A sales forecast model

DATA MINING FINAL PROJECT

Oscar Iannucci s4843592, Nikita Matsnev s1082442
Radboud University, The Netherlands

Jan 9, 2022

ABSTRACT

In this paper we build a model for sales predictions based on a list of video games with sales greater than 100,000 copies by using a python language code. We use 2 algorithms in order to achieve our goal, namely gradient boosting and random forest regression. Finally, with the help of SciKit Learn library we also randomly split the data into training and test sets and further validate. Results show that the 2 algorithms perform differently but both follow the same trend line as the actual values.

*K*eywords: Video games industry · Random Forest regression · Gradient boosting regression · Sales prediction

1 Introduction

Companies are increasingly using data to boost their sales and be more competitive. One of the challenges that they might encounter is to select the right data and attributes to predict their sales and especially formulate accurate predictions. In order to achieve the aforementioned result, different methods can be applied, from easier to more complex ones, but some methods can be better than others.

In this paper we are proposing 2 types of algorithms in order to predict sales from a dataset with video games sales: gradient boosting and random forest regression (described in section 2). Our predicted variable is specifically Global sales and our predicting variables are: Platform, Year of making, Genre and Publisher.

We are using these 2 algorithms and comparing them since they both use different ways of prediction and, additionally, random forest regression can be sensitive to noisy data and create overfitting results. The backup in that case would be gradient boosting.

2 Application domain, research problem & methods description

2.1 Application domain & research problem

Our application domain is the video games industry.

We will apply gradient boosting and random forest regression in order to predict Global_sales and compare the predicted results with actual results to see which algorithm performs better. We will train our gradient boosting tree and random forest regression trees based on the aforementioned attributes.

2.2 Research methods

- Gradient boosting:

  Gradient boosting is a ML technique which is highly recommended for prediction purposes. Specifically, this algorithm belongs to the tree-based algorithm categories and its main strength is that is suitable for weak learners, and it can reduce bias. In other words, this algorithm produces a strong learner with high accuracy by using an additive training method which implies the combination of multiple trees defined as "weak learners". A graphic example of a gradient boosting tree algorithm can be found in figure 1.

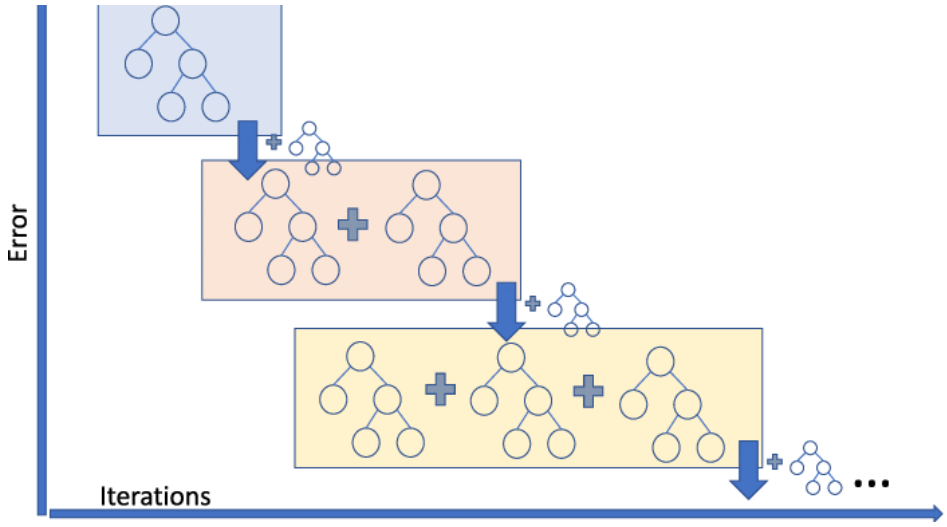

Figure 1: Schematical representation of gradient boosting regression in regards to algorithm iterations (Research Gate, 2021)

  Furthermore, the algorithm is fast, does not need any preprocessing and shows robust results.

- Random forest regression:

  Random forest regression is a model which combines many decision trees, not related to each other, into one unique model, therefore we can define it a meta-estimator. Moreover, this is a bagging technique, therefore it does not use any boosting method. As the name suggests, random forest is based on the classical concept of decision trees algorithm. However, we do know that decision trees generally tend to overtrain data. Random forest solves this issue by averaging multiple depths trees belonging to different parts of the same set, by having the final result of reduced variance. An example of random forest regression algorithm can be found in

figure 2.

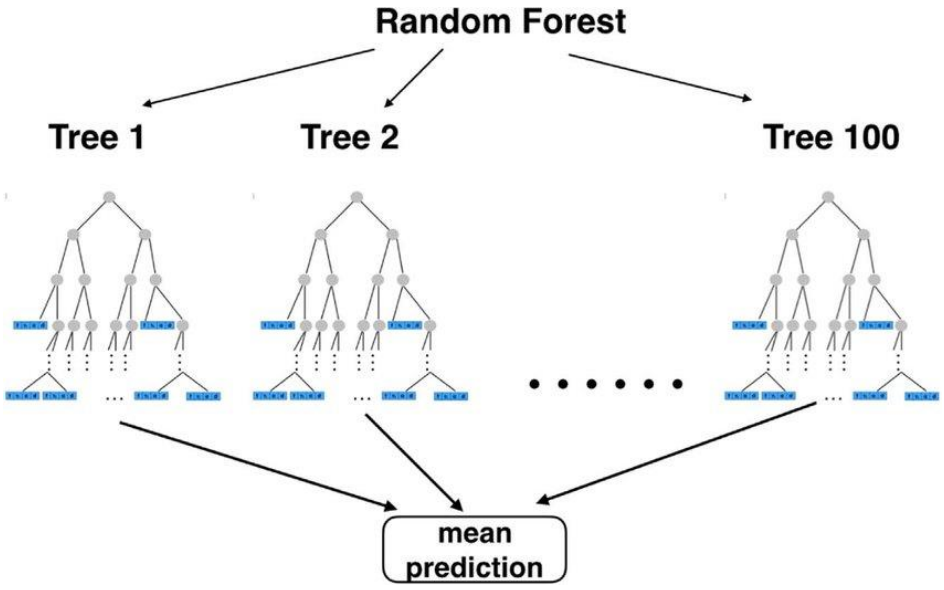

Figure 2: Random Foret Regressor (Research Gate, 2017)

This algorithm advantages include a high accuracy, large data handling and effective estimation of missing data.

3 Related work, similar problems

Several studies already tried to achieve what we are trying to achieve, either with classification problems or regression problems. For example, Henzel & Sikora (2020) use gradient boosting in order to forecast and optimize promotion efficiency in retail. Similarly, Massaro et al. (2021) use gradient boosting in order to predict sales including promotion conditions. The prediction was applied to customer's segments with personalized services according to their purchasing behavior. Also, in other papers (Korolev & Ruegg, 2015; Jain et al., 2015), gradient boosting has been used to predict sales in the pharmaceutical industry based on historical data.

We also have examples of related papers trying to use random forest regression in order to predict sales. For example, Jimenez et al. (2017) used historical available data in order to forecast sales based on random forest regression. Also, Helmini et al. (2019) provides a comparison among different methods for forecasting sales in the retail sector, including random forest regression and gradient boosting.

4 Data set, data collection & preprocessing methods

4.1 Data Set

The dataset includes a list of video games with sales greater than 100,000 copies and the fields included are:

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC,PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)

4.2 Data collection

Data was collected by an archive on Kaggle.com. The dataset of interest was previously generated by a scrape of vgchartz.com. There are originally 16,598 records and 2 records were dropped due to incomplete information. Therefore, we have in total 16,598 rows and 10 columns.

For our research problem we will focus specifically on:

- Nominal attributes: Platform, Year of making, Genre and Publisher.
- Continuous attributes: Global_sales

4.3 Data preprocessing

Before running our 2 selected algorithms, we need to make sure that our data is properly preprocessed:

We selected from the entire set only the attributes we were interested in. Since gradient boosting and random forest regression support only numbers, we converted the string values into integer values and made sure that both training and test data set are in the same numeric format.

The resulting subset would assume that transformed string values are having a unique integer value assigned. We also centered and standardized the data to give better reliability to the whole model.

Furthermore, we created a Global_sales attributed based on the sum of all sales attributes (EU_sales, NA_sales, JP_sales and Other_sales).

We have removed redundant attributes such as regional sales and sorted everything due to the year in ascending order. And then we obtain data set X and expected output set y. It is possible to see graphically explained in figure 3 the whole preprocessing step.

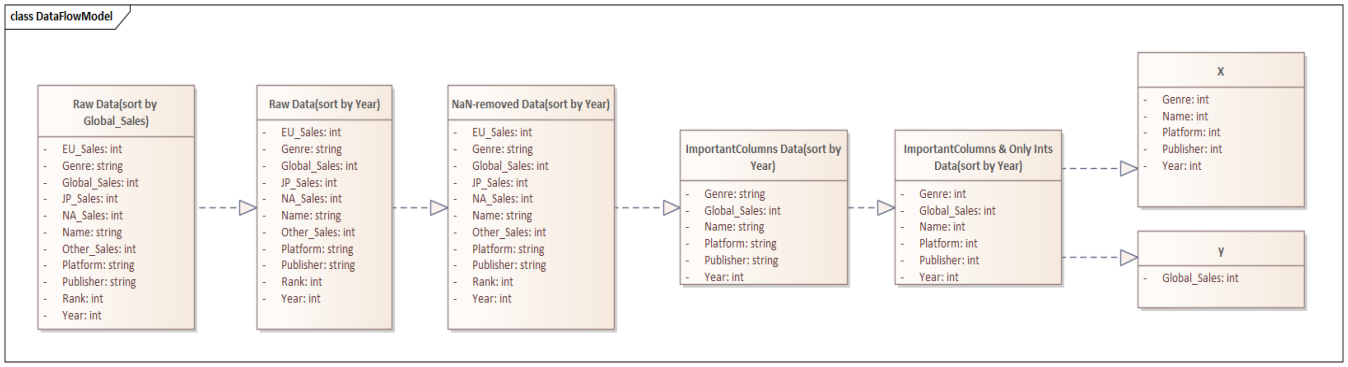

Figure 3: data flow for model

5 Approach/method(s)

5.1 Description

We perform random forest regression and gradient boosting regression on a dataset containing revenues per different world regions of gaming industry, dataset is from Kaggle. As already specified, this is a regression task where we take 4 columns as predictors (namely platform, year, genre, publisher) in order to predict value of 1 column Y (Global sales), with a set consisting of 16597 rows. When plotting the results, we use sales on the Y axis and year of publishing on the X axis, in order to see the sales trends per year of publishing and check how the performance varies YoY.

5.2 Training and test sets

We use the utility of SciKit Learn library “train_test_split” to split given data set in random train and test sets, where we have set up the size of testing set to be 17% as in the large datasets such as we have, it is more relevant to have more training than the default value suggests.

5.3 Accuracy comparison methods

We have used multiple techniques to compare and analyze predictions of these two methods and below in figure 4 we can see the variance importance:

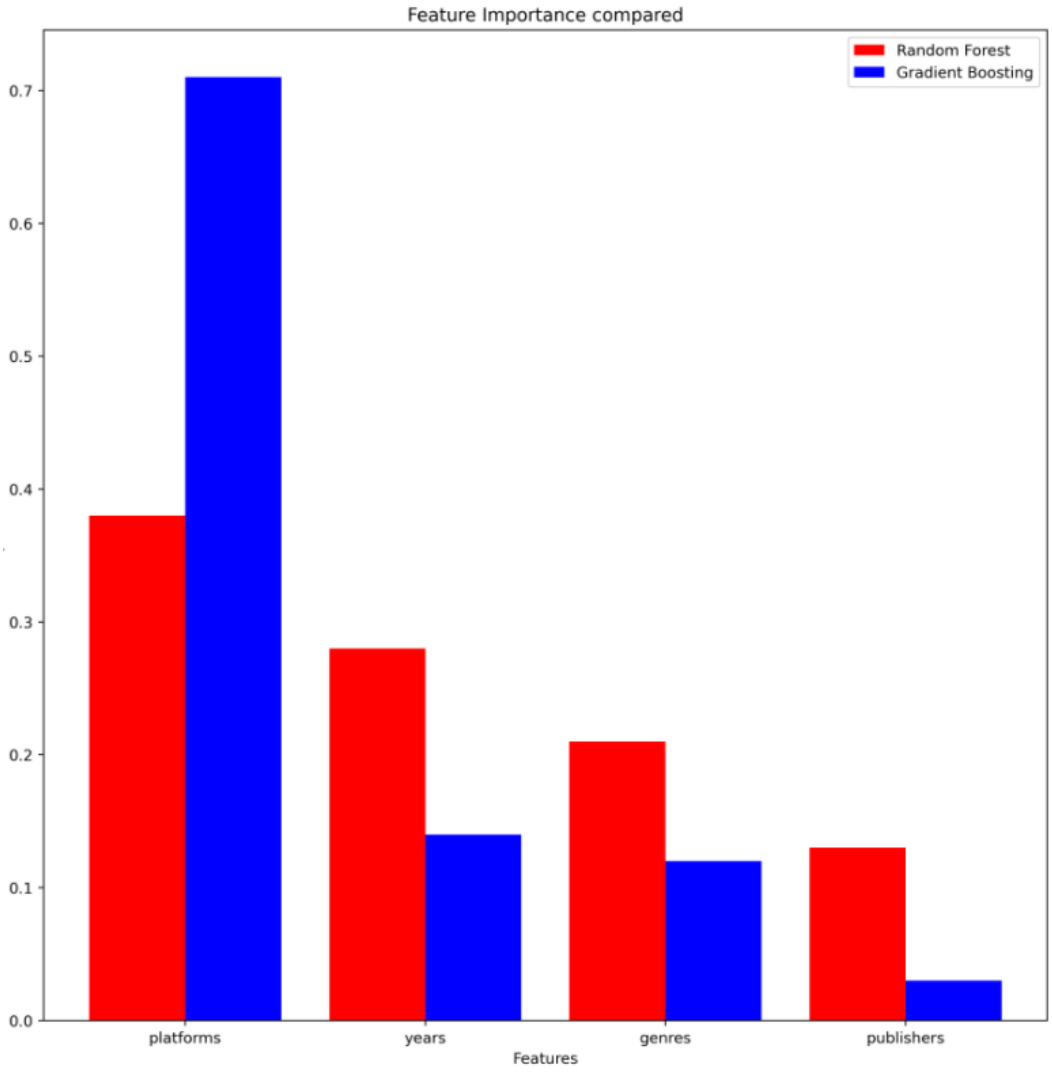

Figure 4: feature importance compared

6 Results

6.1 Random forest regression

By performing random forest regression and plotting the results we can see that trends of predicted values mostly match the actual values. With the regression task, we also defined the importance of each predictor in defining the final prediction:

Variable: years Importance: 0.25
Variable: genres Importance: 0.22
Variable: publishers Importance: 0.15
Variable: platforms Importance: 0.38

Below it is possible to see the plotted results:

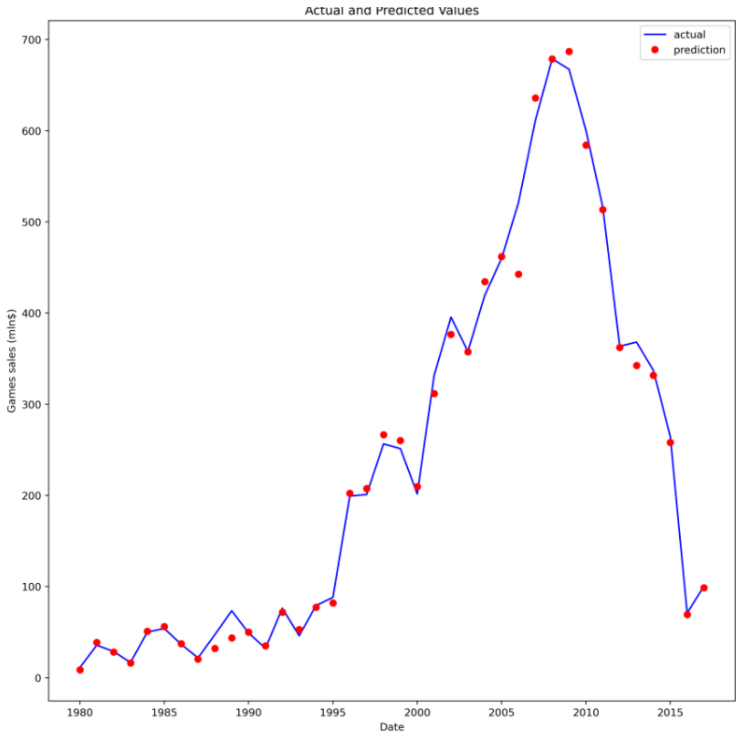

Figure 5. Random Forest Predictions vs. the actual data

6.2 Gradient boosting regression

By performing gradient boosting regression, we can see that it performs mostly based on the publishers' feature instead.

Variable: years Importance: 0.15
Variable: genres Importance: 0.13
Variable: publishers Importance: 0.02
Variable: platforms Importance: 0.70

Below it is possible to see the plotted results:

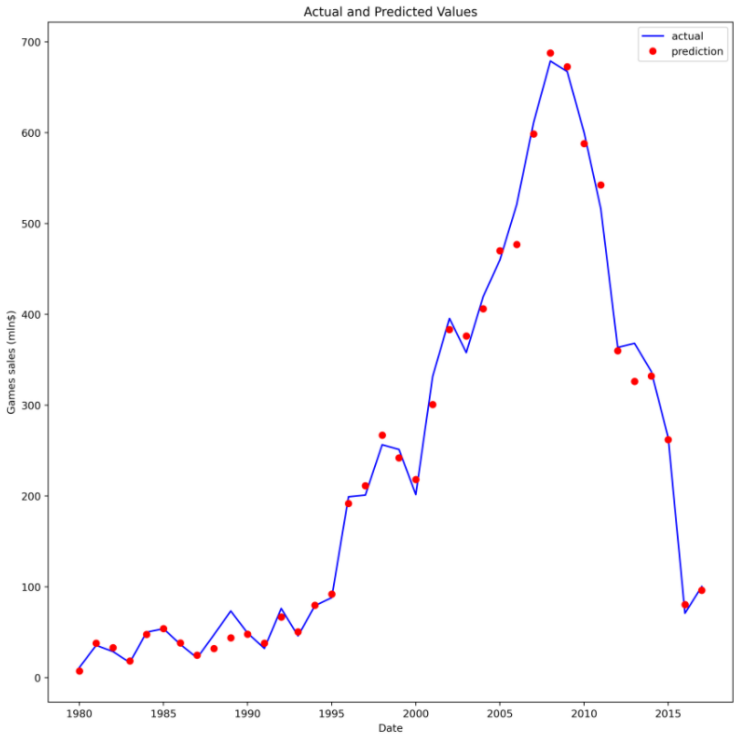

Figure 6. Gradient Boosting Predictions vs. the actual data

7 Analysis of results, conclusions & future recommendations

Surprisingly, both methods perform quiet similar results in all the tests we have managed with some deviations in Coefficient of determination and Explained variance score and also minor fluctuations in MAPE.

	Type	Random Forest	Gradient Boosting
0	Mean squared error	2.265106	2.29512
1	Root mean squared error	1.505027	1.514965
2	Mean absolute error	0.556252	0.558852
3	Mean absolute percentage error	5.47412	5.676698
4	Coefficient of Determination	0.063218	0.050806
5	Explained Variance Score	0.063231	0.050816
6	most important attribute	years	publishers

Figure 7: differences between random forest regression and gradient boosting analysis

To be more specific, and have a clearer picture of differences in performance, let's take a look at the 2

graphs of random forest regression and gradient boosting regression combined:

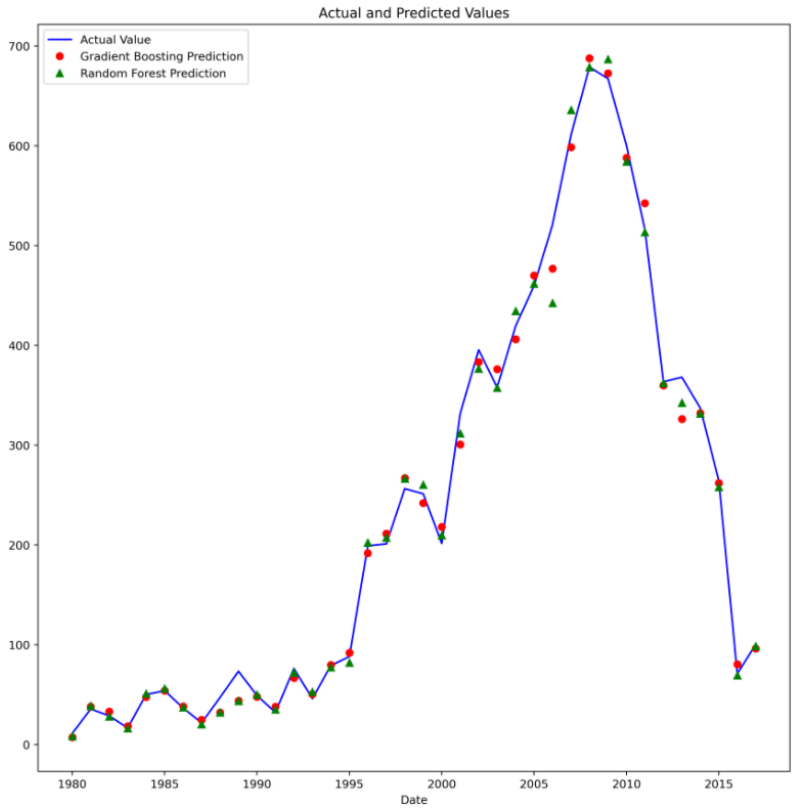

Figure 8: actual and predicted values from random forest regression and gradient boosting regression

We can clearly see that Gradient Boosting, if miscalculating, tends to assume bigger values rather than Random Forest, and that might be due to the fact that gradient boosting bases most of the prediction on only the publishers' attribute, while random forest takes more or less all attributes equally into account.

We can conclude that for this type of dataset with these attributes random forest regression is better than gradient boosting, also due to the fact that RFR calculated the results faster than GB. Therefore, we can recommend random forest regression for similar problems with similar data sets.

What can be done for further research is to compare these 2 algorithms with a classification problem or also with different types of datasets and see if the result of the calculation shows the same/similar results as ours or different outcomes.

To sum it up, we have published our code available here, which is a nice Jupyter notebook guide which can explain you the differences between two techniques more scrupulously and also involve oneself to learn some nice Python code. Henceforth it is worth to mention the libraries we have used: SciPy, SciKit-Learn, Seaborn, Pandas, NumPy, MatPlotLib.

8 References

Video Game Sales. (2016). Kaggle. https://www.kaggle.com/gregorut/videogamesales

Henzel, J., & Sikora, M. (2020). Gradient Boosting Application in Forecasting of Performance Indicators Values for Measuring the Efficiency of Promotions in FMCG Retail. *Silesian University of Technology*. Retrieved from https://arxiv.org/pdf/2006.04945.pdf

Korolev, M., & Ruegg, K. (2015). Gradient Boosted Trees to Predict Store Sales. Retrieved from https://cs229.stanford.edu/proj2015/193_report.pdf

Massaro, A., Panarese, A., Giannone, D., & Galiano, A. (2021). Augmented Data and XGBoost Improvement for Sales Forecasting in the Large-Scale Retail Sector. *Applied Sciences, 11(17)*, 7793. https://doi.org/10.3390/app11177793

Jain, A., Manghat Nitish, M., & Saurabh, C. (2015). Sales Forecasting for Retail Chains. Retrieved from https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/004.pdf

Jiménez, F., Sánchez, G., García, J., Sciavicco, G., & Miralles, L. (2017). Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing, 234*, 75–92. https://doi.org/10.1016/j.neucom.2016.12.045

Helmini, S., Jihan, N., Jayasinghe, M., & Perera, S. (2019, May 8). *Sales forecasting using multivariate long short term memory network models*. PeerJ Preprints. https://peerj.com/preprints/27712/

Random Forest Regressor. (2017b, February). [Graph]. Research Gate. https://www.researchgate.net/publication/313489088/figure/fig3/AS:864415041732616@1583104014685/Fig-A10-Random-Forest-Regressor-The-regressor-used-here-is-formed-of-100-trees-and-the.jpg

Schematical representation of gradient boosting regression in regards to algorithm iterations. (2021, January). [Graph]. Research Gate. https://www.researchgate.net/profile/Ivanna-Baturynska/publication/340524896/figure/fig3/AS:878319096569859@1586418999392/Schematical-representation-of-gradient-boosting-regression-in-regards-to-algorithm.png