Using Machine Learning Methods to Predict Demand for Bike Sharing

Chang Gao[1(✉)] and Yong Chen[2(✉)]

[1] Boston University, Boston, MA 02215, USA
changgao@bu.edu
[2] Ecole Hôtelière de Lausanne, HES-SO, University of Applied Sciences Western Switzerland, Rte de Cojonnex 18, 1000 Lausanne, Switzerland
yong.chen@ehl.ch

Abstract. We applied four machine learning models, linear regression, the k-nearest neighbors (KNN), random forest, and support vector machine, to predict consumer demand for bike sharing in Seoul. We aimed to advance previous research on bike sharing demand by incorporating features other than weather - such as air pollution, traffic information, Covid-19 cases, and social economic factors- to increase prediction accuracy. The data were retrieved from Seoul Public Data Park website, which records the counts of public bike rentals in Seoul of Korea from January 1 to December 31, 2020. We found that the two best models are the random forest and the support vector machine models. Among the 29 features in six categories the features in the weather, pollution, and Covid-19 outbreak categories are the most important in model prediction. While almost all social economic features are the least important, we found that they help enhance the performance of the models.

Keywords: Machine learning · Data mining · Bike sharing · Demand prediction · Seoul

1 Introduction

Over the past two decades sharing economy has not only revolutionized the organization of economic activity but also unleashed the consumption and production potentials of a variety of tourism and hospitality businesses. These businesses include but are not limited to sharing accommodation exemplified by Airbnb, sharing transportation pioneered by Uber and Lyft, as well as various online booking platforms such as Booking.com and OpenTable. There are even more localized sharing businesses, such as bike sharing provided by private enterprises or governments as an alternative to the so-called "last-mile" public transportation. Bike sharing has been popular in many countries, due to the fact that environmental proception organizations proposed environmental sustainability transportation methods such as electric vehicles and bicycles [13]. Bike sharing provides benefits in various aspects and is achieving world-wide popularity [20]. For instance, the number of renters in US was larger than 28 million in 2006 [33]. All these businesses share one commonality, for which consumer demand is upon request. Namely, suppliers need to immediately, if not instantaneously, deploy

J. L. Stienmetz et al. (Eds.): ENTER 2022, *Information and Communication Technologies in Tourism 2022*, pp. 282–296, 2022.
https://doi.org/10.1007/978-3-030-94751-4_25

goods and services as soon as demand is generated. On the one hand, the success of sharing economy lies at such on-demand features; on the other hand, this requires supplies to predict consumers demand on various occasions as accurately as possible in the first place, thereby diverting goods and services to consumers as efficiently and timely as possible.

One telling example is Uber's surge pricing. Uber is capable of striking immediate balance between demand and supply through detecting riders' request in different periods of time, especially when demand fluctuates drastically in small geographical region [8]. In this case and many others, conventional econometric modeling in predicting demand would become less useful because it relies on predictors that usually do not change in the short run. For instance, it is extremely rare, if at all, to model consumer demand on a daily or hourly basis through using social or economic indicators. Of course, both economic indicators, such as income and price and a wide range of social demographics have compelling explanatory power in predicting long-term demand because they are grounded on sound economic theories. They would become useless in predicting instantaneous demand, such as in the case of Uber's surge pricing in which demand changes in a course of a few hours. The reason is that these predictors are constant on a daily basis not to mention on an hourly basis, which renders conventional economic modeling and forecasting obsolete. For this reason, machine learning has gained momentum in predicting demand in these contexts.

While studies using machine learning techniques to predict consumer demand are proliferating in tourism and hospitality, there are very few devoted to predicting demand for bike sharing. A wealth of studies that indeed addressed bike sharing are primarily from the field of computer sciences [5, 14, 26, 27, 34]. In fact, modeling tourism demand is disproportionately devoted to predicting tourist arrivals using either machine learning or a combination of machine learning and search query data [3, 9, 10, 23–25, 30]. However, sharing economy has not only changed the way we model tourism demand but also extended what is modeled to reflect the nature of sharing economy in various areas. In this regard, we aim to use machine learning techniques to predict consumer demand for bike sharing. We also aim to advance previous research on bike sharing by incorporating a wide range of features other than weather to increase prediction accuracy.

2 Literature Review

Machine learning and big data have been increasingly applied to model and predict tourism demand in various domains. This strand of research bifurcates evidently between enhancing the performance of econometric models through incorporating machine learning techniques and using search engine data in prediction algorithms [1, 3, 6, 9, 10, 30, 35]. As a matter of fact, tourism research has focused on predicting tourist arrivals through using both conventional econometric models and machine learning techniques [1, 3, 9, 10]. For instance, Akın [1] used Neural Network models to predict tourist arrivals in Turkey while using conventional econometric techniques, such as autoregressive integrated moving average (ARIMA), as a benchmark. Claveria et al. [9] used machine learning algorithms such as the support vector regression,

Gaussian process regression, and neural network models to predict tourist arrivals in Spain. Similar to Akin [1], they found that machine learning methods improved forecasting performance against the autoregressive moving average (ARMA) model as a benchmark.

On the other hand, researchers have started to realize the importance of big data in predicting tourism demand. In particular, search engine data provides researchers a viable substitute for conventional economic variables as predictors in modeling and forecasting tourism demand. In this respect, search engine data have been extensively used to predict tourism demand and tourist arrivals in particular [23–25, 30, 35]. Sun et al. [30] used kernel extreme learning machine (KELM) models and search results generated by Google and Baidu to forecast tourist arrivals in China. Xie et al. [35] fed search query data (SQD) generated from Baidu to a least squares support vector regression model with gravitational search algorithm (LSSVR-GSA) to predict cruise tourism demand. Many studies concluded that using machine learning coupled with search query data increases the forecasting performance and robustness of the models [25, 30, 35]. This perhaps explains why various search engine data were also used to model and predict tourist arrivals [23, 24], which used to be addressed in conventional econometric models.

One of the advantages of using machine learning is to predict micro-level tourist demand and the facet of demand, such as network effects on the Internet, that cannot be accounted for by conventional economic indicators. This advantage also enables researchers to narrow down the prediction horizon, thereby modeling short-term demand patterns. However, demand modeled in many studies is conventional tourism consumption, such as park attendance, cruise demand, and tourist arrivals [23, 24, 35]. The overriding objective was to improve prediction accuracy through using machine learning techniques. Hence the focus is a matter of model selection while having little to do with modeling on-demand economy, such as car or bike sharing. In fact, bike-sharing modeling entails short-term even almost instantaneous demand prediction. On the other hand, machine learning models need to take into account stational-level variance in bike demand, which would allow suppliers to deploy bikes efficiently across destination to ensure supply. Such deployment requires modeling and forecasting demand across different docking stations on an hourly basis depending on the degree of demand fluctuation.

There is a great deal research devoted to forecasting bike demand in various cities [5, 27]. A majority of these studies modeled bike demand on an hourly basis, aiming to provide policy implications for deploying bikes in a timely manner [14, 32]. For this reason, the features that were used to predict bike demand were exclusively weather conditions, ranging from precipitation, humidity to wind speed and temperature in the course of 24 h. We aimed to predict bike demand by extending the scope of features on a daily basis. Indeed, some studies have shown that the geography of bike-docking stations has impacts on bike demand, which has a lot to do with social and economic situations in which these stations are located. Obviously, hourly-based models with weather conditions as the primary predictors are insufficient to account for such difference. Insofar as policy is concerned, this study can provide implications for the supply of bikes in different districts and the deployment of bikes across stations.

3 The Data

We retrieved the counts of public bike rentals in Seoul of Korea from January 1 to December 31, 2020 from Seoul Public Data Park website [21]. This data set consists of hourly bike rentals recorded from 2,148 docking stations in 25 districts of Seoul. Note that 55 stations that were not functioning in the study period were discarded from the analysis. We ended up identifying a total of 2,093 stations that were active during the whole study period. We aggregated hourly data to compile daily rental counts, giving rise to a total of 9,111 observations, with a daily average of 2029 bike rentals in Seoul in the year 2020.

To predict bike rentals in Seoul, we identified a total 29 features in six categories: (1) weather, (2) air pollution, (3) traffic accidents, (4) Covid-19 outbreak, (5) social and economic factors, and (6) seasonality. These data are retrieved from the website Seoul Open Data [21]. These 29 features are the potential features influencing bike sharing demand. When weather or air quality is bad, people might be reluctant to rent a sharing bike. On the other hand, when traffic is bad, renting a bike will be more efficient. We also suspect that Covid-19 cases and other social economic factors might also influence the demand of bike renting. Note that Covid-19 confirmed cases and deaths were analyzed with a one-day time lag since their influence on bike demand, if any, would take at least one day to emerge. The reason that we delayed one day confirmed and deaths cases is that residents need time to process the news information produced. They might not realize the disease cases immediately after the release of the news on media, and they need some time to process the information. Since the new cases counts updates each day, the case number delayed by one day is more applicable. We aimed to pinpoint the most important features that can accurately predict bike demand.

4 Methods

We performed four machine learning algorithms to predict bike rentals, which are linear regression, the k-nearest neighbors (KNN), random forest, and support vector machine. All of these models were performed on R studio. Since these four models were developed based on different assumptions for identifying the relationships between independent and dependent variables, it is a convention in machine learning to use them complementarily for prediction.

4.1 Algorithms

Linear regression. Linear regression is the most widely used and simplest method to predict demand in various contexts. Due to its simplicity and straightforward economic intuition in explaining the relationship between predictors and the outcome, we use linear regression as a benchmark against which other more advanced models are compared for their predictive power. The linear regression model is given as

$$y = \beta_0 + \sum_i^n \beta_i x_i + \varepsilon \tag{1}$$

where β_i is the coefficient of feature x_i, β_0 is the constant, and ε is the random error [28].

K-Nearest Neighbors (KNN). The k-nearest neighbors (KNN) is a machine learning algorithm used for both classification and prediction. The KNN is a nonparametric technique which provides solution for the curve fitting of unknown shape, and has an advantage for data mining, because it does not assume specific forms of regression functions [2]. For both classification and prediction, explanatory variables take into account the k (a positive integer) closest instances. The parameter k needs to be tuned before modeling and it is crucial for non-parametric regression performance [2]. The calculations of the KNN are based on distances between an instance to its neighbors. The distances used for continuous variables are the Euclidean distance. The Euclidean distance d between two n-dimensional vectors $(p_1, p_2, \ldots, p_n)$ and $(q_1, q_2, \ldots, q_n)$ is given by:

$$d = \sqrt{\sum_i^n (p_i - q_i)^2} \tag{2}$$

The prediction of an observation is the mean of the values of k neighbors that are the nearest when implementing the KNN as the regression model in prediction.

Random Forest. Random forest is a almighty tool which ensembles decision trees and bagging [4].The base learner of random forests is a binary tree constructed by recursive partitioning (RPART) and then developed using classification and regression trees [7]. Binary splits of the parent node of a random forest splits data into two children's nodes and increases homogeneity in children nodes compared to parent nodes. Note that a random forest does not split tree nodes based on all variables; instead, it chooses random variable subsets as candidates to find the optimal split at every node of every tree [7]. Then the information from the n trees is aggregated for classification and prediction [7]. Random forests also provide the importance of each feature by accumulated Gini gains of all splits in all trees representing the variable discrimination ability [19]:

$$impor_j = \frac{1}{\#trees} \sum_{v \in x_j} Gain(x_j, v) \tag{3}$$

where $Gain(x_j, v)$ is the gain of the Gini index of feature x_j combined with node v [32].

Support Vector Machine. Support vector machine (SVM) is a machine learning technique for classification and regression [11]. SVM is suitable for general relationships between explanatory variables and responsive variables. The basic idea of SVM is to map nonlinear explanatory vectors onto a high dimensional space in order to find a linear decision hyperplane. The solution of SVM regression is given as:

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) K(x_i, x) + b \tag{4}$$

where $K(x_i, x)$ is the kernel function that satisfy Mercer's conditions, where α_i and α_i^* are the dual variables larger than or equal to 0 and smaller than or equal to the hyperparameter C [31]. We use the radial basis function (RBF) kernels with the corresponding Kernel equation of

$$K(x_i, x) = \exp(-\gamma ||x - x_i||^2) \tag{5}$$

in which γ is the kernel parameter. The RBF kernel provides solutions when the relationship between features and responsive variables is nonlinear and is computationally easier than polynomial kernels [12].

4.2 Feature Selection

We split the 9,111 observations into a training set with 75% of the cases, or 6,235 observations, while 25% as a test set, or 2,276 observations. The training set was used for feature selection, hyperparameter tuning, and prediction. The test set was used for evaluation and prediction for bike rentals. Prior to selecting features, we explored the Pearson correlation coefficients between the number of bike rentals and the features in each of the six categories. Major findings are summarized here. Most of the pollution features except CO are positively correlated with bike rentals. Covid-19 cases and deaths are negatively correlated with bike rentals. All but two social economic factors, namely the number of markets (−0.09) and number of stores (−0.06), are positively correlated with bike rentals. The population in a district has the strongest correlation with bike rentals (0.35). The number of traffic accidents is positively correlated with bike rentals (0.20). Visibility and humidity are most correlated with bike counts (0.29). Visibility is positively correlated to the number of bike rentals, while humidity, precipitation, and wind speed are negatively correlated with bike rentals.

We proceeded to use Boruta and recursive feature elimination (RFE) to select features. Boruta is a wrapper approach to determine the relevance of features through implementing a random forest classifier. A shadow attribute is created for each feature, and classification is performed based on the feature importance by using all attributes and shadow attributes. These shadow attributes help reduce the distracted impact of random fluctuations [22]. Even though Boruta uses random forest as the base algorithm, this will not increase the accuracy of random forest since the testing set was never exposed to the algorithm. Figure 1 shows the result of the Boruta feature selection on all 29 features but districts and rented bike counts because it is the dependent variable. The blue boxes represent the shadow attributes, green ones are the accepted or confirmed attributes while red attributes are rejected. Thus, the number of deaths in the category of traffic accidents is rejected, so this feature will not be entered in the regression models. Binary variables of traditional holidays and leisure holidays

are not as important as expected, and this result indicates that bike renting demand was not strongly influenced by the indicator holiday. We suspected that most residents rent bikes for many other reasons instead of holiday leisure activities.

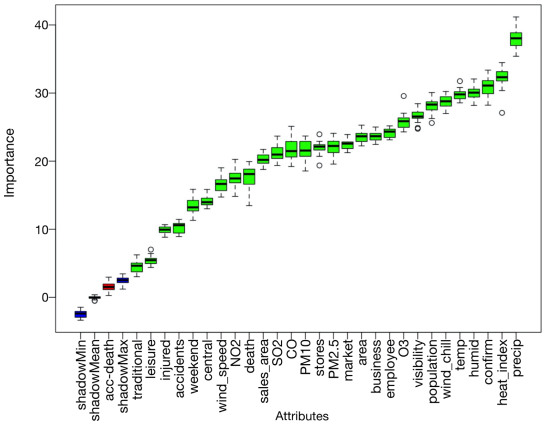

Fig. 1. Boruta feature selection.

While the Boruta algorithm can indicate what features will be accepted or not based on their performance, it does not state the variable's root mean squared error (RMSE). To retrieve lower RMSE, we further used the recursive feature elimination (RFE) to select features that can minimize the RMSE [15]. Like Boruta, RFE is also based on random forests in terms of method of implementation. RFE was implemented along with cross validation repeated three times for training to increase prediction performance. Like Boruta, no testing set instances had been exposed to the RFE algorithm. We identified the threshold number of features with the lowest RMSE is 25. Thus, the first 25 confirmed features are selected, and the excluded features are the number of injuries in the category of traffic accidents and holidays in the category of seasonality.

4.3 Model Development

We used hyperparameter tuning to optimize the performance of each of the four models. Hyperparameters are crucial to the result of machine learning algorithms and can affect the performance of the models [34]. There are several hyperparameter tuning methods, such as manual tuning, random search, and grid search, which can be applied in different contexts. We performed grid search for it is widely implemented and requires less experience and computational efforts. Grid search iteratively assesses over potential hyperparameter values, which are the number of neighbors (k). Figure 2 shows that a search on k value between 1 and 30 is computed, and the optimal k value with the highest coefficient of determination (R-squared) is 12. We identified two hyperparameters: *ntree* and *mtry* of the random forest. *ntree* is the number of trees to grow in the model and *mtry* is the number of variables that are selected as candidates at

liberty during each split [18]. We set *ntree* as the default value of 500, which is large enough to produce stable models and *mtry* in the range from 1 to 15 in the tune grid. Figure 3 shows that 10 is the optimal value of *mtry*.

The support vector machine (SVM) has two essential hyperparameters, sigma and cost, to be tuned. The tune grid of cost ranges from 0 to 120 with the step of 10. The tune grid of sigma uses 0.1, 0.01 and 0.001 as these three values are the conventional learning rate of SVM models. Figure 4 shows that the optimal combination with the highest *R*-squared is a cost of 120 with sigma equal to 0.01.

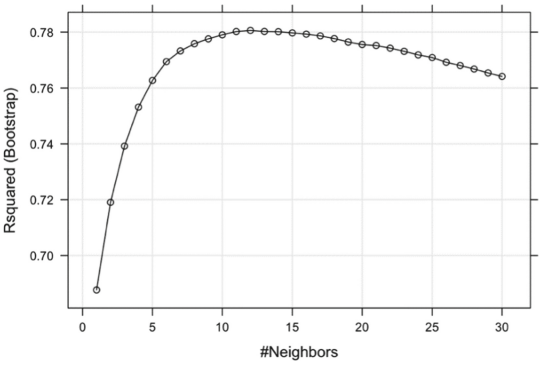

Fig. 2. Grid search of KNN.

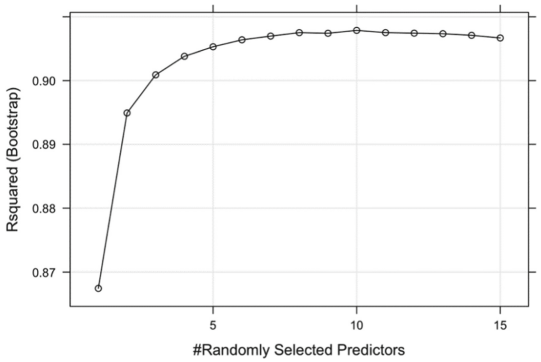

Fig. 3. Grid search of random forest.

5 Results and Discussion

All prediction models were implemented using10-fold cross-validation process repeated for three times during training, which generated a total 30 results for each model. Cross-validation is an approach to increase the performance of the proposed models [29]. The K-fold cross-validation separates the data set randomly into k subsets and one

subset is used for testing while the other k-1 subsets are used for training. The whole process of randomly separating, splitting, training, and testing is repeated several times and the optimal one is identified as having the minimum RMSE.

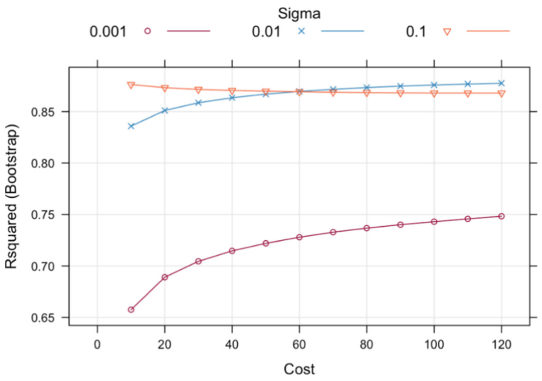

Fig. 4. Grid search of support vector machine.

5.1 Model Performance

We use the R-squared, RMSE, and the mean absolute error (MAE) to evaluate the performance of each of the four models. R-squared is a statistic measure (also called coefficient of determination) of the variation proportion in the responsive variable predicted by the explanatory variable [16]. Higher R-squared suggests better model performance in predicting dependent variable [17, 19]. Model with the highest R-squared, lowest RMSE and MAE is considered having the best predictive power. Table 1 shows that SVM yields the highest R-squared and the lowest RMSE and MAE in the training set. While RF has the same R-squared in training set (0.92), SVM outperforms RF due to its lower RMSE and MAE. However, when it comes to the testing set, RF outperforms SVM in terms of both R-squared and RMSE and MAE values. RF performs slightly better in the testing set than in the training set. Comparing prediction performance in the training and testing sets, RF's R-squared in the testing set is 0.93 while 0.1 lower in the training set. This result suggests that the RF model performs even better in the testing sets. The R-squared of KNN in the testing set decreases by 0.4 than in the training set which is the largest decrease compared with other models. The LM has the worst performance in both the training and testing sets, indicating that the relationship between bike rentals and the explanatory variables is nonlinear.

Table 1. Results of regression algorithms.

Models	Training			Testing		
	R^2	RMSE	MAE	R^2	RMSE	MAE
LR	.48	1065.12	778.87	.46	1062.37	796.30
KNN	.85	589.95	408.74	.81	641.82	448.22
RF	.92	442.25	274.04	.93	399.21	264.40
SVM	.92	415.45	252.40	.90	457.88	306.21

Note: LM = Linear regression, KNN = *k*-nearest neighbors, RF = Random Forest, SVM = Support vector machine, R^2 = *R*-squared, coefficient of determination, RMSE = Root mean squared error, and MAE = Mean average error.

5.2 Feature Importance

As shown, the random forest model performs the best in terms of R-squared, RMSE, and MAE. Figure 5 shows the feature importance of the RF model. As we can see, precipitation is the most important feature in predicting daily bike rentals, followed by Covid-19 confirmed cases and the O3 level of air pollution. Heat index and the levels of PM10 and PM2.5 are also strong predictors. The least important predictor for bike rentals is the number of traffic accidents. The most important social-economic feature is population while the rest are not salient. Table 2 shows the average of the feature importance in different categories of variables for the RF model. The category with the highest average feature importance is Covid-19 (50.37) while the lowest average feature importance category is traffic accidents (14.56). Air pollution and weather have similar average feature importance.

Table 2. Average feature importance by category of RF

Feature category	Average importance of features
Weather	40.31
Air pollution	41.70
Covid-19 outbreak	50.37
Traffic accidents	14.56
Social economic	21.86

Although the SVM has lower performance than RF, the evaluation matrices of SVM is also superior. The SVM in this study implemented RBF kernel. Unlike linear kernel, since RBF does not directly provide feature importance, the relative feature importance is composed by the weight of weight vectors. Features with higher weights indicate higher importance. Figure 6 shows the feature importance generated by the SVM model. The level of O_3 has the highest weights, followed by wind chill temperature, visibility, temperature, and population. The feature for the number of stores in

the district has the least weight. The level of PM10, weekend or not, the number of business and number of employees in the district also have low weights. It is worth noting that the features that are important in RF are not necessarily important in the SVM model, for instance PM10 level and heat index. The number of markets, business and employees are not the strong indicators in both RF and SVM models.

We also calculated the average feature importance in each of the six categories of the variables. Table 3 shows that weather has the highest weight, followed by Covid-19 outbreak and traffic accidents. Social economic features have the lowest weight. Comparing the feature importance of the RF model and SVM model, features in weather and Covid-19 are important in both models. Features in the Social-economic category have less importance in the RF and SVM models. In the RF model, the category of air pollution is more important than traffic accident, while in SVM model, air pollution is less important than traffic accidents. In both models, the level of O_3 ranks top 5 for the feature with high importance or weights.

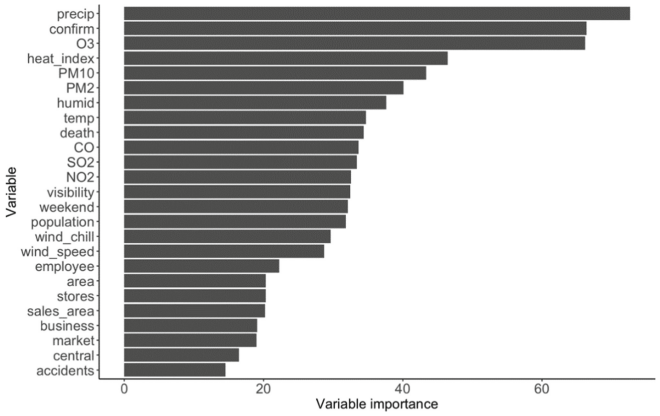

Fig. 5. Random forest model feature importance.

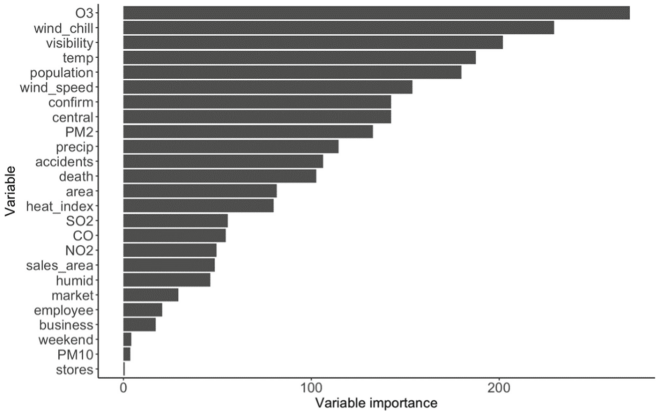

Fig. 6. SVM model feature importance.

Although social-economic features are not important, they did increase the predictivity of both the RF and SVM models. A subset without social-economic features was taken from the data set and implemented in RF and SVM models. Table 4 shows that the RF and SVM models without social economic factors have substantially lower evaluation matrices. For the RF model, the *R*-squared of the model without social economic features decreased by 0.39 in the training set and 0.38 in the testing set compared to the evaluation matrices with the features. As for the SVM model, the *R*-squared of the model without social economic factors also decreased drastically in both the training (0.34) and testing sets (0.37). This result suggests that social economic features are crucial to increase prediction accuracy, even though they may not have high feature importance values on their own right.

Table 3. Average feature importance by category in SVM

Feature category	Average importance of features
Weather	144.77
Air pollution	94.30
Covid-19 outbreak	122.58
Traffic accidents	106.33
Social economic	53.93

Table 4. Results of the RF and SVM models with and without social economic factors

Models	Training			Testing		
	R^2	RMSE	MAE	R^2	RMSE	MAE
RF	0.92	442.25	274.04	0.93	399.21	264.40
RF w/o	0.53	1019.77	722.62	0.55	972.90	693.26
SVM	0.92	415.45	252.40	0.90	457.88	306.21
SVM w/o	0.58	982.13	615.05	0.53	1018.03	649.37

6 Conclusion

While machine learning models are completely data driven, we have attempted to incorporate social economic variables in the models to predict bike sharing demand. Despite the fact that these variables are barely useful in explaining and predicting short-term bike demand because they are constant, they did reveal demand differences between docking stations that are characterized by different social economic conditions. The roles that these variables play are to reveal population and economic activity that may differ across districts where bike docking stations are located. In this regard, bike sharing demand at the station level could perhaps be divided into basic demand, which is determined by social economic factors and induced demand, which changes with weather, pollution as well as a wide range of features that vary in the short term or

even instantaneously. We advanced studies conducted by V E et al. [32] and E and Cho [14] in predicting bike demand in Seoul in the sense that they only addressed the induced demand for bike sharing on a daily basis.

The best model is the random forest model in our study, and the most important features are precipitation, the number of Covid-19 cases, the level of O3, heat index, and the level of PM10. The most important categories of features for the random forest model are Covid-19 outbreak, followed by air pollution and weather. Almost all social economic features are the least important, however they played a role in enhancing the performance of the models. The SVM is also an acceptable model. The features in the categories of weather, Covid-19 outbreak and traffic accidents have highest average weights. These results indicate that weather features such as precipitation, temperature, heat index, wind chill temperature as well as Covid-19 outbreak have huge impacts on bike sharing demand in Seoul. Further research can focus on many other potential features that influence bike sharing demand and many other machine learning algorithms such as Multilayer Perception Model.

References

1. Akın M (2015) A novel approach to model selection in tourism demand modeling. Tour Manage 48:64–72. https://doi.org/10.1016/j.tourman.2014.11.004
2. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46:175. https://doi.org/10.2307/2685209
3. Bi J-W, Han T-Y, Li H (2020) International tourism demand forecasting with machine learning models: the power of the number of lagged inputs. Tour Econ. https://doi.org/10.1177/1354816620976954
4. Breiman L (2001) Mach Learn 45:5–32. https://doi.org/10.1023/a:1010933404324
5. Chang P-C, Wu J-L, Xu Y, Zhang M, Lu X-Y (2017) Bike sharing demand prediction using artificial immune system and artificial neural network. Soft Comput 23(2):613–626. https://doi.org/10.1007/s00500-017-2909-8
6. Chen K-Y, Wang C-H (2007) Support vector regression with genetic algorithms in forecasting tourism demand. Tour Manage 28:215–226. https://doi.org/10.1016/j.tourman.2005.12.018
7. Chen X, Ishwaran H (2012) Random forests for genomic data analysis. Genomics 99:323–329. https://doi.org/10.1016/j.ygeno.2012.04.003
8. Chen Y (2021) Economics of tourism and hospitality a micro approach. Routledge, New York, NY
9. Claveria O, Monte E, Torra S (2016) Combination forecasts of tourism demand with machine learning models. Appl Econ Lett 23(6):428–431. https://doi.org/10.1080/13504851.2015.1078441
10. Claveria O, Monte E, Torra S (2018) Modelling tourism demand to Spain with machine learning techniques. The impact of forecast horizon on model selection. Revista de Economia Aplicada, **24**(72):109–132
11. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/bf00994018
12. Dong B, Cao C, Lee SE (2005) Applying support vector machines to predict building energy consumption in tropical region. Energy Build 37:545–553. https://doi.org/10.1016/j.enbuild.2004.09.009

13. Dora C, Phillips M (2000) Transport, environment and health. World Health Organization, Regional Office for Europe, Copenhagen
14. Sathishkumar VE, Park J, Cho Y (2020) Using data mining techniques for bike sharing demand prediction in Metropolitan City. Comput Commun 153:353–366. https://doi.org/10.1016/j.comcom.2020.02.007
15. Fan C, Xiao F, Wang S (2014) Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy 127:1. https://doi.org/10.1016/j.apenergy.2014.04.016
16. Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning (Vol 1, No 10). Springer, New York, NY. https://doi.org/10.1007/978-0-387-84858-7
17. Glantz SA, Slinker BK (1990) Primer of applied regression and analysis of variance. McGraw-Hill, Health Professions Division
18. Han S, Kim H (2021) Optimal feature set size in random forest regression. Appl Sci 11:3428. https://doi.org/10.3390/app11083428
19. Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY. https://doi.org/10.1007/978-0-387-84858-7
20. Hulot P, Aloise D, Jena SD (2018) Towards station-level demand prediction for effective rebalancing in bike-sharing systems. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. https://doi.org/10.1145/3219819.3219873
21. Seoul Open Data. http://data.seoul.go.kr
22. Kursa MB, Rudnicki WR (2010) Feature selection with the boruta package. J Stat Softw 36:11. https://doi.org/10.18637/jss.v036.i11
23. Law R, Li G, Fong DK, Han X (2019) Tourism demand forecasting: a deep learning approach. Ann Tour Res 75:410–423. https://doi.org/10.1016/j.annals.2019.01.014
24. Li H, Hu M, Li G (2020) Forecasting tourism demand with multisource big data. Ann Tour Res 83:102912. https://doi.org/10.1016/j.annals.2020.102912
25. Li X, Li H, Pan B, Law R (2020) Machine learning in internet search query selection for tourism forecasting. J Travel Res 60:1213–1231. https://doi.org/10.1177/0047287520934871
26. Li Y, Zhu Z, Kong D, Xu M, Zhao Y (2019) Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction. Proc AAAI Conf Artif Intell 33:1004–1011. https://doi.org/10.1609/aaai.v33i01.33011004
27. Liu J et al (2015) Station site optimization in bike sharing systems. In: 2015 IEEE international conference on data mining. https://doi.org/10.1109/icdm.2015.99
28. Saud S, Jamil B, Upadhyay Y, Irshad K (2020) Performance improvement of empirical models for estimation of global solar radiation in India: a k-fold cross-validation approach. Sustain Energy Technol Assess 40:100768. https://doi.org/10.1016/j.seta.2020.100768
29. Sun S, Wei Y, Tsui K-L, Wang S (2019) Forecasting tourist arrivals with machine learning and internet search index. Tour Manage 70:1. https://doi.org/10.1016/j.tourman.2018.07.010
30. Vapnik V (1999) The nature of statistical learning theory. Springer, Berlin. https://doi.org/10.1007/978-1-4757-3264-1
31. Sathishkumar VE, Cho Y (2020) A rule-based model for Seoul bike sharing demand prediction using weather data. Europ J Rem Sens 53(sup1):166–183. https://doi.org/10.1080/22797254.2020.1725789
32. Wang Z, Sun Y, Zeng Y, Wang B (2018) Substitution effect or complementation effect for bicycle travel choice preference and other transportation availability: evidence from US large-scale shared bicycle travel behaviour data. J Clean Prod 194:406–415. https://doi.org/10.1016/j.jclepro.2018.04.233

33. Wong J, Manderson T, Abrahamowicz M, Buckeridge DL, Tamblyn R (2019) Can hyperparameter tuning improve the performance of a super learner? Epidemiology 30:521–531. https://doi.org/10.1097/ede.0000000000001027
34. Xie G, Qian Y, Wang S (2021) Forecasting Chinese cruise tourism demand with big data: an optimized machine learning approach. Tour Manage 82:104208. https://doi.org/10.1016/j.tourman.2020.104208
35. Xu T, Han G, Qi X, Du J, Lin C, Shu L (2020) A hybrid machine learning model for demand prediction of edge-computing-based bike-sharing system using Internet of Things. IEEE Internet Things J 7:7345–7356. https://doi.org/10.1109/jiot.2020.2983089