**Scoping Document for Election Forecast Model**

* The aim is to build an election forecasting model for the UK. The project is inspired by the 538 election forecasting model.
* The overall goal is to build a model that will be capable of predicting UK election. It will be trained on historic UK polling data. This is obtained from <https://www.markpack.org.uk/opinion-polls/> This data is updated roughly one a quarter. Current data is for q4 2018. It starts with the first ever by-election poll in the UK in 1938 and is comprehensive from 2008.
* The first task is to clean the data for analysis. The data set is messy, it does not conform to the tidy data principles as set out in <https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf> by Hadley Wickhams.
* See for different time periods how accurate polling is at predicting election outcomes. For an individual poll this can be calculated by determine the difference between the poll result and the election result. This could then be aggregated to find out the predictive power of polling 1 week, 1 month or 6 months out from an election. The polls could be aggregated simply by using the mean. A better approach would be to weight the polls by their historic accuracy. The polls could be weighted by any variable which is correlated with the outcome variable, in sampling theory this is known as the alpha weigh. Potential options: sample size – larger more accurate probably, sample method- maybe phone or person interview more accurate, and polling company, as different polling companies have better track record than others. This last variable is a good starting point, and captures the variability associated with the different methodologies that polling companies use. Different weighting methods can be explored, by calculating the correlation for each variable or a linear combination. A potential issue is that this correlation will not be invariant to time.
* Once the polls have been aggregated, a linear regression model can be then be used to model the relationship. Other probabilistic modelling techniques can then be explored. A start would be a maximum likelihood linear Gaussian model. This is a form of probabilistic Bayesian linear regression. No distinction is made between a scalar or multivariate case (for practical purposes this has little impact). This model is suitable as we are less interested in the point estimate (the vote share of a party) but the probability density for a prediction conditional on the historic data. These models are for continuous data which we are using. A nonparametric probabilistic regression model using Gaussian processes can also be explored. This model doesn’t rely on parametric modelling assumptions it is instead flexible. This means the Gaussian process adapts to the model complexity as more data arrives, and account for non- nonlinearities. This is an attractive feature, as the data is likely to display non- nonlinearities (such as the Corbyn surge in 2017), the models parameters are not fixed on the historic data. This however may raise an issue with over fitting. In general the training data is not summarised by some parameters but is part of the model like other k-NN machine learning techniques.
* This would be the basic polls model. Historic data will have been used to train the weighs (or beta’s in regression nomenclature) which can then be used to forecast a future election based off the current polling data.
* A second step is to build in fundamentals. These are variables which improve the fit of the model, they increase the explanatory power. These might be political or economic. Political examples could be incumbency or leader popularity ratings. Economic variables are more interesting and high quality data easier to source. Economic growth, per capita income, unemployment, inflation are all likely to increase predictive power, e.g. rising per capita income is likely to favour the incumbent party. These will be incorporated into the model using a multivariate regression model. This is a simple extension of LR- linear regression to more than one explanatory (or independent) variables.
* Other points, the current prediction will incorporate data over various time frames, will need to weigh the polls accordingly, giving more weight to more recent polls. Too high a weight, the prediction will jump around to incorporate the latest polls, too low a weight and will not be sensitive to changes. The forecast will be probabilistic; will need to calculate confidence intervals using historic variability. Another interesting approach would be to use a Bayesian framework, perhaps using a machine learning technique such as naïve bayes.
* Aside, create a rolling polling average. This will aggregate the polls and graph it to show the current popularity of the parties.
* More advanced machine learning techniques can then be tested. As there is a large amount of data PCA, principle component analysis, could be for dimensionality reduction. This could then be fed into a predictive machine learning method. Neural networks are suitable to this task and likely to improve on regression analysis due to them incorporating non-linearity’s into the analysis. Neural networks are also effective at forecasting.
* Note: potential to incorporate twitter data? Use of the software library Tweepy (python) or rtweet (for r) to access twitter’s public API.
* With the model built interactive graphs and charts could be developed. A website incorporating this could be deployed or an app.

Literature

<https://fivethirtyeight.com/features/a-users-guide-to-fivethirtyeights-2016-general-election-forecast/>

<https://fivethirtyeight.com/features/2018-house-forecast-methodology/>

<https://www.markpack.org.uk/opinion-polls/>

<https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf>

<https://www.sciencedirect.com/science/article/pii/S0261379415001109>

<http://eprints.whiterose.ac.uk/87122/2/EFTINVITED.pdf>