

DSC 680

Inman, Gracie

Project 3 White Paper

06/01/2024

### Deciding to Date

According to Clara Ludmir, in the United States people were expected to spend \$25.8 billion on Valentine's Day in 2024 (2024). Most people want to find love and sometimes even turn to apps to find someone. According to the Pew Research Center, "Three in ten U.S. adults say they have ever used a dating site or app" (Vogels & McClain, 2023). According to Yahoo! Finance, "The Brainy Insights estimates that the USD 7.2 billion online dating services market will reach USD 10.8 billion by 2032" (Yahoo!, 2024). What do people look for in someone to date?

The speed-dating dataset was obtained from Kaggle (Mexwell, 2024). The speed dating dataset has several attributes such as gender, age, income, life goal, career, whether it was to decide to match with that person, and ratings on attractiveness, sincerity, intelligence, and how fun the person was. Using these attributes, the goal was to build a model that will be able to determine if the person was matched. In addition, I wanted to determine what is the most important feature in deciding if someone will be matched. The attributes within the dataset are ratings on how attractive, sincere, intelligent, fun, ambitious, if they have shared interests/hobbies, how much do you like this person (1=don't like at all, 10=like a lot), how probable do they think it is that this person will say 'yes' for them (1=not probable, 10=extremely probable), and if they have met this person before.

Before the model was able to be built, the data had to be cleaned. The dataset was visualized and checked for missing values. There were a significant number of values missing out of the 8,378 rows as shown in Table 1 below. The missing values were filled using the mean. After filling in the mean, there were only 89 missing values in the career column that were missing. These values were dropped. After this, the data was split into training and testing for building the models.

Missing values in each column:	
gender	0
age	95
income	4099
goal	79
career	89
dec	0
attr	202
sinc	277
intel	296
fun	350
amb	712
shar	1067
like	240
prob	309
met	375

*Table 1: Shows the number of missing values in each column.*

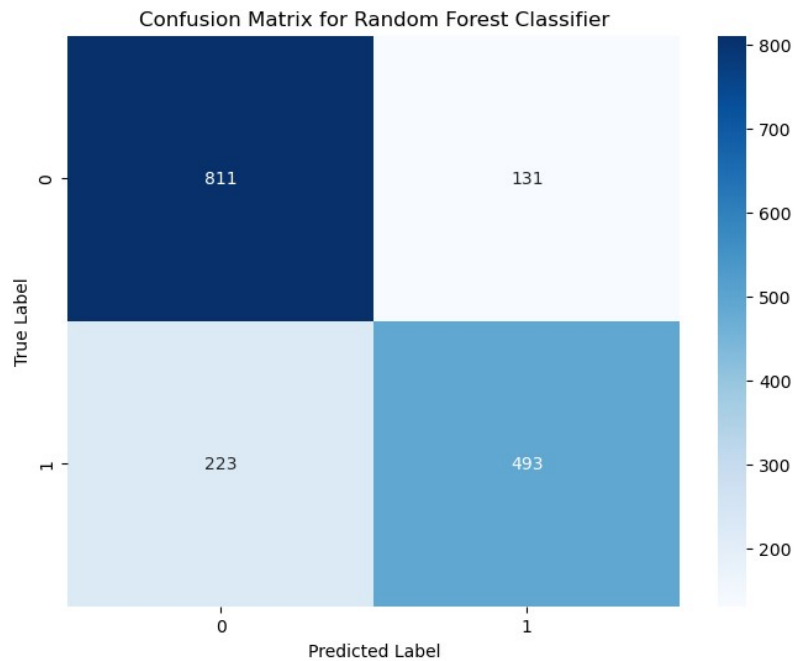
For this project, I have decided to train random forest, KNN, support vector machine classifier, gradient boosting machine, and Naïve Bayes. The models were evaluated based on

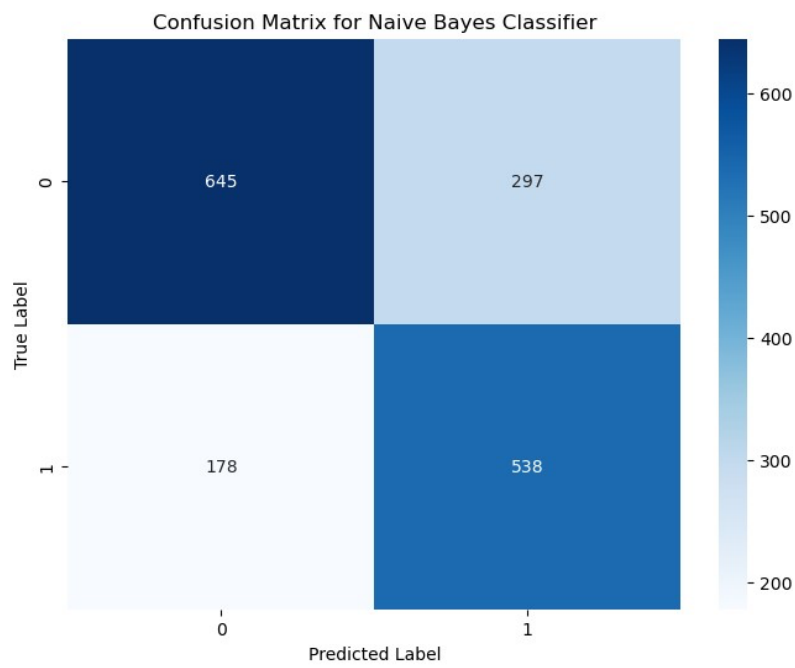
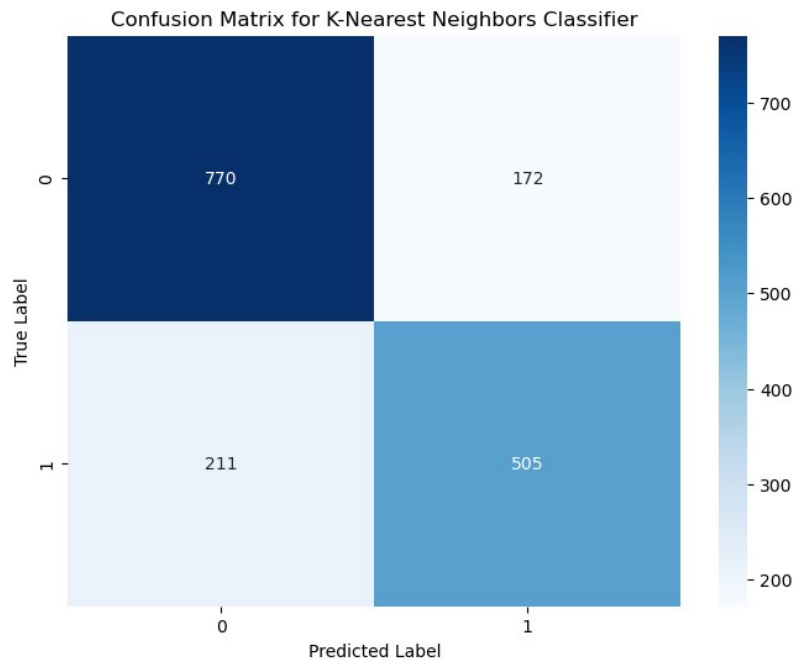
accuracy, precision, RMSE, and confusion matrices. The accuracy, precision, and RSME are shown below in Table 2. The results from the confusion matrix are shown below in

Figures 1-5.

Model	Accuracy	Precision	RMSE
Random Forest	0.786490	0.790064	0.462072
KNN	0.768999	0.745938	0.480626
Naïve Bayes	0.713510	0.644311	0.535247
SVM	0.568154	0.000000	0.657150
GBM	0.817853	0.805310	0.426787

*Table 2: Shows the accuracy, precision, and RMSE scores for each model.*





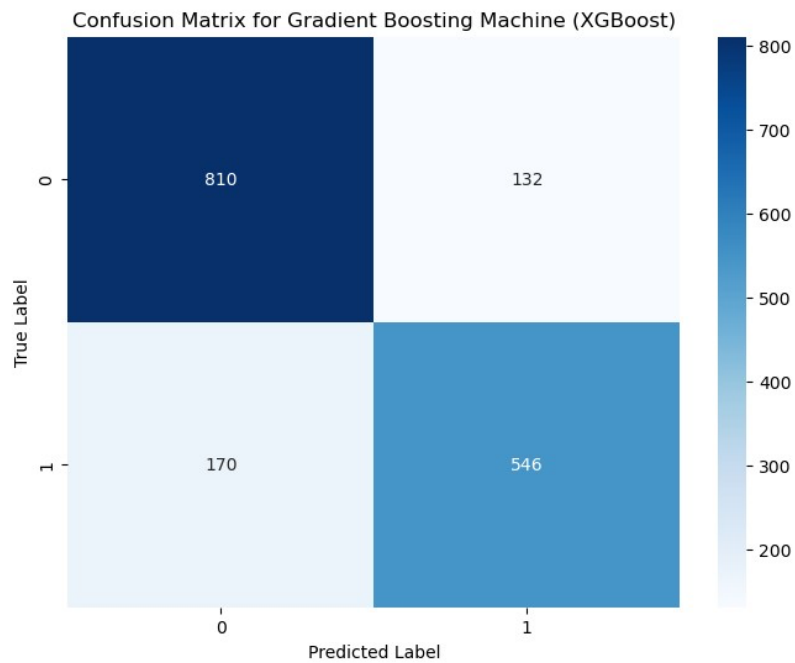
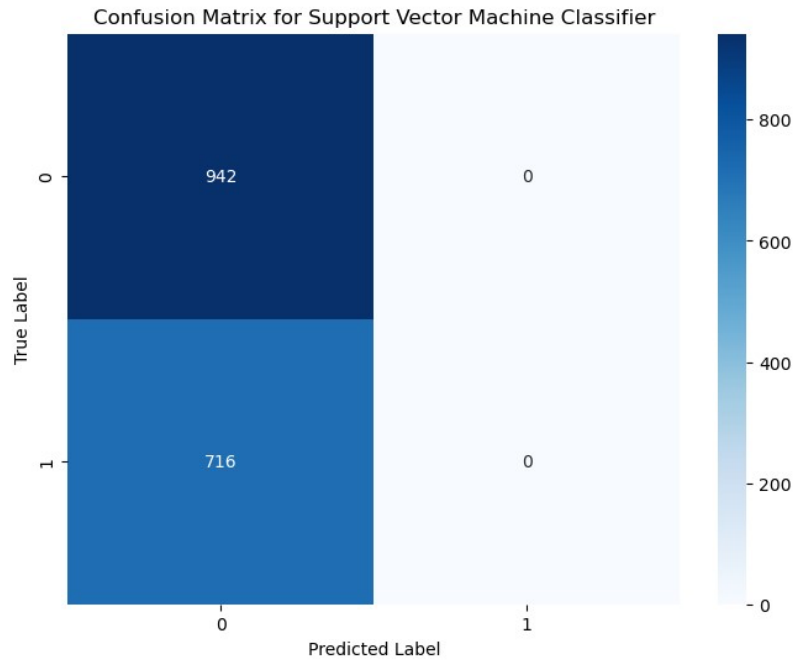
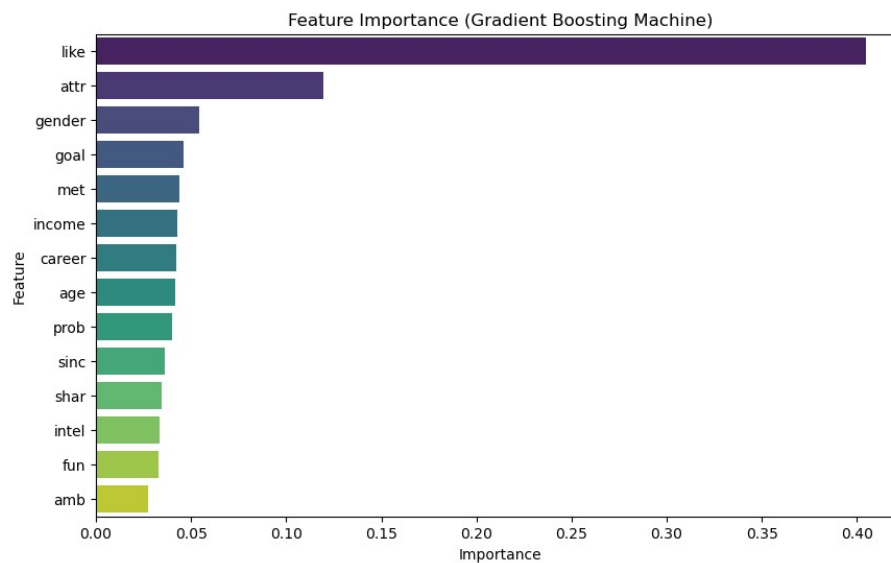


Figure 1 A-E: Shows the confusion matrix for each model.

The best model was determined by looking at all of the metrics together. Based on the confusion matrix and evaluation metrics, the best model for the dataset is the gradient

boosting machine. This was determined by leveraging all the evaluation metrics for each model and determining what worked best for the data. It had the lowest RMSE, but the highest precision and accuracy. It also performed the best when looking at the confusion matrix with the highest number of true positives and negatives. Shown below in Figure 2 is the feature importance of the model. This shows the most important feature for the model is how much the subject likes the other person followed by attractiveness.



*Figure 2 Shows the feature importance for the chosen model.*

A challenge with this data was the number of missing values. The decision was to fill the numeric values and drop the remaining missing values. By filling the missing data, I am assuming that the filled values accurately represent the population in the study. This also assumes that the model will perform well on new non-populated data.

There are limitations of this data that do need to be considered. One limitation of the data is that is solely based on the perspective of the participant. Another limitation is the quality of the

data that was mentioned previously. The assumptions mentioned previously can also be seen as limitations.

Ethical concerns are always important to consider. There are little to no ethical concerns with this data. The data was sourced from a study conducted by Columbia University and has no identifying information. However, a challenge of this dataset is there appearing to be a large number of missing values. This could lead to other ethical concerns based on how I handle the missing data.


Based on the current findings, the model is not currently ready for deployment. This is because while the model performed well, a significant amount of data was missing. The model needs to be tested on new, non-populated data before deployment. While GVM performed the best on the current tests, random forest was not far behind and could perform better on new unseen data. My current recommendation is to test both models with new data and confirm current findings before proceeding.

There are multiple future applications based on this model. One application is to use past ratings on current participants to see if they are likely to be matched. Another application would be to create a new dating site based on the data and have people evaluate the other person (anonymously) after the date to allow for better matching in the future.

#### References:

Ludmir, C. (2024, February 19). *Valentine's Day spending to hit \$26 billion, as "experience gifts" grow in popularity*. Forbes.

<https://www.forbes.com/sites/claraludmir/2024/02/13/valentines-day-spending-to-hit-26billion-as-experience-gifts-grow-in-popularity/?sh=74a7096f74e3>

Mexwell. (2024, April 15).  *speed dating*. Kaggle.

<https://www.kaggle.com/datasets/mexwell/speed-dating>

Vogels, E. A., & McClain, C. (2023, February 2). *Key findings about online dating in the U.S.*

Pew Research Center. <https://www.pewresearch.org/short-reads/2023/02/02/key-findingsabout-online-dating-in-the-u-s/>

Yahoo! (2024, January 8). *Online dating services market size surpassing USD 10.8 billion by 2032, growing at projected 4.2% CAGR*. Yahoo! Finance.

[https://finance.yahoo.com/news/online-dating-services-market-size-140000598.html?guccounter=1&guce\\_referrer=aHR0cHM6Ly93d3cuZ29vZ2x1LmNvbS8&guce\\_referrer\\_sig=AQAAAKiLoJnk28HaZxYUmJcWcVAmylkSdiPiS9TE5crarPQe7IjbOE7sMiurDNFtdl5I-ppRcxvl-tJZdFPqzUTkJLLTRC7Dcwa7nkYPVyVWUO-Rya5WFnu7iLPf9aAoLdIJ8MWf71oQqF4xyD1sfBaIR0liU33f0NlzKm1po80bch](https://finance.yahoo.com/news/online-dating-services-market-size-140000598.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2x1LmNvbS8&guce_referrer_sig=AQAAAKiLoJnk28HaZxYUmJcWcVAmylkSdiPiS9TE5crarPQe7IjbOE7sMiurDNFtdl5I-ppRcxvl-tJZdFPqzUTkJLLTRC7Dcwa7nkYPVyVWUO-Rya5WFnu7iLPf9aAoLdIJ8MWf71oQqF4xyD1sfBaIR0liU33f0NlzKm1po80bch)

## Appendix

### A. Questions

1. How does the survey reported data affect the analysis?
  - a Survey data is not the most accurate data. In this case, survey data relies on people's feelings and opinions which vary from person to person and can make it difficult to perform an accurate analysis.
2. Why did you fill the data?
  - a There was to many missing values to remove the data. I filled it with the average values.
3. Why did you get rid of data after trying to fill it?



- a There were 83 rows with missing values of occupation. I did not want to fill the occupation in case it was an important factor. Since it was only 83 rows I removed them.
4. Why did you choose GVM if random forest also had good evaluation metrics?
- a This is because in this case it was important to look at all the evaluation metrics together and decide which is the best model for the dataset. In this case all the metrics pointed to GVM, but I do recommend testing both models with new unseen data since they were both performing well.
5. Why was the precision for SVM 0?
- a The confusion matrix for SVM shows why the precision was 0, the model predicted everyone as not matching.
6. How would you go about getting more data?
- a A survey would probably be the best course of action with the same questions as the original.
7. Do you have any concerns about obtaining new data?
- a I am concerned that the age of the original data will lead to issues with the current model with new data.
8. Why did you do feature importance, isn't it obvious that liking someone is the most important factor in matching?
- a While it may seem obvious, it is not always the case, and it also shows just how important it is.
9. You talked about the possibility of translating the model into an app. How does speed dating translate to online dating?
- a I mean speed dating is like online dating you look at someone's profile and decide fairly quickly if you will match with them. I feel like it will translate well.
10. Why did you use so many evaluation metrics?

- a It is always best to check how the model is performing in multiple ways, just because the model is accurate doesn't mean it is precise.