Determining the Popularity of Game's Based on Various Factors and Showing Relevant Card to Customers

1st Shirish Sonvane
Student of Masters in Data Analytics
National College of Ireland
Dublin, Ireland
x19164165@student.ncirl.ie

***Abstract*—Objective of this paper is to implement at least 5 different machine learning methods on 3 datasets. The two of them are from Entertaining and Gaming commerce and rest is from the Tesco marketing where for the first two datasets, factors that affect the popularity have been studied, and for the Tesco dataset, what kind of shopping affects to the card that is offered to the customer has been studied. All three datasets are being gathered from the Kaggle named as Steam Games [1], Board Games [2], and Tesco marketing content [3]. On the dataset of Steam Games, C5.0 Decision tree and RIPPER algorithm have been applied. On the second dataset, the M5-Prime model and Multiple Linear regression have been applied to measure the esteem of them considering owners of the game and average rating of the games as dependent variables respectively. And for the last dataset, which card to give to customers predicted, where content1 is being a dependent variable. Along with implementing these machine learning methods on these datasets, a comparison between these methods has also been taken into consideration except for the Tesco dataset. The M5 model should have been performed well compared to linear regression as per Lantz B [4]. however both model have performed almost equal. For the second dataset, the tree has 120 branches where RIPPER has only 6 rules for the same set of variables. The KDD methodology has been used in this project to deal with data.**

***Index Terms*—Machine Learning, Prediction, Response Variable, Linear Regression, Decision Tree, M5 Prime, RIPPER Algorithm, Logistic Regression.**

I. Introduction

The data is a crucial term in this era. It contains a large number of different records in it. We have a large data around us and it is increasing very rapidly day by day. As the data is growing technologies are also taking different modes by and day. These technologies can help wrangle this data and look if there is any pattern between the data. This throughout the world known by data mining and machine learning.

In this paper, we have taken some data from the Gaming and Marketing commerce. The Gaming section from where Steam Games and Board games are acquired. Tesco marketing content dataset has been acquired from Marketing. The purpose behind selecting these domains is to see what the factors that people see before buying the games and factors on which card is being provided. The below questions pops on the mind to which we get the answers by studying data.

"Which attribute has more significance while predicting the popularity? How significant these factors would be? What factors are being evaluated while providing cards to customers?"

Answering the mentioned questions is one part of this paper. Another part of this creating a different model using five machine learning methods and study the factors. This can be achieved by implementing two methods on two datasets and one method on the fifth one.

The Regression model and M5 Prime model are being implemented on the steam game dataset to predict the average rating of the games. Similarly, the RIPPER algorithm and C5.0 Decision tree are being implemented on the board games to distinguish the number of people who have bought the game into four various classes. These two are being know as the classification model. Lastly, whether to give the card to customers or nor is being predicted on Tesco dataset using logistic regression.

The KDD methodology is being used to get the desired outputs. This is a machine learning methodology used on the dataset to get the results by measuring correctness and performance. This methodology involves collecting the data to transform the same so that it should meet the criteria of the model which we are going to apply. Once the data is ready, we apply the algorithms and interpret the model and understand the results.

First of all, related work is being discussed in the next section that has been done not just on these datasets but also on the domain that have performed earlier and try to seek answers. This may help us to understand whether our results are the same or nearby by comparing wherever possible.

Second section is data mining methodology, in that section we will look into KDD. The main part of KDD involves data preparation, selection, data cleaning, adding prior knowledge to data, and interpret the correct results. Extracting meaningful information from the data is the main objective of the KDD.This is achieved by implementing various machine learning and data mining methods. KDD consists of multidisciplinary events. The KDD process has been reached to its top in recent few years.

II. Related Work

In this busy life game and entertainment plays a key role to relax our mind. Relaxation leads to better performance in the work. Media psychological research has found that a great amount of happiness comes from playing games and fills their enjoyable experience [5]. It can be inferred that people purchasing the games and ratings of those games are directly proportional [5]. One more study shows that one can purchase the game if that provides you entertainment as well as any kind of knowledge [6]. An example of this where the game can provide pleasure and some meaningful insights is board game and strategy games [6]. This study concludes that the public does go for video games however more attracted to those games that provide some knowledge as well.

Nowadays, game developing companies are facing various problems due to sudden changes in the demand for games. People are playing various kinds of games every day which causes them problem to focus on developing one game [7]. The game having any kind of achievement or reward are most satisfactory. Thus, it can be said that if that game has any aim to achieve in it then that game is very famous within the folks [7].

There is a new type of genre that has been come to the market called a casual genre. These type of games does nothing but enjoyment and number of these games has increased so much in the market [8]. Most of them have an Application buy option in it. In research has found that some financial models gives an option to pay and forward into the games. The control of anyone's in spending the money to buy this also influence the pleasure and understanding [8]. Hence application buy can help in determining fame of any game.

Gaming technologies have increased to such level that gaming servers are providing the feature to users to create their micro games, customize the same as per their need and play the same [9]. This shows that users have more curiosity to play the games that customization ability.

In this era, it is very tough for the supermarkets to provide customized offers to the users. Providing the users list of shopping based on his/her present requirements [10]. However, present approaches are not up to the mark to take into consideration other different factors that are affecting the user's thought process.

A. Steam Games

The game industry which has an online platform has billions of users across the world with an option of sharing games and their experience. The steam platform is one of the such online platforms which acts as a common platform for the game enthusiast [11]. Gathering data from the steam platform is quite simple. Extracting personalized data is better to begin the analysis for any association.

Different attributes present in the games play the key in selling of any famous game. Many features have a positive effect on selling video games and the newness of the game is one of them [12]. The overall rating, the price, and the mode of the game in terms of whether it is a single or multiple user game affects the fame of games [12]. It can be inferred that a single or multiple player game has some sort of significance on the popularity of the game.

The game features like challenge, tales, fantasy, nostalgia are also significant for one when trying to play the game. Multiplayer games change the way of looking towards it [13]. Mostly in such conditions our view changes to more positive towards those games. This makes one more excited to play these kinds of games where you to play as a team to win the same [13].

B. Board Games

These games give the platform to individuals to increase their problem-solving skills. This problem-solving skill ultimately helps in improving one's perceptive capability [14]. This capability involves one's information processing, awareness, and consideration which leads to the improvement in remembering things [14]. It can say that more people are attracted to this game since it involves enjoyment as well as knowledge.

These kinds of offline games always need to be played in a team of folks. Playing games in the group helps each other to increase social interaction which sometimes turns out to be educational interaction which is crucial for kids [15]. Learning along with playing is a better way to educate children and this ultimately grows their social aids [15]. These games are widely famous within the kids and parents with small children.

These games are just not limited to children. Even adults can play this game. Just like children, adults can also be learned and educate themselves using these games [16]. Along with the enjoyment, these games help to increase mental health and social interaction. It can be inferred that these games are not only famous within children but also in adults since they give facts and some meaningful information [16].

Board games can be helpful in the treatment of psychological diseases such as Alzheimers [17]. The doctor often asks these kinds of patients to play board games as part of the treatment. Thus, it can be inferred that just like kids and grown-ups this game is also famous in the doctors as well.

Every game has some kind of minimum and maximum playtime. Also, the minimum and maximum number of players are much significant towards the popularity of the game [18].

C. Tesco Marketing Content

The human race moving from offline to online rapidly day by day. By providing the cards to users it can be achieved. However, to do so, we need to study what factors need to analyzed so that company can offer the card. The data can be gathered from the organization and divided into various factors. Logistic regression is being used to decide whether to give the card to customers or not. Tesco can use the AI to decide this based on the data they have [19].

Logistic regression gives us the answer either in yes or no format and it is one of the best methods to determine whether the cardholder is in default or not. This analyzes various factors present in the dataset before giving any answer to the question [20]. Target variable has only two factors. Rest variables do not need to have in a categorical format. To be specify did not find much related work for this dataset.

III. Data Mining Methodology

The Knowledge Discovery in a database is the procedure where we select the data, remove unnecessary records from it, and make some enhancements to it. The next step is transformation, in that step we analyze the attributes and extract some meaningful information. Data mining presents a crucial part of KDD since it involves applying machine learning and statistical methods to retrieve patterns, relation, and rules. Interpretation used to check if there is any sense present in the detected pattern [21].

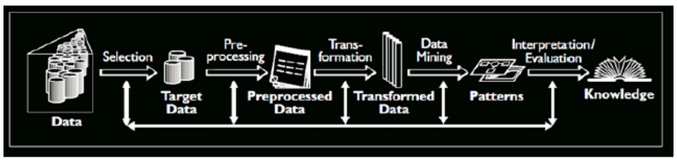
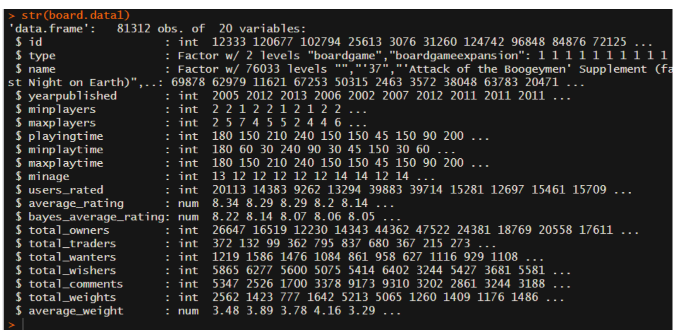


Fig. 1. KDD Methodology

A. Board Games

First of all, we take a look at our raw dataset for the board games data so decide what transformation needs to be done.

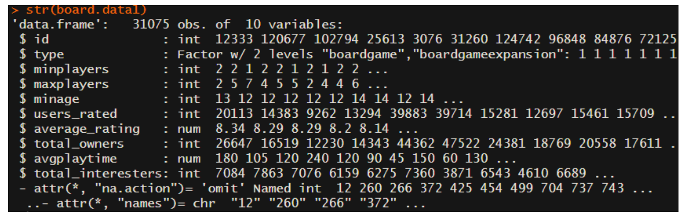


```
> str(board.data1)
'data.frame':	81312 obs. of  20 variables:
 $ id                  : int  12333 120677 102794 25613 3076 31260 124742 96848 84876 72125 ...
 $ type                : Factor w/ 2 levels "boardgame","boardgameexpansion": 1 1 1 1 1 1 1 1 1 1
 $ name                : Factor w/ 76033 levels "","'37","'Attack of the Boogeymen' Supplement (fa
st Night on Earth)",..: 69878 62979 11621 67253 50315 2463 3572 38048 63783 20471 ...
 $ yearpublished       : int  2005 2012 2013 2006 2002 2007 2012 2011 2011 2011 ...
 $ minplayers          : int  2 2 1 2 2 1 2 1 2 2 ...
 $ maxplayers          : int  2 5 7 4 5 5 2 4 4 6 ...
 $ playingtime         : int  180 150 210 240 150 150 45 150 90 200 ...
 $ minplaytime         : int  180 60 30 240 90 30 45 150 30 60 ...
 $ maxplaytime         : int  180 150 210 240 150 150 45 150 90 200 ...
 $ minage              : int  13 12 12 12 12 12 14 14 12 14 ...
 $ users_rated         : int  20113 14383 9262 13294 39883 39714 15281 12697 15461 15709 ...
 $ average_rating      : num  8.34 8.29 8.29 8.2 8.14 ...
 $ bayes_average_rating: num  8.22 8.14 8.07 8.06 8.05 ...
 $ total_owners        : int  26647 16519 12230 14343 44362 47522 24381 18769 20558 17611 ...
 $ total_traders       : int  372 132 99 362 795 837 680 367 215 273 ...
 $ total_wanters       : int  1219 1586 1476 1084 861 958 627 1116 929 1108 ...
 $ total_wishers       : int  5865 6277 5600 5075 5414 6402 3244 5427 3681 5581 ...
 $ total_comments      : int  5347 2526 1700 3378 9173 9310 3202 2861 3244 3188 ...
 $ total_weights       : int  2562 1423 777 1642 5213 5065 1260 1409 1176 1486 ...
 $ average_weight      : num  3.48 3.89 3.78 4.16 3.29 ...
>
```

Fig. 2. Board Game data

Since the multiple regression and M5 prime both regression model, they come under the same family of the linear regression. Hence the data same steps have been followed while cleaning and transforming the board data. Some less important attributes have been removed from the dataset to reduce the size. The name of the game, the year of the game published, playtime, Bayes average rating, total traders, comments, total and average weights. These factors do not hold much significance while predicting the popularity of the game hence removed from the dataset.

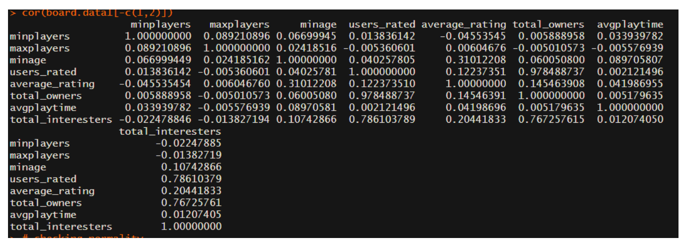
As mentioned above in related work, average playtime is crucial in deciding the popularity of the games. Hence minimum and maximum playtime transformed in average playtime by taking the average of their sum. Similarly, the attributes total wishers and total wanters merged to form one meaningful attribute and named as total interest. There were 15 NA values were present in the dataset thus decided to remove them. There were also a few zero's that causing problems to our assumptions hence removed them as well. The final dataset has the 10 attributes with records 31075 as shown in the figure below. Before applying the regression model to the dataset, we need to check some assumptions. Multicollinearity, Normality, homoscedasticity, linearity are few of them



```
> str(board.data1)
'data.frame':	31075 obs. of  10 variables:
 $ id               : int  12333 120677 102794 25613 3076 31260 124742 96848 84876 72125
 $ type             : Factor w/ 2 levels "boardgame","boardgameexpansion": 1 1 1 1 1 1 1
 $ minplayers       : int  2 2 1 2 2 1 2 1 2 2 ...
 $ maxplayers       : int  2 5 7 4 5 5 2 4 4 6 ...
 $ minage           : int  13 12 12 12 12 12 14 14 12 14 ...
 $ users_rated      : int  20113 14383 9262 13294 39883 39714 15281 12697 15461 15709 ..
 $ average_rating   : num  8.34 8.29 8.29 8.2 8.14 ...
 $ total_owners     : int  26647 16519 12230 14343 44362 47522 24381 18769 20558 17611 .
 $ avgplaytime      : num  180 105 120 240 120 90 45 150 60 130 ...
 $ total_interesters: int  7084 7863 7076 6159 6275 7360 3871 6543 4610 6689 ...
 - attr(*, "na.action")= 'omit' Named int  12 260 266 372 425 454 499 704 737 743 ...
  ..- attr(*, "names")= chr  "12" "260" "266" "372" ...
```

Fig. 3. Board Game Final dataset

1) Multicollinearity: While studying the correlation between attributes, it found that the attribute users rating has a high correlation with the total number of owners and the total number of interested people hence that attribute has removed from the model as shown in fig below.



```
> cor(board.data1[-c(1,2)])
                    minplayers   maxplayers     minage  users_rated average_rating total_owners  avgplaytime
minplayers         1.000000000  0.089210896 0.06699945  0.013836142    -0.04553545  0.005888958  0.033939782
maxplayers         0.089210896  1.000000000 0.02418516 -0.005360601     0.00604676 -0.005010573 -0.005576939
minage             0.066999449  0.024185162 1.00000000  0.040257805     0.31012208  0.060050800  0.089705807
users_rated        0.013836142 -0.005360601 0.04025781  1.000000000     0.12237351  0.978488737  0.002121496
average_rating    -0.045535454  0.006046760 0.31012208  0.122373510     1.00000000  0.145463908  0.041986955
total_owners       0.005888958 -0.005010573 0.06005080  0.978488737     0.14546391  1.000000000  0.005179635
avgplaytime        0.033939782 -0.005576939 0.08970581  0.002121496     0.04198696  0.005179635  1.000000000
total_interesters -0.022478846 -0.013827194 0.10742866  0.786103789     0.20441833  0.767257615  0.012074050
                  total_interesters
minplayers              -0.02247885
maxplayers              -0.01382719
minage                   0.10742866
users_rated              0.78610379
average_rating           0.20441833
total_owners             0.76725761
avgplaytime              0.01207405
total_interesters        1.00000000
```

Fig. 4. Correlation

2) Normality: The dependent variable has to be distributed normally across the independent variable is one of the assumptions. From the figure 5 below, it can be inferred that our dependent variable has normally distributed.

3) Quantity of data: This assumption is also quite significant. The size of data has to be greater 50 than 8 times the quantity of predictor variables [22]. This dataset has records of 31073 which definitely fulfills this assumption.

4) Outliers: Outliers have also been checked and found that this dataset has many outliers. We have used winsorizing method to remove the outliers from the data. With the help Inter-Quartile Range we calculted the upperlimit of the attributes and replaced the same with calculated upper limit.

Assumptions like normality of residual, linearity, and homoscedasticity can be verified only after the model is being built. Hence this can be discussed while evaluating the model in the next section. Before applying the model, the dataset has been split into train and test in the ratio of 70:30. After this multiple regression and M5 prime models were implemented on the dataset considering the most significant independent variables.

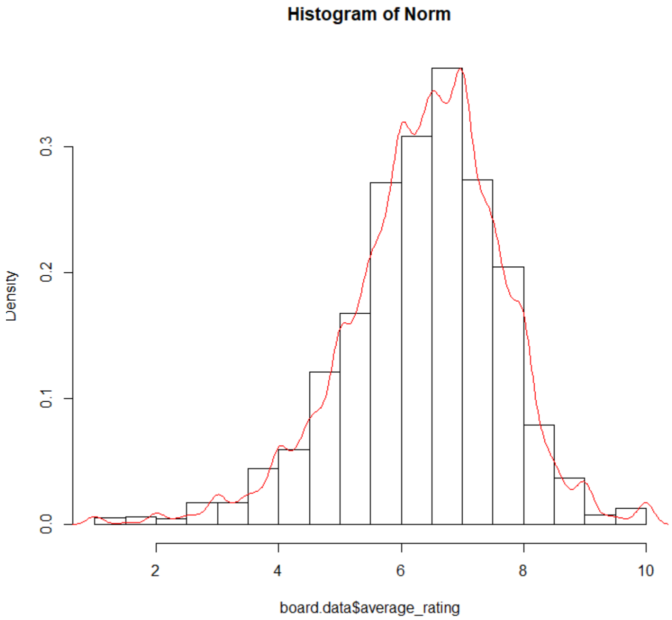


Fig. 5. Normality

B. Steam Games

C5. 0 Decision tree and RIPPER algorithm are both belong to the classification model hence we had to perform data transformations and cleaning once for all this dataset. To begin with we checked all the attributes of the datasets so that we can make the necessary changes or remove unwanted from the sample.

```
> steam.data <- read.csv("D:/PG/Data Mining/Projects/Final Report/steam.csv")
> str(steam.data)
'data.frame':	27075 obs. of  18 variables:
 $ appid           : int  10 20 30 40 50 60 70 80 130 220 ...
 $ name            : Factor w/ 27033 levels "'1st Core: The Zombie Killing Cyborg'",..: 5096 22366 5874 6073 10412 1
8566 10404 5098 10411 10405 ...
 $ release_date    : Factor w/ 2619 levels "1997-06-30","1998-11-08",..: 5 3 10 7 4 5 2 13 7 17 ...
 $ english         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ developer       : Factor w/ 17113 levels "'What Day is it?' Games",..: 15909 15909 15909 15909 6055 15909 15909 1
5909 6055 15909 ...
 $ publisher       : Factor w/ 14354 levels "-","--","-Yodasaurus- Games",..: 13295 13295 13295 13295 13295 13295 13
295 13295 13295 13295 ...
 $ platforms       : Factor w/ 7 levels "linux","mac",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ required_age    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ categories      : Factor w/ 3333 levels "Captions available",..: 330 330 498 330 2487 457 2292 2487 624 3062 ...
 $ genres          : Factor w/ 1552 levels "Accounting;Animation & Modeling;Audio Production;Design & Illustration;E
ducation;Photo Editing;Software Trainin"| __truncated__,..: 3 3 3 3 3 3 3 3 3 3 ...
 $ steamspy_tags   : Factor w/ 6423 levels "1980s;Great Soundtrack;Retro",..: 225 225 2583 225 2554 225 2562 225 255
4 2554 ...
 $ achievements    : int  0 0 0 0 0 0 0 0 0 33 ...
 $ positive_ratings: int  124534 3318 3416 1273 5250 2758 27755 12120 3822 67902 ...
 $ negative_ratings: int  3339 633 398 267 288 684 1100 1439 420 2419 ...
 $ average_playtime: int  17612 277 187 258 624 175 1300 427 361 691 ...
 $ median_playtime : int  317 62 34 184 415 10 83 43 205 402 ...
 $ owners          : Factor w/ 13 levels "0-20000","100000-200000",..: 4 12 12 12 12 12 12 4 12 4 ...
 $ price           : num  7.19 3.99 3.99 3.99 3.99 3.99 7.19 7.19 3.99 7.19 ...
> 
```

Fig. 6. Steam Games Dataset

First of all, we removed unwanted columns like app id, name of the apps, release date of app, developers and publishers of the games, genres of the games, tags and median playtime of the games. These attributes do not hold much of the significance to the sale of the game. After this we converted some variables into factors, for example, the english attribute was in the integer form which we transformed in factor like 'yes' and 'no'. After that required age column also renamed and converted to factor as shown in the figure 7. In order, check the popularity we are going to use the number of the owners as the dependent variable, however, there were already many levels that were present in that attributes, so we had to decrease the same to make some sense out of it. We reduced it to four levels as shown in figure 8. Similarly, we transferred the attribute platform factor while removing special symbol in it.

```
## Column english converted into 'Yes' and 'No' ##
steam.data$english <- factor(steam.data$english, levels = c(1,0), labels = c("Yes", "No"))
##  Column required_age converted into different ages ##
steam.data$required_age <- factor(steam.data$required_age, levels = c(0,3,7,12,16,18),
                                  labels = c("No Age Limit", "3+", "7+", "12+", "16+", "18+"))
```

Fig. 7. Data factorization

```
owner <- function(x){
  print(x)
  if(x == '0-20000'){
    return('<20K')
  }
  else if ((x == '0-20000') | (x == '100000-200000') | (x == '20000-50000') | (x == '50000-100000') |(x== '200000-500000')){
    return('20K to 500K')
  }
  else if ((x == '500000-1000000') | (x=='5000000-10000000') | (x == '2000000-5000000') | (x == '1000000-2000000')){
    return('500K to 10M')
  }
  else
    return('10M to 200M')
}

steam.data$owners <- factor(sapply(steam.data$owners, function(x) owner(x)))
```

Fig. 8. Levels of Owner

```
> levels(df_decision$english)
[1] "Yes" "No"
> levels(df_decision$owners)
[1] "<20K"        "10M to 200M" "20K to 500K" "500K to 10M"
> levels(df_decision$platforms)
[1] "linux"             "mac"               "mac linux"         "windows"           "windows linux"
[6] "windows mac"       "windows mac linux"
> levels(df_decision$required_age)
[1] "No Age Limit" "3+"           "7+"           "12+"          "16+"          "18+"
> 
```

Fig. 9. Levels of Columns

We discussed in the above section that whether the provided game is a single or multiplayer has the utmost significance while determining the popularity. Hence we converted this column into three levels as the single-player, multiplayer, and both. The grep function has been used to pick the necessary text from the given string as shown in the figure 10.

```
## Column categories- It consist multiple categories, however, we need only three
categories <- function(x){
  if ((length(grep('Multi-player', x))>0) && (length(grep('Single-player', x))>0)){
    var = 'Both'
  }
  else if (length(grep('Multi-player',x))>0){
    var = 'Multi-player'
  }
  else if (length(grep('Single-player',x))>0){
    var = 'Single-player'
  }
  else {
    var = 'Not Mentioned'
  }
  return(var)
}
steam.data$categories <- factor(sapply(steam.data$categories, function(x) categories(x)))
```

Fig. 10. Levels of Player

The below figure 11 shows the final dataset after all cleaning and transformation. Again this dataset was split into train and test to perform analysis on it in the ratio of 7:3. Once the split was done, we had verified the same using the dependent variable. We found that the split was equal for both test and train data in relations of proportion as shown in the figure 12. After this model was built separately on the train data and tested using the test data.

C. Tesco Marketing Dataset

Binomial logistic regression also belongs to the regression family, however, the difference here is that the dependent

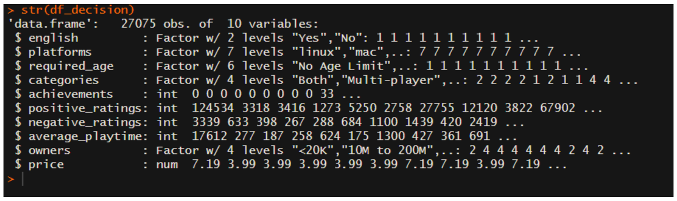


Fig. 11. Dataset for Classification Model

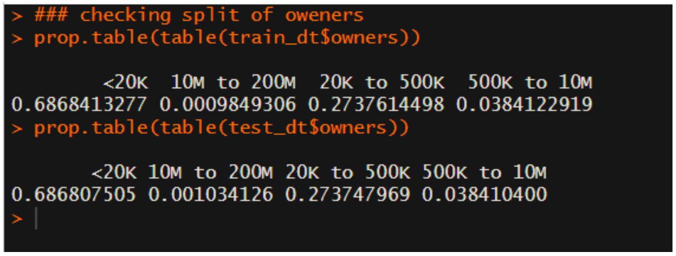


Fig. 12. Distribution of Owners

variable is of categorical type. On this dataset we are applying only one algorithm. We first look at the structure of the dataset so that we necessary steps can be taken to retrieve some sense out of data.

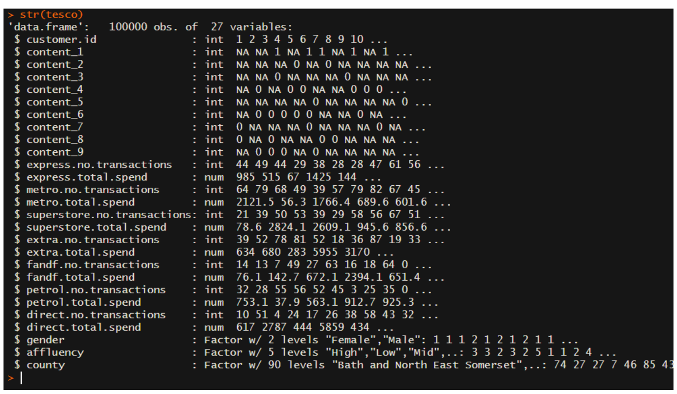


Fig. 13. Tesco Dataset

After this we removed the unwanted columns from the dataset like content2, content3 till content9. These attributes do not hold any significance in predicting if the card is being offered or not. The dependent variable has two levels as 1,0 and NA. 1 means the user has clicked the card, 0 means the user did not click on the card, and NA means the card was never shown to the customer. We needed all factors but the algorithm can not interpret NA as values instead it counts as missing value. Hence we converted the NA to 0 to make it machine-readable.

We will now look into the assumptions. Logistic regression anyway do not much assumptions. Multicollinearity and outliers are the only two assumptions we need to check for this method.

1) Multicollinearity: The figure shows that attributes do not have much correlation between them which satisfies our assumption.

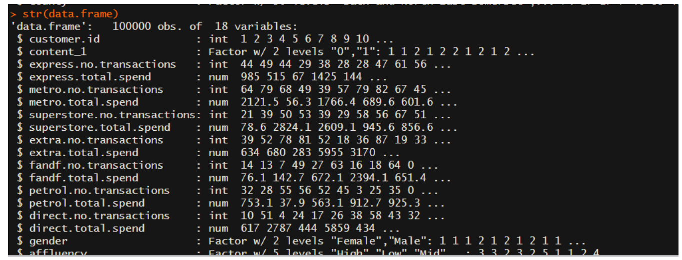


Fig. 14. Final Tesco Dataset

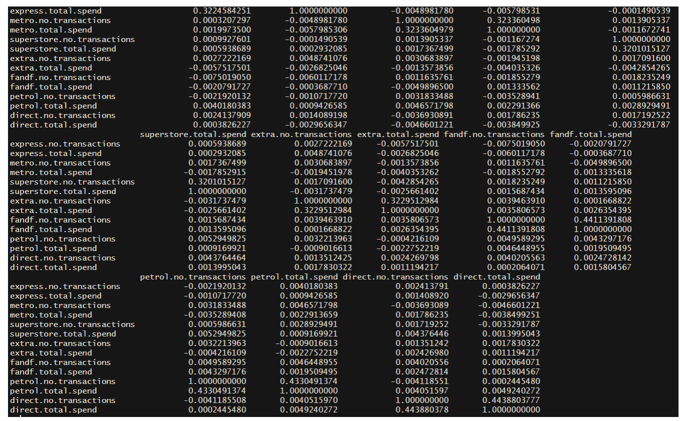


Fig. 15. Multicollinearity

2) Outliers: We have handled the outliers using the winsorizing method as mentioned in the board game section.

The dependent variable also converted into a factor in order, to apply logistic regression. The final dataset with 10000 records after removing some attributes looks like as shown in fig 14.

Once the dataset got ready, we split the dataset again into train and test as 7:3. We applied the model on the training dataset and predicted the results using the test dataset. We found that many of the independent variables do not hold any kind of significance. We have also used k-fold as the sampling method where the value of k has put as 10. We compared both models and did not find much difference between them.

IV. Evaluation of Methods

Two models on two datasets and one is on the rest dataset. In this we look into the results and try to interpret the same.

A. Board Games

From the summary of the multiple regression, we can see that all the independent variables are significant. The performance of this module measured in terms of R2 and the value for the same is 20 percent which not that good. So we can say that this is not the best fit of the model.

From the figure 17 of the summary of the regression model, we can interpret the homoscedasticity and normality of residuals. From this plot, we can see that residuals and

```
> summary(mlm)

Call:
lm(formula = average_rating ~ . - id - type - users_rated, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9207 -0.6067  0.0099  0.6452  4.9161 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       5.166e+00  3.479e-02 148.525  < 2e-16 ***
minplayers       -8.243e-02  1.070e-02  -7.704 1.37e-14 ***
maxplayers       -1.728e-02  3.455e-03  -5.003 5.69e-07 ***
minage            1.081e-01  3.014e-03  35.856  < 2e-16 ***
total_owners     -7.388e-04  4.135e-05 -17.866  < 2e-16 ***
avgplaytime       1.760e-03  1.643e-04  10.713  < 2e-16 ***
total_interesters 1.049e-02  2.378e-04  44.128  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.126 on 24174 degrees of freedom
Multiple R-squared:  0.2058,	Adjusted R-squared:  0.2056 
F-statistic:  1044 on 6 and 24174 DF,  p-value: < 2.2e-16
```

Fig. 16. Multiple Regression Model Summary

actual values are normally distributed which is known as homoscedasticity. The first graph in the plot shows that the data is linearly distributed. From the fourth graph of the plot it can said that there are no outliers in the dataset.

Summary of the M5-Prime model doesn't display the significance between the response and predictor variable. However, we can measure the root mean square error for the same and it is 1.3 %. Since both are the regression models we can compare the performance between them using correlation and mean absolute error.

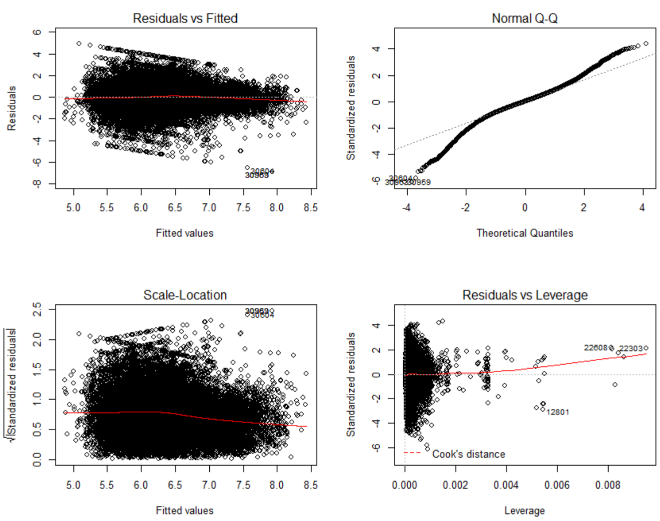


Fig. 17. Assumptions for Regression Model

```
> # checking correlation between predicted and actual values
> cor(pred, test$average_rating)
[1] 0.2633304
> # MAE mean absolute error between actual values and predicted values
> MAE(pred, test$average_rating)
[1] 0.9765997
```

Fig. 18. COR and MAE for Multiple Regression

From the correlation and mean absolute error it can be interpreted that multiple regression model has correlation of 26% and MAE is 0.97 and for the M5-Prime model correlation is 23% percent and surprisingly this model also has the same value of MAE that is 0.97. Based on this results we can say that both the models have performed equally same can be used to predict the average rating of the games. However, we still need to make some changes so that we can increase the performance of the multiple linear regression model.

```
> # checking correlation between predicted and actual values
> cor(pred_m5, test$average_rating)
[1] 0.238663
> # MAE mean absolute error between actual values and predicted values
> MAE(test$average_rating, pred_m5)
[1] 0.9763553
```

Fig. 19. COR and MAE for M5 Prime

B. *Steam Games*

From the summary of the decision tree, we can say that tree shows that the error of 9.7 % while classifying the records. The main attributes for this classifications are negative rating, average playtime, and positive ratings. This tree has created the 120 branches in total. Most of the games have less than 20K owners as shown in the fig 20.

```
Evaluation on training data (20306 cases):

	    Decision Tree   
	  ----------------  
	  Size      Errors  

	   120 1966( 9.7%)   <<


	   (a)   (b)   (c)   (d)    <-classified as
	  ----  ----  ----  ----
	 13335           612          (a): class <20K
	          10            10    (b): class 10M to 200M
	  1054          4418    87    (c): class 20K to 500K
	                 203   577    (d): class 500K to 10M


	Attribute usage:

	100.00% negative_ratings
	 92.17% average_playtime
	 49.75% positive_ratings
	 34.99% price
	 13.22% required_age
	 11.44% categories
	  5.44% english
	  4.51% achievements
	  2.76% platforms


Time: 0.1 secs
```

Fig. 20. Decision Tree Summary

However, the RIPPER algorithm doesn't display significance about the independent variable. Although, it gives the correct rate of classification and which is 68%. From the summary of this method, it can be said that this tree has generated 6 rules and these rules are based on age. These rules are nothing compared to a decision tree. The p-value of this model is very less as we see from the figure. To get more answers in details we need to look at the confusion matrix of both model.

From the confusion matrix, we can see that the accuracy of the decision tree model is 89% where the RIPPER algorithm

```
Call:
OneR.formula(formula = owners ~ ., data = train_dt)

Rules:
If required_age = No Age Limit then owners = <20K
If required_age = 3+           then owners = <20K
If required_age = 7+           then owners = <20K
If required_age = 12+          then owners = <20K
If required_age = 16+          then owners = 20K to 500K
If required_age = 18+          then owners = <20K

Accuracy:
13961 of 20306 instances classified correctly (68.75%)
```

Fig. 21. RIPPER Tree Summary

```
> # confusion matrix
> confusionMatrix(steam_pred, test_dt$owners)
Confusion Matrix and Statistics

             Reference
Prediction    <20K 10M to 200M 20K to 500K 500K to 10M
  <20K        4440           0         362           1
  10M to 200M    0           2           0           0
  20K to 500K  209           0        1444          98
  500K to 10M    0           5          47         161

Overall Statistics

               Accuracy : 0.8933
                 95% CI : (0.8857, 0.9006)
    No Information Rate : 0.6868
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7579

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: <20K Class: 10M to 200M Class: 20K to 500K Class: 500K to 10M
Sensitivity               0.9550          0.2857143             0.7793            0.61923
Specificity               0.8288          1.0000000             0.9376            0.99201
Pos Pred Value            0.9244          1.0000000             0.8247            0.75587
Neg Pred Value            0.8937          0.9992611             0.9185            0.98490
Prevalence                0.6868          0.0010341             0.2737            0.03841
Detection Rate            0.6559          0.0002955             0.2133            0.02378
Detection Prevalence      0.7096          0.0002955             0.2587            0.03147
Balanced Accuracy         0.8919          0.6428571             0.8584            0.80562
```

Fig. 22. Decision Tree Confusion Matrix

```
> confusionMatrix(ripper_pred, test_dt$owners)
Confusion Matrix and Statistics

             Reference
Prediction    <20K 10M to 200M 20K to 500K 500K to 10M
  <20K        4632           5        1836         241
  10M to 200M    0           0           0           0
  20K to 500K   17           2          17          19
  500K to 10M    0           0           0           0

Overall Statistics

               Accuracy : 0.6868
                 95% CI : (0.6756, 0.6978)
    No Information Rate : 0.6868
    P-Value [Acc > NIR] : 0.5059

                  Kappa : 0.0106

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: <20K Class: 10M to 200M Class: 20K to 500K Class: 500K to 10M
Sensitivity              0.99634           0.000000           0.009174            0.00000
Specificity              0.01792           1.000000           0.992270            1.00000
Pos Pred Value           0.68990                NaN           0.309091                NaN
Neg Pred Value           0.69091           0.998966           0.726542            0.96159
Prevalence               0.68681           0.001034           0.273748            0.03841
Detection Rate           0.68430           0.000000           0.002511            0.00000
Detection Prevalence     0.99187           0.000000           0.008125            0.00000
Balanced Accuracy        0.50713           0.500000           0.500722            0.50000
```

Fig. 23. RIPPER Tree Confusion Tree

has an accuracy of 68%. It can be said that the decision tree has performed better compared to the RIPPER algorithm. Other statistics like kappa can also be used to interpret the accuracy of the model. These values for decision and RIPPER algorithm are 0.75 and 0.01 which states that the decision tree has outperformed the RIPPER algorithm.

The area under the curve is one more statistic that is being used to compare the accuracy of the models. The AUC for the decision tree is 89% and for the RIPPER algorithm it is 53% from this as well we can say that the performance of the decision tree is outstanding compared to the RIPPER algorithm a shown in the figure.

```
> c5_roc

Call:
multiclass.roc.default(response = test_dt$owners, predictor = steam_pred_prob)

Data: multivariate predictor steam_pred_prob with 4 levels of test_dt$owners: <20K, 10M to 200M, 20K to
0K, 500K to 10M.
Multi-class area under the curve: 0.8983
> ripp_roc

Call:
multiclass.roc.default(response = test_dt$owners, predictor = ripper_pred_prob)

Data: multivariate predictor ripper_pred_prob with 4 levels of test_dt$owners: <20K, 10M to 200M, 20K t
00K, 500K to 10M.
Multi-class area under the curve: 0.5395
>
```

Fig. 24. AUC for Decision and RIPPER

C. Tesco Marketing Dataset

```
> summary(logistic_model)

Call:
glm(formula = content_1 ~ . - customer.id, family = "binomial",
    data = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.169  -1.111  -1.089   1.245   1.302

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.557e-02  4.896e-02  -1.748   0.0805 .
express.no.transactions  -2.040e-04  4.253e-04  -0.480   0.6315
express.total.spend      -1.007e-05  1.289e-05  -0.781   0.4349
metro.no.transactions     3.037e-04  4.249e-04   0.715   0.4748
metro.total.spend         7.186e-06  8.712e-06   0.825   0.4095
superstore.no.transactions -1.778e-04  4.265e-04  -0.417   0.6767
superstore.total.spend    8.952e-07  3.540e-06   0.253   0.8004
extra.no.transactions    -7.755e-04  4.277e-04  -1.813   0.0698 .
extra.total.spend         7.933e-06  3.547e-06   2.237   0.0253 *
fandf.no.transactions     2.211e-04  4.832e-04   0.458   0.6472
fandf.total.spend        -6.330e-06  9.537e-06  -0.664   0.5069
petrol.no.transactions   -4.022e-04  4.840e-04  -0.831   0.4059
petrol.total.spend       -1.176e-05  1.745e-05  -0.674   0.5004
direct.no.transactions   -3.095e-04  4.842e-04  -0.639   0.5226
direct.total.spend        1.115e-07  3.512e-06   0.032   0.9747
genderMale                1.891e-02  1.517e-02   1.247   0.2126
affluencyLow             -3.872e-02  2.402e-02  -1.612   0.1070
affluencyMid             -5.181e-02  2.015e-02  -2.571   0.0102 *
affluencyVery High       -6.882e-02  3.770e-02  -1.826   0.0679 .
affluencyVery Low        -5.966e-04  3.813e-02  -0.016   0.9875
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 96597  on 69999  degrees of freedom
Residual deviance: 96574  on 69980  degrees of freedom
AIC: 96614

Number of Fisher Scoring iterations: 3
```

Fig. 25. Logistic Regression Summary

The binomial logistic regression falls under the regression model. This model checks the significance level with each of the independent variable. The variable can be removed with p-values less than 0.05 stating that independent variable is not statistically significant. From the summary of this model we can say only two variables are significant i.e. extra total spend and affluency. Rest variables are not that significant and hence can be removed from the model. The iteration in the summary displays that there were total 3 iteration though which we have got the best maximum likelihood estimation. Also, AIC has to be less in order to fit the model better.

```
> confusionMatrix(pred.results, test$content_1)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 16193 13807
         1     0     0

               Accuracy : 0.5398
                 95% CI : (0.5341, 0.5454)
    No Information Rate : 0.5398
    P-Value [Acc > NIR] : 0.5024

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.5398
         Neg Pred Value :    NaN
             Prevalence : 0.5398
         Detection Rate : 0.5398
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

Fig. 26. Logistic Regression Confusion Matrix

We can see that the accuracy for this model is around 54% from the confusion matrix which is half half. Means half of time our prediction is going to to be right and half of the time it isn't. Also, it can see that kappa statistic is 0 means this model is not good fit at all. The result whatever we are going to get has to be false positive or false negative. Sensitivity of the model is 1 which is very rare. This shows that results are classified correctly. So it can be concluded that this model is not a best fit.

One of the sampling method k-fold also has been applied on this dataset. However, the same results has been shown by the model. K value selected was 10 still the accuracy got was 54% with kappa statistics 0.

V. Conclusion and Future Work

All 5 machine learning methodologies have been applied on the 3 datasets. The results of all the models are not up to the mark and there is a lot of scope to improvement. M5-Prime and multiple regression model has almost performed equally well. However, when it comes to the classification model, it can be observed that Decision tree has outsmart the RIPPER algorithm. There are different factors like age, language, ratings, playtime has quite statistical significant in determining the average ratings and owners of the games. Standardization of the data or taking log of data can help in improving the model in certain extent.

These factors can help game developers to understand what kinds of games need to be created. By applying other machine learning methodologies, the significance levels of these factors can be measured which ultimately going to help companies those produces the games. The better the model prediction the more the chances of buying good game for one.

References

[1] N. Davis. Steam store games. [Online]. Available: https://www.kaggle.com/nikdavis/steam-store-games

[2] Aryan. Board games prediction data. [Online]. Available: https://www.kaggle.com/centipede148/board-games-prediction-data

[3] A. Bayowa. Tesco marketing content. [Online]. Available: https://www.kaggle.com/linkonabe/tesco-marketing-content

[4] B. Lantz, *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing Ltd, 2019.

[5] D. Possler, A. S. Kümpel, and J. Unkel, "Entertainment motivations and gaming-specific gratifications as antecedents of digital game enjoyment and appreciation." *Psychology of Popular Media Culture*, 2019.

[6] M. B. Oliver, N. D. Bowman, J. K. Woolley, R. Rogers, B. I. Sherrick, and M.-Y. Chung, "Video games as meaningful entertainment experiences." *Psychology of Popular Media Culture*, vol. 5, no. 4, p. 390, 2016.

[7] R. Staewen, P. Trevino, and C. Yun, "Player characteristics and their relationship to goals and rewards in video games," in *2014 IEEE Games Media Entertainment*. IEEE, 2014, pp. 1–8.

[8] M. Soroush, M. Hancock, and V. K. Bonns, "Self-control in casual games: The relationship between candy crush saga™ players' in-app purchases and self-control," in *2014 IEEE Games Media Entertainment*. IEEE, 2014, pp. 1–6.

[9] K. Fathoni, R. Y. Hakkun, N. Ramadijanti, A. Basuki, and R. Trisanjaya, "Online game server framework for creating platformer games," *International Journal of Simulation–Systems, Science & Technology*, vol. 19, no. 5, 2018.

[10] R. Guidotti, G. Rossetti, L. Pappalardo, F. Giannotti, and D. Pedreschi, "Personalized market basket prediction with temporal annotated recurring sequences," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2151–2163, 2018.

[11] R. Becker, Y. Chernihov, Y. Shavitt, and N. Zilberman, "An analysis of the steam community network evolution," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2012, pp. 1–5.

[12] H. S. Choi, M. S. Ko, D. Medlin, and C. Chen, "The effect of intrinsic and extrinsic quality cues of digital video games on sales: An empirical investigation," *Decision Support Systems*, vol. 106, pp. 86–96, 2018.

[13] J. H. Lee, R. I. Clarke, H. Cho, and T. Windleharth, "Understanding appeals of video games for readers' advisory and recommendation," *Reference & User Services Quarterly*, vol. 57, no. 2, pp. 127–140, 2017.

[14] B. Türkoğlu, "A mixed method research study on the effectiveness of board game based cognitive training programme." *International Journal of Progressive Education*, vol. 15, no. 5, 2019.

[15] E. B. Kirikkaya, S. Iseri, and G. Vurkaya, "A board game about space and solar system for primary school students." *Turkish Online Journal of Educational Technology-TOJET*, vol. 9, no. 2, pp. 1–13, 2010.

[16] I. Boghian, V.-M. Cojocariu, C. V. Popescu, and L. M, "Game-based learning. using board games in adult education," *Journal of Educational Sciences and Psychology*, vol. 9, no. 1, 2019.

[17] S. Noda, K. Shirotsuki, and M. Nakao, "The effectiveness of intervention with board games: a systematic review," *BioPsychoSocial medicine*, vol. 13, no. 1, p. 22, 2019.

[18] A. d'Astous and K. Gagnon, "An inquiry into the factors that impact on consumer appreciation of a board game," *Journal of Consumer Marketing*, 2007.

[19] Z. Ju and Y. Li, "Analysis on internet of things (iot) based on the" subway supermarket" e-commerce mode of tesco," in *2011 International Conference on Information Management, Innovation Management and Industrial Engineering*, vol. 2. IEEE, 2011, pp. 430–433.

[20] Y. Sahin and E. Duman, "Detecting credit card fraud by ann and logistic regression," in *2011 International Symposium on Innovations in Intelligent Systems and Applications*. IEEE, 2011, pp. 315–319.

[21] H. M. Safhi, B. Frikh, and B. Ouhbi, "Assessing reliability of big data knowledge discovery process," *Procedia computer science*, vol. 148, pp. 30–36, 2019.

[22] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman, *Using multivariate statistics*. Pearson Boston, MA, 2007, vol. 5.