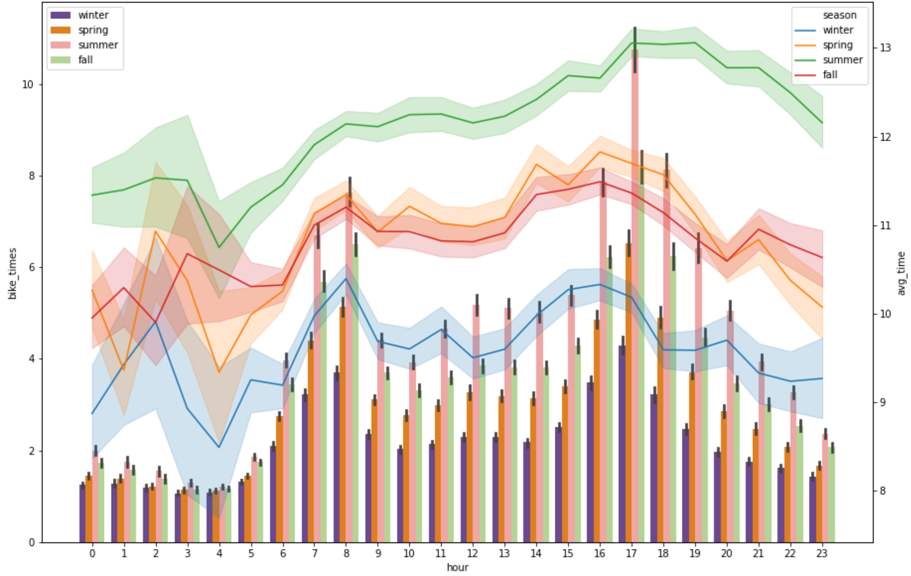


Figure 5: Divvy Bike Usage on Weekdays and Weekends



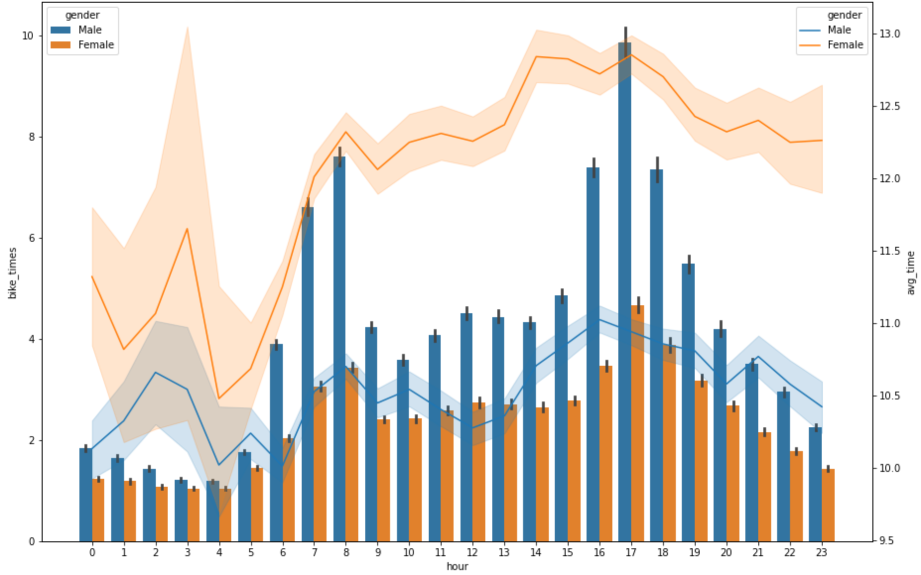
In different seasons, as shown in Figure 6, we found that the frequency and average riding time of people using Divvy Bikes in summer increased significantly, while the winter usage and riding duration were the least. In the spring and fall, people use Divvy Bikes more often in the fall, but the average riding time between the two seasons is not much different.

Figure 6: Divvy Bike Usage on Different Seasons



As shown in Figure 7, we found that men use Divvy Bikes more often, but the average riding time is significantly shorter than women. At different times of the day, the change in the number of rides by men is significantly greater than that by women.

Figure 7: Divvy Bike Usage on Different Genders



Through research on the different consumption plans chosen by consumers, we found that Subscribers use Divvy Bikes more frequently, but their riding time is relatively fixed, basically concentrated on short distances. The number of Dependents and Customers is relatively small, and the number of Divvy Bikes is significantly less. Nevertheless, when they use Divvy Bikes, their riding time is relatively long. This indicates that Dependents and Customers may only use Divvy Bikes centrally in certain situations.

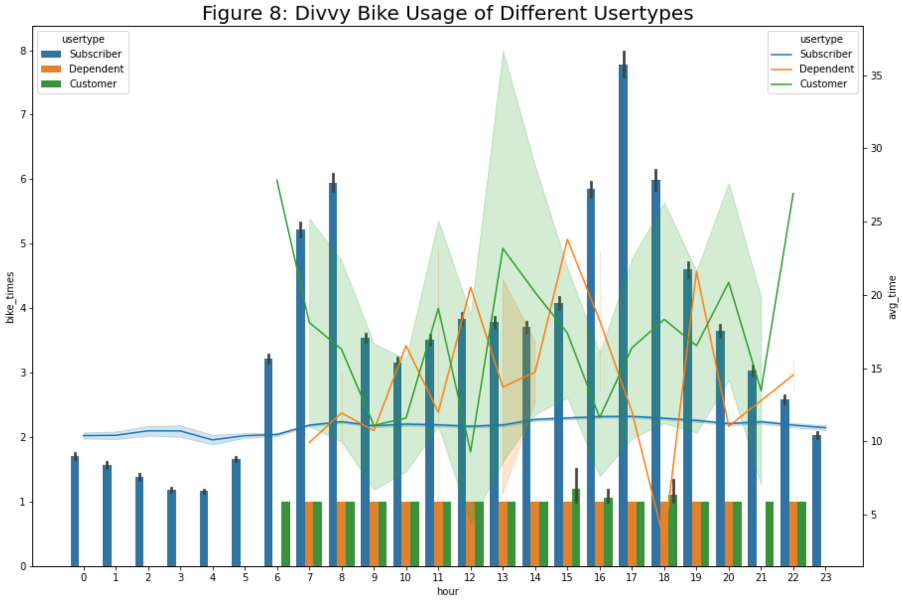


Figure 8: Divvy Bike Usage of Different Usertypes

Time series analysis of Divvy Bikes usage

By studying the influence of different factors on the two dependent variables, we found that these variables are not enough to help us accurately predict the usage of Divvy Bikes. Therefore, we hope to explore the changes in Divvy Bikes usage through time series. The advantage of the time series model, in this case, is that we can get people's demand for Divvy Bikes without considering the influence of different factors.

In Figure 9 and Figure 10, we predicted the use of Divvy Bikes in 2018. According to the model we have established, in 2018, the usage of Divvy Bikes will continue to rise steadily, but the average riding time of people remains at a fixed level.

At the same time, in Figure 11 and Figure 12, we can see how each component affects the usage of Divvy Bikes. We can find that the effect of different periods on Divvy usage in the time series model is very similar to the conclusion we reached in the previous regression phase.

We can see that the effect of the trend on people's average riding time is gradually stable, but as time goes by, more and more people will use Divvy Bikes. People use Divvy Bikes less often on weekdays. In summer, people prefer to use Divvy Bikes.

To verify whether the time series model we established can accurately predict the use of Divvy by Chicago citizens, we compared the predicted data with the factual data, as shown in Figure 13 and Figure 14.

Figure 9: Divvy Bike Usage Prediction (Times)

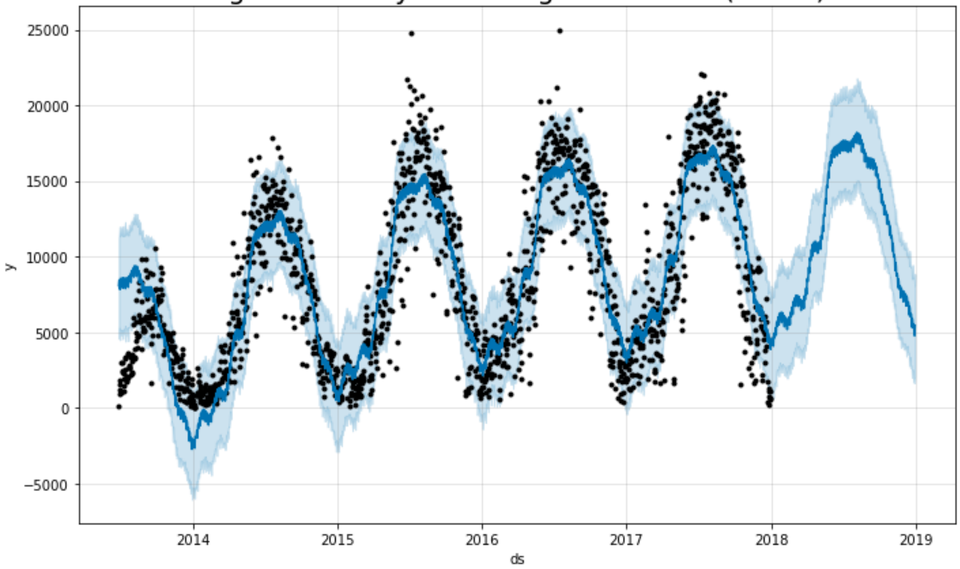


Figure 10: Divvy Bike Usage Prediction (Average Time)

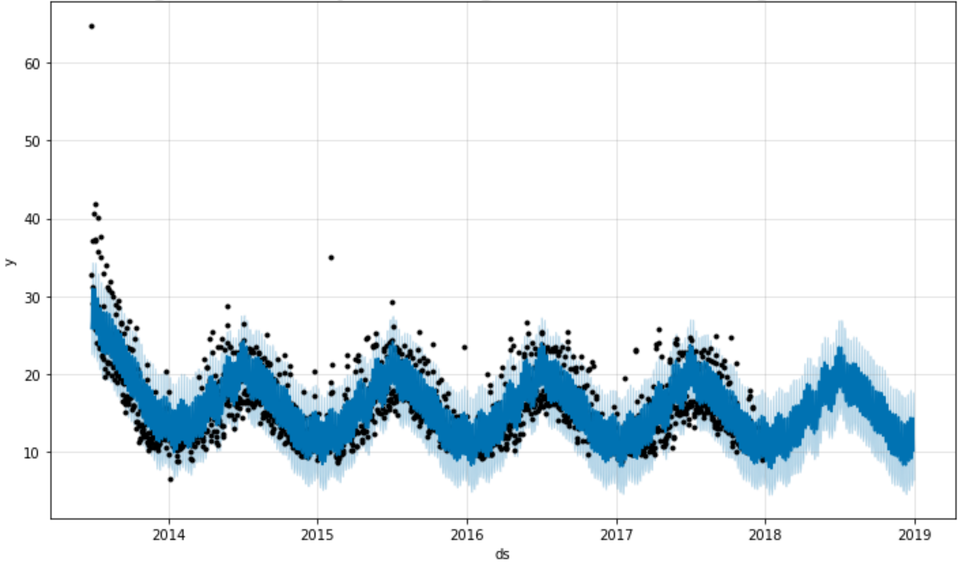


Figure 11: Time Series Components (Times)

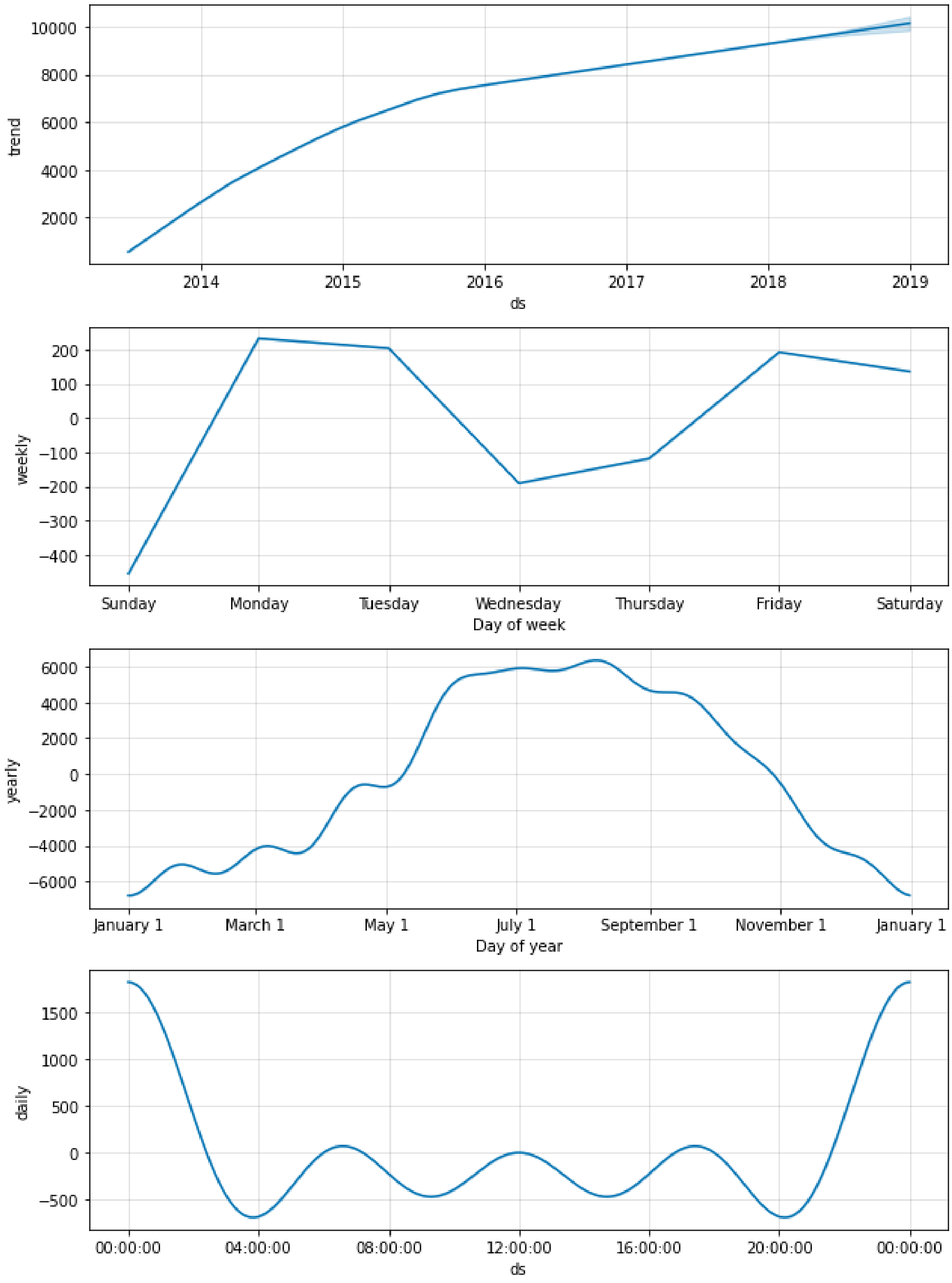
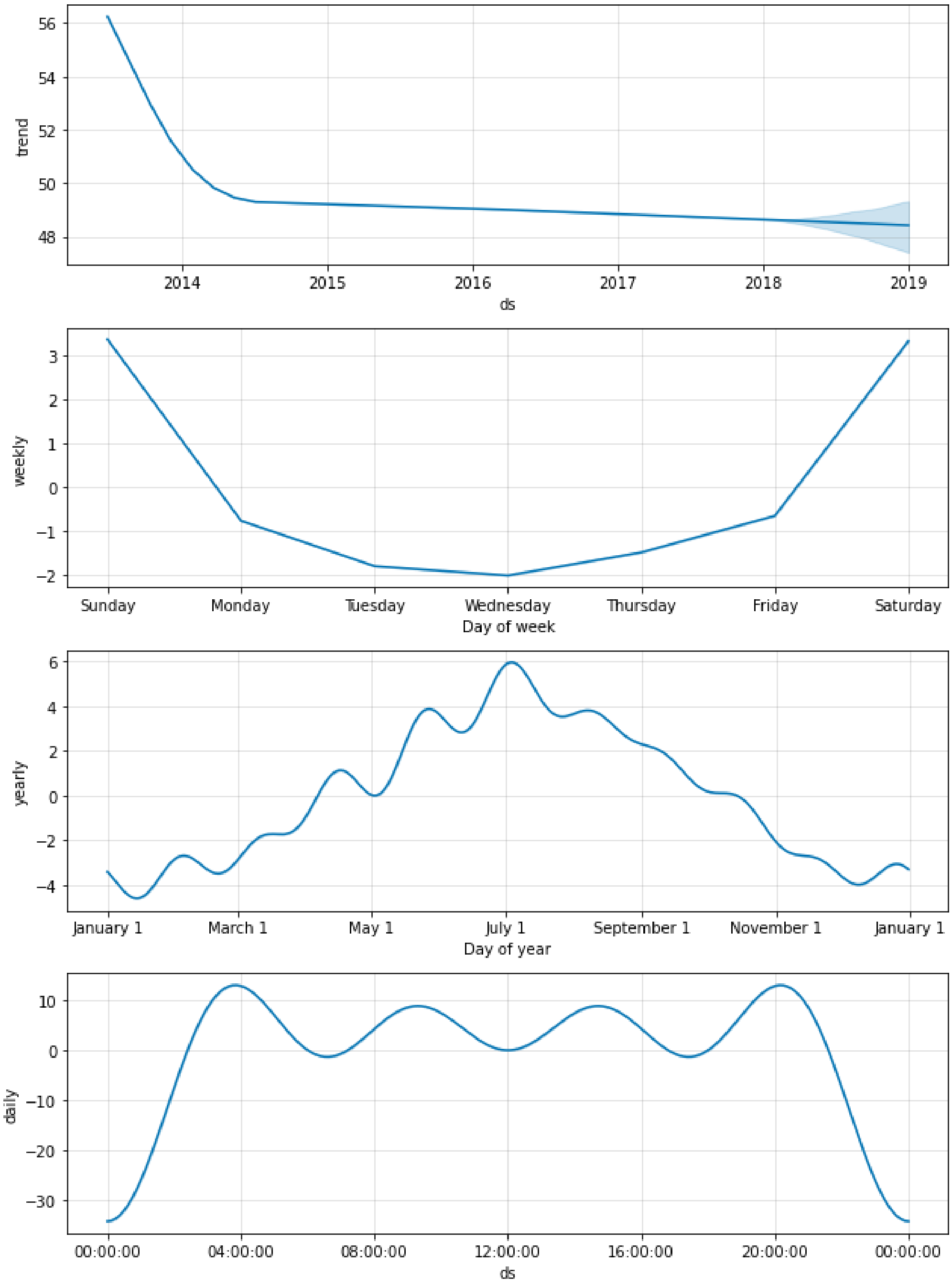


Figure 12: Time Series Components (Average Time)



We found that the time series model we established can predict the trend of people using Divvy Bikes. But in the actual situation, people's consumption data have greater variance. This is more obvious in predicting the average riding time of people. We can find that in 2018, the average riding time of people is higher than our prediction model as a whole, especially on some days from January to April, the average number of users is high. Therefore, we need to consider the impact of some special events on people's preference for Divvy Bikes when building time series models. For example, from January to April in 2018, Chicago may have witnessed some collective events, or the weather conditions during this time prompted people to ride bikes.

Figrue 13: Divvy Bike Usage in 2018 (Times)

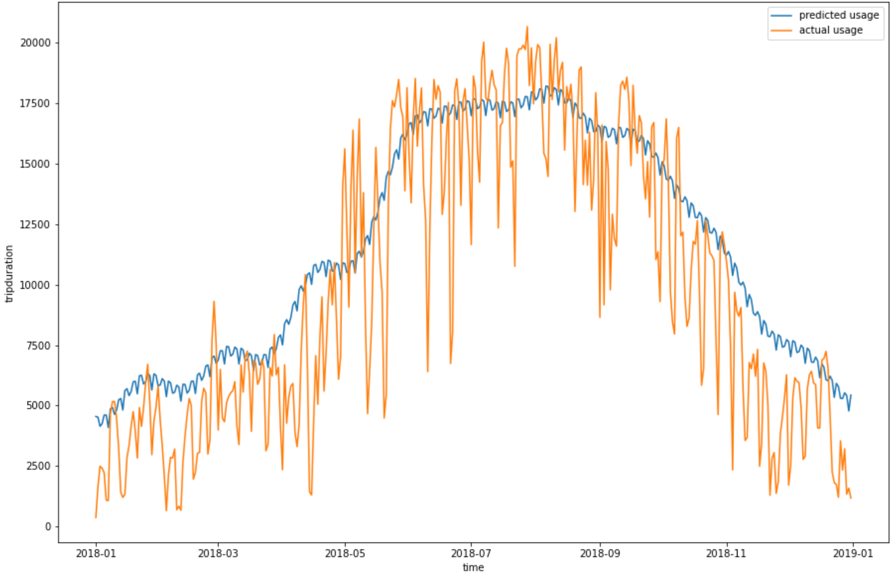
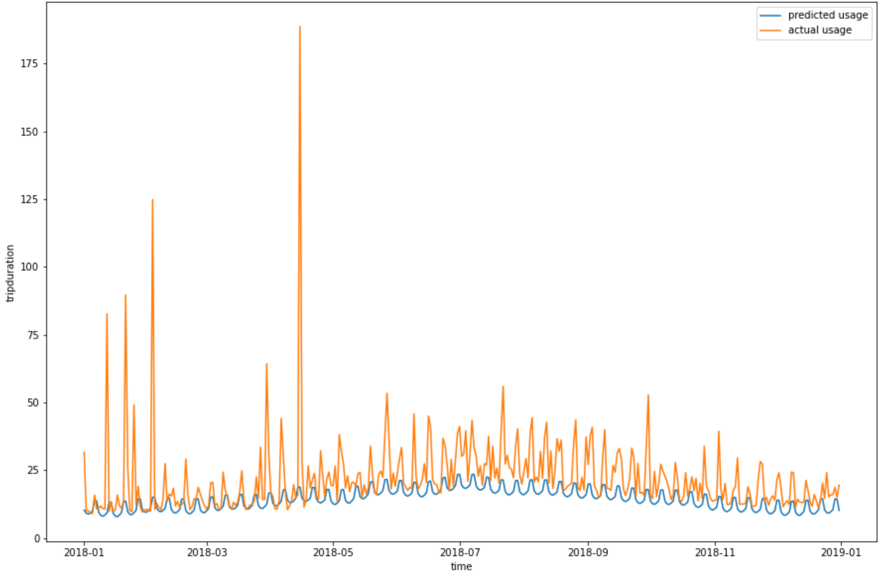


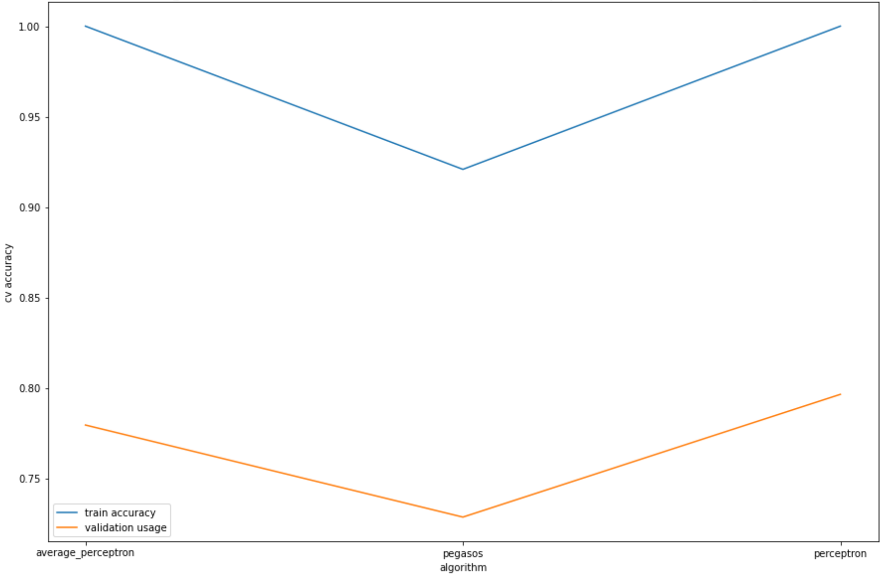
Figure 14: Divvy Bike Usage in 2018 (Average Time)



Content Analysis on Divvy Bikes Reviews from Users

We try to use the Perceptron algorithm, Average Perceptron algorithm, and Pegasos algorithm to classify the sentiment of reviews. Before training hyperparameters, we found that if these three algorithms are used, the accuracy of the model will be relatively high, as shown in Figure 15. Therefore, we conducted hyperparameter training on these three models, hoping to get higher accuracy.

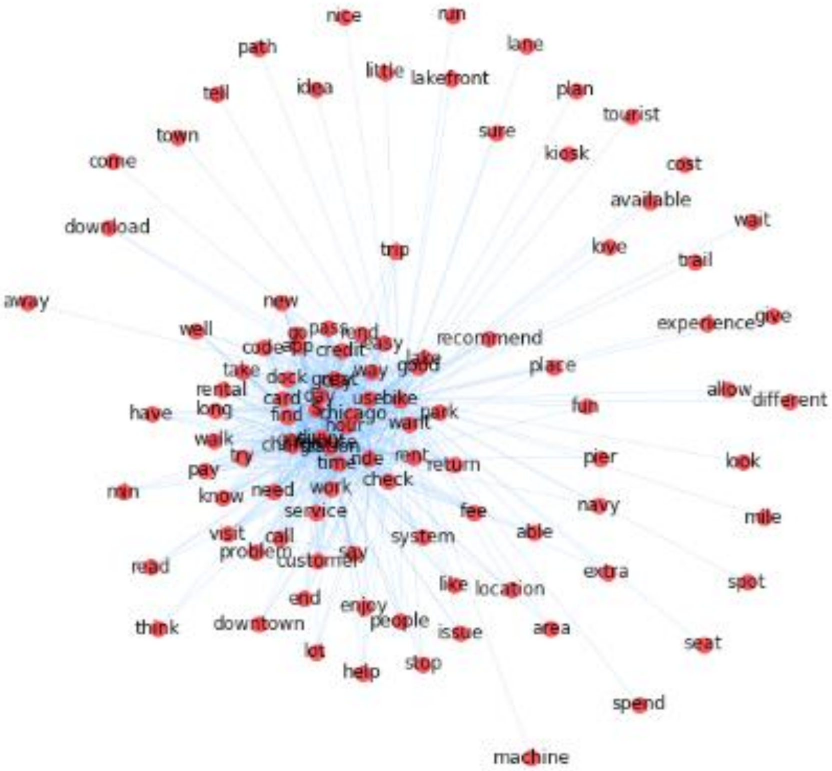
Figure 15: Accuracies of Different Preceptrons (before Tuning)



After tuning hyperparameter training, we decided to use the Average Perceptron algorithm. This algorithm achieved 75% accuracy in the test set data. According to the test results, we separated the ten words that best reflected positive comments, including ‘easy’, ‘way’, ‘app’, ‘park’, ‘around’, ‘times’, ‘rode’, ‘per’, ‘stations’ and ‘plan’. We found the ten words that best reflected negative comments including ‘another’, ‘customer’, ‘walked’, ‘docked’, ‘got’, ‘point’, ‘walk’, ‘said’, ‘told’, and ‘check’.

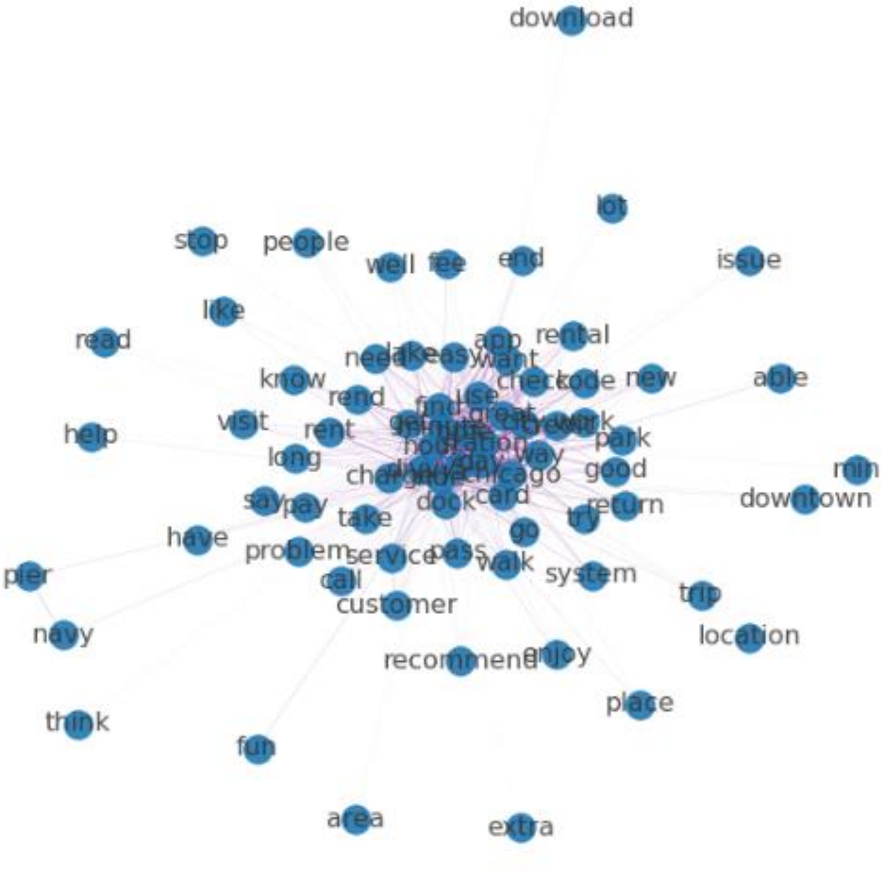
With communication networks, we picked out the core words in reviews, as shown in Figure 16. We can see that divvy and bike are in a very central position, and they are very closely related to things that are very related to the use of Divvy Bikes service, such as payment methods (‘credit’), usage methods (‘pass’, ‘app’, etc.), riding experience (‘lake’, ‘easy’, etc.). What we learned from this is that consumers are more inclined to decide whether they use Divvy Bikes based on the actual product experience. This point is also clearly shown in the specific text. Some consumers who gave negative reviews criticized Divvy Bikes for using a very fancy idea, that is, the sharing economy, but there are still problems in the specific billing method and bicycle quality.

Figure 16: Central Words in Reviews



We selected words related to bike from the entire communication network, and we hope to understand people's evaluation of Divvy Bikes more intuitively. From Figure 17, we can see that the words that are closely related to ‘bike’ are similar to the state presented in the complete communication network. These words cover all aspects of the product experience, such as usage, convenience, and related issues. According to these tips, we found that many negative reviews mentioned that there is a delay in the Divvy Bikes billing system, which leads to consumers often having to pay higher fees than they actually should pay. Besides, the app on the mobile phone of Divvy Bikes is not satisfying, where sometimes the consumers found it hard to get the necessary information. The locations of the stations in some areas are not reasonable. There are a large number of Divvy Bikes stations in some areas of Chicago, and these stations are often very close. Some consumers think that such an intensive design is not necessary. In other regions, it is difficult for consumers to find Divvy Bikes stations, which causes great inconvenience for them to stop using Divvy Bikes and complete their commute plan.

Figure 17: Central Words to bike



Discussion

Evaluation of the Project Design

Big Data in the Project

The good characteristics of big data include big, always-on, and nonreactive (Salganik, 2018). In the project, the datasets include huge volumes of data. For example, the first dataset includes 23 columns that each describe one feature of one observation of Divvy Bikes usage, and it contains nearly 9.5 million observations. This not only allows us to infer more accurate models but also allows us to explore some subtle differences. Furthermore, since we only used data from 2013 to 2018, the data will not change in the future. Finally, whether we do this project or not, the usage of Divvy Bikes in Chicago will not change.

The bad characteristics of big data include incomplete, inaccessible, nonrepresentative, drifting, algorithmically confounded, dirty, and sensitive (Salganik, 2018). Although these characteristics are sometimes unavoidable in social science research with big data, we have solutions to reduce their negative impact on this project.

1. In the first dataset, some specific information about the consumer and the specific day when the consumer used the Divvy Bikes is added. However, some information is lost. Therefore, not all observations from the complete records from Divvy Bikes company are contained in the first dataset. It does have the problem of incompleteness. Nevertheless, this dataset is the best one that we can get about the comprehensive information that we need, and the number of useful observations is still very huge. Consequently, this can well solve the problem of incompleteness.
2. The datasets are all open data that are accessible online. Therefore, we do not have the problem of inaccessibility.
3. Our datasets do not contain any personal information. Thus, they are not sensitive.

The Anticipated Internal/External Validity

Validity means how much the results of our project can deduce general conclusions. Internal validity focuses on if researchers conduct the experimental procedures properly, and external validity measures how much the conclusions of our project can be generalized to the population (Salganik, 2018). In this project, the internal validity is determined by how truthful the consumers of Divvy Bikes are. They are consumers confirmed by the system, so there is no question of authenticity. The external validity is determined by how we select unbiased samples of records of usage of Divvy Bikes and respondents to give feedback online. Since the first two datasets we selected are all data sets that can be obtained by observation, we believe that our results can be generalized. However, in the first data set, we do not know whether the data that was dropped because of the missing data is biased. Because the values of the two dependent variables in the first data and the second data are relatively close, we believe that the first data set is also unbiased. For reviews, because of the small number, we suspect a biased issue.

Evaluation of the analysis results

This research has achieved many constructive results that give the picture of the current development of Divvy Bikes, or bike-sharing market, in Chicago. We found many different specific periods, and the use of Divvy Bikes by Chicago residents will show regular changes. Based on the usage records from 2013 to 2017, we can more accurately predict the use of Divvy Bikes by Chicago residents in 2018. Through natural language processing technology, we can find that in user reviews, most of the comments on Divvy Bikes are positive, which can be matched with the fact that the usage of Divvy Bikes is increasing year by year. At the same time, we also found some problems that currently exist in Divvy Bikes based on the results of natural language processing analysis. This is very helpful for Divvy Bikes to determine the next development strategy.

However, this study also has some problems. First of all, we were not able to build a model that predicts Divvy Bikes usage. We think there are two main reasons:

1. We did not consider the influence of geographical factors. Under different location conditions, people's demand for public bicycles is different. We should incorporate the differences in socio-economic and geographical conditions in different regions into the prediction model.
2. Bike-sharing is a concept that has just emerged in recent years. Divvy Bikes was founded in 2013 and has been continuously expanding and growing in recent years. The use of Divvy Bikes by Chicago residents has not stabilized, so there are many disturbance factors.

In the time series model, to improve the accuracy of the model, we should add the seasonal factors of the Chicago area in 2018 (for example, the winter duration of the Chicago area is longer), holiday factors, and some special events to the model in more detail in.

In sentiment analysis and communication networks, we regret that the number of review samples used is less than 300. We believe that if there are more review samples, the analysis in this part will be more accurate and the results obtained will be more detailed.

The value of this study is mainly the analysis of Divvy Bikes in a detailed and targeted manner and the findings of the common problems in the current product design of Divvy Bikes. This study will provide a reference for Divvy Bikes to analyze the usage of Divvy Bikes under different location conditions in the future. However, in the future, this research can expand the dimensions. We hope to incorporate economic factors (such as the per capita income level of the surrounding region of a station) and spatial factors (location factors) into the variables that affect the usage of Divvy Bikes. Studies have proved that people in different geographic locations have different demands for shared bicycles. For example, citizens in Hangzhou, China, prefer to use shared bicycles at stations close to home or work (Shaheen et al., 2011). In the meanwhile, people with different economic conditions have different uses for shared bicycles. For instance, in Hangzhou, with personal income growth, people may use bicycles less, while in Washington, D.C., people with different income levels have no significantly different preferences in the use of sharing bicycles (Chen et al., 2015). Additionally, we hope to get more user reviews over time, making our content analysis results more convincing.

Conclusion

In this study, we found that the combination of variables such as time, road conditions, weather factors, user's gender, and account properties cannot accurately predict the average time Chicago residents use Divvy Bikes and the usage of Divvy

Bikes in a single day. But we found that if these factors are analyzed separately, we can find the usage pattern of Divvy Bikes in the Chicago area. People prefer to use Divvy Bikes on weekends, but whether it is weekends or workdays, the use of Divvy Bikes appears during commuting hours on weekdays. Among them, the usage of Divvy Bikes in the evening peak period is significantly higher than that in the morning peak period. Women use Divvy Bikes less frequently than men, but the average riding time is significantly higher than men. People especially like to use Divvy Bikes in the summer, but in winter, the use of Divvy Bikes is significantly reduced. In different account types, Subscribers use Divvy Bikes more frequently and relatively fixed. The average duration of use is not very long. It can be seen that they mainly use Divvy Bikes in daily short-distance commuting. Consumers and Dependents will focus on some Use Divvy Bikes in commuting situations where sex is relatively long.

The time series model can better predict the trend of Divvy Bikes' future usage, but the magnitude is slightly lower than the actual data in 2018. We found that the frequency of people using Divvy Bikes has increased over time, but the average riding time has remained at a stable level.

According to the users’ reviews, some popular words can reflect the users’ experience of riding Divvy Bikes, like ‘cheap’, and also describe problems of the current product, like ‘rent’ and ‘walk’. In sentiment analysis and communication networks, we found that most users are satisfied with Divvy Bikes products and services but also raised some representative questions. These affirmations and criticisms of Divvy Bikes mainly discuss the product quality of Divvy Bikes. This shows that the main solution for Divvy Bikes to expand its market share is to improve product quality, for example, perfecting Divvy Bikes app and adjusting the allocation of Divvy Bikes stations, rather than through marketing methods and price policies.

Reference

Addie, J. P. D. (2013). Metropolitics in motion: The dynamics of transportation and state reterritorialization in the Chicago and Toronto city-regions. *Urban Geography, 34*(2), 188-217.

Chen, L., Zhang, D., Pan, G., Ma, X., Yang, D., Kushlev, K., ... & Li, S. (2015, September). Bike sharing station placement leveraging heterogeneous urban open data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 571-575).

Downs, A. (2000). *Stuck in traffic: coping with peak-hour traffic congestion.* Brookings Institution Press.

Farmer, S. (2011). Uneven public transportation development in neoliberalizing Chicago, USA. *Environment and Planning A, 43*(5), 1154-1172.

Furrer, O., Thomas, H., & Goussevskaia, A. (2008). The structure and evolution of the strategic management field: A content analysis of 26 years of strategic management research. *International Journal of Management Reviews*, *10*(1), 1-23.

Griffin, G. P., & Sener, I. N. (2016). Planning for bike share connectivity to rail transit. *Journal of public transportation*, *19*(2), 1.

Guo, C., Ge, Q., Jiang, H., Yao, G., & Hua, Q. (2020). Maximum Power Demand Prediction Using Fbprophet With Adaptive Kalman Filtering. *IEEE Access*, 8, 19236-19247.

Jia, H., Miao, H., Tian, G., Zhou, M., Feng, Y., Li, Z., & Li, J. (2020). Multiobjective Bike Repositioning in Bike-Sharing Systems via a Modified Artificial Bee Colony Algorithm. *IEEE Transactions on Automation Science and Engineering, Automation Science and Engineering, IEEE Transactions on, IEEE Trans. Automat. Sci. Eng*, *17*(2), 909–920. https://doi-org.proxy.uchicago.edu/10.1109/TASE.2019.2950964

Jin, J., & Rafferty, P. (2018). Externalities of auto traffic congestion growth: Evidence from the residential property values in the US Great Lakes megaregion. *Journal of Transport Geography*, *70*, 131–140. https://doi-org.proxy.uchicago.edu/10.1016/j.jtrangeo.2018.05.022

Lomendra, V., Sharmila, P., Ganess, D., & Vandisha, N. (2018). Assessing the Causes

& Impacts of Traffic Congestion on the Society, Economy and Individual: A Case of Mauritius as an Emerging Economy. *Studies in Business & Economics, 13*(3), 230–242. https://doi-org.proxy.uchicago.edu/10.2478/sbe-2018-0045

Salganik, Matthew J., *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, 2018.

Shaheen, S. A., Zhang, H., Martin, E., & Guzman, S. (2011). China's Hangzhou public bicycle: understanding early adoption and behavioral response to bikesharing. *Transportation Research Record*, *2247*(1), 33-41.

Sukaryavichute, E., & Prytherch, D. L. (2018). Transit planning, access, and justice: Evolving visions of bus rapid transit and the Chicago street. *Journal of Transport Geography*, *69*, 58–72. https://doi-org.proxy.uchicago.edu/10.1016/j.jtrangeo.2018.04.001

Wallsten, S. (2015). The competitive effects of the sharing economy: how is Uber changing taxis. *Technology Policy Institute*, *22*, 1-21.

Wakimoto, T. (2019). Ensuring the Safety of Commercial Space Transportation through Standardization: Implications of the Chicago Convention and ICAO Standards. *Space Policy*, *49*. https://doi-org.proxy.uchicago.edu/10.1016/j.spacepol.2019.05.004

Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., & Moscibroda, T. (2016, June). Mobility modeling and prediction in bike-sharing systems. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (pp. 165-178).

Zhou, X. (2015). Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago. *PloS one*, *10*(10).

Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (n.d.). Big Data Analytics in Intelligent Transportation Systems: A Survey. (2019). *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 20*(1), 383–398. https://doi-org.proxy.uchicago.edu/10.1109/TITS.2018.2815678

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, *15*(2), 265-286.