Social Relationship Recommendation

Nivetha Jayakumar
Department of Computer Engineering
San Jose State University
Student-Id : 013758667
San Jose, CA

***Abstract*—The social relationship is something that extends beyond the normal connection that we have with others. This paper discusses on what and how influential the features and other attributes of interests and background plays a role in making a successful match. Random forest, Support Vector Classifier and Gradient boosting are the models built for the analysis.**

***Keywords*—** Social recommendation, Random forest, Gradient boosting, Support Vector Classifier, Gradient Boosting, Dating

I. Introduction

The social relationship recommender is built to analyze what features influences to have a passion towards someone or to maintain a healthy relationship with the opposite sex. The few factors that were considered to approach this were how different are the people responding to the racial background, how important are the shared interests among them and how important did the attractiveness of the partner plays a role.

II. DataSet Overview

Data were collected from participants on an event of experimental dating dates back to 2002-2004 where the attendees would be introduced to their first four-minute date with an attendee of the opposite sex. They were also asked to rate their date on six features namely Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. The dataset also includes questionnaire data gathered from participants at different points in the process. These fields include demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information. The data had an attribute the location, of which the data set had redundant data such as duplicates of a location in the form of both abbreviation and acronym as such like NY and New York, NYC and New York City. The data was further categorized removing all these inconsistencies and the data was refined for a better predictive model.

III. Data Preprocessing

The types of preprocessing are furnished as below,

1) Data Cleaning
2) Data Visualization

A. Data Cleaning

Data preprocessing is a process of cleansing data for acquiring more knowledge about it. We are surrounded by immense data for which we lack knowledge as there exists a lot of inconsistencies in data. The data is available in various formats such as files, databases or comma-separated values. The data obtained can have missing values, incorrect data or misspellings during data entry. These inconsistencies are reviewed and the outliers are removed before the data is exposed to further processing. Data preprocessing certainly increases the quality and readability of the data.

B. Data Visualization

The next step in preprocessing is data visualization. "A picture is worth a thousand words" and once the dataset is refined, it is exposed to visualize the patterns, trends and correlations. Our data was visualised to find the relationship between certain features for a better understanding of the dataset before building a model.

The following are the relationships observed between few attributes,

1) Age differences between the matches: Once the data cleaning is done, the refined data is plotted visually to see the difference in the age of the partners that matched. This is visualized with the help of histograms. The histogram

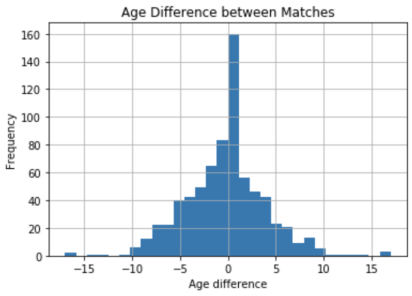

Fig. 1. Age differences between the partners

distribution with thirty bins clearly shows that there were more matches with zero differences or less than 5 years of differences. From this, we could infer that age plays an important role in choosing a successful date partner.

2) Distribution of gender: To check if the data is unbiased, the distribution of male and female ratio was visualised. The dataset seems to have a good distribution of the gender as we see that the male to female ratio is nearly equal which is just a little less than the total number of men and not a huge disparity. This gives an idea of how the male and female ratio

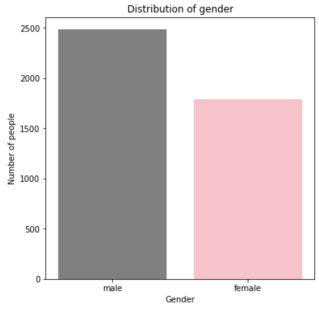

Fig. 2. Distribution of male and female

are distributed for matching to a partner on a date night.

3) Racial Distribution: The data for the distribution of races were plotted. The races in the dataset were Black/African American, European/Caucasian-American, Latino/Hispanic American, Asian/Pacific Islander/Asian-American, Native American and other races. From the above, out of all the

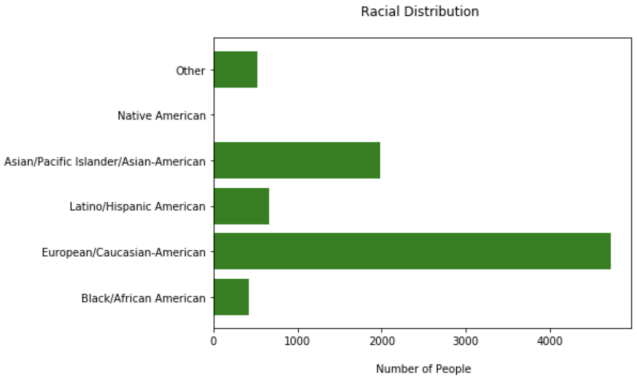

Fig. 3. Racial Distribution

races the highest and the popular was European/Caucasian-American followed by Asian/Pacific Islander/Asian-America and others. However the races were distributed, how much is race given among the people has to be noted.

4) Racial Importance: The racial importance was plotted to see how important the race played a role to match to a partner. On a scale of ten, the importance of race was visualised from the dataset. From the above, it is inferred that most of the people rated 1.0 on a scale of 10.0 for their personal opinion on whether the partner has to be of the same race or not for a match.

5) Importance of religion: The importance of religion is visualized and depicts that it gives nearly no importance as such as racial distribution. From the plot on a scale of 10, most

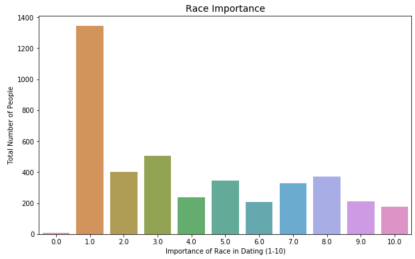

Fig. 4. Racial importance on a scale of 1-10

of the people have rated 1 on 10 for its importance in matching to a partner for a date. Hence from the above all visualizations,

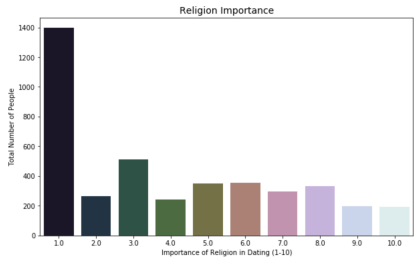

Fig. 5. Religion importance on a scale of 1-10

the general idea of the dataset was known. The features from the recommendation models can be filtered based on the data inferred from the visualizations.

IV. Recommendation Model

In this paper, I have chosen three models: Random Forest, Support Vector Classifier and Gradient boosting.

1) Random Forest: Random forest is a supervised learning and a easy to use algorithm. It consists of many decision trees predicting and one is chosen based on the voting. The forest with more trees is considered to be more robust algorithm. For the dating dataset, the estimators were taken as hundred for a better prediction that included the features chosen for the target variable 'match'. Before fitting the data into the model the data was split into test and train set using train test split library.

- Importance of features

From the above bar graph, the importance of the features considered for classification is seen. Among the features considered which is income, goal, location, age of the partner and attractiveness of the partner, the income stays at the top gaining the most importance.

The random forest model is analysed for its metric and they are furnished as below,

- Accuracy Score : 0.80
- Precision Score : 0.39
- Recall Score : 0.70

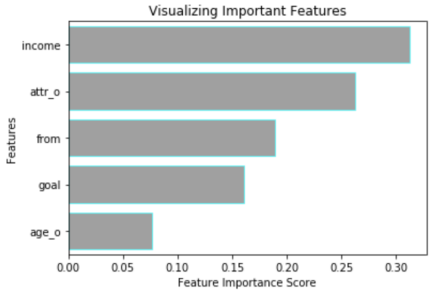

Fig. 6. Importance of features considered for classification

2) Support Vector Classifier: Support Vectors Classifiers are simple compared to logistic regression and decision trees. The basic idea of support vector classifier is that it finds a hyperplane which divides the dataset into with an acceptable margin between them. Hyperplane is the decision line that separates the set of objects on a dataset. Support Vectors are those points that are closer to the the decision making marginal hyperplane.

SVC does not work well for larger datasets as it takes a lot of time to manipulate it. If the dataset cannot be divided by the decision line then for non linear data points, it mis moved to a larger dimension inorder to classify. It is also called as discriminative classifier.The dataset is divided into features and labels for the support vector classifier for which the label is the target variable 'match'. The data set is separated into test and train set to classify the model and the further analysis is done.

The support vector classifier model is analysed for its metric and they are furnished as below,

- Accuracy Score : 0.73
- Precision Score : 0.31
- Recall Score : 0.36

3) Gradient Boosting: Gradient Boosting is an algorithm that can be used for both classification and regression. The idea of adjusting the weak learners is boosting. This algorithm is efficient than the others as in every stages of boosting, weak learners are added to the existing weak learners.

Gradient boosting is used to analyze instances that cannot be predicted accurately. For our data set, the features considered were same as such of random forest for which the classification is performed. Ensembles are added in stages. Ensembles are a combination of separate models. The data is fit into the model and the performance is calculated for the features selected. The most popular algorithm in gradient boosting is XGBoost.

The gradient boosting classifier model is analysed for its metric and they are furnished as below,

- Accuracy Score : 0.78
- Precision Score : 0.35
- Recall Score : 0.22

V. Future work

The future work of this paper is to have a deeper analysis on the data to find better correlation of the features and to increase the accuracy by choosing the right features based on its importance.

VI. Conclusion

In this paper, I have successfully implemented three models for the dataset namely Random forest, Gradient boosting and Support Vector Classifiers. The dataset was preprocessed for redundancies and noisy data were eliminated. The missing values were either removed or the median values were updated in the place of NaN values. To have better understanding of the preprocessed data, the data visualization was done. The metrics of three models are calculated to have known the best model for the future work on data analysis and implementation.

VII. Acknowledgment

We would like to thank Professor Shih Yu Chang and our Teaching Assistant Mr. Surya Sonti for guidance and support in successfully executing this project.

References

[1] JS House, KR Landis, D Umberson - Science, 1988 Social relationships and health
[2] N Gharachorloo, A Moshrefi, A Nasir - US Patent 9,489,698, 2016 Media content recommendations based on social network relationship
[3] P Resnick, HR Varian - Communications of the ACM, 1997 Recommender systems
[4] G Adomavicius, A Tuzhilin - IEEE Transactions on Knowledge Data, 2005 Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions
[5] J Bobadilla, F Ortega, A Hernando - Knowledge-based systems, 2013 Recommender systems survey