Capital Bikeshare Strategy: Based on a Support Vector Regression Methodology

Mingyuhui Liu (Jane), Master of Science Candidate in Data Analytics

Advisor: Benjamin Harvey, PhD

December 7, 2017

School of Engineering and Applied Sciences

The George Washington University

Table of Contents

Table of Figures and Tables

Acronyms

atemp	Actual Temperature
CB	Capital Bikeshare
DC	Washington, D.C.
EC	Empirical Covariance
MCD	Minimum Covariance Determinant
MSE	Mean Square Error
OCSVM	One-class Support Vector Machine
R^2	R-Square
rbf	Radial Basis Function
Std	Standard Deviation
SVM	Support Vector Machine
SVM	Support Vector Regression
TTR	Training Vs Testing Ratio
UCI	University of California Irvine Repository

Abstract

In light of recent market penetrations from dockless bikeshare companies, the traditional bikeshare service provided by Capital Bikeshare (CB) in Washington, D.C. area is threatened. To address this problem and to increase the profit for CB, this research investigates the detailed hourly sharing bike data from CB in a 2-years' period (January 2011 to December 2012) in Washington DC. The research takes a Support Vector Regression (SVR) approach to find the best-fitted model to predict hourly rental count based on weather, season, holiday information, month, etc. After outlier detections, model comparisons and parameter tunings, an SVR model, with dependent variables of actual temperatures, working days, hours, weather situations, seasons, months, and windspeeds, is selected. Based on the model, demand elasticities from different consumer groups are calculated, from which business strategies are provided for CB.

This paper provides the detailed model selection analysis, and the reasoning behind the business strategy suggestions.

Introduction

As of 2015, Washing, D.C. (DC) resided more than 672,000 people, while welcomed more than 22 million visitors (DC Press, 2017). Such huge population flow put burdens on the transportation systems. Many people started to rely on bikeshare services given its economic and convenient nature.

Capital Bikeshare (CB), starting operation in September 2010, is a bicycle sharing company that serves Washington, D.C. and several surrounding counties. It has more than 440 stations and 3,700 bicycles. As of July 2016, CB provided different options for instant rentals at the station: riders may purchase a single trip ($2), a 24-hour pass ($8), or a 3-day pass ($17) (Capital Bikeshare, 2017). Figure 1 shows an example of a station inside DC, at Eastern Market Metro, 8th St & Pennsylvania Ave SE.

Figure 1. Capital Bikeshare station. This figure shows an example of a station. (Ben Schumin, 2010)

In September 2017, dockless bike-sharing services started to penetrating DC's bikeshare market (Dent, S., 2017). Its convenience for instant renting and returning has brought threats to the traditional bikeshare businesses. To address the threats, this research provides possible solutions for a profit growth for CB. By looking into the detailed hourly bike sharing data, this project serves the goal from the following aspects: a. to find a best Support Vector Regression (SVR) model to predict the hourly rental counts; b. to extract business information from the model.

With the emerging threats from the recent smart bike-sharing services in DC area, the results from this research can be of great value for stakeholders from CB. And the general approach to address such threats used in this research can be a good example for other traditional bikeshare business.

Data Description and Pre-processing

Descriptive Statistics

CB's rental data is extracted from University of California Irvine Repository (UCI) (UCI, 2013). The original data contains 17,379 instances, spanning in 17 columns, and is stored as CSV format. The rental data ranges from January 2011 to December 2012, and is integrated with other useful information, such as holiday, weather situation, temperature, windspeed, etc. Table 1 explains the definitions of each column, and Table 2 shows briefly the first few examples from the original dataset.

Table 1

Definitions of Columns in the Original Dataset

Column Name	Explanation
instant	Record index
dteday	Date
season	1:spring, 2:summer, 3:fall, 4:winter
yr	0: 2011, 1:2012
mnth	month (1 to 12)
hr	hour (0 to 23)
holiday	0: not a holigay, 1: holiday
weekday	Day of the week
workingday	0: either weekend or holiday, 1: otherwise
weathersit	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist Cloudy, Mist Broken clouds, Mist Few clouds, Mist 3: Light Snow, Light Rain/Thunderstorm, Light Rain 4: Heavy Rain, Ice Pallets, Thunderstorm with Mist, Snow with Fog
temp	Normalized temperature in Celsius. Function: (t-t_min)/(t_max-t_min)
atemp	Normalized feeling temperature in Celsius. Function: (t-t_min)/(t_max-t_min)
hum	Normalized humidity
windspeed	Normalized wind speed
casual	Count of casual users
registered	Count of registered users
cnt	Count of total rental bikes including both casual and registered

Table 2

Examples of First 7 Rows from the Original Dataset

instant	dteday	season	yr	mnth	hr	holiday	weekday
1	1/1/2011	1	0	1	0	0	6
2	1/1/2011	1	0	1	1	0	6
3	1/1/2011	1	0	1	2	0	6
4	1/1/2011	1	0	1	3	0	6
5	1/1/2011	1	0	1	4	0	6
6	1/1/2011	1	0	1	5	0	6
7	1/1/2011	1	0	1	6	0	6

Examples of First 7 Rows from the Original Dataset (Continued)

workingday	weathersit	temp	atemp	hum	windspeed	casual	registered
0	1	0.24	0.2879	0.81	0	3	13
0	1	0.22	0.2727	0.8	0	8	32
0	1	0.22	0.2727	0.8	0	5	27
0	1	0.24	0.2879	0.75	0	3	10
0	1	0.24	0.2879	0.75	0	0	1
0	2	0.24	0.2576	0.75	0.0896	0	1
0	1	0.22	0.2727	0.8	0	2	0

Examples of First 7 Rows from the Original Dataset (Continued)

cnt
16
40
32
13
1
1
2

Given the purpose of this research, many columns are dropped and several new columns are created. After the cleaning and reorganizing the data, 10 variables are kept, including the dependent variable. Table 3 gives a brief description with counts, means, standard deviation (Std), minimum and maximum.

Table 3

Description Statistics for Cleaned Data

	season	hr	workingday	weathersit	temp
Mean	2.50	11.55	0.68	1.43	0.50
Std	1.11	6.91	0.47	0.64	0.19
Minimum	1	0	0	1	0.02
Maximum	4	23	1	4	1.00

Description Statistics for Cleaned Data (Continued)

	atemp	hum	windspeed	cnt	month
Mean	0.48	0.63	0.19	189.46	12.57
Std	0.17	0.19	0.12	181.39	6.88
Minimum	0.00	0.00	0.00	1	1
Maximum	1.00	1.00	0.85	977	24

Noted that “month” has been recoded from 1 to 24 representing months from January 2011 to December 2012, which is why the variable “yr” from the original dataset is dropped.

To understand the data more comprehensively and to grasp potential correlations among the variables, a matrix of graphs has been created. The matrix is shown in the following Figure 2.

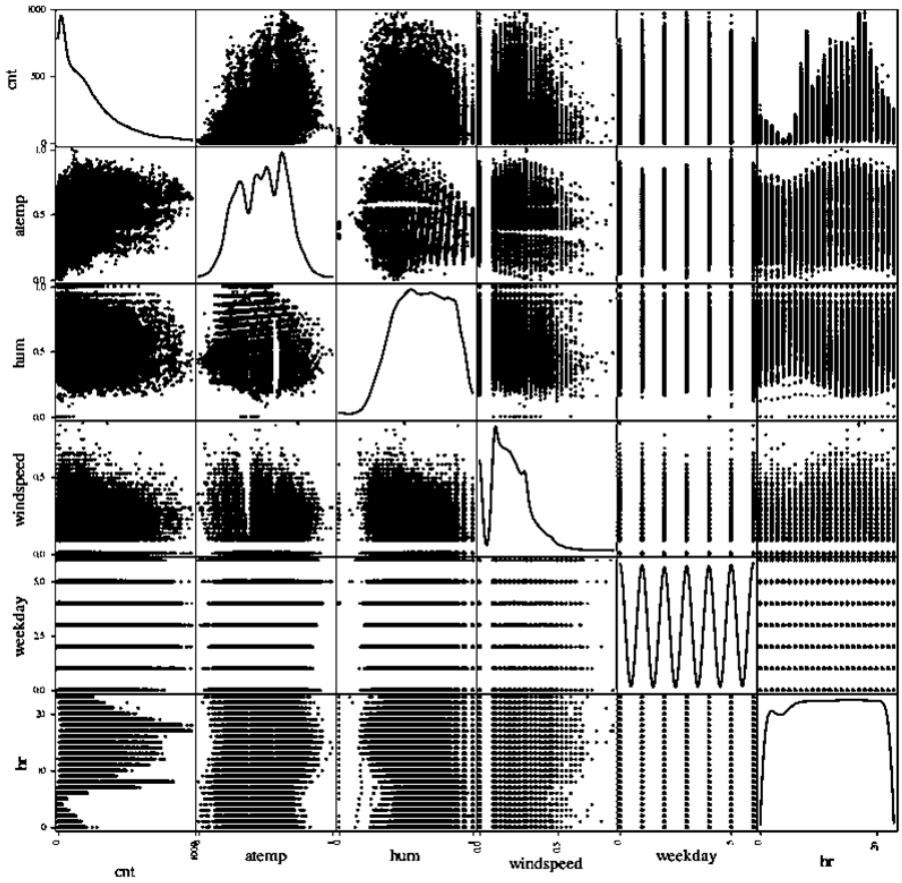

Figure 2. Scatter matrix for cleaned data. This figure shows the general idea of correlations among variables.

From Figure 2, it is obvious there are patterns between rental counts and other variables, however, without clear linear correlations. Another takeaway from Figure 2 is that there are both continuous variables and ordinal variables. However, since this research is using SVR, it is reasonable to treat all the variables as continuous, meaning that there is no additional processing for ordinal variables.

Outlier Detections

Outlier detections are crucial for accurate and efficient model construction. To choose the best outlier detection method for the dataset, this research designs 3 outlier detection methods: empirical covariance (EC), minimum covariance determinant (MCD), and one-class Support Vector Machine (OCSVM) (Scikit-Lear, n.d.). Figure 3 shows one of the examples this research conducts on the dataset, with rental counts over normalized temperature. Noted that the contaminations set for the EC and MCD methods are both 0.01, which indicates 1% of outliers should be detected.

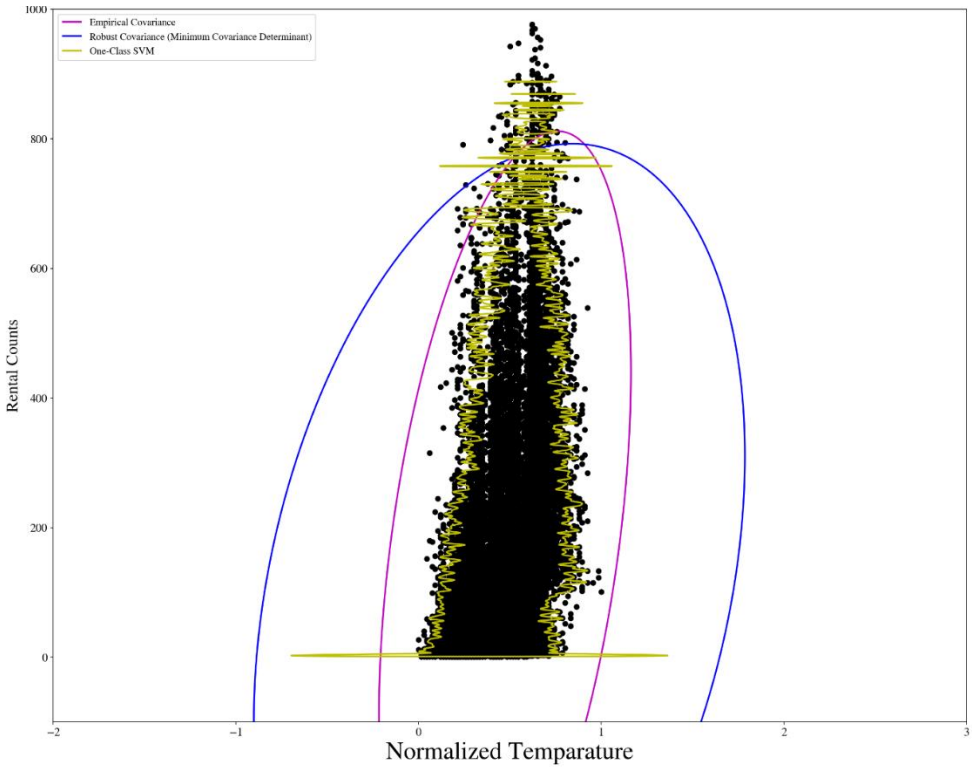

Figure 3. 3-way outlier detections on rental counts over temperature. This figure shows the outlier detections for the dataset.

From Figure 3, both EC and MCD are only able to detect the highest 1% of rental counts, while OCSVM has a more sensitive cut over the x-axis.

Considering the dependent variable in the dataset, which is the bikeshare rental counts in DC, outliers should only be taken out after close-up detailed investigation on each ruled-out data point. The rationale behind this lies in that there are possibilities that the rental counts can change dramatically because of occasional special events in the city. Therefore, EC and MCD are deemed to fail to classify the outliers since they only detected large rental counts, and OCSVM is complicating the outlier elimination. To make it more complicated, if any of these 3 methods were used, it would force the research to conduct the detection on rental counts over every single potential explanatory variable.

Given the reasons in the previous paragraph, and to simplify the outlier detection process, this research finally uses Median Absolute Deviation (MAD) method to conduct outlier detections on the dependent variable (Ley, Klein, Bernard & Licata, 2013). In the end, a total of 267 outliers are detected, based on the equation for calculating MAD from Figure 4.

$$MAD = \frac{\sum_{i=1}^{n} | xi - \bar{x} |}{n}$$

xi : Performance Value for Period i
$\bar{x}$: Average Value
n : Number of Data

Figure 4. Equation for MAD. This figure shows the equation for calculating MAD (ASPROVA, n.d.).

However, even though MAD simplifies the outlier detection process significantly, it is still a huge burden to conduct detailed investigation on the 267 outliers individually. To further simplify the situation, this

research fits both the datasets with and without outliers the same model, and compared the results to decide whether to take out the outliers.

SVR and Model Selection

Support Vector Regression Methodology

The SVR is performed with Python, SKLearn package, over the cleaned dataset. The steps for SVR include variable selections, dividing the cleaned dataset into training set and testing set for regression, kernel selection and validation.

Variable Selection. The first step to perform a SVR is to decide which independent variables to use in the model.

Recalling that there are 9 variables to be featured in the model potentially: temperature (temp), actual temperature (atemp), season, hour (hr), working day or not (workingday), weather situation (weathersit), humidity (hum), windspeed, month. Based on common sense, the research takes atemp, weathersit, workingday and hr as the main variables of interest. That being said, the model selection process is built up on these main variables. That leaves month, windspeed, season and humidity to be picked additionally.

In order to select the variables, the research first builds a list in Python for the main variables. And then a function to randomly choose random numbers of variables from the remaining variables is created. This function allows Python to loop through all the combinations on variables and to conduct a thorough SVR on all possible combinations.

Training Vs. Testing. Normally, with the increase of the training size in SVR, the accuracy of the predictions would increase. However, there might be abnormalities, and sometimes there would be a turning point where the increase of the training size will cause a decrease in the accuracy. Thus, this research compares different training Vs testing ratios (TTR) in order to choose the best ratio for training.

In order to select the training ratio, this research designed a function to allow SKLearn to loop through 0.12 to 0.33 ratios to perform the SVR.

Kernel Selection. The research uses kernel for SVR. In order to choose a better model, the research tests "linear" and "Radial Basis Function (rbf)" kernels, by embedding a loop for kernels inside the TTR selection function.

Noted that the loop for kernel selection needs to be inside the loop of the training ratios. That being said, the TTR needs to be chosen first and use that TTR to perform both linear and gaussian regression. Since every time when a new loop for TTR begins, it will allow SKLearn to randomly choose a certain ratio of training data from the whole dataset, which will cause bias if the kernels are chosen from new TTR loop.

Validation. The research uses Mean Square Error (MSE) and R-squared (R^2) to decide which model is performing better. The MSE is calculated following Equation 1.
Equation 1

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

n = size of the dataset;

$\hat{Y}_i$= predicted dependent variable;
Y_i= true dependent variable.

MSE checks the accuracy of the predicted data by mapping the distances between the predicted data and the true data, while R^2 represents the percentage of variances the model can represent.

Model Selection

Based on the methodology described above, the research chooses 7 variables as the dependent variables set: actual temperature, weather situation, working day, hour, season, month, windspeed, with a TTR of 0.12 and an "rbf" kernel, which means the testing data contains 12% of the whole dataset. After the initial model selection, the research runs the same model again on the dataset without rental count outliers, which are taken out by MAD method. Equation 2 is a quick reference for the model selected.

Equation 2

Rental Count = F (actual temperature, weather situation, working day, hour, season, month, windspeed);

F represents an SVR.

Initial Model Selection. Table 4 shows the validations of the chosen model.

Table 4

Validation for the Chosen Model with different TTRs and Kernels.

Model	TTR	rbf		linear	
		MSE	R^2	MSE	R^2
	33.33%	14561.58	0.5478	22953.20	0.2872
	25.00%	14009.37	0.5676	23207.64	0.2837
'temp', 'weathersit', 'workingday', 'hr', 'season', 'month_all', 'windspeed'	20.00%	13496.47	0.5770	22915.55	0.2818
	16.67%	13357.53	0.5810	22976.38	0.2792
	14.29%	13326.43	0.5846	23303.85	0.2736
	12.00%	13309.08	0.5870	23140.17	0.2749

MSEs are abnormally large, which might result from the un-logged dependent variable. Recalling Equation 1 for MSE, since rental count is not logged, which can result in huge $(\hat{Y}_i - Y_i)^2$ values.

Based on R^2, 12.00% TTR is chosen with "rbf" kernel. This model returns a 13309.08 in MSE and 0.5870 in R^2, which indicates a 58.70% of variances can be represented by the model.

Outliers Vs No-outliers. When runs the model again on the dataset without outliers, the research obtains MSE and R^2 as 11431.54 and 0.5830 respectively. Since the un-logged rental counts might contribute to the large number of MSE, the research favorites a larger R^2. That being said, the initial model with outliers included is chosen.

Parameter Tuning

Equation 3 represents the finally chosen model, which is ready for parameter tuning to obtain a better representation on the dataset.

Equation 3

Rental Count = SVR(kernel='rbf', degree=3, gamma=auto, coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=-1)

C = penalty parameter C of the error term;
gamma = Kernel coefficient for 'rbf', 'poly' and 'sigmoid'. If gamma is 'auto' then 1/n_features will be used instead. (Scikit-Lear, n.d.)

For parameter tuning, the research uses GridSearchCV in Python. Parameters of interest in this research include C and gamma. The research performed a parameter tuning to find the best C and gamma for this model. The alternative C and gamma are set as 10 and 0.1 respectively.

In the end, both C and gamma values are using the alternative values provided. Namely, the new SVR model is as Equation 4.

Equation 4

Rental Count = SVR(kernel='rbf', degree=3, gamma=0.1, coef0=0.0, tol=0.001, C=10.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=-1)

The MSE and R^2 for the new model bear the values of 6646.88 and 0.7937, which features a 50.06% decrease in MSE and a 35.28% increase in R^2. The model's efficiency and accuracy increases after the parameter tuning.

Results and Discussion

Results Visualizations

Unlike the linear regression, SVR cannot provide a clear linear equation for the rental counts and 7 explanatory variables. Instead, SVR predicts. That being said, a closer look at the relationships of between the rental count and each independent variable is needed.

Recall Equation 2 for the model selected. Figure 5 shows a matrix for 2-dimensional demonstrations of the coverage of the predictions from the model of 6 out of 7 independent variables over the testing rental counts.

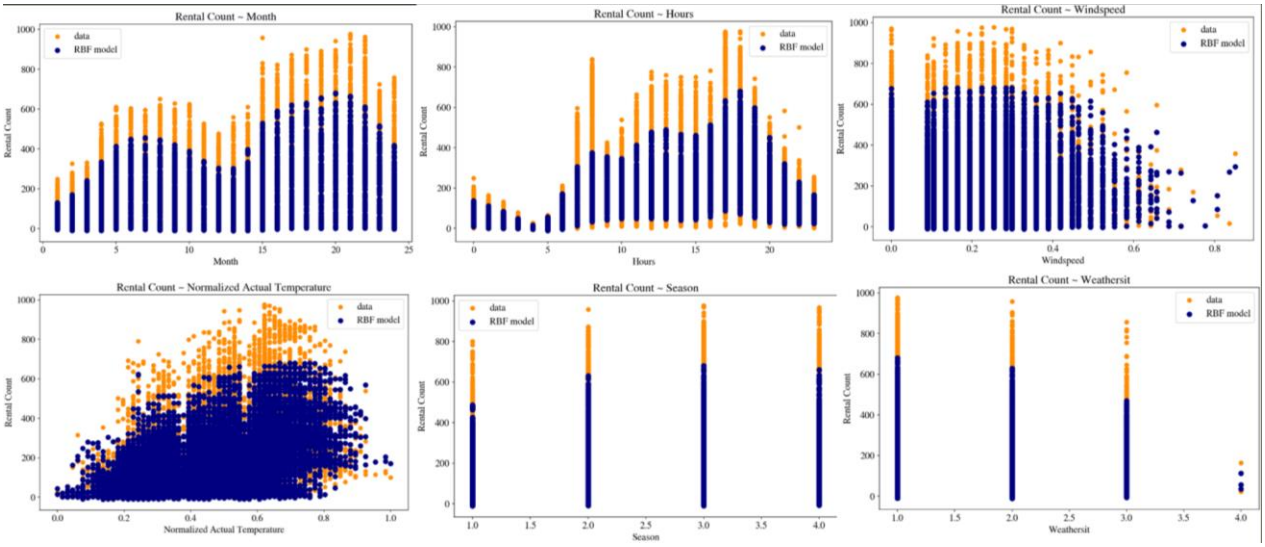

Figure 5. 2 by 3 predicted data plotted with testing data. This figure shows the model prediction coverage over the testing data.

From the matrix above, there is still no clear simple linear correlations between any of the independent variables and the rental counts. But certain patterns in several explanatory variables can be spotted. For example, the rental counts vary from hour to hour, and the peak is shown during both morning and evening rush hours. And as the coverage shows, the model failed to grasp the blast of the rental count during the morning rush hour. Another pattern could be spotted with the rental counts over month and weather situation. When the months locate in winters, or the weather situations get worse, the rental counts drop. These patterns are expected and in line with the common sense.

Furthermore, Figure 6 shows the predicted data coverage over the testing data between rental count and workingday.

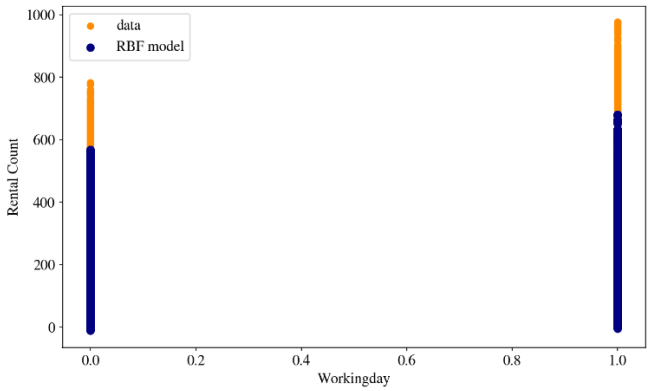

Figure 6. Rental count over workingday predicted data plotted with testing data. This figure shows the model prediction coverage over the testing data.

From the Figure 6, the pattern observed from the workingday is not in line with the common sense. From the figure, it is seemed that when the day is not a working day, the rental counts have a higher mean. However, given the high volumes of visitors in DC, it might be another way around. And it is presumed that the commuters would prefer more reliable transportation, such as metro and bus, other than the bike rentals from a dock which you will not be certain if there is bike or dock left at the precise location. This is the information worthy of further investigation.

Rental Count over Workingday

To examine if there is truly a difference between the workingday and non-workingday, the research conducts a one-way ANOVA on the data. Null hypothesis in the ANOVA test is that the mean of the rental counts grouped in workingday (workingday = 1) equals to the mean of the rental counts grouped in non-workingday (workingday = 0).

The research utilizes Scipy in Python to conduct the ANOVA test. A p-value of 6.52E-05 is returned by the ANOVA, which means it is statistically significant that the null hypothesis should be rejected. Namely, there is indeed difference between rental counts grouped by workingday and non-workingday.

Other ANOVA tests are also run to examine whether there are discrepancies between rental count grouped with workingday over other ordinal variables. For example, when rental count grouped by both workingday and season, the ANOVA shows 2.65E-10 significance, which again implies differences exist.

Noted that the research does not concern about which one group truly has a higher rental count mean, all it needs is whether there is difference, so that business strategy suggestions could be made based on this fact that they are different.

Business Strategy Discussions

Assumptions and Rationales

As discussed in the previous section, the rental counts grouped by workingday shows different patterns over other variables.

When considering the ANOVA test run on the rental count grouped by both workingday and season or wheathersit, the differences shown from the working day rental count vs non-working day rental count inspires the research to take into account the concept of demand elasticity.

First assume that workingday renters are mostly from commuters, and non-working renters are from tourists. This assumption is fair given a common sense, which can be illustrated by the following Figure 7, which is a photo taken in a busy business district during a weekend. It implies less commuters during non-working day.

Figure 7. A usually busy district during the weekend in DC. This figure helps explain the rationale behind the assumption that during the non-workingday less commuters are in DC.

From this first assumption, the research treats the working and non-working rental counts as commuter and tourist rental counts. Secondly, when it comes to the differences between traveling and working mobility needs, it is reasonable to assume that the commuters would have less elasticity of bike rental demands than the tourist over other factors, such as season, weather, etc. That being said, for example, when the weather becomes harsh, the commuters who need a bike to get to work or school are likely to still rent a bike, while tourists can easily change their visiting plan accordingly to the bad weather, meaning, not to rent a bike anymore.

Based on these reasonings, the research checked the elasticities of demand for the commuter group and tourist group.

Elasticities

In order to check the elasticity, the research has to feed the selected SVR model with self-defined new dataset, and calculate the elasticity based on the predictions.

The creation of the self-defined new dataset features a Control Variates method. For example, to check the elasticity of the commuter and tourist rental counts over the actual temperature (atemp), the new dataset puts in different atemp values and control other variables constant except for workingday (since this is how the research groups the commuters and tourists). Table 5 shows the elasticities tested for the commuters and tourists over atemp and weather situations (weathersit).

Table 5

Elasticities for Commuter and Travelers

Group	atemp	% less elasticity	weathersit	% less elasticity
Commuter	0.28732	13.18%	-22.2223	40.43%
Traveler	0.330947		-37.3074	

Note. "% less elasticity" describes the percentage of the less elasticity from commuters compared to that of travelers'.

Since the ANOVA tests mentioned previously show significance in the discrepancies between the two groups, it is only logical to deem that the 13.18% and 40.43% of differences in elasticities can also be significant.

Strategy Suggestions

Now that the research concluded less elasticities from the commuter group, it is natural to provide the suggestions from the following two perspectives: a) ways to utilize the inelasticity of commuters to gain more profit; b) ways to motivate the tourist to rental bikes regardless of harsh situations.

Examples for option a include: promote more powerful membership programs to attract more commuters to join the CB services; locate the docks nearer to business buildings to boost the number of users from the commuter's group; create bonus for registered every-day users; and etc.

Examples for option b include: locate more docks near tourist places to target specifically at the visitors; locate more docks near metro stations, so that even with harsh weather the users will still feel safe and convenient; build raincoat vending machine in the docks; etc.

Note that these are merely suggestions based on the analysis conducted from this research. To validate the efficiency and feasibility of these suggestion, more data will be needed.

References

ASPROVA. (n.d.). *Mean Absolute Deviation MAD.* Retrieved December 24, 2017 from the ASPROVA: https://www.asprova.jp/mrp/glossary/en/cat256/post-777.html

Ben Schumin. (2010, 19 December). Retrieved December 24, 2017 from the Wikipedia: https://en.wikipedia.org/wiki/Capital_Bikeshare

Capital Bikeshare. (n.d.). Retrieved December 24, 2017 from the Capital Bikeshare: https://www.capitalbikeshare.com/

DC Press. (n.d.). Retrieved December 24, 2017 from the DC Press: https://washington.org/DC-information/washington-dc-facts

Dent, S. (2017, September 21). *Mobike's dockless bike-sharing service comes to Washington, DC.* Retrieved from: https://www.engadget.com/2017/09/21/mobike-dockless-app-bike-sharing-washington-dc/

Ley, C., Klein, O., Bernard, P. & Licata, L.(2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*, 764-766.

Scikit-Learn. (n.d.) *Novelty and Outlier Detection.* Retrieved December 24, 2017 from the Scikit-Learn: http://scikit-learn.org/stable/modules/outlier_detection.html#novelty-and-outlier-detection

Scikit-Learn. (n.d.) *sklearn.svm.SVR.* Retrieved December 24, 2017 from the Scikit-Learn: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

University of California Irvine Machine Learning Repository (UCI). *Bike Sharing Dataset.* (2013). Retrieved from: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset