The Interdisciplinary Evaluation of Machine Learning Algorithms versus Regression Type Analyses in Political Science Research

Cosmin Andrei ARTIMOF
andrei.artimof@gmail.com
M.Sc. in Big Data

UNIVERSITY of STIRLING

INTRODUCTION

- The goal of social science is to **understand** the behaviour of people, groups, companies, societies, cultures, and economies.
- Quantitative social research is grounded in inferential statistics and employs statistical models to nearly exclusively test causal hypotheses.
- In political science, statistical models are used almost exclusively for causal **explanation** and models that have high explanatory power are often assumed to inherently possess predictive power.
- The type of statistical models used for testing causal hypotheses in the social sciences are almost always **association-based** models applied to observational data. Regression models are the most common example.
- Machine learning is a subfield of computer science and artificial intelligence that concerns itself with discovering predictive models and how to learn meaningful patterns from data. As such, data-driven algorithms **induce** models from the data.
- Machine learning algorithms typically do not assume (normal) distributions and therefore are considered **non-parametric**. While models are far more flexible, because less is known about the data, non-parametric algorithms (e.g. decision trees) are typically iterative and don't guarantee that the optimal solution has been found.

PROBLEM STATEMENT

- Although machine learning methods have spawned some modicum of interest in the political methodology literature, they are not very prominent in applied political science research. There's **opportunity** for empirical inter-disciplinary research.
- Superficially, machine learning methods are seen as appropriate for **atheoretical** tasks ("black box" methods) and not very useful for deriving "substantive" insight. Many problems in social sciences entail a combination of prediction and explanation.
- Clearing the current ambiguity between **explanation and prediction** is critical for proper statistical modelling. Making social science researchers aware of the alternative ML algorithms will open them to a wealth of research opportunities by enabling them to explore the digitized social data flowing around us.

RESEARCH OBJECTIVES

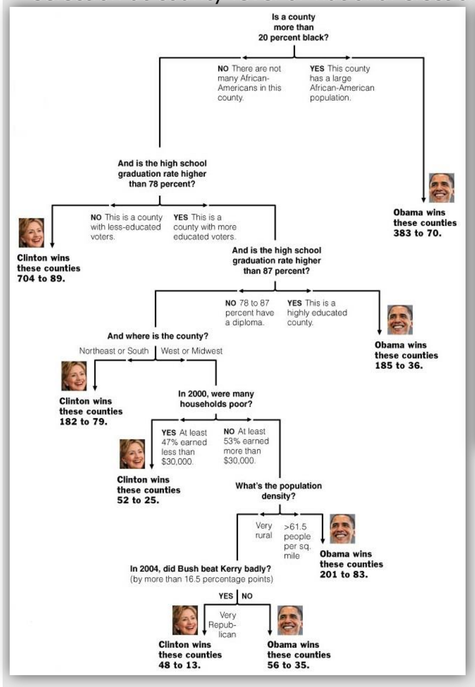
- Exploring an **empirical** and methodological question. Can ML algorithms and predictive analytics practices (feature selection, cross-validation) improve the quality of political science research?

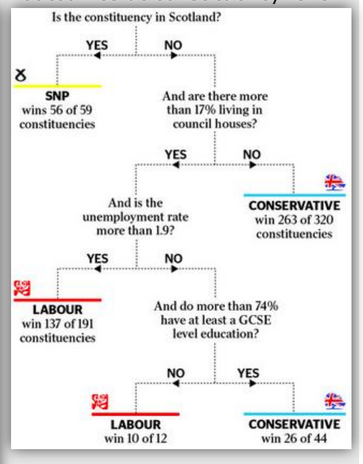
RESEARCH METHODOLOGY

- The study is a **comparative analysis** of the predictive performance of artificial neural networks (ANN) and decision trees ML algorithms as opposed to linear multivariate and logistic regression analyses, predominant in social sciences.
- The proposed analytical framework is the Cross-Industry Standard Process for Data Mining (**CRISP-DM**).
- Will focus on the **data preparation** stage (dealing with outliers and missing values, discretizing or normalizing variables depending on which algorithm is used, and tackling the dimensionality of data by extracting features with different feature selection methods) and the **data modelling** stage (establishing baseline performance, balancing the training dataset, cross-validation and overfitting analysis).
- Will do a **replication study** of a political science journal article and one original piece of research using the **CSES** (Comparative Study of Electoral Systems) datasets. The purpose is to present to social researchers the steps and software used in the field of predictive analytics.
- The 4 CSES modules include **258.461** observations from 158 post-electoral surveys administered in over 40 countries in the period spanning between 1996 and 2016.

DECISION TREE CLASSIFICATION

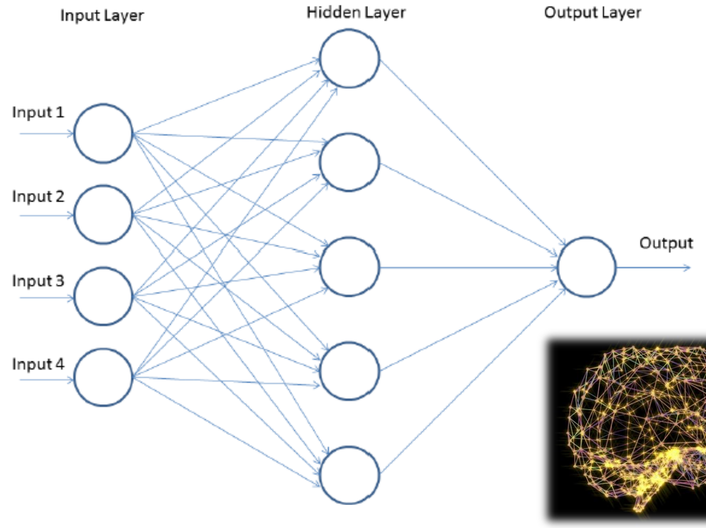
- Decision trees apply a "**divide-and-conquer**" approach to the problem of learning from data.
- The feature that best divides the training data is used to **partition** the records into smaller subsets.
- Decision trees are **highly interpretable**, can deal with any kind of input data, and can also cope with missing values.
- The learning process of decision trees is usually **fast** compared to algorithms like neural networks (ANN), support vector machines (SVM).
- In **political science**, decision trees can be used to **predict** party candidate selection at county level or national election outcomes at constituency level.





Other possible applications include: predict **partisan leanings**, intention to take part in elections, support for policies, etc..

NEURAL NETWORKS FORECASTING



- Artificial neural networks (ANN) were inspired by attempts to simulate **biological** neural systems, or the behaviour of the human brain.
- The human brain can **learn** by changing the strength of the **synaptic** connection between neurons upon repeated stimulation by the same impulse.
- ANN can handle **redundant** features and are able to deal with any kind of input data, though a couple of assumptions about the error terms, present in statistical modelling too, have to be met.
- The training of ANN is usually very time consuming and ANN are also sensitive to the presence of **noise** in the training data.
- Contrary to that, their classification speed is rather **fast**.
- Social and political relationships are generally characterized by **nonlinearity** and complexity and are usually of unknown functional forms.
- Neural network models are capable of approximating arbitrary functional forms under general conditions and can handle rich patterns of nonlinearity in the utility functions. They are therefore potentially **better** suited to typical social science data as compared to linear or logistic regression type analyses.
- ANN should be able to accurately predict voter turnout, or judges' decisions in constitutional courts on legislation.

Future work

- There are currently no viable, simple-to-use, machine learning algorithms that can replace multilevel modelling techniques in social sciences (analysis of clustered data). An alternative to multilevel-modelling would be to aggregate data at the higher level and employ a Random Forrest algorithm.