Understanding dating algorithms

Computing a matching model based on a real study

Universitat Oberta de Catalunya

Full name of the student
Àlex Franco Granell

Type
Predictive Model

Project supervisor
Francesc Julbe López

Coordinating professor
Ignasi Fernandez Hinojosa

Date of submission
01 – 2024

SUMMARY OF THE FINAL PROJECT

Title of the project:	Understanding dating algorithms: Computing a matching model based on a real study
Author name:	*Àlex Franco Granell*
Project supervisor:	*Francesc Julbe López*
Coordinating professor:	*Ignasi Fernandez Hinojosa*
Date of submission (MM/YYYY):	*01 – 2024*
Name of the degree:	Data Science
Topic of the final project:	*Predictive Model*
Language:	*English*
Keywords:	*Dating app, attraction, machine learning*

Abstract

This study looks at how dating apps work by replicating an algorithm trained on data from an actual speed dating study. Specifically, by analyzing freely published data from the Fisman et al., 2006 study, where they analyzed which characteristics made men and women choose to stay with a person after a four-minute interaction. My intention is to expand the knowledge of free dating apps as it is a growing market.

The work has been carried out mainly by analyzing the data with two Python notebooks. First, an exploratory analysis of the data is done to infer general trends in the dataset; then the second workbook discusses what is the best approach to a predictive algorithm and finally shows an example application to be able to test it with new data.

This paper verifies, through a neural network, that current dating app algorithms require data from the user's interaction with the app to be fully effective; with more interaction providing more predictive ability. Matching people based only on their interests and preferences is not enough for the algorithm to perform well.

Finally, it is shown how the provided algorithm could be improved in order to be applied in a real implementation of a dating application.

Index

List of figures

List of tables

1. Introduction

1.1. Context and motivation

Since the early study of human behavior, in the beginning of the XX century, by the first behaviorists, The Academia has been interested in the relationship between our surroundings and the way we behave [1]. Why do people choose some paths instead of others? How important are the things that we experience, or our background; like how we are raised, in our day-to-day lives?

This was latter researched into by some of the famous experiments of the Behavioral sciences. Firstly, the empirical experiments of Ivan Pavlov showed us that there was an explicit correlation between external, physical, stimulus and conduct [2]. Later, Burrhus Frederic Skinner differed in the way that the stimulus system operated. He stated that the stimulus happened after the response, and not the other way around. This meant that you learn patterns, based in previous experiences, to achieve desired future goals (stimulus), even if the reward is not instantaneous [3].

These theories had a lot of repercussions, specially in the education field [4], but severely impacted an emerging field of study: the advertising industry. Although agencies tried to influence consumer markets with their products since the very first beginning of the mass communication era, these theories showed a clear way to validate, throughout empirical experiments, previous prejudices and even create new ones [5]. Ultimately, this lead to an increased funding of behavioral studies in search of better ways to influence and increase consumption [6].

Over the years, the hyper individualistic nature of liberal markets and a strong focus on consumerism have leaded us to a liquid modernity, as Zygmunt Bauman put it in his work, where social structures are no longer rigid and stable in benefit of personal exploration and a constant desire of instant satisfaction [7]. Furthermore, these tendencies have only been increased by the appearance of the internet, where we achieved constant stimulus that was previously unseen.

All this trends: the expansion of advertisement; hyper consumer markets; the constant desire of stimulus; and the individualistic, or egotistical, exposure that the internet has granted; converged into the industry that this paper intends to explore empirically: the dating app industry.

With more than 300 million people using dating apps worldwide, about 20 million paying for premium features, and an estimated market value of 4.94 billion dollars in 2022 [8], there's no doubt that this industry has come to stay. Furthermore, in a study conducted by the Forbes magazine among more than 5000 participants, they found that the youngest users (Gen Z) preferred to establish relationships in a shorter time frame (one to three months) than it was common by the other users (four to six months) [9]. So the search for instant satisfaction may be a trend in the coming generations, even in dating.

For everything said before, we can be sure that there's a growing interest in the field of dating algorithms, and the search for the patterns that people follow to find a partner.

Right now the recommendation algorithms are proprietary and there's little data aviable to conduct precise studies open to the public. This Master's thesis intends to provide some insights of the current tendencies of dating, and a free and open prediction model that works with matching data.

For this, we will study a public dataset of speed dating data [10] conducted between 2002 and 2004. In which, participants were given four minutes to meet another participant of the opposite sex and were asked if they would like to see their date again. Also, the study collected different attributes and interest of the couples to infer the data [11]. Although the data isn't from dating apps, because they weren't invented yet, the similarity in the study may lead us to similar insights.

1.2. Goals

As stated above, our overall goal was to extract some insights from this dataset and build a model that can predict a match. Going into detail, the goals were:

General objectives	Description	Priority
O1	Extract all the knowledge that this dataset can provide about attraction patterns.	Medium
O2	Create a prediction algorithm that can tell whether or not a person will have a match with a certain type of person of the opposite sex.	High
O3	In the case of finding an effective match prediction algorithm, devise a possible implementation of it	Low

Table 1: General objectives

Going into the details of each objective, with the first one, O1, the intention was to explain the general trends that occur in the data set retrieved from the previous study. An exhaustive search for possible inferences from this data set was not performed because it has already been previously analyzed [11]. The analysis of the data has been done with the idea of being able to adapt the data for training the final algorithm if necessary.

For the second objective, O2, extensive research was done on which was the best way to achieve an effective prediction model. Given the priority of this point, and its relevance in the work, several approaches to an effective model have been tested in a specific notebook.

Finally for the third objective, O3, the intention was to build a scheme of general application in the case of finding favorable results. Finally, it has been decided to only do the theoretical implementation in this memory.

In the end, the intention of all three objectives is to increase the general knowledge about dating apps through different approaches.

1.3. Sustainability, diversity, and ethical/social challenges

1.3.1. Sustainability

Regarding this work, for all its development there is no impact on the sustainability of what it produces, because it is based on an external study and of an established industry that mainly operates through the internet; and a field, behavioral science, which does not apply to this area.

1.3.2. Ethical behaviour and social responsibility

On this point, however, it is necessary to explain its effects and how we can try to take advantage of them to do good.

As it is easy to infer, this work can serve as a basis for worsening the egocentric hyperfixation of people and promote low self-esteem of those who do not conform to the norm. Although not the focus of this study, it must be admitted that this is a possible outcome. There are really few limitations that can be placed on this problem as it is inherent to the work. I could restrict access to my code, but that would go against the mission of this paper: to provide free knowledge on the subject.

On the other hand, this work can also serve to improve the understanding of algorithms by society and the administration. As AI has entered the public sphere, the debate about how these algorithms affect people have increased, with some calling for their operation to be regulated and monitored [12]. That's why having an open description of how they work, in such a sensitive area as finding a partner, can be useful in creating regulations if necessary.

1.3.3. Diversity, gender and human rights

The data has been treated with the utmost care in order to ensure that no erroneous conclusions are drawn about the different tastes of different sexes and communities. In the case of finding specific correlations that point in this direction, It is mandatory to remark that this study is only an approximation of the case, and is not sufficient to draw conclusions.

1.4. Approach and methodology

In order to achieve the proposed objectives, the planning of the work has taken into account the hardware to be used, as well as the software and data necessary for the training of the final model. Everything organized in a limited time frame. Specifically, the CRISP-DM methodology, [13], has been used to organize the general set of tasks and processes to be carried out.

At first, public datasets were searched in order to have data for the analysis and training of the model. At the same time, a study proposal was created. Once the speed dating data was located, an initial exploratory analysis of the data was conducted and information about the algorithms of the dating apps was sought.

With a general understanding of the business environment, the available data was analyzed as well as cleaned and prepared. Meanwhile, possible gender, race and age biases were checked, in order to have balanced data.

When a clean and analyzed data set was obtained, the final algorithm was modeled using several approximations. At this stage, modeling and model evaluation operated simultaneously in order to find the best approximation to the final model. After all necessary tests were done, the final model was saved and all findings were summarized in this report.

1.4.1. Resources

This section describes everything that has been used, both data and software, in order to achieve a final model. As for the hardware, there is not much to say, the code was executed in a personal computer with a python environment manager, access to the internet and windows 11.

As for the data, on the other hand, there's a lot to cover. The body of this paper revolves around the analysis of the dataset published on OpenML.org, [10], from the study on the behavior of both sexes in speed dating [11].

This data was gathered from participants in experimental speed dating events from 2002-2004. During the events, the attendees would have a four-minute "first date" with every other participant of the opposite sex. At the end of their four minutes, participants were asked if they would like to see their date again. They were also asked to rate their date on six attributes: 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 [10]. This dataset will be explored in more detail in the results section.

Regarding the software choices, two main blocks can be described: the work environment and the libraries used.

The work environment has been designed to be easily reproducible. The code was run with a Python environment manager, Anaconda (version 2.5.1), where a specific environment was created for this work. A copy of the Conda virtual environment has also been added to the repository where the code is saved for execution and testing in order to recreate the study with the same libraries.

In addition, it was decided from the start to generate the code in two python notebooks in order to modify or adapt this work to new implementations, if necessary. Although untested, this code should run in distributed environments that allow Python code execution, such as Google Colab, because it doesn't implement hard-to-access libraries/code.

Speaking of libraries, essentially the following have been used:

Libraries	Function	Version
Pandas	Data analysis	2.1.4
NumPy	Scientific Computation	1.26.2
Scipy	Data ingestion from arff file	1.11.4
Matplotlib	Data visualization	3.8.2
Tensorflow	Modeling neural networks	2.15.0
Scikit-learn	Testing different classification models	1.3.2
Seaborn	Data visualization	0.13.0
imbalanced-learn	Resampling the data	0.11.0
Keras	Saving and testing final model	2.15.0

Table 2: List of libraries used in the code

1.4.2. Tasks description

This segment describes the different tasks in which we have broken down the work according to the goals set. The following table shows their descriptions as well as the relationship of each one with the corresponding general objective. In addition, to order them, the prioritization detailed in section 1.2 has been taken into account in order to carry out the most important tasks first, while also respecting the CRISP-DM scheme detailed at the beginning of the section 1.4.

General objective	Task Num.	Task	Description
O1	1	Data gathering	Download the data from the repository
O1	2	Exploratory analysis	Analyze the data searching for patterns
O1	3	Check missing values	Study missing values and correct if necessary
O1	4	Variable selection	Choose which attributes will use the model
O1	5	Check variable weights	See if some balancing is needed in the final dataset
O1	6	Data transformation	If necessary, adapt the data to the final model
O1	7	Perform correlation exploration	Check correlations to infer possible biases and adapt the model to them
O1	8	Perform importance test	Perform importance test on the variables of the final dataset to reevaluate all previous tasks and decisions
O2	9	Model search	Test different models to search for the best one
O2	10	Model tuning	Tune the most interesting model to further improve it.
O3	11	Theorize/implement the model	In case the model is good enough, try to implement it to an actual real case
-	12	Write the memory	Write the memory of the paper
-	13	Public defense	Public defense of the Master's thesis

Table 3: List of tasks

1.5. Schedule and evaluation

In this section, all the work was distributed in temporary spaces using a Gantt diagram. As you can see, some tasks were planned in parallel, since there are tasks that were closely linked or repeated with different importance. Temporarily, they have been organized according to the monitoring of the tutor's evaluation.

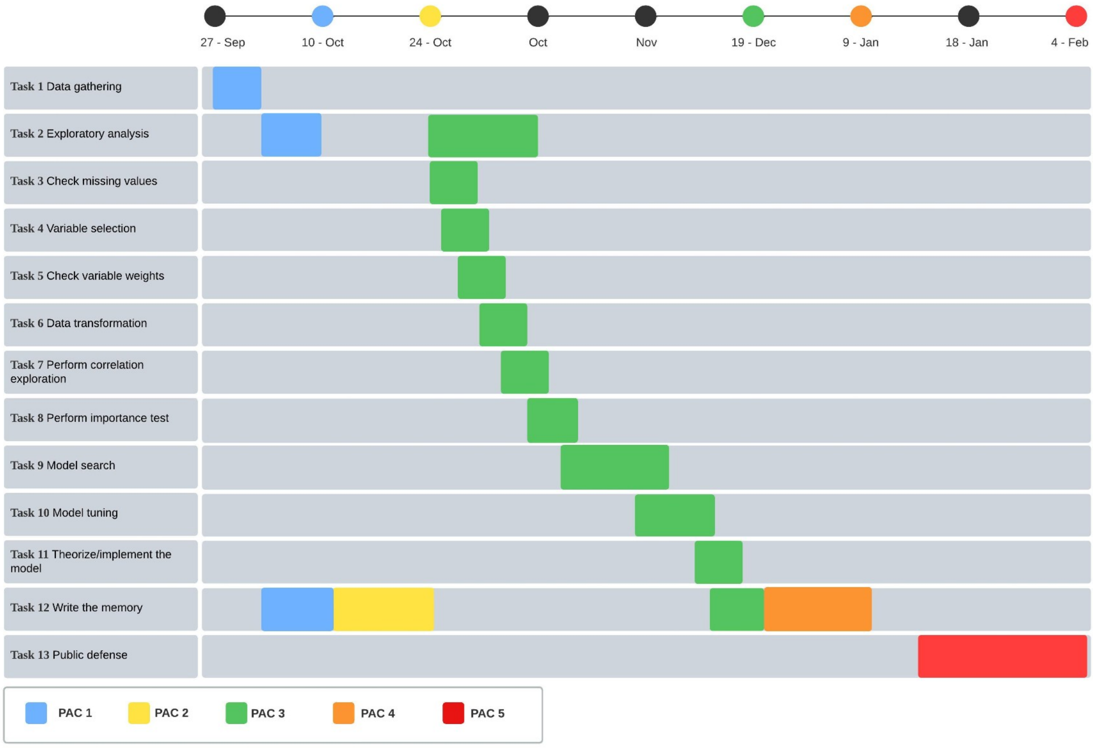


Figure 1: Gantt chart of planned work

As can be seen, the tasks were distributed to accomplish the objectives in a staggered and orderly manner, ensuring at each moment that all tasks would be performed according to their order of appearance or priority. Also, you can see some tasks overlapped with others because the tutor's control required it (for example, you can't do a study proposal without writing a memory guide and exploring the data), or because they were related tasks.

Now let's compare this planning with an approximation of the work that has actually occurred:

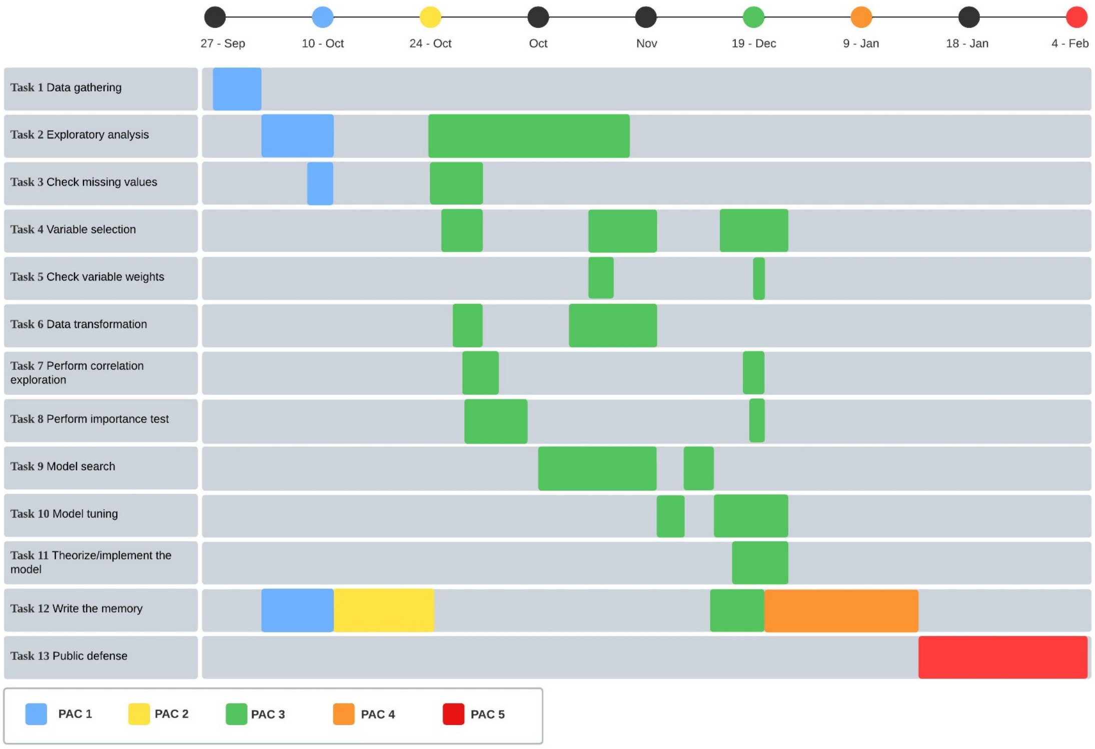


Figure 2: Gantt chart of real workflow

As you can see, the planning has been followed fairly closely for releases 1, 2, 4 and 5. The problem was in release 4, where the implementation of the code had to be done. In summary, as we will go into much more detail when the results of the data analysis are explained, the main problems were found in the modelling.

First, an analysis of the data was done and it was decided to model the continuous data. The performance of the model was seen to be poor. Subsequently, a selection of variables was made again and the analysis of these was extended, finding a great imbalance.

Finding the imbalance, in the target variable, it was decided to work on the discretized data because there were more records after cleaning. These would prove to be more effective, especially after applying balancing techniques, but the model still performed poorly.

Finally, the best model of the data was saved and the search for a better model continued after that, to see what went wrong. This trial and error of modelling the data is what has delayed the implementation of the work.

1.6. Summary of the outputs of the project

After finishing this project, essentially two products have been saved as python notebooks. One for data exploration and another for model exploration. Likewise, each one has its own results:

I. Exploratory analysis of the data: This notebook provides a detailed exploration of the analyzed dataset and generates up to four files depending on the chosen run.
   A. *SpeedDating_Original.csv*: The dataset of the original study, unaltered.
   B. *SpeedDating_clean.csv*: The dataset with no missing values.
   C. *SpeedDating_processed.csv*: The dataset with clean and discrete data
   D. *SpeedDating_continuous.csv*: The dataset with clean and continuous data

II. Prediction models: This notebook provides an extensive analysis of different tuned models and also has an output:
   A. *prediction_model.keras*: The final model trained by the notebook.

1.7. Brief description of the remaining chapters of the report

In the following chapters it is explained in detail the way in which it has been worked, starting from a theoretical analysis, until the implementation of the work, to return to a conclusion of a more theoretical nature. Specifically, you will see the following:

Chapter 2 analyzes the world of dating apps with an exhaustive search of material that is free and available to be analyzed. Specifically, this chapter delves theoretically into what are the current trends that are being followed by various applications, as well as the conclusions that are being drawn by academics about the uses made by the users of these programs. The general intention of this chapter is to lay a theoretical foundation on which to discuss, at the end of the analysis, how dating apps operate, and why they might operate that way.

Chapter 3 shows in detail the evolution of the research as well as the best implementation of the prediction model. It is mainly divided into the section on the results of the work, as well as the future work that could be done to improve the current implementation. The intention of this chapter is to show all the possible avenues of analysis that have been attempted, and to outline theoretical material to resume the debate started in chapter 2.

Finally, chapter 4 summarizes everything previously explained in points 2 and 3, and draws a theory of how dating application algorithms currently work, and which behaviors are linked to these implementations.

Then there are chapters 5, 6 and 7 which are more related to the memory than to the work itself. Chapter 5 is the glossary of terms used during the writing of this report. Chapter 6 contains the bibliography used in its preparation. And finally, in chapter 7 are the appendices that could not be included in the report due to lack of space, or due to their lack of relevance.

2. State of the art

It's important to clarify from the start that this research builds upon previously utilized data. In the initial study, this data was employed for statistical analysis, leading to certain conclusions regarding the mating patterns of both genders [11]. Notably, this contribution lies in the application of a data science approach, involving the development of a predictive model, to analyze the case, which represents a novelty and an opportunity to understand dating algorithms.

Regarding the data analysis, it's crucial to highlight a recurring issue that was initially addressed in the original article: the limitation of the study's sample when it comes to drawing broader conclusions about the entire population. This limitation arises from the fact that the sample comprises American college students, characterized by a lack of diversity in terms of interracial representation and a distinct economic profile that may not accurately reflect the broader population.

Going deeper into this matter, the article 'The Weirdest People in the World' [14] underscores the ongoing debate within academic behavioural analysis concerning the need for more inclusive and representative sampling. While acknowledging the challenges in accessing diverse data, the article recommends enhancing sampling methodologies to avoid potentially flawed conclusions.

This concern is further accentuated by recent studies, as 'The Economics of Hypergamy' [15] investigation revealed that economic disparities significantly influence mate selection in certain cultures. In particular, the study focused on the observation that Norwegian men had to earn more than their potential partners to increase their chances of securing a mate.

Specifically, hypergamy is a concept in behavioural analysis where an individual's desirability as a partner is often associated with their financial status, predominantly affecting poor men. This phenomenon was also noted in Spanish society by the online newspaper 'Xataka' [16], aligning with the findings of the Norwegian study.

Additionally, factors like age play a pivotal role in data analysis, given that men typically seek younger partners. It's worth noting that demographic shifts in many countries have led to an inverted population pyramid, giving rise to the 'marriage squeeze' phenomenon [17]. For our particular case, it should be noted that the set of data to be analyzed only includes young people, with some samples of a large age difference. Rather than modeling according to age, the focus of this research has been on age difference.

We can also be sure that this dataset, collected from personal encounters, has the potential to offer insights into the mechanics of dating websites. This claim is supported by a study conducted last year, as reported in the Universal Wiser magazine [18], which demonstrated the striking similarity between romantic attraction and the experience of finding a partner, whether through dating apps or in-person encounters. Despite participants expressing reservations about dating apps, it was revealed that personal beliefs and idealizations played a more significant role in shaping their experiences than the specific method of meeting a partner.

As we delve into the realm of dating apps, it becomes apparent that the behaviours of both genders vary significantly when it comes to seeking romantic partners. In this study, we will primarily focus on the widely popular app, Tinder, which not only enjoys a prominent place in contemporary dating culture but has also been a subject of extensive academic research in the field of online dating networks.

First and foremost, the study 'Are men intimidated by highly educated women? Undercover on Tinder' [19] reveals that it's not possible to make a generalization regarding a preference for a specific educational level in most cases. The data suggests that, in general, only women tend to seek partners who are intellectually equal or superior, while this hypothesis does not hold for men. Additionally, the assumption that individuals primarily look for partners with similar educational backgrounds can be refuted.

These findings are valuable as they offer insights into potential correlations among the preferences of the participants in our dataset. Simultaneously, they help mitigate concerns about the limited representativeness of our data. If our analysis fails to reveal any significant results regarding the significance of educational levels in forming couples, we can consider excluding this parameter from the final model.

Simultaneously, 'A first look at user activity on Tinder' [20] offers valuable insights into the distinct strategies employed by both genders to optimize their chances of finding a match. The study highlights that women tend to adopt a more passive approach, amassing matches relatively quickly, whereas men typically invest more time in the app before securing matches and respond promptly once they do. Moreover, men display a broader range of profiles that pique their interest, while women tend to initiate conversations, albeit with slightly delayed response times.

The research specifically concludes that men often refrain from using the 'likes' feature to filter potential female candidates, instead filtering once a match is established. Notably, it underscores the significance of a user's biography in enhancing the match ratio.

Similar findings were also reported in the article 'Assessing Attractiveness in Online Dating Profiles' [21], emphasizing the importance of both images and textual content within user profiles. However, this study did not identify statistically significant fixed parameters, like those found in our dataset, that directly led to a match. It's worth noting that despite this absence of direct relevance, the recommendation algorithm does utilize such data to link profiles.

In this sense, the article 'Online dating system design and relational decision-making' [22] explains that of all the types of algorithms (see-and-screen, algorithm based, and blended) the most effective approach involves a fusion of background algorithms, which link individuals, with the interactive capabilities provided by matchmaking systems. This synergy enables users to initiate conversations, thereby confirming the compatibility predicted by the algorithm and leading to better decision.

Finally, it is also important to highlight the impact of pre-established beliefs when evaluating the effectiveness of an algorithm. The study [23], published in the journal

Sage, highlighted that the perceived effectiveness of matchmaking algorithms was more important than their actual performance. Participants' belief in the algorithms was correlated with better first dates, highlighting the role of perception in the effectiveness of compatibility. This must be kept in mind when evaluating whether the result of this study is fruitful or not.

Likewise, there are already other works that have attempted an approach to the correlation of tastes in order to create algorithms capable of guaranteeing matches. Specifically, this work [24] highlighted its importance in establishing long-lasting relationships, but with the focus of a new application, not the description of the model itself.

We will resume all these points once conclusions are drawn from the dataset we are going to analyze in the next chapter.

3. Results and future work

This section explains all the work done to find a better prediction algorithm based on the data provided in the Speed Dating study. The development of this section is described according to the chronological order of the various actions taken. It is important to note that during the preparation of the work, several problems have been found that have made this implementation complicated.

First of all, the data from the repository was downloaded through a connection via request from the data exploration notebook itself. This was done to ensure the reproducibility of the analysis. To process these, they first had to be converted to pandas Dataframe format because they were stored in the repository in arff format. Once the conversion was done, they were saved as the original dataset in a csv format. An initial exploration was done with external programs, Excel, to see all the data and their typologies.

This original dataset has a high complexity of columns, with a moderate number of records. Its dimensions are 123 columns by 8378 rows. Of these, it should be noted that many of the variables were previously discretized, by previous analysis, to facilitate their understanding, and this results in many of the attributes being duplicated. The first design decision came to find the best way to approach the data for the most complete study of it. To fully understand the problem, let's summarize the variables in the following table (discrete values marked as 'd_'):

Column	Description	% of nuls
has_null	If has nul	0
wave	Wave number related to experiment	0
gender	Gender	0
age	Age of user	1.13
age_o	Age of other	1.24
d_age		0
d_d_age		0
race	Race of user	0
race_o	Race of other	0
samerace	Are same race?	0
importance_same_race	Rate importance of same race	0.94
importance_same_religion	Rate importance of same religion	0.94

d_importance_same_race		
d_importance_same_religion		0
field	Field of work	0
pref_o_attractive	I. of atractive in a couple (other)	1.06
pref_o_sincere	I. of sincere in a couple (other)	1.06
pref_o_intelligence	I. of intelligence in a couple (other)	1.06
pref_o_funny	I. of funny in a couple (other)	1.16
pref_o_ambitious	I. of ambitious in a couple (other)	1.27
pref_o_shared_interests	I. of interests in a couple (other)	1.53
d_pref_o_attractive		0
d_pref_o_sincere		0
d_pref_o_intelligence		0
d_pref_o_funny		0
d_pref_o_ambitious		0
d_pref_o_shared_interests		0
attractive_o	Rating of the other on the user	2.53
sinsere_o	Rating of the other on the user	3.42
intelligence_o	Rating of the other on the user	3.65
funny_o	Rating of the other on the user	4.29
ambitous_o	Rating of the other on the user	8.61
shared_interests_o	Rating of the other on the user	12.84
d_attractive_o		0
d_sinsere_o		0
d_intelligence_o		0
d_funny_o		0
d_ambitous_o		0
d_shared_interests_o		0
attractive_important	I. of attractive in a couple (user)	0.94
sincere_important	I. of sincere in a couple (user)	0.94
intellicence_important	I. of intellicence in a couple (user)	0.94
funny_important	I. of funny in a couple (user)	1.06
ambtition_important	I. of ambtition in a couple (user)	1.18
shared_interests_important	I. of interests in a couple (user)	1.44
d_attractive_important		0
d_sincere_important		0
d_intellicence_important		0
d_funny_important		
d_ambtition_important		0
d_shared_interests_important		0
attractive	Self-perception of the user	1.25
sincere	Self-perception of the user	1.25
intelligence	Self-perception of the user	1.25
funny	Self-perception of the user	1.25
ambition	Self-perception of the user	1.25
d_attractive		0
d_sincere		0
d_intelligence		0
d_funny		0
d_ambition		0

attractive_partner	Score of the user over the other	2.41
sincere_partner	Score of the user over the other	3.30
intelligence_partner	Score of the user over the other	3.53
funny_partner	Score of the user over the other	4.17
ambition_partner	Score of the user over the other	8.49
shared_interests_partner	Score of the user over the other	12.73
d_attractive_partner		0
d_sincere_partner		0
d_intelligence_partner		0
d_funny_partner		0
d_ambition_partner		0
d_shared_interests_partner		0
sports	User interest from 0 to 10	0.94
tvsports	User interest from 0 to 10	0.94
exercise	User interest from 0 to 10	0.94
dining	User interest from 0 to 10	0.94
museums	User interest from 0 to 10	0.94
art	User interest from 0 to 10	0.94
hiking	User interest from 0 to 10	0.94
gaming	User interest from 0 to 10	0.94
clubbing	User interest from 0 to 10	0.94
reading	User interest from 0 to 10	0.94
tv	User interest from 0 to 10	0.94
theater	User interest from 0 to 10	0.94
movies	User interest from 0 to 10	0.94
concerts	User interest from 0 to 10	0.94
music	User interest from 0 to 10	0.94
shopping	User interest from 0 to 10	0.94
yoga	User interest from 0 to 10	0.94
d_sports		0
d_tvsports		0
d_exercise		0
d_dining		0
d_museums		0
d_art		0
d_hiking		0
d_gaming		0
d_clubbing		0
d_reading		0
d_tv		0
d_theater		0
d_movies		0
d_concerts		0
d_music		0
d_shopping		0
d_yoga		0
interests_correlate	Correlation between both interests	1.88
d_interests_correlate		0
expected_happy_with_sd_people	Expectations of the study (happy)	1.20

expected_num_interested_in_me	Expected number of interests (user)	78.51
expected_num_matches	Expected number of matches	14.0
d_expected_happy_with_sd_people		0
d_expected_num_interested_in_me		0
d_expected_num_matches		0
like	Rate the other person from 0 to 10	2.86
guess_prob_liked	Rate perception of likeness	3.68
d_like		0
d_guess_prob_liked		0
met	Have both partners met before?	0
decision	Do you want to date the other?	0
decision_o		0
match	Match after decision	4.47

Table 4: Table of original data

As can be seen from the extensive table above, most of the missing values are in variables that are not discretized. This led to the main problem of data preprocessing: which sets were better to take? At the same time, another problem was added. By visually scanning the data and counting unique values, 199 values were found to be saved as "?".

These encounters complicated the choice of variables because there was uncertainty. Why was the discretized data completed? What criterion had been followed? Why were values saved as '?'? Since there is no official explanation, it can be inferred that if there are missing values it is because the truest data, to the experiment, is the continuous one, right? Well, that's what was assumed at first.

The problem with data saved as '?' was solved by removing these records as they were few. It was interpreted that previous analysts would thus save records for which they did not know the data, while those data that had not been answered, or indeed had been lost, were saved as missing values. Once the values saved as ‘?’ were removed, the cleanup of continuous data began and discrete data were discarded.

The approach to the treatment of continuous data would prove to be quite complicated to implement. The treatment of nulls first began by exploring which variables were suitable to be inferred through substitution techniques, such as sample means, but with some, such as preferences or interests, it was difficult to choose a consistent approach.

The problem came, as it would later be demonstrated with a visual exploration of the data, that the pairs had been made with the same users. In other words, there were people who appeared once per round as user, and then could appear several times as other. If the data were inferred, the dataset could be biased towards the preference of the majority in those variables that presented a higher proportion of nulls.

Because of all this, the approach that was chosen in the end was that of the direct elimination of all the records that presented null values. This technique guaranteed that the data were as truthful as possible and no complexity was introduced to the problem. The downside is that we were only left with 5790 records.

The analysis of the continuous data continued and ended even with tests of several models with them. To keep this memo brevity, however, I'm going to focus only on the final implementation of discrete data. The remains of this implementation are saved as a possible notebook output from the data exploration, as *SpeedDating_continuous.csv*.

To finish the analysis of the continuous data, it is necessary to clarify that the main reason why this approach was discarded was because of the poor performance of the tested models. Unfortunately at this stage of the implementation no visual records were kept, and therefore we cant provide complex samples of the performance of the algorithms, but it can be stated that this was around 0.55 of the area of the ROC curve and overall bad accuracy and recall.

At first, the poor performance was attributed to the scarcity of training data. At the same time, unbalanced data was also searched and a strong imbalance was found in the target. The conscious decision was as follows: the continuous data may be more accurate than the discrete data, but it leaves a lot of variability in the final models; this was a conservative approximation. On the other hand, the discrete data were less accurate because it was not possible to know how it had been decided to discretize the nulls or missing values, or if simply the data were saved directly as discrete data, but these records presented much more variability; this approach was understood as flexible. Finally, the flexible approach was chosen to ensure sufficient variability in the data for model training.

Once the discrete data approximation was chosen, redoing the work was not so much complicated as it was tedious. The discrete data did not present nulls, therefore this aspect did not have to be worked on, however, there were columns that have not yet been discussed that did present them, and many. These were also removed in the continuous data, but we will to discuss them now.

The problem with the data, when it comes to selecting attributes, is that it comes from a study done in person on people in a university environment, and that, in addition, variables specific to the study are kept. Columns like whether two people had met before (met), or what their decision was at the end of the test (decision), were not valid columns for trying to build a model that applies to dating apps. It was necessary to choose which columns to use in the final modeling and analyze them.

Without giving too much away, the variables that were removed that are linked to the study are: has null, wave, expected_happy, expected_num_interested, expected_matches, like, guess_prob_liked, met, decision and decision_o. Of all these, the most interesting to comment on are like, met, and decison; the rest can be seen in the description of the previous table that they were attached to the study.

Like is decided to be removed because a prediction algorithm used to recommend matches cannot use information related to attraction without interaction between the two couples. This variable, however, will be explored later. Met and decision are removed for the same reason as like, but were not tested.

With these data removed and with the selection of discrete variables, the data set was completed for in-depth analysis. If desired, the result of this selection can be obtained as an optional output from the Exploratory analysis of the data notebook, saved as *SpeedDating_clean.csv*. From this moment on, we will continue with the analysis of the data carried out in this notebook.

First it was decided to explore the main variables of the dataset to see how they were distributed. These, such as age, gender and race, were not discretized per se, they could not be except for age, and were analyzed without any transformation. Below I show them:

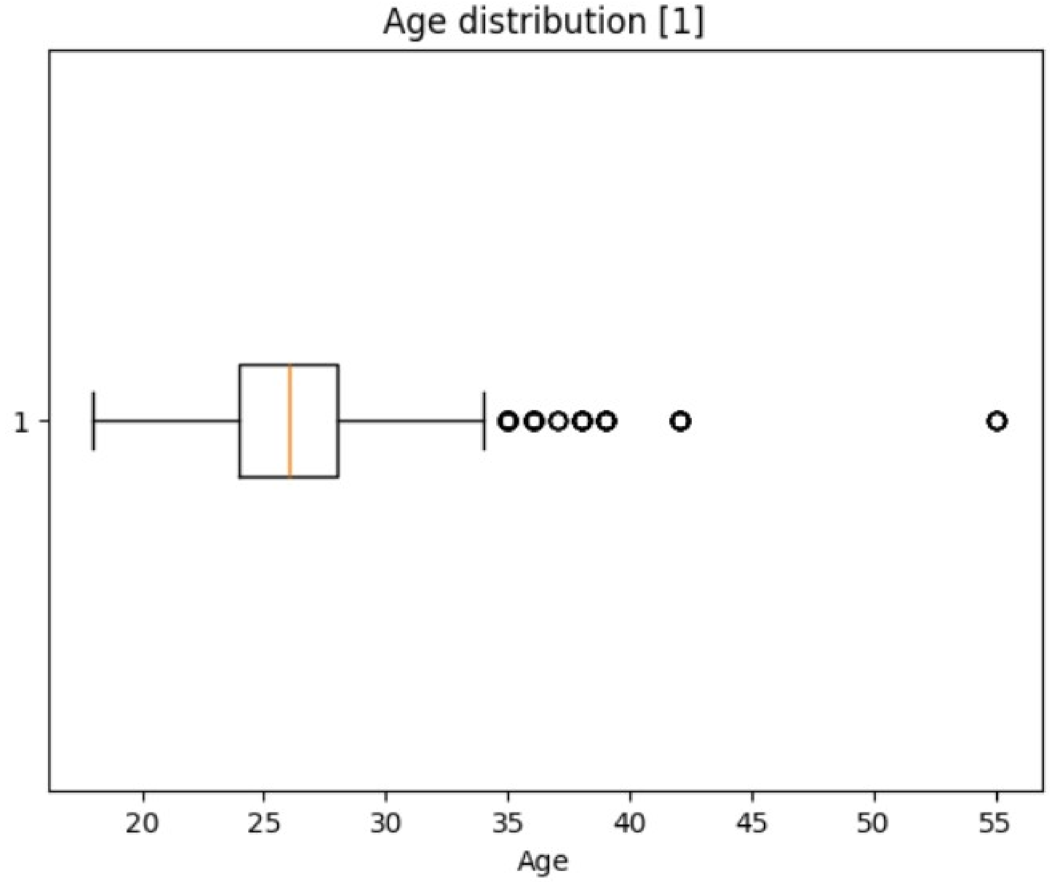


Figure 3: Age distribution in dataset

	Mean	Std	Min	25%	50%	75%	Max
age	26.36	3.57	18	24	26	28	55

Table 5: Age distribution values

As can be seen in the distribution, most of the data is between 18 and 28 years of age. Outliers who are over 35 years old were not considered a problem because the age pattern is also analyzed through the discretized difference between participants. Therefore the model receives the age of both participants and a discretized age difference column that helps the prediction. In the end, it was decided to keep the three columns to give the model the maximum variability in terms of age possible.

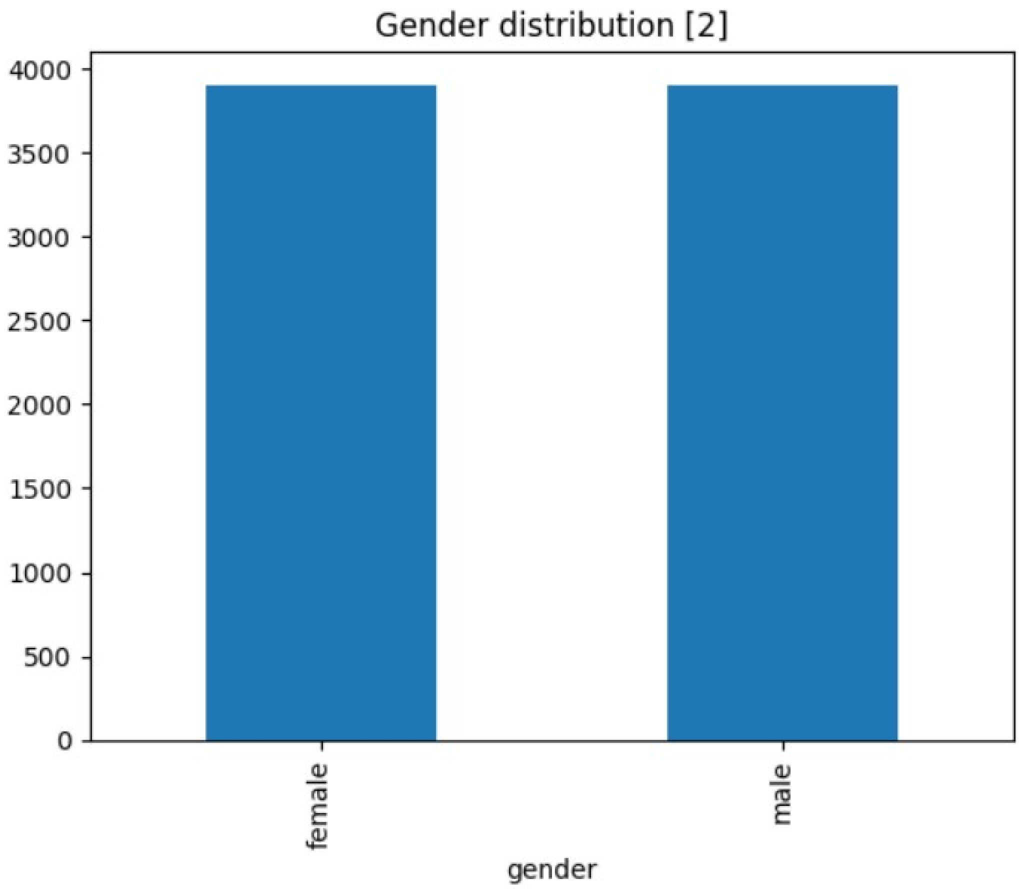


Figure 4: Gender distribution

Fortunately the gender data were balanced. So far the most stable variables of all the general characteristics. Race was then analyzed and it is necessary to explain how this variable was carried out.

The race column was comprehensively analyzed when the continuous data analysis was done. In this, it was found that the race values were sufficiently unbalanced and a first attempt to correct them was made in order to ensure the same representativeness of the data. This was decided following the recommendations previously cited in section 2, where the emphasis was on treating ethnic minorities correctly [14]. This approach was maintained until the redo of the study, when the discrete variables were chosen. The reason was that in the analysis of the importance of the model, it was seen that the race of the people mattered little to explain the model, and the variable of whether they were the same race, 'samerace', became more relevant.

That is why in the end it was decided not to do a treatment of the column, in order not to introduce complexity to the final analysis and to keep the maximum number of records. Since a conservative treatment of this reduced the total number of records to balance the category. However, due to its small importance in the model, it is decided not to treat it but to leave all the columns in the final model so that it has enough variability to be able to train. That is why its distribution is not shown; you just need to know that most of the data is from white/Caucasian people.

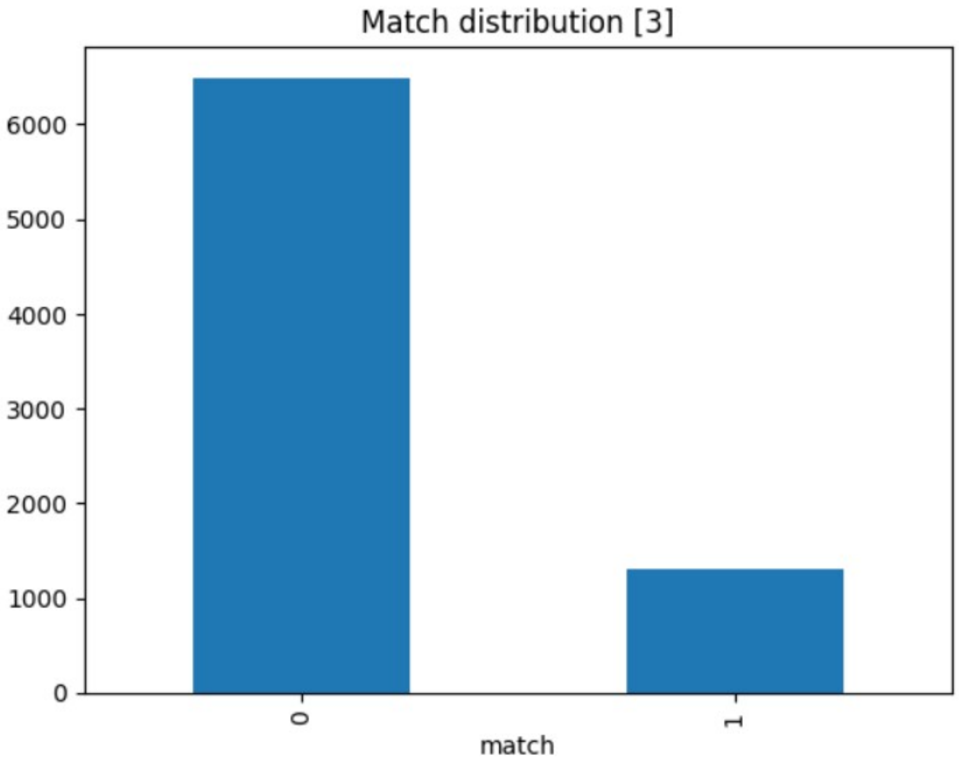


Figure 5: Match distribution

	% of 0	**% of 1**
Match	83.2	16.8

Table 6: Percentage of matches

After analyzing the race variable, and having done the analysis of the continuous data, it was decided to do the balanced data search again. Out of all of them, the match variable was found to be very unbalanced: as you can see, only 16.8% of the data was assigned as a match. The treatment of this column, however, would be left for the modeling section, not for the final data set. With the main variables analyzed, we proceeded to analyze the attributes that made up the body of the model: the participants' interests, self-perception and score. It should be kept in mind that these graphs show only the values that have matched, according to the proportion of each category.

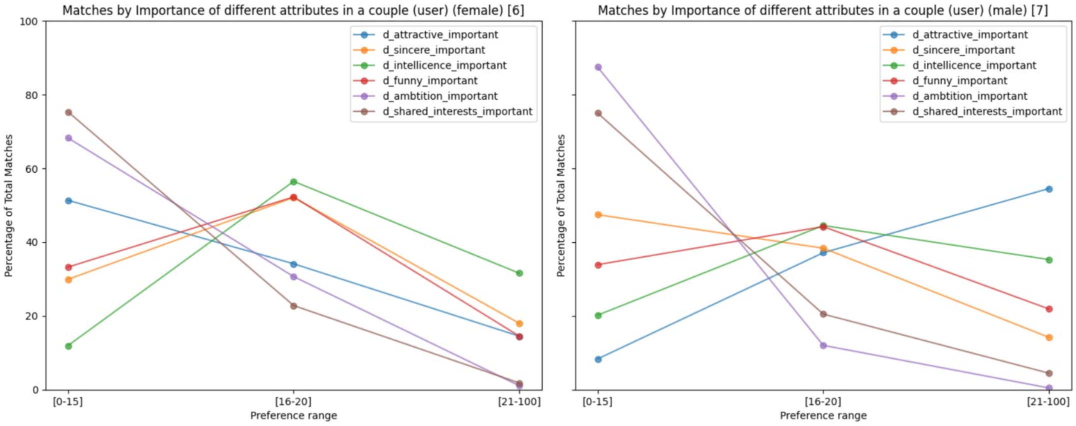


Figure 6: Matches by importance of different attributes in a couple

The first thing that was found in these variables was, as discussed above, that the pairs had been made between the participants themselves. That is why we will only analyze the data of the users, and not the data of the couple, because they are the ones that are correctly labeled. You can find the other plots in the appendices or in the notebook.

If we look at the data on the importance of attributes [plots 6-7], we can see that for both sexes it is quite important that the partner is intelligent, as well as funny and sincere. Also for both sexes, it is seen that it is not important that the partner is ambitious or shares interests. Men stand out because they focus considerably more on the partner's attractiveness, while women focus on the man's intelligence, and value the partner's ambition more.

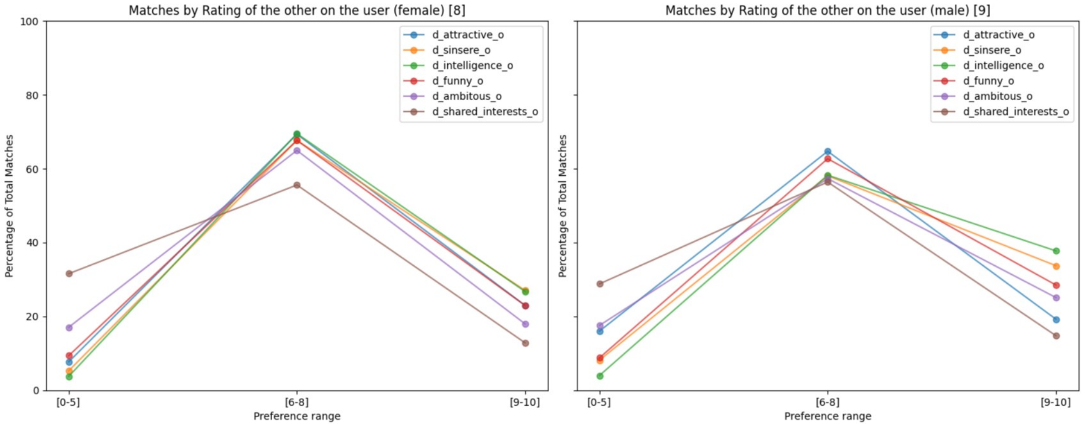


Figure 7: Matches by rating of the other on the user

Regarding the other person's ratings [8-9], it should be noted that most matches, for both sexes, rate the partner between 6 and 8, showing a conservative behavior when it comes to short dates. The main difference between both genders is that men tend to score higher than women. If the score above 5 is added, it can be seen that women focused more (positively) on the attributes of intelligence, sincerity and fun, as expected; while men scored more sincerity, intelligence and that they were funny. Interestingly, men value attractiveness more in a partner, but in the experiment they did not leave such high scores.

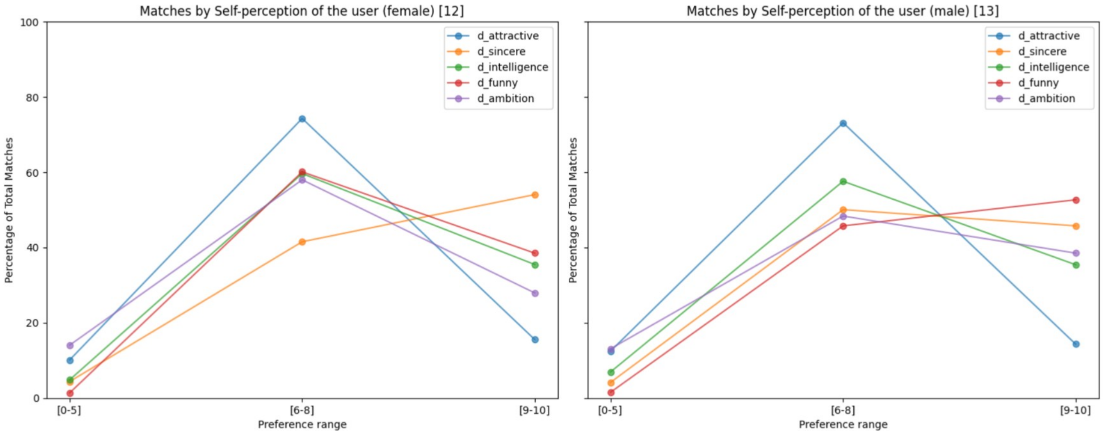


Figure 8: Matches by self perception

If we look at the self-perception graphs [12-13], we see that most matches, for both sexes, are rated between 6 and 8, but they differ from the ratings perceived by others that many tend to rate themselves above 8. The best matching data for both genders is that the funnier you were, as well as the more sincere you were, the better. The difference between the genders is that women who have more matches tend to consider themselves very honest; while men tend to consider themselves more ambitious and funny.

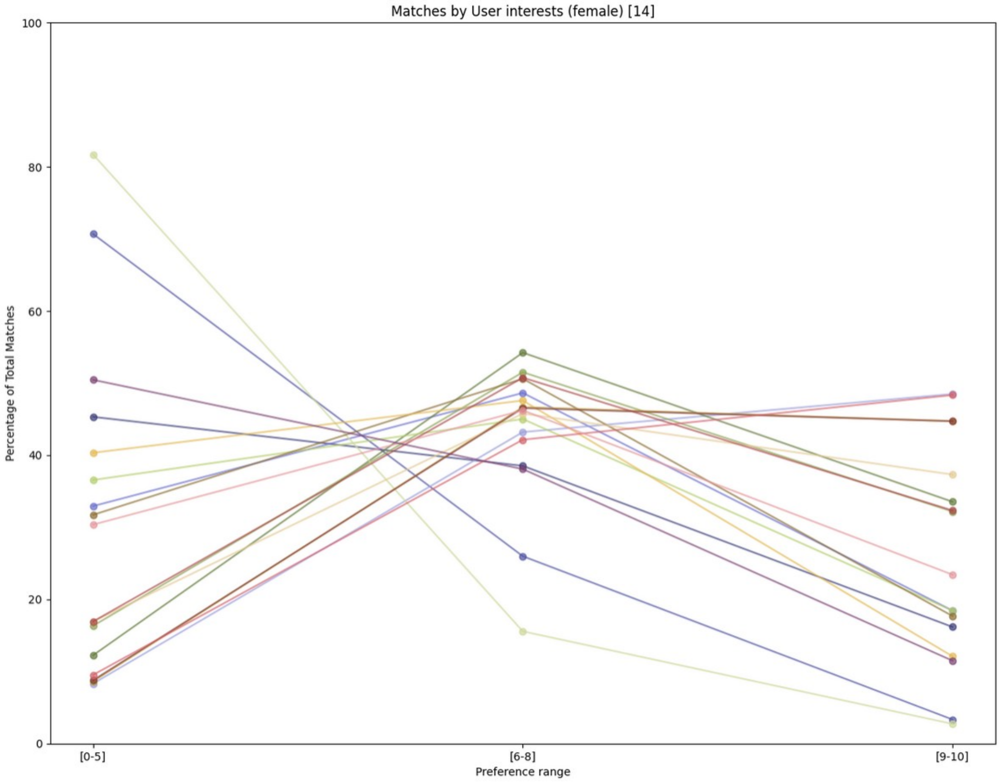


Figure 9: Matches by user interest (male)

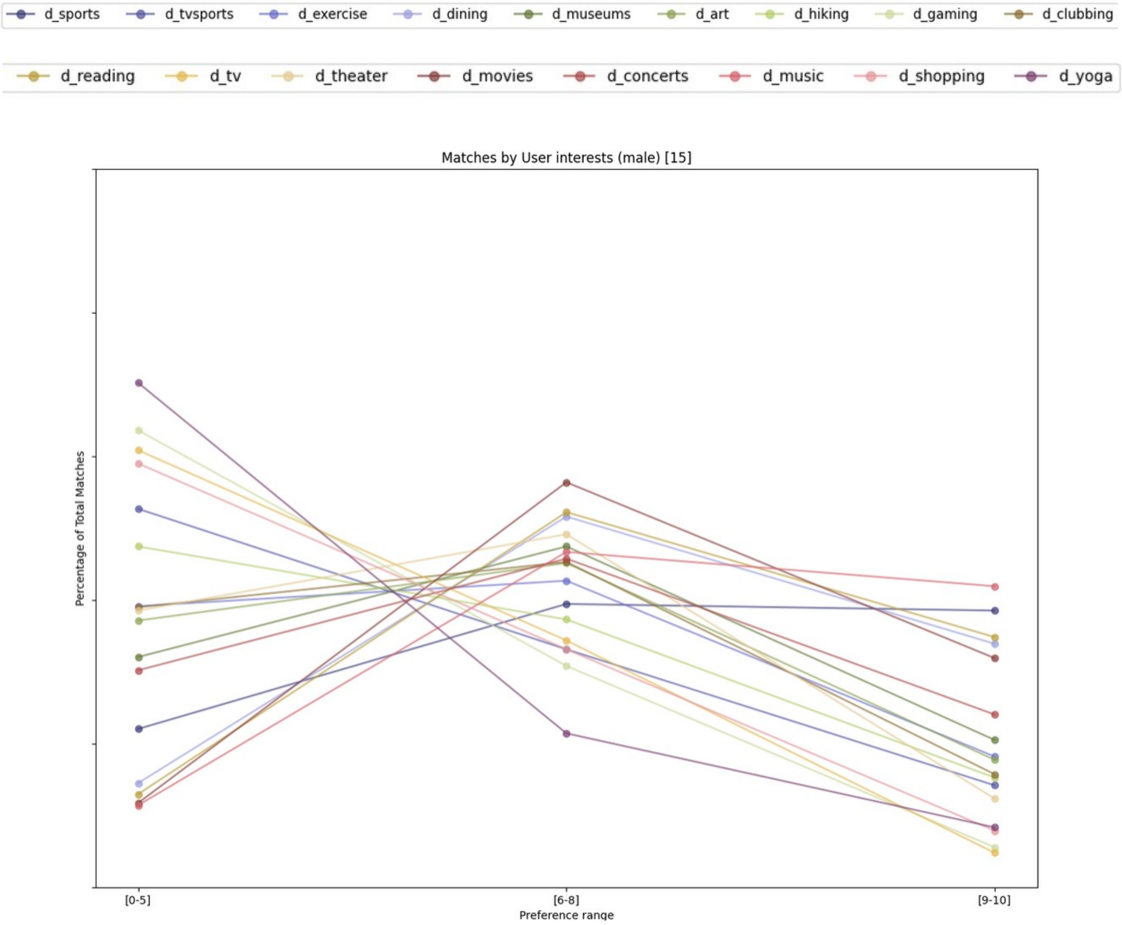


Figure 10: Matches by user interest (female)

If we look at the matches by interests [14-15], we see that there are many variables that could mean the same thing, which is why I will focus only on those people who are minimally interested in the category that appears represented [+5]. Interested in 6 to 8, women tend to get more matches if they are interested in sports, music, shopping and art. Men, on the other hand, are more successful if they are interested in sports, shopping, music and watching movies. As for those who are very interested in a subject, women are more successful when it is for dining, watching movies or music. Men for sport and shopping. In short, it means that the majority is more successful in the activities most common to the rest of the participants, that is to say, that their interests are linked as it can be seen below.

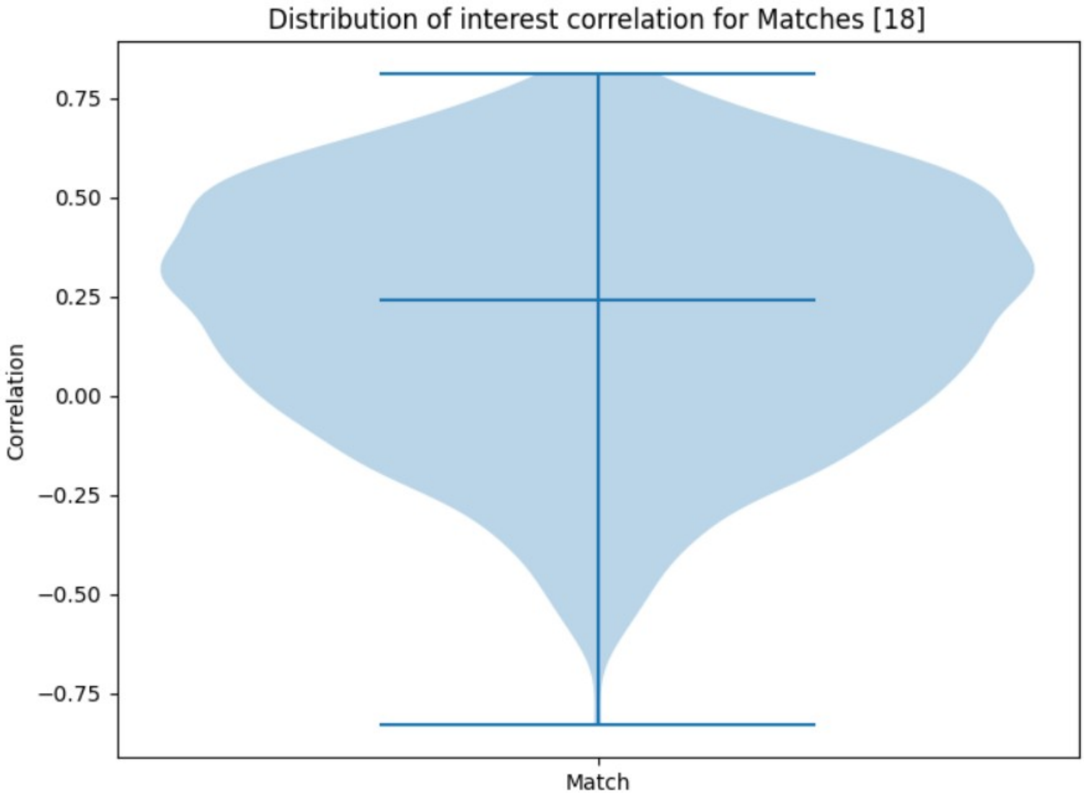


Figure 11: Interest correlation for matches

You can see in the graph [18] that the more interests the couples share, the more matches there are. Even if they don’t consider it important.

Finally, the jobs were analyzed, and it should be noted that this was complicated to interpret. This column was cleaned up to group the variables a bit, but still as they are saved there is too much variability. The good part of this approach is that it keeps a lot of detail, but for a professional model may be interesting to simplify it. As has been said, the treatment was light due to the large number of types. That is why it was decided to only analyze the professions that accumulated the most matches.

The conclusions that were found were: women tend to prefer men who work as entrepreneurs, lawyers, in marketing, engineering or finance. Men tend to prefer women who work as lawyers, social workers, entrepreneurs, psychologists, doctors, in politics or in public administration. In general, women tend to prefer less variety of specializations than men.

These encounters can be linked to the theoretical conclusions established in chapter 2, where it is said that women tend to prefer fewer types of men and with higher paying professions; while men choose more variedly. Nevertheless, it should be noted that this analysis is not sufficient to draw conclusions about the general population, but it is curious to see that the patterns are similar.

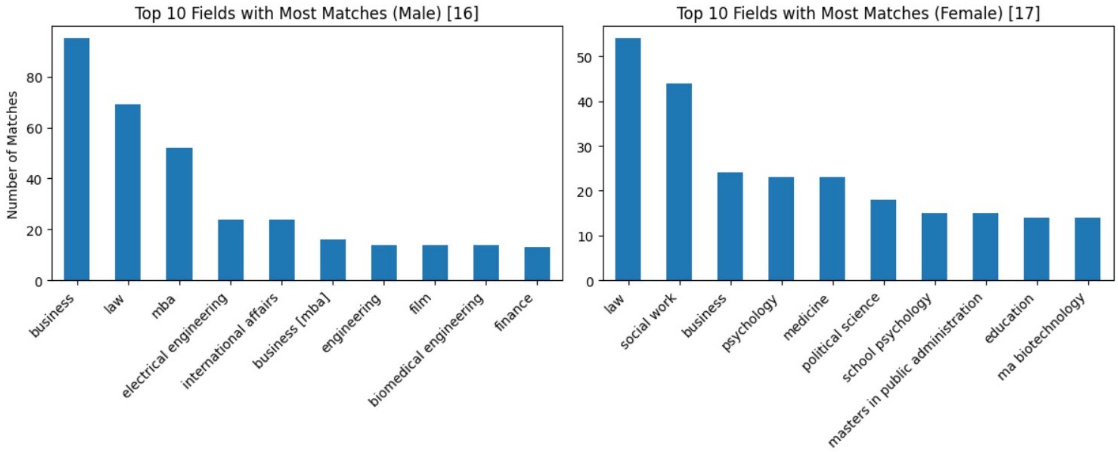


Figure 12: Top 10 fields with most matches

Up to this point are the findings that were made by analyzing the original data. Next what was done is a final transformation of the data to save it with numerical values, so that the algorithms can interpret it. This result is saved as *SpeedDating_processed.csv* and is the final dataset that will be used for model searching later. Before that, however, an analysis of the current data was done to look for correlations and test the significance of the variables in an example model.

Of this analysis, however, we will not go into too much detail because the conclusions are best seen in the modelling section. It is only necessary to point out that the correlation matrix was useful to show the low explanatory power of the variables of personal interests, as well as the cleaning of variables done in the modelling. These points will be more relevant later, so there is no need to elaborate now. If you want to consult the correlation matrix, it can be consulted in the appendices. The same can be said for the importance tests made in this segment. Also, this section can be reviewed in the notebook itself for more details. After all, at the end of the exploratory data analysis, the dataset contained 58 columns and 7802 records.

The modelling part could be described in extensive detail, but I think it would be counterproductive due to the large amount of tests that have been done. Instead, I will summarize the main problems encountered in this section, as well as what decisions have been made to solve them.

First, as explained above, continuous data proved to be ineffective in training a model. At this point I was still choosing which models to test, so it was decided to include balanced models to see if it improved the performance. When it improved slightly, that's when the analysis was redone, as already mentioned.

Once the discrete data had been processed, it was decided to make the final selection of variables and their rescaling in the modelling section. Several strategies were to be tried for the processed data set, but the final choice of variables had to be settled first.

Just as the variables linked to the study had not been included in the model, data remained that could not be included in a real implementation of the code. These were the columns that kept one person's score over the other in various aspects. As you can imagine, a recommendation algorithm cannot recommend based on already established opinions, because it cannot know them.

However, the poor performance of the model required trying to maintain them and several approaches were tried. In the end it was decided to be consistent and eliminate them, as well as the variables of interest. These, although they would be useful in a real model, it was seen in the correlation matrix and in the significance analysis that they could be eliminated and leave only the result of the correlation of the couple's interests, as it stored the same data.

This dataset finally presents 28 training columns and the target column, with a total of 7801 records. Of all there is, it only remains to say that the interest correlation was decided to be taken as continuous data instead of the discrete column because after doing several tests it showed better performance that way. Although variable selection will be important in the future work chapter, the final dataset is:

Column	Description
gender	Gender
age	Age of user
age_o	Age of other
d_d_age	Age difference
race	Race of user
race_o	Race of other
samerace	Are same race?
d_importance_same_race	Rate importance of same race
d_importance_same_religion	Rate importance of same religion
field	Field of work
d_pref_o_attractive	Importance of attractive in a couple (other)
d_pref_o_sincere	I. of sincere in a couple (other)
d_pref_o_intelligence	I. of intelligence in a couple (other)
d_pref_o_funny	I. of funny in a couple (other)
d_pref_o_ambitious	I. of ambitious in a couple (other)
d_pref_o_shared_interests	I. of interests in a couple (other)
d_attractive_important	I. of attractive in a couple (user)
d_sincere_important	I. of sincere in a couple (user)
d_intelligence_important	I. of intelligence in a couple (user)
d_funny_important	I. of funny in a couple (user)
d_ambition_important	I. of ambition in a couple (user)
d_shared_interests_important	I. of interests in a couple (user)
d_attractive	Self-perception of the user
d_sincere	Self-perception of the user
d_intelligence	Self-perception of the user
d_funny	Self-perception of the user
d_ambition	Self-perception of the user
interests_correlate	Correlation between both interests
match	Match after decision

Table 7: Final dataset

Once it was clear which variables should be used, the approach of rebalancing them remained. As can be seen in the presented code, it was programmed in such a way that it allowed to do several tests with various approximations. For this report, we will focus on the final execution.

Before that, however, you need to know the tested methods. The unbalanced dataset was tested first, as mentioned, and performed poorly. Then a resampling approach was chosen, starting from the target variable, in order not to introduce fictitious data into the training. The good part of this approach was that it was trained with real data, the bad part is that a lot of variability was lost in the training, and the number of misassigned data increased considerably. After some research, the implementation was finally tested with SMOTE, which is a rebalancing technique that adds inferred data to the training set.

Although the different tests with different balancing methods are listed in the appendices, in the end it was seen that the best approach was to use SMOTE. Keeping the number of records and balancing it with inferred data ensured a better accuracy of the algorithms, with a slight performance improvement. The resampling method is considered that could have been correct if more data was available, since it also improved the performance but increased the assignment of false negatives by a great margin.

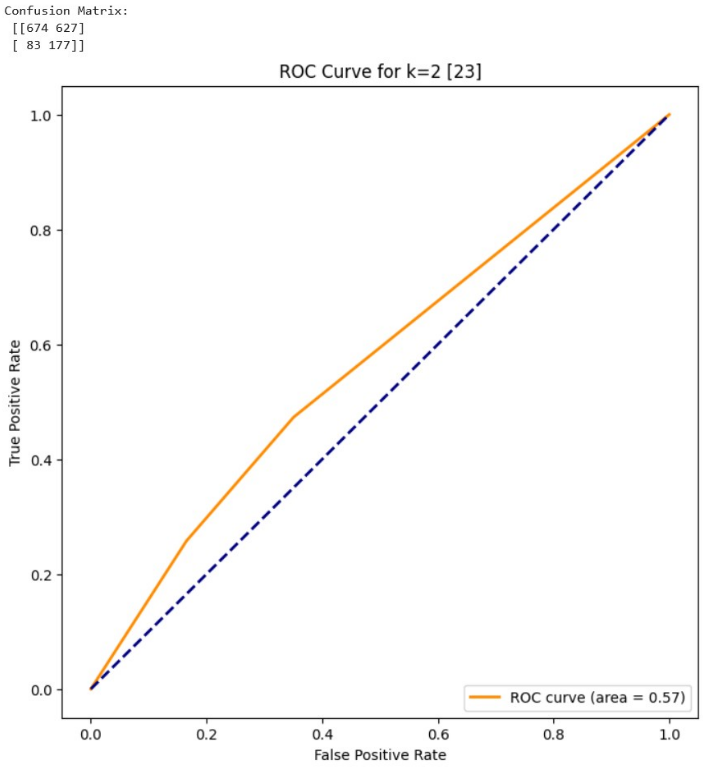
All in all, the final dataset was rebalanced using the SMOTE technique for the training sets. Now it remained to be seen which model performed better according to business interests. The question was, what were these? At first, the aim was to find a model with the highest possible accuracy, but as the poor performance of the models was seen, the approach was changed. In the end the choice of model was made based on the recall of the evaluation.

It was interpreted that in a real implementation, a real business case, given several models with poor performance, the one that would always guarantee that the number of false positives was minimal would be chosen. This is because the company wants to ensure that it always recommends the largest number of matches possible; as also because in a real application the number of false positives can be beneficial.

It should be kept in mind that dating apps make money the more time the user spends on it. At the same time, it is in their interest to ensure that those users who spend more time on it have a better chance of getting a match. If the number of false positives increases, this would imply that less compatible people would be recommended to the user, which would increase the user's time in the application. Otherwise, it is more interesting for the application to guarantee that the user will always have options than to guarantee that they get a match. Therefore, false positives were not interpreted as a problem, but as a feature. This will be resumed in the future work chapter.

All in all, these are the models that were tested:

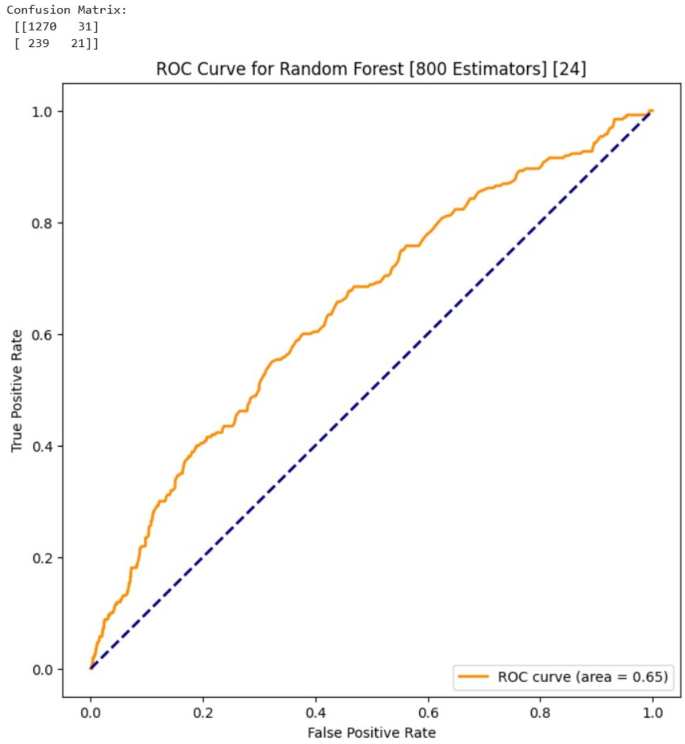
Knn algorithm:



```
Confusion Matrix:
 [[674 627]
 [ 83 177]]
```

Figure 13: Final Knn with SMOTE

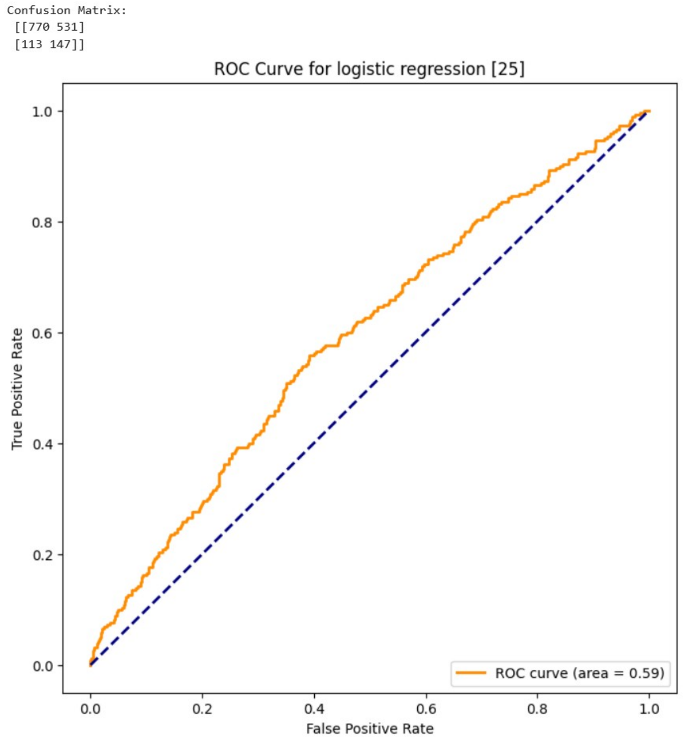
RandomForest algorithm:



```
Confusion Matrix:
 [[1270   31]
 [ 239   21]]
```

Figure 14: Final RandomForest with SMOTE

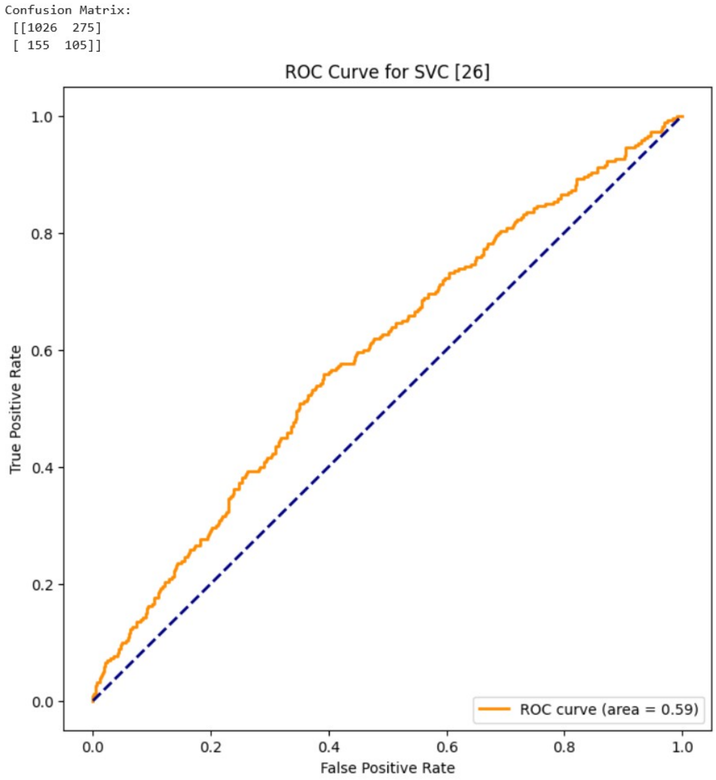
Logistic regression algorithm:



```
Confusion Matrix:
 [[770 531]
 [113 147]]
```

Figure 15: Final logistic regression with SMOTE

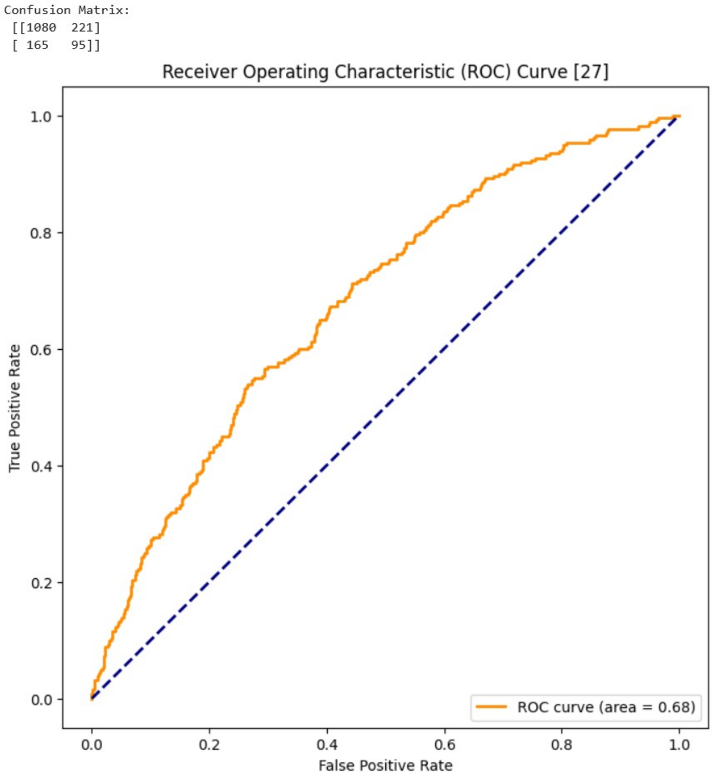
SVC algorithm:



```
Confusion Matrix:
 [[1026  275]
 [ 155  105]]
```

Figure 16: Final SVC with SMOTE

Neural network algorithm:



```
Confusion Matrix:
 [[1080  221]
 [ 165   95]]
```

Figure 17: Neural network test with SMOTE

In addition to these tests, two other balanced ones are included that I will not discuss in the memory because with SMOTE the results do not change.

As you can see in the previous images, of all the algorithms the ones that present the best data according to our interests are those of SVC and Neural Networks. The one from knn does not have enough accuracy, the Random Forest has a tendency to false positives, the logistic regression is not bad by itself but others have more accuracy, and finally we are left with the two mentioned.

After tests with several rebalancing methods, and with several runs for each, it was seen that the results of the neural network were consistently better than those of the SVC algorithm. Neural networks were generally more accurate, with a similar tendency to false negatives. That is why it was decided to finally fine-tune this model.

For the tuning of the model, a series of steps were followed to guarantee the best choice of final hyperparameters. First, a neural network was created with 512 neurons fully connected to a second layer of 256, until it is concentrated to a single output that assigns the probability of a record to present a match or not.

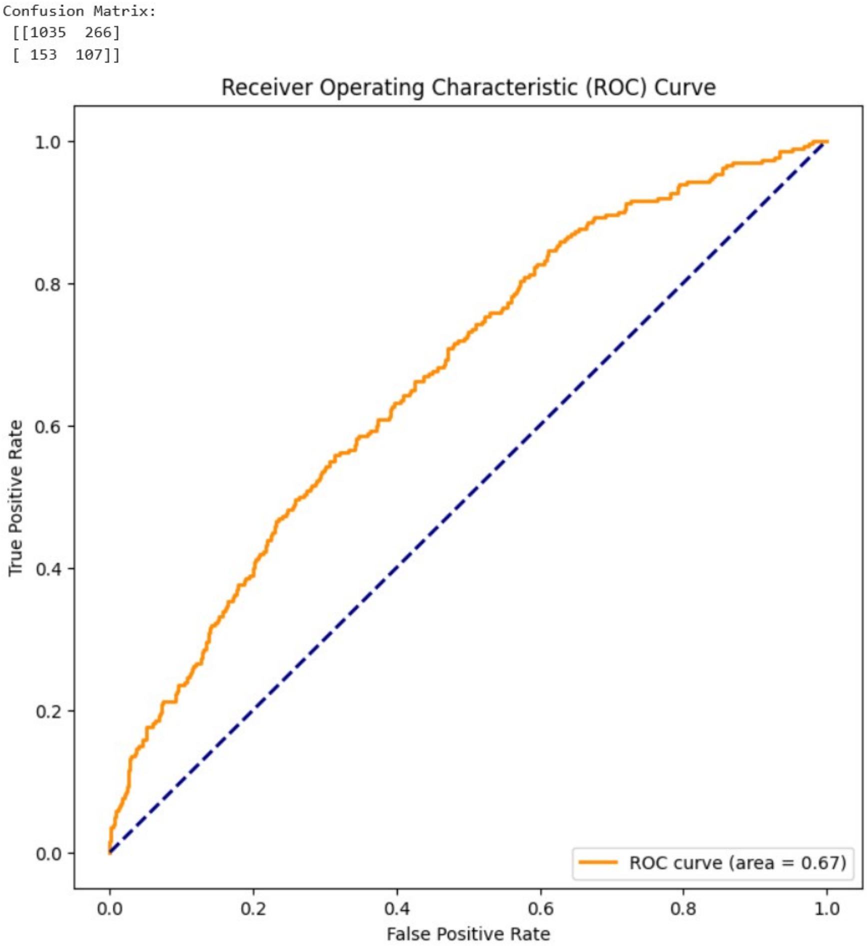
Dropout layers were added between the different layers to prevent overfitting of the model. This layer was put specifically after choosing to operate with SMOTE, because some models presented it. Dropout is assigned between layers with a 50% probability.

Finally the model was created according to the selected optimizer, with a binary crossentropy loss function, suitable for binary problems. It should be noted that the neurons go from relu, to a sigmoid final neuron to properly fine-tune the final selection.

The set was then split into training and test data with an 80/20 ratio. For the selection of hyperparameters, a grid search was made with several options of epochs and batch_size. Various optimizers were also tested, such as adam (for general problems), rsmprop (for binary problems) and sgd (which tunes the learning rate more precisely). Of all of them, the best parameters were 50 epochs, 16 batch_size and the rmsprop optimizer. In total the search for hyperparameters takes just over 3 hours, searching for the model with the best recall, as already explained.

As a detail, it is shown in the notebook that the SMOTE application balances the data around 5000 records per typology. In other words, the records are increased to more than 10000.

Once the best parameters were found, the model was trained with early stopping to guarantee the best version of it. At first, due to low performance, the limit parameter that assigns a match as positive was modified. Usually neural networks assign as positive from a probability of 50%. In the first tests, this variable was lowered, up to 45%, to guarantee better results despite the reduction in accuracy. In the final model, however, it remains as the standard value with the possibility of being able to modify it if desired.



```
Confusion Matrix:
 [[1035  266]
 [ 153  107]]
```

Figure 18: Final prediction model

As can be seen, it does not improve much more than the test neural network, but it guarantees better accuracy with some bias towards false positives. Finally, a test was added to the notebook to be able to test the predictive capacity of the algorithm on demand. This model was saved as *prediction_model.keras*.

Since it is a neural network, the interpretation of the importance of the features is complex because it must be analyzed through its weights. That is why in this sense it has been chosen to analyze the importance of the characteristics for the Random Forest test. Although they are certainly not equivalent, and this algorithm tends towards false negatives, trying to infer from this alternative model is simpler than from the different weights of the neural network.

Feature	**Importance**
d_d_age	‘0.088’
d_pref_o_funny	‘0.071’
interests_correlate	‘0.065’
age_o	‘0.064’
d_pref_o_sincere	‘0.057’
field	‘0.053’
age	‘0.05’
d_funny_important	‘0.041’
d_pref_o_intelligence	‘0.039’
d_pref_o_shared_interests	‘0.039’
d_pref_o_attractive	‘0.035
d_importance_same_race	‘0.034’
race_o	‘0.033’
d_funny	‘0.029’
d_sincere_important	‘0.027’
d_importance_same_religion	‘0.027’
d_intellicence_important	‘0.026’
d_pref_o_ambitious	‘0.025’
d_sincere	‘0.024’
d_shared_interests_important	‘0.024’
race	‘0.023’
d_intelligence	‘0.023’
d_ambition	‘0.023’
d_attractive_important	‘0.023’
d_attractive	‘0.02’
samerace	‘0.013’
d_ambtition_important	‘0.013’
gender	‘0.01’

Table 8: Final SMOTE importance

Of all the variables, the most important seem to be: the age difference, the interest in a partner being funny, that they have the same interests (as could already be predicted with the analysis of the correlations shown previously), the age of the other person, the preference for the partner to be honest, the field of work and the same age. In other words, the most important thing is the age, that the couple is funny and sincere, that they have the same interests and a good field of work.

3.1. Future Work

During, and after, the implementation of this work, several problems have arisen that have hindered an optimal development of this prediction algorithm proposal. However, in a real application I find that some parts of this code could be adapted to find a better implementation. In this section we will superficially explore some of the decisions taken and how they could be improved.

In the selection of variables, it was chosen to generate an algorithm based on the perception of the user, the partner's expectations, the users interests and sociological data. This implementation has been the most correct for the analyzed problem, but it can be improved much more.

Today's dating app algorithms use user behavior to gauge whether or not they are interested in the profile they are viewing [21], [22]. In this way it can be inferred how applications like Tinder guarantee users that the more time they spend on the application, the more chance they have of having matches. In the algorithm presented, however, this aspect cannot be evaluated.

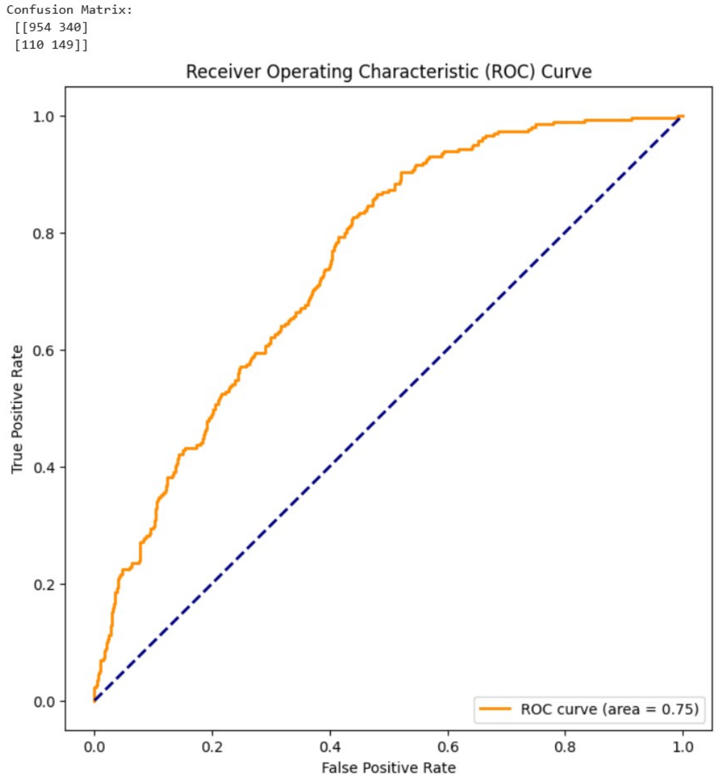
During the test training, it was tried to incorporate at first the variables of ratings of users on others. Clearly the results were much better. Also, in the analysis of importance in data exploration, it was applied to the whole set in general, and it was seen that the most explanatory variables were those that were linked to the user's experience (linked also to the experiment). Variables such as decision or like were closely linked to the probability of a match.

It is for all this that I wanted to do a last training with the best model, just adding the experience control variable 'like'. This not only improved the accuracy of the algorithm, but also showed in the confusion matrix that the model improved on false positives and favored false negatives. Being considered an improvement without a doubt on the final model presented.

What can be extracted from all this? Apparently, because there isn't enough evidence to say it outright, algorithms based on personal interests and preferences have a limited ability to match people, and it's when you check the user experience that you see that the model improves considerably. This was also in line with the conclusions of previous researchers [21], [22].

In this final experiment the ‘like’ variable was used because it was the most easily adaptable attribute for a real code implementation. It was stored in values from 0 to 10, and could be implemented as an average aggregation of variables such as: time spent viewing a profile, time spent reading a biography, time spent to give a like (so it applied to subsequent similar matches, creating precedents from which to retrain the neural network), etc. This variable was the one that gave the most flexibility for a real adaptation of the code. However, it goes without saying that if this code is adapted to a real case, the programmer's intention is to grab as much user data as possible to train the model.

As a final statement, we can see below the proposed trained model, with the same parameters as the final one, but using the 'like' variable:



```
Confusion Matrix:
 [[954 340]
 [110 149]]
```

Figure 19: Final model trained with SMOTE and like column

4. Conclusions

This work has tested an approach to a real business case that aimed to generate a recommendation model for a dating application. At the same time, the intention was to contrast everything found during the implementation with the theoretical research proposed in chapter 2 of the report. All with the will of creating free knowledge around dating apps to make them easier to be understood.

The objectives of this work were therefore to explore the data, to create a prediction model that could be implemented in a real case of a dating application, and to theorize about its implementation. With several nuances about each objective, I consider that a good research and approach to the problem has been achieved.

The analytical exploration of the data has been carried out correctly, although it is true that it has not been statistically in-depth. The decision not to focus on deep analysis of the data and infer trends about the population came because this dataset has already been analyzed previously [11]. Without an increase in the sample, nor a substantial change in the methodology, it was considered that delving into the statistical trends did not bring anything new. Instead, it was decided to analyze the data to prepare it in detail for the presented problem, the creation of a predictive model.

Regarding the exploration of several models, as well as the refinement of the final model, It is considered that they have been done according to the possibilities. Although the data collected is not the best to make an example implementation, a model has been achieved that is able to predict a match moderately well. It is in the theorizing of a real implementation, in chapter 3.1., that a good predictive model has been visualized. Although it escaped the limits proposed in the study.

Despite having tried to follow the planning proposed in the introduction, the choice of variables has turned out to be a sufficiently complex problem that it has lasted until the end of the study. This has meant that a lot of time has been invested in redoing the study and not improving, if possible, the implementation of the final model. However, It is considered to be correct and complete.

For the entire development of chapter 3, especially chapter 3.1., where future work is discussed, it can be concluded that the creation of dating algorithms requires saving user interaction data with it in order to guarantee good performance of the algorithm. Algorithms based on interests or preferences only serve to make some initial recommendations, for the rest the algorithm should save user data for a good implementation.

Based on this, and with the course of the implementation, it can be verified that the conclusions presented in the various studies in chapter 2 are true:

- The behavior in the selection of partners is different in men and women: men preferring more variability while women choose fewer professional profiles [15] [20].

- The importance of the picture or biography is essential in a dating app, as they allow to save more data of the user's interaction with the app [21] [22].

- Age is very important in the selection of partners. Especially the age difference [17].

- Chatting is essential for evaluating the performance of an algorithm, as it allows measuring more interaction variables, such as conversation time, number of messages, etc [22].

In short, the algorithms of dating applications do not only require a person's interests and preferences; they also require user interaction data to increase predictive power. Linking these findings to the conclusions of Chapter 1.3.2, this should be taken into account if restrictions are to be applied to the uses of algorithms in this area.

This study, unfortunately, could only be carried out for the race data collected in the dataset and published freely; as well as only heterosexual relationships could be analysed. It is necessary to expand the work on the subject.

5. Glossary

- **Accuracy**: A measure of the overall correctness of a model, calculated as the ratio of correctly predicted instances to the total instances.
- **Algorithm**: A step-by-step procedure or set of rules followed to solve a specific problem or perform a particular task in computing or data analysis.
- **Balanced Data**: A dataset where the distribution of classes or categories is roughly equal, preventing bias in model training.
- **Behaviour studies**: Research that examines and analyzes human behavior, often used in fields such as psychology, sociology, and marketing to understand patterns and trends.
- **Complexity**: The level of intricacy or difficulty in a system, model, or problem.
- **Continuous Variables**: Variables that can take any value within a given range.
- **Correlation**: A statistical measure that describes the extent to which two variables change together, either positively or negatively.
- **CRISP-DM** (Cross-Industry Standard Process for Data Mining): A widely used process model for data mining and analytics projects.
- **Dataset**: A collection of data, typically organized in tabular form.
- **Discrete Variables**: Variables that can only take distinct, separate values.
- **Discretize**: The process of converting continuous variables into discrete categories or bins.
- **Fine-tune**: The process of adjusting hyperparameters or model parameters to optimize performance.
- **Hypergamy**: The practice of marrying someone of a higher social or economic class, often associated with seeking a mate of higher status.
- **Implementation**: The process of putting a plan or decision into effect, referring to the practical application of algorithms or systems.
- **Jupyter Notebook**: An open-source interactive web application for creating and sharing documents containing live code.
- **Knn** (k-Nearest Neighbors): A supervised learning algorithm that classifies a data point based on the majority class of its k-nearest neighbors.

- **Logistic Regression**: A statistical method used for binary and multiclass classification, modeling the probability of an instance belonging to a particular class.
- **Mach**: A term used to describe traditional masculine attitudes or behavior.
- **Marriage Squeeze**: A demographic phenomenon where a shortage of eligible partners in a particular group or population leads to difficulties in finding suitable marriage partners.
- **Missing Values / Nulls**: Absent or undefined values in a dataset.
- **Neural Network**: A computational model inspired by the structure and function of the human brain, used for machine learning and pattern recognition.
- **Performance**: The effectiveness or efficiency of an algorithm.
- **RandomForest**: An ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes as the prediction.
- **Recall**: Also known as Sensitivity or True Positive Rate, it measures the ability of a classification model to capture all relevant instances.
- **ROC Area**: The Area Under the Receiver Operating Characteristic curve, a graphical representation of a model's ability to discriminate between positive and negative classes across different thresholds.
- **Sigmoid**: A type of activation function commonly used in neural networks, producing output values between 0 and 1.
- **SMOTE** (Synthetic Minority Over-sampling Technique): A technique used to address class imbalance in machine learning datasets by generating synthetic examples for the minority class.
- **SVC** (Support Vector Classification): A type of support vector machine algorithm used for classification tasks.
- **Tinder**: A popular mobile dating app that allows users to anonymously swipe to like or dislike other users based on their profiles and interests.
- **Value Weights**: Assigning different levels of importance to values or features in a dataset.
- **Variable Importance:** A measure indicating the contribution of each variable or feature to the predictive power of a model.

6. Bibliography

[1] CLEVENLAND STATE UNIVERSITY and R. F. Rakos, 'JOHN B. WATSON'S 1913 "BEHAVIORIST MANIFESTO":SETTING THE STAGE FOR BEHAVIORISM'S SOCIAL ACTION LEGACY', *Rev. Mex. Análisis Conducta*, vol. 39, no. 2, pp. 99–118, Sep. 2013, doi: 10.5514/rmac.v39.i2.63920.

[2] P. I. Pavlov (1927), 'Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex', *Ann. Neurosci.*, vol. 17, no. 3, pp. 136–141, Jul. 2010, doi: 10.5214/ans.0972-7531.1017309.

[3] E. A. Plazas, 'B. F. Skinner: la búsqueda de orden en la conducta voluntaria', *Univ. Psychol.*, vol. 5, no. 2, pp. 371–383, 2006.

[4] A. G. Zepeda, 'Aportaciones de la psicología conductual a la educación', *Sinéctica*, no. 25, Art. no. 25, 2004, Accessed: Oct. 18, 2023. [Online]. Available: https://sinectica.iteso.mx/index.php/SINECTICA/article/view/264

[5] S. Sánchez Pérez, 'El consumo de tabaco como símbolo de libertad femenina: análisis de las estrategias publicitarias utilizadas por Virginia Slims', 2018, Accessed: Oct. 18, 2023. [Online]. Available: https://uvadoc.uva.es/handle/10324/29340

[6] A. Bartholomew, 'Behaviorism's Impact on Advertising: Then and Now', *Theses Coll. Journal. Mass Commun.*, Dec. 2013, [Online]. Available: https://digitalcommons.unl.edu/journalismdiss/37

[7] E. Palese, 'Zygmunt Bauman. Individual and society in the liquid modernity', *SpringerPlus*, vol. 2, no. 1, p. 191, Apr. 2013, doi: 10.1186/2193-1801-2-191.

[8] D. Curry, 'Dating App Revenue and Usage Statistics (2023)', Business of Apps. Accessed: Oct. 18, 2023. [Online]. Available: https://www.businessofapps.com/data/dating-app-market/

[9] J. Booth, 'Dating Statistics In 2023', Forbes Health. Accessed: Oct. 18, 2023. [Online]. Available: https://www.forbes.com/health/dating/dating-statistics/

[10] J. Vanschoren, 'Speed Dating', OpenML. Accessed: Oct. 18, 2023. [Online]. Available: https://www.openml.org/search?type=data&sort=runs&id=40536&status=active

[11] R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson, 'GENDER DIFFERENCES IN MATE SELECTION: EVIDENCE FROM A SPEED DATING EXPERIMENT*', 2006.

[12] J. Pastor, ‘La Unión Europea quiere dominar la revolución IA con regulación. Para los expertos es una mala idea’, Xataka. Accessed: Jan. 09, 2024. [Online]. Available: https://www.xataka.com/robotica-e-ia/estallido-ia-se-ha-encontrado-viejo-enemigo-regulatorio-a-vuelta-esquina-efecto-bruselas

[13] ‘CRISP-DM IBM’. Accessed: Oct. 24, 2023. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

[14] J. Henrich, S. J. Heine, and A. Norenzayan, ‘The Weirdest People in the World’, Mar. 2009.

[15] I. Almås, A. Kotsadam, E. R. Moen, and K. Røed, ‘The Economics of Hypergamy’, 2019.

[16] A. Sanchis, ‘Los hombres más pobres tienen más probabilidades de estar solteros. La hipergamia, en un gráfico’, Xataka. Accessed: Oct. 24, 2023. [Online]. Available: https://www.xataka.com/magnet/hombres-pobres-tienen-probabilidades-estar-solteros-hipergamia-grafico

[17] D. S. Akers, ‘On Measuring the Marriage Squeeze’, *Demography*, vol. 4, no. 2, pp. 907–924, Jun. 1967, doi: 10.2307/2060328.

[18] D. Selterman and S. Gideon, ‘Experiences of Romantic Attraction Are Similar Across Dating Apps and Offline Dates in Young Adults’, *J. Soc. Psychol. Res.*, pp. 145–163, Sep. 2022, doi: 10.37256/jspr.1220221542.

[19] B. Neyt, S. Vandenbulcke, and S. Baert, ‘Are men intimidated by highly educated women? Undercover on Tinder’, *Econ. Educ. Rev.*, vol. 73, p. 101914, Dec. 2019, doi: 10.1016/j.econedurev.2019.101914.

[20] G. Tyson, V. C. Perta, H. Haddadi, and M. C. Seto, ‘A first look at user activity on tinder’, in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2016, pp. 461–466. doi: 10.1109/ASONAM.2016.7752275.

[21] A. Fiore, L. Taylor, G. Mendelsohn, and M. Hearst, ‘Assessing Attractiveness in Online Dating Profiles’, presented at the Conference on Human Factors in Computing Systems - Proceedings, Apr. 2008, pp. 797–806. doi: 10.1145/1357054.1357181.

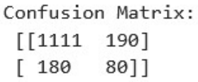
[22] S. T. Tong, J. T. Hancock, and R. B. Slatcher, ‘Online dating system design and relational decision making: Choice, algorithms, and control’, *Pers. Relatsh.*, vol. 23, no. 4, pp. 645–662, Dec. 2016, doi: 10.1111/pere.12158.

[23] L. L. Sharabi, ‘Exploring How Beliefs About Algorithms Shape (Offline) Success in Online Dating: A Two-Wave Longitudinal Investigation’, *Commun. Res.*, vol. 48, no. 7, pp. 931–952, Oct. 2021, doi: 10.1177/0093650219896936.

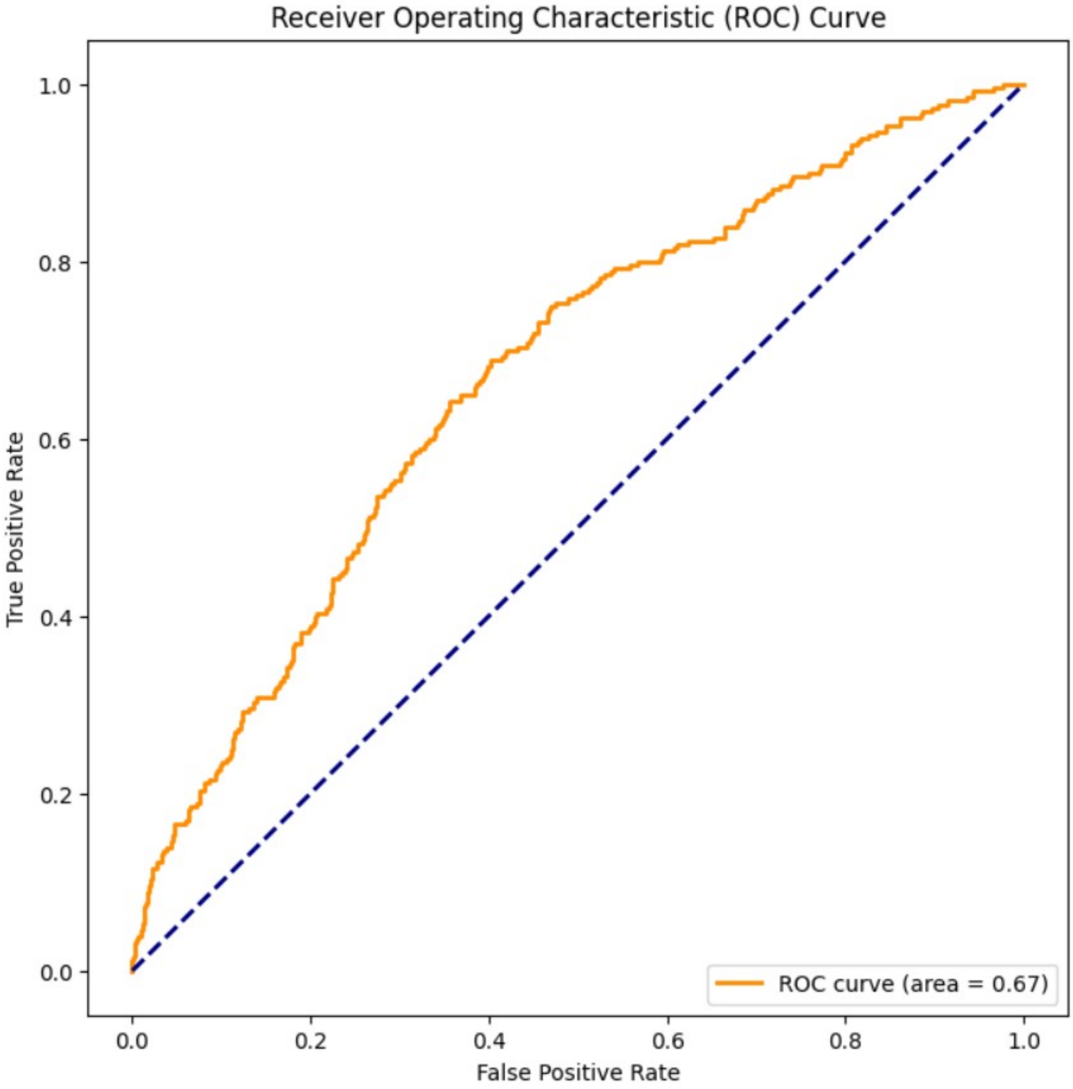
[24] C. A. Maya Rodríguez, ‘Designing a dating application which helps to predict enduring relationships’, masters, E.T.S. de Ingenieros Informáticos (UPM), 2020. Accessed: Oct. 24, 2023. [Online]. Available: https://oa.upm.es/65425/

7. Appendices

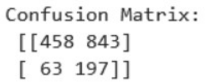
FINAL MODEL ACCURACY SMOTE:



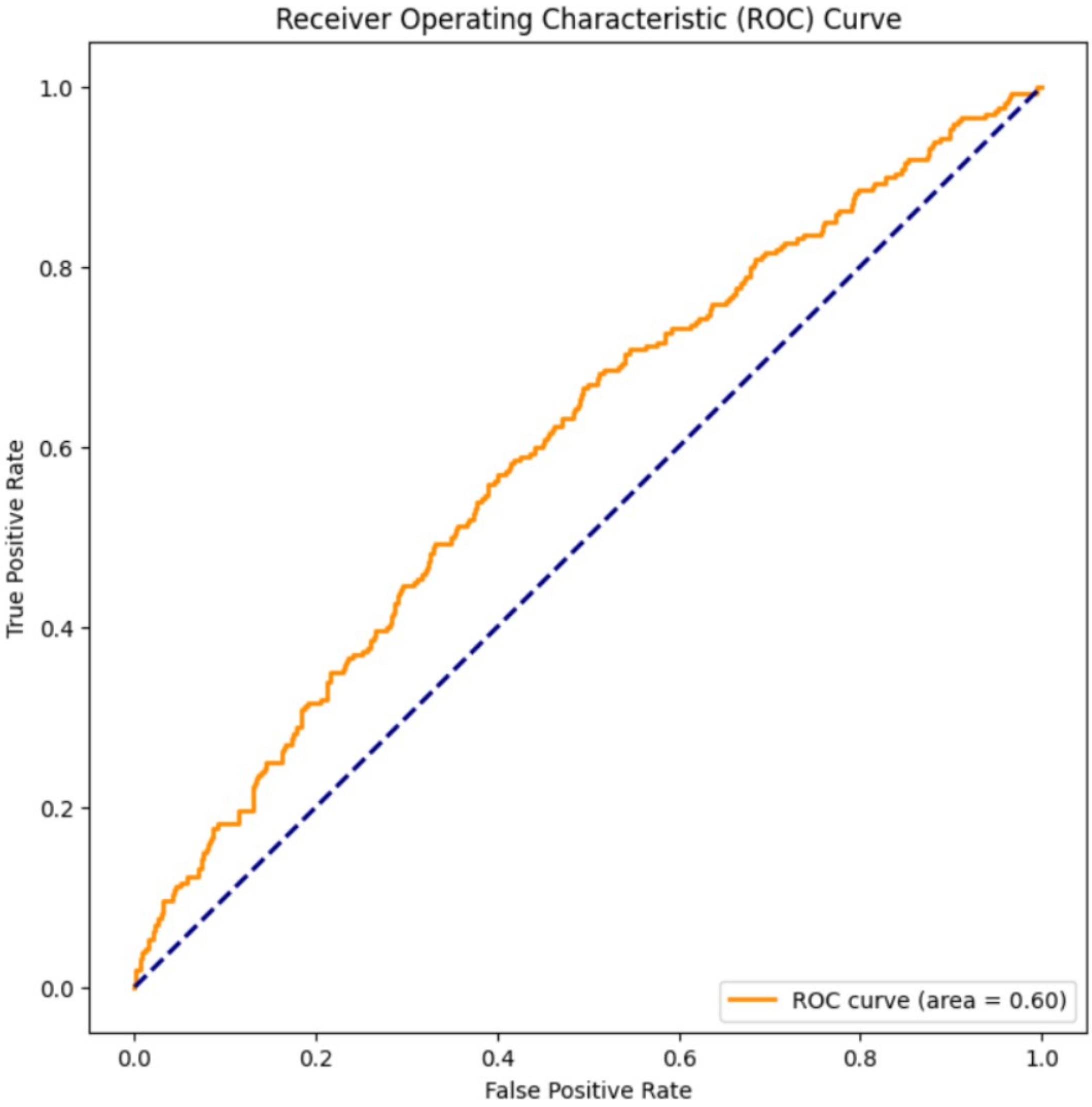
```
Confusion Matrix:
 [[1111  190]
 [ 180   80]]
```



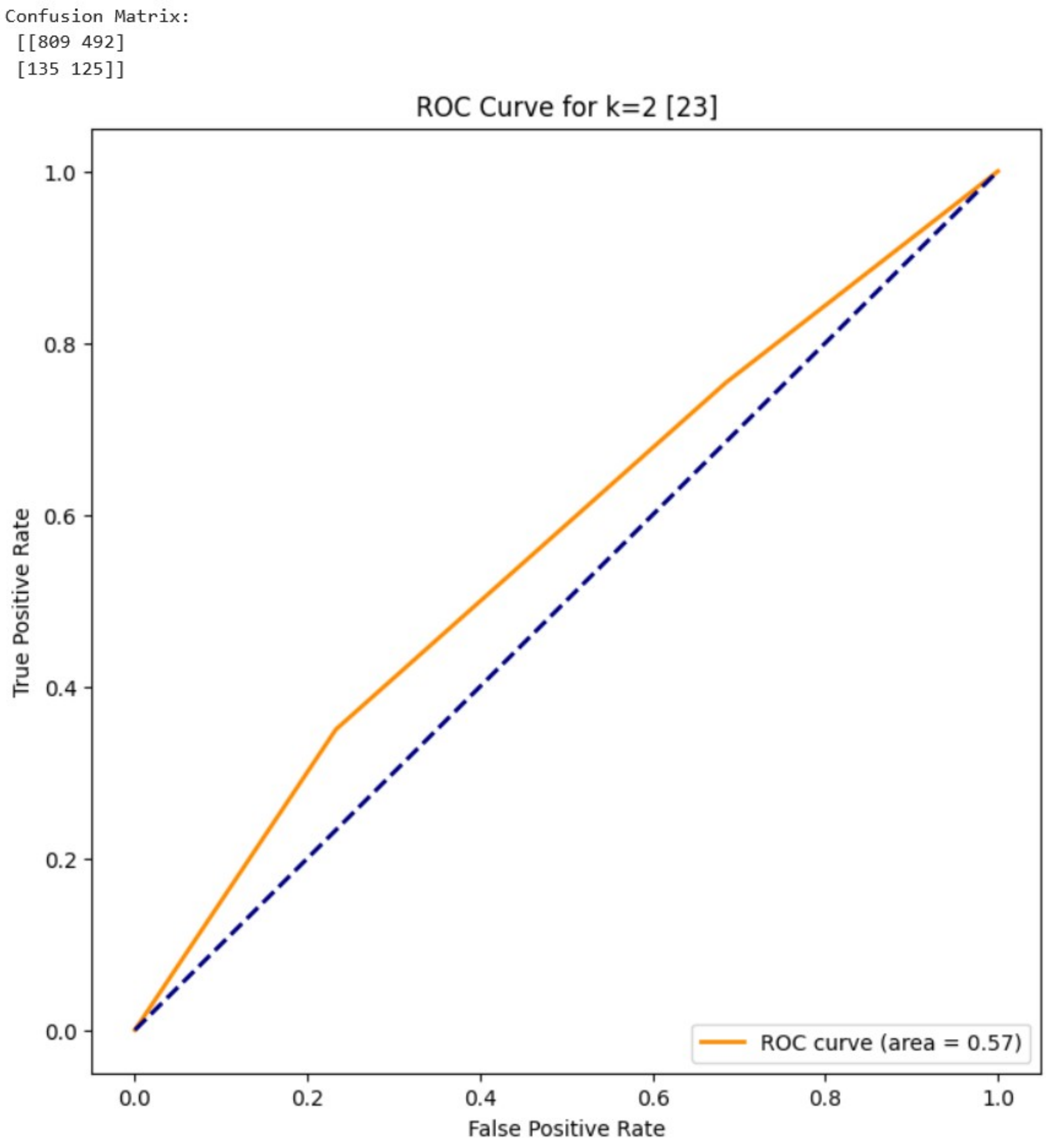
FINAL MODEL RESAMPLE RECALL:



```
Confusion Matrix:
 [[458 843]
 [ 63 197]]
```

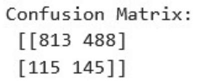


KNN RESAMPLE:

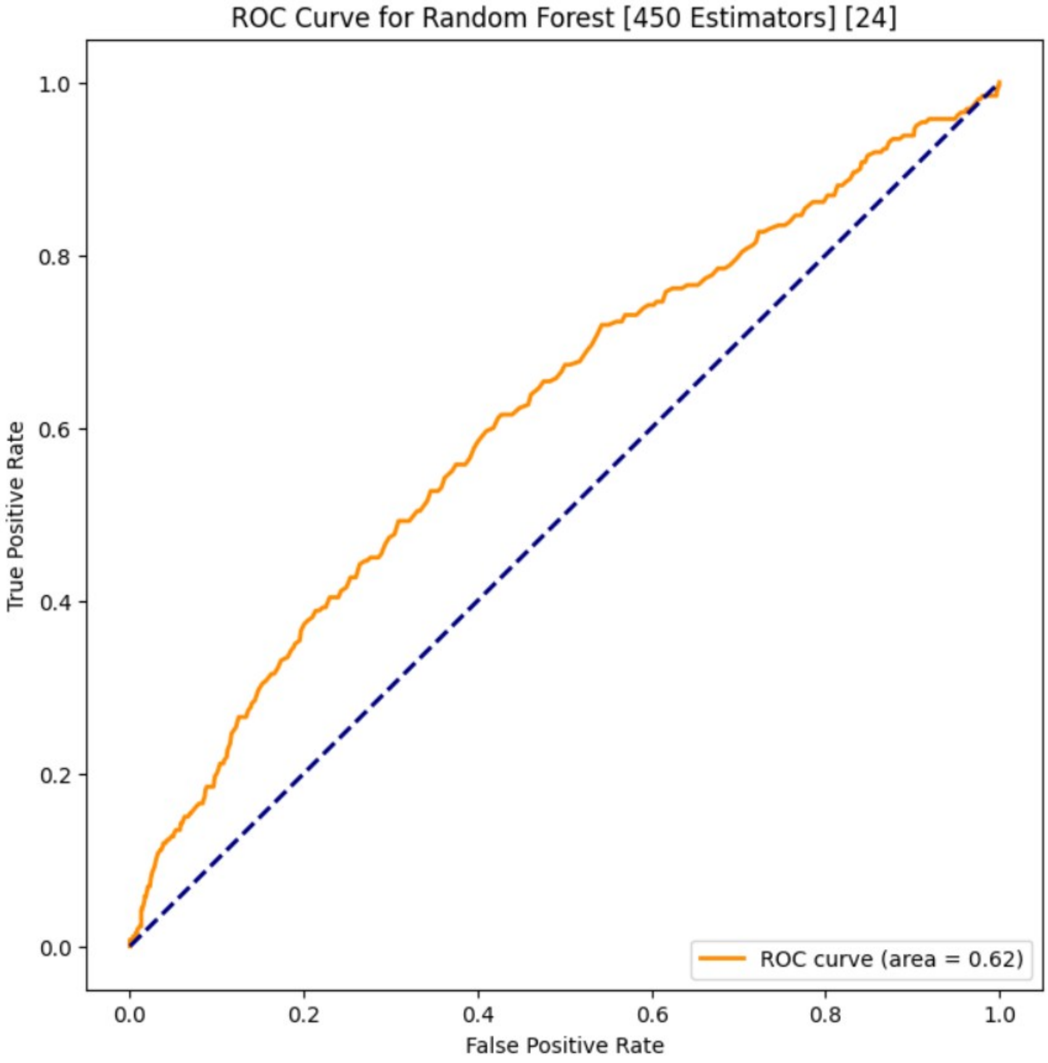


```
Confusion Matrix:
 [[809 492]
 [135 125]]
```

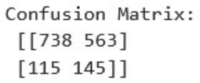
RANDOM FOREST RESAMPLE:



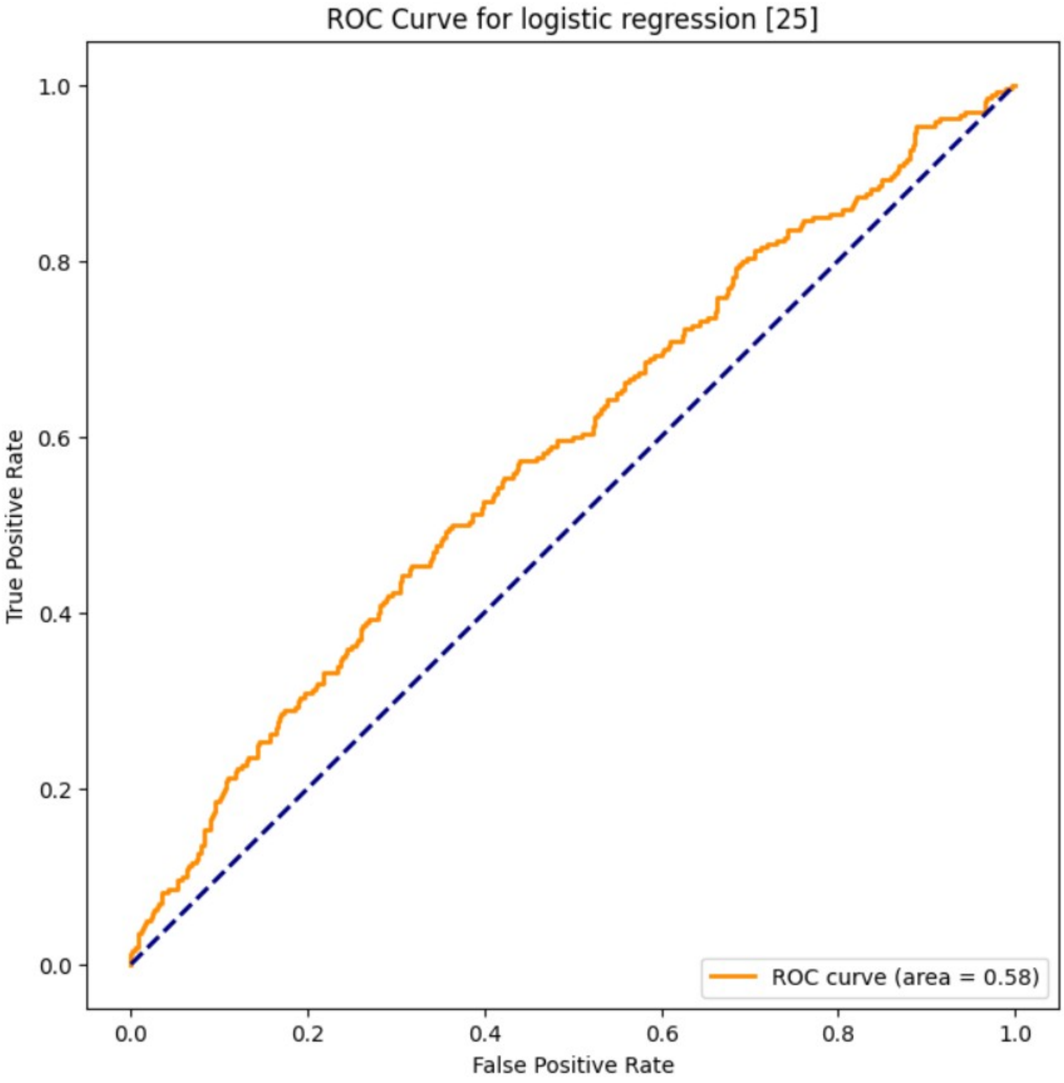
```
Confusion Matrix:
 [[813 488]
 [115 145]]
```



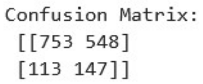
LOGISTIC REGRESSION RESAMPLE:



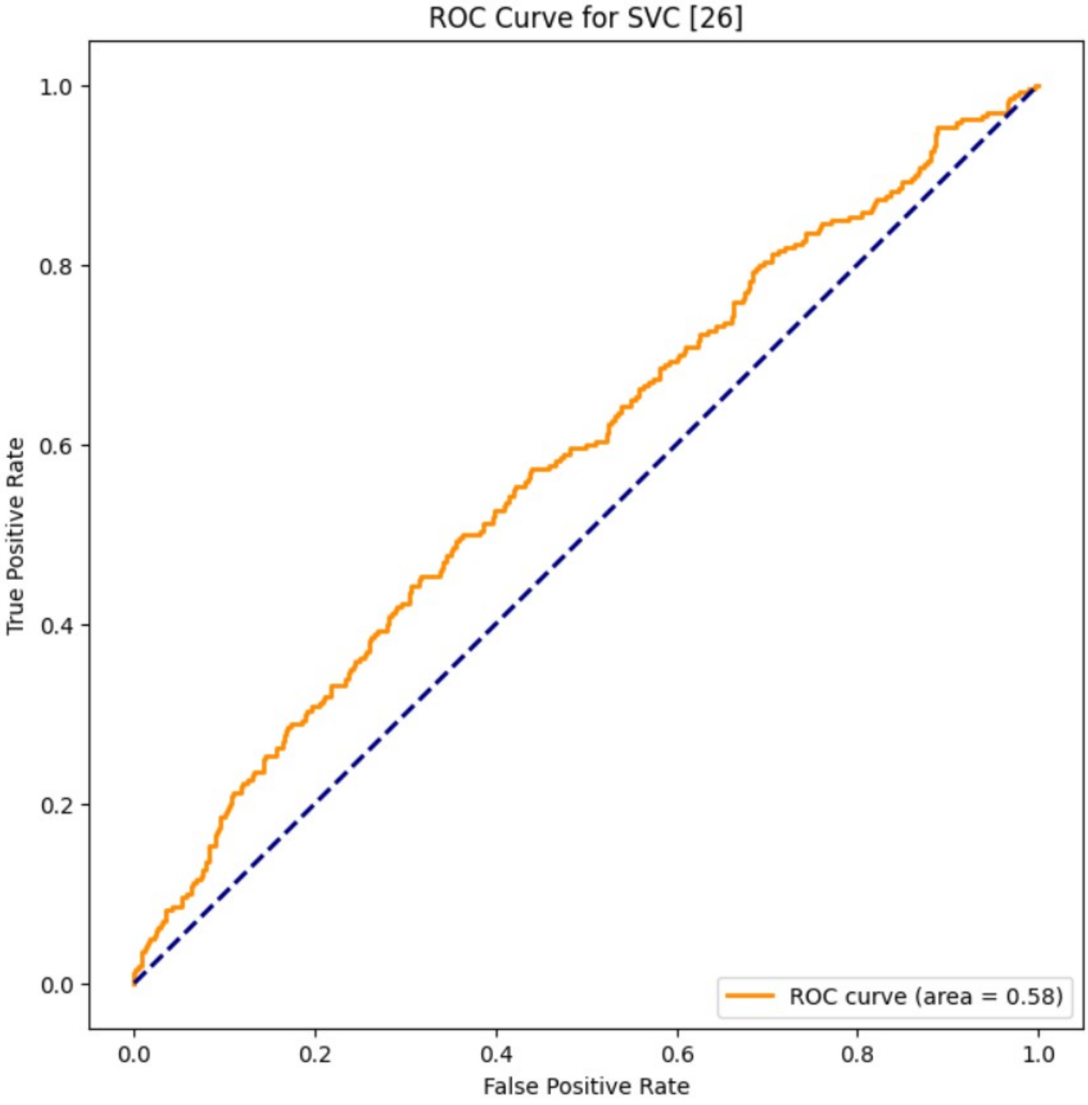
```
Confusion Matrix:
 [[738 563]
 [115 145]]
```



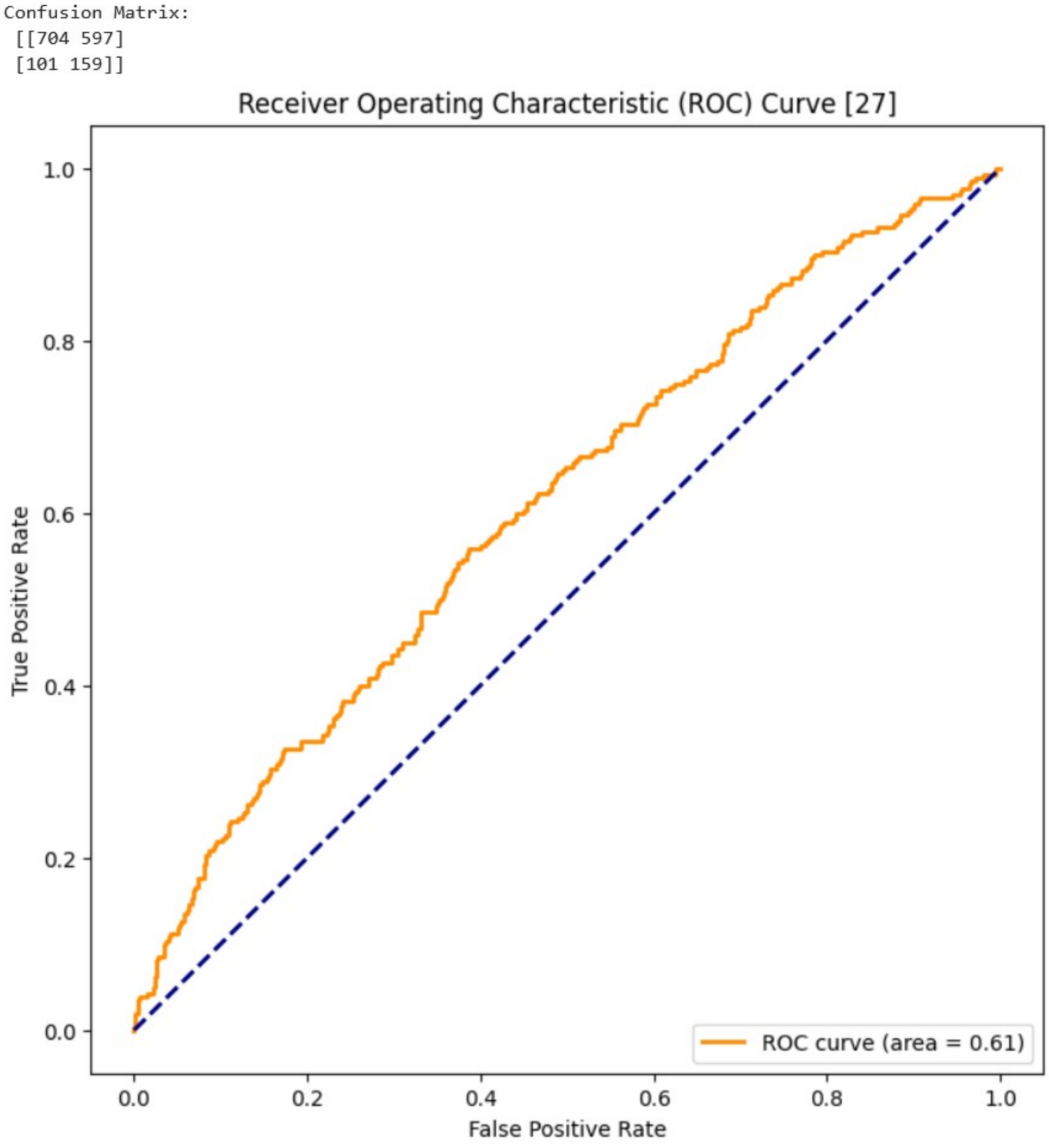
SVC RESAMPLE:



```
Confusion Matrix:
 [[753 548]
 [113 147]]
```

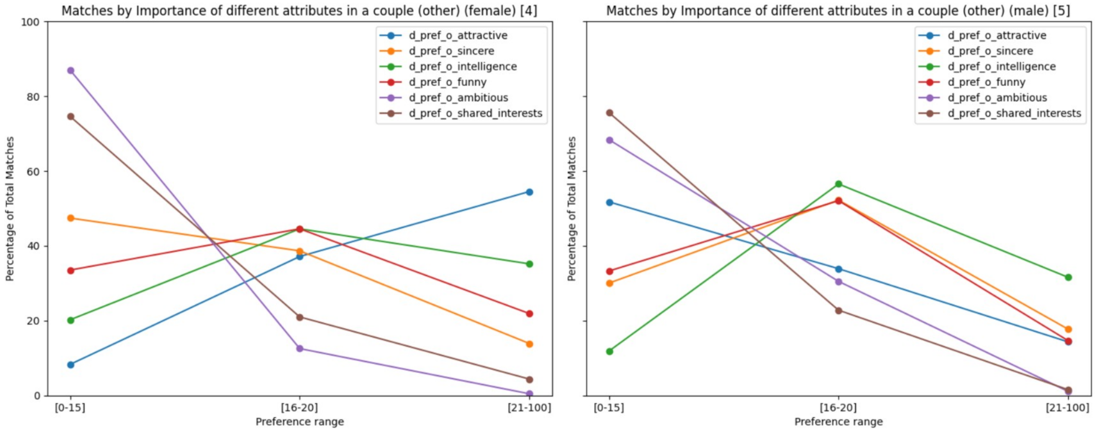


UNTUNED NEURAL NETWORK RESAMPLE:

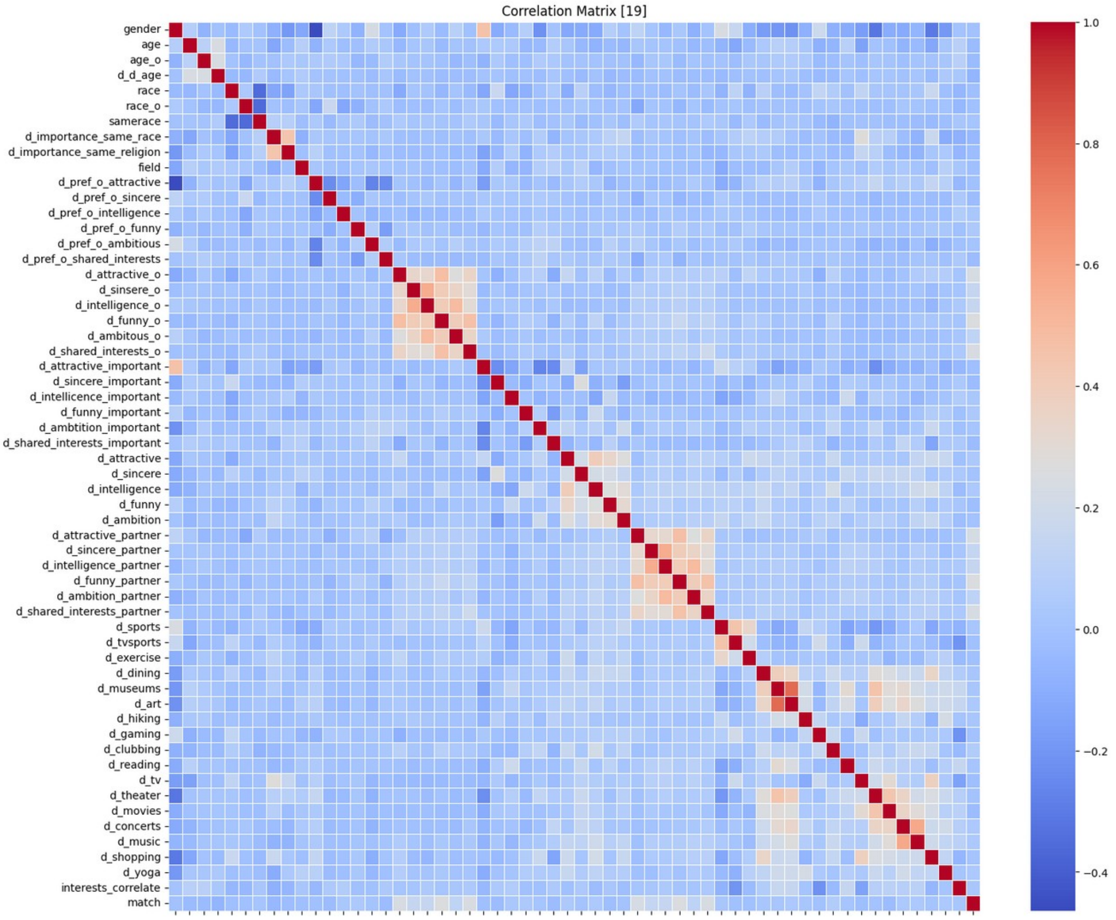


```
Confusion Matrix:
 [[704 597]
 [101 159]]
```

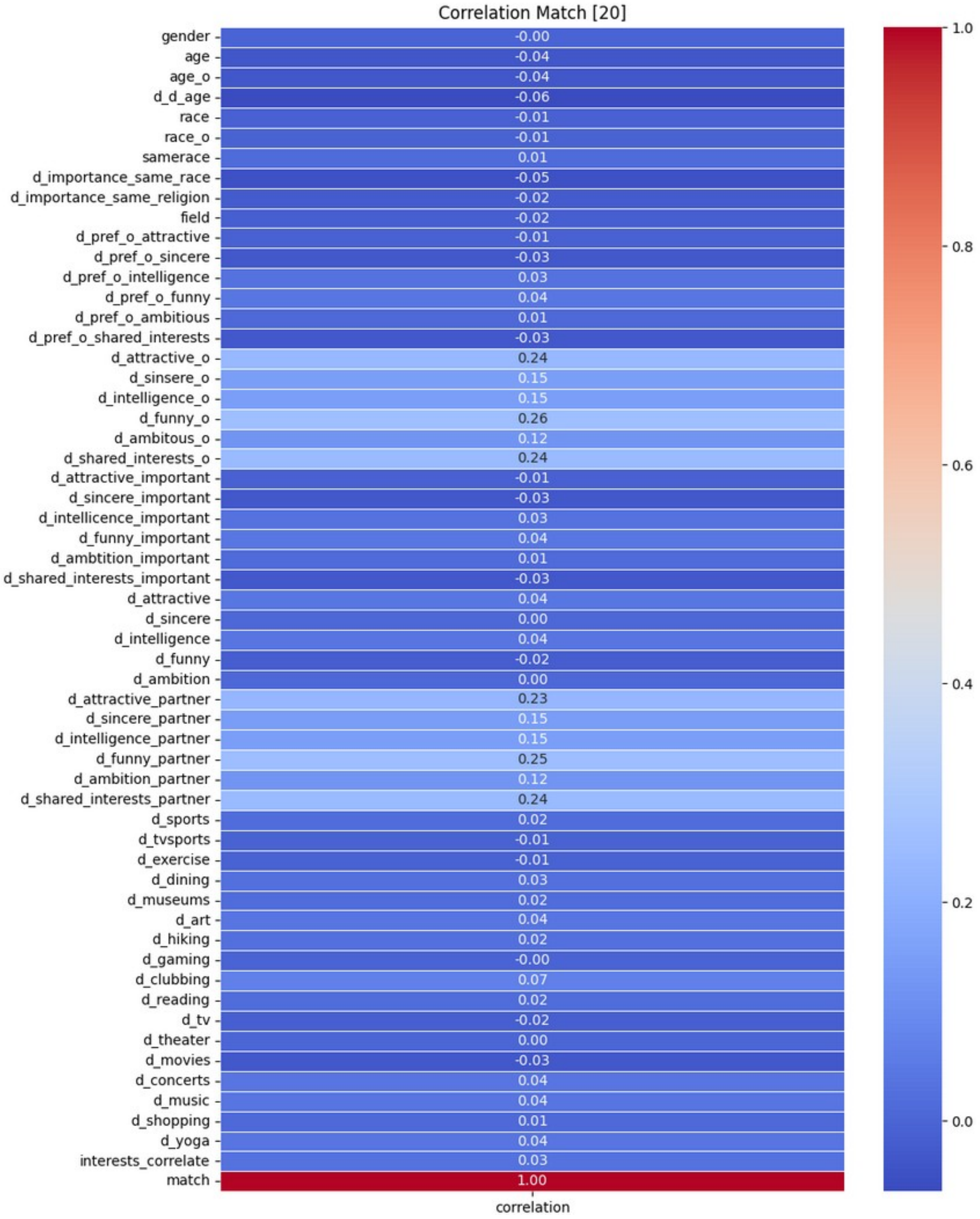
MATCHES OF THE OTHER BY IMPORTANCE



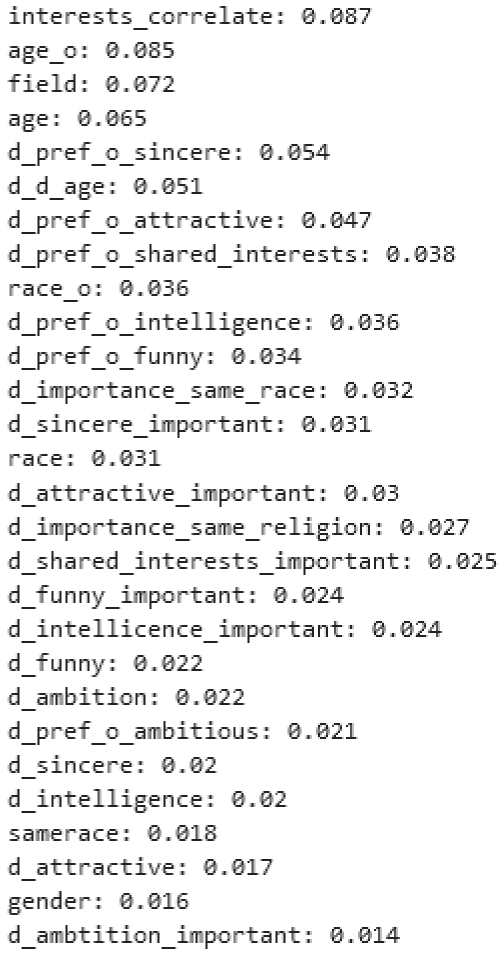
CORRELATION MATRIX



FINAL CORRELATION MATRIX MATCH

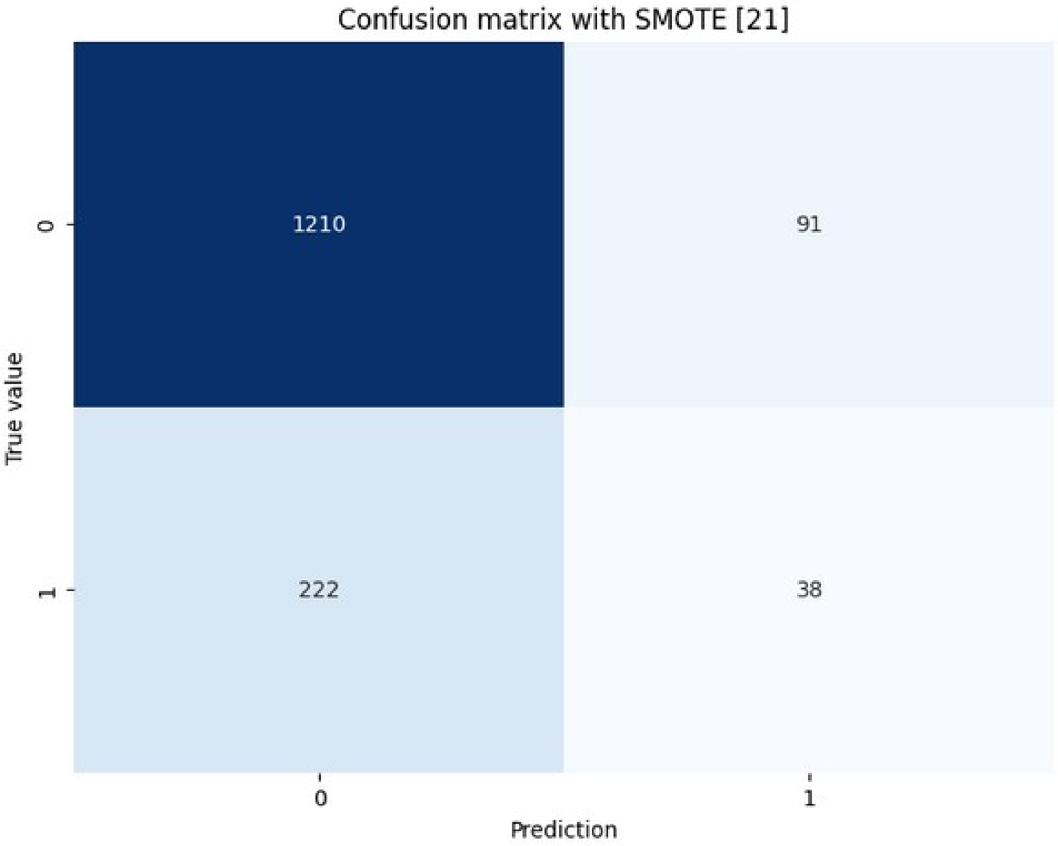


IMPORTANCE ANALYSIS OF FINAL ATTRIBUTES

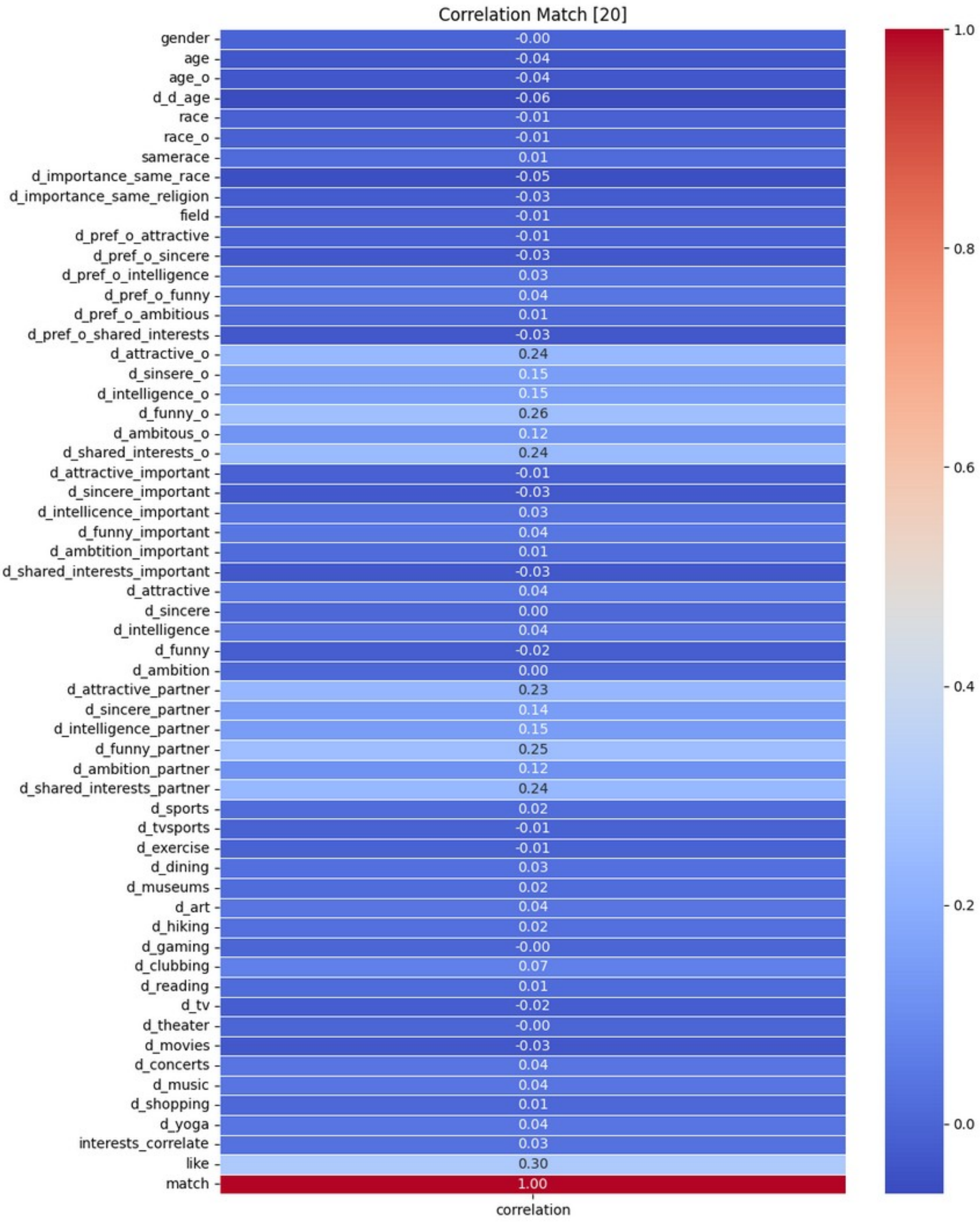


```
interests_correlate: 0.087
age_o: 0.085
field: 0.072
age: 0.065
d_pref_o_sincere: 0.054
d_d_age: 0.051
d_pref_o_attractive: 0.047
d_pref_o_shared_interests: 0.038
race_o: 0.036
d_pref_o_intelligence: 0.036
d_pref_o_funny: 0.034
d_importance_same_race: 0.032
d_sincere_important: 0.031
race: 0.031
d_attractive_important: 0.03
d_importance_same_religion: 0.027
d_shared_interests_important: 0.025
d_funny_important: 0.024
d_intellicence_important: 0.024
d_funny: 0.022
d_ambition: 0.022
d_pref_o_ambitious: 0.021
d_sincere: 0.02
d_intelligence: 0.02
samerace: 0.018
d_attractive: 0.017
gender: 0.016
d_ambtition_important: 0.014
```

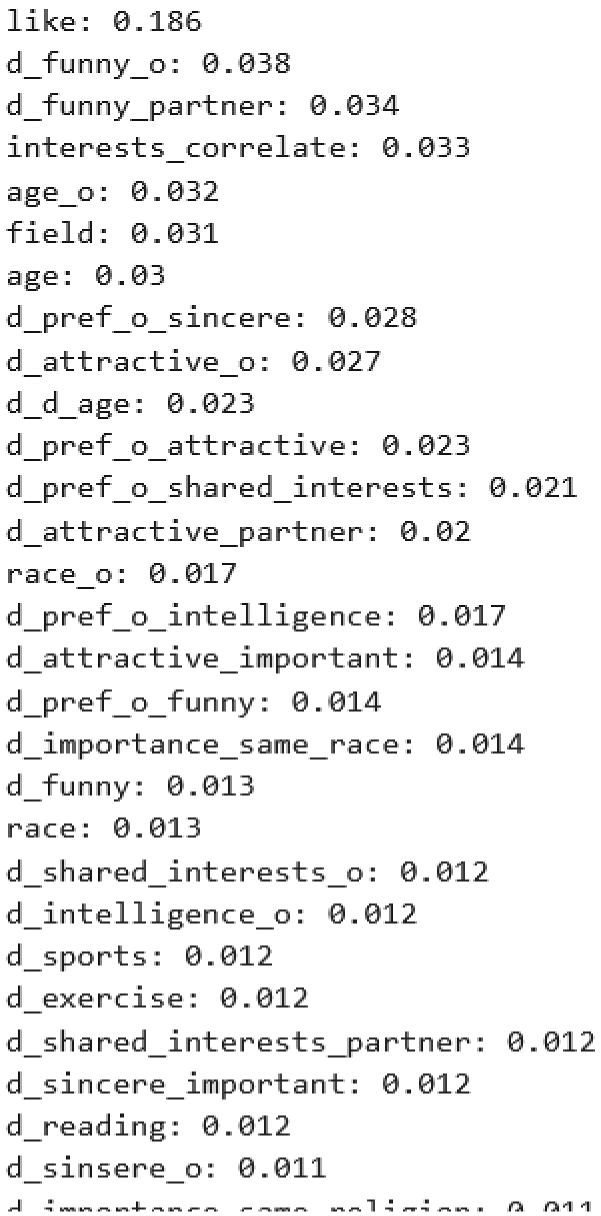
IMPORTANCE ANALYSIS CONFUSION MATRIX



CORRELATION MATRIX MATCH WITH LIKE

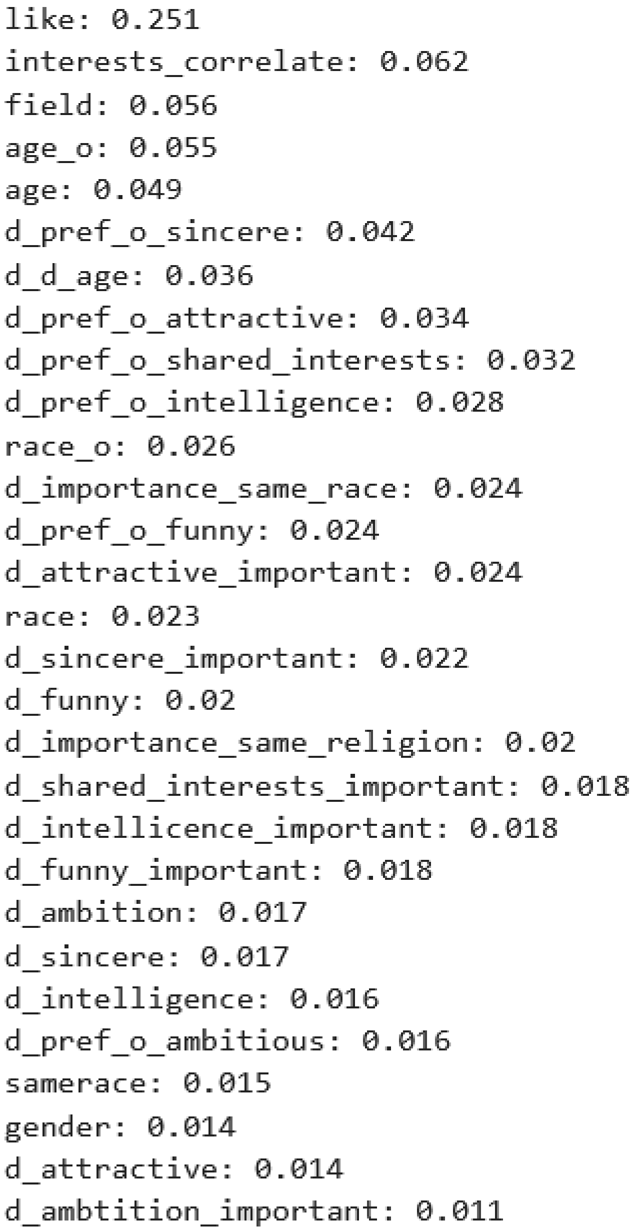


VALUE IMPORTANCES LIKE NOT CLEAN



```
like: 0.186
d_funny_o: 0.038
d_funny_partner: 0.034
interests_correlate: 0.033
age_o: 0.032
field: 0.031
age: 0.03
d_pref_o_sincere: 0.028
d_attractive_o: 0.027
d_d_age: 0.023
d_pref_o_attractive: 0.023
d_pref_o_shared_interests: 0.021
d_attractive_partner: 0.02
race_o: 0.017
d_pref_o_intelligence: 0.017
d_attractive_important: 0.014
d_pref_o_funny: 0.014
d_importance_same_race: 0.014
d_funny: 0.013
race: 0.013
d_shared_interests_o: 0.012
d_intelligence_o: 0.012
d_sports: 0.012
d_exercise: 0.012
d_shared_interests_partner: 0.012
d_sincere_important: 0.012
d_reading: 0.012
d_sinsere_o: 0.011
d_importance_same_religion: 0.011
```

VALUE IMPORTANCES LIKE CLEAN:



```
like: 0.251
interests_correlate: 0.062
field: 0.056
age_o: 0.055
age: 0.049
d_pref_o_sincere: 0.042
d_d_age: 0.036
d_pref_o_attractive: 0.034
d_pref_o_shared_interests: 0.032
d_pref_o_intelligence: 0.028
race_o: 0.026
d_importance_same_race: 0.024
d_pref_o_funny: 0.024
d_attractive_important: 0.024
race: 0.023
d_sincere_important: 0.022
d_funny: 0.02
d_importance_same_religion: 0.02
d_shared_interests_important: 0.018
d_intellicence_important: 0.018
d_funny_important: 0.018
d_ambition: 0.017
d_sincere: 0.017
d_intelligence: 0.016
d_pref_o_ambitious: 0.016
samerace: 0.015
gender: 0.014
d_attractive: 0.014
d_ambtition_important: 0.011
```