**Personality traits as feature to improve machine learning accuracy in marketing strategies**

Raimundo Otávio Nogueira Dias; Guilherme André Peleglini Rocha

**Personality traits as feature to improve machine learning accuracy in marketing strategies**

Summary

This work investigates the use of the Big Five personality traits with machine learning analysis to improve advertising strategies. Traditional advertising methods can be inefficient in reaching potential customers that will benefit and purchase the product or service. Neuromarketing emerged as a new area that focuses on measuring brainwaves to understand human behavior, but the disadvantage of using medical devices for this purpose is their high cost and hard access. Therefore, alternative methods such as machine learning have been created to reduce costs of irrelevant advertising. The ADS-16 dataset was used, containing 300 advertisements from 20 categories and three types (text, image, and rich media), ranked by 120 individuals on a scale of 1 to 5, and including personal information and Big Five Personality scores of the participants. Unsupervised machine learning was used to cluster individuals based on their personality scores, and supervised machine learning (Random Forest, Gradient Boosting, Neural Networks) was applied to predict consumer behavior based on feedback to advertisements. The study reveals that personality traits analysis can significantly influence marketing practices and this research contributes to the understanding of modeling consumer behavior using data analysis techniques with machine learning.

**Key-words:** K\_Mean Cluster; Random Forest; Gradient Boosting; Neural Networks; OCEAN Score

1. Introduction

According to John Wanamaker “Half of the money I spend on advertising is wasted. The problem is, I do not know which half” (Schreiner *et al*., 2019). This statement shows the biggest problem in advertisement that is to efficient allocate resources to reach the effective potential customers that will benefit and purchase the product or service.

How to find the wasting costs and have the most efficient advertisement strategy to the company? This question is not trivial. A lot of techniques have been developed to understand consumer behavior. The concept of “Rational Human” that means consumers are taking rational decisions during consumption has been contested from researches showing that humans tend to make rather emotional decisions according to subconscious reactions. (Murphy *et al*., 2008).

From the market necessity to improve marketing strategies and to avoid wastes to consumers on pointless advertisement that they do not want to see, a new area emerged called “Neuromarketing”. That scientific area is a combination of science of neurology and marketing and focuses on the emotional side of the brain by measuring brainwaves to understand human behavior. It allows the research of effective parameters that raise attention and positive emotions on people’s mind (Gómez *et al*., 2016).

Neuromarketing is possible with use of medical devices that measure which parts of the brain are affected for a particular buying behavior such as Positron Emission Tomography (PET), ElectroEncephaloGraphy (EEG), Eye-Tracking, Transcranial Magnetic Stimulation (TMS), and several other techniques (Racine *et al*., 2005).

The disadvantage of using those techniques is the high cost and hard access to the gauges to conduct the studies. With this knowledge advertisement models have been created for recommendation algorithms that target to show the right customer the most relevant products in which the buying experience can be facilitated and reduce costs of irrelevant advertising. Those techniques are the following but not exclusive Logistic Regression (LR), Support Vector Regression with radial basis function (SVR-rbf) and L2-regularized L2-loss Support Vector Regression (L2-SVR) (Fan *et al*., 2008).

Other alternatives to study the human brain behavior is the use of Artificial Neural Networks (ANN) which cheaper hardware, faster calculation, adaptability, availability and with possibility of deeper analysis. The ANN method can produce consistent and reliable methods when datasets with relevant inputs are available (Renvoise *et al*., 2007).

The Big Five factors are defined by five measurable personality traits as Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism also known by acronym OCEAN (B. Rammstedt *et al*., 2007). The Big Five can represent unique psychological characteristics of the individual as for example impulsiveness, intellectual curiosity, orientation to novelty and variety, confidence, sensitivity, extraversion, creativity and others which affect personal motivations and behavior tendencies.

The personality traits analysis has the potential to contribute a lot for the marketing practices since they are good predictors of impulsive behavior and reaction to the marketing experience (Turkyilmaz *et al*., 2015). A previous study conducted by Roffo and Vinciarelli (2016) using the ADS-Dataset of real advertisements and participants rating each of them, show efficacy on inclusion of personality traits as parameter for advertisement recommendation.

The objective of the study is to use a dataset with relevant input of ad categories and types; users personal information containing OCEAN personality score and the interaction between them with feedback from each user to each ad as output of ad performance. We used techniques of unsupervised machine learning to study the preferences of OCEAN personality clusters and used supervised machine learning methods as Random Forest [RF], Gradient Boosting [GB] and Neural Networks [NN] with the goal to find the best model to predict user feedback of advertising.

The research contributes to the knowledge of modeling consumer behavior without use of expensive medical devices for neuromarketing but instead use of data analysis techniques with help of machine learning to take conclusions of relevant dataset.

1. Material and Methods

Using the ADS-16 dataset from Roffo and Vinciarelli (2016) of 300 advertisements from 3 types: text, image and rich media as seen in Figure 1. They are divided in 20 categories and each advertisement is ranked by 120 non-acquainted individuals from 1(negative) to 5 (positive). The dataset also contains personal data of the subjects as their preference for websites, movies, music, tv programs, books, hobbies, demographic information as age, nationality, gender, home town, type of job, weekly working hours and financial situation. Finally, we have available the measured Big Five Personality score of each individual.

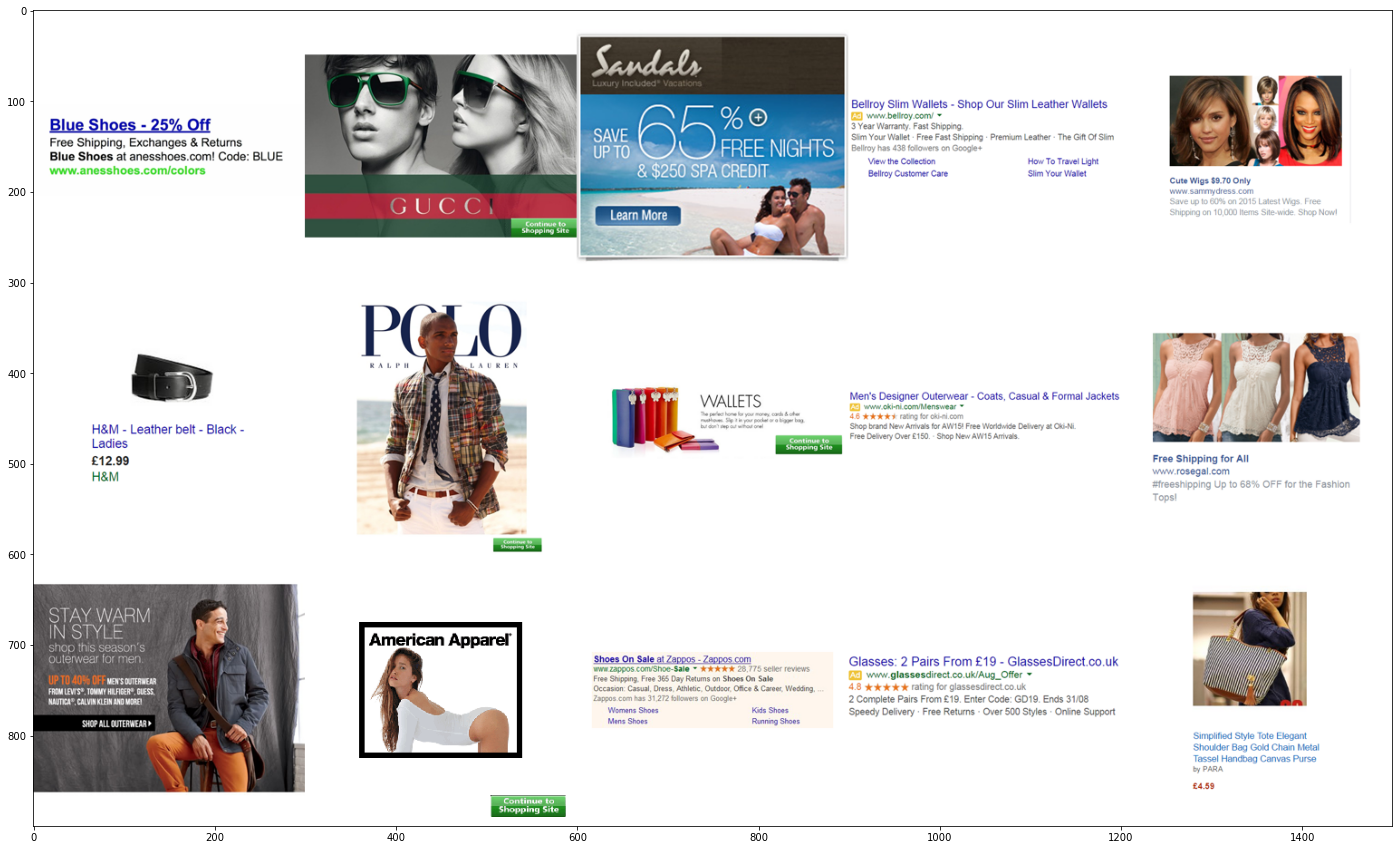


Figure 1: Example of the three Ad types. Rich Media (left), text (middle), image(right)

Source: Original research data

Using Python, we applied in database unsupervised machine learning by organizing the subjects in clusters with similar personality traits according to their OCEAN score. From the clusters of personalities, we could draw their profiles and analyze behavior to the advertisements.

Finally, we applied supervised machine learning as RF, GB and NN to predict consumers behavior on feedback of advertisements. We tested which method is more accurate and the importance of each parameter to the final prediction capacity.

1. Results and Discussion

First step was to collect the data from the dataset from different .csv files to one. Content of full dataset:

* There are 20 Ad categories —> 300 Ads and 120 Users in total.
* Each Ad Category folder contains a collection of 15 image files.
* 300 Ads —> 100 Rich Media Ads, 100 Image Ads and 100 Text Ads. Each ad receives a rating for each user between 1 to 5 for how willing they are to click in the ad.
* An Ad Category is considered to be "clicked" if it contains an advert that is rated “4” or “5”.
* Each User folder (for e.g. U0001) contains 6 CSV files:
* (CSV #1) U0001-INF.csv (14 cols, 1 row) contains personal information (e.g. Gender, Age, Income...).
* (CSV #2) U0001-PREF.csv (5 cols, 1 row) contains preferences (e.g. Most visited websites, most read books...). Each field is a CSV of categories (e.g. Comedy, Horror, Mystery...)
* (CSV #3) U0001-B5.csv (3 cols, 10 rows) contains answers to Big Five Inventory-10 personality test.
* (CSV #4 & #5) Both U0001-IM-POS.csv & U0001-IM-NEG.csv(5 cols, 2 rows each) refer to the contents of the respective folders and user reactions (e.g. "my cats" for an image in POS and "violence" for an image in NEG) to the same.
* (CSV #6) U0001-RT.csv (20 cols, 2 rows) contains user rating for each ad in each ad category along a Likert Scale ranging from +1 to +5. +4 and +5 corresponds to a "click" in the paper.

Since content of CSV #4 and #5 are not relevant for the work, data was left out of further steps.

Using Pandas to recover the path of each .csv file and extract all information to series, the work proceeded with creation of one single data frame from all series from CSV files #1, #2, #3 and #6. Output in Table 1:

Table 1. Head of data frame with all information from users and their grade to each advertisement in each category. Ad 1 from Category 1 for example is in column Cat1\_1

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| user | E\_score | A\_score | C\_score | N\_score | O\_score | Gender | Age | Income | Cat1\_1 | Cat1\_2 | … |
| U0001 | 5 | 10 | 10 | 5 | 8 | F | 62 | 1 | 1 | 1 | … |
| U0002 | 3 | 4 | 4 | 10 | 8 | F | 26 | 1 | 3 | 1 | … |
| U0003 | 9 | 9 | 8 | 3 | 8 | M | 22 | 1 | 4 | 3 | … |
| U0004 | 5 | 8 | 8 | 6 | 8 | F | 24 | 1 | 1 | 1 | … |
| U0005 | 5 | 8 | 7 | 5 | 9 | F | 34 | 1 | 3 | 3 | … |

Source: Original research results

* 1. **Descriptive statistics**

From the database we extracted descriptive graphs to better understand our data. The OCEAN score for users varies between 2(lowest score) to 10(highest score). The scores frequencies for all 120 users are presented in Figure 2 below:

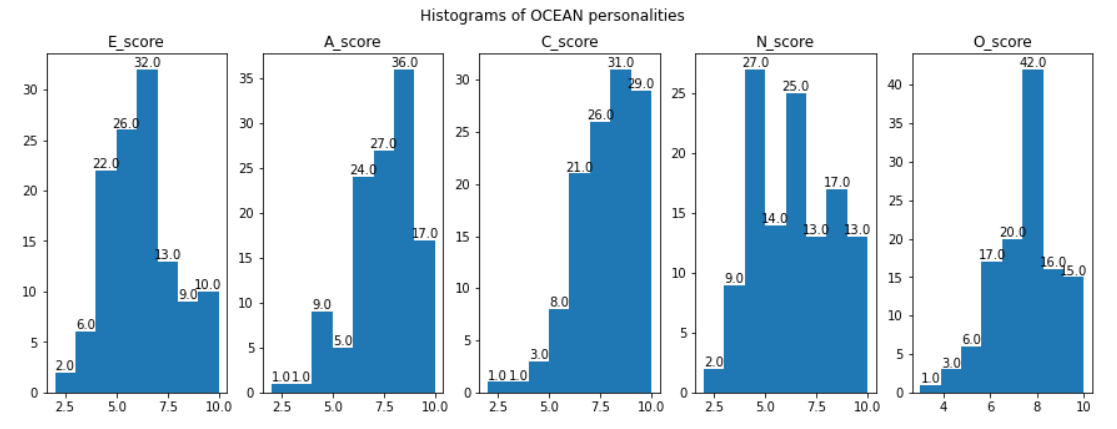


Figure 2. Histogram of scores for each of OCEAN personalities (E: Extroversion, A: Agreeableness, C: Conscientiousness, N: Neuroticism, O: Openness)

Source: Original research results.

From histograms of Figure 2 we observe the distribution of each personality factor in our database according to the OCEAN test done by users. While Extroversion has a distribution with normal curve shape, users tend to score high in Conscientiousness and Agreeableness.

We could also Extrapolate from data the percentage of positive clicks (4 and 5) for each category as shown in Table 2 below:

Percentage of positive feedbacks per category (4 and 5 ratings)

|  |  |
| --- | --- |
| Categories | Positive feedbacks (rates 4 or 5)% |
| 6\_Consumer Electronics | 20.61% |
| 7\_Console and Video Games | 19.89% |
| 5\_Media | 19.33% |
| 10\_Grocery | 18.83% |
| 1\_Clothing and Shoes | 17.00% |
| 17\_Computer Software | 16.67% |
| 13\_Jewelery and Watches | 15.78% |
| 15\_Stationary and Office Supplies | 15.33% |
| 4\_Health and Beauty | 14.83% |
| 9\_Outdoor Living | 14.72% |
| 18\_Sports | 14.00% |
| 19\_Toys and Games | 13.94% |
| 11\_Home | 12.17% |
| 2\_Automotive | 10.61% |
| 14\_Musical Instruments | 9.50% |
| 16\_Pet Supplies | 9.39% |
| 8\_Tools and Hardware | 9.00% |
| 3\_Baby | 8.17% |
| 12\_Betting | 4.67% |
| 20\_Social Dating Sites | 4.50% |

Source: Original research results

The favorite category from users was “Consumer Electronics” with the most positive feedbacks superior of 20%. The “Social Dating Sites” had the least positive feedbacks with only 4,5%, from which we can infer that this is a “niche” category.

* 1. **Unsupervised machine learning**

The previous data frame is ready for use in Unsupervised Machine Learning. To analyze the group of 120 people a clusterization was done to reduce number of attributes according to the OCEAN personality score.

To identify the best quantity of clusters, as showed in Figure 3, it was applied the Elbow Method based on the 'inertia': distance between each observation and the center of the cluster. In this method it is possible to observe the gain of variance for each new additional cluster created.

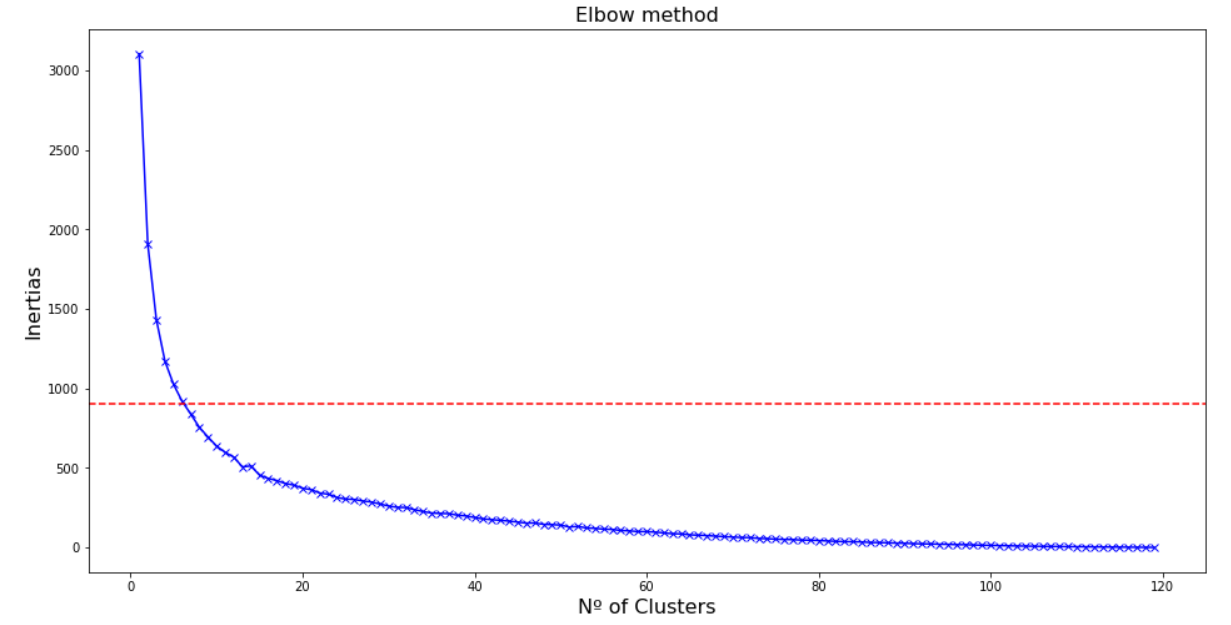


Figure 3. Elbow Method with border line on the sixth cluster. The first 6 clusters have the biggest gain of information

Source: Original research results

Using K Means method which calculates the distance between each observation and the closest center, it was set up 6 clusters with 6 centroids as basis for clusterization. The centroids are visible in Table 3 with a center value for each of the OCEAN scores that help us identify the cluster central personality trait.

Table 3. Cluster centroids

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kmean cluster | E\_score | A\_score | C\_score | N\_score | O\_score |
| 1 | 5.39 | 6.47 | 6.91 | 8.13 | 8.69 |
| 2 | 5.83 | 6.16 | 6.58 | 5.33 | 5.95 |
| 3 | 6.04 | 7.13 | 8.43 | 4.52 | 8.47 |
| 4 | 8.81 | 8.18 | 8.12 | 3.50 | 7.87 |
| 5 | 3.46 | 6.46 | 5.06 | 7.86 | 7.40 |
| 6 | 4.89 | 8.26 | 9.00 | 5.89 | 7.47 |

Source: Original research results

For visualization we plot every cluster centroid in the radar map. The radar map displays the cluster centroid with 5 edges representing the scale of each OCEAN score in Figure 4.

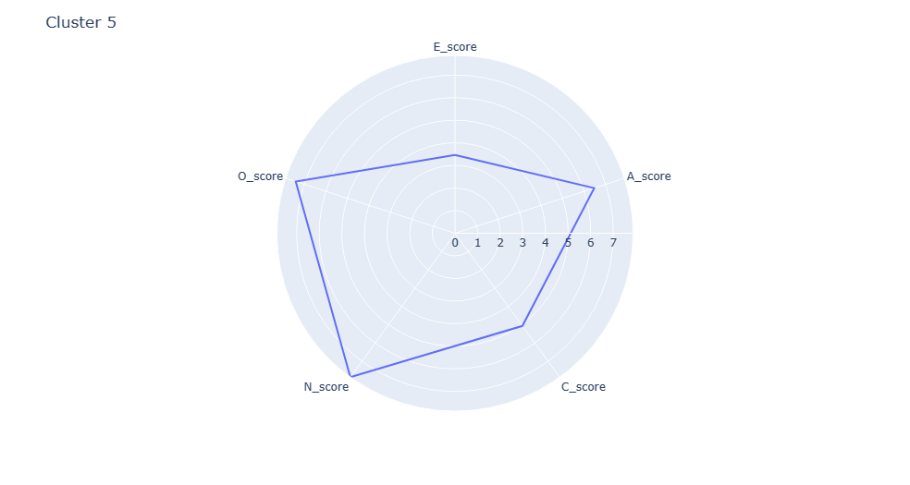
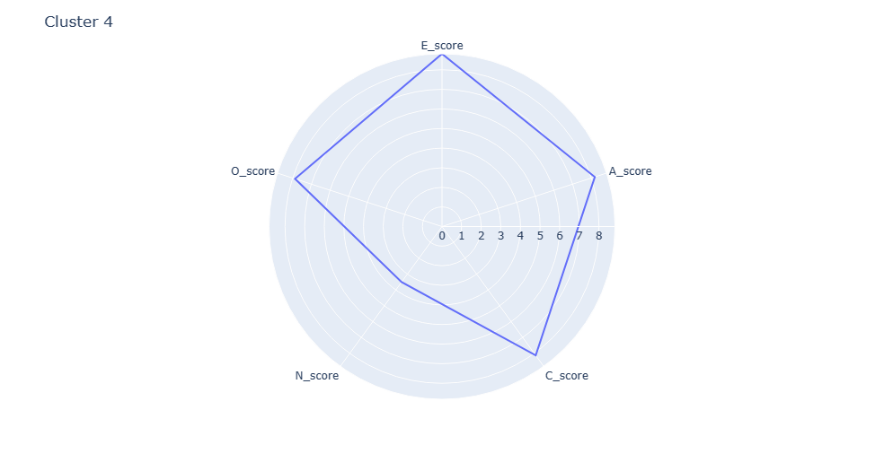
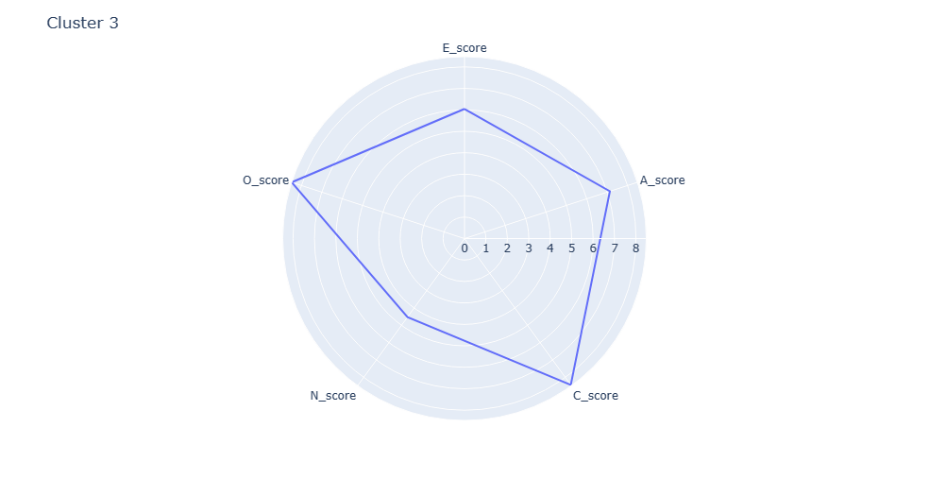
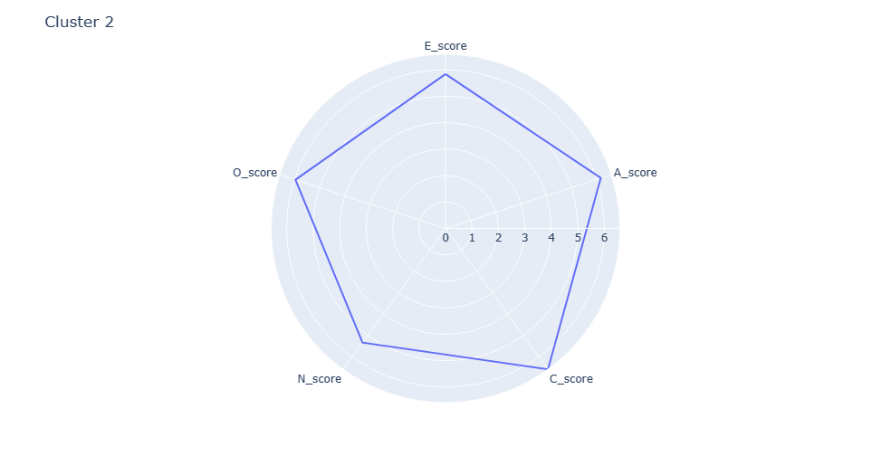
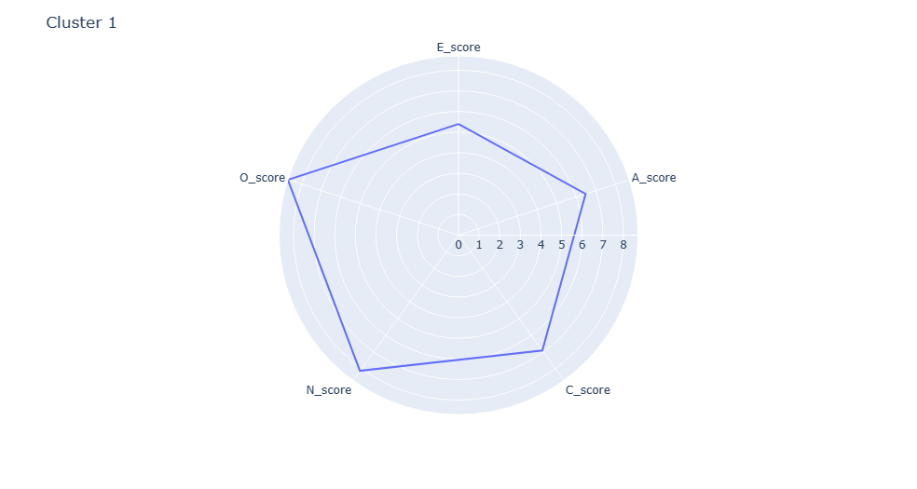
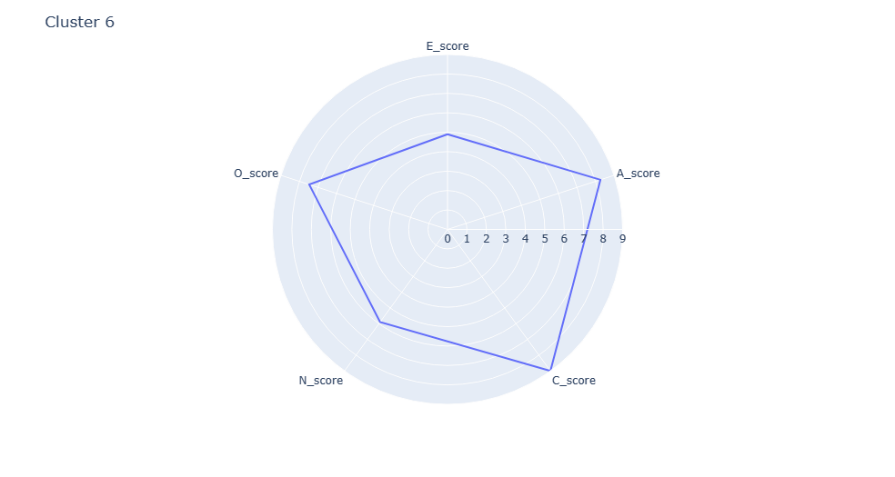
 

Figure 4. Radar map of all centroids of OCEAN personality clusters.

Source: Original research results

We can observe cluster 6 represent users with highest Agreeableness and Conscientiousness but lower Extroversion compared with clusters 3 and 4. The cluster 2 represents users with average scores in all sides. Cluster 5 represents a group with high Openness and Neuroticism but lower Extroversion and Conscientiousness compared to cluster 1.

With the pairplot in Figure 5 we can see the map of density of each group in each OCEAN score in the main diagonal plus marks for each bidimensional plot between 2 OCEAN scores.

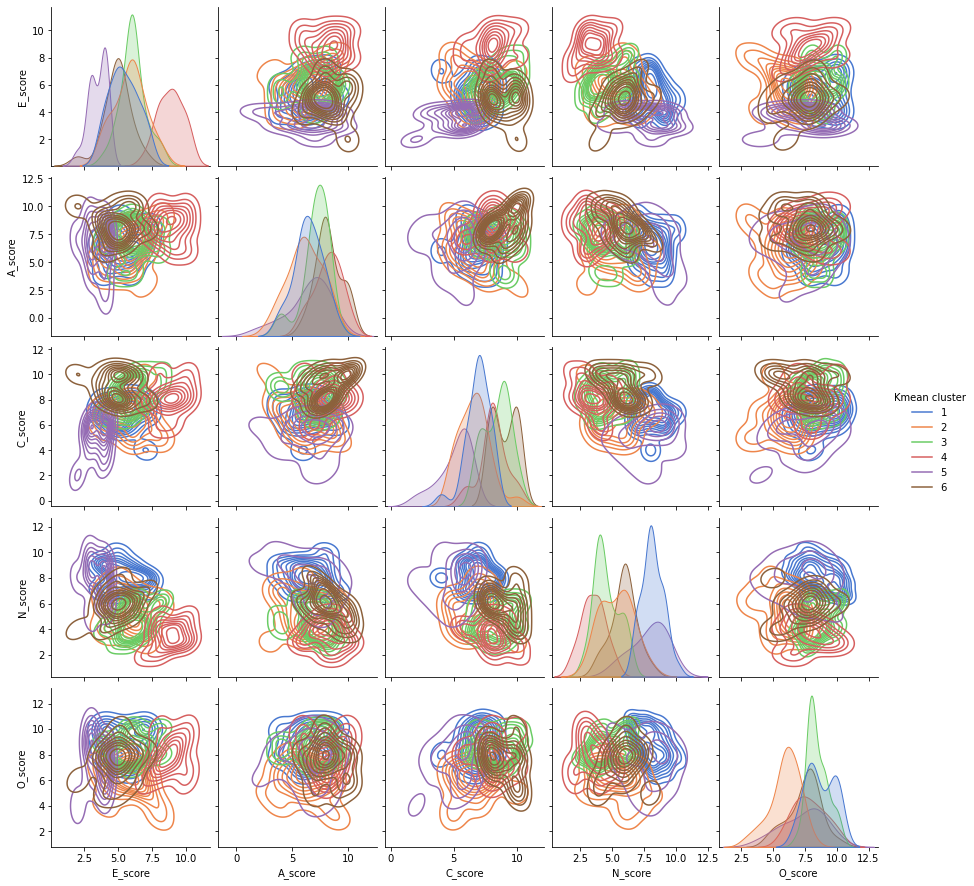


Figure 5. Pair plot with correlation between all OCEAN scores and the density of each of the 6 clusters (color) in every bi-dimensional plot between 2 scores.

Source: Original research results

In the figure we can see characteristics of each cluster, as for example, in the first main diagonal plot for E\_score the cluster 4 (pink) represents users with highest scores and cluster 5 (purple) with lowest scores as observed also in radar plot. In pair plot the density between 2 scores is also displayed as for example between C\_score and E\_score cluster 5 (purple) represents the users with lower score in both of them and cluster 4 (pink) represents users with high score in both of them.

For further analysis of user’s characteristics in each cluster we extract the percentages of gender, age and income in Table 4.

Table 4. Users characteristics per cluster for average of Age, proportion of Gender and proportion of Income

|  |  |  |  |
| --- | --- | --- | --- |
| OCEAN cluster | Age | Gender (F/M) | Income (0: <11kUSD, 1: 11~50kUSD, 2: 50~85kUSD, 3: >85kUSD) |
| 1 | 32.0 | 69.57% / 30.43% | 0: 34.78%----- 1: 56.52%----- 2: 8.70%----- 3: 0.00% |
| 2 | 31.8 | 54.17% / 45.83% | 0: 16.67%----- 1: 45.83%----- 2: 29.17%----- 3: 8.33% |
| 3 | 36.0 | 60.87% / 39.13% | 0: 21.74%----- 1: 47.83%----- 2: 21.74%----- 3: 8.70% |
| 4 | 32.5 | 50.00% / 50.00% | 0: 12.50%----- 1: 62.50%----- 2: 25.00%----- 3: 0.00% |
| 5 | 22.2 | 73.33% / 26.67% | 0: 33.33%----- 1: 26.67%----- 2: 26.67%----- 3: 13.33% |
| 6 | 33.3 | 78.95% / 21.05% | 0: 21.05%----- 1: 42.11%----- 2: 21.05%----- 3: 15.79% |

Source: Original research results

With Table 4 the cluster data can be deeper analyzed. The cluster 3 has the older users while the cluster 5 the younger ones. Cluster 6 contains more users with high income and cluster 1 has the users with lowest incomes in comparison.

We also extrapolated in Table 5 the favorite categories for each personality cluster by the best rated categories. From the table we can correlate characteristics of each personality profile and the categories that are more attractive to each of them.

Table 5. Top 5 categories with most positive feedbacks for each cluster (continue)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| OCEAN cluster | 1º rated | 2º rated | 3º rated | 4º rated | 5º rated |
| 1 | Category 7 (17.39%) | Category 5 (15.07%) | Category 10 (13.04%) | Category 4 (11.59%) | Category 13 (10.72%) |
| 2 | Category 10 (25.56%) | Category 5 (23.61%) | Category 6 (23.06%) | Category 7 (21.11%) | Category 1 (20.0%) |
| 3 | Category 9 (26.67%) | Category 6 (22.61%) | Category 1 (19.42%) | Category 7 (18.84%) | Category 10 (17.39%) |

Table 5. Top 5 categories with most positive feedbacks for each cluster (conclusion)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| OCEAN cluster | 1º rated | 2º rated | 3º rated | 4º rated | 5º rated |
| 4 | Category 7 (30.83%) | Category 6 (27.5%) | Category 13 (27.5%) | Category 10 (25.83%) | Category 1 (25.0%) |
| 5 | Category 6 (21.78%) | Category 5 (20.89%) | Category 17 (19.56%) | Category 7 (17.33%) | Category 1 (16.0%) |
| 6 | Category 6 (23.16%) | Category 5 (18.25%) | Category 17 (17.54%) | Category 7 (15.44%) | Category 10 (15.44%) |

Source: Original research results

We can make a parallel with the user demographics. For example, both cluster 5 and 6 are more attracted to Consumer Electronics, media and computer software. Characteristics between clusters:

* Cluster 5 represents users with high neuroticism and high openness, while cluster 6 represents users with high Conscientiousness and openness.
* Both clusters have the most percentages of users with highest incomes,
* Cluster 5 represents younger users with 22 age average and 6 older users with 33 age average.
  1. **Supervised machine learning**

For supervised machine learning we modified the dataset to prepare data for the machine learning. We keep the OCEAN personality clusters as our parameter of personality, other user characteristics are maintained as “Gender”, “Age” and “Income”. “Gender” is converted to dummy “Gender\_M” with value 1 for masculine and 0 for feminine. “Income” variable is qualitative, so we converted from numeric representation “int” to “category” type.

Also as independent variables we extracted from the data frame information about the ads as “Media type” (text, image, rich media) and “Category” (1 to 20). We create dummies for Media type by splitting all types of media in 3 columns with “1” value on the media type of the ad and “0” value in all other values. “Category” variable is qualitative, so we converted from numeric representation “int” to “category” type.

The dependent variable is the feedback rating of advertisement. The dependent variable was also dummyfied with “Negative click” or “0” for ratings “1”, “2” and “3”, and “Positive Click” or “1” for ratings “4” and “5”. The machine model should predict with base on user characteristics and ad parameters what is the output feedback. If positive or negative.

The original dataset was melted to have one variable (or dummy variable) per column. The data frame head is displayed in Table 6.

Table 6. Melted data frame head with all variables in their own columns. Output variable in column “Feedback” with all 36000 ad ratings (120 users x 300 ads)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ad | Age | Income | OCEAN cluster | Ad category | Gender\_M | Ad type\_Image | Ad type\_Rich Media | Ad type\_Text | Feedback\_Pos |
| 1 | 62 | 1 | 4 | 1 | 0 | 0 | 0 | 1 | 0 |
| 2 | 62 | 1 | 4 | 1 | 0 | 0 | 0 | 1 | 0 |
| 3 | 62 | 1 | 4 | 1 | 0 | 0 | 0 | 1 | 0 |
| 4 | 62 | 1 | 4 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5 | 62 | 1 | 4 | 1 | 0 | 0 | 0 | 1 | 0 |

Source: Original research results

We separated the data into target set as “labels” with the values we want to predict and the features which are the independent variables used to train the machine as the inputs for prediction of target values. The target variable is “Feedback\_Pos” as 1 for positive and 0 for negative. The features are all the other columns “Age”, “Income”, “Kmean cluster” “Ad category”, “Gender\_M” and “Ad type”.

With defined parameters we prepared the cross-validation of the models by splitting the dataset into training sample with 75% of data to be used for development of model and test sample with 25% of the data to be use for evaluation of the model. Finally, the shape of our data is prepared for optimal understanding of machine and development of models with all values in numerical data and samples for training and test defined.

The following machine learning models are trained by randomized techniques as bootstrapping. This can cause problems on reproducibility of the data. Different machines with different capacities and different parameters can achieve different results. In this work we will use techniques to optimize our parameters in each technique to achieve optimal results in all of them.

* 1. **Random Forest (Supervised machine learning)**

The first method of machine learning to explore is a variation of decision trees. The decision tree is a sequence of binary segmentations of independent variables of data with goal of achieving a homogeneous response variable. The model will use the training data to check which input variables lead to the output “Feedback\_Pos” from users and select criterions from those variables to split data. Example of small decision three with 3 levels in Figure 6.

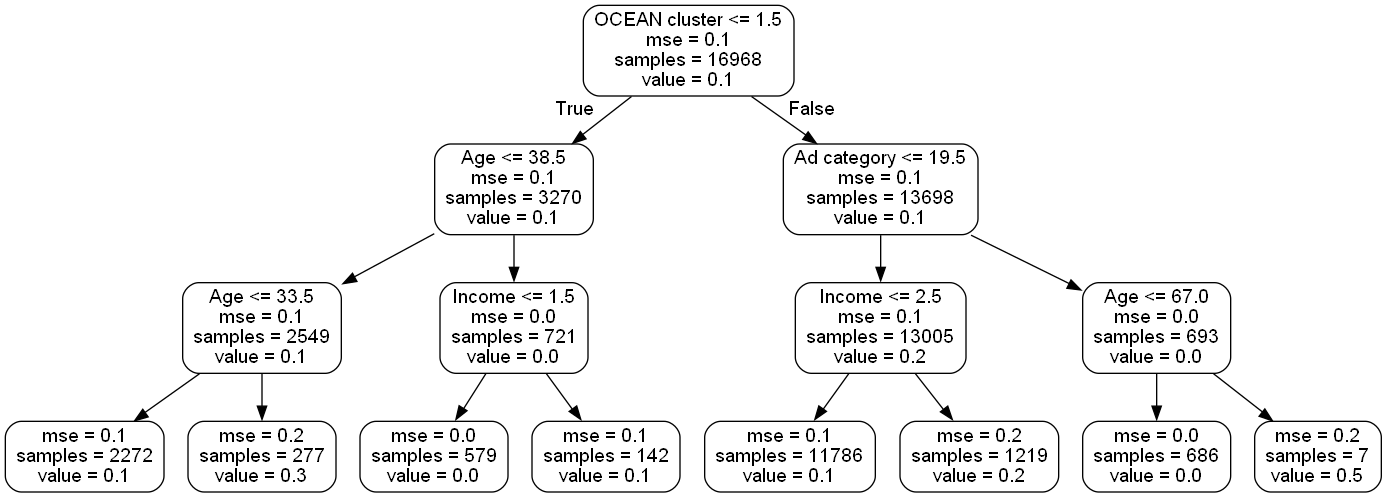


Figure 6. Decision tree example with first split leaf OCEAN cluster\_1 <= 1.5. If case is ‘False’ the decision tree moves right and if ‘True’ moves left to next split leaf. Every leaf has a predict value that is closer to ‘0’ or closer to ‘1’ in which will define the predicted value of the model

Source: Original research results

During each split in the tree building process, the algorithm tries to find the best feature to split on randomly from the entire set. The function Mean Squared Error [MSE] is used to define the best quality of split.

To increase accuracy of model we used techniques as “Bagging” which is an “Aggregation” that takes the average of all predictions of same algorithm where each prediction is built several times on a random subset of training data with replacement (bootstrapping). Applying the “Bagging” method with “Decision trees” we have the “Random Forest” algorithm which can introduce additional randomness and reduce overfitting of the model.

The “Random Forest” algorithm has some hyperparameters which influence its performance in avoiding overfitting and accurately predict new data. Some of these hyperparameters are:

* n\_estimators: This hyperparameter controls the number of decision trees in the boosted ensemble. It's typically set to a higher value for more complex datasets, and a lower value for simpler datasets, to prevent overfitting.
* max\_depth: The maximum depth of each decision tree in the forest. A deeper tree may capture more complex relationships in the data, but can also lead to overfitting.
* min\_samples\_split: The minimum number of samples required to split an internal node. Increasing this parameter can reduce overfitting by requiring a larger number of samples to make a split.
* min\_samples\_leaf: The minimum number of samples required to be at a leaf node. This parameter also helps to control overfitting by preventing the model from creating leaves with too few samples.

Other hyperparameters can also be used to change performance of algorithm with the better combination of them. Using Skicit-learn library in python we performed a random search from samples of hyperparameters with “RandomizedSearchCV” function for random search and “RandomForestRegressor” for Random Forest model.

The random search algorithm is a hyperparameter optimization technique that samples from a distribution of hyperparameters to find the best combination that maximizes the performance of the model. The hyperparameters searched:

* n\_estimators: [50, 100, 300].
* max\_depth: [18, 22, 27].
* min\_samples\_split: [20, 25, 30].
* min\_samples\_leaf: [2, 4, 6].

We performed randomized search over the specified hyperparameter space, using 3-fold cross-validation and running 100 iterations. With output of 'n\_estimators': 100, 'min\_samples\_split': 20, 'min\_samples\_leaf': 4, 'max\_depth': 18 from random search we use same parameters in our “Random Forest” model in training data and we test performance in the test data.

The performance of the model can be determined by the True Positive Rate [TPR] also known as sensitivity which represents the proportion of correctly positive cases identified by the model, True Positive[TP] ratio to True Positive +False Negative [FN]. Other parameter of performance is False Positive Rate [FPR] represents the proportion of actual negative cases that are incorrectly classified as positive, False Positive[FP] in ration to False Positive +True Negative [TN]. The graphical representation of performance is expressed in the Figure 7 with ROC Curve that plots the trade-off of TPR and FPR for different cut-offs for the same model.

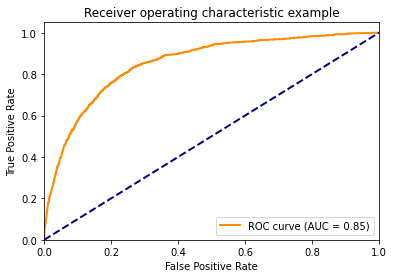


Figure 7. ROC curve for Random Forest Model

Source: Original research results

We can validate performance of model by Area Under Curve [AUC]. The bigger the area the best tradeoff between sensitivity and sensibility. The AUC-ROC ranges from 0.5 (random guessing) to 1 (perfect classification). Other metrics of model performance are presented in Table 7 with a cut-off of 0.5 (If probability is 50% or higher the prediction of “Feedback\_pos” is considered as event, if lower as non-event):

* Precision = TP / (TP + FP), is the fraction of true positives out of the total predicted positives.
* Recall = TP/ (TP + FN), Recall is the fraction of true positives out of the total actual positives.
* F1-score = 2 \* (precision \* recall) / (precision + recall). F1-score is the harmonic mean of precision and recall, and provides a balanced measure of the two metrics.
* Accuracy = (TP + TN) /(TP + TN + FP + FN). Accuracy is the fraction of correct predictions out of the total number of predictions.

Table 7. Final metrics of Random Forest model with 0.5 cut-off

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | sample size |
| 0 | 0.89 | 0.98 | 0.93 | 7800 |
| 1 | 0.64 | 0.25 | 0.36 | 1200 |
| accuracy |  |  | 0,88 | 9000 |

Source: Original research results

With Random Forest model we could achieve 88% accuracy with f1-score 93% for negative feedbacks and 36% for positive feedbacks. It is harder for the model to predict the positive feedbacks with 64% true positives out of total predicted positives but only 25% real positives were correctly predicted by the model.

We can extract from the model the relative importance of each variable to the prediction using the Skicit-learn library with function “feature\_importances\_”. The Table 8 show the contribution of each variable in our dataset for the model prediction:

Table 8. Importance of variables for Random Forest Model

|  |  |
| --- | --- |
| Variable | Importance |
| Ad category | 0.36 |
| Age | 0.26 |
| OCEAN cluster | 0.14 |
| Income | 0.11 |
| Ad type\_Rich Media | 0.04 |
| Ad type\_Text | 0.04 |
| Gender\_M | 0.03 |
| Ad type\_Image | 0.02 |

Source: Original research results

We see that ad category has the biggest representation for prediction. That is also because the parameter has 20 inputs and each of them influence the model, where “Ad category” is the sum of all those influences. We also see age of user has a big influence followed by the OCEAN cluster with 14%. The sum of all Ad types influences around 10% of prediction.

* 1. **Gradient Boosting (Supervised machine learning)**

Boosting methods are sequential models in which the response variable of an iteration is the ’error ’of the previous one. This allow the algorithm to improve the error of the previous model. The “Gradient Boosting” uses the “Decision Tree” algorithm with “Boosting” iteration of the error of previous model.

The hyperparameter that determines the contribution of each tree in the ensemble is the “Learning rate”. the learning rate shrinks the contribution of each tree by a factor between 0 and 1. A lower value means each tree contributes less, which makes the learning slower but can lead to better generalization. A higher value means each tree contributes more, which makes the learning faster but can lead to overfitting.

We use random search algorithm for hyperparameter optimization and fit best parameters. The hyperparameters are:

* n\_estimators: [50, 100, 500, 1000]. It is the number of boosting stages to perform. The higher the number of estimators, the more complex the model becomes, and it can lead to overfitting.
* Learning rate: [0.01, 0.1, 0.5]. It is the rate at which the boosting algorithm shrinks the feature weights.
* max\_depth: [3, 5, 7]. It controls the maximum depth of each tree in the ensemble.
* min\_samples\_split: [2, 5, 10]. It is the minimum number of samples required to split an internal node.
* min\_samples\_leaf: [1, 2, 4]. It is the minimum number of samples required to be at a leaf node.

We performed randomized search over the specified hyperparameter space, using 3-fold cross-validation and running 100 iterations. With output of 'learning\_rate': 0.1, 'max\_depth': 7, 'min\_samples\_leaf': 4, 'min\_samples\_split': 10, 'n\_estimators': 100, from random search we use same parameters in our “Gradient Boosting” model with function “GradientBoostingClassifier” in training data and we test performance in the test data to extract ROC curve of Figure 8.

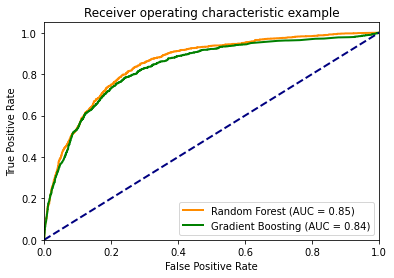


Figure 8. ROC curve for Random Forest and Gradient Boosting Model

Source: Original research results

The output AUC was similar with slightly worse output compared to Random Forest. Other metrics of model performance shown in Table 9 with a cut off of 0.5.

Table 9. Final metrics of Gradient Boosting model with a 0.5 cut off

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Sample size |
| 0 | 0.88 | 0.99 | 0.93 | 7800 |
| 1 | 0.69 | 0.12 | 0.21 | 1200 |
| accuracy |  |  | 0.88 | 9000 |

Source: Original research results

With Gradient Boosting model we could achieve 88% accuracy with f1-score 93% for negative feedbacks and 21% for positive feedbacks. It is harder for the model to predict the positive feedbacks with 69% true positives out of total predicted positives but only 12% real positives were correctly predicted by the model.

We extract from the model the relative importance of each variable to the prediction using Skicit-learn library. The Table 10 show the contribution of each variable in our dataset for the model prediction:

Table 10. Importance of variables for Gradient Boosting Model

|  |  |
| --- | --- |
| Variable | Importance |
| Ad category | 0.39 |
| Age | 0.24 |
| OCEAN cluster | 0.14 |
| Income | 0.12 |
| Ad type\_Text | 0.05 |
| Ad type\_Rich Media | 0.03 |
| Gender\_M | 0.02 |
| Ad type\_Image | 0.02 |

Source: Original research results

In GB model the order of importance of input variables is similar than RF model. Here Ad category gain even more importance to the prediction. The personality variable has same importance as RF model of around 14%.

* 1. **Artificial Neural Networks (Supervised machine learning)**

Artificial Neural Networks (NN) is a machine learning algorithm that has its structured based on the neural network of the brain. The algorithm is trained with inputs and outputs of a dataset by modeling through layers of neurons. Each neuron receives information from inputs or previous layers, processes information with an activation function, multiplies the information by weights and send to next layer of neurons or output.

The values of weights start random, but after obtaining predicted output, the error with real output by an error function like mean squared error is taken and optimized by an algorithm based on gradient descent. In last step of loop the function is back propagated to the network updating all parameters to obtain minimum error possible in predicted output.

From Skicit-learn and tensorflow libraries we performed a random search function to optimize best parameters of Neural Network model and use of function “keras” to create model. The hyperparameters to be optimized are:

* num\_layers: [1, 2, 3]. Number of layers of hidden neurons in the model,
* num\_neurons: [32, 64, 96]. Number of neurons in each hidden layer,
* activation: ['relu', 'sigmoid']. Activation function to use in hidden layers,
* optimizer: ['adam', 'sgd']. Optimization algorithm for training of model.

We performed randomized search over the specified hyperparameter space, using 3-fold cross-validation and running 100 iterations. The output from random search is:

* 'optimizer': 'adam' indicates that the optimizer used during training is the Adam optimizer. Adam is a popular optimization algorithm used to update the parameters of neural networks during training.
* 'num\_neurons': 96, indicates that the number of neurons in each hidden layer is 96.
* 'num\_layers': 2, indicates that the number of hidden layers in the neural network is 2.
* 'activation': 'relu'. indicates that the activation function used in the hidden layers is the rectified linear unit (ReLU) activation function.

We use same parameters in our “Neural Network” model using the “KerasClassifier” object in training data and we test performance in the test data with 100 epochs (100 iterations of backpropagation and error reduction) in Figure 9.

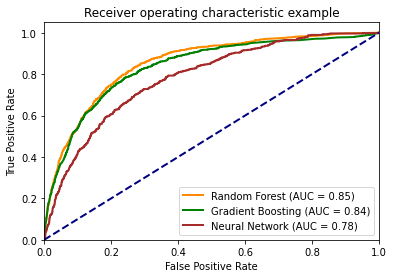
****

Figure 9. ROC curve for Random Forest, Gradient Boosting and Neural Network Model

Source: Original research results

The NN model have the worst prediction performance from all models with 0,78 AUC. Other metrics of model performance in Table 11 with a cut off of 0.5.

Table 11: Final metrics of Neural Network model with 0.5 cut off

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | sample size |
| 0 | 0.87 | 1.00 | 0.93 | 7800 |
| 1 | 0.74 | 0.04 | 0.08 | 1200 |
| accuracy |  |  | 0.87 | 9000 |

Source: Original research results

With Neural Network model we could achieve 87% accuracy with f1-score 93% for negative feedbacks and 8% for positive feedbacks. The model had the worst performance to predict the positive feedbacks with 74% true positives out of total predicted positives but only 4% real positives were correctly predicted by the model. Even if it was the case that 100% of negative feedbacks were correctly predicted by the model, the general performance of prediction was worse

We cannot extract from Keras library for the NN model the table of importance for input variables. As our worst performing model, it will be considered the baseline. Our desired model should have improved performance upon baseline.

1. Conclusions

With the robust data frame of users, ads and their characteristics we could use unsupervised and supervised machine learning to extend our understanding on marketing and psychology according to the objectives of this work. We could find clusters of common personality traits based on OCEAN score and explore their preferences, consumer behavior and other group demographic characteristics.

For supervised machine learning we used 3 methods of machine learning (DF, GB and NN) to model the behavior of consumers for advertisement preference. The hyperparameters of each model were optimized to increase accuracy of the model and avoid overfitting. We found explanatory algorithms and created machine intelligence which understand the correlation between consumer personality and consumer response to different types of ads.

The best performing method was the Random Forest. While all 3 models had similar accuracy, prediction of negative feedbacks, the RF model had the best performance to predict the positive feedbacks and best AUC from ROC curve.

The knowledge of this work can be used for case studies to improve the engagement of target public to value and explore outreach strategies with more efficacy aligned to the personality, needs and interests of the customer. Future work could extend the analysis of this study by applying regression models as for example Multinomial logistic regression or Multilevel Regression Modeling to improve efficiency of prediction capacity. Also, the hyperparameters of each machine learning model could be further optimized for a more robust model and understand if further improvements can be achieved using more machine capacity.

1. **Acknowledgment**

I acknowledge the important help of Dr. Riad Rashid to the conclusion of this work. With his guidance this research could be better developed in direction of science in data field.

1. References

Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., and Lin C.-J.. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 2008.

Gómez-Adorno, H.; Markov, I.; Sidorov, G.; Posadas-Durán, J.P.; Sánchez-Pérez, M.; Chanona-Hernández, L. Improving feature representation based on a neural network for author profiling in social media texts. Comput. Intell. Neurosci. 2016, 2016, 1638936. [CrossRef]

Murphy, E.R.; Illes, J.; Reiner, P.B. Neuroethics of neuromarketing. J Consum Behav. 2008, 7, 293–302. [CrossRef]

Racine, E.; Bar-Ilan, O.; Illes, J. FMRI in the public eye. Nat. Rev. Neurosci. 2005, 6, 159–164. [CrossRef]

Rammstedt B. and John O. P. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality, 2007.

Renvoise, C.; Morin, P. Neuromarketing: Understanding the "Buy Button" in Your Customer’s Brain; Thomas Nelson Publishers: Nashville, TN, USA, 2007.

Roffo, G., & Vinciarelli, A. (2016, August). Personality in computational advertising: A benchmark. In 4 th Workshop on Emotions and Personality in Personalized Systems (EMPIRE) 2016 (p. 18).

Schreiner, M.; Fischer, T.; Riedl, R. Impact of content characteristics and emotion on behavioral engagement in social media: Literature review and research agenda. Electron. Comm. Res. 2019, 21, 329–345. [CrossRef]

Turkyilmaz C. A., Erdem, S. and Uslu. A. The effects of personality traits and website quality on online impulse buying. 2015. International Conference on Strategic Innovative Marketing.