Predictive Analysis of Bike Sharing System Using Machine Learning Algorithms

Alekhya Bhupati
x18132634
MSc Data Analytics
National College of Ireland
Dublin, Ireland

Jhanavi Govinda Gowda
x18128998
MSc Data Analytics
National College of Ireland
Dublin, Ireland

Sushant Parte
x18137440
MSc Data Analytics
National College of Ireland
Dublin, Ireland

***Abstract*—In the recent years transportation has seen a tremendous growth. The technology has improved so fast that the people can gather all the information and check for the arrival and departure of the bus, train just by using the app. Now a day, Bike sharing system is gaining more and more popularity as people are tending to use public transportation because of low cost and low maintenance and easy to use. Bike sharing system is one such service provided for the people which rents bikes to the people for free or with some charges and allows the people to use bikes for some duration from one bike station to another bike station. In this project, the analysis has been made on the bike sharing system in analyzing the number of bikes rented per hour, per day, in weekdays and weekends. And check the day and the peak hours of the people and increase the number of bikes so that it increases the companys business and yields more profit. The implementation of machine learning algorithms has given outstanding prediction models with good efficiency. Here, the popular algorithms like, Linear Regression, XGBoost, SVR, decision tree, Random Forest, PCA have been applied to check the accuracy, sensitivity of each model compare them and suggest the model with high accuracy. ystem has been given from various research work done in this field.**

Key words: Renting, SVR, regression, prediction, Grid search and Random search

I. Introduction

Bike sharing system is a kind of service provided to the users which allow people to use shared bicycles for free or by paid for a short-term use. Many bikes sharing system allows people to use bicycles from one bike racks, which is also called as Dock and after 30min or 1 hour to return it in another dock of the same system. There are even smart phone mapping apps which manages the bike sharing system. It was started by Luud Schimmelpennink in Amsterdam by implementing 50 bicycles. But within a month the bicycles where last due to poor management which failed in keeping track of bicycle being used. In order to overcome this problem a new innovation called smart technology was adopted. The users will be provided with a card, to which they can load money, it acts as the access for the bicycles and record the login and logout time of user at one station and record if the user delays to return the bicycle and applies charges to it. But in the business perspective if the company will know the users pattern of using bicycles in weekdays and weekends, hours and days, it becomes easy for the company to analyze the peoples behavior and draw conclusions. If there are more people who rent bicycle in weekends, then the company can focus more on weekends and may increase the number of bikes on those days. There was a major necessity for the automation of the system which gives the information regarding the customers, the bicycle being used and should adopt a model which predicts the number people using bicycle on a daily basis, percentage of people renting bikes hourly, daily, in weekdays and in weekends. It should analyze the fluctuations in the usage and the demand of the user and place more bikes at the place where there is more demand and where the crowd is more. And should also predict the same as this will help to improve the business.

This led to development of machine learning in bike sharing system, in recent years machine learning has attracted number of researchers to carry out research in this field. It has implemented many algorithms like SVM, KNN and so on. But know the concentration has been diverted to the use of deep learning system which is a part of machine learning system. This deep learning model extract different features from the dataset and finds the pattern which user follows using training data and predicts the system which gives way to the company in making profit.

In this project, a predictive model has been built using both machine learning and deep learning algorithms like SVR, decision tree, random forest XGBoost. The output is compared between different types of algorithms applied and checks for accuracy, sensitivity and specificity of each system. The system will implement the algorithm which is the best in the above-mentioned features.

Research Question

How well the machine learning models evaluate the performance for bike renting system?

II. Related work

The article reviewed [1] the existing methods for bike sharing and categorized them according to the usage in each state of the continent reviewed in the article. The author refers to the effect of socio environmental factors that affects the behavioural trend of bike sharing which considers the model for emission free transportation. The statistical data analysed,

and past data evaluated knowledge on growing body of bike sharing. The four generation of bike sharing reviewed resulted in bike sharing has enlarged in lot of different continents. The statistical analysis can be reviewed in the article [2]. The results predict daily prediction of adults usage of bike sharing in United States of America where the dataset used has categorical variables according to the age and hourly and monthly usage of bike sharing with the area variable which has city and location. The result predicted gives information as comparison of metropolitan regions and small areas. The accuracy evaluated predicts the percentage of bikers in the area proposed. Evaluation was proposed on regression metrics as R square was calculated. The models implemented in the project were analysed from the article [3] for the study area of bike sharing in San Francisco. The models used in the proposed articles are of two categories multivariate (Random forest and Least Square Boosting) and univariate (Partial Least Square algorithm and multivariate regression algorithm). The results showed an effective evaluation as univariate models have low error prediction as compared to multivariate. The factors which were considered in the monitoring data were population, noise and greenhouse gas emissions. The evaluation metrics used were MAE (Mean average error). For the further modelling like prediction of the future bike rentals article [4] can be reviewed to predict the return and rental of the bikes in study area of Austria. The models proposed in the article are Poisson, Negative binomial and hurdle models. The models were compared depending on the factors as wind speed, temperature and hours where Poisson model proved to be best for forecasting the future rentals. The article [5] predicts in the bike sharing system the usage of bikes using four models SVR (support vector Regression), Ridge Linear Regression, Random forest and gradient boosting. The models are being evaluated using the Root mean square logarithm error (RMSLE) and cross validation was performed for comparison of the models where it was found that random forest was predicted as the best in both terms prediction accuracy and training time. The proposed article [6] specifies the XGBoost as most powerful statistical models which is used to detect the nonlinear patterns in the data sets having missing values and outliers. The data exhibits classification in two categories as healthy patients versus patients with epilepsy. The model is evaluated with to curves as Area under the curve (AUC) which predicted 100% of score and receiver operating curve (ROC) was plotted with true positive. Random forest was implemented for the regression model implemented in the article [7] to make resources of the dataset to derive result and to derive the maximum depth of the dataset. The evaluation result proved that random forest predicts good performance results.

III. Exploratory Analysis and Hypothesis

Exploratory analysis is the process of performing critical analysis and doing initial investigations over the data.
The graph shows the record of number bikes rented per month.
The factors which influence the bike sharing system is:

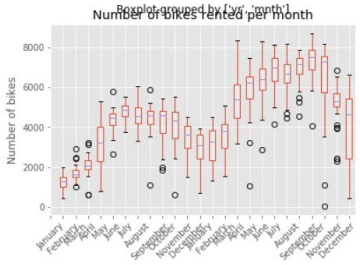


Fig. 1. Number of Bikes Rented Monthly

- **Daily trends:** The comparison is made between casual and registered users. It is giving the correlation between the total bike Rentals and years.

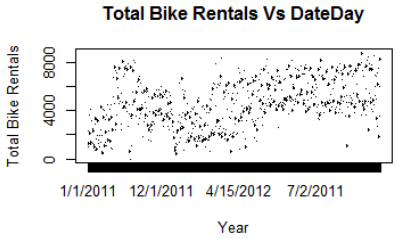


Fig. 2. Total Bike Rented as Per Date

- **Temperature:** It gives the correlation between the rental bikes and how it is distributed for different temperature.

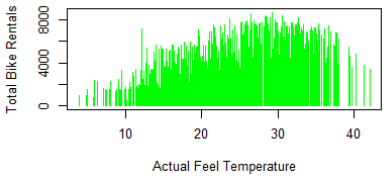


Fig. 3. Actual Temperature vs Bike Rented

- **Wind Speed:** The graph gives the relation between rental bikes and the wind speed. If wind speed is increased, then the number of bikes taken for rent is less.

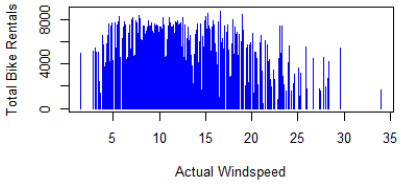


Fig. 4. Actual Windspeed vs Bike rented

- **Season:** The graph of seasons and rental bikes shows that, the number of rental bikes in winter is very less when compared summer as the temperature drops.

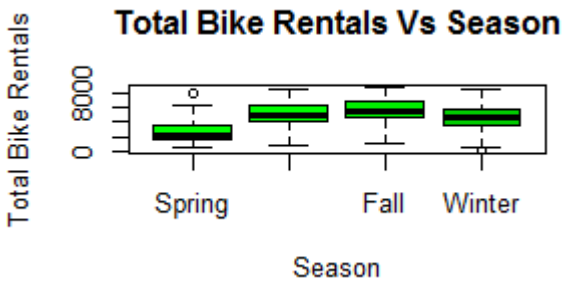


Fig. 5. Total Bike Rental vs Season

- **Rain and snowfall:** The relation between rain and snowfall show that during rainy and winter season the renting of bikes is gradually decreased and in normal days it is normal. The graph shows that when the weather is good, clear and sunny, the number of bikes rented is more. And when it rains the number of bikes rented is less.

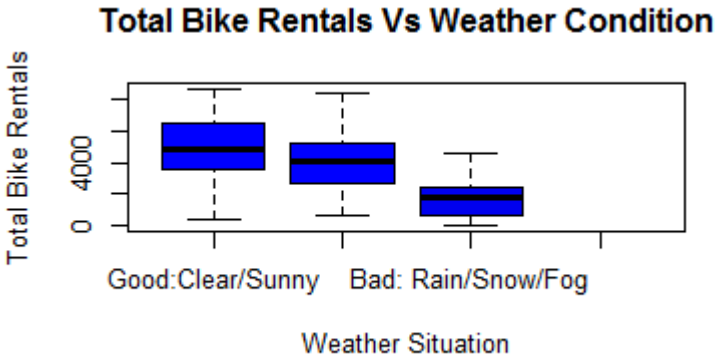


Fig. 6. Total Bike Rental vs Weather condition

IV. Methodology

It provides the mechanism of the predictive models applied on the dataset and the approach which is carried out to fetch different output from different models. The project follows CRISP-DM methodology.

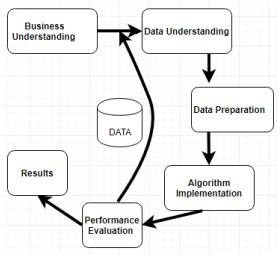


Fig. 7. CRIPSP-DM Methodology

A. Business understanding

Bike sharing system provides service to the people which rents bikes and allows the people to use bikes for some duration from one bike station to another bike station. Analyzing the system will help the company to check the peak hours, weekdays and weekend business and the peoples pattern of renting bike which help the growth of the business.

B. Data understanding

The dataset taken for the analysis contains daily rental data from 2011- 2012. Each algorithm which is applied will divide the training and test data accordingly. The training dataset contains attributes like customer details, bike details and temperature, humidity, wind speed which mainly determines the peoples pattern of renting bikes. It has registered users means the users who are the members and regularly use bike sharing system and casual users who are not the members of the system.

The Dependent variable is count which gives the total bikes rentals and independent variables in the dataset are temperature, humidity, wind speed, weekend and weekdays which gives the count of number of bikes rented.

V. Implementation

The algorithm such has Random Forest, SVR, XGBOOST, Decision Tree, Linear Regression have been implemented for the dataset and it is compared. The dataset is first preprocessed with the process of dropping of few columns like date and instant, then the data is split into train and test by using train-test-split function from SKlearn package.

```
In [14]: M #Benchmark

         # Training SVR
         svr = SVR()
         lr = LinearRegression()
         dtr = DecisionTreeRegressor()
         rfr = RandomForestRegressor()
         gbr = GradientBoostingRegressor()

In [15]: M fit1 = lr.fit(X_train,y_train)#Here we fit training data to Linear regressor
         fit2 = dtr.fit(X_train,y_train)#Here we fit training data to Decision Tree Regressor
         fit3 = rfr.fit(X_train,y_train)#Here we fit training data to Random Forest Regressor
         fit4 = gbr.fit(X_train,y_train)#Here we fit training data to Gradient Boosting Regressor
         fit5 = svr.fit(X_train, y_train)#Here we fit training data to Support Vector Regressor

         C:\Users\Alekhya Bhupati\Anaconda3\envs\newenvt\lib\site-packages\sklearn\ensemble\forest.py:245: FutureWarning: The defaul
         t value of n_estimators will change from 10 in version 0.20 to 100 in 0.22.
           "10 in version 0.20 to 100 in 0.22.", FutureWarning)
         C:\Users\Alekhya Bhupati\Anaconda3\envs\newenvt\lib\site-packages\sklearn\svm\base.py:193: FutureWarning: The default value
         of gamma will change from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly t
         o 'auto' or 'scale' to avoid this warning.
           "avoid this warning.", FutureWarning)

In [16]: M print("Accuracy Score of Linear regression on train set",fit1.score(X_train,y_train)*100)
         print("Accuracy Score of Decision Tree on train set",fit2.score(X_train,y_train)*100)
         print("Accuracy Score of Random Forests on train set",fit3.score(X_train,y_train)*100)
         print("Accuracy Score of Gradient Boosting on train set",fit4.score(X_train,y_train)*100)
         print("Accuracy Score of SVR on train set",fit5.score(X_train,y_train)*100)

         Accuracy Score of Linear regression on train set 79.72724777012029
         Accuracy Score of Decision Tree on train set 100.0
         Accuracy Score of Random Forests on train set 97.71981929150658
         Accuracy Score of Gradient Boosting on train set 96.73430804419 62
         Accuracy Score of SVR on train set 0.18969341012881058
```

The algorithms which are taken for the analysis are imported from SKlearn package. The trained data is fed into these algorithms and the accuracy for each modelled is been examined.

XGBoost: It is one of the dominant algorithms in machine leaning which gradient boost the decision tree which is fast and gives high performance. It gives highly versatile and flexible for huge data. The portability and compatibility feature of XGBoost will allow the user to implement on any system like, Windows, Linux and so on. The objective function of XGBoost is given by:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Real value (label) known from the training data-set

Can be seen as f(x + Δx) where x = $\hat{y_i}^{(t-1)}$

Random Forest: It is the method for classification, regression and another task. It is the group of decision trees which operates as an ensemble. Each individual tree will give an output that is the prediction class. The one with highest vote will become the predictive model. The principle behind Random Forest is wisdom of crowds. Huge number of uncorrelated models will operate in a group and outperform the individual models, because trees will protect one another from the individual errors. Some trees may be right and giving proper output and some other trees may be going in wrong direction. But as a group they will all perform well.

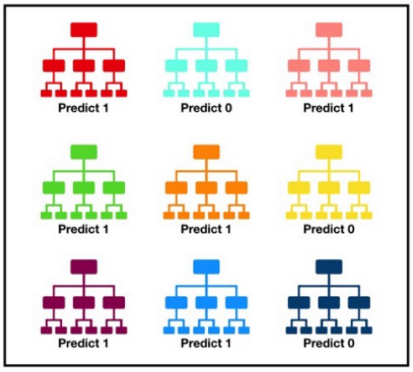


Fig. 8. Prediction class of Random Forest

The prerequisites for the analysis are, the features should have some actual signals so that there can be some random guessing and there should be low correlation between the predictions of individual trees.

Linear regression: The basic machine learning algorithm is being proposed in the project as create a foundation for the project. In linear regression model is implemented to find the relationship between the two or more variables (independent) which is dependent on the continuous variable (dependent). the dependences of the variable categorize the model in 2 parts as univariate as linear regression and multivariate termed as multiple regression. The following equation can be used to represent the model-

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

For training the model into certain data fit line must be implemented. The Predicted values and the observed values must have minimum error rate. The line can be drawn which shows the least error value for the regression model called as best fit line with the error values (residuals).

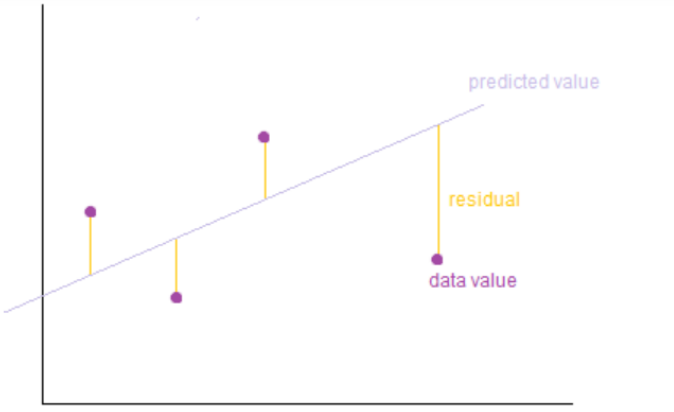


Fig. 9. Linear Regression Fit LIne

The performance can be evaluated using Root mean square error (RMSE) and r square (r2) which determines the level of variance between the observed and predicted values.

Decision Tree: The Decision tree model is influenced a lot of influences in machine learning as it is implemented in classification and regression. The model supports for making decision analysis using a decision tree for which the tree drawn upside down with its root on the top.

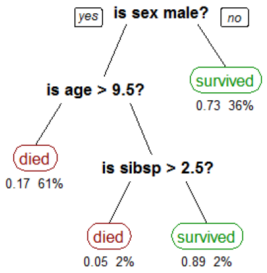


Fig. 10. Decision Tree Model

The above example predicts the proper classification models with black portion as internodes with edges or branch split. The end of the tree is the decision proposed or well termed as leaf. The model is simple to understand with simple dataset but faces challenges in complex data set where the leaf structure does not predict any fruitful information. Decision trees are unstable due to slit variation in data. Proposed project has stable dataset, so it is being implemented to validate the dataset and generate accuracy.

Support Vector Regression: Support vector regression is termed same as support vector machine implementing with linear regression. The model can be used for regression, classification and detection of outliers. The classifier is built to derive a margined to differentiate the hyperplane achieved same as linear regression. This predicts that higher the distance

from the point of class for the training data lower would be the generalization error for the classifier.

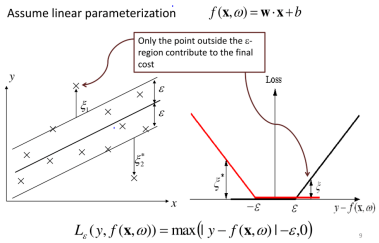


Fig. 11. Support Vector Regression Fit

The above diagram explains as the margin of tolerance is set in approximation to the SVM which already predicted the problem. The algorithm is more complicated therefore is taken into consideration for the proposed project. The validation accuracy for the SVR was evaluated very less therefore to improve the performance Grid Search and Random search was performed on the model SVR.

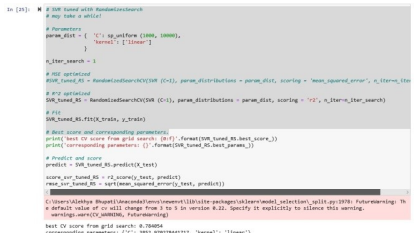
Grid search: The model grid search is a hyperparameter which is a characteristic of the model which external. The hyperparameter is used to improve the performance of SVR model whose value was validated very low as. The hyperparameter cannot be predicted so it comes under the hidden layer of the implementation of the model. Grid search are more reliable in low dimension space. The grid search was used in the data set to predict the accurate results with improvement the model. The model provides grid hyperparameter values for each combination to provide the best combination and derive the evaluated results.

Random Search: The model grid hyperparameter which selects random values from the training model and provide results to the model with selecting close enough values for fewer iteration which effects the boosting the performance. Random search is same as grid search but more efficient than grid search. The efficiency calculated can be determined as same for both hyperparameters models, but random search proves toe more efficient.

Evaluation matrix: The scores obtained by applying linear Regression, Decision Tree, Random Forest, XGBoost and SVR are 0.74, 0.72, 0.83, 0.86 and -0.08 respectively.

When all the algorithms are compared, the scores show that the value of SVR is very low which constitutes about 0.2309 with high Root Mean Square error. Hence it requires fine tuning of the model by applying Grid search and Randomized search. In Grid search model is built for all the combinations of hyperparameters which is specified and evaluates each model and in Random search, different combinations of hyperparameters. The cross validation for Grid search obtained is 0.76

And for random search is 0.78.



These two models help in increasing the accuracy of the system. And increase the efficiency of the system. Upon comparison of different models, the output obtained shows that XGBoost best fits the system with a score of 0.86.

VI. Conclusion

The proposed project, bike sharing system has been implemented with 5 different algorithms like Random Forest, SVR, XGBOOST, Decision Tree and Linear Regression. The accuracy of these algorithms obtained is 0.74, 0.72, 0.83, 0.86 and -0.08 respectively. SVR has the lowest accuracy with high RMS error which needs to be addressed. Hence, it was fine-tuned by applying Grid search and Randomized search. In Grid search model is built for all the combinations of hyperparameters which is specified and evaluates each model and in Random search, different combinations of hyperparameters. These two models help in increasing the accuracy of the system. The cross-validation search obtained for these two searches is 0.76 and 0.78 respectively. Upon all the models implemented, XGBoost gives the highest score of 0.86 and best fits the system. The future work in this project can concentrate on implementing different deep learning algorithms and considering some more factors like traffic which also affect the number of bikes rented.

References

[1] Krizek, B., Barnes, G. and Krizek, K. (no date) Estimating Bicycling.

[2] Ashqar, H. I. et al. (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). IEEE, pp. 374378. doi: 10.1109/MTITS.2017.8005700.

[3] Rudloff, C. and Lackner, B. (2014) Modeling Demand for Bikesharing Systems, Transportation Research Record: Journal of the Transportation Research Board, 2430(1), pp. 111. doi: 10.3141/2430-01.
[4] Shaheen, S. A., Guzman, S. and Zhang, H. (2010) Bikesharing in Europe, the Americas, and Asia, Transportation Research Record: Journal of the Transportation Research Board, 2143(1), pp. 159167. doi: 10.3141/2143-20.
[5] Yin, Y.-C., Lee, C.-S. and Wong, Y.-P. (no date) Demand Prediction of Bicycle Sharing Systems.
[6] R. R. Parmar, S. Roy, D. Bhattacharyya, S. K. Bandyopadhyay, and T.-H. Kim, Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions, IEEE Access, vol. 5, pp. 71567163, 2017.
[7] M. R. Segal, Machine Learning Benchmarks and Random Forest Regression, 2003.