Machine Learning Mini Project: Speed Dating and The Match-Determining Factors

Esra Kantarcı and Ayça Yiğit

20160808023 - 20160808051

1 Introduction

The dataset is Speed Dating Experiment from Kaggle. Data for this dataset was gathered from participants in experimental speed dating events. During which the attendees would have a date with every other participant and at the end 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 consists of participants' demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information. For example, they are wanted to score their hobbies in specific interest areas(hobbies) from 1 to 10. So, actually, it is not just a dataset for matching, but also a dataset for analyzing the human behavior, interest areas, expectations, and choices by gender, age, race, etc.

With this project, we wanted to look up at what criteria would matter in a date and some interesting correlations in data.

2 Data Preprocessing

At first, it is needed to split the data with "," seperator as we read it into the dataframe. We could also use ISO-8859-1 encoding while reading the csv file.

Before starting to work we had to take a look at the dataset and the columns inter-correlations. This is mainly because the dataset is too huge that it does not fit in the screen. So we needed to split the columns into new sets to evaluate the hypothesis we had in mind.

After splitting the columns we checked the whole dataset and found out that it has too much missing data, which we replaced with corresponding mean values of their respective columns to overcome the effects.

Later on, we looked at correlation between features we had split for readability purposes and chose the most useful ones.

We did not need normalization, standardization, or label encoding utilities in the features we decided to use. So, in the data-preprocessing stage, we split the columns, filled null values with respective means and splitted the data into test-train values by using sklearn functions by 0.2 factor.

3 Methodology

Once the data was preprocessed, we used some common algorithms to train our model and see how accurate they are. These algorithms were implemented:

Random Forest Linear Regression Support Vector Machine (SVM) Decision Tree K – Nearest Neighbour (KNN) AutoML Regression

Since the results were not accurate as expected, we had changed the column sets and try the built-in training methods again and again after checking the correlations heatmap of the sets we splitted.

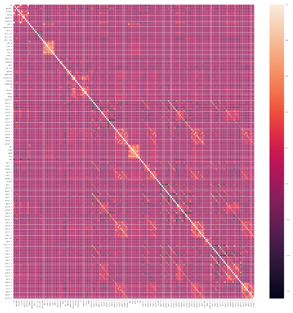

Figure 1: Heatmap of All Features Which Lead Us to Split Columns

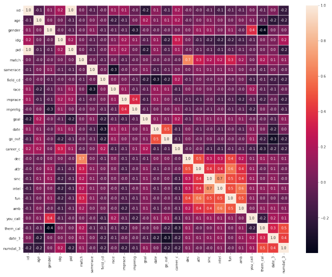

Figure 2: Personality Attributes Correlations Heatmap with Better Hints

The first sets we had created were about personal choices, personal hobbies and personal skills+call history. As outputs, we considered different values: like, prob, dec, match, them_cal, date. The most efficient and accurate results were given on dec and match cases.

You can check the dictionary of the columns we attached to the homework for the further information on the features. What we initially thought as hypothesis was the personal skills and preferences (as they are shown in the columns_fundamental) should be determining the probabilities of liking, going on a date, getting a call for a date, matched at the speed dating. We checked the accuracy of the training results, then for the best ones, we got precision-recall-f1 scores as well.

4 Experiment Results

We tried to find the most relevant features and applied previously mentioned regression and classification algorithms on them. Some of the correlations were just as expected, for example if you like music, you are more likely to go concerts.

If you are women, you tend to like shopping than the man. You can check the correlations in attached figures.

Here is the key points of the results:

- If you are woman, you are more likely to get call for date and less likely to call for date.
- If you are woman, you are more likely to say no for the further dates, but if you are man going on date or not going on a date probabilities are just too close to each other.
- As you get older, you expect more happiness from the partner. But here is an interesting plot twist: after age of 35 you are more probable to go out with the matched partner. But as it is expected, when you get older, you got less matches.
- Attractiveness is the most important feature, and there is positive 0.6 correlation with attractiveness and fun personality, which indicates people often think you are attractive if you are fun person or find you funny if you are attractive. So, let’s hope your partner laughs at your jokes!
- You do not necessarily go on a date if you call the partner, but if you got called by them, you are likely to go on a date.
- Lawyers, Academic-Researchers, Banking-Consulting-CEO-Entrepreneur-Admins and people who work for International-Humanitarian Affairs are more likely to go on further dates, possibly due elocutionary skills and persuasion experiences.
- If you think your partner is probable to say “yes” to you, you are more likely to go on further dates, which indicates you actually hope that s/he says “yes” to you.
- The most important analysing histograms are on the attachment, using columns: ’age’,’gender’,’idg’, ’pid’,’match’,’samerace’, ’field_cd’,’race’,’imprace’, ’imprelig’, ’goal’, ’date’,’go_out’, ’caree_c’, ’dec’,’attr’, ’sinc’,’intel’,’fun’,’amb’ ,’you_call’,’them_cal’, ’date_3’, ’numdat_3’,’prob’,’like’,’dec’

For the number of dates and probability of liking prediction, we had used regression. For matching or deciding to date purposes, we used classification (since it is 0 or 1 only.)

4.1 Some Example Outcomes

Input Features: Personality Assets, Call Features, Matching Results, Careers
Target Value: Number of Dates

- Decision Tree Model Error Rate: 0.16
- Naive Bayes Model Error Rate: 0.25

- Logistic Model Error Rate: 0.14

Input Features: Expectations, Gender, Age Target Value: ”match”

- Logistic Model Error Rate: 0.17 (Predicted all 0)
- Decision Tree Model Error Rate: 0.17 (Predicted 28 as match, but misclassified some)
- Naive Bayes Model Error Rate: 0.17 (Predicted 39 as match but misclassified some)

The best and accurate results were given by logistic regression on ”decision to date” classification. Input Features: Personality Assets, Call Features, Matching Results, Careers
Target Value: ”dec” (Decision to go on a date after speed dating meeting)

- Logistic Model Error Rate: 0.10
- Neural Network with 5-2 Hidden Layers Model Error Rate: 0.35

On weighted version, Classification using Logistic Regression scores are; precision: 0.899, recall: 0.898, fscore: 0.899

Auto Sklearn had failed, because there were 395 matches in 2095 records, so generally it learnt the majority of the outputs (which is 0) with giving DummyScore warnings.

5 Conclusion

In conclusion, it is important to choose the features correctly in this kind of big datasets. Not always automated learning protocols are suitable for the better solutions.

Logistic Regression is more accurate than Random Forest, Decision Trees, Naive Bayes algorithms, Neural Network and KNN on this problem set.

It is normal to be turned down at speed dating, but if you are attractive or funny, you are luckier than the others. If you call for further dates, you increase your chances to get the date; but it is more preferable if you get the calls. If you are a woman and call the partner for the date, you have a higher chance to get the date than the men who called the partner.

Skills of oratory and persuasion and having a reputable career will get you to the further dates. If you are between 25-30 old, we recommend you to go try speed dating because the matching correlation is positive. Try to invest in your career, speech, jokes, appearance and fashion; because these are mainly what partners look for.

More comprehensive evaluation will be put on Kaggle, this was just a summary due to page-constraints of the project task.

6 Tools and Utilities that are Used

The tools we had used in the analysis and training are:

- Google Colab: https://colab.research.google.com/drive/16f57Dl3E6vlF0Sfrkhee-fRufeUTO19S?usp=sharing
- Kaggle: Dataset: https://www.kaggle.com/annavictoria/speed-dating-experiment and Similar Notebooks
- Models: Random Forest, Naive Bayes, Logistic Regression, Decision Trees, Neural Networks, Support Vector Machines, Auto-ML.
- Libraries: Sklearn, MatPlotLib, Seaborn, Auto-Sklearn, Pandas, Numpy
- Scoring: R2, Sklearn: Precision, Recall, F1 scores.

References

[1] Kaggle - Speed Dating Experiment Dataset
`https://www.kaggle.com/annavictoria/speed-dating-experiment`

[2] Kaggle - Ugly Truth of People Decisions in Speed Dating
`https://www.kaggle.com/jph84562/the-ugly-truth-of-people-decisions-in-speed-dating`

[3] Kaggle -
`https://www.kaggle.com/berkesun/earthquakes-in-turkey-1910-2017-basic-eda`

[4] Kaggle - The Secret to Getting The Second Date
`https://www.kaggle.com/aeshen/the-secret-to-getting-the-second-date`

[5] Kaggle - Data Science Book of Love
`https://www.kaggle.com/lucabasa/the-data-science-book-of-love`

[6] Kaggle - What Matters Most in Four Minutes
`https://www.kaggle.com/zurfer/what-matters-most`

[7] KDNuggets - Easy Quide to Data Preprocessing in Python
`https://www.kdnuggets.com/2020/07/easy-guide-data-preprocessing-python.html`

[8] Data Flair Training - Data Preprocessing, Analysis and Visualization
`https://data-flair.training/blogs/python-ml-data-preprocessing/`

[9] Jim Frost - How to Interpret R-Squared in Regression Analysis
`https://statisticsbyjim.com/regression/interpret-r-squared-regression/`

7 Figures for Further Analysis if Interested

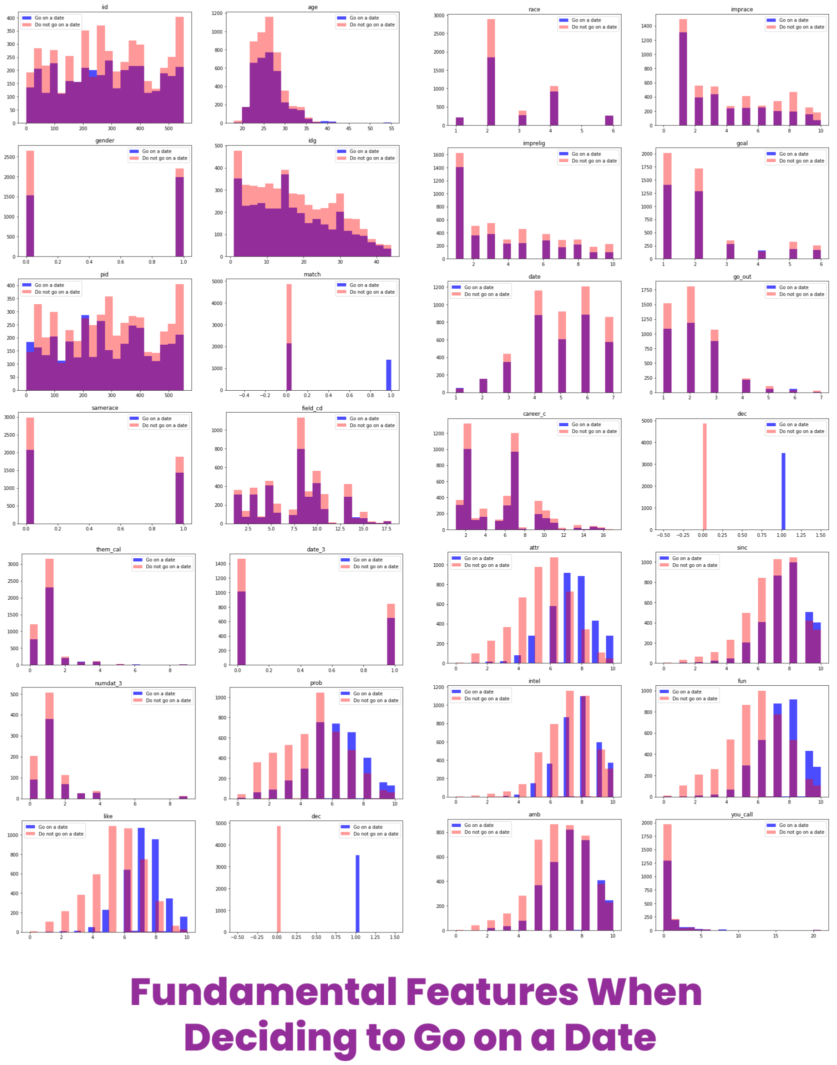

Figure 3: The Histograms of Features When Deciding To Go Out with Partner

Better Resolution on Google Colab

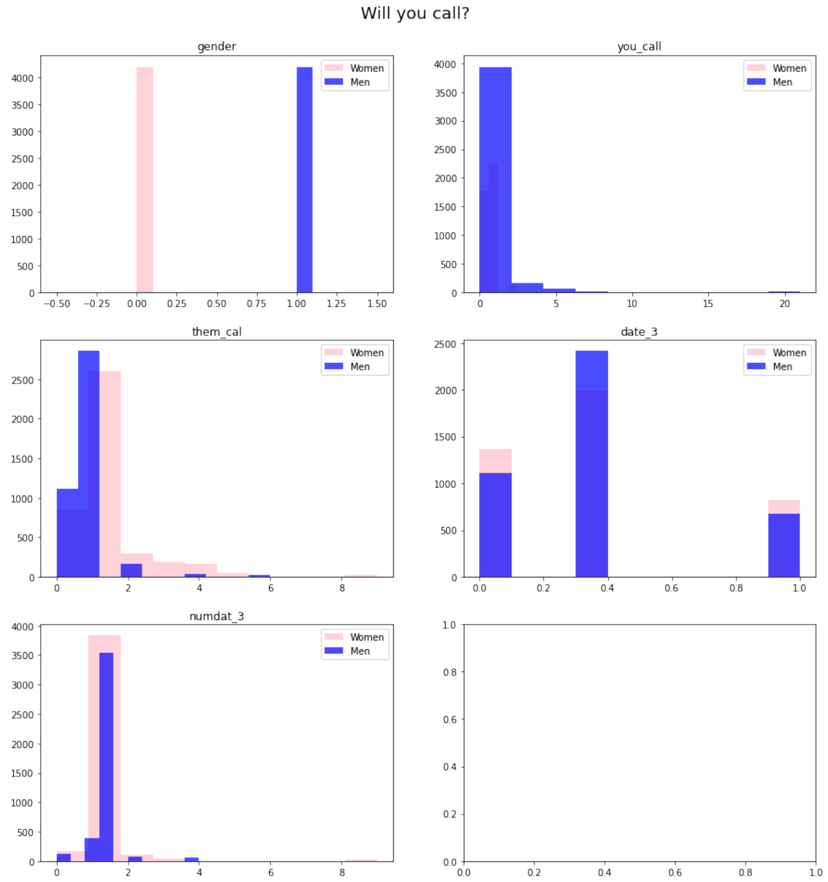

Figure 4: Who Calls? Does the Call Leads to Date?

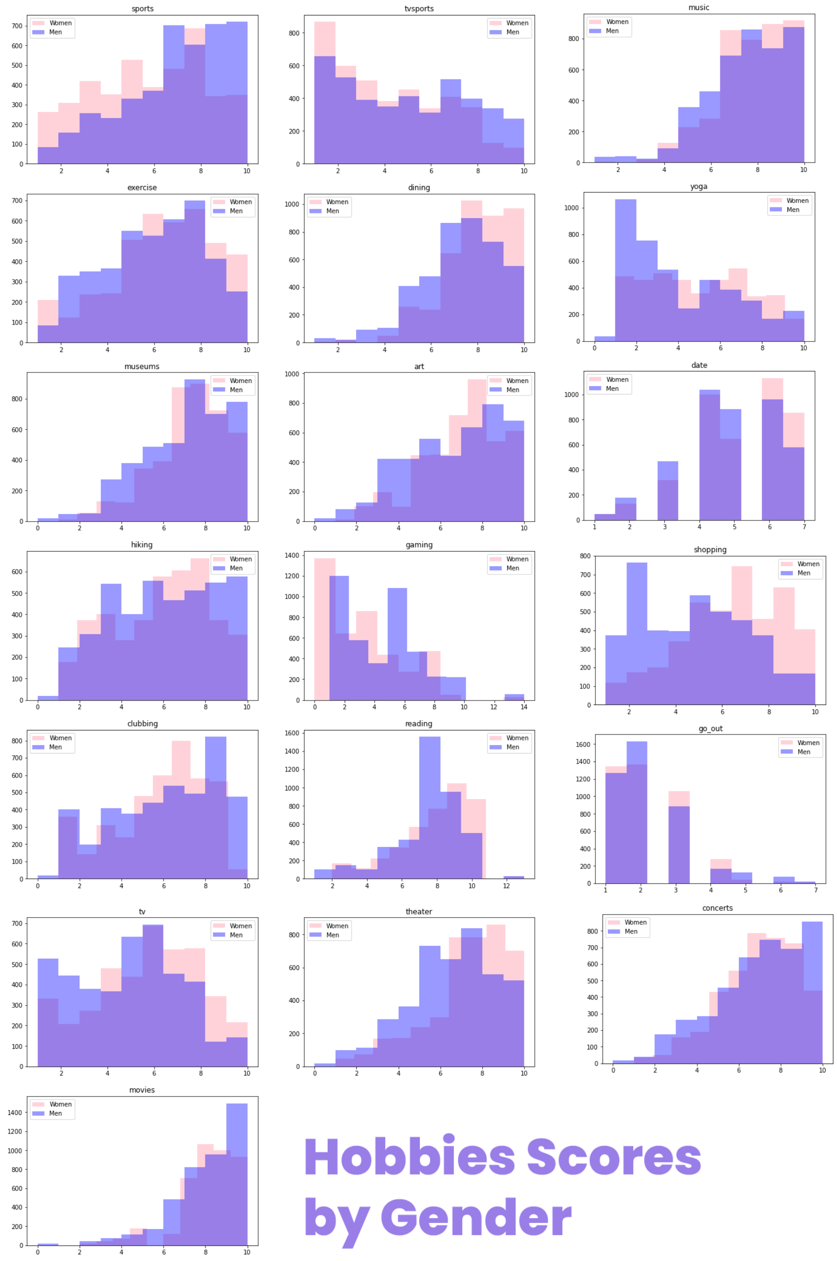

Figure 5: Hobbies by Genders, Better Resolution on Google Colab

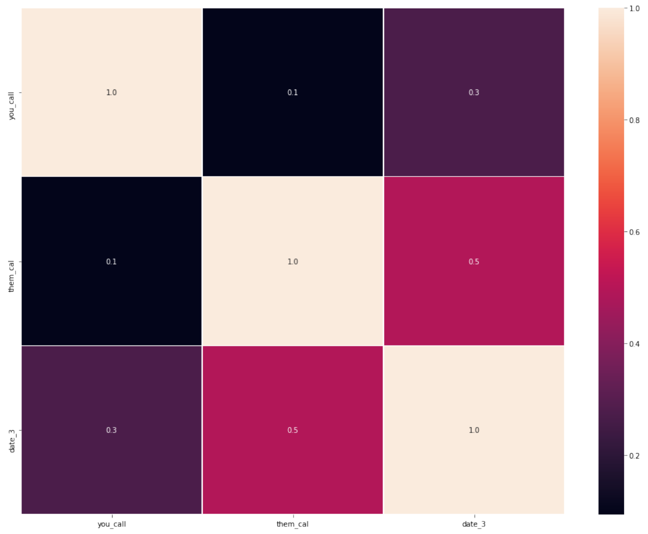

Figure 6: HeatMap of Calls-Date Correlation

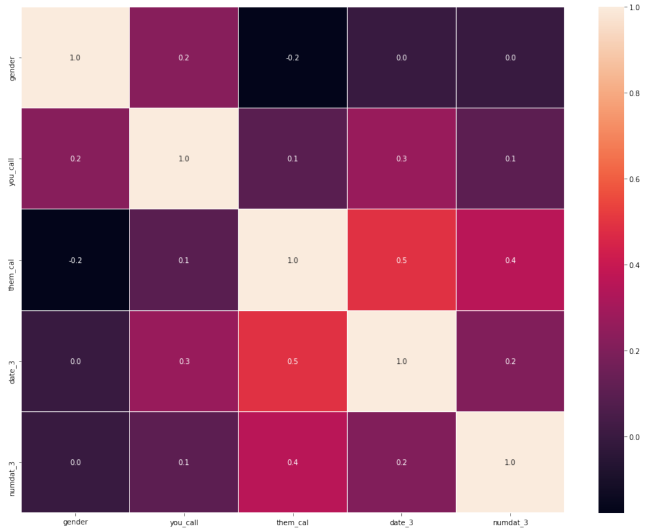

Figure 7: HeatMap of Calls-Gender Correlation

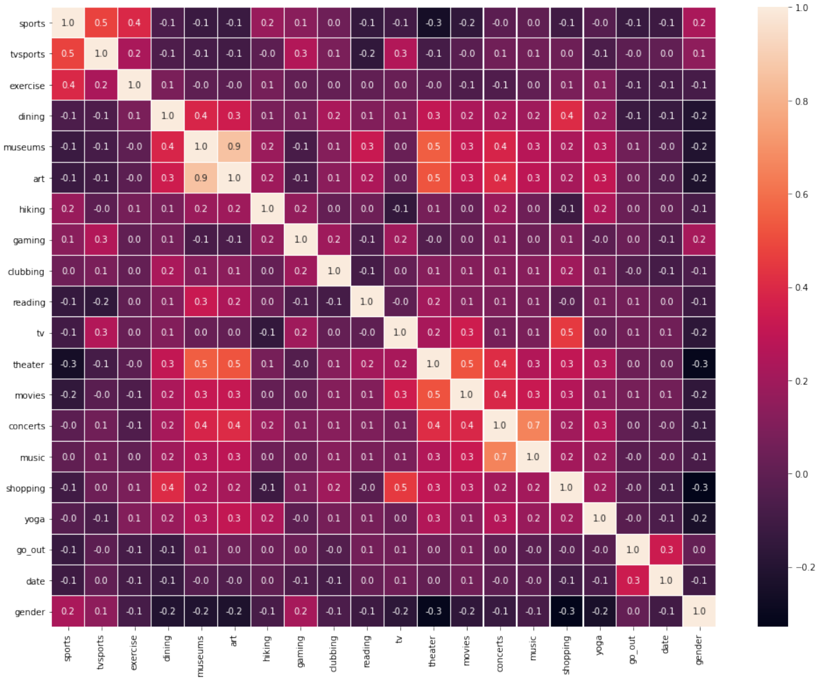

Figure 8: HeatMap of Hobbies Correlation

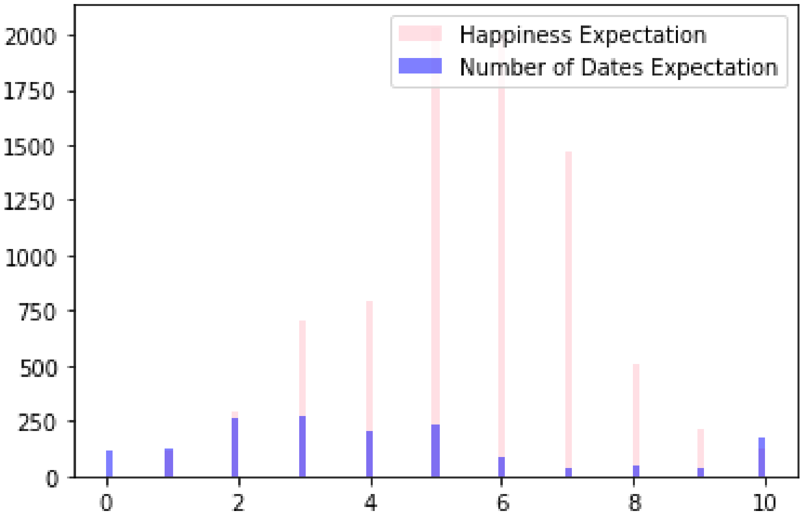

Figure 9: Expectations Distribution