Predictive Analytics WS2020 Project

Task: Predict if in a speed dating event someone will get a second date

The data:

An introduction to the data and to its analysis can be found here:

https://www.kaggle.com/annavictoria/speed-dating-experiment

https://www.kaggle.com/aeshen/the-secret-to-getting-the-second-date

General goal: Predict with highest accuracy for every participant if he/she will get a second data

Approach:

- Build teams of 3-4 students and divide the work amongst you
- Use python and iPython notebooks to implement and document your work

Subgoals:

Exploratory data analysis:

- What is the distribution of gender for different age groups?
- What is the distribution of race for the two genders?
- Are there differences in gender, age and race in the likelihood to get a second date?
- What is the correlation of ones interests with the chance for a second date?
- What is the correlation between ones own opinion on ones attributes (attractive, sincere, intelligent, fun, ambitious) with the chance of getting a second date?
- Perform an analysis of which features have missing or invalid values
- Visualize the results
- Clean the data by either removing invalid entries or filling them with suitable values

Prediction

Compare

- Logistic Regression with non-linear features of your choice
- Support Vector machines with non-linear features of your choice
- Other classifiers of your choice (e.g. random forest, …)

Approach

- Investigate effects of regularization
- Perform nested cross validation for hyperparameter tuning and performance prediction. Use 10 fold cross validation for outer loop
- Evaluate the importance of the different features using permutation importance (https://scikit-learn.org/stable/modules/permutation_importance.html) or also additional approaches

Creativity

- Try different ideas to select, transform (e.g. by combining them to a new feature) the features to increase performance
- Try to gain insights why and what is happening

- Try unique visualizations to gain these insights

Presentation

Prepare a 10 min presentation with subsequent questions to show on January 19th or 21st 2021 during the lecture where each team member presents part of the work and answers to the questions. Code has to be handed in on the evening of the 18th, latest.