Recommendation of Movie Genre Based on Hobbies and Interests

Sanjana Moudgalya
B.Tech, CSE Dept.
PES University
Bangalore, India
sanjanam1998@gmail.com

Anusha S Rao
B.Tech, CSE Dept.
PES University
Bangalore, India
anusha.s.rao1999@gmail.com

Aninditha Ramesh
B.Tech, CSE Dept.
PES University
Bangalore, India
aninditha12@gmail.com

***Abstract*—People spend a lot of time on their hobbies (reading, writing, dancing, sports or even socializing). This research is designed to examine the correlation between these interests and the genre of movies preferred. Sometimes fears also affect the movie enjoyed. The aim of this project is to recommend a movie genre to a user after considering a few attributes like hobbies and fears.**

The users can be clustered based on their likes and dislikes. The similarity between the users in a particular cluster can be used to determine the movie genres for a new user. Agglomerative Kmodes was used for clustering and collaborative filtering was used to calculate similarity and recommend genres.

The accuracy of this recommendation system was calculated using K-fold cross validation.

***Index Terms*—personality prediction, Agglomerative Kmodes, collaborative filtering, movie recommendation, visualization, K-fold cross validation.**

I. Introduction

Movie recommendation is a comprehensive and complicated task which involves various tastes of users, various genres of movies, and so forth. Therefore, lots of techniques for recommendation have been proposed to solve the problems. Each technique has its own advantage in solving specific problems. Considering the usage of online information and user-generated content, collaborative filtering is supposed to be the most popular and widely deployed technique in recommender system.

Due to the ambiguity of perceiving human interests / personality, predictive modelling in the field of psychology can be challenging. Many psychologists have argued that researchers need to pay more attention to ordinary aspects of people's daily lives. Entertainment is undoubtedly important to people and with the ease of internet access, there is a significant growth in the amount of time spent to watch movies of different genres. The movie watched may reflect the personality of a human.

As an example, some people develop certain hobbies such as video games addiction in order to become in control of the imaginary world these games allow them to live in. When some people feel powerless or not in control they might look for another realm that can make them feel in control such as an imaginary world provided by a certain game. That's also the same reason why some people like action and fantasy movies. Hence a recommendation of movies based on hobbies and interests might come handy. The present research provides a gateway to explore this growing domain.

II. Literature Survey

A. A Movie Recommender System: MOVREC [2]

- This paper is based on collaborative filtering approach that makes use of the information entered by the users, analyzes them and then recommends the movies that is best suited to the user at that time. The recommended movie list is sorted according to the ratings given to these movies by previous users.
- K-means is the algorithm used to recommend the movies to the user.K initial centroids where chosen where K is the desired number of clusters.Each point is then assigned to the nearest centroid. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.
- Using this as a reference we plan on recommending movies to the user based on their hobbies or interests, instead of collaborative filtering that analyzed the rating of the previous users.

B. Listening, Watching, and Reading: The Structure and Correlates of Entertainment Preferences [1]

- The primary aim of this research was to inspect the personality or structure based on entertainment preferences. Preference ratings of 3,227 participants from the three samples combined, using all 108 music, film, book, and television genre labels were considered as variables.
- Five Entertainment-Preference Factors (communal, aesthetic, dark, thrilling and cerebral) has been considered to categorize the entertainment preference variables. Then correlations of these preferences with personality factors and demographics vary between 0.22 and 0.64.
- One of the limitations of this research was that all the data analyzed were based on self-reports. This might have caused a problem as a user might enter the details based on what is considered trendy/desirable. Thus, the results might not be the same if the preferences were collected through some other sources which reflect their personality (eg. Movies, Books, CD).

- An improvement to considering only personality ratings would be to consider the actual hobbies, interests, fears and a mentioned taste in different movie genres.

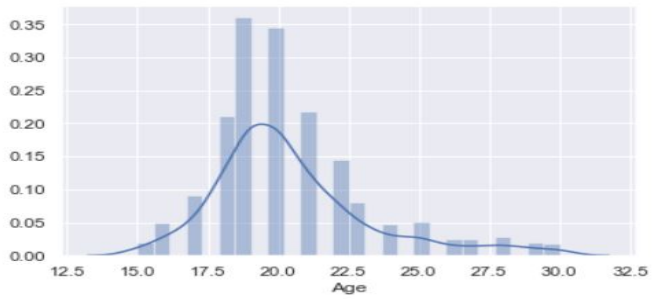
C. Personality Based Music Recommendation System [3]

- The main aim of this thesis was to discover the impact of personality traits in collaborative filtering, which is a recommendation system technique that relies on the user's historical preferences
- It used users' social media accounts to to obtain an insight into their personalities. The personalities were modelled based on the "Big Five Model", also known as the "Five Factor model". The FFM suggests five broad dimensions commonly used to describe the human personality, namely Openness to Experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism.Two models were used for personality classification of a user, the Naive Bayes classifier and Logistic regression. This classifies users into the categories mentioned above.
- The RMSE values for these models was used as a measure of comparison. The RMSE for the collaborative filtering model was 3.04.
- An improvement for this model would be to consider emojis used since they are used by users to express their feelings. An insight into the user's age may also be helpful. Since music and movies go hand in hand, We would like to predict movies based on user's personality.

III. Problem Statement

The aim of this project is to build a movie recommendation system based on human hobbies and interests.
The 'Young People Survey' dataset (1010 x 150) has been chosen from Kaggle. This data has been collected in a survey conducted on the youth of Slovakia. There are two csv files, responses.csv and columns.csv. The former file contains ratings of youth to various questions asked about their personal interests and demographic data. The latter contains the detailed description of the questions asked. The attributes include movie genre preferences, music genre preferences, hobbies, interests, fears, education, age, weight, height. Most of the attribute entries are in the form of levels ranging from 1-5 (1 indicates strongly disagree and 5 indicates strongly agree).

Fig. 1. Distribution of Age



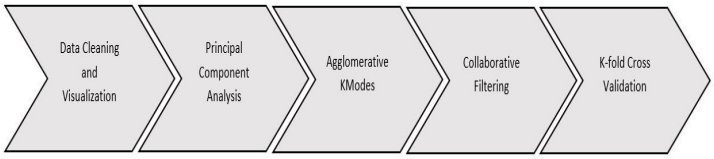
The assumption made for this project is that the Slovakian data is a fair representation of all the youth in the world. The hobbies and fears considered were limited. This was due to unavailability of data through the survey.
The biggest constraint was that all the data considered was categorical. The way each user looks at and perceives the scale of ratings (1-5) might be different. The responses may be biased due to Hawthorne effect where subjects may change their responses due to their consciousness.

Category	*Number of Columns*
Music Preferences	19
Movie Preferences	12
Hobbies and Interests	32
Phobias	10
Health Habits	3
Spending Habits	7
Demographics	10

TABLE I
GROUPING OF VARIOUS COLUMNS

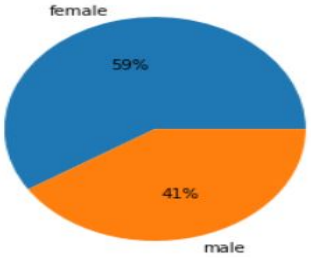
IV. Methodology

Fig. 2. Block Diagram of Methodology



- ***Cleaning and Visualization***
  Data cleaning and visualization play a major role in Exploratory Data Analysis (an approach to analyzing data sets to summarize their main characteristics). Since most of the columns contained categorical variables, only the missing values had to be filled in. Missing values are those rows for which data is not available due to various reasons like unwillingness of the user who took the survey. These values were filled in by taking median value of that particular column. Few columns had only two categories as entries (eg. male and female, village and town, left handed and right handed). Values were encoded into 0 or 1 based on their category. Based on the ratio of the categories, missing values were filled proportional to the calculated ratio.

Fig. 3. Pie Chart showing Male:Female Ratio



The demographics like age, height and weight had outliers in the data. Example of outlier in the height column - all the values are entered in 'cm' but one value had 60 as an entry. This is not possible for a human being. Hence, this value was changed to 'inch' (cm = inch*2.54). This method was followed for the weight column too (kg = pound*0.454). Due to the varied range of height and weight in humans, all outliers were not replaced as they maybe a good representation of the extremes.

Fig. 4. Box Plot of Height Column Before Removing Unlikely Values

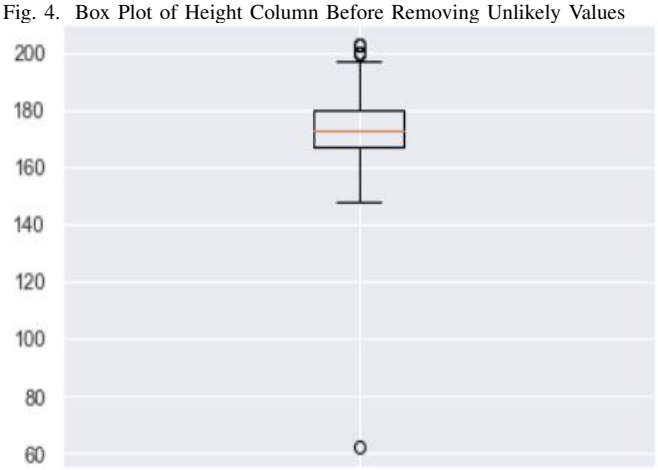
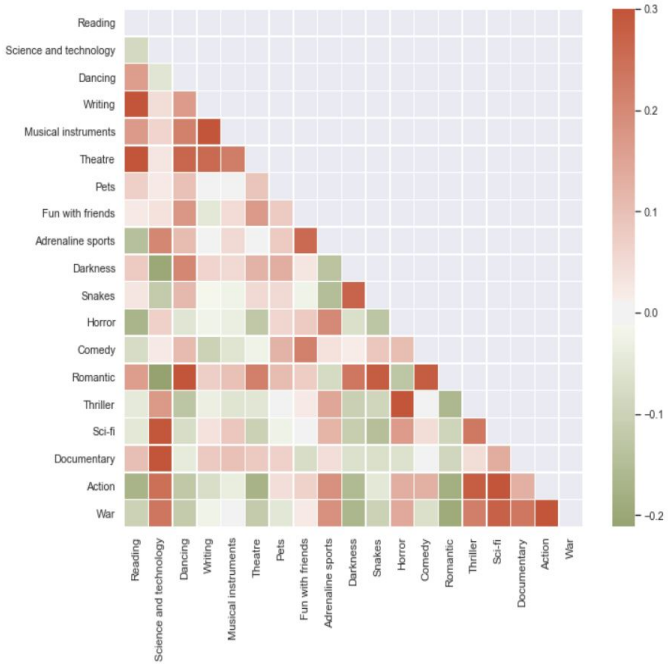


Fig. 5. Box Plot of Height Column After Removing Unlikely Values

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
Histograms, box plots, pie charts and correlograms were plotted to visualize data and get a better understanding of the columns. Since only movie genres, hobbies and phobias were used for further research, a correlogram of those columns was plotted. It was inferred that most columns have weak correlations.

Fig. 6. Correlogram of Relevent Columns



Few observed correlations are -
A person who likes writing generally enjoys reading too
Since horror and thriller genres are similar, they are correlated
Romantic genre is correlated to comedy genre (RomCom) and dancing

- Principle Component Analysis
  It is technique of linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space.Since the data set had a large number of columns, dimensionality reduction necessary to get a better understanding of the data.

- ***Models***
  - Linear Regression
    Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). All the hobbies were used as independent variables, to predict the one of the movie gene('Comedy').The 'LinearRegression' function from sklearn.linear_models library was used. R-squared value of 0.028 was obtained. This wasn't good enough to proceed with the model. This might be due to the fact that the data being purely categorical might not be linearly dependent.

  - Kmeans
    K-means clustering is one of the simplest

and popular unsupervised machine learning algorithms.Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. In order to use this, the data was divided into two clusters based on Age. The 'Kmeans' function from sklearn.cluster library was used. As there was no clear distinguishing factor between clusters, this model wasn't used in the future. The data being categorical was the main disadvantage.

- Agglomerative Kmodes
  To remove the numeric-only limitation of the k-means algorithm,the k-modes algorithm which extends the k-means algorithm was developed by using a simple matching dissimilarity measure for categorical attributes, modes in place of means for clustering and a frequency-related strategy to update modes to minimize the clustering cost. Initially data was divided into 2 clusters based on age. Since this clusters were not an appropriate representation of the data, Agglomerative Kmodes based clustering was implemented. It is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. Initially the data was divided into 500 clusters.The number of clusters were gradually reduced until a minimum cost was obtained.Thus the optimal number of clusters was found to be 8. Few inferences made from these clusters are -
  Somebody who likes socializing, spends a lot of time on the internet.
  In general, people who like learning foreign languages aren't afraid of public speaking.
  Youth in Slovakia do not prefer any trivial educational subjects like Mathematics, Biology, History, etc.
  In this new era of internet, with youth addicted to it, extra hobbies and interest in studies are fading away.

- ***Collaborative Filtering***
  Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user. It mimics user-to-user recommendations and predicts users preferences as a linear, weighted combination of other user preferences.After dividing the data into clusters, the similarity of the users was calculated using cosine similarity with hobbies of the users as a n dimensional vector-

$$\text{Similarity} = \cos(\Theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

For every new user added,the five users in the dataset who are most similar to the new user are considered for recommendation. The ratings of the these five users, are used to predict the rating of the new user for a particular genre.

$$\text{Rating} = \frac{\sum_{i=1}^{n} similarity_i * A_i}{\sum_{i=1}^{n} similarityi}$$

After predicting the ratings for different genres, the genres with the top three rating are recommended to the user.

V. Evaluation Metrics

Validation of a model is important as it gives us an idea of how it will behave with unseen data. Based on these results, one can find out whether the model over-fits or under-fits the data, or is a generalised representation. K-fold cross validation was used to evaluate the model. Cross validation is a re-sampling procedure used to evaluate a model when there is limited data. A portion of the data is kept aside. The model is trained on the remaining data. The model is then evaluated on the portion of data that was left out. In K-fold cross validation, the model is tested 'K' times. The data is first divided randomly into 'K' folds or partitions. The model is fit on (K-1) folds. It is then evaluated on the K'th fold and the errors are noted down. This process is repeated until every K'th fold serves as the test set. The average of the accuracies recorded serves as the performance metric for the model.
Five fold validation was performed on the model. The sklearn library was used, which contained the KFold function. The parameters for the function included number of folds, a Boolean value for whether the data should be shuffled and the seed for the pseudo-random number generator used prior to the shuffle. The boolean parameter was set to true and the seed value was set to one.

VI. Results

The following table shows the accuracy of the model after performing 5-fold cross validation.

Fold	***Accuracy*(%)**
Fold 1	33.16
Fold 2	28.71
Fold 3	33.33
Fold 4	31.51
Fold 5	34.81
Average accuracy	**32.31**

TABLE II
Accuracy measure

Accuracy varies with each fold, since the performance of the model on different portions of the data are different. Thus, the average of the accuracy of each fold is an appropriate representation of the model.

VII. Conclusions

The accuracy of the model was 32.31%. There are many possible reasons for a low accuracy. One of the main reasons are that human preferences cannot be captured accurately by a model. The human brain is very complex. The genre that a person prefers watching can depend on other factors such as their mood. Changing trends may cause the popularity of one genre to shoot up. Due to the pressure of keeping up with recent trends, users may prefer watching other genres.
A larger data set could have increased the accuracy of the model. Improvements include considering user's mood while recommending genres. Other accuracy metrics such as Dissimilarity and Coverage could also be used. They consider what percentage of the predictions made were wrong.

VIII. Contributions

- Team Member 1 - Aninditha Ramesh
  -Principal Component Analysis
  Since the data set had a large number of columns, dimensionality reduction was done using PCA.
  -Kmeans
  Inorder to divide the users into categories, K-means clustering was explored.
  -Kfold cross validation
  The accuracy of the recommendation system was calculated using cross validation techniques.

- Team Member 2 - Sanjana Moudgalya
  -Visualization
  Various visualization techniques were used to analyse the data and get a better idea about its representation.
  -Kmodes
  As the data was categorical, as an alternative to Kmeans, Kmodes was used to divide the users into clusters.
  -Agglomerative Kmodes
  Since Kmodes was not giving the desired results, agglomerative kmodes was used. The optimum clusters was found out and then data was clustered.

- Team Member 3 - Anusha S Rao
  -Cleaning
  Missing data and outliers was taken care of in most columns. In addition, pre-processing was done on the categorical data to encode it.
  -Linear Regression
  Regression was the first approach tried. The categorical data might not be linear and hence did not give us good results.
  -Collaborative Filtering
  User-user similarity was used to recommend top 3 movie genres for users in the test data.

References

[1] Rentfrow, Peter J et al. "Listening, watching, and reading: the structure and correlates of entertainment preferences." Journal of personality vol. 79,2 (2011): 223-58.

[2] Kumar, Manoj, D. K. Yadav, Ankur Singh, and Vijay Kr Gupta. "A movie recommender system: Movrec." International Journal of Computer Applications 124, no. 3 (2015).

[3] Abhishek Paudel, Brihat Ratna Bajracharya, Miran Ghimire, Nabin Bhattarai. "Personality Based Music Recommendation System"

[4] Klein, D. N., Kotov, R., Bufferd, S. J. (2011). Personality and depression: explanatory models and review of the evidence. Annual review of clinical psychology, 7, 269–295.

[5] https://www.verywellmind.com/music-and-personality-2795424

[6] https://pypi.org/project/kmodes/

[7] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[8] https://machinelearningmastery.com/k-fold-cross-validation/

[9] https://realpython.com/build-recommendation-engine-collaborative-filtering/