Factors Affecting Target Audiences
For
Movie Promotions

Charu Girdhar – 218256
ANLY 502-51, Fall -2017
Professor Dr. Ali Motamedi
Harrisburg University of Science and Technology

A**bstract**

The primary objective of the research is to identify the variables that affect the preference of movie selection in a designated region or county. This research makes use of a subset of survey data collected by The GroupLens Research Project for their recommender system. The underlying research will be performed on 3118 records of 100 movie goers, who were asked to rate movies from 7 major genres. Ratings variable will help in identifying and concluding impact the factors for this research with the help of few statistical methodologies and R studio programming.

Introduction

In the current dynamics of world, where data analytics is vital to every business, it is crucial for entertainment industry to make use of data science to maximize profit using intelligent marketing. There are several ways to achieve that using focused advertising. If target audience is known and factors are statistically backed, advertising money can be put to use in best places and will become more lucrative. This research paper focusses on obtaining, examining and analyzing the factors which affect the preference of customer's movie choices based on several factors like age, occupation, location, gender, genres and movie ratings etc. The variable movie ratings will be used to create a dummy variable called "Liking" which will have a value 'like' if the rating is 3 and above else its value will be 'not like'. This dummy variable will act as the direct outcome variable in this research which will help us achieve the conclusion. This whole research is based upon a survey conducted by The GroupLens Research Project for a movie recommender system developed in 1992. The survey data collected by GroupLens research project will be modified, tidied, and statistically analyzed in this paper to answer the main question of how and what factors affect the movie choices of a customer and which ones are most and least effective for the target advertising.

Research Objective

This research will help examining the various elements of a populations demographics which impacts the commercial movie market. It can be used as a baseline for decision making while planning for promotions and maximizing the impact on the target audience.

The key questions that will be answered in this research are:

- Does age affect the choice of movies for a customer? If yes, how big a factor is it?
- How important is the occupation of a customer which makes him/her the target audience?
- Gender is definitely a factor for audience selection, but how big a factor is it?
- Genre is a major factor for audience selection, but how big a factor is for the selected region?

Several statistical methods and analysis techniques like summarization, EDA, regression and hypothesis testing will be used and implemented in R studio to arrive at a conclusion for the research.

Data Summary and Methodology

The Database used in this paper was created through a survey conducted by The GroupLens Research Project. The main objective of their research was to develop information filtering, collaborative filtering, and recommender systems using historical data for prediction. The information generated by these systems is used at smaller level to predict and recommend movies to visitors of a web page. The main focus in this research was to identify how a customer has given rating to a movie and then provides recommendations based on individual's choice. But it does not focus on group trends and does not categorically factor the key elements affecting the choice; which will be the main objective of this paper.

A subset of original database will be fetched out with the conditional requirement that the data belongs to one particular region (pin code). Data for 100 customers from that region, who have rated the 7 major genres of films including Action, Romance, Comedy etc. has been filtered out for this research. The modified dataset has 3118 records with movie ratings, having values 1 to 5. For the purpose of our research we will add a dummy variable liking with values Like and Not Like. The new response variable is a categorical variable, which will help us in identifying the patterns in data. The data set has 4 predictor variables age, occupation, genres and gender which again are categorical variables with a fixed set of values.

Data Summary:

Variable	**Categories**	**Proportion %**	**NA's**	**Role in ongoing research**
Liked	Not Like	44	None	Response
	Like	56		
Age	below 18	22	None	Factor
	19 to 35	54		
	36 to 50	18		
	51 above	6		
Occupation	A – Artist	17	None	Factor
	H – Homemaker	5		
	P – Professional	58		
	S- Student	18		
Gender	F- Female	35	None	Factor
	M- Male	65		
Genre	Action	5	None	Factor
	Animation	2		
	Comedy	43		
	Drama	40		
	Horror	3		
	Romance	1		
	Sci-Fi	2		
Total Observations (n)	3118			
Survey Participants	100			

Summary table shows the categories in all the response and factor variables in the dataset. It shows the proportion percentage falling under that category from whole dataset. It also displays if any record has NA’s in any other column for that category. Finally, it shows the role of (factor or response variable) that categorical variable in the ongoing research.

Methodology:

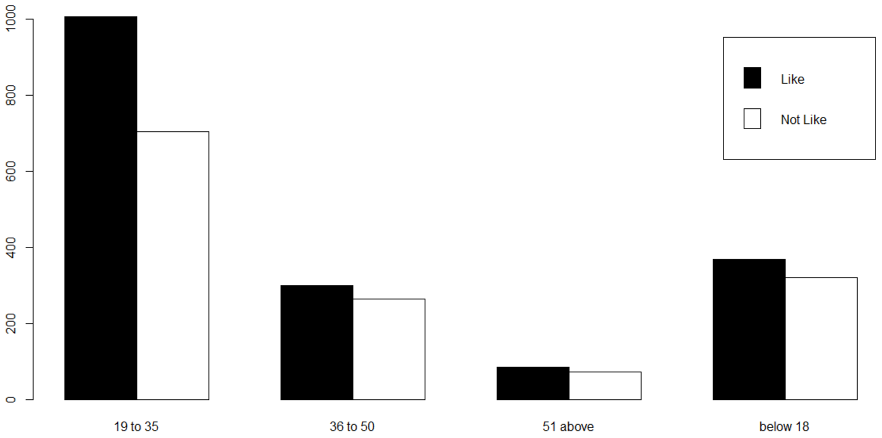
This research will focus on 2 methods of Data analysis to identify and weigh the relationship between various factor variables (Age, Occupation, Gender, genre) and outcome variable (Liking). The first methodology will focus on the aspect of association among variables. The second method will calculate the weight of each association with respect to the various categories of a factor variable.

Since the variables in our research are all categorical variables, we will use Pearson Chi-Squares Test of Independence between each of those pairs, which is perfect for Identifying the association between a pair of 2 categorical variables. After it is established whether “Liking” response is associated with Age, Gender, Occupation and Genre, we will perform regression analysis on each pair to identify which value of each categorical variable increases the number of likings for the movies in the region under research. Since our outcome variable is binary (like/not like) we will use logistic regression with family = binomial. A significance level of 0.05 will be assumed in both the methods.

- Pearson Chi-Square Test of Independence.

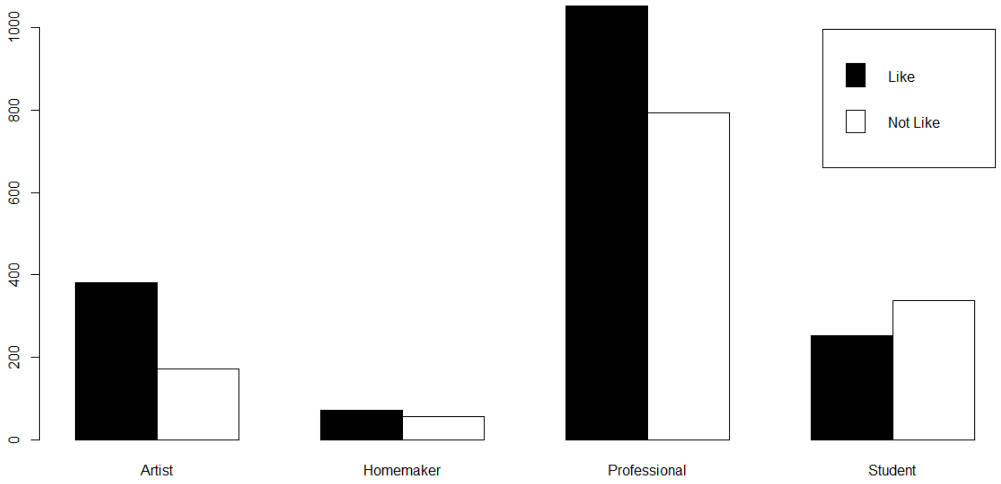
Age – The variable age has been divided into 4 broad categories as mentioned in below table. Pearson Chi Square test of independence test checks if movie likes are associated with the age of the participant or not. P value for the test is 0.05723(0.06), which is marginally greater than our significance level of 0.05. This means we cannot reject the null hypothesis that age and liking are independent of each other. The histogram of age and liking also depicts the same conclusion, as we can see the proportion of like and not like in each age group is almost similar.

Age ->	**Below 18**	**19 to 35**	**36 to 50**	**51 above**
Liked	369	1005	300	84
Not Like	320	703	263	73
		P-value = 0.05723		



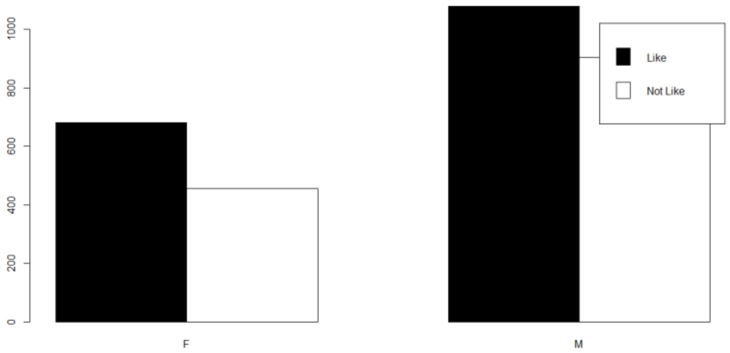
Occupation – The variable Occupation has 4 major categories which have been put in a contingency table to perform Pearson Chi Square Test of Independence. The test result shows a p -value < 2.2e-16. Which means we can reject the null and accept the alternative that occupation of a person affects the number of likes. In the next step of regression, we will be able to identify which occupations is more likely to enhance the outcome variable. The histogram gives us an idea that a professional may be most likely to like a movie, which we will confirm later using logistic regression.

Occupation ->	**Artist**	**Homemaker**	**Professional**	**Student**
Like	382	72	1052	252
Not Like	171	56	794	338
		P-value < 2.2e-16		



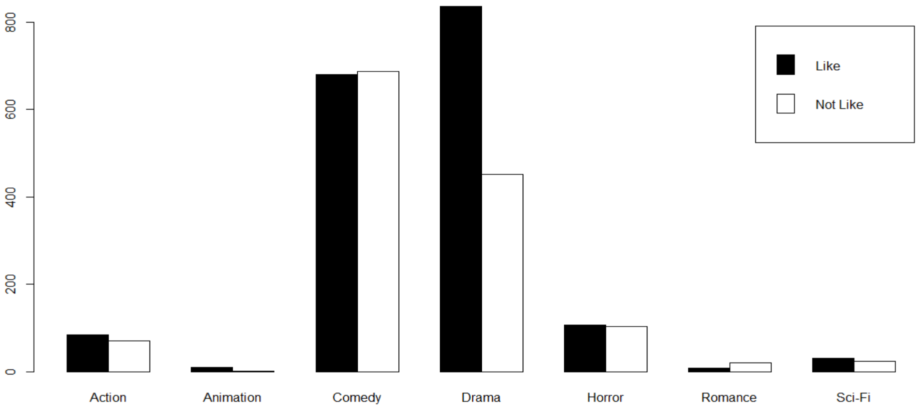
Gender – Pearson chi-square test of Independence will check if there is any association between a person being male or female affects the chances of liking a movie in the region or not. p-value of 0.003115 tells us it does affect. Low p-value tells us to reject the null hypothesis and accept the fact that the gender and liking are in fact associated.

Gender	**Female**	**Male**
Like	680	1078
Not Like	455	904
	P-value = 0.003135	



Genre – This a categorical variable with 7 major categories of movies (Action, Animation, Comedy, Drama, Horror, Romance, Sci-Fi), This may be the most interesting variable for the purpose of target advertising. Pearson Chi Square Test shows a p-value of 1.135e-14. Which is less than 0.05, hence the variable occupation is closely associated with the outcome variable. Looking at the histogram, we can see the variation in proportion of like and not like in different genres.

Genre	**Action**	**Animation**	**Comedy**	**Drama**	**Horror**	**Romance**	**Sci-Fi**
Like	85	10	681	836	104	8	31
Not Like	70	2	687	452	104	20	24
			P-value =	1.135e-14			



- Logistic Regression.

  Age- In the previous section we proved that the 2 variables age and liking are independent and there is no association between age of a person and his/her liking of the movies. Although we need not perform the logistic regression on this variable, but for the sake of confirming our findings, logistic regression is performed on age categorical variable. The following table depicts the coefficients of logistic regression in its exponential form. Since glm () function in r gives the logit (log of odds) value of coefficient, exponentials of those log odds can be used to find the actual value of impact on response variable. Here we can see that all age groups affect the outcome variable equally, and there is no one category of age, which could significantly increase of decrease the probability of liking or not liking the movie.

Liking ~ Age

Exponential of Logit coefficients:

Below 18	0.70
19 to 35	0.88
36 to 50	0.87
51 above	0.87

Occupation- Exponential values of coefficients of logistic regression for occupation are shown below. A value of 3.0 depicts that increase in student as a customer may increase the probability of liking the movie by 3 times. Stating it in simple words, student is more likely to like a movie as compared to an artist, professional or a homemaker. Here we can see the importance of regression analysis over simple EDA using based on counts. Histogram evaluation showed us that professionals have high impact, but regression analysis showed us a better picture, with student having most impact on probability of outcome variable.

Liking ~ Occupation

Exponential of Logit coefficients:

Artist	1.45
Homemaker	1.74
Professional	1.69
Student	3.00

Gender – Logistic regression on gender as factor variable shows a strong relationship gender of the person and outcome variable. Coefficients in exponential form for female and male are 0.67 and 1.25 respectively. This shows that having a male customer increase the probability of liking a movie 1.25 times better than a female customer.

Liking ~ Gender

Exponential of Logit coefficients:

Female	0.67
Male	1.25

Genre: Performing logistic linear regression on genres gives the following coefficients for each category of movie type. Looking at the exponential values of coefficients we can deduce that Romantic movies are most likely to be liked by people of this region. After Romance, Comedy and horror are most likely to succeed. Animation is the least liked genre and Action and Sci-Fi did mediocre in our survey.

Liking ~ genres

Exponential of Logit coefficients:

Action	0.82
Animation	0.20
Comedy	1.01
Drama	0.54
Horror	0.97
Romance	2.50
Sci-fi	0.77

Conclusion

The intent of this research was to identify the factors which can affect the movie preferences of residents of a region under research. This research was performed on subset of big research data set for a movie recommender system. This research subset of 3118 records had 6 variables, out of which ID variable was not used in the research. A Categorical variable “Liking” with 2 values “Like and “Not Like” was considered as the Dependent/Response variable. Rest of the 4 variables namely Age, Occupation, Gender & Genre were used as the predictor variables. All the variables used in this research were categorical variables.

The research was conducted in 2 parts. In part 1 association between each pair of Factor-Response variables was performed using Pearson Chi-Squared test of Independence. And In part 2 logistic regression analysis was performed on the same pair, to quantify that relationship and also to identify the category from that factor variable which is most important for the purpose of advertisement. A significance level of 0.05 was assumed to perform test of independence between Liking-Age, Liking-Occupation, Liking-Gender and Liking-Genre.

The test of Independence gave us some surprising results. For e.g. Liking-Age pair, one might assume that there is a high probability of some amount of association between Age and Liking. But the test of Hypothesis showed us there is no association, and all the customers like/not like the movies equally, this was proved using output of logistic regression where each age group had almost equal coefficient values. Test of independence for rest of the factor-response pairs was positive, meaning occupation, gender and genre do have some amount of association with liking variable. Performing logistic analysis on rest of the 3 pairs gave us the conclusion that a student customer has high effect on increasing the liking probability than a professional or a homemaker, but artists are least likely to affect the outcome. Similarly, male respondents are big contributors in increasing the liking factor than the female ones. Finally, logistic regression of the most interesting factor of genre showed us that a romance movie multiplies the number of likes by 2.5 times. Comedy and horror follow the romance genre and rest of the genres are not that effective.
This research has answered all our key questions efficiently, and we can use these interpretations in promotions and target advertising to effectively use our marketing cost.

There are several more robust, accurate and explanatory statistical methods for these types of analysis, and this research can be extended by using those methods, including more participants from the same region, or may be adding more factor variables to the dataset.

Works Cited

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: Historyand Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4,
Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872

References

http://www.math.csi.cuny.edu/Statistics/R/simpleR/stat006.html
https://www.youtube.com/watch?v=EocjYP5h0cE&t=2186s
http://stattrek.com/chi-square-test/independence.aspx?Tutorial=AP

R- Code

```
movie = read.csv(file.choose(),header=T)
head(movie)
str(movie)
summary(movie)
library(UsingR)
occu = table(movie$Liking,movie$Occupation)
prop.table(occu)
barplot(occu,beside=T,legend.text = T, col = c("white","Black"))
chisq.test(occu)
gen = table(movie$Liking,movie$Gender)
prop.table(gen)
barplot(gen,beside=T,legend.text = T, col = c("black","White"))
chisq.test(gen)
age= table(movie$Liking,movie$Age)
barplot(age,beside=T,legend.text = T, col = c("black","White"))
prop.table(age)
```

```
chisq.test(age)
genre = table(movie$Liking,movie$genres)
barplot(genre,beside=T,legend.text = T, col = c("black","White"))
prop.table(genre)
chisq.test(genre)
result = glm(Liking ~ genres , family ="binomial",movie)
summary(result)
round(exp(cbind(extimate = coef(result),confint(result))),2)
result = glm(Liking ~ Gender-1, family ="binomial",movie)
summary(result)
round(exp(cbind(extimate = coef(result),confint(result))),2)
result = glm(Liking ~ Age - 1, family ="binomial",movie)
summary(result)
round(exp(cbind(extimate = coef(result),confint(result))),2)
result = glm(Liking ~ Occupation, family ="binomial",movie)
summary(result)
round(exp(cbind(extimate = coef(result),confint(result))),2)
```