**Abstract**

Here we explore the Ad data to find various insights on how certain features affect an Ad’s success., and assess the scope of revenue that can be generated by a proposed ad. Further, we predict whether the revenue generated will cover costs to produce and air the ad. This will help guide decision making for the firm, as they will want to pursue ads that are likely to generate a net gain for their clients— thereby bolstering the advertising firm’s reputation.

1. **Introduction**

Marketing is the bridge between the product and the customer. During the holiday cheer, retail brands, big and small, wish to earn considerable profits, and therefore, invest significantly in advertising. In the hope of an increased footfall in stores, these organizations communicate with their current as well as potential customers through advertising campaigns so as enhance the buyer’s response to their offerings and hence achieve profitable sales in the long run. An understanding of the effectiveness of advertisements holds great significance and contributes to the productivity of advertisers in terms of effective allocation of their marketing fund and improves their customer base too.

1. **Background**

Authors in [1] have analyzed the differences between online and traditional advertising. They have used an Ad evaluation system to understand the effectiveness of on advertisement on the audience, which is based on four dimensions: psychological effectiveness, communication effectiveness, economic effectiveness and social effectiveness. We observe in our analyses too that people who connect to an emotionally, are likely to generate a net gain for the company.

Authors in [2] have explained the importance of video advertisements and have outlined an idea to predict the effectiveness of the ad, based on the historical data as well as the effectiveness of similar kinds of video ads. They have propped a multi model mixture-based algorithm to predict the effectiveness automatically, before it is aired on the multimedia platforms.

Paper [3] studies the consumer response to sponsored advertisements that are displayed on social media. The authors have assessed the effects of user perceptions of privacy risk, intrusiveness concerns and utilities of sponsored advertising on consumer attitudes and their purchase intent.

1. **Methodology**

3.1 About the Dataset

The Dataset was taken from [4]. Dataset consists of 12 variables out of which 9 are factor type variables. Our target variable is netgain, which is a binary variable containing values True and False. There are no missing values present in the data set.

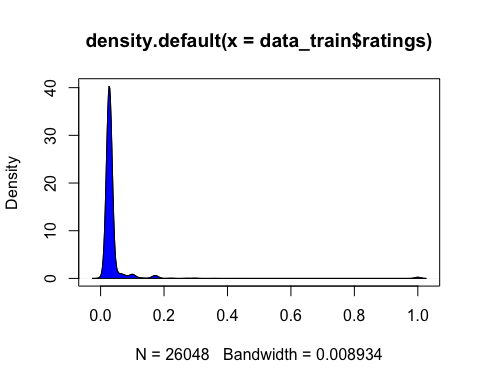
Table 1:Dataset Description

|  |  |  |
| --- | --- | --- |
| **Data** | **Data Description** | **Data Type** |
| id | Unique id for each row | Factor |
| ratings | Metric out of one which represent show much of the targeted demographic watched the advertisement | Numeric |
| air location | Country of origin time when the advertisement was aired | Factor |
| air time | Time at which the advertisement was aired | Factor |
| average\_runtime (minutes per week) | Minutes per week the advertisement was aired | Integer |
| targeted\_sex | Sex that was mainly targeted for the advertisement | Factor |
| genre | The type of advertisement | Factor |
| industry | The industry to which the product belonged | Factor |
| relationship\_status | The relationship status of the most responsive customers to the advertisement | Factor |
| expensive | A general measure of how expensive the product or service is that the ad is discussing | Factor |
| money\_back\_guarantee | Whether or not the product offers a refund in the case of customer dissatisfaction | Factor |
| net gain (target) | Whether the ad will incur a gain or loss when sold | Factor |

3.2 Exploratory Data Analysis

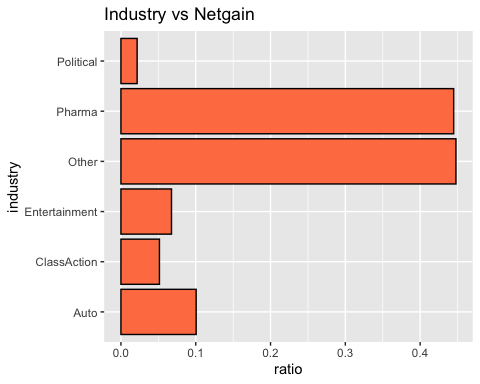
*The following graphs were coded with the help of* [5]

*a) Analyzing the reach of the ads.*



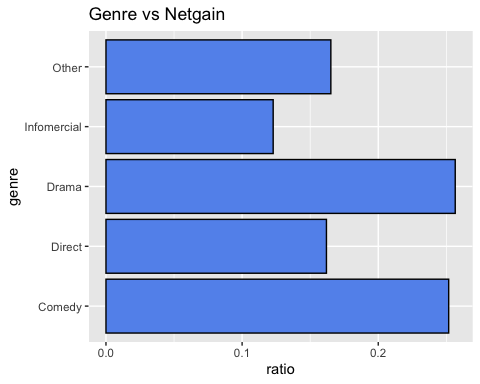
**Inference:** We observe that the density plot of ratings is left skewed, which means that the targeted audience did not watch the advertisement. We can infer from this observation that the audience was not targeted efficiently. The company needs to invest in online as well as offline advertisements, to increase their reach.

*b) What industry is performing the best with respect to netgain?*



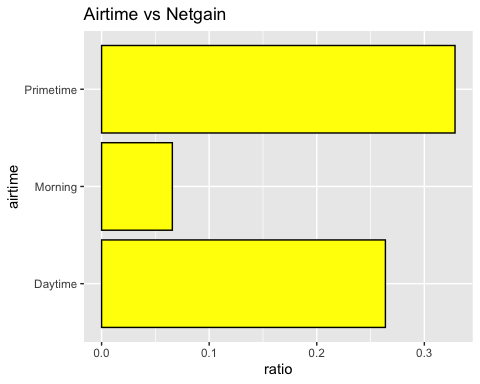
**Inference:** We observe that the best performing industry comes under the ‘Others’ category, followed by the Pharma Industry. It means that people respond well to advertisements promoting different brands of medicine. The Auto industry gets fairly covered on it’s marketing campaign, whereas there is least return on investment on political advertising campaign.

*c) What genre is performing the best with respect to netgain?*



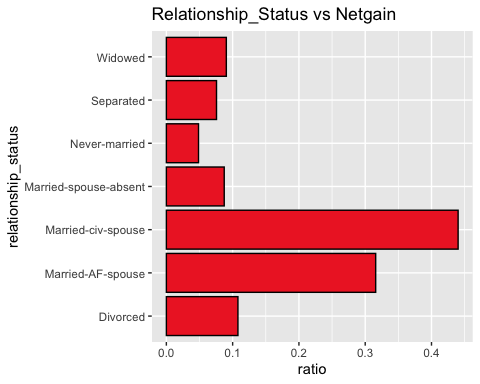
**Inference:** Here we see how genre affects the netgain for their clients. It can be seen that people are attracted towards dramatic and funny advertisements more than those commercials that try to present a product in a direct manner(lacking creativity) or tries to present some information. It is generally observed that viewers tend to spend on a product or an idea when they are emotionally moved by the Ad, and this is clearly demonstrated with the netgain in the genres Drama and Comedy.

*d) What is the best time to air the ad to generate most netgain?*



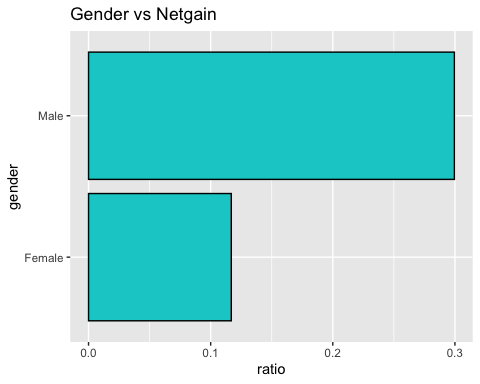
**Inference:** We observe that brands can cover advertising costs if they air the ad during Primetime. This is because the number of people using their TV sets during primetime is the highest. This is further followed by the Daytime viewership, and the ad is the least viewed during Morning hours. This could be possible because people rush for their work in the morning, and usually read the newspaper instead of using their TV sets. Thus the apt time to target their customer base is during Prime time.

*e) What type of audience should be targeted?*



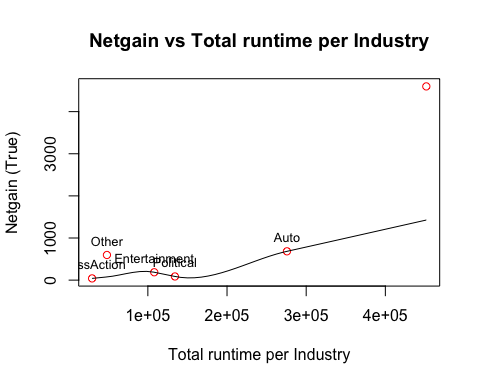
**Inference:** Here we observe that the Married-civ-spouse should be targeted the most in their marketing campaign as they are responding well to the Ads. Next in line who could be targeted are Married-AF-spouse. Rest of the people are resulting approximately an equal amount of netgain to no netgain ratio, thus they need a better incentive to respond to an Ad. For example, the Never-married people are tend to make impulse purchases, thus they could be targeted with better incentives and discounts.

*f) To analyze which gender is responding the best*



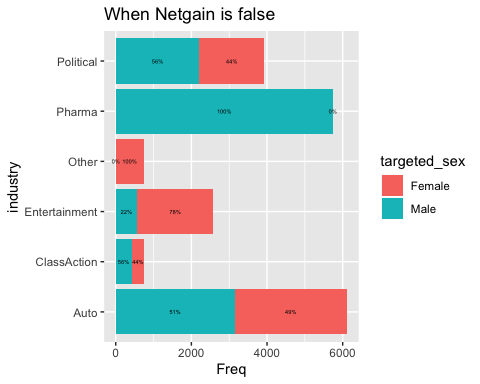
**Inference:** We observe from the above plot that female to male ratio is very low. This means that the probablity of a netgain is higher, when males are spending on the products, which further means that either both the genders are not equally targeted for the Ad campaigns i.e the product is not communicated well to the females, or the product does not possess a good incentive for females.

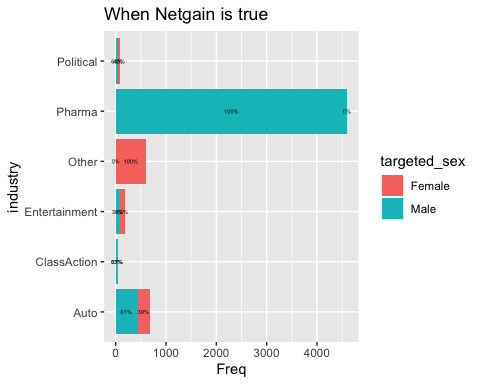
*g) Does runtime affect netgain?*



**Inference:** We observe that as the total runtime per industry is increasing, the netgain is increasing too. This implies that if duration of the commercial is long enough, it tends to leave an impact on the viewers, which has a higher chance of resulting in a netgain. Maximum netgain is demonstrated in the Pharma industry.

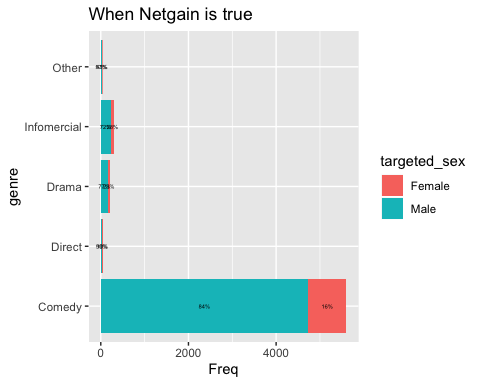
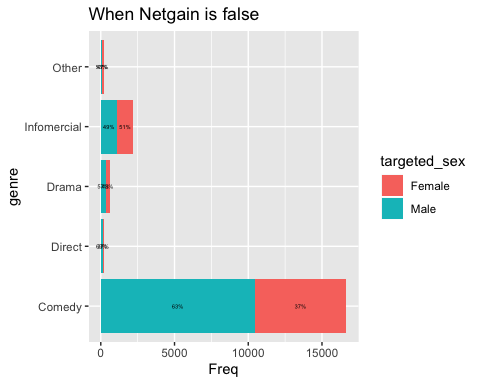
*h) Combined influence of industry and gender on netgain*





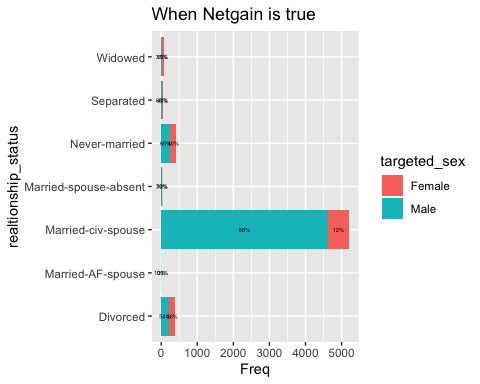
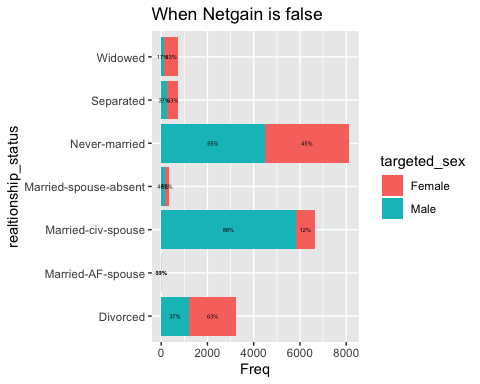
**Inference:** Here we have two plots depicting return on investment with respect to the targeted sex in various industries. We observe a peak in netgain when it comes to the Pharma industry. We see that 44% males out of the targeted population are resulting in a netgain for the Pharma Industry. This means that males are more likely to spend on a medicine based on watching an advertisement about it. We also see that Females don't generate netgain for the Pharma Industry. The second-best performing industry is Auto industry. Out of 682 people who are resulting in a netgain, males have a 61% share in their profit whereas females contribute 39%. Overall, when we observe from both the graphs, we see that Males tend to respond better towards ads than females, except the 'Others' category.

*i) Combined influence of genre and gender on netgain*



**Inference:** Above two plots show that maximum netgain is observed under the genre of comedy. We see that 5603 out of the targeted population are attracted towards Comedy and result in a netgain. Males have an 84% share in the company’s profit whereas females contribute only 16% in the Comedy category. On comparing the above two plots, we can say that males tend to respond more to the advertisements than females, hence they could be targeted to generate profit. Second highest netgain is observed in the 'Infomercial' category and lowest netgain occurs in the 'Direct' category.

*h) Combined influence of relationship status and gender on netgain*



**Inference:** We see that netgain is highest under the Married-civ-spouse category. Among the 5213 Married-civ-spouses, 88% are males. We also observe that never married people do not respond well to advertising campaigns, thus they need better incentives, so that their purchases result in a netgain. Divorced people were targeted too, but only 0.01% respond, among which males demonstrate higher netgain.

Verifying association between categorical variables using chi square test:

*Code and result interpretation for the chi-square test were taken from* [6]

H0: There is no significant association between the variables under study

H1: There is a significant association between the variables under study

chisq.test(data\_train$genre,data\_train$netgain)

##   
## Pearson's Chi-squared test  
##   
## data: data\_train$genre and data\_train$netgain  
## X-squared = 223.26, df = 4, p-value < 2.2e-16

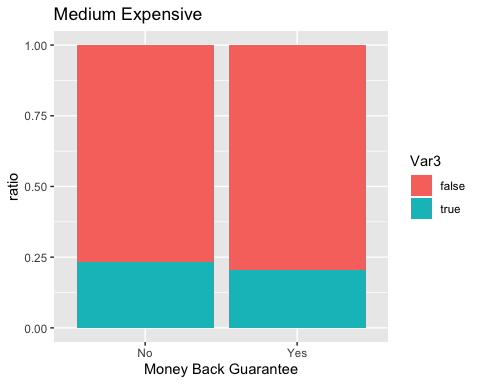
**Inference:** Since the p-value is less than 0.05, we reject H0. Therefore, there is a significant association between genre and netgain.

chisq.test(data\_train$expensive,data\_train$netgain)

##   
## Pearson's Chi-squared test  
##   
## data: data\_train$expensive and data\_train$netgain  
## X-squared = 2.7477, df = 2, p-value = 0.2531

chisq.test(data\_train$money\_back\_guarantee,data\_train$netgain)

##   
## Pearson's Chi-squared test  
##   
## data: data\_train$money\_back\_guarantee and data\_train$netgain  
## X-squared = 0.037506, df = 1, p-value = 0.8464



**Inference:** Since the p-value for both money-back-guarantee and expensive is greater than 0.05, we accept H0. Therefore, there is no significant association between money-back-guarantee and netgain AND expensive and netgain.

This also implies that a highly expensive product that ensures a money back guarantee, or a low-cost product that does not give any money back guarantee, does not make a difference to the overall netgain of the company.

chisq.test(data\_train$airtime,data\_train$netgain)

## Pearson's Chi-squared test  
##   
## data: data\_train$airtime and data\_train$netgain  
## X-squared = 2110.6, df = 2, p-value < 2.2e-16

**Inference:** Since the p-value is less than 0.05, we reject H0. Therefore, there is a significant association between airtime and netgain.

chisq.test(data\_train$targeted\_sex,data\_train$netgain)

## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: data\_train$targeted\_sex and data\_train$netgain  
## X-squared = 1071.9, df = 1, p-value < 2.2e-16

**Inference:** Since the p-value is less than 0.05, we reject H0. Therefore, there is a significant association between targeted sex and netgain.

chisq.test(data\_train$industry,data\_train$netgain)

## Pearson's Chi-squared test  
##   
## data: data\_train$industry and data\_train$netgain  
## X-squared = 5104, df = 5, p-value < 2.2e-16

**Inference:** Since the p-value is less than 0.05, we reject H0. Therefore, there is a significant association between industry and netgain.

3.3 Model Building (Classification)

We now have 4 models and accuracy estimations for each. We need to compare the models with each other and select the most accurate.

*a) Logistic regression*

Logistic Regression is a classification algorithm . It is used to predict a binary outcome (1/0, Yes/No, True/False) given a set of independent variables. To represent binary/categorical outcome, we use dummy variables. [7]

It can be considered as a special case of linear regression when the outcome variable is categorical and it predicts the probability of occurrence of an event by fitting data to a logit function.

sample.ind <- sample(2,nrow(results),replace = T,prob = c(0.6,0.4))  
train <- results[sample.ind==1,]  
test<- results[sample.ind==2,]  
  
outcome = test[, 'netgain']  
test = test[, -13]  
  
logReg = glm(netgain ~ ., data = train, family = binomial)  
pred <- predict(logReg, newdata = test, type = "response")  
pred <- ifelse(pred > 0.5,1,0)  
pred = as.factor(pred)  
outcome = as.factor(outcome) confusionMatrix( data=pred, reference=outcome,positive='1')

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted Class** | |
| False | True |
| **Actual Class** | False | 7564 | 1531 |
| True | 420 | 863 |

**Accuracy**: 81.2%

**Inference:** On applying Logistic Regression for binary classification, we see that 7564 observations were correctly predicted, giving an 81.2% prediction accuracy.

*b) Decision Tree*

A Decision Tree is a supervised learning predictive model that uses a set of binary rules to calculate a target value.

It is used for either classification (categorical target variable) or regression (continuous target variable). Hence, it is also known as CART (Classification & Regression Trees). [8]

Here we have a binary target variable, thus a Classification tree will be generated.

sample.ind <- sample(2,nrow(data\_train),replace = T,prob = c(0.6,0.4))  
train <- data\_train[sample.ind==1,]  
test<- data\_train[sample.ind==2,]  
outcome = test[, 'netgain']  
test = test[, -12]

decisionTree = rpart(netgain ~ ., data=train, method = "class", parms = list(split="gini"), control = rpart.control(cp = 8.042895e-04, maxdepth = 2))  
  
pred <- predict(decisionTree, newdata = test, type = "class")   
library(caret)  
confusionMatrix(data=pred,reference=outcome,positive='true')

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted Class** | |
| False | True |
| **Actual Class** | False | 7856 | 1918 |
| True | 7 | 484 |

**Accuracy**: 81.25%

**Inference:** The Decision tree model gives us 81.25% prediction accuracy which is similar to the logistic regression model built.

*c) Random Forest*

Random forests improve predictive accuracy by generating a large number of bootstrapped trees (based on random samples of variables), classifying a case using each tree in this new “forest”, and deciding a final predicted outcome by combining the results across all of the trees. [9]

dat.d <- sample(1:nrow(data\_train),size=nrow(data\_train)\*0.60,replace = FALSE) #random selection of 60% data.  
   
train <- data\_train[dat.d,] # 60% training data  
test <- data\_train[-dat.d,] # remaining 40% test data  
  
varNames <- colnames(data\_train[,-c(7,11)])   
varNames1 <- paste(varNames, collapse ="+")  
rf.form <- as.formula(paste("netgain", varNames1, sep = "~"))

data\_rf<-randomForest(rf.form, test ,ntree=78,importance=T)  
pred <- predict(data\_rf,test)

confusionMatrix(data=pred,reference=test$netgain,positive='true')

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted Class** | |
| False | True |
| **Actual Class** | False | 7439 | 1184 |
| True | 489 | 1308 |

**Accuracy**: 83.94%

**Inference:** Using the Random Forest model, we can say that 83.94% of the observations were accurately predicted. The Random forest model has been the best model so far.

*d) K-Nearest Neighbor (KNN) Algorithm*

KNN which stand for K-Nearest Neighbor is a Supervised Machine Learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points.

KNN is a lazy algorithm, this means that it memorizes the training data set instead of learning a discriminative function from the training data.

It can be used for solving both classification and regression problems. [10]

dat.d <- sample(1:nrow(data\_train1),size=nrow(data\_train1)\*0.7,replace = FALSE) #random selection of 70% data.  
   
train <- data\_train1[dat.d,] # 70% training data  
test <- data\_train1[-dat.d,] # remaining 30% test data  
  
train\_netgain <- data\_train1[dat.d,10]  
test\_netgain <-data\_train1[-dat.d,10]

sqrt(NROW(train\_netgain))

## [1] 135.0296

knn.135 <- knn(train=train, test=test, cl=train\_netgain, k=135)  
knn.136 <- knn(train=train, test=test, cl=train\_netgain, k=136)

confusionMatrix(table(knn.135 ,test\_netgain))

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted Class** | |
| False | True |
| **Actual Class** | False | 5896 | 581 |
| True | 21 | 1317 |

**Accuracy**: 92.3%

confusionMatrix(table(knn.136 ,test\_netgain))

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted Class** | |
| False | True |
| **Actual Class** | False | 5904 | 584 |
| True | 13 | 1314 |

**Accuracy**: 92.36%

**Inference:** We have 18233 observations in our training data set. The square root of 18233 is around 135.029, therefore we’ll create two models, one with k value as 135 and the other with 136. Both the models (with k-value 135 and 136) give us almost the same accuracy i.e., 92.3%.

We see that the KNN model gives the highest prediction accuracy as compared to other models. This is because KNN is a non-parametric model which means that it does not make any assumptions about the data set. This makes the algorithm more effective since it can handle realistic data.

1. **Conclusion**

From the above analysis and findings, we can conclude the following.

* 1. The industries must try various different methods to increase their demographic reach.
  2. The most effective ads that resulted in maximum netgain had the following findings –

i) ‘Pharma’ and ‘Others’ industry did the best among all the other industries.

ii) Comical and dramatic ads lead to most netgain.

iii) Best time to air the ad is primetime i.e., between 20:00 – 23:00 hours.

iv) Married-CIV-Spouse and Married-AF-Spouse have the maximum response towards netgain.

v) Targeting males resulted in more netgain than females. This can imply two things

* The female audience weren’t targeted effectively.
* The female audience didn’t respond well i.e., the ads weren’t female centric. (Only the ‘Others’ industry did well with respect to netgain and targeting females)

vi) Airing your ad for a greater number of minutes per week will result in high netgain given the ad satisfies the above observations

vii) More males can be targeted for the pharma industry since maximum netgain is seen from the males.

c) We performed hyper parameter tuning to improve the accuracy of models. And amongst all the models, KNN did the best in predicting whether an ad will result in netgain or not.

# References

|  |  |
| --- | --- |
| [1] | L. X. G. Huini Zhou, "Evaluating and Analyzing the effectiveness of Online Advertising," *IEEE,* 2018. |
| [2] | S. S. S. M. V. S. Qi Lou, "Content-based Effectiveness Prediction of Video Advertisements," *IEEE,* 2018. |
| [3] | T. K. Carolyn A Lin, "Predicting user response to sponsored advertising on social media via the technology acceptance model," *Elsevier,* 2016. |
| [4] | "https://www.hackerearth.com/challenges/competitive/predict-ad-success-hackerearth-machine-learning-challenge/," [Online]. |
| [5] | "https://www.r-graph-gallery.com/," [Online]. |
| [6] | "http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r," [Online]. |
| [7] | "https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc," [Online]. |
| [8] | "https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/," [Online]. |
| [9] | "https://towardsdatascience.com/understanding-random-forest-58381e0602d2," [Online]. |
| [10] | "https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn-1f2c88da405c," [Online]. |