Speed Dating Final Report

CIS 9660 Data Mining for business analytics

Professor:Qiang(David) Gao

Group 6:
Shiqi Lin
Monica Dabas
Anna (Yat Sen) Tse
Christopher Perinpanayagam

TABLE OF CONTENTS

Executive Summary

As a team, we have chosen a study done by the Columbia School of Business as a business case for this project. The study concerns speed dating and its user's attributes and preferences of partners. The project extrapolates this study into a potential business opportunity whereby we can offer different services to current and new users.

Data collected includes demographic information, participants preferences, self evaluation and decision factors. In total there are 195 attributes and 8378 items in total.

For our project, we are trying to tackle three business problems for online dating experience.

First of all, we predicted whether or not two people on one date will elect to go on another date with each other by using classification model. Deep learning algorithm worked best for this classification.

Secondly, we provide a list of users from the pool of daters who will have high likelihood of matching with a particular user. To achieve this we applied k-means clustering to identify different groups of people based on their preferences & attributes. Further, we studied the dynamics of matches between and within clusters to come up with recommendations. To solve this problem we also developed a classification model using Deep learning algorithm to classify new users into a cluster.

Finally, we wanted to provide a premium service to customers that would indicate the attributes they may need in order to match with someone fitting their defined preferences. For this we used gap analysis of our clusters attributes and the user's attributes to come up with recommendations.

The report discusses the business idea, business problems, solution to each problem, project limitations, improvements and key takeaways.

Business Idea

Singles make up the majority of adults in America. According to 2015 census data, 107 million of American age 18 or above are reported single, which is 45% of the population. In addition, research done by the Pew Research Center points out that most people still find marriage ideal that is they still want to find their significant others. Hence, we would like to leverage this as an opportunity by helping people not only find their partners but their perfect match through providing premium matching advice services. The premium service will generate revenues for the business.

The idea will be implemented through a website portal where anyone above 18 yrs of age can be a member by creating a free account. The following information will be required from new members while creating an account.

- Personal Information: Location, education, career, income, interest in 16 activities (sports, museums, art, clubbing, hiking, etc)
- Preference for attributes in partner: 6 attributes (attractive, ambitious, intelligent, sincere, shared interests)
- Individual's rating on the 6 attributes

Once people become a member of our website portal, they can send dating request to other people for free. If someone accepts the request, the two can go on a date. However, we expect that some people may not be able to get a date for whom we will have the paid premium service to help them get a partner.

Business Problems

The business needs to solve 3 problems to be able to implement the business idea. The problems are explained below.

1. Predict if two people will go on a date with each other. That is, if two people happen to accept their request for a date, we should be able to predict if they will both like each other or not.
    a. User benefits from insight about whether to pursue the match
    b. Business benefits from additional data on matches to improve the premium service model
2. Model a recommendation system to suggest probable dates for a user i.e. who all in the pool of our members will have a high likelihood of matching with a given member.
3. Provide premium matching advice for users to know what they need to improve in order to match with someone fitting their defined preferences.

Data

To answer the above questions, we gathered a dataset from Columbia University's School of Business which consists of 8379 rows of data each representing a different 4 minute date between 552 unique individuals found in one csv file. (Speed Dating Data.csv (362.28 KB)). The data has 195 different attributes all defined and described here: (Speed Dating Data Key.doc (158 KB)). There is also a sample of the questionnaire given to each participant in the attached document to provide further context for the data.

Out of 195 attributes, we used 10-50 to solve each problem. The data consists of following types of attributes.

- User Preferences: What do they want in their partner. Rate the 6 personality traits on varied scale (0-10/ 0-100). The 6 attributes are attractive, intelligent, ambitious, fun, shared interests
- User Attributes: Individual's own rating on the above 6 personality traits, their rating on 16 hobbies (art, movies, hiking, exercise, etc) based on their interest in them.
- User Demographics: This includes attributes like individual's location, age, income, career, etc
- Results of each date conducted: Scores given to each other on the 6 personality traits, decision (whether they liked the other or not)

Solution

Problem 1: Predict Match using Classification

Predict if two people will go on a date with each other. That is, if two people happen to accept their request for a date, we should be able to predict if they will both like each other or not.

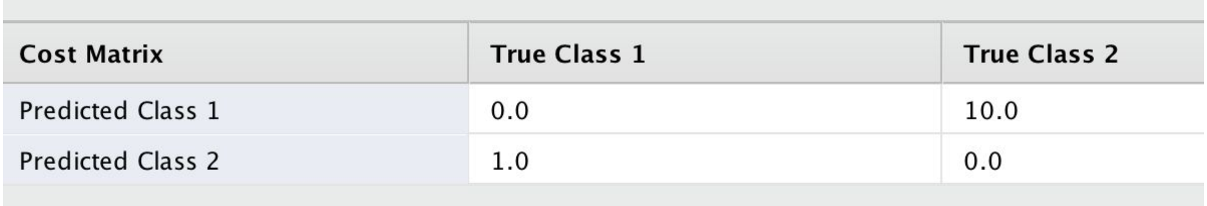
We have tried to solve it using classification technique of data mining.

Data Preparation:

- The scales of rating were different, so we normalized them
- Created a column called Match with value 1 if both the persons on a date has decision as 'yes', i.e, they both would like to date the other, else the value of match is 0.
- Converted numerical attributes to polynomial
- Used stratified sampling to split the data into training and test set.

Attributes selected:
Predictors: Selected a subset of attributes (demographic details of the individuals going on a data, their preferences about the attributes they look for in their partner)
Label: Match

Process: Since the data was skewed towards no-match (0) class and the important class was match class (1), we added a cost matrix to control the results of confusion matrix.
Cost Matrix used: First class = No Match (0), Second Class = Match (1)



Cost Matrix	True Class 1	True Class 2
Predicted Class 1	0.0	10.0
Predicted Class 2	1.0	0.0

Classification methods applied: Decision trees, Naive Bayes, CHAID, Random Forest, Deep learning

See Appendix A for the modeling process and results of Deep learning, Naive Bayes and Decision Tree Models.

Best Performance by Deep Learning Model: From the Deep learning results in Appendix A, we can see that the model's accuracy is high given high precision and recall for the Class 1. The lift chart also suggests that the model is able to predict the important class (Match = 1) early on. Hence, this model has done the classification with high confidence.

Since our data was skewed towards a class, we also tried undersampling to balance our data and applied multiple classifiers. Decision tree worked best on the balanced data, however, the deep learning model discussed above still proved to perform better.

Problem 2: Find Matching Clusters

Model a recommendation system to suggest probable dates for a user i.e. who all in the pool of our members will have a high likelihood of matching with a given member.

We have tried to solve this problem using two step method.

1. **Clustering:** Find groups of people by using clustering based on their preferences & attributes; check if matches are intra-cluster or inter-cluster to identify which clusters match with each other
2. **Classification:** Create a model to predict the cluster an individual belongs to. Once we have the cluster of an individual, we can know people from which cluster should be recommended to this person.

Clustering Model:

Attributes selected: Person's preferences and the individual's features
Cluster method used: k-means with k = 3
Other clustering methods like Hierarchical clustering, SVM clustering etc, were used too but they were not able to perform better than k-means.

See Appendix B for modeling process and the results of k-means clustering.

Important attributes are the ones that differentiate between clusters. For this model, important attributes are from (individual's location), attr1_1 (individual's preference for partner's attractiveness), amb1_1 (individual's preference for partner being ambitious).

See Appendix C for more details on the differences in preference attributes among clusters.

We also tried this model to include other attributes about individuals activities like sports, tv, yoga, museums etc that shows their interest is such activities. However, we saw that the clusters had almost similar values for these attributes and it did not make any difference to clustering results. For the ease of representation, we removed those attributes to be considered for clustering model.

By looking at the clusters and the matches between and within them, we came up with the following recommendation rules.

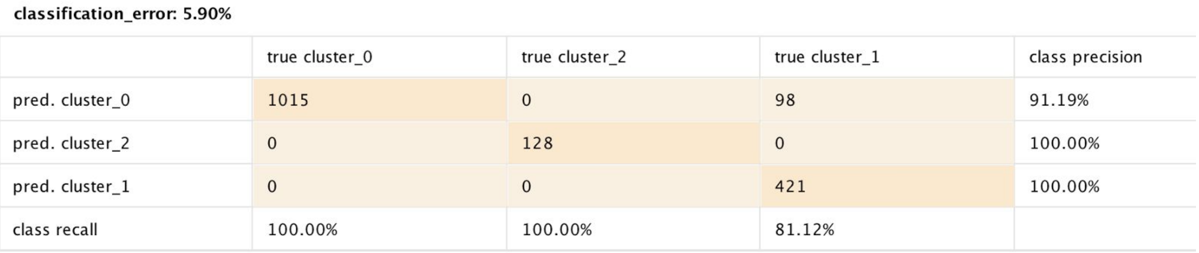
Rule 1: If a person belongs to Cluster 0, we recommend people from Cluster 0
Rule 2: If a person belongs to Cluster 1, we recommend people from Cluster 0 and Cluster 1
Rule 3: If a person belongs to Cluster 2, we recommend people from Cluster 0

See Appendix D for detailed process about formulating these rules.

Classification Model:
The second part of the solution was to create a classification model to classify people into a cluster identified in previous step.

To do this we mapped the individual's with their cluster (as identified by the clustering model) in our data sheet and ran a few models and found decision tree worked best with 5.9% classification error but the data had a lot of duplicates since an individual went on multiple dates.

Result of Decision Tree Model

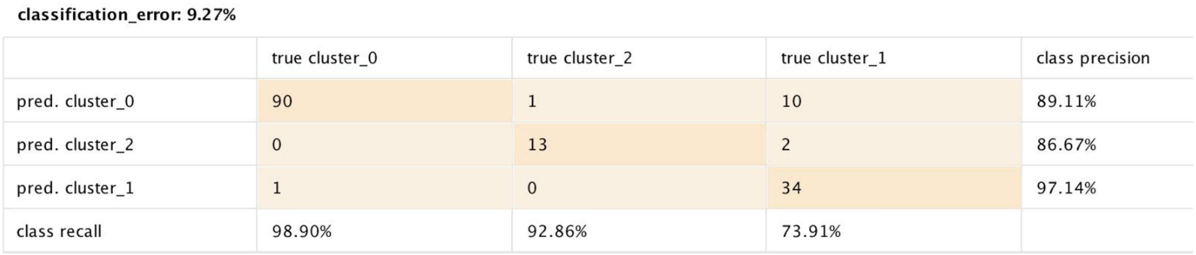


classification_error: 5.90%

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	1015	0	98	91.19%
pred. cluster_2	0	128	0	100.00%
pred. cluster_1	0	0	421	100.00%
class recall	100.00%	100.00%	81.12%	

After removing the duplicates, we ran the model again and it did not work so we tried it with Deep Learning and other classification models. Deep learning worked the best with 9.27% classification error.

Result of Deep Learning Model



classification_error: 9.27%

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	90	1	10	89.11%
pred. cluster_2	0	13	2	86.67%
pred. cluster_1	1	0	34	97.14%
class recall	98.90%	92.86%	73.91%	

From this confusion matrix we can see that the model is able to identify people into Cluster 0 with high accuracy followed by Cluster 2 and Cluster 1.

This further validates our earlier finding about the differences in Clusters that people in Cluster 1 cannot be differentiated from those in Cluster 0 due to very less inter cluster distance.

Problem 3: Recommendations for Improvement

Provide premium matching advice for users to know what they need to improve in order to match with someone fitting their defined preferences.

With this we are trying to generate revenue for the business by providing a paid premium service. This service is targeted to individuals who are not able to find partners who match their preferences.

One possible reason for a person to not match with the people who satisfy the preferences criteria of this person is that this person does not have the attributes that his/ her preferred matches look for in their partner.

Hence, we tried to solve this problem by finding out the gap between the attributes of the premium service user and the average of attributes of people in the cluster who matches with the cluster this person belongs to.

Evaluation Process:
Step 1: Identify the cluster of Premium Service User
Step 2: Identify the cluster (s) that matches with the cluster of Premium Service user
Step 3: Find the average of differentiating attributes for the matching Clusters
Step 4: Find the percentage gap between the attributes of the user and the clusters
Step 5: If the percentage gap is above 10%, we recommend improvement on that attribute

For clarity, we have shown the solution through an example in Appendix E.

Limitations

Limitations of data:

- Imbalanced Data: Our data was skewed towards No match class (802 records as match, 4034 as no-match in training data). Since most models do not work well with imbalanced data, we could not leverage the advantages of many models.
- Sample did not represent population: Since our data was collected by conducting dates among Graduate Students, it was not diverse and did not represent the general population. Due to this, some Personal Information attributes (Age, Career, Income, Hobbies) were very similar which could have been important factors had the data been diverse.
- Data Loss due to undersampling: As the data was only about 8000 records, applying undersampling to balance the data led to a loss of huge data and we could not base our decisions on the smaller data.

Limitations of Rapidminer: For data with skewed classes, Hellinger Distance based decision trees has been proven to work well but Rapidminer does not have this function.
See Appendix F for details about limitations.

Improvements

- Collecting larger and more diverse sample can improve the model and its usability to a great extent
- The dataset uses the individual's own scores on different attributes which is not a good measure. Hence, scoring people on different attributes based on a uniform tool like Personality Tests will be beneficial
- The problems can be better solved by using other tools for modeling like Python so that we can apply more models and explore the data extensively

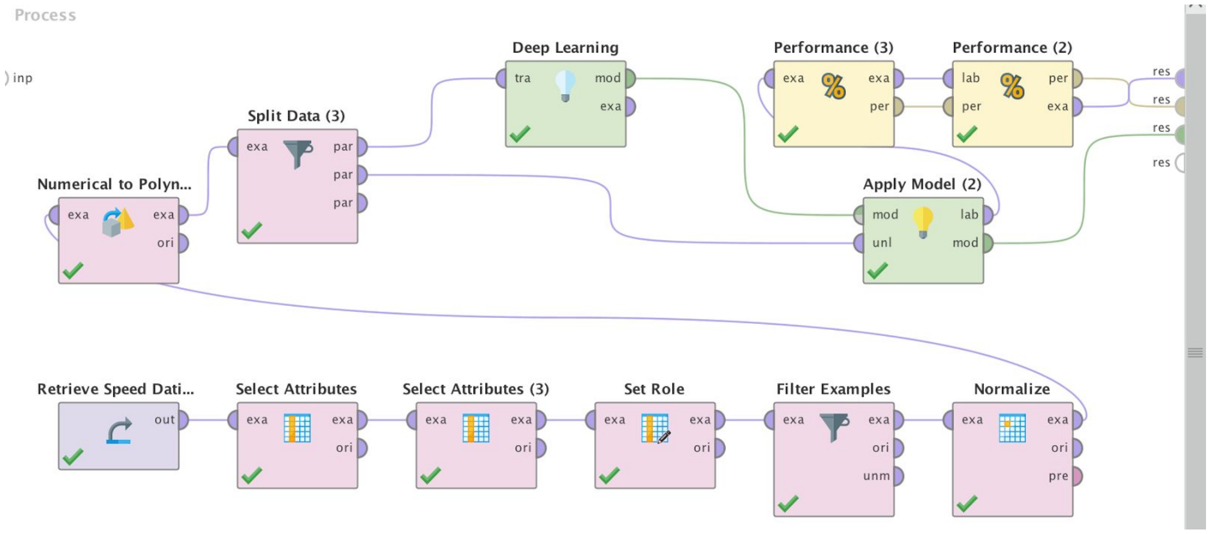
Key Takeaways

- Understanding the business problems and data is imperative to any analysis
- Well thought through data preprocessing can save a lot of time
- A problem can be solved by multiple models so we should try all applicable models before suggesting the best model
- Solution to some problems may be a simple analysis instead of application of a model
- Clustering and Recommendation Decisions can be subjective

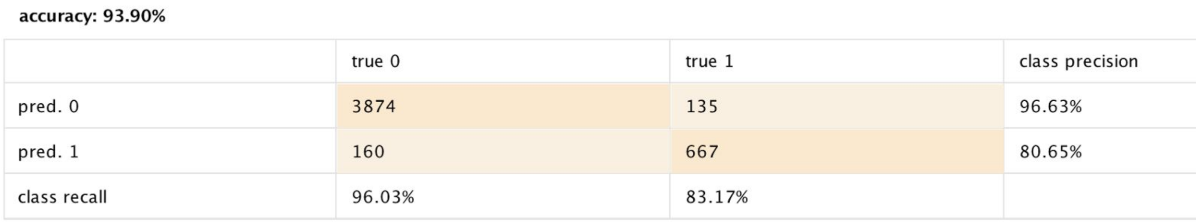
Appendices

Appendix A

Modeling Process used to solve Problem 1 to predict if two people going on a date will like each other or not.:

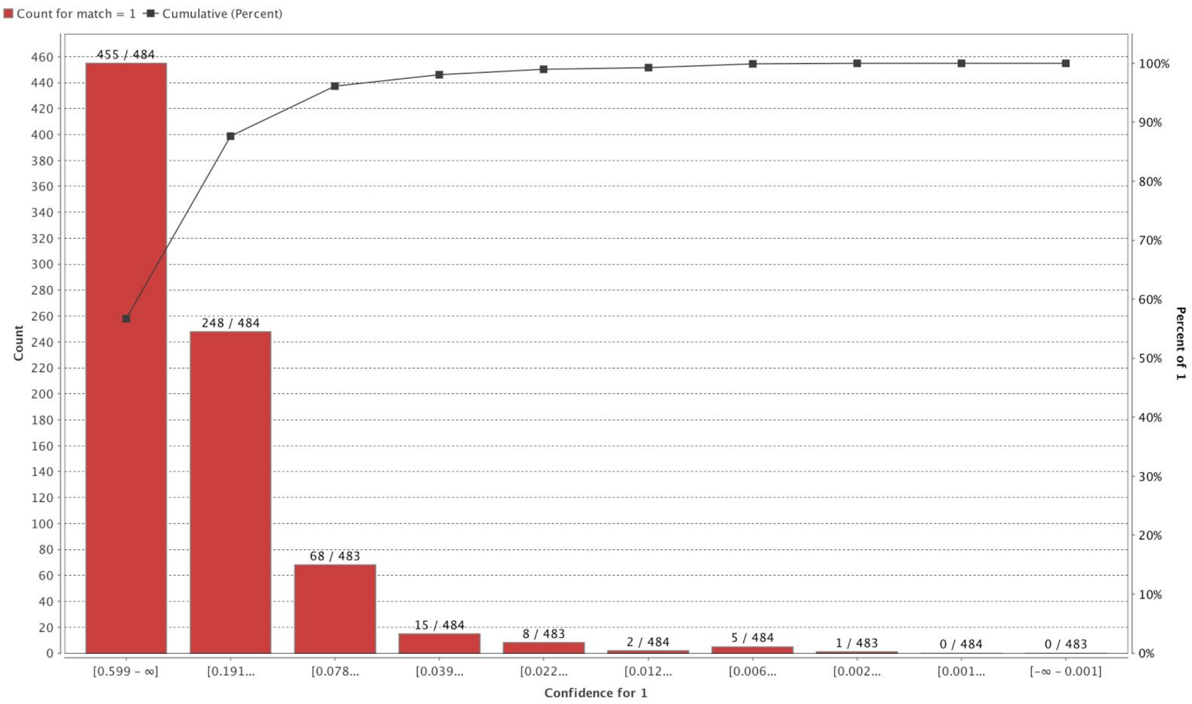


Results for Deep Learning model

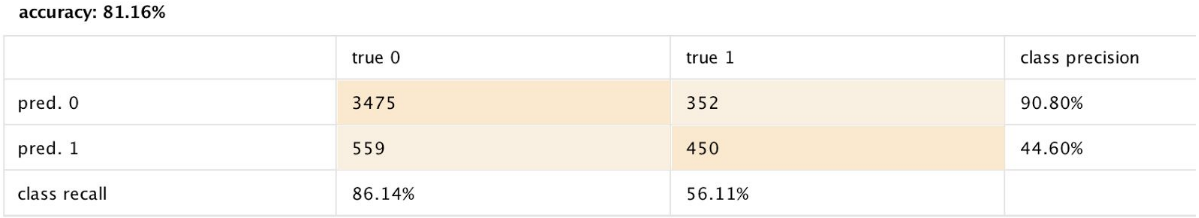


accuracy: 93.90%

	true 0	true 1	class precision
pred. 0	3874	135	96.63%
pred. 1	160	667	80.65%
class recall	96.03%	83.17%	

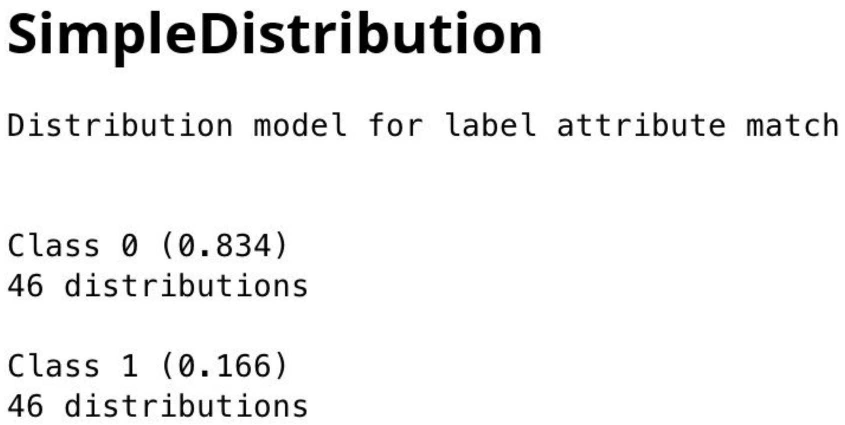


Results for Naive Bayes Model:



accuracy: 81.16%

	true 0	true 1	class precision
pred. 0	3475	352	90.80%
pred. 1	559	450	44.60%
class recall	86.14%	56.11%	

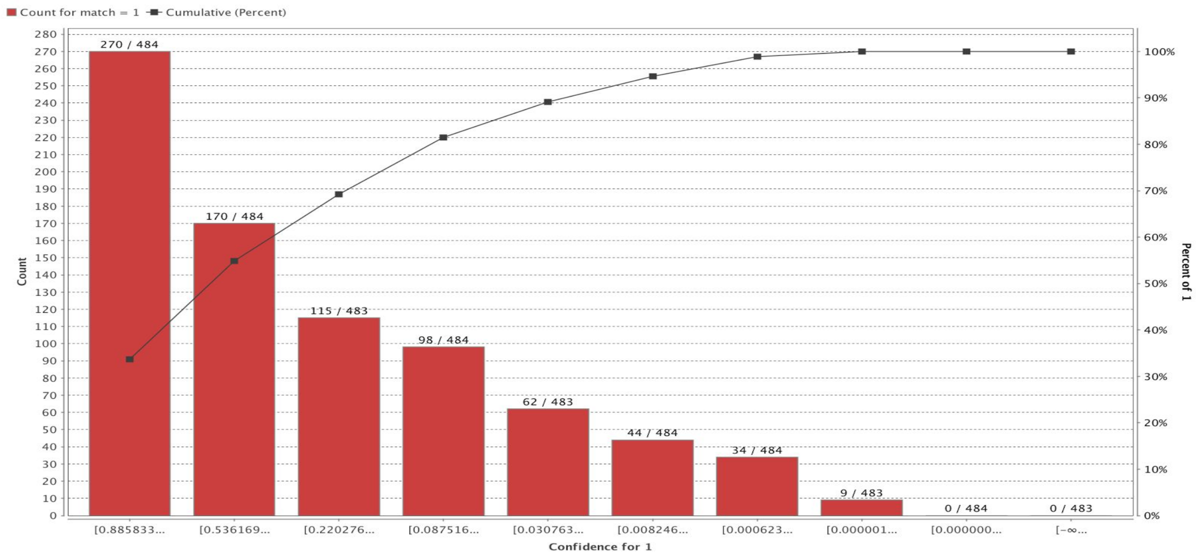


SimpleDistribution

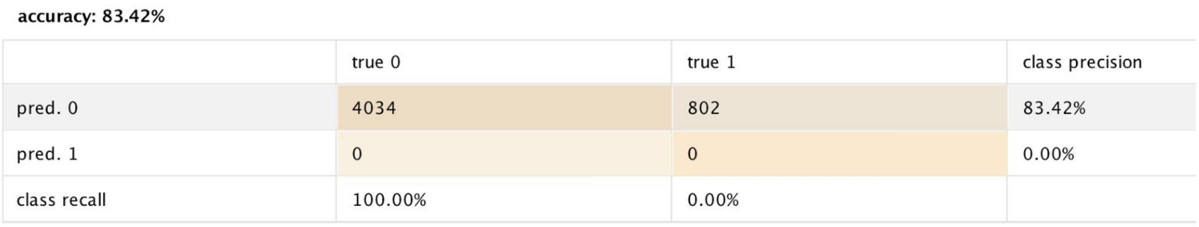
```
Distribution model for label attribute match


Class 0 (0.834)
46 distributions

Class 1 (0.166)
46 distributions
```



Worst Performing Model: Decision Trees as it cannot work on skewed data
It predicted every record to a non-match (the highly represented class in dataset)

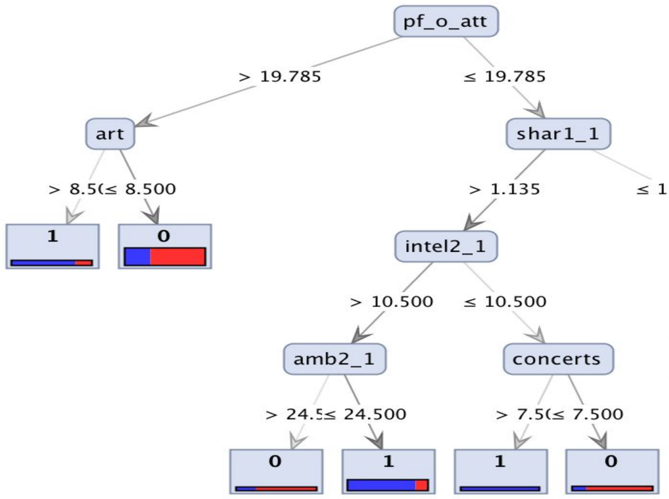


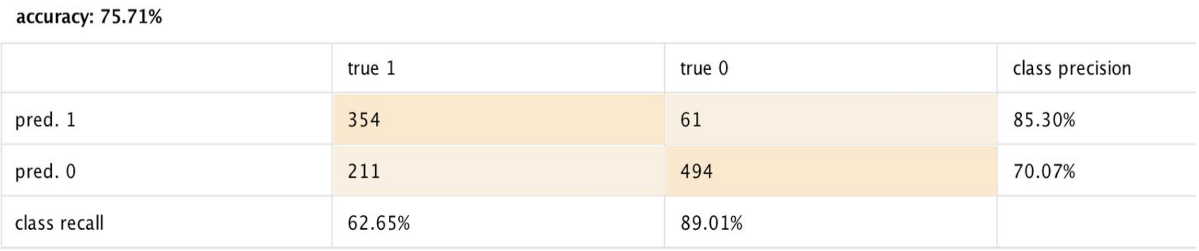
accuracy: 83.42%

	true 0	true 1	class precision
pred. 0	4034	802	83.42%
pred. 1	0	0	0.00%
class recall	100.00%	0.00%	

However, when we balanced data by random undersampling the records of over represented class i.e. no-match (matches = 0), decision tree seems to work better with purity criteria as accuracy.

Below are the results for decision tree with balanced data.



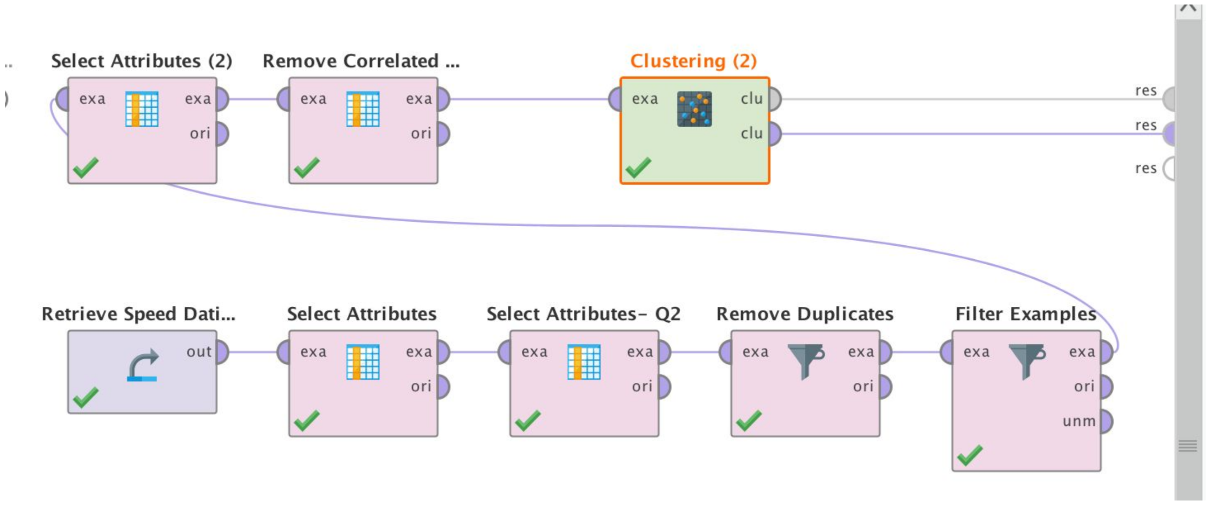


accuracy: 75.71%

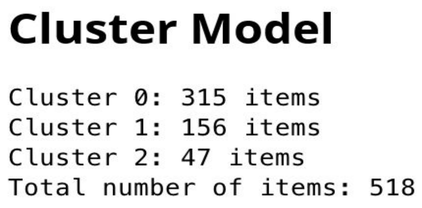
	true 1	true 0	class precision
pred. 1	354	61	85.30%
pred. 0	211	494	70.07%
class recall	62.65%	89.01%	

Appendix B

Modeling Process used to identify clusters to solve Problem 2



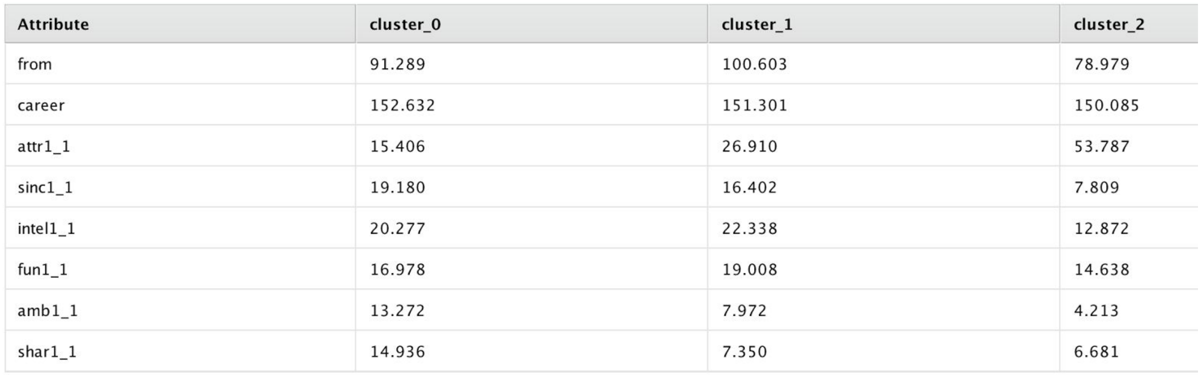
Results of k-means clustering used to solve Problem 2.



Cluster Model

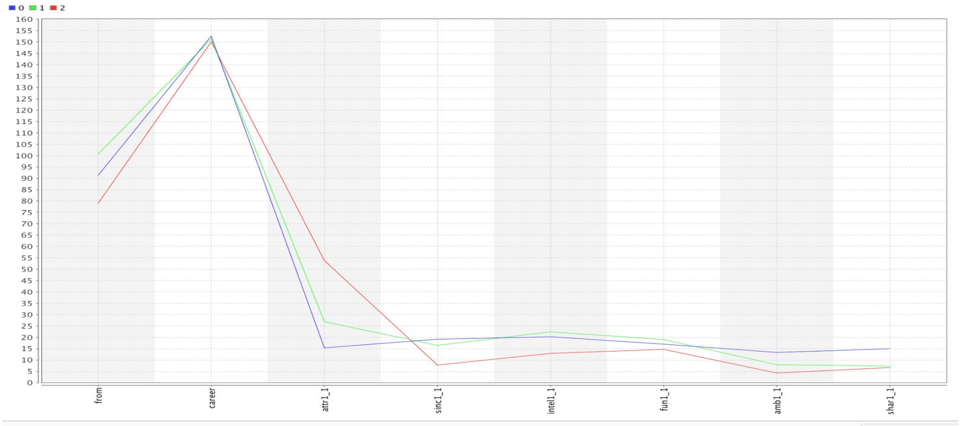
```
Cluster 0: 315 items
Cluster 1: 156 items
Cluster 2: 47 items
Total number of items: 518
```

Centroid table



Attribute	cluster_0	cluster_1	cluster_2
from	91.289	100.603	78.979
career	152.632	151.301	150.085
attr1_1	15.406	26.910	53.787
sinc1_1	19.180	16.402	7.809
intel1_1	20.277	22.338	12.872
fun1_1	16.978	19.008	14.638
amb1_1	13.272	7.972	4.213
shar1_1	14.936	7.350	6.681

Plot of centroid Table



Other Visualizations for clusters

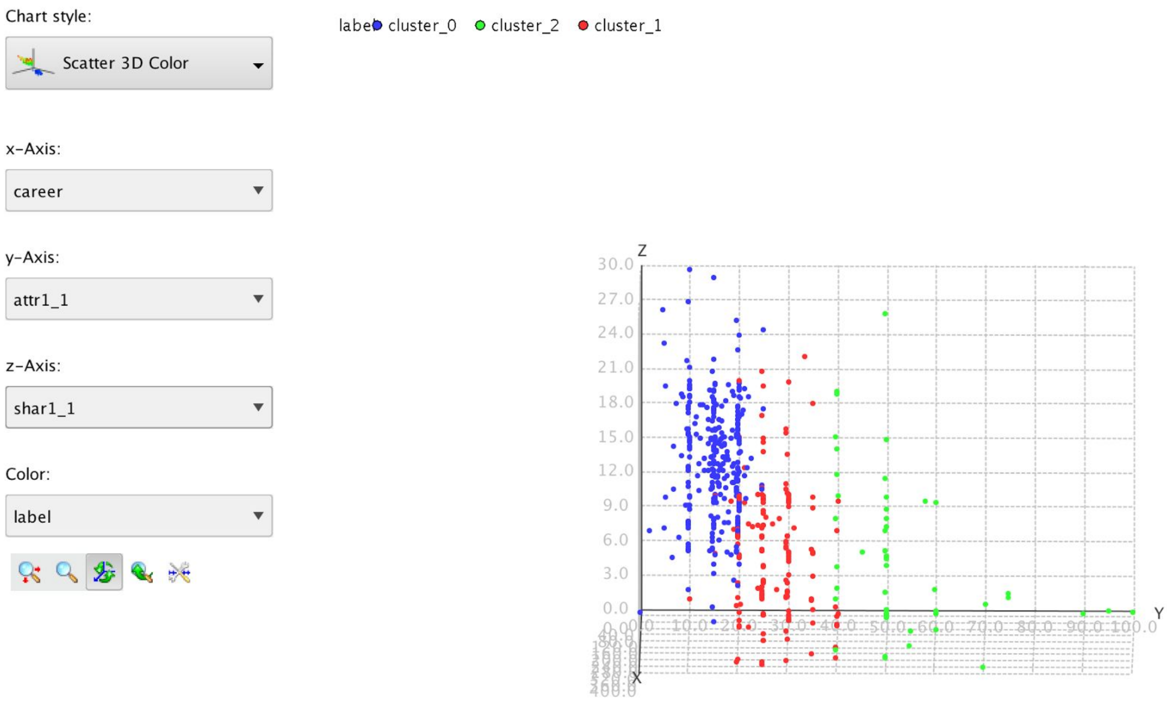


Chart style: Scatter 3D Color

x-Axis: career

y-Axis: attr1_1

z-Axis: shar1_1

Color: label

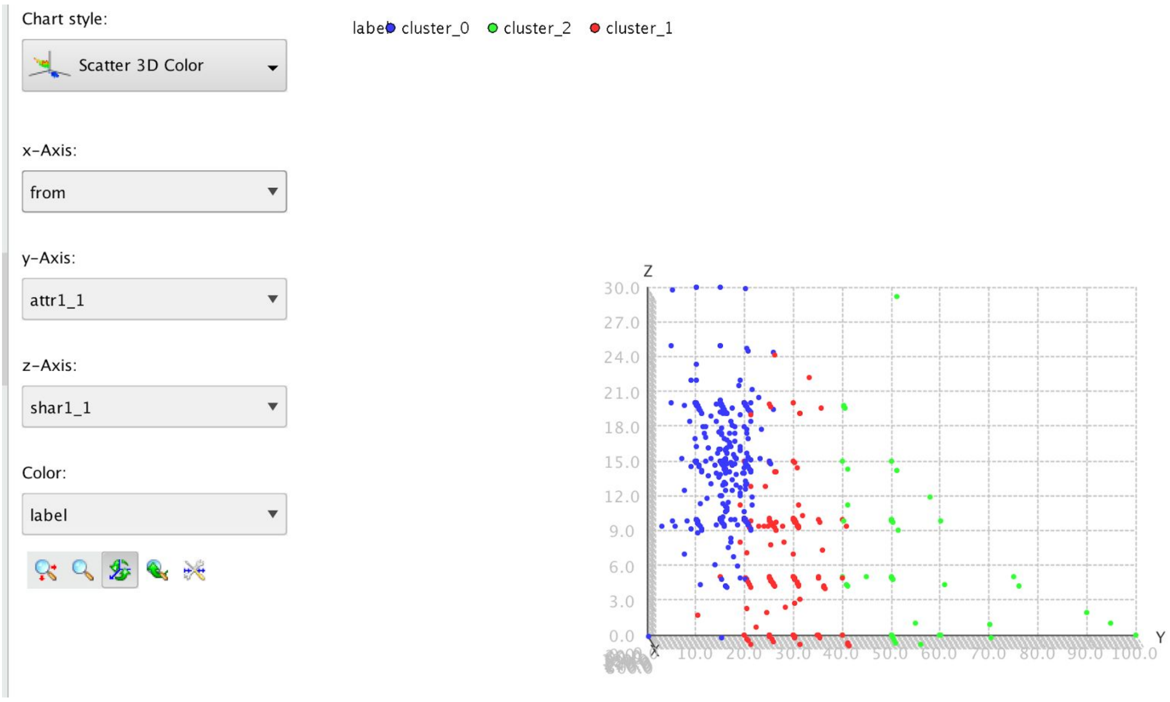


Chart style: Scatter 3D Color

x-Axis: from

y-Axis: attr1_1

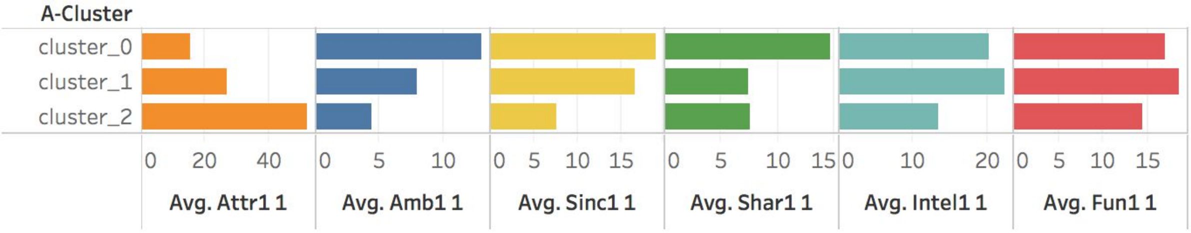
z-Axis: shar1_1

Color: label

Both the graphs show that Cluster_0 and Cluster_1 are close enough while Cluster_0 and Cluster_2 are far apart. Even Cluster_1 and Cluster_2 can be identified separately.

Appendix C

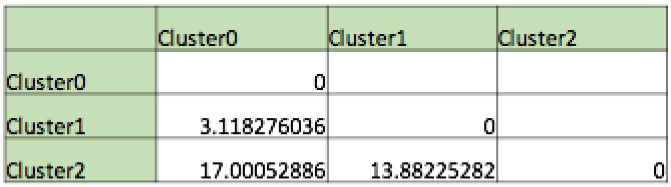
Graph showing the differences in average rating of each personality trait the clusters prefer in their partners.



It shows that the two main attributes that can be used to differentiate between clusters are Attractive and Ambitious.
To differentiate between cluster 1 and 2, sincerity and intelligence can also be used.
Fun seems to be the least important attribute as it is not able to differentiate any cluster clearly.

Appendix D

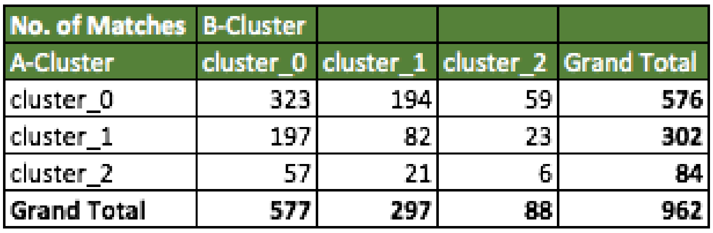
Explanation of the process to come up with the cluster matching rules as a solution to Problem 2.
The clustering process explained in Appendix B gave us the clusters. When we matched the clusters by percentage of matches, we found that the percentages are very similar. Since, we are not able to identify which cluster match with which other cluster we wanted to see if our clusters are good. So we checked the inter cluster distances.



	Cluster0	Cluster1	Cluster2
Cluster0	0		
Cluster1	3.118276036	0	
Cluster2	17.00052886	13.88225282	0

This table shows that inter-cluster distance between Cluster 0- Cluster 2 and Cluster 2- Cluster 1 are comparatively higher than Cluster 0- Cluster 1, so our model should be able to distinguish between Cluster 0 and Cluster 2 but may confuse records from Cluster 1 to belong to Cluster 0 or Cluster 1.

We then decided to look at absolute number of matches and found out the following number of matches between clusters.



No. of Matches	B-Cluster			
A-Cluster	cluster_0	cluster_1	cluster_2	Grand Total
cluster_0	323	194	59	**576**
cluster_1	197	82	23	**302**
cluster_2	57	21	6	**84**
Grand Total	**577**	**297**	**88**	**962**

Based on the above table we decided the following rules for recommendation according to the highest values. For Cluster 1 we recommend Cluster 1 also as it was highest according to the percentages.

Rule 1: If a person belongs to Cluster 0, we recommend people from Cluster 0
Rule 2: If a person belongs to Cluster 1, we recommend people from Cluster 0 and Cluster 1
Rule 3: If a person belongs to Cluster 2, we recommend people from Cluster 0

Appendix E

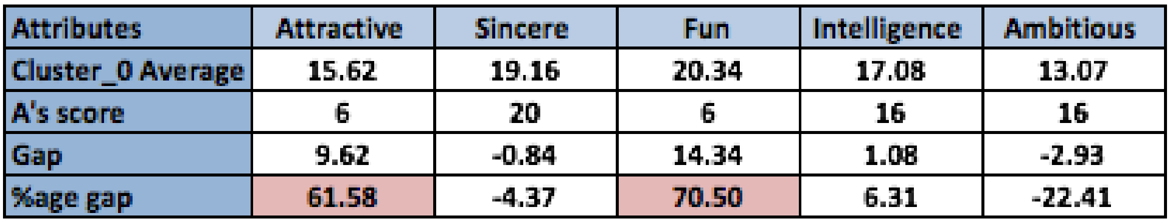
Example of the Premium Service recommendation system (Problem 3).

We randomly picked an individual from the dataset who had less number of matches.
Person A: Identified as Global ID (iid) 78 in the dataset

Step 1: Identify the cluster of Premium Service User
Person A belongs to Cluster_0.
Whenever a new person creates an account on our website, we classify them into a cluster based on the model used in Problem 2.

Step 2: Identify the cluster (s) that matches with the cluster of Premium Service user
According to the conclusion of Problem 2, people in Cluster_0 matches with people in Cluster_0.

Step 3: Find the average of differentiating attributes for the matching Clusters
Step 4: Find the percentage gap between the attributes of the user and the clusters



Attributes	Attractive	Sincere	Fun	Intelligence	Ambitious
Cluster_0 Average	15.62	19.16	20.34	17.08	13.07
A's score	6	20	6	16	16
Gap	9.62	-0.84	14.34	1.08	-2.93
%age gap	61.58	-4.37	70.50	6.31	-22.41

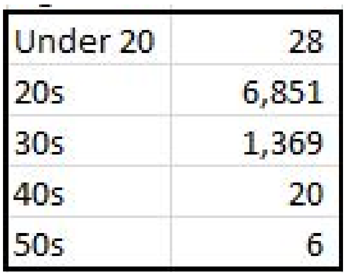
Step 5: If the percentage gap is above 10%, we recommend improvement on that attribute
From the table in Step 4, we found that Person A needs to improve on two attributes: Attractive and Fun

Recommendations:

- Try to be more presentable.
- Improve on interpersonal skills so that people perceive you as a fun person
- Take on some interesting hobbies

Appendix F

During our data mining process we discovered that there is a limitation in attribute variance. For instance, the participants were vastly in their 20’s and 30’s and only six participants were in 50’s as shown in below. Moreover, participants in similar age groups also shared similar habits such as leisure activities, watch tv and sports etc.. Therefore, when we were creating the clusters many attributes were not effective in differentiating participants.



Under 20	28
20s	6,851
30s	1,369
40s	20
50s	6

Moreover, this empirical study was used to support or reject hypotheses in traditional statistical techniques. Statistical methods typically do not scale well to very large data sets. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population. On the other hand data mining methods are suitable for large data sets. In fact, data mining algorithms often require large data sets for the creation of quality models. As mentioned before, our dataset is unbalanced, mostly no-match (over 6000 observations of total 8374 records), that hindered decision tree analysis. In fact classification models do not work well with skewed datasets. Undersampling that is removing samples from the majority class (no-match) using an undersampling algorithms e.g.One Sided Selection (OSS), Edited Nearest Neighbors (ENN). Random Undersampling was adopted in our project.