COMP5310 Principles of Data Science, 2019 sem2
Assignment 2 - Experiment, Quantify, Report

Lavanshu Agrawal
490617999
lagr7305

Manisha Gupta
490537365
mgup6878

1a. Research Problem:

Bike sharing is a new trend that is widely popularising in metro cities as it provides environment friendly and economic ways of transportation in the current time of high fuel prices and rising congestion allowing an individual to rent bikes for their work or leisure purposes. The following report outlines an analysis of a research done over the bike sharing systems in order to predict the demands of the bicycle (dataset provided by Capital Bikeshare) as against to features like day of the week, time, season, temperature, wind speed, humidity, etc (Kaggle, 2015). This would help the companies running the bike sharing systems to ready their inventories and further logistics as against the variable bike-rental demands depending over several factors.
The following research objectives are sought to be answered with the analysis of the dataset:

1. Forecasting the demand of the rental bikes.
2. Analysis of different regression algorithms and computation of their performance metrics.

In addition to the above research questions, the study also examines the following hypothesis:
Null hypothesis: Linear Regression gives better performance than the other models.
Alternative hypothesis: A better performance can be achieved than that obtained by the Linear Regression.

1b. Evaluation Setup:

The problem is hand is a typical multivariate regression problem.The parameters that have been used to measure the performances of the models are the mean-squared error, mean-absolute-error and the r-square value. Mean-squared-error calculates the average of the squares of errors between the actual values of the target variable and the values predicted by the model. Mean-absolute-error calculates the average of the difference between the true and the predicted values of the target variable. The r-square value represents the goodness of fit of the model, i.e., the variance of data explained by the model (Swalin, 2018). These estimators are used together to assess if the models applied provide a reliable prediction of the target variables.
Dataset: The 'train.csv' dataset has been used for this assignment (Kaggle, 2015). To prepare it for further analysis, first the missing values were checked (did not contain any missing values) and outliers were adjusted. The dataset has also been tweaked to contain only the useful features by performing feature selection and removing the features that do not have any significant contribution towards the target variable. Further, the dataset has been sub-divided into train set and validation set for the purposes of training and validating the performance of the various models.

2. Approach Description:

The research questions deal with the regression problem of forecasting the demand (count) of the bikes as against to different features. The dataset was initially analysed using exploratory data analysis (EDA), processed and feature-engineered (steps mentioned below).
After getting the data to the desired format (containing training and validation sets), the assumed benchmark model **Linear Regression** model, as per the null hypothesis, was applied to the training set and its performance on the validation set was observed to be showing an **MSE of ~7746.80**

MSE, MAE of ~66.365 and **R-squared value of ~0.675** which was taken as benchmark. Linear Regression models the data using a linear relation between the output variable and the input variables. Various other multivariate regression models like **Support Vector Regression, Bagging Regression, Random Forest Regression** and **Keras Sequential model** were then employed in order to check whether they can outperform the existing benchmark model based on the performance metrics.

The approach followed is described below:

> **Understanding data and handling missing values:** EDA was done (previous assignment) in order to understand the features contained by the data and the correlation they hold with the target variable. There were no missing values in the dataset.

> **Feature engineering:** It was observed that datetime column would not add much value while predicting the count. Hence it was split into year, month, day and hours. Based on the EDA, the casual and registered columns were also removed.

> **Outlier analysis:** EDA demonstrated that the data was skewing, so outliers were removed using the **Interquartile Range (IQR) rule** (i.e., upper bound = 0.75 quartile + 1.5*IQR and lower bound = 0.25 quartile - 1.5*IQR).

> **Handling categorical values:** Categorical features in the dataset 'season', 'month', 'hour', 'holiday', 'day', 'workingday', 'weather', 'year' have been **one-hot encoded** so as to facilitate the model in correctly interpreting the numerical values of these features.

> **Feature selection and tuning:** Lasso regression has been used in order to identify and eliminate the features that do not have a significant contribution towards predicting the target variable. The regularisation property has been used where lasso regression assigns high value of λ using cross validation in order to minimise the regression penalty (Starmer, 2018).

> **Train and validation sets:** The pre-processed data is divided in order to train the models using 'training' set and then evaluate their performance using the 'validation' set.

> **Models implementation:** The performance of benchmark model was first calculated and thereafter all the other models were trained using train set.

> **Evaluating the models:** The validation set was parallely used to evaluate the remaining models using validation set. The values of MSE, MAE and r-square were calculated for each of the model.

3a. Result Presentation:

The following table (Table 1) enlists the results observed using the different models. The mean-squared-error, mean-absolute-error and the r-squared values obtained by the different models are also enlisted below. It could be observed that the linear models do not perform as good as compared to the other models and have a high difference between the values of their evaluation parameters:

Table (1) Evaluation results of different models

Model Name	**MSE**	**MAE**	**R-squared**
Linear Regression *(Benchmark)*	~7746.80	~66.36	~0.68
Linear SVR	~9535.53	~69.38	~0.60
Bagging Regressor	~3177.97	~37.02	~0.87
Random Forest Regressor	~2731.93	~34.43	~0.89
Keras Sequential	~1541.73	~25.41	~0.93

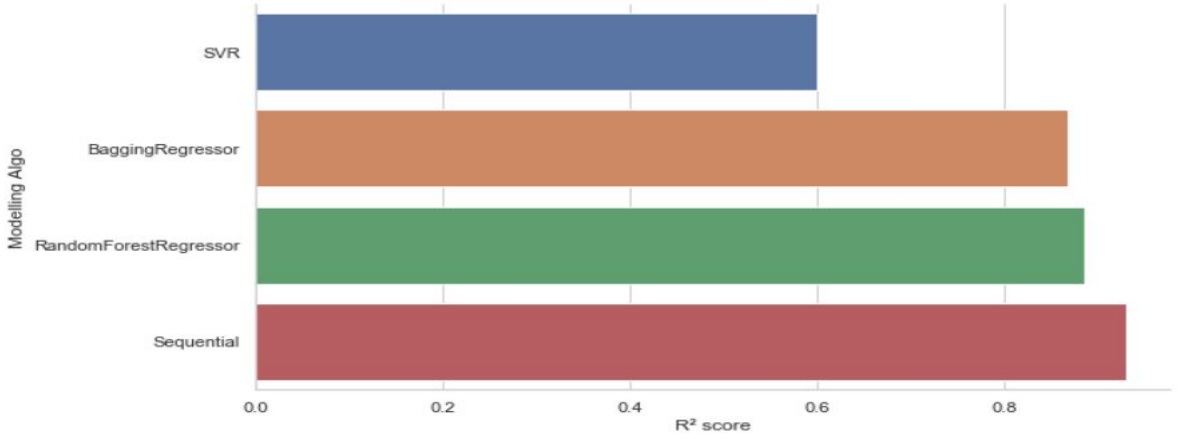

Fig (1) R-square values for different modelling algorithms (Factor plot)

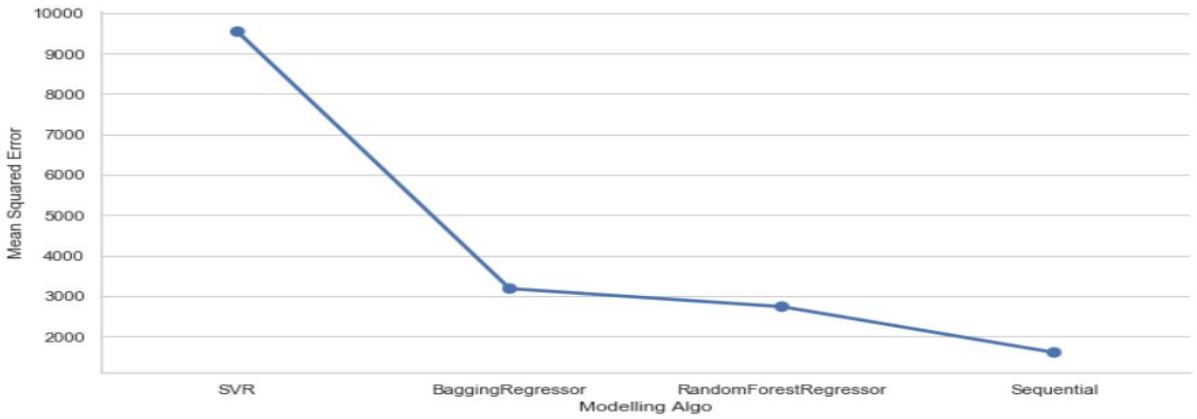

Fig (2) Mean squared errors for different modelling algorithms (Point plot)

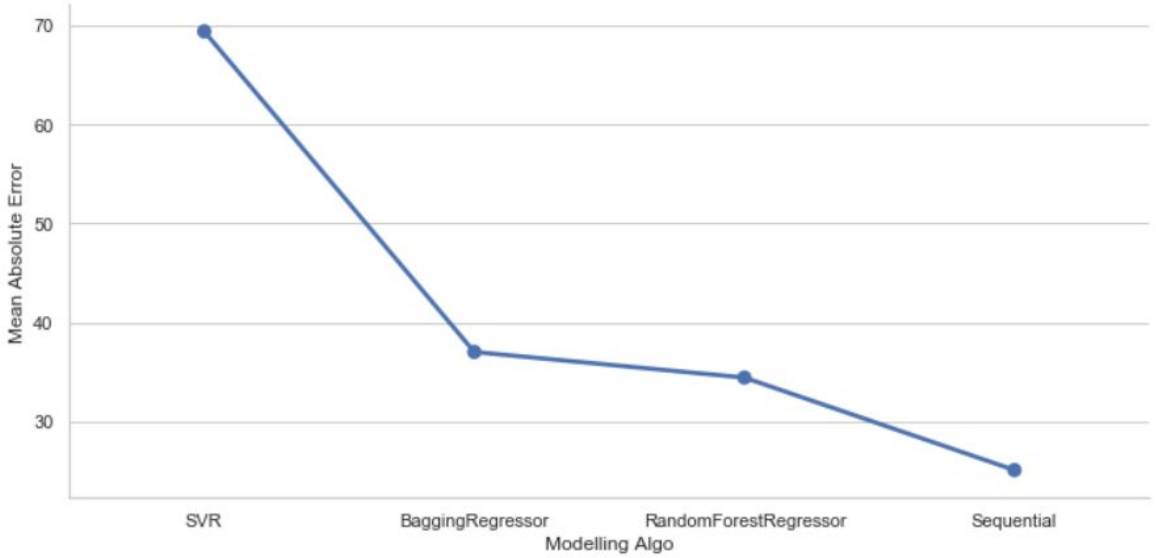

Fig (3) Mean absolute errors for different modelling algorithms (point plot)

3b. <u>Result Analysis:</u>

Out of all the models applied, the Linear SVR performs worse than the benchmark model. All the other models, i.e., Bagging Regressor, Random Forest Regression and Keras Sequential model perform much better than the benchmark model. The Keras Sequential model gives the best

performance out of all the models applied with a low MSE of 1541.73, MAE of ~25 and high r-square value of ~0.93.

Underlying function: The linear regression is observed to outperform the linear SVR, while all the other regressors give a significant high performance than the linear regression models. This implies that the dataset has a non-linear underlying function that relates the features with the output variables. It can also be concluded that the higher the complexity of model, more generalised is the model which made the Sequential Model outperform other models.

Limitations of data: The data available for the problem included a high number of skewed data-points. To overcome these outliers, a direct approach could not be implemented to remove the outliers; and thus the IQR rule was implemented to choose the significant data-points lying close to the median value.

Setup Analysis: The strength of the setup was the feature engineering. The Lasso Regression helped to identify the significant features from the large number of features obtained after the one-hot encoding. This enabled to gain a high performance as against to earlier case where Linear Regression (benchmark) was giving R-squared value of only 0.40.

Possible Improvement: The project has room for improvement by using a more efficient way of handling the skewed data (outlier removal). Also, as Keras Sequential model is observed to be giving the best results, further neural networks' models could be employed to check if the performance increases.

4. Conclusion and Recommendations:

The Keras Sequential model gives the best values of the performance metrics and is observed to clearly outperform the other models. The benchmark model (Linear Regression) seems to be not suitable for the data as the non-linear models outperform the linear models by a great extent and thus we can reject the null hypothesis.

As a Data Scientist, this study gave an insight towards the working of multivariate regression models and their evaluation techniques. It allowed to learn the importance of feature selection and how it can help improve a model's performance. The resultant Keras Sequential model can be applied to any new set of test data to generate a reliable output as it is a more generalized model. Thus the predicted values on a new test set could be relied upon as the model gives a very good performance. For instance, the 'test.csv' has been predicted at the end of the code.

References:

1. Kaggle, 2015. Bike Sharing Demand, Kaggle, viewed 19 October 2019, https://www.kaggle.com/c/bike-sharingdemand/overview
2. Swalin, A., 2018. Choosing the Right Metric for Evaluating Machine Learning Models — Part 1, Data Institute, viewed 21 October 2019, https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4
3. Starmer, J., 2018. *Regularization Part 2: Lasso Regression*, available at https://www.youtube.com/watch?v=NGf0voTMlcs, accessed on 27 October 2019.