Ron Sailer & Adam Erdreich – feedback

Final grade – 115

Documentation:

Good documentation which helped a lot to appreciate your work. I liked the story you told using the notebook and visualizations.

Imputation:
Well done! You've done a very nice job in your analysis of the dataset and in the consequent actions.

Your observations regarding the amount of complete rows is important. Using feature redundancy for imputation is also a very good approach.

Outlier Detection:

Nice work. Indeed it is good to measure the significance of your work. As you've demonstrated by measuring the classification error.

Normalization:

Why did you decide to normalize all your features to a normal distribution?
This makes sense if a feature is somewhat Gaussian. However, for a uniform distribution, for example, this might be a bad choice. Such normalization changes its distribution, which may affect the final predictions.

Splitting the dataset:

Clean and short code. Very nice.

Bonus:

Very true. The Relief algorithm is indeed computationally demanding despite its advantage of taking the joint distribution into account.
Another disadvantage is that the notion of "miss" and "hit" are hard. That is, one outlier may disproportionally affect the closest match.

Regarding the implementation of all 3 algorithms: Your code is nicely written and well documented. I hope to encounter such code more often.

Features selection:

ID	Selected Features	Actual Description	Valuable (1, 0)
1	Avg_Residancy_Altitude		1
2	Avg_government_satisfaction	Random numbers in the range[0,10]	0
3	Avg_monthly_expense_on_pets_or_plants	Linear combination of Yearly_ExpensesK Yearly_IncomeK	
4	Avg_monthly_household_cost	Linear combination of Yearly_ExpensesK, Yearly_IncomeK and Overall_happiness_score	0
5	Avg_size_per_room	Polynomial function of Yearly_IncomeK	0
6	Financial_agenda_matters		1
7	Last_school_grades	Linear function of "Most_Important_Issue	
8	Married	Linear function of "Most_Important_Issue"	0
9	Most_Important_Issue		1
10	Occupation	Uniform discrete distribution 1-5	0
11	Overall_happiness_score		1
12	Phone_minutes_10_years	Polynomial function of Yearly_ExpensesK	0
13	Political_interest_Total_Score	Linear combination of Yearly_ExpensesK, Yearly_IncomeK and Overall_happiness_score	0
14	Weighted_education_rank	Polynomial combination of Yearly_ExpensesK Yearly_IncomeK Overall_happiness_score	0
15	Will_vote_only_large_party		1
16	Yearly_ExpensesK		1
17	Yearly_IncomeK		1

Looks like you were able to maintain all of the 7 relevant features while keeping an additional 10 features.

Nice idea to use a cascade of feature selection algorithms.
However, you've used only wrapper methods without any filter methods.
In the imputation section you found interesting insights regarding the data's distribution, yet for some reason (lack of time?) didn't use it.
By running a few extra plots, or calculating the correlation between features and labels you could have easily found redundancy in the features you've selected and reached a tighter set of features.
Your work is a fine demonstration why using one family of algorithms, even when done thoroughly, is not enough.

All in all, very nice work. I recommend you perform better normalization for the subsequent exercises.