Social Network Advertisements Prediction Using Unsupervised Machine Learning

N. Bala Sundhar Ganapathy[1], S. B. Subash[2], P. Shuriya[3*], S. Srijith Krishnan[4], I. Salman Faizul[5]

[1]*Associate Professor, Department of Information Technology, Panimalar Engineering College, Chennai, India*

[2,3,4,5]*Student, Department of Information Technology, Panimalar Engineering College, Chennai, India*

*Corresponding author: shuriyapalanikumar@gmail.com

***Abstract*: Social media is forming an increasingly central part of how companies communicate their marketing strategies to their customers and it provides an empirical analysis of the impact social media communication has on brand equity and purchase intention. The aim is to investigate given social advertisement datasets by unsupervised machine learning methods like K-Means and to predict whether a user purchased a particular product or not on social media network advertisement. The analysis of the given dataset to capture information like, variable identification, missing value treatments, and analyze the data validation, data cleaning/preparing, and data visualization will be done. It describes a comprehensive guide to unsupervised machine learning method of parameters about find accuracy calculations and the result shows that GUI application of user behaviours.**

***Keywords*: Dataset, K-Means, Python.**

1. Introduction

We live in an era where everyone is interested to start business by their own. In order to give success to the products they develop there is a need for the method to reach out maximum number of customers. In order to do that Marketing plays a huge role. It can be done through various ways but the most successful way is delivering advertisements through any Social media platforms because most of the people in this generation use social media platforms frequently and also in an effective manner and purchasing a product by seeing an advertisement through social media become as casual. So it is the best advised method to obtain success.

Social Media platforms obtains enormous users of different age groups and different kinds of people who purchases a product frequently and some who does not purchase at all so sending advertisements to all kinds of users in social media is useless and it may raise to financial issues in a company because to deliver an advertisements to large amount of users it costs high this is because of the older algorithms we follow that does not send advertisements according to users interest so to overcome this problem we use Machine Learning algorithms to reduce the count of delivering advertisements to users. This overcomes financial issues too. The main objective of using Machine Learning algorithms is instead of sending advertisements to the users who have low probability of purchasing we can send advertisements to the users who have higher probability of purchasing a product.

Machine learning is to predict the long run from past knowledge. Machine learning (ML) may be a sort of computing (AI) that has computers with the flexibility to be told while not being expressly programmed. It can be a method of computing (AI) that provides computers with the ability to be told while not being expressly programmed. The tactic of coaching job and prediction involves use of specialized algorithms. It feed the coaching job data to associate rule, and additionally the rule uses this coaching data to supply predictions on a latest check data. It is often roughly separated into 3 classes.

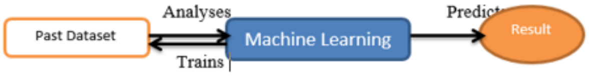


Fig. 1. Process of machine learning

They measure supervised learning, unattended learning and reinforcement learning. Supervised learning program is each given the input data and additionally the corresponding labelling to tell data must be tagged by somebody's being beforehand. Unsupervised learning isn't any labels. It provided to the educational rule. This rule must decipher the clump of the input data. Finally, Reinforcement learning dynamically interacts with its atmosphere and it receives positive or feedback to spice up its performance. At a high level, these completely different algorithms are often classified into 2 teams supported the approach they learn" regarding knowledge to create predictions: supervised and unattended learning. Classification is that the method of predicting the category of given knowledge points. Categories square measure generally known as targets/labels or classes. Classification prophetical modelling is that the task of approximating a mapping operate from input variables(X) to distinct output variables(y). In machine learning and statistics, classification can be a supervised learning approach during which the laptop program learns from the data input given to it then uses this learning to classify new observation. This data set might just be bi-class (like distinguishing whether or not or not the person is male or female or that the mail is spam or non-spam) or it's about to be multi-class too. Some samples of classification problems are: speech recognition, handwriting recognition bio metric identification, document classification etc. Supervised learning needs that the info wont to train the rule is already labelled with correct answers. Supervised learning issues are often any classified into classification issues. The distinction between the 2 tasks is that the undeniable fact that the dependent attribute is numerical for categorical for classification. A classification model tries to draw some conclusion from discovered values. Given one or a lot of inputs a classification model can try and predict the worth of 1 or lot of outcomes. A classification downside is once the output variable may be a class, like “red” or “blue”.

Unsupervised learning may be a machine learning technique during which there's no direction needed instead the model is allowed to figure on its own to find data. It primarily deals with the untagged knowledge. It permits you to perform lot of advanced process tasks compared to supervised learning. Unattended learning is often a lot of unpredictable compared with different natural learning ways. The most effective time to use unattended machine learning is once we don’t have knowledge on desired outcomes, like decisive a target marketplace for a completely new product that your business has ne'er sold before.

2. Data Preparation Process

Validation techniques in machine learning area unit want to get the error rate of the Machine Learning model, which may be thought-about as on the point of actuality error rate of the dataset. If the information volume is massive enough to be representative of the population, you will not would like the validation techniques. However, in real-world situations, to figure with samples of knowledge that will not be a real representative of the population of given dataset. To finding the missing worth, duplicate worth and outline of knowledge sort whether or not it's float variable or number. The sample of knowledge want to offer associate degree unbiased analysis of a model work on the coaching dataset whereas calibration model hyper parameters.

A. Data Validation/ Cleaning/Preparing Process

It is first step of the process to check the datasets whether it has any mistake or missing data because the raw data is gathered from different types of source which might cause wrong or incorrect results. A pre-processing helps to find the mistake present in the data that can be missing values and repeated values. This types of data can mislead the process so it has to be omitted from the datasets. The missing can be taken as null to drop that particular individual data from the dataset. This action is also taken for the repeated values which present in the multiple times in the dataset. And align the dataset into new format for the analysis.

B. To split the dataset for training purpose

Data mental image is a vital ability in applied statistics and machine learning. Statistics will so target quantitative descriptions and estimations of knowledge. Information mental image provides a vital suite of tools for gaining a qualitative understanding. This will be useful once exploring and reaching to apprehend a dataset and may facilitate with characteristic patterns, corrupt information, outliers, and far a lot of. With a bit domain data, information visualizations is accustomed specific and demonstrate key relationships in plots and charts that square measure a lot of visceral and stakeholders than measures of association or significance. Information mental image and beta information analysis square measure whole fields themselves and it'll suggest a deeper dive into some the books mentioned at the top.

3. Comparison of Supervised Machine Learning Algorithm (SMLT)

A. Logistic Regression

It is a method for associate degree analyzing a knowledge set within which there are one or a lot of freelance variables that verify an outcome. The end result is measured with a divided variable. The goal of logistical regression is to seek out the most effective fitting model to explain the link between the divided characteristic of interest and a collection of freelance variables. Logistical could be a Machine Learning classification formula that's wont to predict the likelihood of a categorical variable quantity. In logistical regression, the variable quantity could be a binary variable that contains information coded as one or zero.

Confusion Matrix result of Logistic Regression is:

[[166 0]. [1 0]]

B. Decision Tree

It works for each continuous moreover as categorical output variables. At the start, we tend to contemplate the entire coaching set because the root. Attributes are assumed to be categorical for info gain, attributes are assumed to be continuous. On the premise of attribute values records are distributed recursively. We tend to use applied math strategies for ordering attributes as root or internal node.

This method is sustained on the coaching set till meeting a termination condition. It's made in a very top-down algorithmic divide-and-conquer manner. All the attributes ought to be categorical. Otherwise, they ought to be discretized before. Attributes within the high of the tree have additional impact towards within the classification and that they are known mistreatment the data gain thought. A call tree are often simply over-fitted generating too several branches and will replicate anomalies thanks to noise or outliers.

Confusion Matrix result of Decision Tree is:

[[166 0] [0 1]]

C. Random Forest

Random call forests correct for call trees’ habit of over fitting to their coaching set. Random forest could be a sort of supervised machine learning formula supported ensemble learning could be a sort of learning wherever you be a part of differing types of formulas or a similar algorithm multiple times to create an additional powerful prediction model. The random forest formula combines multiple algorithms of a similar sort i.e. multiple call trees, leading to a forest of trees, thus the name "Random Forest". The random forest formula is used for each regression and classification tasks.

In the case of a regression downside, for a replacement record, every tree within the forest predicts a worth for Y (output). The ultimate price is calculated by taking the common of all the values foretold by all the trees within the forest. Or, just in case of a classification downside, every tree within the forest predicts the class to that the new record belongs. Finally, the new record is allotted to the class that wins the bulk vote.

Confusion Matrix result of Random Forest is:

[[166 0] [1 0]]

D. Support Vector Machines

A classifier that categorizes knowledge set by setting associate optimum hyper plane between data. I selected this classifier because it is unbelievably versatile within the variety of various kernelling functions will be applied and this model can yield a high foregone conclusion rate. Support Vector Machines square measure maybe one in every of the foremost in style and talked concerning machine learning algorithms. They were very in style round the time they were developed within the Nineteen Nineties and still be the go-to methodology for a high-performing algorithmic rule with a bit standardization.

Confusion Matrix result of Support Vector Classifier is

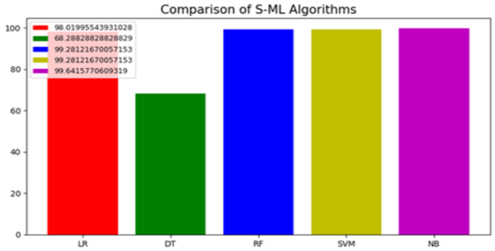
[[166 0][1 0]]

E. Naive Bayes algorithm

Naive Thomas Bayes could be an applied math classification technique based mostly it's one in every of the best supervised Learning algorithms. Naive Thomas Bayes classifier is Accurate and reliable algorithmic rule. Naive Thomas Bayes classifiers have high accuracy and speed on massive datasets. Naïve Bayes classifier assumes that the result of a specific feature in an exceedingly category is freelance of different options. For example, a loan human is fascinating or not looking on His/her financial gain, previous loan and dealings history, age and location. Although these options square measure mutually beneficial, these options square measure still thought of severally. This assumption simplifies computation, and that is why it's considered as naive. This assumption is termed category conditional independence.

Confusion Matrix result of Naive Bayes is:

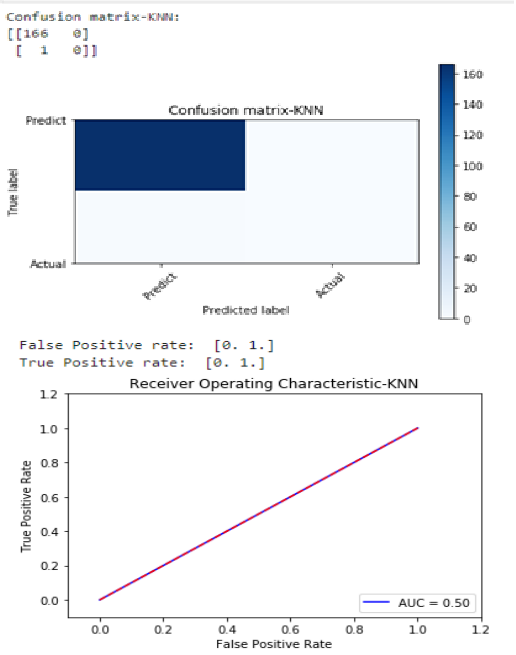
[[166 0] [1 0]]



F. K-Nearest Neighbour (KNN)

K-Nearest Neighbour may be a supervised machine learning rule that stores all instances correspond to coaching information points in n-dimensional house. Once associate degree unknown separate information is received, it analyses the nearest k variety of instances saved and returns the foremost common category because the prediction and for real-valued information it returns the mean of k nearest neighbours. Within the distance-weighted nearest neighbour rule, it weights the contribution of every of the k neighbours consistent with their distance mistreatment the subsequent question giving larger weight to the nearest neighbours.

Usually KNN is powerful to shire information since it's averaging the k-nearest neighbours. The k-nearest-neighbours rule may be a classification rule, and it's supervised: it takes a bunch of labelled points and uses them to be told a way to label alternative points. To label a replacement purpose, it's at the labelled purposes nearest thereto new point (those area unit its nearest neighbours), and has those neighbours vote, thus whichever label the foremost of the neighbours have is that the label for the new purpose Makes predictions regarding the validation set mistreatment the whole coaching set. KNN makes a prediction a few new instance by looking through the whole set to search out the k "closest" instances. "Closeness" is decided employing a proximity measure (Euclidean) across all options.



```
Confusion matrix-KNN:
[[166   0]
 [  1   0]]
```

```
False Positive rate:  [0. 1.]
True Positive rate:  [0. 1.]
```

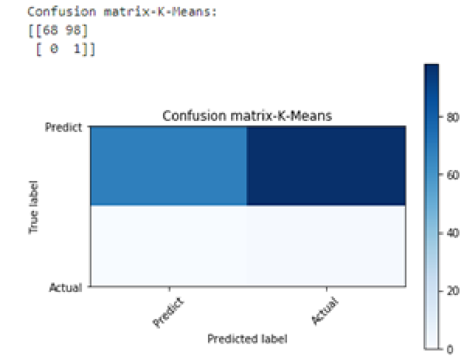
4. Performance of Unsupervised Machine Learning Algorithm (USMLT)

A. K-means clustering algorithm

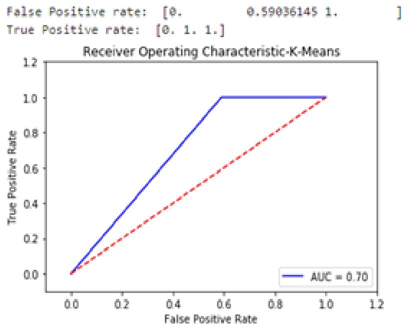
K-means is one in every of the only unsupervised learning algorithms that solve the well-known cluster drawback. The procedure follows an easy and straightforward thanks to classify a given information set through an exact variety of clusters (assume k clusters) fastened apriori. the most plan is to outline k centres, one for every cluster. These centres ought to be placed in a very crafty means thanks to completely different location causes different result. So, the higher alternative is to put them the maximum amount as potential isolated from one another. Succeeding step is to require every purpose happiness to a given information set and associate it to the closest centre. Once no purpose is unfinished, the primary step is completed associated an early cluster age is finished. At this time, we'd like to re-calculate k new centroids as barycentre of the clusters ensuing from the previous step. when we've got these k new centroids, a brand new binding needs to be done between identical information set points and therefore the nearest new centre. A loop has been generated. As a results of this loop we tend to could notice that the k centres amendment their location step by step till now a lot of changes area unit done or in different words centres don't move any further.

B. *K-Means Clustering Algorithm steps*

1. Indiscriminately choose ‘c’ cluster centres.
2. Calculate the space between every datum and cluster centres.
3. Assign the info purpose to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.
4. Calculate the new cluster centre.
5. Calculate the space between every datum and new obtained cluster centres.
6. If no datum was reassigned then stop, otherwise repeat from step three.



```
Confusion matrix-K-Means:
[[68 98]
 [ 0  1]]
```



```
False Positive rate:  [0.         0.59036145 1.        ]
True Positive rate:  [0. 1. 1.]
```

5. Proposed System

A. *Exploratory data analysis of social network ads*

The traditional and existing system generate advertisement to any type of user without any analysis of the user data or applying any specific algorithm to sort the ads to reach a specific type of user rather than reaching an irrelevant user. To overcome these problem and increase cost effectiveness in providing ads this system is proposed

In this multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

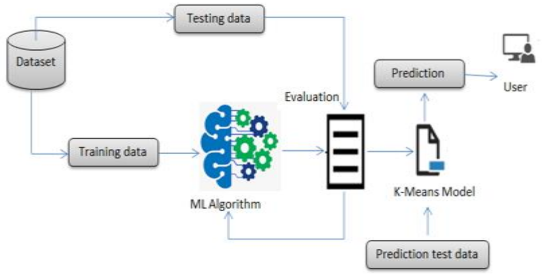
Training the system in the basic and much important phase of the machine learning process. To the train the system proper training data should without any anomaly. The data cleaning should be done in order that all the redundant data is removed.

Data wrangling is the section where report will be load in form of the data, Checks for cleanliness, and then trimming and cleaning of the given dataset for analysis is carried out. Made sure that the document steps carefully and justify for cleaning decisions.

The data set collected for predicting given data is split into Training set and Test set. Generally, 70:30 ratios are applied to split the Training set and Test set. These training and test data are passed into various machine learning algorithms by comparing the results and accuracy each algorithm the best one suitable for the training the system is system. After training the system KNN algorithm which is any unsupervised machine learning algorithm that allows the system to perform accordingly to dynamic nature and real-time data, this algorithm predicts the result for the given input dataset.

KNN or K Nearest Neighbor algorithm that simply processes the stores all the available cases and classifies based on the similarity and mapping cardinalities. KNN works by the means of finding the distance between given query and available sample data, by selecting the k number of near nodes and then votes for the highest frequency cases thus predicting the tendency of the user the user with low tendency to buy via social media is predicted and removed from the ads provider. Thereby increasing the efficiency of the ads and decreasing cost of expense in providing advertisements.

B. *Architecture*



Our technique could find a solution in providing the advertisements beneficially in cost wise and provide needful ads for the user by the means of better training algorithm and unsupervised learning algorithm which filters out the odd data and provides improved results.

6. Conclusion

Social ads square measure one among the fastest and only ways in which to attach with our target market. These ads bring profitable opportunities in digital promoting formats. These tiny ads build the foremost of all the data users share on social media to supply them content as extremely personalized as potential, and, as such, conversion opportunities. a lot of and a lot of brands have gotten on board with social ads on social media. The probabilities and types of ways in which to succeed in dead set our audience appear endless, with on the face of it nice samples of advertising on social media and personalization. With this technique Social Ads square measure associate degree unbelievably profitable and versatile advertising channel that provides United States of America the flexibility to create specific campaigns on social media to fulfil a range of various business goals at comparatively low prices. The prevailing system analyses the influence of client call and it doesn't concentrate on client purchase intention. However, the planned system collects multiple datasets from completely different sources that may be combined to make a generalized dataset, and so completely different machine learning algorithms would be applied to extract patterns and to get prediction results with most accuracy. Therefore, the system provides economical results that helps the digital promoting to be profitable and more practical.

References

[1] Xiaokang Zhou, "Multi-Modality Behavioral Influence Analysis for Personalized Recommendations in Health Social Media Environment."

[2] Mia Angeline, Stephany Chandra" Digitalize Your Brand: Case Study on How Brands Utilize Social Media Platforms to Achieve Branding and Marketing Goals."

[3] Aryo Bismo, "Application of Digital Marketing (social media and email marketing) and its Impact on Customer Engagement in Purchase Intention: a case study at PT. Soltius Indonesia."

[4] Mouzhi Ge, "Factoring Personalization in Social Media Recommendations."

[5] Xiao Liang, "Enhancing Content Marketing Article Detection with Graph Analysis."

[6] Chen Wang, Yutong, "Soc. Inf: Membership Inference Attacks on Social Media Health Data with Machine Learning."

[7] Xiaoyu Sean Lu, "Clustering-Algorithm-Based Rare-Event Evolution Analysis via Social Media Data.”

[8] Guillermo Mondragon, "Evaluation model of the digital experience in the retail sector using customer journey."

[9] Daniel (Yue) Zhang, "On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications".

[10] Jana Prodanova, "How Beneficial Is Social Media for Business Process Management? A Systematic Literature Review."

[11] Mengling Qiao, "Fine-grained Subjective Partitioning of Urban Space Using Human Interactions from Social Media Data."

[12] Basit Shahzad, "Quantification of Productivity of the Brands on Social Media with Respect to Their Responsiveness."

[13] Yuxia Xue, "Predicting Platform Preference of Online Contents Across Social Media Networks."

[14] Dingqi Yang, "Privacy-Preserving Social Media Data Publishing for Personalized Ranking Recommendation."