Predicting bike sharing demand with machine learning

Aleksa Jelic
STUDENT NUMBER: 2024812

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Supervisor: Dr. Sharon Ong
Second Reader: Dr. Paula Roncaglia

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2021

Preface

I would like to thank Dr. Sharon Ong for helping me to find data for my thesis and guiding me to complete my thesis on bike sharing demand data set. I strongly believe that mobility share markets should create predictive model for more quicker and better experience for its customers. I am glad that I have opportunity to study in Netherlands which is the land of bikes, and this topic has connection to.

Predicting bike sharing demand with machine learning

Aleksa Jelic

The objective of this research is to predict the number of bikes at station level for bike sharing systems. Majority of previous work use computational models which are complex. An alternative approach to predict the number of bikes is via machine learning. This thesis explores traditional regression methods and autoregression. The auto-regression model provided the highest R2 with the lowest RMSE among the algorithms. The most important variables were temperature and time. Results of this research benefit all mobility share companies and their customers by enabling them to use their vehicles more effectively and efficiently, without delay.

1. Introduction

The aim of this research is to develop machine learning models to predict the number of bike required in a Bike Sharing System. Bike-sharing systems allow anyone to hire bicycles from one of the city's numerous automated rental stations, ride it for a short distance, and then return it to any station in the city. Many cities across the world have recently implemented similar systems to encourage citizens to utilize bicycles as an environmentally sustainable and socially equitable means of transportation, as well as a good complement to other forms of public transportation (mode-sharing) (Raviv et al., 2013). First country to implement this model is the Netherlands, following their example bike sharing systems are widespread throughout the world (Shaheen et al., 2010).The ability of a bikesharing system to meet the variable demand for bicycles and vacant lockers at each station is critical to its effectiveness. This is accomplished by a repositioning operation that involves taking bicycles from some stations and transporting them to other stations with the help of a specialized fleet of trucks (Raviv et al., 2013). From the economical point of view, this process is highly intensive and expensive for bike sharing companies, therefore another optimal solution is needed. To solve such a problem machine learning methods will be used.

There is a big difference between the concepts of bikesharing and bike rental. In bikesharing, all bicycles are picked up from the same place and dropped off at the same place. The bikesharing concept, on the other hand, allows its customers to move freely around the city and pick up or return their bikes at any docking station (Sathishkumar et al., 2020). In (Qiu & He, 2018) study, bike-sharing models increase bicycle use and mobility options, reduce traffic congestion and energy consumption, and improve public health.

1.1 Motivation

Transportation plays important role in human everyday activity, therefore, it is very important part of the research. Improving mobility share systems, machine learning methods could be executed for fast and efficient way of better transportation. The

population is growing (Roser et al., 2013), so speed and efficiency is key to prosperity. Bike sharing systems solve a variety of problems, from CO2 reduction, affordable transportation prices, traffic congestion, etc. But how good is their overall model, how many bikes should they install at each location, will their customers use their service again if they have to walk to a docking station further away. Nowadays, people use different types of transportation every day and not only bicycles are affected by this problem but also other types of shared mobility vehicles. There must be a system that can estimate enough bicycle supplies at certain docking stations, which will save customers' time and allow them to pick up the bicycle at the nearest bicycle station. With an efficient and accurate model that can solve the bike supply problem, machine learning is very helpful in solving such a problem. .By using machine learning methods, this problems can be resolved to some extent.

The research is relevant in various groups of society especially for bike sharing companies. Forecasting the rental bike demand is very crucial part not only for individuals, but as an example for other mobility share companies that are yet to come. All this research vastly contributes to more sustainable, scalable, and efficient businesses.

From scientific point of view, more and more different kinds of mobility share markets are coming into play, and with that more difficulties are set for scientist to deal with. The scientific goal is to maximize the accuracy of the predictive models in order to make accurate predictions on bikes. It is important to consider the correlation of variables and the distribution of data when determining the level of accuracy. Present research is using historical data. Plenty of research uses time-series models (Noussan et al., 2019) in order to deal with historical data. This opens space of timer-series predictive models such as Autoregression which will be used for this study.

There is little research on predicting bicycle demand at the station level, as well as applying machine learning methods to this matter. Also little resreach have used time-series prediction models in order to predict count on mobility share vehicles .In this way, study will contribute to the current research and help in further research.

1.2 Research Questions

In this study, multiple regression algorithms will be used to predict bike count. An evaluation metric will be used to measure the performances of the regressors, and the best regression will be selected. The main research question is as follows:

"To what extent we can predict bike count required for the stable supply of rental bikes using machine learning"

Since the goal of every research is to find the best model that will output highest accuracy on data set, therefore the first sub-research question is:

"Which machine learning algorithm gives the highest prediction accuracy?"

In addition, there are many features in the data such as temperature and time which can be used to predict bike counts. It would be interesting to see which fatures are more importance for prediction. Therefore, a second sub question is

"Which features are the most important for bike count prediction?"

Due to data being historical, it is of great significance to predict numbers of bikes at each hour. Each hour's bike count may depend on the bike count of the previous hour,

so data may not be independent and equally distributed. A regression technique that accounts for this Autogression. Therefore, a third sub-research question is:

"To what extent Autoregresion can predict bike count required at each hour for the stable supply of rental bikes"

Main research question will be answered by use of supervised machine learning and check to what extent bike count can be predicted. In regard to first sub research questions, these models will be evaluated using metrics such as $RMSE$ and R^2 to find the best performing one. For second sub-research question, feature importance table will be shown regarding which variable accounts for most of the importance. Last sub-research question will use Autoregression algorithm that is capable of predicting time-series that such the one current research use.

1.3 Summary of findings

This research found that the best performing algorithm was Autoregression with results of R^2=92% and RMSE=0.19. Further more feature importance techniques were implemented, where feature temperature and team accounted for most important variables. These results will be introduced in section Results and Discussion.

2. Related Work

Recently, bike sharing became popular all over the world, which opened a whole new field of research, not only for bike sharing demand, but also for other mobility sharing models.

To predict the number of bikes, many different algorithms have been applied by different researchers; Random Forest (RF) and Partial Least-Squares Regression (PLSR) (Ashqar et al., 2017), K-nearest neighbors (KNN) and Conditional Inference Tree (CIT) (Sathishkumar et al., 2020), Linear Regression (LM), Gradient Boosting Machine, Support Vector Machine (SVM) and Boosted Trees (Sathishkumar et al., 2020).

Many of these algorithms are used today in current work projects, in the study (Feng & Wang, 2017) Washington D.C. bike-sharer network used multiple linear regressions and random forest algorithms to predict demand for rental bicycles. Similarly, in the Chinese city of Suzhou, Long Short-Term Memory. (LSTM) networks and gated recurrent units (GRUs) to predict the availability of bicycles.

In the study (Nair & Miller-Hooks, 2014), a model was created that can determine where bicycle docking stations should be placed. This model is written as a two-level stochastic dynamic program, where the upper-level problem is to place the stations in a way that improves streamlined user flow. The lower level problem is to reduce the transportation cost of each consumer by choosing the transportation mode and travel route based on the station locations. Each consumer solves the problem independently. To solve the extensive network of locations, a heuristic solution approach based on a genetic algorithm is created. The model uses historical data of bike-sharing in Washington DC.

In the study of (Sathishkumar et al., 2020) they proposed a rule based model for Seoul Bike data which this resreach uses as well. They have worked with 5 different machine learning algorithms such as: CUBIST (regression model), Regularized Random Forest, classification and Regression Trees, KNN and Conditional Inference Tree. Out of the 5 different algorithms, CUBIST showed the best prediction results based on variance (R^2).

In (Tomaras et al., 2018), a framework called SmartBIKER was developed to model patterns of bicycle requests on significant occasions. The system uses a pattern prediction model to evaluate low or attractive bike stations and build a migration technique that reduces overall movement costs while improving user stations.

In addition to bike sharing, a study (Waserhole & Jost, 2016) vehicle sharing systems on top of the existing system is proposed. It is assumed that a demand curve is created for each trip. The vehicle sharing system is described as a distributed queuing network with unlimited buffer capacity and Markovian demand. The objective is to optimize utilization by setting prices and rewards for each potential trip under certain constraints. The problem is solved heuristically by combining the total distribution on the initial graph with a greedy algorithm approach for an integer program.

The bimodal Gaussian Inhomogeneous Poisson (BGIP) algorithm was (Huang et al., 2018) proposed to predict the number of bicycles. They proved that the process of taking and returning bicycles can be described by an inhomogeneous Poisson distribution, and approximated this intensity of the process with a bimodal Gaussian function. The number of rental bicycles is then predicted by estimating the mean check-in and check-out of rental bicycles. Their method outperformed time series regression methods such as. Autoregressive Moving Average Models (ARMA).

(Vishkaei et al., 2020) used a Jackson network in which a generic algorithm was developed to obtain the correct number of features for balancing the number of bikes (inventory). Their bike sharing system (Public Bicycle Sharing Systems (PBSS) is slightly different from the work in (Tomaras et al., 2018) because in their case, the customer provides the location via app before renting the bike. PBSS provides the client with the estimated time and availability of the bike counts.

In the study (Almannaa et al., 2020), dynamic linear models (DLM) were proposed to predict the bike availability ina bike sharing system. First and second order polynomial models were used and compared with the least squared boosting algorithm as a basis. Dynamic linear models outperformed the baseline algorithm, achieving a prediction error of 0.37 for a 15-minute forecast and 1.1 bikes per station for a 2-hour forecast horizon.

This paper focuses on findings from previous research and builds regression models using bicycle data from Seoul. Most methods use computational models, while my approach is based on machine learning. Due to the availability of data today and the ever increasing amount of data that can be collected.

3. Methods

3.1 Data set

For this research, the dataset of bike-sharing demand in Seoul was used. The dataset was retrieved from UCI Machine Repository and contains the number of public bicycles. It is composed of 14 attributes: Date, Rented Bike Count, Hour, Temperature(°C), Humidity (%), Wind Speed (m/s), Visibility (10m), Dew Point Temperature(°C), Solar Radiation (MJ/m2), Rainfall (mm), Snowfall (cm), Seasons, Holidays, Functioning Day with 8760 instances. Each 24 instances represent one day in a year. The target is the number of bicycles rented per hour. Table 3.1 shows a brief description of all variables in the dataset.

Variable	Variable Description	Type
Date	The day of the 365 days	str
Rented Bike Count	Number of rented bikes	int
Hour	The hour of the day	int
Temperature (C)	Temperature of the day	float
Humidity	Humidity in the air	int
Wind speed (m/s)	Speed of wind	float
Visibility (10m)	Visibility in range of 10m	int
Dew point temperature (C)	Temperature at the start of the day	float
Solar radiation (MJ/m^2)	Solar radiation per unit on horizontal area	float
Rainfall (mm)	Amount of rainfall	float
Snowfall (cm)	Amount of snowfall	float
Seasons	Season of the year	str
Holiday	If it is a holiday	str
Functioning day	if it is a function day	str

Table 1
Description of variables

3.2 Data exploration and pre-processing

In this section, methods for data exploration and preprocessing will be presented in order to improve prediction for machine learning models.

The correlation matrix provides a better understanding of which features to focus on. A correlation matrix of the features in the dataset can be seen in Figure 1. When comparing the features to target variable, it is clear that temperature and time of day have the highest correlation. When comparing other features, the features "temperature (°C)" and "dew point temperature(°C)", are highly correlated, which causes multicollinearity. This can cause problems when fitting the models, so the dew point temperature is excluded from the dataset

It is crucial to handle categorical data during the pre-processing phase, as in the data having 3 categorical features, namely Seasons, Holidays, and Functioning Days. Machine learning can seldom deal with categorical data. Hence, statistical models must be converted into numerical ones. One-hot encoding, creating dummy variables, and many other methods can be used for categorical variables. Based on current work, dummy variables were created for each of the 3 features. Seasons range from 1 to 4, 1- Spring, 2- Summer, 3- Autumn, 4- Winter. Holidays and function days contain binary numbers that range from 0 to 1.

The next step is to verify that the data is normal. Depending on the distribution of the data, parameters will be decided. As can be seen in Figure 2 (left), target variable distribution does not follow a normal distribution. Thus, data was transformed to closely resemble a normal distribution (see Figure 2) (right).

The box-cox transformation assumes that target variables are not normalized, so that errors will be normalized through the transformation. Using this transformation can improve the predictive power of machine learning algorithms.

Figure 1
Correlation matrix

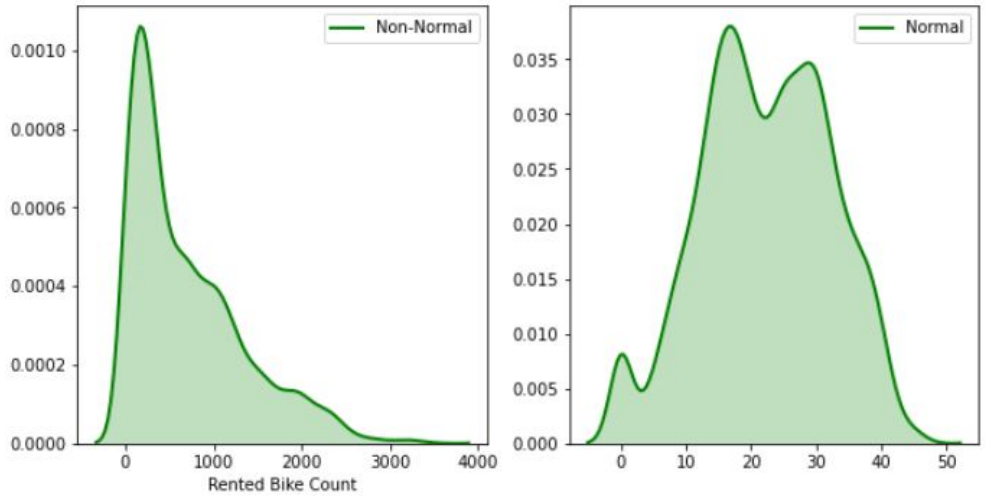

Figure 2
On the left side *Figure 2* distribution of target variable before normalizing, right side shows normalized distribution of target variable

3.3 Regression Algorithms

To be able to achieve research goals, this research is trying to find best possible regression algorithm in order to predict bike sharing demand count. For purpose of this study 6 traditional regression algorithms and 1 times-series algorithm will be used, namely Linear Regression (LR), Lasso and Ridge regression, Support Vector Regressor

(SVR), Decision trees (DT), Random Forest Regressor (RFR) and Autoregression (AR). Figure 3 provides list of algorithms and their description.

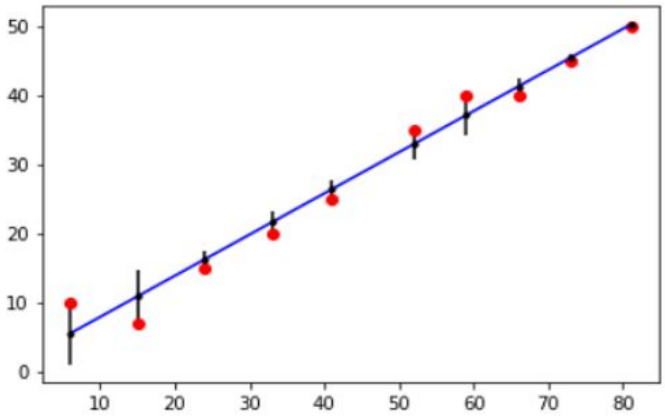

Figure 3
Fitted regression line with error margins

3.3.1 Linear Regression. In the world of machine learning, linear regression is one of the most simple and widely used models. When the dependent variable is continuous (e.g., bike count), linear regression is a good option.

Linear regression assumes that the bike counts are linearly correlated to the features in the dataset such as temperature. This assumption may be valid because as temperature increase, more people may go out and rent bikes. Linear regression fits a linear model with coefficients for each feature to minimize the mean square error.

Figure 3 shows a fitted regression line over the points shown in red. The mean square error is the mean of square of the residuals. For the purpose of the research, root mean squared error (RMSE) will be used. RMSE entails same formula as MSE, but root of the MSE value is taken into account. When RMSE is used, RMSE penalizes large errors more, and so can be more appropriate in calculating how many bikes is our model off.

Conversely, in the linear regression approach, outliers can have a significant impact on the regression. Linear regression assumes that the independant variables (features) are linearly correlated with the target variable. It also assumes that attributes are independent of one another. Furthermore, linear regression may be prone to overfitting as features with higher coefficients may skew or bias the prediction.

3.3.2 Ridge and Lasso regression. Ridge and Lasso regression are popular regularization techniques which are used to prevent overfitting. Ridge regression applies an L2 penalty, which penalized coefficients with higher weights.

Lasso regression employs the L1 normalization technique which penalizes less relevant features in the dataset by setting coefficients zero and hence eliminates them. As a result, you get of feature selection and a more efficient model. However, Lasso regression may be used to remove features that high correlated by dropping one of them. From the correlation matrix (shown in Figure 1), if two features here correlated, Lasso regression will remove one of them by setting the coefficients to zero.

Overall, Ridge and Lasso regression were chosen because the data contains multiple features which are correlated such as.. Research shown that these algorithms perform better when dealing with lower numbers of features such as our dataset (Pavlou et al., 2015). Both of these algorithms are optimized for prediction rather then inference which present research aims to.

3.3.3 Support vector regression (SVR). Support Vector Regression (SVR) utilizes the same equation as linear regression, but instead of regression line, a hyperplane with margins is implemented. The margins allow for some error to be tolerated. Data points that are nearest to either side of the hyperplane are called Support Vector, and they are used to map the boundary line(Üstün et al., 2007).

Linear regression assumes that the features have a linear relationship with the target variable. SVRs allow for non linear relationships by mapping the features to a higher dimensional feature space using a non-linear mapping function (kernel function) (Üstün et al., 2007).

However, SVR does not perform well on complex and larger data sets due to its training inefficiency. This dataset in this work is relatively small and therefore compatible for SVR (Awad & Khanna, 2015).

3.3.4 Decision Tree Regressor. Decision trees are statistical models that measures a target value using a collection of binary rules. A decision tree (Figure 4) generates an approximation by "asking" the data a series of questions between 2 nodes, each of which narrows the range of possible values until the model is positive to make a single prediction. The model determines the order of the questions as well as their content.

Following the training process, each tree calculates the best 'question' as well as the order in which they should be executed to make the most precise estimate possible. For more accurate prediction, model should be fed the same data format in order to make a prediction.

In order to make a decision between 2 nodes, decision trees use mean-squared error (MSE) to decide to split a node to more sub-nodes. A method for determining the best split is to test each variable and all possible value of that variable to determine which variable and value produces the best split. The model will stop with creating trees when the data is fully split or when hyperparameter constraints are met. Hyperparameters refers to parameters such as maximum depth or maximum leaves.

The advantage of this algorithm is that less pre-processing requred as data scaling nor data normalization is necessary. However this model can be computationally expensive if the data has a lot of features. For the purpose of the research this algorithm is selected based on data set size which is not to complex nor simple for such an algorithm. Previous work achieved high prediction results using trees algorithms (Sathishkumar et al., 2020)

3.3.5 Random forest regressor. Decision tree can completely fit the training data but tend to overfit on the test data. An alternative to deal with overfitting in decision trees is to use random forests regressors.

Random forest works by training a large number of decision trees and then calculating the mean prediction of the individual trees. Notion of random implies to randomly created decision trees. Random decision trees are created on different subsets of the features and datapoints.

For accurate predictions, random forest regressors can be optmized by hyperparameter tuning to ensure that model does not depend too heavily on any single feature and

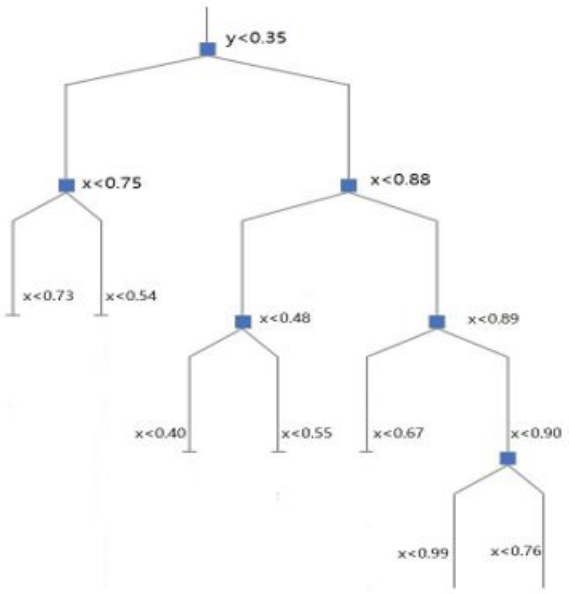

Figure 4
Example of decision tree

that all potentially predictive features are considered equally. Also due to previously mentioned random creation of decision trees, adding randomness prevents overfitting.

Enormous advantage that random forest regression provides is feature importance estimate. This is very crucial for this research to answer sub-question regarding which features are the most important for bike bike prediction. Using feature importance, effort can aid in a deeper understanding of the solved problem and, in some cases, contribute to model improvements.

3.3.6 Autoregression. This research is deals with historical data over certain period of time. Therefore, predicting bike counts can be considered a tiem series forecasting problem. One of the widely used time-series algorithm is Autoregression.

Autoregression uses previous observations to input into a regression equation in order to predict the value of a future time step. The autoregression model makes the assumption that the observations at previous time steps can be used to predict the value at the next time step. This relationship between variables is referred to as autocorrelation. A positive correlation occurs when two variables change in the same direction (for example, when they both rise or fall at the same time). Negative correlation occurs when two variables move in opposite directions during a change in value (one going up and the other going down). The correlation between output variables and previous values can be calculated using statistical measures at various time lags. An autoregression model can place more weight on a specific lagged variable if there is a higher correlation between the outcome variable and the lagged variable.

In this case, as well, it is called an autocorrelation because the variable itself is used to calculate it. Since the time series data are in sequence, it is also referred to as serial

correlation. In addition to helping to select which lag variables are useful for a model, correlation statistics can also help to determine which variables are not. Surprisingly,all lag variables that have little or no relationship to the output variable may suggest that the time series problem isn't predictable.

In addition, due to its time-series capability, Autoregrssion will predict the number of bikes at each hour which will give much better results then just predicting bike at stationary level.

3.3.7 Implementation and Evaluation Metrics. In this research, the data is split with ratio 70:30. Each model will be tested against the baseline regression model which is Linear Regression. All models will be optimized for better accuracy using Grid Search. Grid-search is used to find the model's optimal hyperparameters that provide the most 'accurate' predictions. The function is fed with different hyperparameters, model takes each parameter and calculates accuracy for it. Such parameters include kernel type in SVMs, number of k neighbors in KNNs and the number of hidden layers in neural networks.

To execute the model, Python 3.7 computer programming language will be used. Data exploration and pre-processing part will be done using pandas and NumPy packages. To generate the models ScKit-Learn will be used. In order to run creat Autoregressin tensor board pytorch will be used .All the graphs and figure will use Matplotlib package.

Models will be evaluated based on its accuracy using R^2 metric following with Root Mean Square Error (RMSE) metric. Root mean square error accounts for absolute number indicating how far the expected results differ from the actual number. In this particular case RMSE will be used to estimate how far off are the predicted values from the actual data. R^2 will be used to measure accuracy of the model. In regression model R^2 represent the goodness of fit, the goal is to get closer to 1 in order to get better accuracy.

4. Results

The results of the regression algorithms presented in the Methods section are presented in this section. For this study, 7 different models were trained. Each algorithm was fine-tuned using Grid Search, with the exception of the baseline Linear Regression algorithm. All models are evaluated based on their R^2 and RMSE. The performances of the models are shown in Table 2.

Model	Hyperparameters	R2 score	RMSE
LR	None	49%	7.02
Ridge	Alpha=3	48%	7.02
Lasso	Alpha=0	48.7%	5.02
SVR	C=1000	61%	6.11
DT	Max Depth=7	59%	6.28
RFR	Max Depth=50	64.9%	5.82
AR	Lag=359	92%	0.19

Table 2
Results of Regression Algorithms

Linear regression (LR) was set as the baseline model with an R^2 of 47% and an RMSE of 7.23 on the training set and 49% R^2 on the test set with 7.02 RMSE. The goal of this model was to minimize the RMSE and bring R^2 (accuracy) close to 1. Based on 6394 values, the predictive model did not perform well.

The ridge regression model (L2 regularization) was trained using grid search with different alpha values. Out of 9 different alpha values, the best R^2-train result is 47% with alpha=3 and RMSE 7.23, regarding the R^2-test result is 48% with alpha=3 and RMSE 7.02. The scores are not very different from the baseline; therefore, the lasso regression (L1-regularization) was trained similarly to the ridge regression. In the training set, the model scored 47% with alpha=0 and RMSE 7.23. The test set performed slightly better with an accuracy of 48.7% with alpha=0 and RMSE 7.02. Neither model caused a drastic change in R^2.

Support Vector Regression (SVR) was trained on 12 different kernels. Using grid search, the following results were obtained. The SVR model R^2 achieved 58% with kernel 1000 and an RMSE of 6.01 on the training set. On the test set, SVR performed slightly better with R^2 61% and an RMSE of 6.11 with kernel 1000. SVR performed better than the previous models due to feature scaling, which allows all features to be equally important.

Decision trees (DT) were the next model applied. The data remained scaled from the previous model. A grid search was performed to find the best parameters. For the training set, the accuracy of the decision trees was 57% and 5.93 RMSE with *max depth=7*. For the test set, the accuracy of decision trees was 59% and 6.28 RMSE with *max depth=7*.

Pre-last model to implement was Random Forest Regressor (RFR). Data was still scaled from the previous models. Grid search was made by tuning 'max depth' ranging from 50 to 100 with increment of 10. Grid search was fitted on X and y training data. The best model was kept with the best estimators. For the training set bet score was 64.5% and RMSE 2.18, for test score best score was 64.9% and RMSE 5.82. This is the best model out of the six.

The last model to be implemented was Auto-Regression (AR). Based on bike usage historical data autoregression accounted for 87 % of R^2 and 0.27 RMSE before grid search, which is drastic change in regards to previous traditional algorithms.In order to optimize autoregression algortihm grid search was used. Parameter "lag" was chosen which represents when the outcomes of one time period have an impact on subsequent time periods.

Optimized autoregression achivned 92% of R^2 and 0.19 RMSE. Figure 5 shows Autoregression results comparison between actual data and model prediction.

In Figure 6, feature importance graph is displayed. Feature importance is used to measure the utility of all the variables in the data. Using feature importance from skicit-learn model in this research indicates how specific variable enhance the prediction. Temperature is the most important feature with value of 0.36, second most important feature is Hour or the time with result of 0.22. These results will be discussed in discussion section.

Using hyper-parameter tuning technique grid search best performing algorithm was Autoregression. Best traditional algorithm after grid search was random forest regressor. Table 2 shows results after grid search was implemented.

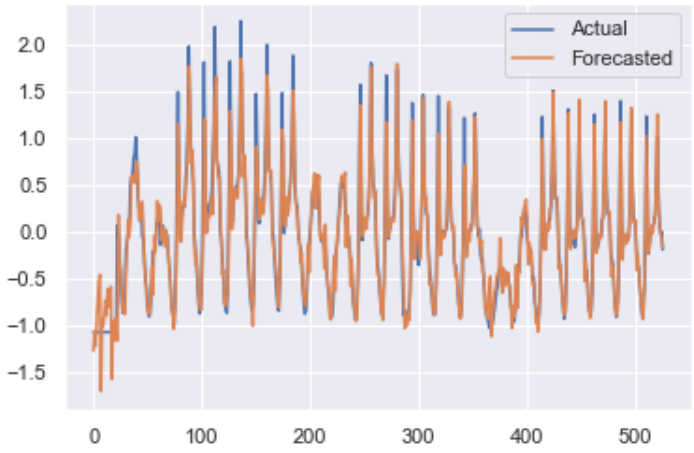

Figure 5
Autoregression model output

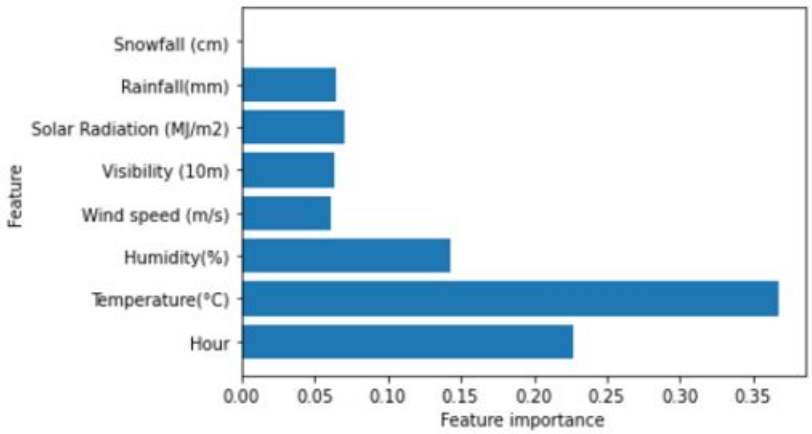

Figure 6
Feature importance graph from Random Forest Regressor

5. Discussion

The objective of the research was to predict the bike count required for the stable supply of rental bikes. The linear regression was chosen for baseline regressor. Applying gird search on proposed algorithms results were obtained. The best model is the model which has highest R^2 score and lowest RMSE, in this case Autoregression performed the best. Random forest regressor and SVR performed best out of traditional machine learning algorithms. Compared to other prediction models Lasso and Ridge performed worst.

Since the Lasso and Ridge regression model is worst, we can conclude that baseline algorithm was not beaten by these two, but the other 4 algorithms successfully outperformed Linear regression. This indicates that the relationship between bike counts and the features are non-linear. Random forest regressor has outperformed every traditional model which is not a big surprise. Studies on bike sharing data proves that this model has highest accuracy in their predictions (V E & Cho, V E & Cho).

Reflecting on research question "To what extent we can predict bike count required for the stable supply of rental bikes using machine learning", multiple regressors were compared and evaluated with R^2 and RMSE scores. With the best model, Autoregression achieved accuracy of 92% and RMSE of 0.19, which indicates that machine learning indeed can predict bike counts required for stable supply of rental bikes.

It is not unexpected that Random forest regressor performed the best ot ouf 6 traditional on the data set. In paper (V E & Cho, V E & Cho) which is also based on bike sharing demand data, random forest regressor has produced the highest accuracy. The reason why this algorithm produced good results is becuase radnom forest regression is ensemble learner, this means that training points are randomly sampled when producing trees. Reflecting on (Sathishkumar et al., 2020) which used same dataset as present research proposed data mining techniques in order to predict bike sharing demand. Their best performing model was XGBTrees with R^2 of 91%, present resreach manged to beat their algorithm with R^2 of 92%. It is clear that time-series algorithms works better on historical data where the target variable is continuous.

Based on current discussion, the answer for first sub-research question, "Which machine learning algorithm gives the best prediction values" is already touched upon. From the results of the 7 algorithms, we can infer that Autorregression has the highest accuracy therefore the best prediction values.

Next sub-research question is "Which features are the most important for bike count prediction". Feature importance was introduced in results section, Figure 6. Feature importance was extracted from the Random forest regressor importance variable module. Feature importance method uses permutation data to quantify each variable's effect on the established model's overall predictive output. Calculating the drop in prediction accuracy produced from random permutation values of a variable, yields the variable's value. The higher the accuracy of prediction, the more important the element, and vice versa. The importance of features can be used to choose the most important features in RF, allowing the user to focus on the most important issues while still recognizing the relationships between independent and dependent variables (V E & Cho, V E & Cho).

Looking back at feature importance graph, temperature accounts for the highest importance. That is not a surprise, study (Horanont et al., 2013) showed how different weather parameters correlate to people's mobility and variation in activity patterns over time. Therefore, people of Seoul are more likely to stay home during colder days than warmer ones. Depending on the feature importance of temperature and study on peoples activity associated with whether,temperature is the most important factor when it comes to sharing bikes.

Second most important feature is Hour in the day. Referring to (Horanont et al., 2013) the findings show that various weather conditions influence people's activity habits during the day in each section of each weather parameter. For example, study shows that between 8AM and 9AM time frame there is high activity most likely due to people going to their jobs or school. Considering this information bike, sharing demand companies should put more bikes in more active time periods then inactive periods such as lunchbreak. Therefore, time is very important factor in bike sharing demand market and it is necessary feature for our model to make right predictions.

The findings show that the suggested feature importance based on permutation may provide a clear understanding of the relative feature importance of each feature to the RF model's overall prediction output.

None of the related work used time-series regression algorithms which in present research made significant progress. In order to answers last sub-research question *To what extent we can predict bike count required at each hour for the stable supply of rental bikes"* Autoregression (AR) was proposed. Previously said, AR achieved highest R^2 of 92%. The reason AR performs this good is autoregression in self which accounts that the current value is reliant on the value that came before it. For instance, prediction is based on previous hour of bike sharing usage. Similar studies were proposed, in (Cheng et al., 2021) autoregressive negative binomial time series model was used in order to investigate the effects of public transit closures on bike sharing demand in Washington, D.C.

5.1 Limitations

Conducted research made valuable insights and results regarding machine learning algorithms in order to predict bike count on stationary and hourly level.

Autoregression performed the best and achieved high accuracy, but its limitation would be predictions made on unseen problems, such as global pandemic (COVID-19) which already had an impact on bike sharing mobility markets. Study (Shokouhyar et al., 2021) proposed creation of new recovery plan for shared mobility after impact of pandemic. This suggests that current research on mobility share markets should take unseen outcomes into account and create models to minimize the consequences of such a problem.

Based on importance feature figure introduces in results section only two variables accounted for highest importance. They could suggest that data is missing features that could boost regression model. In study (Nair & Miller-Hooks, 2014) goal was to estimate the best location for bike docking station, but in the case best location was not estimated, feature like activity around station would give insight on how many bikes should be placed in it.

Speaking of feature importance, research takes into account overall season of the year. Separating season might yield better predictions while consider variable influence on selected season.

5.2 Future work

To prevent unseen occurrences such as earthquake or global pandemic study (Efendi et al., 2018) proposed fuzzy random auto-regression. The focus of their paper is a triangular fuzzy number of data preparation technique for building an enhanced fuzzy random auto-regression model for forecasting purposes using non-stationary stock market data. Reflecting on their study bike count could be predicted on non-stationary level, data could be taken while the customers is riding the bike.

Besides machine learning algorithms this research could be extended towards deep learning models. Deep learning models use neural network architectures that learn features directly from the data without the need for manual feature extraction are used to train deep learning models. Present research showed that non-linear models have higher R^2 and deep learning accounts for non-linear models.

Regarding only two variables that had the most importance on the models, in future more variables could be added to check if the model accuracy will improve. But, for the current data, data could be split in seasons where for each season predictive models would be calculated. Data distributes bikes usage throughout the year, it shows that data oscillates depending on season.This might produce better prediction then on overall data.

The work presented in this paper suggests that future research should focus on better representing data and its underlying structure using cutting-edge techniques.

6. Conclusion

Bike sharing demand is a crucial way of transportation in today's world and its market facing problem on insufficient bike supplies at their docking stations. The aim of the research was to solve that problem using machine learning methods. Machine learning methods were used to predict number of bikes using Seoul bike share demand data set. To recap, predicting bike count in bike sharing system using machine learning algorithm is possible. Best model is Autoregression algorithm that account for the highest R^2 out of 7 trained models. These findings give good contributions to all research in predicting rental bike use of a mobility share systems and increase the number of computational modeling algorithms focused on the prediction of bike sharing demand.

7. Self Reflection

I will summarize my good and bad points during the thesis process in the final section of my thesis, which will indicate my views about conducting research.

Writing a thesis is a challenging process, and I struggled knowing where to start. My supervisor, Ms. Ong, helped me find bike sharing demand data that I was able to use for my research.

My exploration of the dataset and literature on the bike sharing market led me to realize that the market is enormous and that computational models, in addition to machine learning models, play a significant role in improving its speed and effectiveness.

My thesis had several problems as it was written in Academic writing, as someone who was not very skilled in writing academic papers, it was difficult to get through the writing process. With Ms. Ong's feedback, I was able to fix these issues and approach them better.

I decided to step up on this one after my first submission was unsuccessful. My current approaches use a more complex algorithm called Autoregression, which is one of many skills learnt in Master's programs. Previously, I used traditional machine learning methods that were taught in bachelor Machine Learning class.

In terms of algorithms, I found the most satisfaction in creating them. I have a strong foundation in programming, especially Python. My most proud achievement was creating the Autoregession algorithm, which took me about 5 days with the help of scientific blogs and stackoverflow.

I am grateful to my supervisor, Sharon Ong, who guided me through the thesis. Her numerous comments, feedback, and all the work she did for me made me a better writer, programmer, and researcher at the same time. In addition to helping me, she was on hand for the whole group of us who were under her supervision.

Through this project, I learned that planning ahead is the key, the more you do, the less you have to do. Don't take feedback for granted, moral support is vital, especially during a pandemic when everything is online.

Since I started writing my thesis, I have made significant improvements in my writing. I greatly benefited from constant feedback on my mistakes. On the coding aspect I have learned to run Autoregression which was very demanding and time consuming. In the future I would like to do Deep learning methods on the same dataset.

References

Almannaa, M. H., Elhenawy, M., & Rakha, H. A. (2020). Dynamic linear models to predict bike availability in a bike sharing system. *International journal of sustainable transportation*, *14*(3), 232–242.

Ashqar, H. I., Elhenawy, M., Almannaa, M. H., Ghanem, A., Rakha, H. A., & House, L. (2017). Modeling bike availability in a bike-sharing system using machine learning. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, (pp. 374–378). IEEE.

Awad, M. & Khanna, R. (2015). *Support Vector Regression*, (pp. 67–80). Berkeley, CA: Apress.

Cheng, L., Mi, Z., Coffman, D., Meng, J., Liu, D., & Chang, D. (2021). The role of bike sharing in promoting transport resilience. *Networks and Spatial Economics*, 1–19.

Efendi, R., Arbaiy, N., & Deris, M. M. (2018). A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. *Information Sciences*, *441*, 113–132.

Feng, Y. & Wang, S. (2017). A forecast for bicycle rental demand based on random forests and multiple linear regression. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, (pp. 101–105). IEEE.

Horanont, T., Phithakkitnukoon, S., Leong, T. W., Sekimoto, Y., & Shibasaki, R. (2013). Weather effects on the patterns of people's everyday activities: a study using gps traces of mobile phone users. *PloS one*, *8*(12), e81153.

Huang, F., Qiao, S., Peng, J., & Guo, B. (2018). A bimodal gaussian inhomogeneous poisson algorithm for bike number prediction in a bike-sharing system. *IEEE Transactions on Intelligent Transportation Systems*, *20*(8), 2848–2857.

Nair, R. & Miller-Hooks, E. (2014). Equilibrium network design of shared-vehicle systems. *European Journal of Operational Research*, *235*(1), 47–61.

Noussan, M., Carioni, G., Sanvito, F. D., & Colombo, E. (2019). Urban mobility demand profiles: Time series for cars and bike-sharing use as a resource for transport and energy modeling. *Data*, *4*(3), 108.

Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., Elliott, P., King, M., & Omar, R. Z. (2015). How to develop a more accurate risk prediction model when there are few events. *Bmj*, *351*.

Qiu, L.-Y. & He, L.-Y. (2018). Bike sharing and the economy, the environment, and health-related externalities. *Sustainability*, *10*(4), 1145.

Raviv, T., Tzur, M., & Forma, I. A. (2013). Static repositioning in a bike-sharing system: models and solution approaches. *EURO Journal on Transportation and Logistics*, *2*(3), 187–229.

Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2013). World population growth. *Our world in data*.

Sathishkumar, V., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, *153*, 353–366.

Shaheen, S. A., Guzman, S., & Zhang, H. (2010). Bikesharing in europe, the americas, and asia: past, present, and future. *Transportation research record*, *2143*(1), 159–167.

Shokouhyar, S., Shokoohyar, S., Sobhani, A., & Gorizi, A. J. (2021). Shared mobility in post-covid era: New challenges and opportunities. *Sustainable Cities and Society*, *67*, 102714.

Tomaras, D., Boutsis, I., & Kalogeraki, V. (2018). Modeling and predicting bike demand in large city situations. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, (pp. 1–10). IEEE.

Üstün, B., Melssen, W., & Buydens, L. (2007). Visualisation and interpretation of support vector regression models. *Analytica chimica acta*, *595*(1-2), 299–309.

V E, S. & Cho, Y. Season wise bike sharing demand analysis using random forest algorithm. *Computational Intelligence*, *n/a*(n/a).

Vishkaei, B. M., Mahdavi, I., Mahdavi-Amiri, N., & Khorram, E. (2020). Balancing public bicycle sharing system using inventory critical levels in queuing network. *Computers & Industrial Engineering*, *141*, 106277.

Waserhole, A. & Jost, V. (2016). Pricing in vehicle sharing systems: Optimization in queuing networks with product forms. *EURO Journal on Transportation and Logistics*, *5*(3), 293–320.