Predictive Analytics

Speed Dating Dataset

Predicting the chance of a second date

Dozent:
Prof. Dr. M. Heckmann

Gruppe 4:
Dario Leon (80152)
Matias Volman (81595)
Christian Gunzelmann (82493)
Pierre Bonnin (83385)

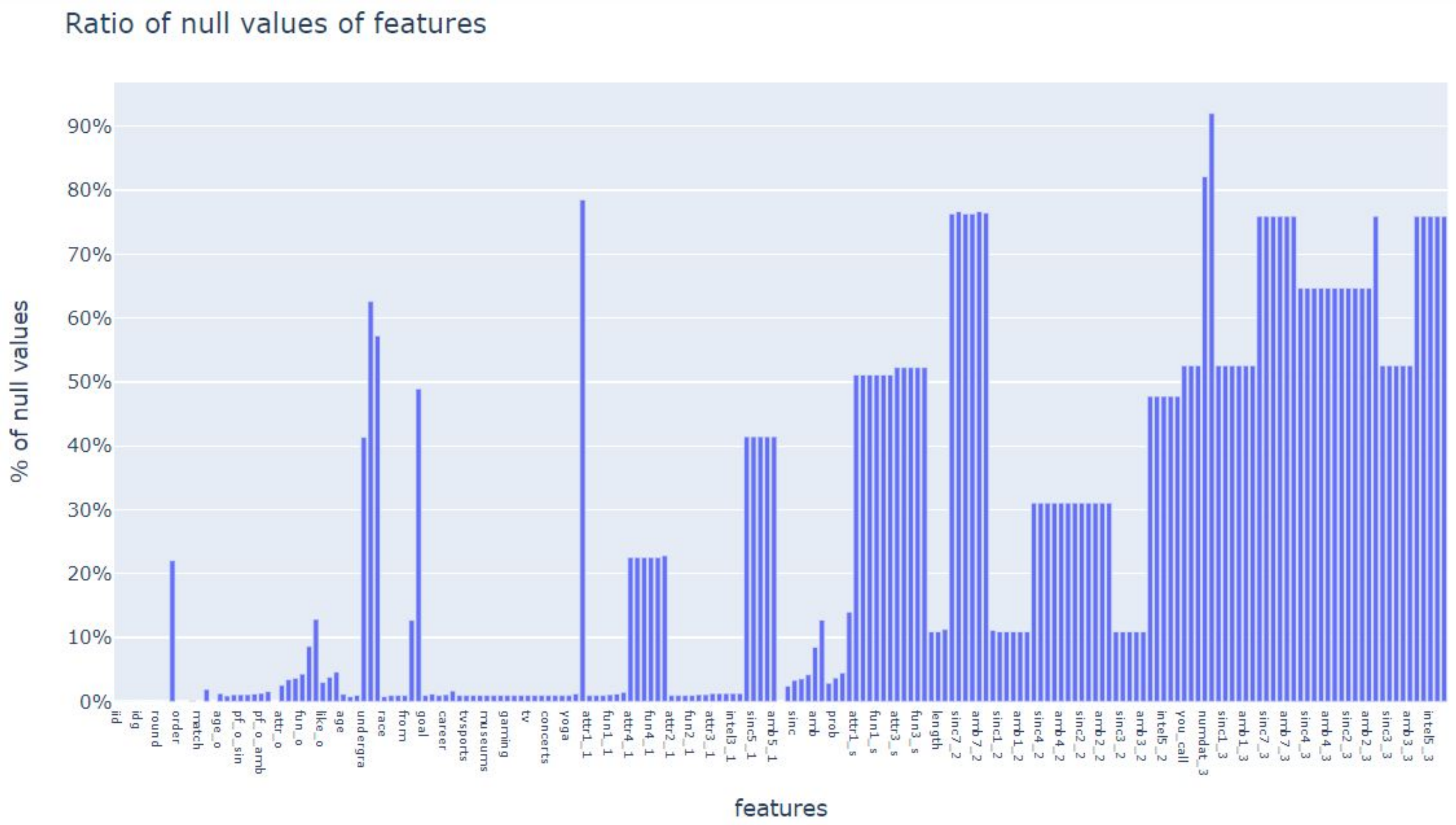
19.01.2021

Agenda

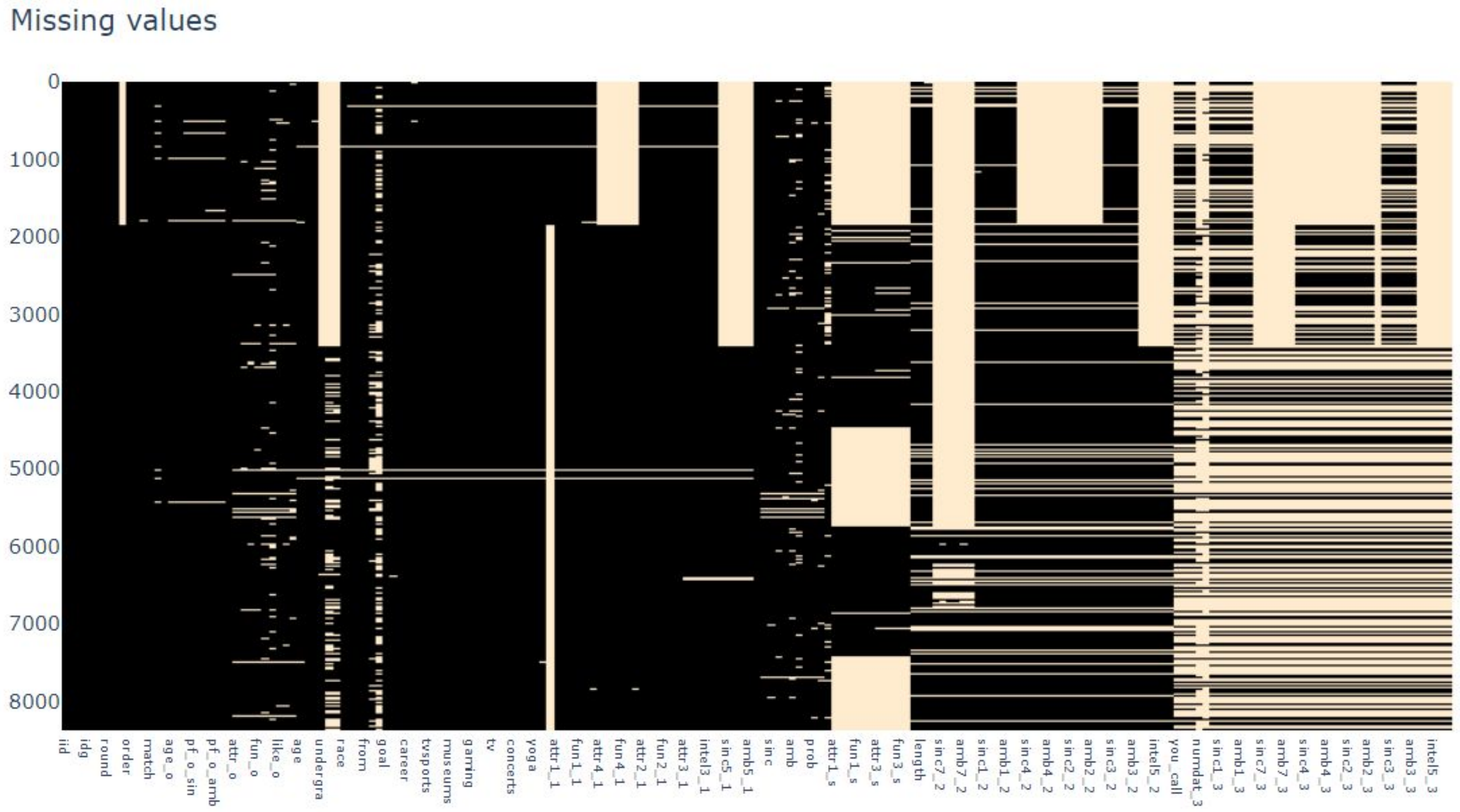
1. Data cleaning
2. Exploratory Data Analysis
3. Prediction

1. Data cleaning

1. Data cleansing and missing values



1. Data cleansing and missing values



1. Data cleansing and missing values

Initial feature cleaning & selection:

- We are only using information which the respondent could reasonably have after the end of the date

- We are deleting all the ids, which can consider as metadata: they don't reflect the information which makes up a person

- Likewise, we are removing features related to organizational matters of the speed dating event

1. Data cleansing and missing values

Initial feature cleaning & selection:

Free text fields

There are 2 reasons why we can drop free text fields:

- The information is duplicated in other categorical-coded features, which makes it redundant ;

- Participants have entered some nonsense inputs which is not suitable for predictions.

```
In [8]: dating_clean["career"].iloc[8373]
        executed in 4ms, finished 19:58:54 2021-01-18
Out[8]: "assistant master of the universe (otherwise it's too much work)"
```

1.2. Analysis of missing values

Time 1

Looking at the the heatmap of missing values of the questions in time 1, we can make 2 observations:

- **questions 4 & 5 were left unanswered for the first few waves**. This leaves out quite a few missing values, and should be dropped.

- **scorecard questions** (i.e *intel1_s*) **show even more missing values** and it would seem that they were left unanswered by many waves, hence are removed.

- other **features with over than 50% of missing values are also discarded** (*tuition, mn_sat, expnum, etc*).

1.2. Analysis of missing values

Time 2 & 3

We are dropping Time 2 features because:

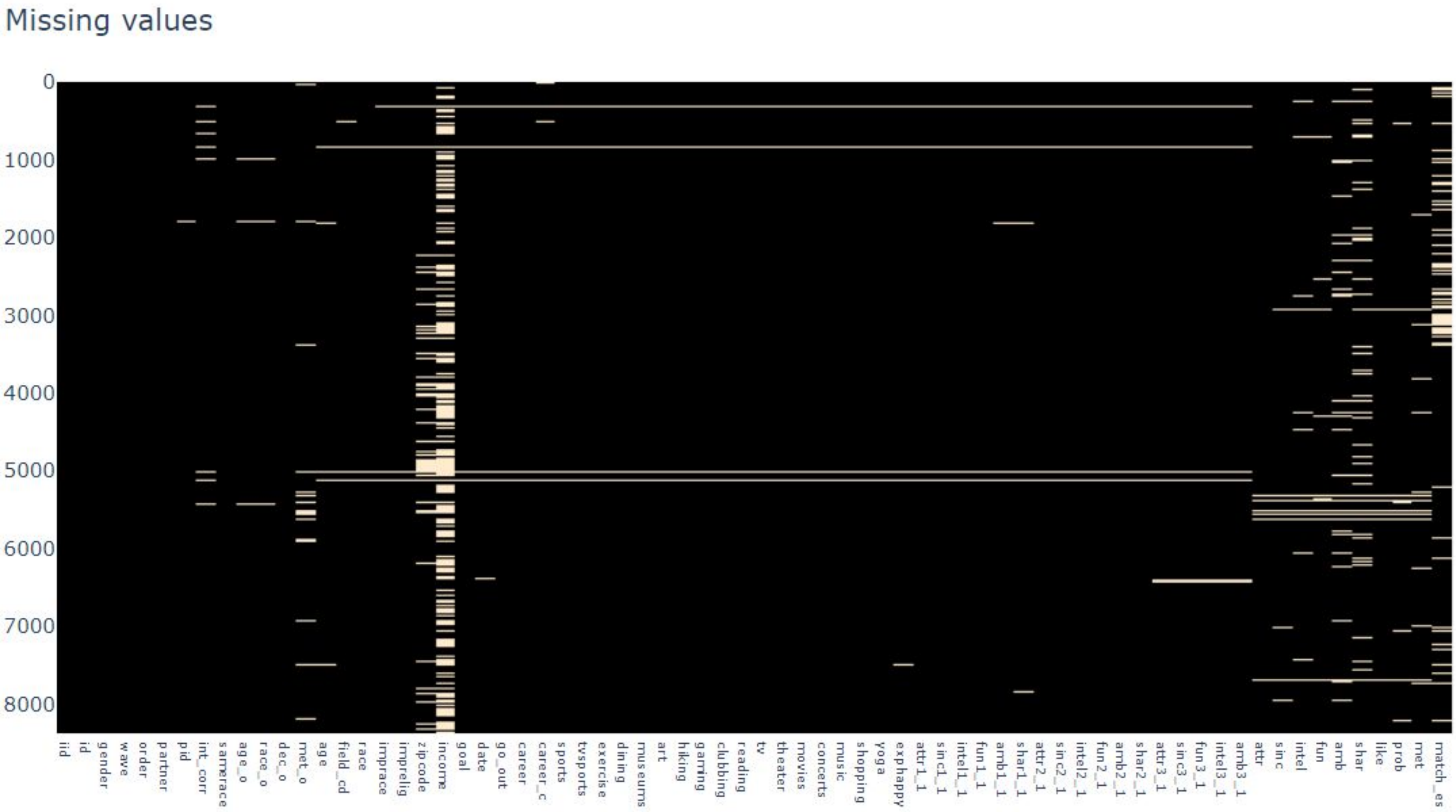
- the information was submitted *after* the first date, so it should not influence dec_o ;
- there are many missing values which are difficult to predict, and we consider it *safer to drop that many missing values than to impute them* ;
- the questions are the same as in time 1. This we can also *avoid redundancy.*

We are also dropping Time 3 features because:

- likewise, the information is submitted *after* the date ;
- even more missing values than Time 2 ;
- the questions are the same as in time 1 and time 2. Hence makes sense to remove them and *only keep time 1.*

1.2. Analysis of missing values

Time 2 & 3



1.2. Analysis of missing values

Age and age_o

- **age** and **age_o** are mutually related.

- Missing values in age can be found in age_o and vice-versa.

- If the age of a person cannot be found, **the mean of the wave**, where the person belong, is used to estimate the age of such person.

Why the mean value of the wave?

⇨ According to the description of the waves, *the wave 5 was performed with only undergraduates*

⇨ The mean value of the wave would be a better approximation than the mean age of the whole dataset.

1.2. Analysis of missing values

Scorecard questions

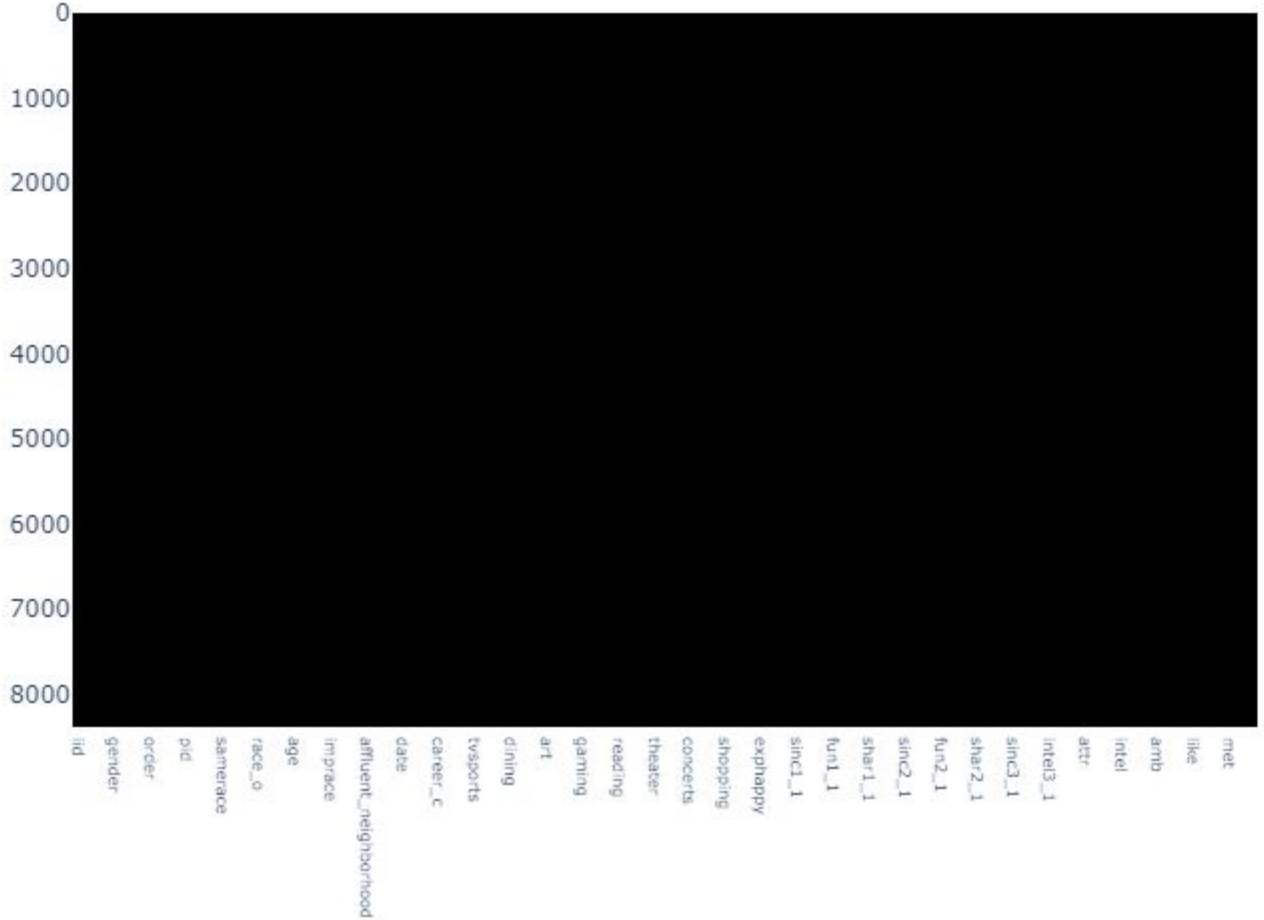
Scorecard questions: *"attr","sinc","intel","fun", etc.*

- For the participants who did not fill *some* of the values, *the median of the features of that person* is used to fill all the missing values.

- This makes sense since some people, for example, evaluate their partners with relatively high values, so a missing value is more likely to be also high.

- For participants who did not fill *any* feature of the scorecard, *the mean of the values of the same wave assigned by the candidate's dates* is used as an estimate.

No more missing values!



1.2. Analysis of missing values

Zip code:

- Has a lot of missing values
- How to use this as a meaningful input feature?

Income:

- Even more missing values
- Median household income based on zip code

Introduction of a new Feature:

- **“**Women exhibit a preference for men who grew up in affluent neighborhoods.”
  Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment Raymond Fisman et al.
- Feature affluent_neighborhood

$$\text{affluent\_neighborhood} \begin{cases} 1 & > \text{income}_{3\ \text{Quartile}} \\ 0 & =< \text{income}_{3\ \text{Quartile}} \end{cases}$$

1.2. Analysis and correction of questions at time 1

attr1_1
Attractive
sinc1_1
Sincere
intel1_1
Intelligent
fun1_1
Fun
amb1_1
Ambitious
shar1_1
Has shared interests/hobbie

1.2. Analysis and correction of questions at time 1

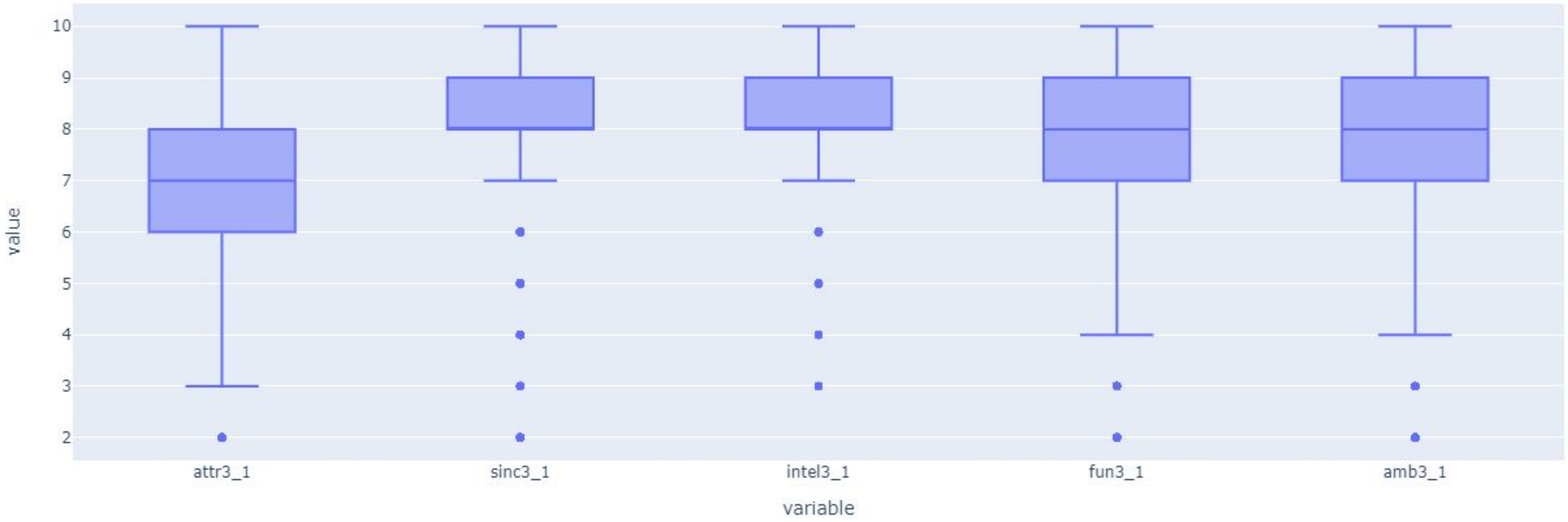
- Sometimes the features are scaled between 1 and 10 and sometimes the sum of all features of one question had to be 100

- **Problem:**
  - Comparing those features is not easy
  - Some features might not match the constraints

Wave #	Date	Preference Scale	Variations	#
1	October 16^{th} '02	100 pt alloc.		1
2	October 23^{rd} '02	100 pt alloc.		1
3	November 12^{th} '02	100 pt alloc.		1
4	November 12^{th} '02	100 pt alloc.		1
5	November $20^{th.}$ '02	100 pt alloc.	undergrads	1
6	March 26^{th} '03	1-10 scale		5
7	March 26^{th} '03	1-10 scale		1
8	April 2^{nd} '03	1-10 scale		1
9	April 2^{nd} '03	1-10 scale		2
10	September 24^{th} '03	100 pt alloc.		9
11	September 24^{th} '03	100 pt alloc.		2
12	October 7^{th} '03	100 pt alloc.	Budget: only allowed to yes	1

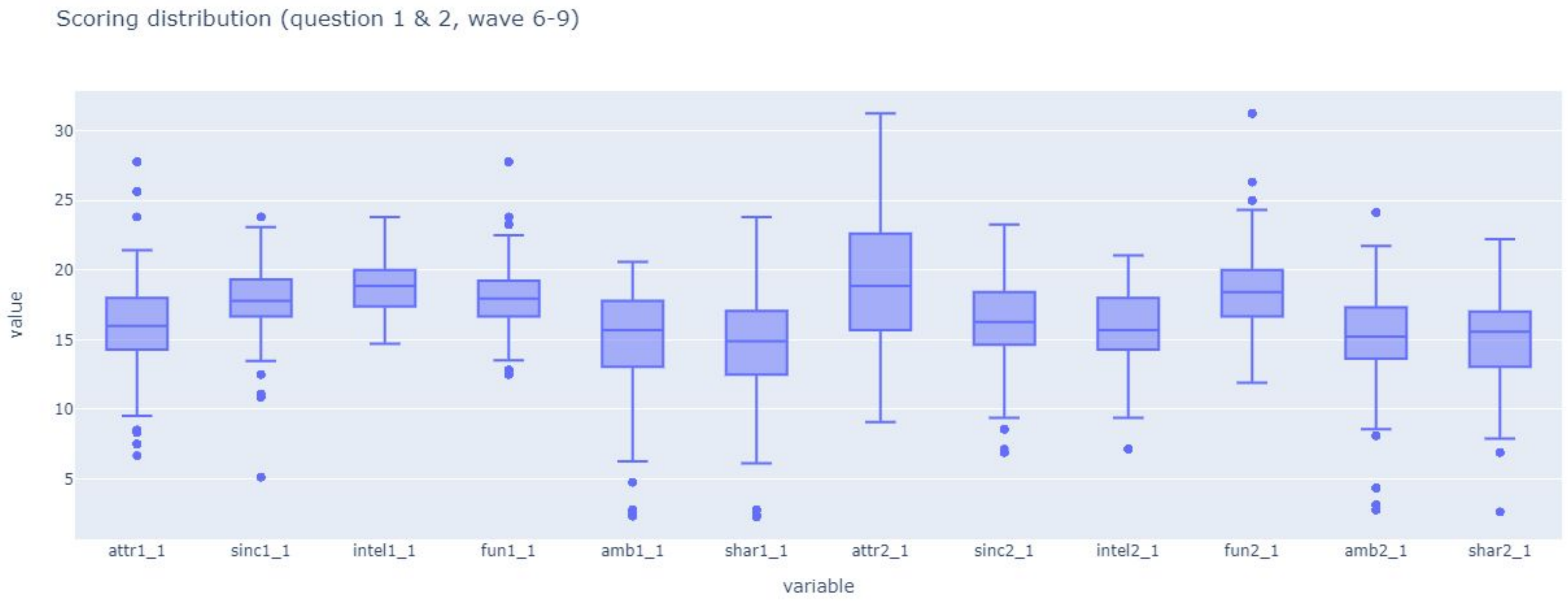
1.2. Analysis and correction of questions at time 1

- Distribution of answers for Question 3 (all waves)
- **Expectation:** All values between 1 and 10



1.2. Analysis and correction of questions at time 1

- Distribution of answers for Question 1 and 2 (wave 6 to 9)
- **Expectation:** All values should be between 1 and 10



1.2. Analysis and correction of questions at time 1

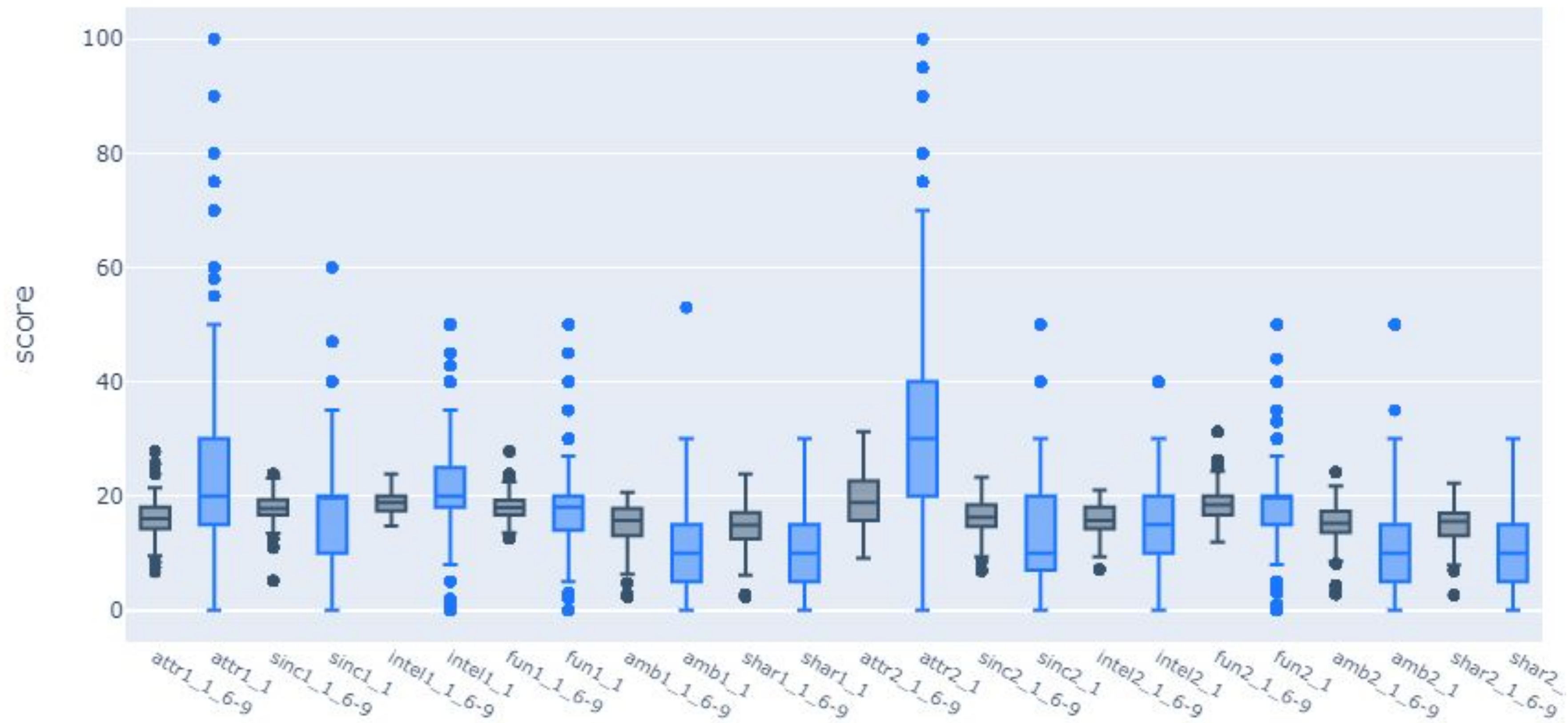
- **Assumption:** The features are already transformed to add up to 100. To be comparable.
- They add up to 100 but with a slight offset (e.g. 100.02)
- Decimal points provide evidence that these values are converted

```
Sum of ['attr1_1', 'sinc1_1', 'intel1_1', 'fun1_1', 'amb1_1', 'shar1_1']
1846    100.02
1847    100.02
1848    100.02
1849    100.02
1850    100.02
```

1.2. Analysis and correction of questions at time 1

Question: Are the transformed features of wave 6 to 9 comparable to the features of the other waves?

Scoring distribution per wave type



1.2. Analysis and correction of questions at time 1

Conclusion: Since the distributions are significantly different due to the transformation, the features of the different waves might not be comparable. This could lead to a bad performance of a classifier.

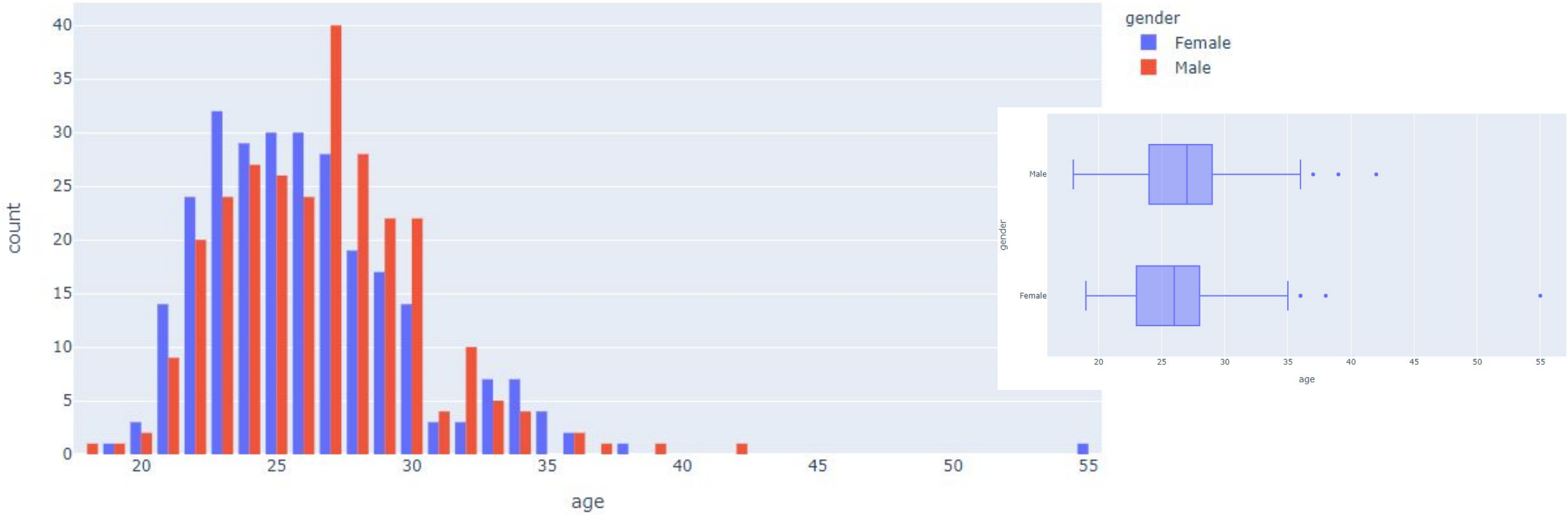
Impacts on prediction part: Test classifiers *with* and *without* these questions

2. Exploratory Data Analysis

2.1. What is the distribution of gender for different age groups?

Outliers: age > 35

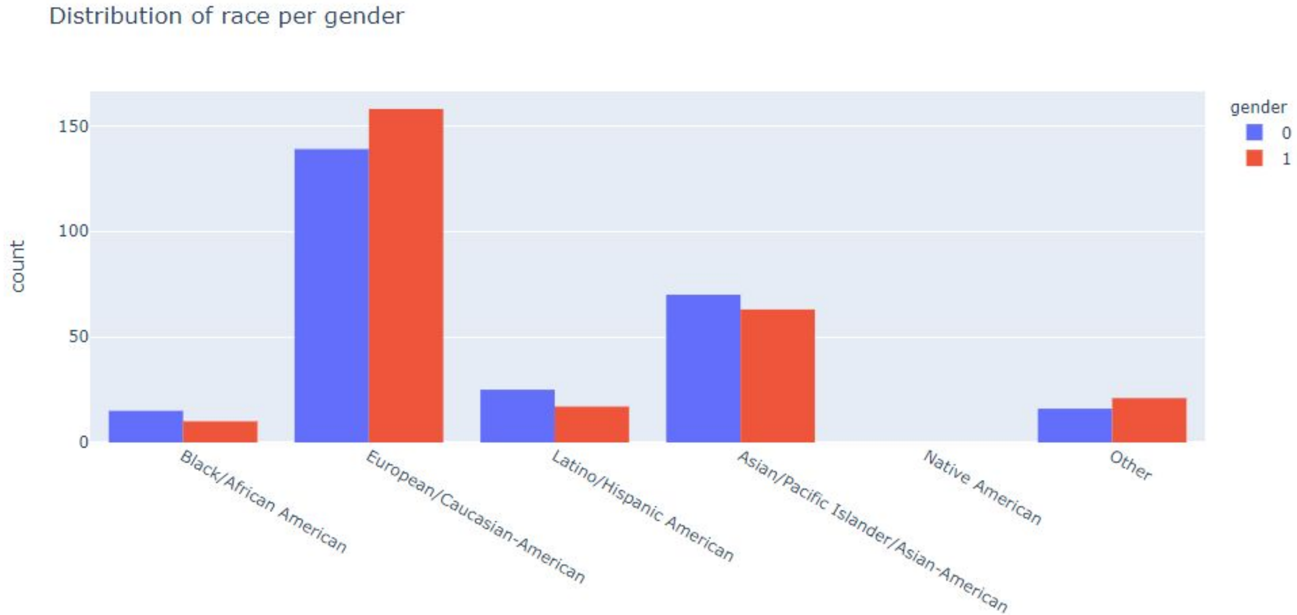
Men are half a year older than women



2.2. What is the distribution of race for the two genders?

The race majority is European/Caucasian-American, across the spectrum ;

The Black/African American race accounts for a minority.



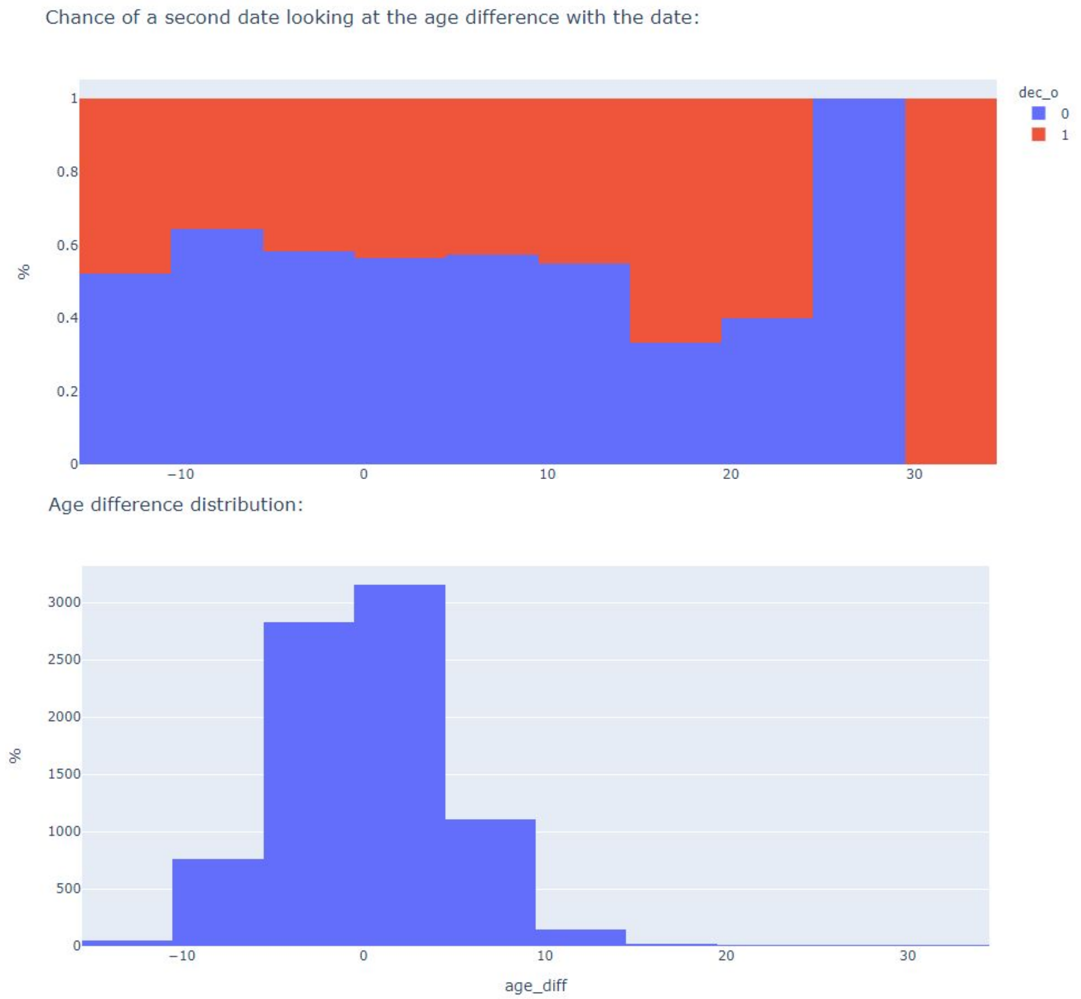
2.3. Are there differences in **gender**, age and race in the likelihood to get a second date?

Chi-squared test for independence:

H0 -> Getting a second date is independent from gender
H1 -> Getting a second date is dependent from gender
a = 5%

H0 rejected
P value: 1.63e-22

2.3. Are there differences in gender, **age** and race in the likelihood to get a second date?

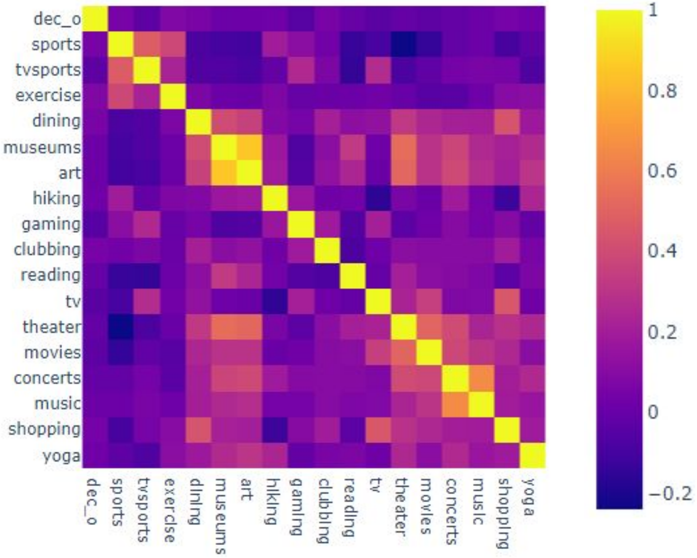


2.3. Are there differences in gender, age and **race** in the likelihood to get a second date?

2.4. What is the correlation of ones interests with the chance for a second date?

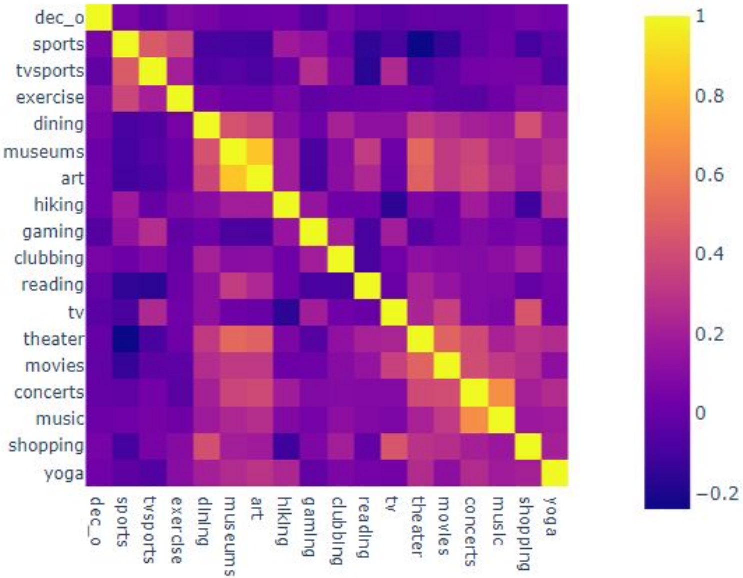
Pearson

Exercise = 0,083



Spearman

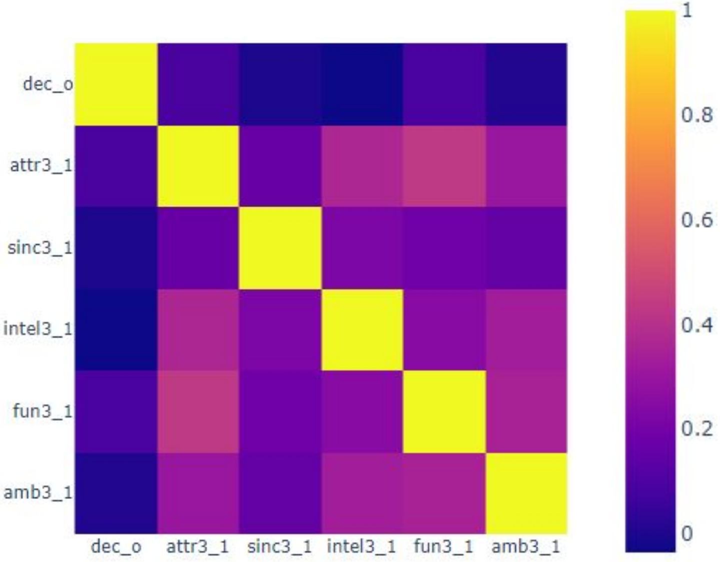
Exercise = 0.086



2.5. What is the correlation between ones own opinion on ones attributes (attractive, sincere, intelligent, fun, ambitious) with the chance of getting a second date?

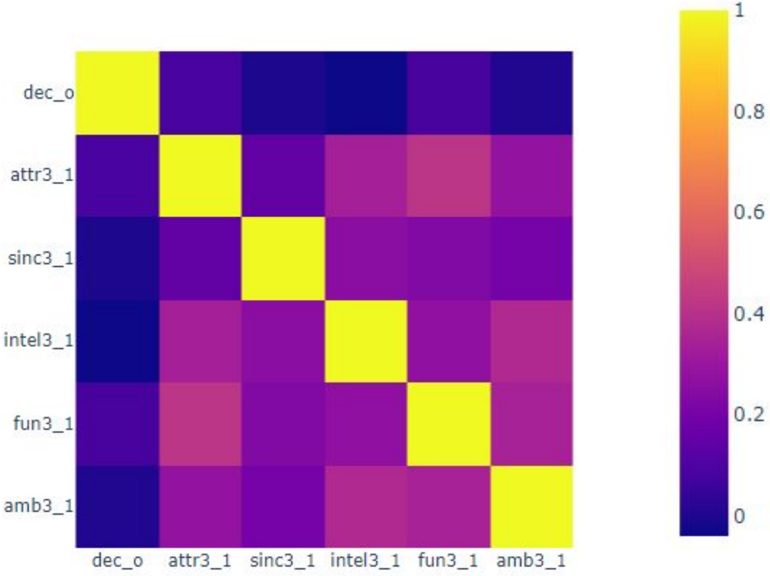
Pearson

attr3_1 = 0,093



Spearman

attr3_1 = 0,085



3. Prediction

Logistic Regression

LogisticRegression (in nested CV, 10 fold)	**Accuracy** (%)
With PCA	62.8
With regularization	63.7
Without regularization	63.7
Removing question 1 & 2	63.4

Conclusion:

- The process of applying **PCA does not seem to improve the prediction**.
- **Same goes for regularization penalties**
- Finally, both models perform equally well without the features of questions 1 & 2. According to **Occam's razor principle**, we assume it **safe to drop these features for future estimators**.

Polynomial Features

- Using all features to create polynomial features would result in over 3000 features.
- We therefore apply a method of feature selection (SelectKBest)
- We convert these features to polynomial features of order 2 and select only the 15 most important features.
- This results into 135 new features.

Polynomial Features(in nested CV, 10 fold)	**Accuracy (%)**
Order 2, 15 most important features	63.4

Conclusion:

- Using polynomial features doesn't help to improve the performance of the classifier.
- However it shows that **only using 15 of the original features** gives **the same accuracy** as using all of them.

Permutation/Feature Importance

- Defined to decrease the model score when a single feature value is randomly shuffled
- It reflects how important this feature is for a particular model, breaking it’s relationship with the target.

Conclusion:

- The attributes: *attr2_1, race_o_2, prob, attr, gender* are always defined as important features for the model.
- Nevertheless, **the accuracy of the model** after using only these important features **is not always improved.**
- Getting **the most important features does not improve the accuracy** of the model for this dataset.

Support Vector Machines

SVC (in nested CV, 10 fold)	**Accuracy** (%)
Linear SVC (with regularization)	63.4
With kernel approximation	58.2
Non-linear SVC	68.2
SVC with SGD	63.9

Conclusion:

- Non linear SVC seems to perform best.
- Linear classifiers, even with a kernel approximation of the feature map does not measure up

Decision Tree Classifier

	Accuracy (%)
Decision Tree (in nested CV, 10 fold)	62.7

Conclusion:

- We encounter an accuracy score which is n**ot better than the other classifiers.**
- **The hyperparameters vary a lot.**
- We will not further look into feature selection and polynomial features since decision trees already select features and combine them.

4. Conclusion

Conclusion

- With just classifying every instance as dec_o = 0 we would get an accuracy of 57,7%.
- Looking at the classifiers we can **surpass this margin only by around 10%**.
- Although **we applied many techniques to transform and select the features**, there was **no method which stood out as a major solution**.
- We therefore conclude that **based on information only the participant would know at the time of the date**, it is difficult to predict the outcome to a high degree of precision.
- Consequently, **self-assessment is not enough to confidently predict the chance of a second date.**