Bike Rides Data Analysis Project

Sai Keerthi Mettu, *200416252,keertthimyname@gmail.com*

Department of Computer Science, University of Regina

Abstract— **With the fast-paced growth of technology development, the transport systems are contributing to a lot of pollution and global warming for the environment. Mindful organizations and governments are taking measures to encourage eco-friendly and affordable transport means for citizens through bike-sharing and rental schemes. The data set available in Kaggle [1] helps to recognize the facts of how far this bike rental system is helpful for the customers and its usage concerning different times in the year 2010 and 2011. The bike rides recorded each day presents the influence of environment and climatic settings on them. Based on the identified trends in the dataset, a machine learning algorithm is applied to discover the flow of rides during usual and typical seasons in a year. A proven model is demonstrated to predict the rides count for any conditions on a day. written by Matthias Bussonnier.**

I. Problem Statement

The number of bike rentals occurring each day would vary based on different factors like holidays, off-seasons and other normal working days. The main objective of this analysis is to define the uncertainty of these rides against the features of bike rides that have a major impact and correlation in the ride count determination statistically.

II. Background

There will be everyday activities like to perform services, troubleshoot the hardware/software problems and keeping track of all records. This is important for these bikes sharing business organizations for their growth and revenue. Maintenance people are responsible for placing a certain count of bikes available for customers in various stations (especially from service warehouses to busy stations) every day. They are responsible for reckoning the number of bikes that are supposed to be placed at the busier stations from the normal bike parking stations. The ambiguity in bikes count can be estimated based on the bike rides data trends. A real-time traditional bike rides dataset was captured in a city by the UCI researchers and stored it in their archive repository [2]. In order to work on this problem, previously an attempt was made to look into the live data following in the project when I was a working teammate in a software company. But the exact features were not possible to capture that directly affects the bike rides count per day. Now an opportunity has come to distinguish trends and able to implement a best functional model using machine learning algorithms on this bike rides data.

III. Development of Model

As we are trying to predict a numerical quantity of data, i.e., the bike rides in a locality, the scientific approach leads to the regression-based model algorithms. They are also of the type Supervisory Machine Learning algorithm.

The bike-sharing dataset follows a Guassian distribution (normal distribution) of data trend when plotted the rides count against the months in a year. Fig. 1 For any given dataset, the best algorithm is declared on the basis of greater accuracy percentages on data fitting with reduced anomaly rectifications and data normalizations. Thus, an ideal model is elected for bike-sharing data (providing minimal manipulations on the original data with increased model efficiency), based on the assessment of list of regression models to predict a numerical feature in this data set. Random Forest regression is the only algorithm that worked like a paramount on it. Fig. 2

Random Forest regression algorithm is known best for its ensemble type of decision tree results. It averages the list of results that are derived from the estimated decision trees unbiasedly and draws an output [5].

The following are the main steps in the development process of RF regression model.

1) Loading the required python libraries specific to the algorithm model.
2) Importing the scrubbed data set that is ready to fit within a model.
3) Classifying the entire dataset into two parts i.e., training data and test data in such a way that data has to be separated as labelled sets and unlabeled sets respectively. This classification should be in the ratio of 80:20.
4) Train the algorithm with the labelled instances.
5) Predict the unlabeled scenarios in the test data,their accuracies and compare the predicted outputs with the actual outputs.

IV. Evaluation of Model

Random Forest Regression is a model that can be best evaluated through the accuracy percentages contrast with the training data set and the acquired predictions on the test data set. It is defined that RF is an ensemble type algorithm gets the unbiased and stable results from the decision trees as the final output [3]. The only drawback of this regression model is that it is not possible to visualize the data fitting to the model clearly as that of other models (E.g. Linear Regression). On fitting the model, its been seen that the training data accuracy stood at 98.4% while the accuracy of the predictions at 86.3%. Although there is a deviation of 12% between the actual and predicted results, this is the only algorithm that contributed to the outstanding accuracies of all the other models. The deviation in accuracies are visually seen in the Fig. 3 and Fig. 4 for the predictions against test data set and the train data set respectively.

Apart from accuracy, performance is another evaluation metric of a model by finding the Mean Absolute Error (MAE) and Root Mean Square Error values (RMSE) [4]. But these values identification cannot be applicable as valuation factors on this dataset, because of the reason that there are nine independent features (have normalized values ranging between 0-12) that controls a single dependent feature (values ranging between 25-8800). Thus, the only limitation of this model is that the MAE and RMSE values would be relatively high and cant be used as a judging criterion for the model evaluation [3]. Fictional instances of any given weather scenario on a particular season day would be an elite way to cross verify the predicted bike rides count results with the actual day types similar results.

V. CONCLUSION

It is eternally agreeable that bicycle rides in a city are totally impacted by its environmental conditions. In this document, the Random Forest Regression model is explained to forecast the number of bicycles rides count and how it is preferred as the finest of any other models for this bike-sharing dataset. The advantages and disadvantages of this model usage are fairly reviewed. Further improvements like change in the number of decision trees and re-estimating the model accuracies can be performed.

VI. APPENDIX

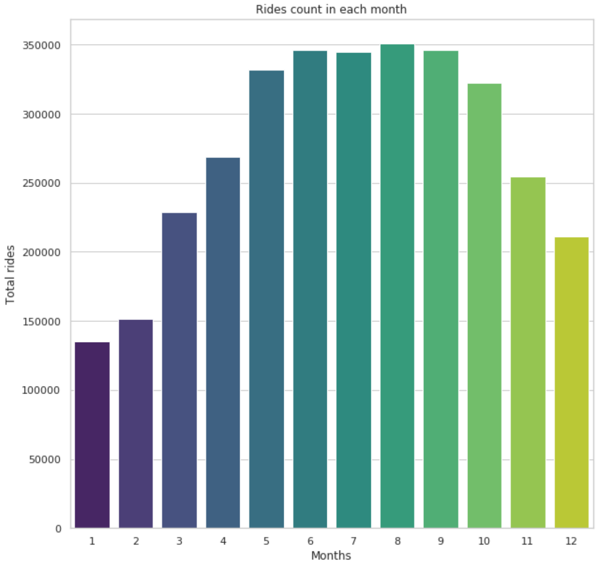

Fig. 1. Bike Rides count vs Months.

REFERENCES

[1] N, Lakshmipathi. UCI Machine Learning Repository: Bike Sharing Dataset Data Set. Uci.Edu, 2011, archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. Accessed 27 Aug. 2019.

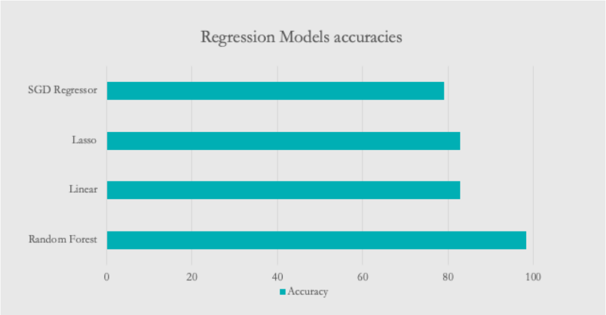

Fig. 2. Regression Model Accuracies

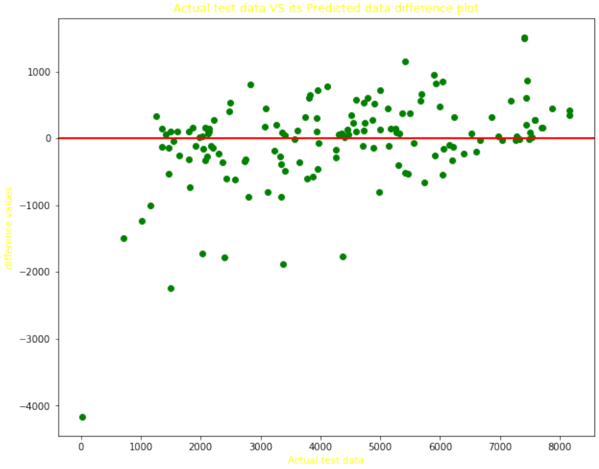

Fig. 3. Actual test data VS its Predicted data difference plot

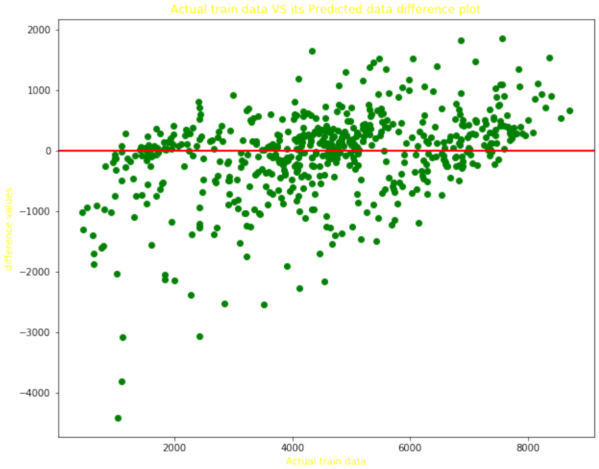

Fig. 4. Actual train data VS its Predicted data difference plot

[2] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, doi=10.1007/s13748-013-0040-3, Springer Berlin Heidelberg. Accessed 27 Aug. 2019.

[3] "3.2.4.3.2. Sklearn.Ensemble.RandomForestRegressor Scikit-Learn 0.20.3 Documentation." Scikit-Learn.Org, 2018, scikit-learn.org/stable/modules/generated/ sklearn.ensemble.RandomForestRegressor.html. Accessed 27 Aug. 2019.

[4] Malik, Usman. Random Forest Algorithm with Python and Scikit-Learn. Stack Abuse, Stack Abuse, 13 June 2018, stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/. Accessed 27 Aug. 2019.

[5] Koehrsen, Will. Random Forest in Python. Medium, Towards Data Science, 27 Dec. 2017, towardsdatascience.com/random-forest-in-python-24d0893d51c0. Accessed 27 Aug. 2019.