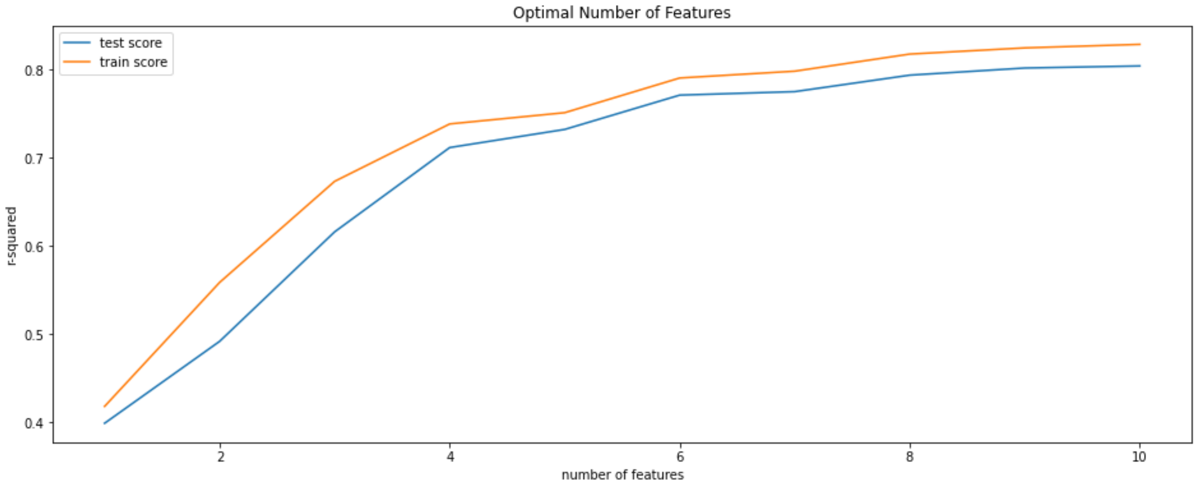
Due to the corona pandemic, many businesses worldwide were shut down, and many are still facing a considerable drop in revenues. One such example is BoomBikes, a US-based bike provider for renting. The bike company is finding it difficult to sustain itself in the present market. Therefore, they want to change the strategy of their business plan to accelerate the revenue collections. Specifically, they tried to know the factors most affecting the demand for bike usage in the US market.
So the findings will be
i) significant variables for the business
ii) How well the model describes the business.
With the previous year's record of bike-sharing, the company made a dataset on the daily bike demands based on some features of bike rent.
The dataset was downloaded from the Boombike dataset from Kaggle [1]. The data features were taken from the data dictionary.

Methods:

The model was built by taking the 'cnt' as a target variable. The variables 'weathersit' and 'season' have numeric data 1, 2, 3, 4 …, with specific levels described in the data dictionary. So, converted these feature values into categorical data. The recursive feature elimination (RFE) process was used to select the features. Then, the cross-validation (CV) with linear regression will fit the model for a good score for the business model.
The linear regression model was chosen to determine the 'variables' strength of connections. The regression analysis tells how much the variables are explained in the model by measuring the R-square value. It also helps to know the most statistically significant variables for the model. 70% of training and 30 % of test data were split from the primary dataset.

To select training and test sets, we can follow a) simply split into train and test: but the result is dependent on the train-test split. b) Split into train, validation and test sets: it will reduce the number of training data. And finally, c) Cross-validation: Split into train and test, and train multiple sets by sampling the train data. The CV is used as a limited sample for better model performance. GridSearchCV is finding the optimal factors with fitting five folds for each of 10 candidates.

Results:



In the above graph the train and test score are very low with 1 to 4 features. As the number of features are increased the R-square score also goes high. So, with the 10 number of factor the R-square value is quite high. Therefore for the linear regression model n_features_optimal = 10 taken. And fit the model with train data. Finally, the trained model is used to predict and calculate the R-square score for the test dataset. That gives 0.83, which is a very good score for the train and test datasets.

Conclusion:

The R-square value is satisfactory and the most significant feature to describe the model are 'yr', 'holiday', 'atem p', 'hum', 'windspeed', 'season_2', 'season_4','mnth_8', 'mnth_9' and 'weathersit_3'.

Bibliography/References

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

```
@article{
        year={2013},
        issn={2192-6352},
        journal={Progress in Artificial Intelligence},
        doi={10.1007/s13748-013-0040-3},
        title={Event labeling combining ensemble detectors and background knowledge},
        url={http://dx.doi.org/10.1007/s13748-013-0040-3},
        publisher={Springer Berlin Heidelberg},
        keywords={Event labeling; Event detection; Ensemble learning; Background knowledge},
        author={Fanaee-T, Hadi and Gama, Joao},
        pages={1-15}
}
```