Prediction of Social Media Consumer Behaviour Using Machine Learning

Supervised by Dr. Priyanka Shukla
M. Sajid, P. Kumar, P. Ray, S. Majhi, S. Das, V. Kumar
Department of Mathematics, Indian Institute of Technology Madras, 600036, India

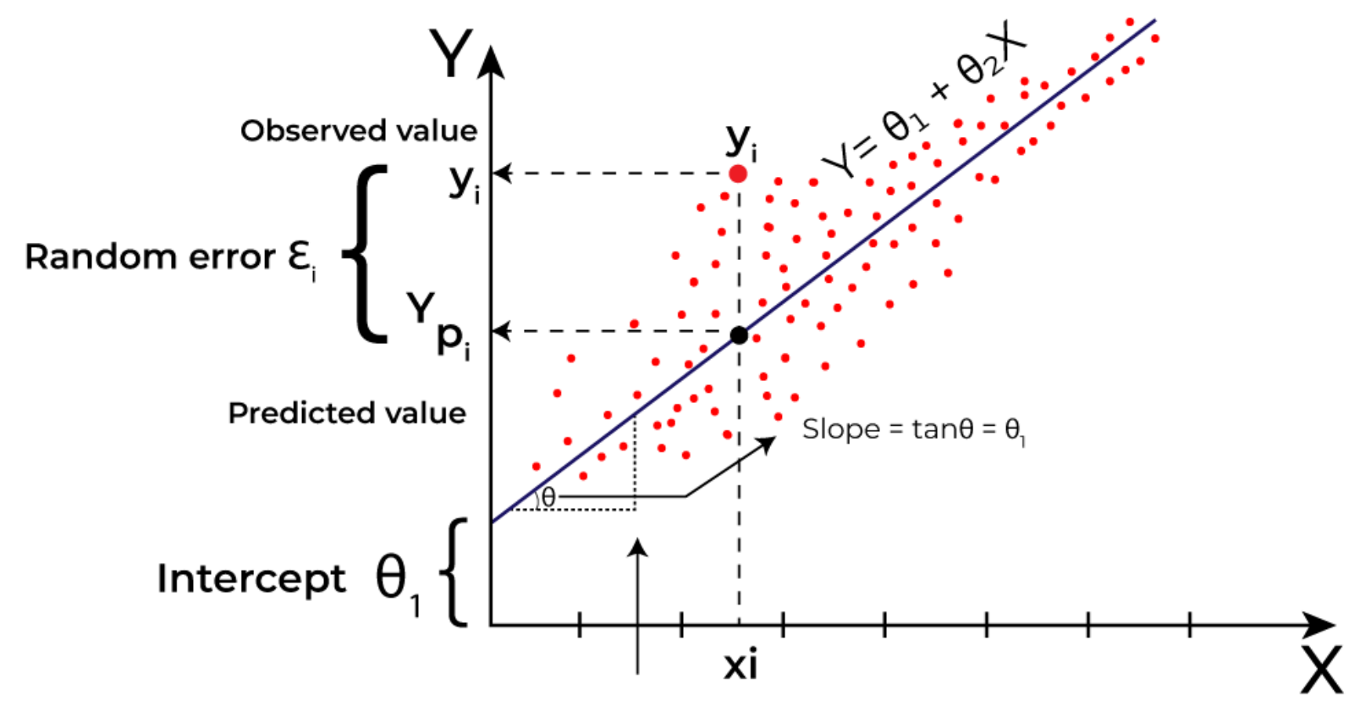
Motivation

Why do we need to predict social media behaviour?

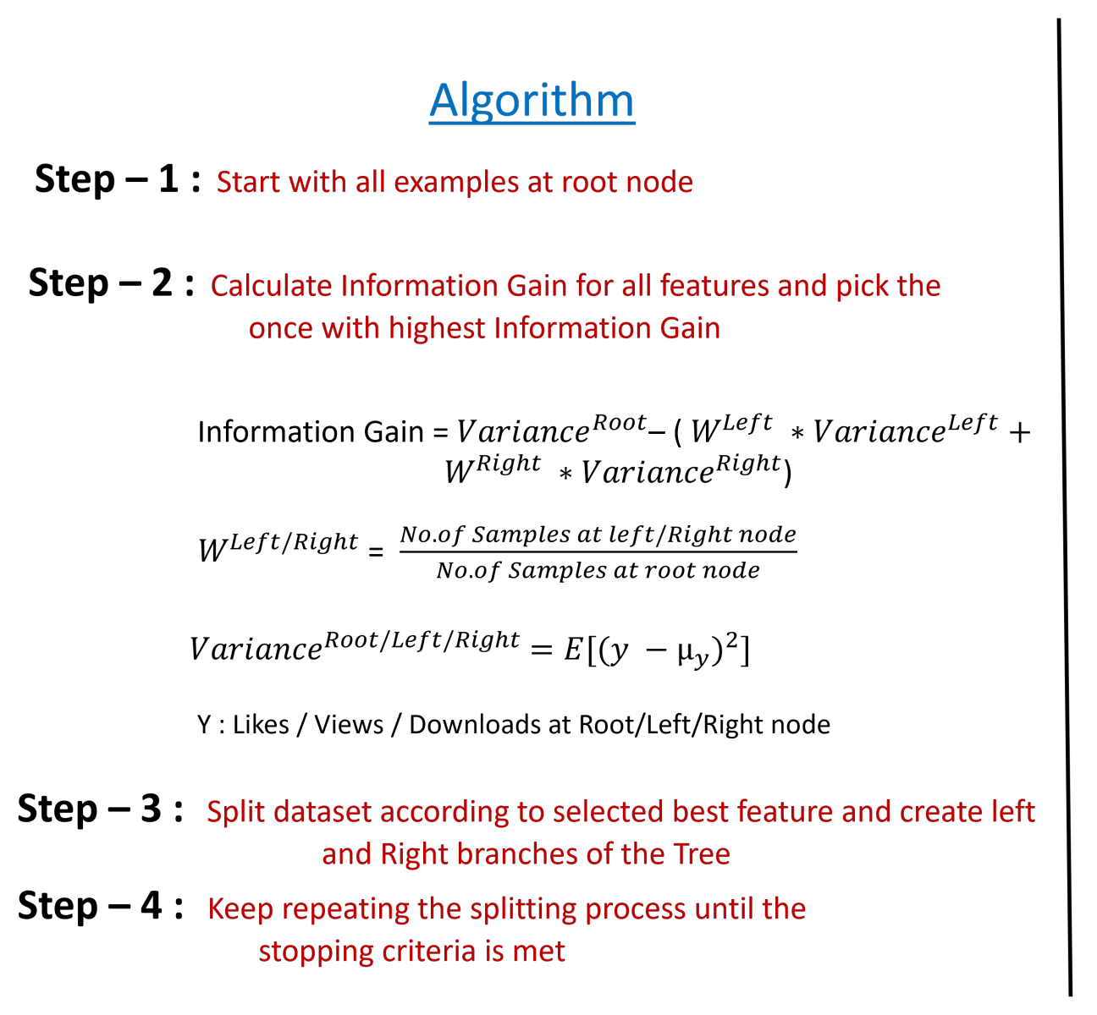
In present world , almost every person irrespective of their age have been engaging themselves in social media. So if we want to reach to a large number of population, social media must be one of the easiest way.By analyzing social media data using machine learning techniques, marketers gain deeper insights into consumer behavior, preferences, and sentiments. Machine learning models can predict individual user preferences based on their social media interactions. So we have proposed some machine learning based mathematical models. Those models will predict the consumer behaviour with different accuracy based on the machine learning approaches that we are using.

Methodology

- Data Preprocessing: We have removed outliers, errors from the data. After cleaning to do the predictions of consumer behaviour, we have used the following machine learning methods.
- Linear and Polynomial Regression: The formula for prediction using linear regression is $\hat{y} = w.x + b$, where $\hat{y}$ is predicted value of consumer behaviour. To find the parameter we will minimize the cost function or mean squared error $J(b,w) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - (b + w.x_i))^2$
- The prediction for behaviour by polynomial regression is $\hat{y}(x) = b + w_1x + w_2x^2 + \cdots + w_nx^n$, and we have calculated the parameters same as linear regression.


Y
X
Observed value
y_i
Random error ε_i
Y_{p_i}
Predicted value
Intercept θ_1
y_i
$Y = \theta_1 + \theta_2X$
Slope = tanθ = θ_1
θ
xi

- Decision Tree: It does not have any parameter. It has a tree like structure,that consists of a root node, branches, internal nodes and leaf nodes.We start with the root node. Then Calculate information gain for all possible features and pick the one with highest information gain. We finally calculate the error between actual and prediced value by mean squared error.



Algorithm

Step – 1 : Start with all examples at root node

Step – 2 : Calculate Information Gain for all features and pick the once with highest Information Gain

Information Gain = $Variance^{Root} - (W^{Left} * Variance^{Left} + W^{Right} * Variance^{Right})$

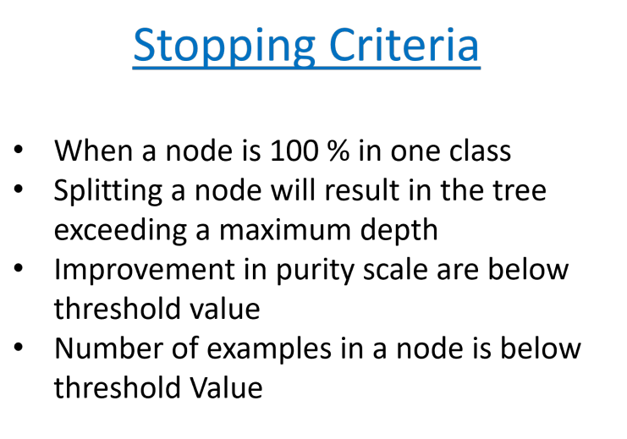
$W^{Left/Right} = \frac{No.of\ Samples\ at\ left/Right\ node}{No.of\ Samples\ at\ root\ node}$

$Variance^{Root/Left/Right} = E[(y - \mu_y)^2]$

Y : Likes / Views / Downloads at Root/Left/Right node

Step – 3 : Split dataset according to selected best feature and create left and Right branches of the Tree

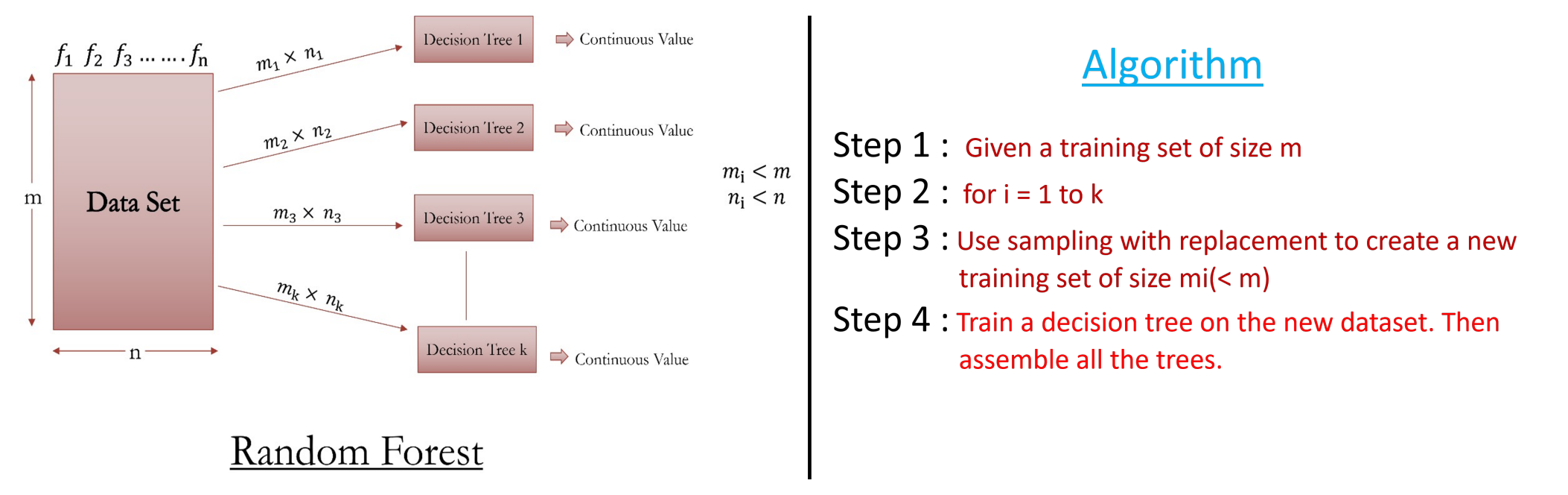
Step – 4 : Keep repeating the splitting process until the stopping criteria is met



Stopping Criteria

- When a node is 100 % in one class
- Splitting a node will result in the tree exceeding a maximum depth
- Improvement in purity scale are below threshold value
- Number of examples in a node is below threshold Value

- Random Forest: Why? We will deal with overfitting of decision tree & combine output of multiple decision trees and get a single result. We will take a subset m_i from total feature m, & train those new $m_i(< m)$ dataset. The average of the outputs of the newly built decision trees will be the final prediction.



Random Forest

Algorithm

Step 1 : Given a training set of size m

Step 2 : for i = 1 to k

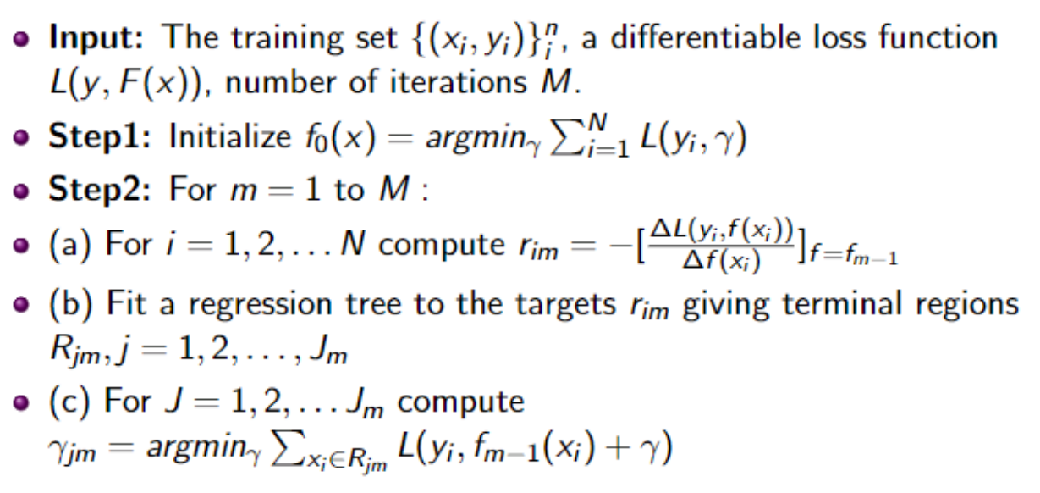
Step 3 : Use sampling with replacement to create a new training set of size mi(< m)

Step 4 : Train a decision tree on the new dataset. Then assemble all the trees.

- XGBoost: It is an optimised distributed gradient boosting library. In this method we calculate similarity gain & similarity score while determining root and leaf node. We define residual by difference between observed value & predicted value.

similarity score(SS) = $\frac{\text{(sum of residuals)}^2}{\text{number of residuals} + \lambda}$,
similarity gain = $(SS_l + SS_R) - SS_{root}$

Methodology(continue)



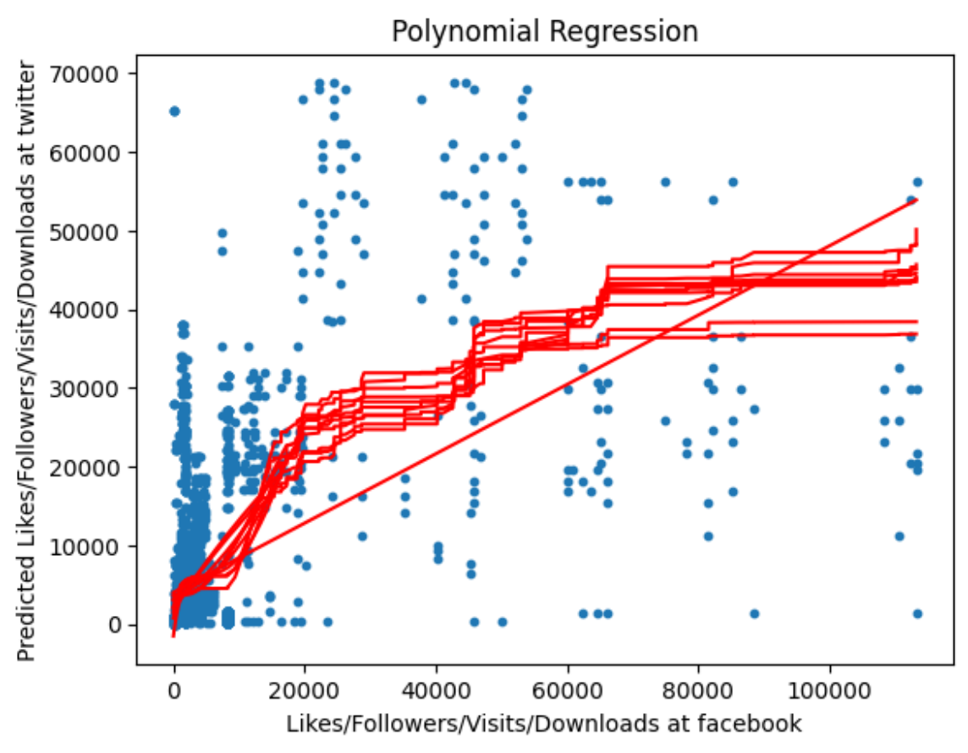
- **Input:** The training set $\{(x_i, y_i)\}_i^n$, a differentiable loss function $L(y, F(x))$, number of iterations M.
- **Step1:** Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
- **Step2:** For $m = 1$ to M :
- (a) For $i = 1, 2, \ldots N$ compute $r_{im} = -[\frac{\Delta L(y_i, f(x_i))}{\Delta f(x_i)}]_{f=f_{m-1}}$
- (b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1, 2, \ldots, J_m$
- (c) For $J = 1, 2, \ldots J_m$ compute $\gamma_{jm} = argmin_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$

- We use one hot encoding of the dataset features for last three algorithm.

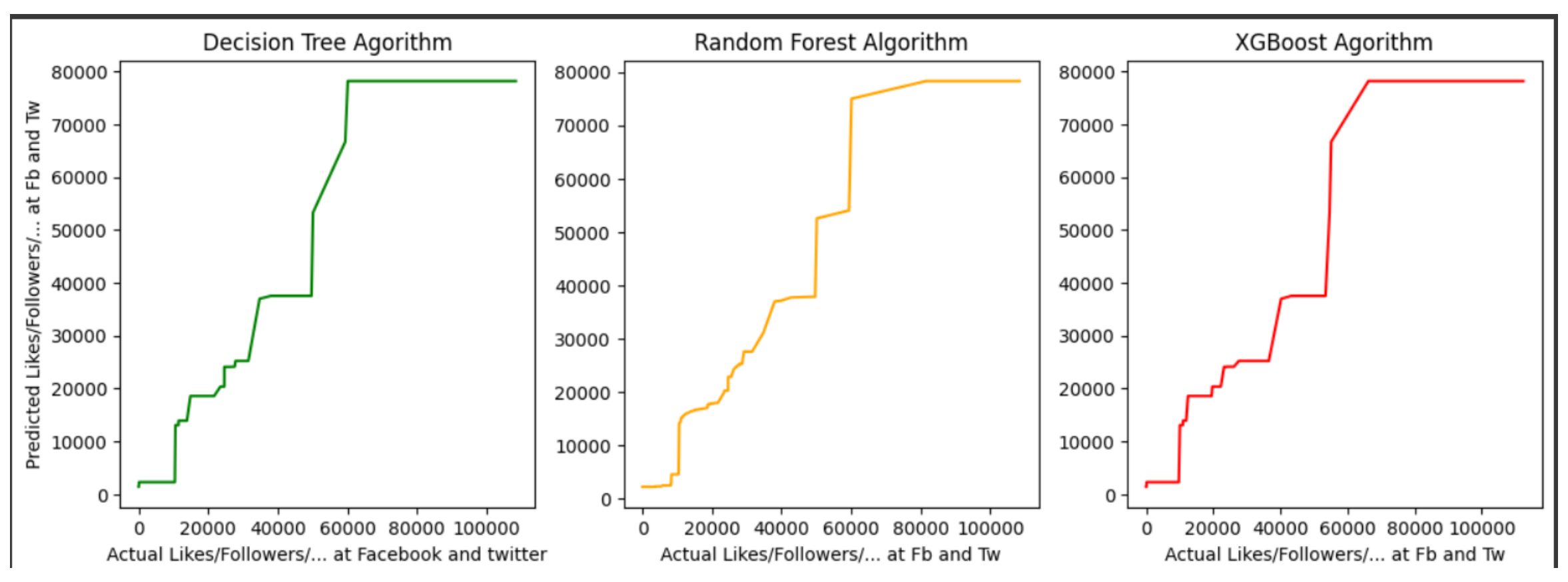
Results & Discussion

After applying above algorithms in Python we are getting the following results, where in all the graphs X axis represents the actual activity of customer in some platform, whereas Y axis represents the predictions for the same. For codes, click on `https://github.com/pritamraymng/Project_2nd_sem`

- Linear & polynomial Regression:



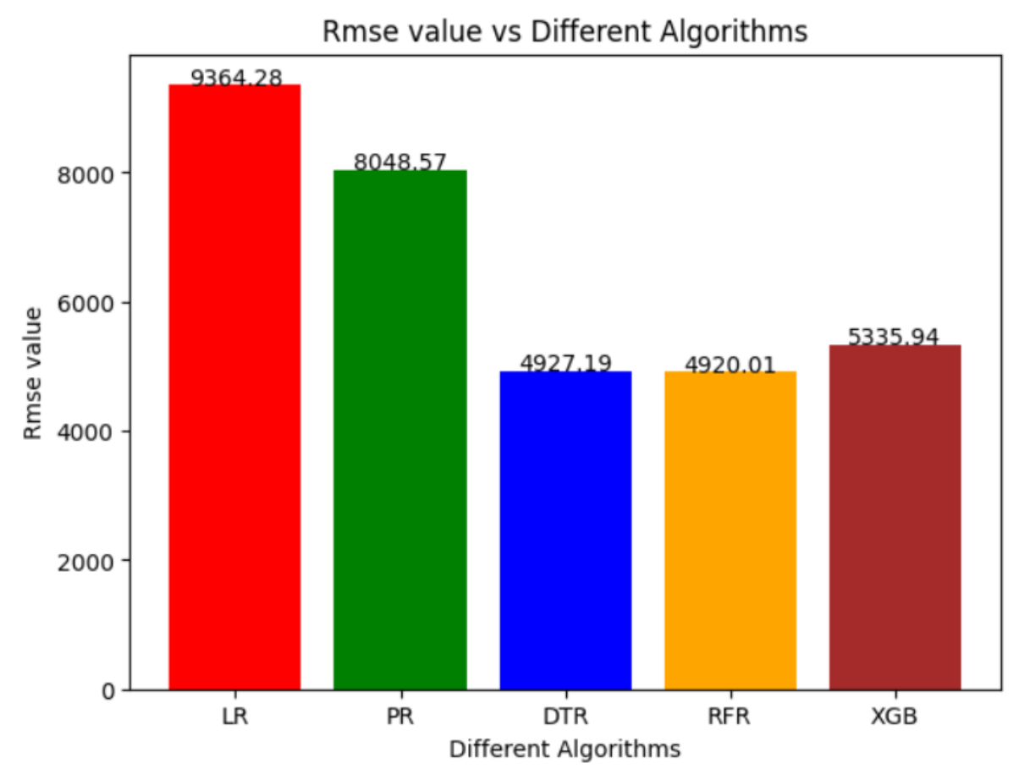
- We are getting the accuracy in consumer behaviour prediction for linear as 28%, for ten degree polynomial accuracy is 47%, & due to overfitting of curve in eleven degree polynomial accuracy is 44%.
- Decision Tree, Random Forest & XGBoost:

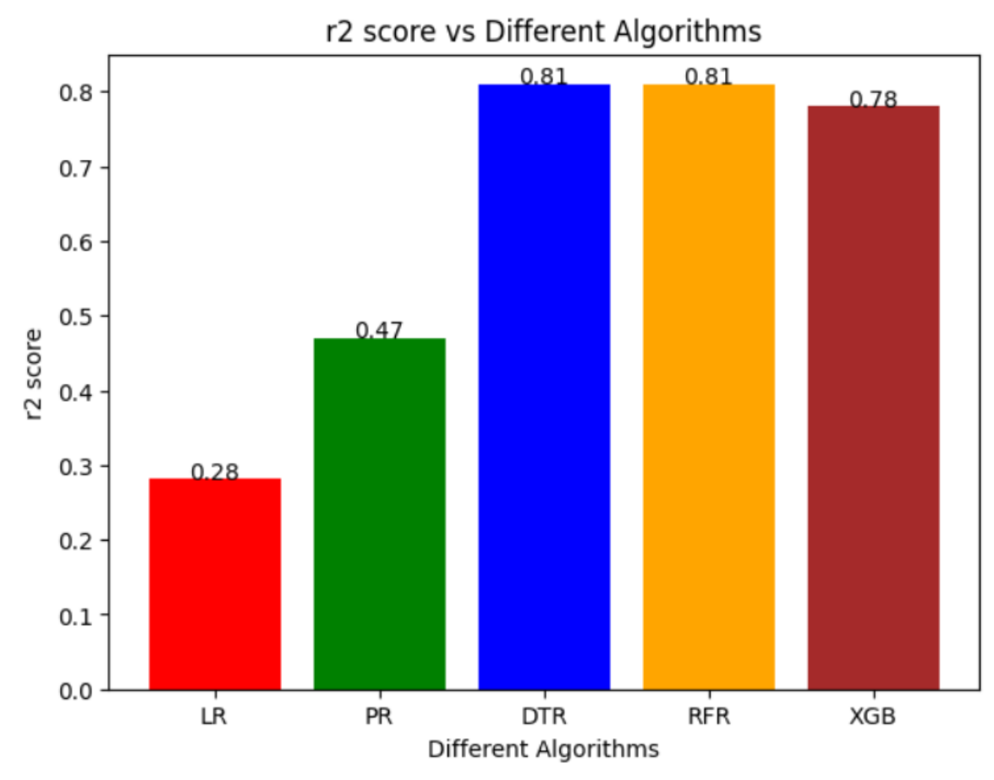


- We are getting the accuracy of prediction in decision tree as 80.89%, in random forest as 81%, and in XGBoost as 78%.

Conclusion

- We can see that random forest algorithm gives the best accuracy in predicting consumer activity as 81%. So, we canselect random forest algorithm as best algorithm. Here are some bar graphs.
- The first graph gives us the comparison between root means square & different algorithm.





- The second graph gives us the comparison between R2 score & different algorithm.

References

- Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics,Kiran Chaudhary, Mansaf Alam , Mabrook S. Al-Rakhami and Abdu Gumaei, Springer Open,2021.
- A machine learning-based approach to enhancing social media marketing,Senthil Arasu, B.JonathBackia Seelan , N. Thamaraiselvan , Department of Management Studies, National Institute of Technology, Tiruchirappalli, India.
- XGBoost: A Scalable Tree Boosting System, Tianqi Chen, University of Washington, tqchen@cs.washington.edu