Ad-Campaign Analysis and Sales prediction using K-means Clustering andRandom Forest Regressor

Ghosh Madhumita[1], Ravi Gor[2]
[1,2]Department of Mathematics, Gujarat University, India

Abstract
A campaign is a series of ads that is aimed to fulfill a purpose, such as generating leads or increasing the number of app installs. Machine learning techniques are used in ad campaigning. In machine learning, there are many types of algorithms, such as classification algorithm, regression algorithm, clustering algorithm, etc. Among the clustering algorithms, K-means algorithm is widely used because it is a simple and fast algorithm. Random forest algorithms are used for classification and regression both. In this paper clustering and regression both techniques are used for ad campaigning analysis and sales prediction. First,ad groups are created using the K-means clustering algorithm then Random Forest Regressor algorithm is used to optimize sales conversion and predict future sales.

Keywords: *Supervised learning, Clustering, Regression, K-means, Random Forest, advertising data*

Date of Submission: 07-03-2022 Date of Acceptance: 23-03-2022

I Introduction

To raise awareness about the brand among the people, to increase sales or to improve communication within a specific market, advertising campaign strategy is used. Several kinds of ways are used for advertising campaign but digital marketing is popular way. An online advertising agency is specifically focused on digital marketing because it is the best way to aware the people about the product. That's why businessman prefers online platforms for advertising campaigns. There are many models available to predict future sales for advertising campaigns. Nowadays the Machine Learning model can also be used to predict future sales.

In Machine Learning there are three kinds of learning methods supervised learning, unsupervised learning and reinforcement learning. Classification and Regression are part of supervised learning. Clustering and Association are part of unsupervised learning.

Clustering is used to classify the data points into a number of groups and this grouping is based on similar traits. Clustering algorithm does not use training data set. In clustering, there is a no labeled data. There are many different clustering algorithms such as, Hierarchical clustering, K-means clustering, Mean-Shift clustering etc. Among the clustering K-meansclustering algorithm is widely use because it is a simple and fast algorithm. In K-means clustering Elbow method is used to find the optimal number of clusters.

Regression is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Regression algorithms are used when there is a relationship between the input variable and the output variable. There are some popular Regression algorithms: Linear Regression, Regression Trees and Polynomial Regression etc. Random forest regressor produces better results when the large dataset is given. Random forest regressor is an ensemble of decision tree. Each tree makes its own individual prediction. Then it calculatesthe average of individual predictions.

II Literature Review

Balabantaray et al. (2013) analysed two clustering algorithms K-means and K-meloids using the refinement clustering method. Both algorithms are trained by a dataset of hundred documents for clustering. For this analysis they applied Euclidean and Manhattan methods to calculate distances in the K-mean algorithm. At the end, they concluded that the K-means algorithm gave better results than the K-medoids because the clusters formed by the K-means algorithm are more efficient than the K-medoids.

Alsmadi and Alkhami (2015) applied K-means clustering and support vector machine algorithms in a large dataset of email for the purpose of folder and subject classification. The model was trained by a large dataset of emails. In this paper they first created eight clusters based on the similarity of emails such as personal, job, business etc. and then classified the subject and folder.

Shreyas et al. (2017) predicted the popularity of articles using Random Forest regression model. They also predicted share counts of articles from the low share count and dense band region. Furthermore, they calculated R^2-score to check the accuracy of model. They compared Random Forest regression model with 13 other models and concluded that this model gives 88% accurate prediction.

Syakur et al. (2018) identified the best customer profile group by using the integration of K- means clustering and Elbow method. In this model they used Elbow method to determine the value of K and the distances between each data point and centroid is calculated by Sum of Squares method.

Gao et al. (2019) predicted the employee turnover in industries by using Weighted Quadratic Random Forest (WQRF) which is based on the traditional random forest algorithm combined with data characteristics. In employee turnover dataset, different employees had some similar characteristic, so they divided the model in two steps: First the random forest algorithm is used to order the feature according to their importance and reduce dimensions. Second, F- measure values are calculated for each decision tree and this value is assigned as a weight for each decision tree. They also compared this prediction with the result of random forest and logistic regression algorithms. Finally, they concluded WQRF yields significant result.

Yuan and Yang (2019) studied K-value selection for the K-means clustering algorithm, as well as they analyzed four K-value selection methods, namely Elbow Method, Gap Statistic, Silhouette Coefficient and Canopy. For this study they used an iris data set and verified the results obtained by four algorithms. Finally, concluded that the Elbow method is best for K- value selection.

III Methodology

The data for the advertising campaign has been collected from Kaggle. In this model impression, click and spent are taken as an independent variable to predict total conversion.

1. Impressions: the number of times the ad was shown.
2. Clicks: number of clicks on for that ad.
3. Spent: Amount paid by company xyz to Facebook, to show that ad.
4. Total conversion: Total number of people who enquired about the product after seeingthe ad.

Two different techniques Regression and Clustering are applied for ad campaign. K-means clustering and Random Forest regressor algorithms are used for Clustering and Regression. Here, both techniques (algorithms) are explained briefly.

Clustering

Clustering is a machine learning technique that divides unlabeled data points into groups. Out of many clustering algorithms, K-means clustering algorithm is selected.

K-means algorithm:

Step-1. Determine the value K; where K represents the number of clusters.

Step-2. Select K number of centroids (i.e. center of cluster) in such a way that they are as farther as possible from each other.

$$c_i = \frac{1}{M}\sum_{j=1}^{M} x_j$$

Step-3. Calculate the distance between each data point and the centroid by using Euclidean distance formula.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Step-4. Assign each data point to that cluster whose center is nearest to that data point.

$$a_{ij} = \{\begin{matrix} 1 & d = \min\{D(x_i, c_i) \\ 0 & otherwise \end{matrix}$$

Step-5. The centroid is calculated by taking mean of each data point in cluster.

$$J = \sum_{i=1}^{n}\sum_{l=1}^{k} a_{ic} D(x_i, c_1)^2$$

Step-6. Repeat step-3 to step-5 until any of the following stopping criteria is met-

- Centre of newly formed clusters do not change
- Maximum number of iterations are reached

These two methods can be used to find K in K-means:

1. The Elbow Method
2. The Silhouette Method

Here, Elbow method is applied to find the optimal number of clusters, which is used by Within Cluster Sum of Squares concept.

$$WCSS = \sum x_{i\,in\,Cluster1}\,distance(x_i C_1)^2 + \sum x_{i\,in\,Cluster2}\,distance(x_i C_2)^2 + \sum x_{i\,in\,Cluster3}\,distance(x_i C_3)^2$$

Where, $\sum x_{i\,in\,Cluster1}\,distance(x_i C_1)^2$ is sum of the square of the distances between each data point and centroid within one cluster and the same for the other two clusters. This distance is calculated by using Euclidean method.

Regression:
Regression is a supervised learning technique used when the output variable has a real or constant value, such as salary, weight etc.

Random Forest:
Random Forest (RF) algorithm can be used for classification and regression both. This algorithm is an ensemble of decision tree. The prediction of the random forest is based on the collection of the trees. In this paper Random Forest regressor is applied to predict the future sales.

Random Forest Regressor:
Step-1. Select the sample randomly from the training data set.
Step-2. Apply the decision tree algorithm individually on the collected sample.
Step-3. Calculate the average of the predictions made by the individual decision tress.

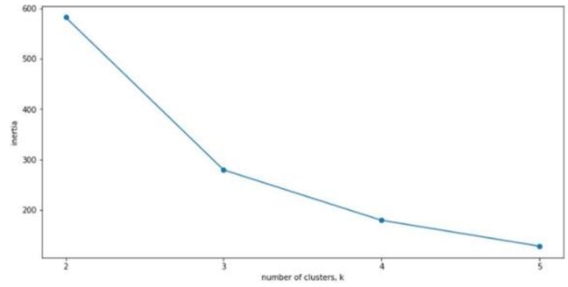
$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$$

Where, T = Decision trees in the Random Forest
$\hat{y}_t$ = Predictions made by each decision tree

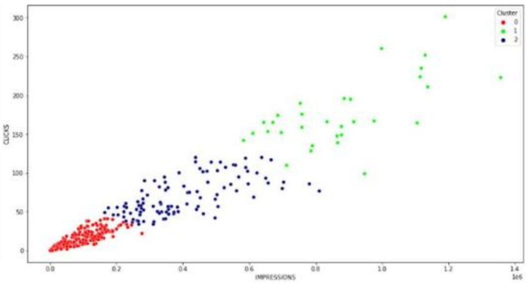
The model is built in two steps.
Step-1: The ad groups are created using the K-means clustering algorithm and these ads aregrouped, based on impressions, clicks and spent.
Step-2: Future sales are predicted using a random forest regressor algorithm.

IV Result and Discussion

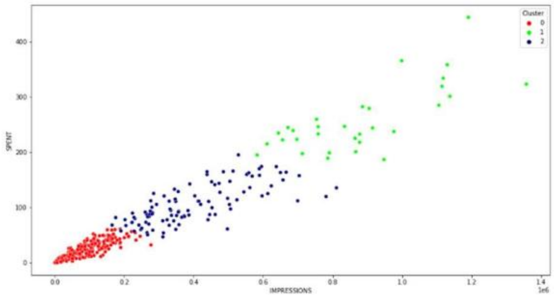
The value of K for K-means clustering is calculated using elbow method as shown in graph 1,where K=3 represent number of clusters.

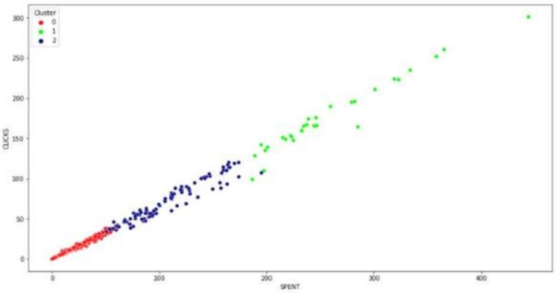


Graph 1: Elbow method to calculate K



raph 2: Relationship between impression and click





Graph 3: Relationship between impression and spent

Graph 4: Relationship between spent and click

The relationship between impression and click is represented by graph 2. There are three clusters which are shown in red, blue and green colours. Red cluster interprets that when the ad is shown for a shorter period of time, the ad is clicked less often. In similar way, blue and green clusters represent the ad shown for a moderate and longer period of time respectively.

The relationship between impression and spent is represented by graph 3. There are three clusters which are shown in red, blue and green colours. Red cluster interprets that when company paid less amount to show the ad, the ad is shown for a shorter period of time. In similar way, blue and green clusters represent the high and medium amount paid by company respectively.

The relationship between spent and click is represented by graph 4. There are three clusters which are shown in red, blue and green colours. Red cluster interpreted that when company paid fewer amount to show the ad, the ad is clicked less often. In similar way, blue and green clusters represent the high and medium amount paid by company respectively.

From these three graphs it can be analyzed that the number of clicks is more when thenumber of impressions and spent is high and the number of clicks is less when the number of impressions and spent is less. From this it can be concluded that impression, spent and click are interdependent. Therefore, these three parameters i.e. impression, click and spent are taken as independent variables for predicting future sales.

This model predicts the number of people who enquire about the product after seeing the ad. 80% of data is used for training and 20% of data is used for testing purpose.

The error between actual future sale and predicted future sale is calculated by Mean Absolute Error (MAE) method. The Mean Absolute Error is 0.991 and the R-squared value is 0.753 which means the accuracy of model is 75.3%.

Table:1 Mean Absolute Error and R-squared

Random Forest Regressor	
Mean Absolute Error	0.991
R-squared value	0.753

V Conclusion

The future sale is predicted by using K-means clustering and Random Forest regressor algorithms. The future sale predicted by this model is much closer to the actual sale for the chosen dataset. Hence, the integration of two algorithms K-means clustering and Random Forest regressor gives permissive result.
In future, this type of problems can be solved with other supervised learning techniques.

References

[1]. R. Balabantaray, C. Sarma and M. Jha, "Document Clustering using K-means and K-Medoids", *International Journal of Knowledge Based Computer System*, 1(1), pp. 7-13, 2013.

[2]. I. Alsmadi and I. Alhami, "Clustering and classification of email contents", *Journal of King Saud University – Computer and Information Sciences*, 27(1), pp. 46-57, 2015.

[3]. R. Shreyas, D.M Akshata, B.s Mahanand, B. Shagun and C.M Abhishek, "Predicting Popularity of Online Articles using Random Forest Regression", *Institute of Electrical and Electronics Engineers*, 8(13), pp. 1-5, 2017.

[4]. M. Syakur, B. Khotimah, E. Rochman and B. Satoto, "Integration K-means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster", *Institute of Physics : MaterialScience and Engineering*, 336(1), pp. 1-6, 2018.

[5]. X. Gao, J. Wen and C. Zhang, "An Improved Random Forest Algorithm for Predicting EmployeeTurnover", *Hindawi : Mathematical Problems in Engineering*, 2019(4), pp. 1-12, 2019.

[6]. C. Yuan and H. Yang, "Research on K- Value Selection Method of K-means Clustering Algorithm", *Multidisciplinary Scientific Journal*, 2(16) pp. 227-235, 2019.

[7]. https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-K-means-708505d204eb

[8]. https://www.cs.cmu.edu/~aarti/Class/10701_Spring14/assignments/hw3_solutions.pdf
[9]. https://www.gatevidyalay.com/tag/K-means-clustering-numerical-example-pdf/

Ghosh Madhumita. "Ad-Campaign Analysis and Sales prediction using K-means Clustering and Random Forest Regressor." *IOSR Journal of Mathematics (IOSR-JM)*, 18(2), (2022): pp. 05-09.