Group Project: Predicting the outcome of the Brexit Referendum

For the group project, you should download the 'Brexit' vote data, that has local authrity-level information on the Brexit vote of 2016, along with numerous area characteristics (the region, demographics, migration, recent political outcomes, educational levels, employment levels and trends and others.). Your task is to explore these data and build prediction models where the variable of interest is the Brexit vote outcome - leave or remain? You should try a range of supervised learning methods e.g. OLS, OLS with regularisation, Logistic (note that it too has regularisation parameters), and/or random forests. You are free to explore other methods that we have not covered in lecture or seminar but are part of Scikit learn (there are plenty of free online resources that explain the methods lucidly.). However, try to focus on the methods we have covered, and not spread yourself too thinly. You should conclude the analytical portion by evaluating and comparing the performance of the various models in terms of their ability to predict outcomes.

Note that the data file is in STATA, and each variable is clearly labelled. For a more detailed description of the construction of this data see Becker, Fetzer and Novy (2017). You will have to import the file (after conversion to csv or txt) for analysis in a Jupyter Notebook. You should copy over the data labels from Stata. Note also that the data are high-dimensional; there are many features (variables) and not very many observations. Pre-processing the data is therefore going to be a critical part of the project, and you may wish to explore methods that emphasise feature selection.

In the discussion (presentation and the Notebook), you might wish to consider questions such as:
1. Which model(s) you would pick and why?
2. Which features matter most for prediction accuracy?
3. Would you rely on the selected model(s) for predicting outcomes in a second referendum if that were to happen?
4. What are some key lessons about applying ML methods to a small high-dimensional dataset?

Guidelines for presentation:
1. Aim to talk for about 10 minutes (and leave 5 minutes for Q&A); so please keep the number of slides to 6-7.
2. Avoid spending time on descriptive statistics (all groups have the same data so it would be repetitive)
3. Compare the performance of the various learning models that you fit to the data.
4. Discuss your responses to the questions above.
4. Keep the Jupyter Notebook with your project code handy in case you have to explain something during Q&A.

Guidelines for the Notebook submission:
1. Should contain the final code with clear annotation, figures, tables, and text organised into readable sections (motivation, data and methods, performance evaluation, discussion and conclusion).
2. The code should be working; we will check your Notebooks.
3. Be precise.

Finally, note: This assignment gives you an opportunity to work with real-world data on a topical issue. The outline above gives some guidance but is flexible enough that you can structure your project as you see fit. You will have to make several choices along the way e.g. should I use method A or B, should I retain or drop some features, which performance metrics should I use etc. There are no unambiguously right or wrong answers to many of those questions - only trade-offs. Those are your choices - make them as a group after due thought and deliberation. So try not to approach us with these questions. This is part of the learning process and if and when you carry out a data science task in the future, you will confront similar questions.

Note the presentations will take place on **Wednesday 8th May, 9 – 11am**. Please upload your slides to Canvas by 4pm on Tuesday 7th May.

Reference:

Sascha O Becker, Thiemo Fetzer and Dennis Novy (2017) "*Who Voted for Brexit? A Comprehensive District-level Analysis*." Economic Policy, Volume 32, Issue 92, Pages 601–650. https://doi.org/10.1093/epolic/eix012