Evaluating the Application of Machine Learning Algorithms to Predict Constituency Voter Turnout of the 2019 UK General Election

Thesis By:
2003816

Course:
MSc Social Data Science

Institution:
University of Essex

Supervisor:
Professor Philip Leifeld

Date:
02/09/2021

Abstract

Machine learning algorithms provide social and political scientists with new tools to create prediction models regarding electoral politics. This paper ambitiously investigates whether a regression-based machine learning model can be utilised to accurately predict on a constituency level the electoral turnout of the 2019 UK general election. Despite parameter tuning and trialling numerous machine learning algorithms that encompassed many different demographic, economic and electoral features of data from recent general elections in the UK, all models failed to predict constituency turnout of the 2019 elections to a satisfying degree of accuracy. However, even though the prediction results were unsatisfactory, this paper outlines the future potential of Artificial Neural Networks as well as highlighting numerous data features used that could be worth implementing in future electoral turnout prediction models.

Contents

1. **Introduction**

Building a model to accurately predict voter turnout on a subnational level is a notoriously difficult undertaking. The reasons for this are vast and there is no clear consensus on what specific factors drive voter turnout (Geys, 2006, pg 638). As the challenge of predicting voter turnout remains relatively unsolved, it provides ample opportunity for the domain to be tackled through relatively new technology and techniques that can better utilize data around the subject. Machine learning has remained a relatively new and underused tool for political and social scientists. This is likely due to the highly technical and specialised nature of producing machine learning models, normally tasks completed by data scientists or computer scientists (Chen et al. 2018. pg 92-93). The goal of this research seeks to provide an investigation on whether supervised machine learning algorithms can be used to accurately predict the electoral turnout of the 2019 general election in the UK on a constituency level. Specifically, this paper will look at the addition of using Support Vector Machine algorithms, Tree-based algorithms and Artificial Neural Networks alongside the more traditional regression algorithms such as multiple linear regression for this prediction task. This research could ultimately lay the blueprint for further investigations of using machine learning algorithms to build voter turnout models. The 2019 election was chosen as it is the most recent to take place in the UK, therefore ensuring the model constructed is relevant to recent and modern elections. Firstly, in the literature review, there will be an overview of the current state of machine learning in political science and predictive model building. The second part of the literature review will contain an outline and discussion of various factors that have been found to affect voter turnout. The factors discussed in the literature review will be the basis for the data that will be used and pre-processed in the methodology. The methodology will outline on a step by step basis how the models were constructed from the data, including the selection of the machine learning algorithms, as well as the hyperparameter tuning that takes place. The results section reveals the highest-scoring algorithms according to the selected evaluation metrics used. All models will be directly compared to one another from on the training dataset, as well as the 2019 holdout dataset. The discussion section provides an analysis of the results as well as going over the limitations of the methodological process in constructing the models from the selected data. Finally, the conclusion will summarise the findings of the paper to determine the utility of using machine learning algorithms to predict voter turnout of the 2019 UK general election, as well as the further implications of the limitations of using machine learning algorithms to solve prediction tasks in social science.

2. Literature Review

i. Machine Learning for Building Predictive Models in Social Science

Machine learning as a tool for social science is seeing a more common and expanded use to gain a better understanding of social phenomena. Machine learning use in social science has a clear distinction from standard model building familiar in the social sciences. One key distinction is that certain machine learning models abandon the linear narrative of the relationship between variables, as machine learning is deemed to have more flexibility within its functions in comparison to more traditional methodologies of model building in social science, such as typical OLS regression (Grimmer et al. 2021. pg 396-399). This robustness is useful for constructing a predictive model in regards to voter turnout, as voter turnout has been a difficult phenomenon to find definitive and generalisable relationships with variables. Therefore, by deploying the use of machine learning algorithms we may learn of relationships through variables that are nonlinear and how they interact with the other feature sets. An advantage of deploying a machine learning model over traditional linear regression models is the machine learning models tend to have superior predictive power in comparison to a standard linear model (Hindman. 2015. pg 48-49). Prediction power is especially relevant to this paper as the goal of how successful a model is will be evaluated on a holdout set of the 2019 general election turnout dataset. The ability to utilize holdout sets is a beneficial aspect machine learning algorithms have over standard linear models used in social science for finding relationships between variables. Holdout sets further allow models constructed to be more robust and replicable for other researchers as they allow easier diagnostics of the sensitivity of outliers and specification fragility, further leading to the construction of more valid models as holdout sets can combat overfitting that occurs in models that are not tested on holdout samples (Hindman. 2015. pg 60-61).

Despite the relatively new adoption of machine learning in the academic domain of social science, there has been a selection of electoral prediction models built. A model predicting the electoral results of a House of Representatives election using mainly demographic data obtained from the US Census Bureau predicted to a high degree of accuracy the political party that would win the district. The conclusion pointed out that machine learning algorithms based on even weak predictor variables can still produce a model that has a high prediction accuracy (Richardson. 2020). Machine learning algorithms can also be used to predict electoral results using non-direct data streams such as Twitter data. Many models based on Twitter data have been used to predict electoral results to a high degree of accuracy (Tsai et al. 2019.), (Bansal & Srivastava, 2018, pg 351-352). This suggests that machine learning models have a wide degree of versatility, in which data from many different sources can be obtained and preprocessed to create an accurate election prediction model. Despite the literature examples not being directly relevant to predicting electoral turnout as these models were set out to predict electoral results,

this set of the literature suggests machine learning may likely be a useful tool to build prediction models involving many aspects of electoral politics.

Another underutilized tool regarding the construction of predictive models in political science is the use of Artificial Neural Networks. Artificial Neural Networks can be used for supervised machine learning tasks, such as regression prediction tasks. Artificial Neural Networks are unique over other machine learning algorithms as the architecture of Artificial Neural Networks uses multiple layers or 'hidden' layers of neurons to mathematically compute outputs based on inputs from prior layers in the network. This approach is supposed to mimic the way the human brain breaks down and interprets information in a mathematical sense (Di Franco, & Santurro, 2020, pg 1010). Similarly to other machine learning algorithms such as Random Forest Trees, Artificial Neural Networks are also effective at finding non-linear relationships in the datasets which make them effective tools for prediction and forecasting models (Di Franco, & Santurro, 2020, pg 1022-1023). A benefit of using Artificial Neural Networks over other machine learning algorithms is that Artificial Neural Networks are more customisable due to the researchers' ability to specify the number of neurons, hidden layers and activation functions in the architecture of the model. This means the researcher can construct a tailored made Artificial Neural Network that is useful to the dataset and learning problem at hand. On the other hand, this same strength also underlines the current challenge when implementing Artificial Neural Networks which consists of the issue of the researcher having to utilize a 'trial and error' approach when constructing an optimal Artificial Neural Network for a certain task. Moreover, unlike machine learning algorithms such as random forest trees, there is no effective strategy for automating parameter tuning and architecture tuning in the same vein as GridSearchCV, which makes constructing a useful Artificial Neural Network model a more time-consuming approach than deploying other machine learning algorithms. Another drawback to deploying Artificial Neural Networks is depending on the complexity of the model they can take much longer to train on the datasets than other machine learning algorithms (Yegnanarayana, 2006, pg 134-135).

Ultimately, there has so far been no published attempts in using machine learning algorithms to build a complete voter-turnout prediction model for a general election in the UK. This provides a unique suggestion this paper will help uniquely contribute to the surrounding literature In terms of using machine learning algorithms for directly predicting electoral turnout. There have been a limited number of models built based on using machine learning algorithms that have been tested on predicting voter turnout. However, these turnout models that use machine learning are mostly based on US election turnout. Whether the types of data selected and used as features would apply to an accurate UK general election will be discussed in the next section. A model created by Bloomberg mainly used machine learning algorithms to predict the turnout of the 2020 presidential election. The Bloomberg electoral turnout model had notable structure within its methodology such as accruing data on prior elections, as well as testing and tuning the model on several different elections, using them as holdout test sets (McCartney et al, 2020). As the Bloomberg model was built before the occurrence of the 2020 election the model

accounted for different scenarios, such as low turnout, medium turnout and high turnout scenarios. While this increases the robustness of any predictions made by the model it also constrains the final predicting power of the model as a predisposed setting needs to be computed in the model, undermining the preciseness of the predictive capabilities of the model. Machine learning has been implemented on numerous occasions to find explanations for the voter turnout phenomena in different nations. Machine learning algorithms have been used to identify the strength of demographic factors in voter turnout in the 2016 US elections (Kim et al, 2020, pg 978-979), as well as the use of Random Forest algorithms being used in determining demographic differences in turnout in an Australian State election (Hoffman & Lazaridis, 2013, pg 34-37). Despite these contributions, the current literature is mainly limited to using machine learning algorithms to find explanatory variables rather than constructing accurate predictive models.

ii. Variables that Affect Voter Turnout

Variables that affect voter turnout will be vital in the construction of this voter turnout model. Rather than randomly collecting data to add to the model the approach taken in this paper will seek to apply a series of unique features based on the literature that has been found to affect voter turnout in UK elections. One factor that has been found to influence voter turnout is the number of significant terrorist attacks that have occurred before an election. Furthermore, the frequency of terror-related attacks was concluded to have a positively correlated statistically significant relationship with voter turnout (Robbins et al, 2013, pg 502-404). One other factor that has been observed to have an impact on voter turnout is electoral competitiveness. This variable has been found to normally have a statistically significant relationship with voter turnout. The closer the election, the higher the electoral turnout usually is (Geys, 2006, pg 652-653). Electoral competitiveness is also outlined as a contributing factor to a general decline in overall electoral turnout since the 1950s (Pattie et al, 2018, pg 101).

Along with the non-demographic features outlined factors, there will be demographic-based features added to the model to attempt to enhance the model predictive capabilities. A trend that influences the UK general election turnout is age. It is an often observed trend that members of the electorate who are members of older age brackets are more likely to vote than electorate members in the lower age brackets (Smets, 2012, pg 407-409). This trend would suggest that electoral constituencies with a higher percentage of older age groups will likely have a higher electoral turnout than constituencies with a higher proportion of younger age groups. The population of a constituency has also been found to relate to voter turnout. A trend discovered equates to a lower population having a higher voter turnout rate in general elections (Geys, 2006, pg 652-653). However, the demographic data added to the model will not include the ethnic population of each constituency as it is suggested that the diversity of an area has no overall effect on voter turnout (Fieldhouse & Cutts, 2008, pg 545-546). This coupled with the limited access of precise data on ethnic and minority populations per

constituency in between elections makes ethnicity a feature that will not be likely to be an effective predictor in voter turnout. Another notable feature absent from having a notable relationship with voter turnout figures is the demographic of gender. Gender has been found to play a minor role in political participation. Although it has been observed there is no overall difference between gender and voter turnout (Childs, 2004, pg 423-424). Based on this, it seems unnecessary to include gender as a data feature in our electoral turnout model.

Economic factors such as unemployment may play a role in affecting voter turnout. Studies have shown unemployment, especially long-term unemployment leads to political disengagement, and therefore a decreased likelihood of electoral participation (Uberoi, & Johnston, 2021, pg 25-27). Furthermore, studies on European countries suggest that although unemployment does not affect turnout significantly overall on a national level, unemployment affects turnout on a subnational level (Azzollini, 2021). Therefore, unemployment may be a useful indicator in determining a constituencies voter turnout rate due to its more localised relation with voter turnout. Interestingly, in terms of other economic phenomena, such as income inequality there is evidence to suggest that it has no impact on electoral participation and voter turnout (Stockemer, Scruggs, 2012). The literature suggests economic factors, in general, do not play a large role in determining voter turnout and other factors such as social, structural and demographic factors are more important in determining electoral turnout.

In terms of education levels and their effect on voter turnout, it has been found additional education plays no overall role in voter turnout. Only a slight role in political participation was found with additional education levels (Pelkonen, 2010, pg 68). Because of the established weak influence education plays in voter turnout, there will be no education features included in the models.

While the literature does provide conclusions on factors that contribute to voter turnout, the literature on UK elections ultimately lacks the sheer number of studies on its elections compared to the US. The literature, in general, is rather US-centric with many studies conducted on what drives voter turnout in US elections specifically. Aside from discussing current machine learning models on US elections, this appears to avoid using studies conducted on the US elections to extract features that may be prudent in constructing the UK general election turnout model due to the inherent cultural, political and economic difference between the two nations. While an argument can be made that both nations are Western democracies and therefore there is an underlying similarity between voters, overall it would not be wise to use US election based literature to extrapolate features for UK election based voter turnout as there are still underlying differences between the voters of both countries. Furthermore, models that model both US and UK elections using similar prediction features for the outcome find discrepancies between the prediction accuracies of the UK general elections and the US general elections (Lauderdale et al, 2020)

Another conclusion that can be justified from the literature analysed regards the breadth of data we should use when constructing data features of the model. As analysed in the literature many different methods and data sources can be used for building a predictive machine learning

model. This paper mainly seeks to utilize data that is either demographical, historical or economical as there has not been an attempt to build an inclusive model utilizing all these forms of data whilst using a wide range of machine learning algorithms to attempt to predict constituency turnout in the UK, despite similar methods being used to predict election results, mostly in the US.

Other factors have been discovered to influence voter turnout such as personality-driven factors, cognitive factors, personality traits (Denny & Doyle, 2008, pg 306-310) and behavioural factors (Cravens, 2020, pg 10-11). Different models of voter turnout have also been proposed, such as models that involve game theory (Dhillon & Peralta, 2002, pg 334-335). These features and models, while correlated and have meaningful influences on voter turnout, won't be discussed at length in this paper due to the practical implausibility of acquiring the data to use as features or incorporate other models in the models used in this paper. Prior features discussed in this section can be incorporated into the model as they are features with accessible data.

3. Methodology

i. Data collection and Preprocessing

To construct a model suited for predicting the electoral turnout of the 2019 election, high-quality data needs to be collected and combined into a Dataframe. Necessary justifications on compromises made for practicality will be addressed in the methodology of constructing the voter turnout model. Data collection for building the model that will be trained will involve data from the last three general elections (2010, 2015 and the 2017 elections in the UK). The last three elections were specifically chosen for two reasons. Reason one is the data available for the model is more limited for prior elections, therefore if we were to take older elections into account there would be an inevitable loss of features in the model. The second reason consists of the idea that to account for variability that occurs between elections and to keep the data trend more in line with modern elections, the training data is, however, limited.

Most of the data used for building the model was acquired from the House of Commons data Constituency dashboard (Watson et al, 2020). This dashboard contained lots of data relevant to the construction of this model, most importantly, the data was formatted to pertain relevant information to each constituency. The population feature was taken into account with the electorate data feature as the number of people registered in the electorate correlates with the voting age population of constituencies. A decision was made to use data results from the prior election as electoral competitiveness would be difficult to directly model for the standard model built as it would require opinion polls for each constituency. While this may not quite capture the immediate level of electoral competitiveness in an election cycle, this data is still sufficient to determine what seats are more likely to be ‘safe’ seats in an election and therefore indicate the likely competitiveness of a constituency, as with safe seats in general, there is an unlikelihood

that they will vote in another party within the next election. Election statistics and data involving constituencies electoral turnout and electoral results by party from the 2005 election was obtained (Watson et al, 2020). This data was then merged with data pertaining to the constituency, electorate and electoral turnout of the 2010 election (Watson et al, 2020). Unfortunately, due to the method in which the data identification of the constituencies uses a different system between the two elections, there was data loss in this instance of 216 constituencies. The electoral results data of the 2010 election, as well as the turnout, was then subsequently merged with the electorate data and turnout data of the 2015 election (Watson et al, 2020). This process was repeated with both the 2017 and 2019 datasets (Watson et al, 2020), there was no instance of data loss throughout the merging of these electoral datasets due to all these datasets using the standardised ONS_ID format to identify the respective constituencies. The data format most data columns contained was in the format of a percentage value. To normalise the data, the percentage values were converted into a decimal format of 4 significant figures.

Once the electoral data were collected and merged, the Country/Region column in the DataFrame was converted into dummy variables for each election dataset. This was implemented for the model to account for the geographical location of the constituency, to observe whether it would have an impact on the prediction model.

Once this was completed the other auxiliary data features were collected. Data relating to the population of each age group in each constituency was retrieved from the House of commons Constituency data dashboard for the years 2015, 2017 and 2019 respectively (House of Commons Library, Constituency data: Population, by age., 2021). These datasets were separated and converted into DataFrames of each of these years. The age groups were orientated and converted as the columns while the ONS_IDs were formatted as the rows of the dataset, this was done to allow for merging with the respective election datasets. The values of the data in the age dataset were also converted from a standard percentage format into a decimal format of 4 significant figures to keep data values standardised consistent. The data for the 2010 elections was not available from the House of commons Constituency dashboard, therefore the data in another format was collected from the House of Commons library (Park, 2020). This data required editing for it to be formatted the same as the other age-related datasets used. The editing required the column groups that were partitioned into age groups every 5 years to be combined into age groups that were partitioned into age group splits of every 10 years, to be consistent with the rest of the data collected. The data values regarding the 2010 population groups per constituency were also incompatible as the values were not in a percentage format. To rectify this the values were divided by the column regarding total population per constituency to then obtain the percentage value in a decimal format. The columns of each Age dataset were then merged with their corresponding election year on constituency ID. There was again data loss that occurred in the 2010 election.

The next feature to be added to the model is the number of terror attacks that happened before the election. Finding a global terrorism database, (The National Consortium for the Study

of Terrorism and Responses to Terrorism, 2021), only attacks that happened 12-months before the election will be included as values. Determining the number of terror attacks to be included in the features will rely on data from the Global Terrorism database and will include the number of violent attacks that have occurred that involve injury or death. The reason for setting this definition of terror attacks that will qualify for the model is that these attacks receive national media coverage and are therefore relevant to be included in the model. The number of attacks are then recorded and computed up to 12 months before each election and added as a data column which is then merged to the respective dataset As terror attacks receive national coverage and have a statistically significant effect on turnout the value is added to each constituency, rather than just the constituency the attack happened in as covered terror attacks have national significance.

The last feature added to the dataset will be the unemployment rate of each constituency recorded at the time closest to the elections. The data on constituency unemployment was also retrieved from the House of Commons constituency data dashboard (House of Commons Library, Constituency data: People claiming unemployment benefits, 2021). The dataset was separated into data for 2010, 2015, 2017 and 2019 election datasets with the unemployment rates being retrieved matching to a month before each election. The unemployment rates were then merged with the election datasets using the ONS_ID data column.

When all the features were combined to each election dataset, the datasets of the 2010, 2015 and 2017 elections were appended to create a DataFrame which will then be ready to be utilised as the training dataset. The 2019 election Dataframe is left out due to its status as the holdout test set for the trained model to be used to predict the turnout.

Once all the DataFrames had been merged there was the issue of NAN values being present in numerous features mainly concerning the political party results in the prior election. The NAN values were all set to zero regarding the context of the data being voted for a political party that was not present in the constituency. For example, the SNP only has results for Scottish constituencies as it is a political party based exclusively in Scotland. Therefore setting the NAN values to 0 is the most optimal way of handling this data problem.

ii. Model Selection

The machine learning algorithms that will be trained on the training dataset and tested to predict the electoral turnout of the 2019 election include multiple linear regression., Decision Tree Regression, Random Forest Regression, Support Vector Machine Regression, Gradient Boosting Regression and an Artificial Neural Network. Multiple linear regression attempts to fit the coefficients using a linear approximation to minimize the residual sum of squares with the dataset and the target variable, which in this case is the turnout feature (Géron, 2019, pg 112-114). Decision Tree Regression utilizes observations that are split into ‘branches’ to predict a continuous target variable that makes up the ‘leaf’ component of the algorithm’s structure (Géron, 2019, pg 183-185). Random Forest Regression is a more elaborate form of the Decision

Tree algorithm in which a random forest is a meta estimator that fits many classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (Géron, 2019, pg 198-199). Gradient Boosting Regression is a unique algorithm that allows for the optimization of arbitrary differentiable loss functions. In each stage, a regression tree is fit on the negative gradient of the given loss function. Finally, Support Vector Machine Regression is a versatile algorithm that can encompass both linear and non-linear forms of regression based on the kernel in which margin bounds are used and fitted on with the purpose of there being limited margin violations (Géron, 2019, pg 203-206). All the algorithms aside from the Artificial Neural Network will be implemented using their respective SciKit-Learn library. The Artificial Neural Network on the other hand will be built using the Keras TensorFlow library. This specific selection of models has been chosen as it provides a diverse range of machine learning algorithms that can be compared with one another to tackle predicting voter turnout. This selection also allows the comparison of more elaborate non-linear algorithms such as Random Forest Trees, Support Vector Machines and Artificial Neural Networks to the more traditional models utilised in political science such as the multiple linear regression model, as outlined in the literature.

Once the data has been collected and preprocessed per the first part of the methodology, training and testing of the model is then completed. Rather than using a traditional train-test split as what is commonly used in machine learning model building to split the data, K-Fold Cross-validation will be the splitting method of choice. The advantage of using K-Fold cross-validation is it validates a model over multiple folds of the data to provide a more constant value of the overall performance of the model by taking the average of all folds, rather than just one split which is traditionally observed with the train test split. Selecting the optimal K value for the cross-validation is important for the training of the turnout model, with the optimal value for K pertaining to be 10 folds as that is the optimal number in terms of the tradeoff between computing time with bias and variance on a reasonably sized dataset (Kohavi, 1995, pg 1140).

The evaluation metric to judge the quality of the model on the test data and holdout sets were selected. The evaluation metric is the coefficient of determination (R^2 score). The justification for using this metric is the R^2 scores of each respective model give a clear normalised comparison of how well the models perform over the set of predictions. The R^2 scores further indicate how well the model predictions correlate or fit the true values.

iii. Hyperparameter Tuning and Artificial Neural Network Architecture

From the results of the model selection methodology, The hyperparameter tuning for the Random Forest Regression, Gradient Boosting Regression, Decision Tree Regression, Support Vector Machine Regression and Multiple Linear Regression was done using the GridSearchCV process to determine the optimal hyperparameters of each model respectively. The hyperparameters tuned will be specific to each model. Only the most important hyperparameters will be tuned to avoid unnecessarily long training times.

Building the Artificial Neural Network with optimal parameters was a more challenging undertaking. This is because, unlike the other machine learning algorithms, as outlined, constructing the optimal architecture and parameter tuning for Artificial Neural Networks required a manual trial and error approach, experimenting with different neural network layers and activation functions. Upon trailing many different architectures, the one most suited for this regression task had the following structure outlined in Figure 2, the activation function within the Artificial Neural Network hidden layers was Rectified Linear Unit (ReLU). ReLU was used as it has been commonly used to fit non-linear data. Other advantages also include a faster training time when compared to other activation functions. The R^2 scores were recorded as the mean of 10 iterations of the neural network being trained on 1000 epochs. As the model involves numerous hidden layers of many neurons, a dropout layer was added between each hidden layer. Dropout layers were added due to their effectiveness in reducing the potential of the neural network to overfit the training data (Srivastava, et al, 2014, pg 1951-1952). Dropout layers were set to a value of 0.2 for every hidden layer.

4. Results

i. Model Results

The Results are split into two separate parts. In the first part of the analyses of the results the R^2 scores values, each model produced on the training set. This result was obtained to see how well the models perform predicting turnout within the dataset before it is trailed on the unseen 2019 election dataset. Table A shows the result.

Model	**R^2 Score on Training Data**
Multiple Linear Regression	0.1100
Decision Tree Regression	0.3801
Random Forest Regression	0.4253
Support Vector Regression	0.1191
Gradient Boosting Regression	0.4021
Artificial Neural Network	**0.7594**

Table A - Training Data Turnout Scores

The results of Table A show that within the training the Artificial Neural Network performs rather well with an R^2 score of 0.7594. This score is significantly higher (0.3573 higher than Random Forest regression) than the other machine learning algorithms trained on the same data. The second, third and fourth highest scoring algorithms, Random Forest Regression, Gradient

Boosting Regression and Decision Tree Regression respectively R^2 score values close to 0.4. While still clearly falling short of the Artificial Neural Network, both models still obtained an R^2 score in which there is a clear link coefficient between the data and target value of voter turnout. On the other hand, the worst-performing models on the training data according to the R^2scores were Multiple Linear Regression and Support Vector Machine Regression, both garnering an R^2 score that barely indicates a significant relationship between the feature variables and the target variables since any R^2 score below 0.1 can indicate the relationship between feature variables and target variables is statistical noise. Based on the results of the training data, the better prediction models are better suited to more non-linear relationships between the data features and target variable, which explains why the Artificial Neural Network and Tree-based algorithms performed better than the linear regression algorithm.

The results in Table B below show how well the models performed predicting the turnout values of the 2019 election for each respective machine learning algorithm trialled.

Model	**R^2 Score on 2019 Election Prediction Dataset**
Multiple Linear Regression	-0.1561
Decision Tree Regression	-0.4102
Random Forest Regression	-0.1826
Support Vector Regression	-0.1268
Gradient Boosting Regression	-0.0049
Artificial Neural Network	**0.1902**

Table B - 2019 Election Turnout Scores

The results in Table B show an unexpected deviation in R^2 scores from the results seen in the training data. All the models perform markedly worse in predicting constituency turnout from the 2019 election. The most superior model is still the Artificial Neural Network. However, the R^2 score dropped significantly compared to the training data (0.5692 difference). Moreover, the other algorithms went from having positive R^2 scores to negative ones, making them all weak algorithms for predicting a turnout of this particular prediction task. The biggest difference can be observed in the R^2 values of the Tree-based regression algorithms as the training dataset contained substantially different and better R^2 scores than in the results of the 2019 test dataset. Although a slight drop in model effectiveness was to be expected due to the differences in the context of the unseen attribute of the 2019 dataset, this large drop in the model R^2 scores suggests the data features may not have been high quality enough to build a prediction model for the 2019 general election.

The Following Table C below shows the cumulative overall prediction for the turnout of the 2019 election in each respective model. This was calculated by finding the mean turnout percentage of all the constituencies combined, since the dataset used did not predict any raw values on turnout. The recorded turnout of the 2019 election was found to be 67.3%, according to official figures from the UK government (McInnes, 2021).

Model	**Overall 2019 Election Turnout Prediction**
Multiple Linear Regression	66.62%
Decision Tree Regression	68.90%
Random Forest Regression	68.49%
Support Vector Machine Regression	66.10%
Gradient Boosting Regression	67.76%
Artificial Neural Network	**66.92%**

Table C - Overall Turnout Prediction 2019 Election

As observed in Table C, the Artificial Neural Network based model comes the closest to predicting the overall electoral turnout of the 2019 election, falling short of predicting overall turnout by 0.38%. Furthermore, the Gradient Boosting regression model is a very close second in predicting the overall turnout of the 2019 election scoring by 0.46% above the actual value. The rest of the models miss the recorded turnout value more substantially, with the worst prediction coming from the Decision Tree Regression model, missing the actual value of prediction by 1.6%. The results in Table C suggest that these algorithms may have a better ability to predict overall voter turnout in an election, rather than turnout on the smaller constituency level. This is due to the fact the artificial Neural Network and Gradient Boosting models prediction values were reasonably close to the actual value, despite the fact the R^2 values for constituency prediction suggest the model is far off from being reliable prediction models on a constituency basis.

As the Artificial Neural Network model was the highest performing model according to the R^2 values on both the training data and the 2019 prediction dataset, there was a comparison of the predicted constituency turnout values to the recorded constituency turnout values on 2019 general election. Figure 1 below presents a geographically mapped comparison between the two values for each subsequent constituency. This map was constructed to visualise the constituencies the model predicted the best in and which constituencies the model performed the worst at predicting.

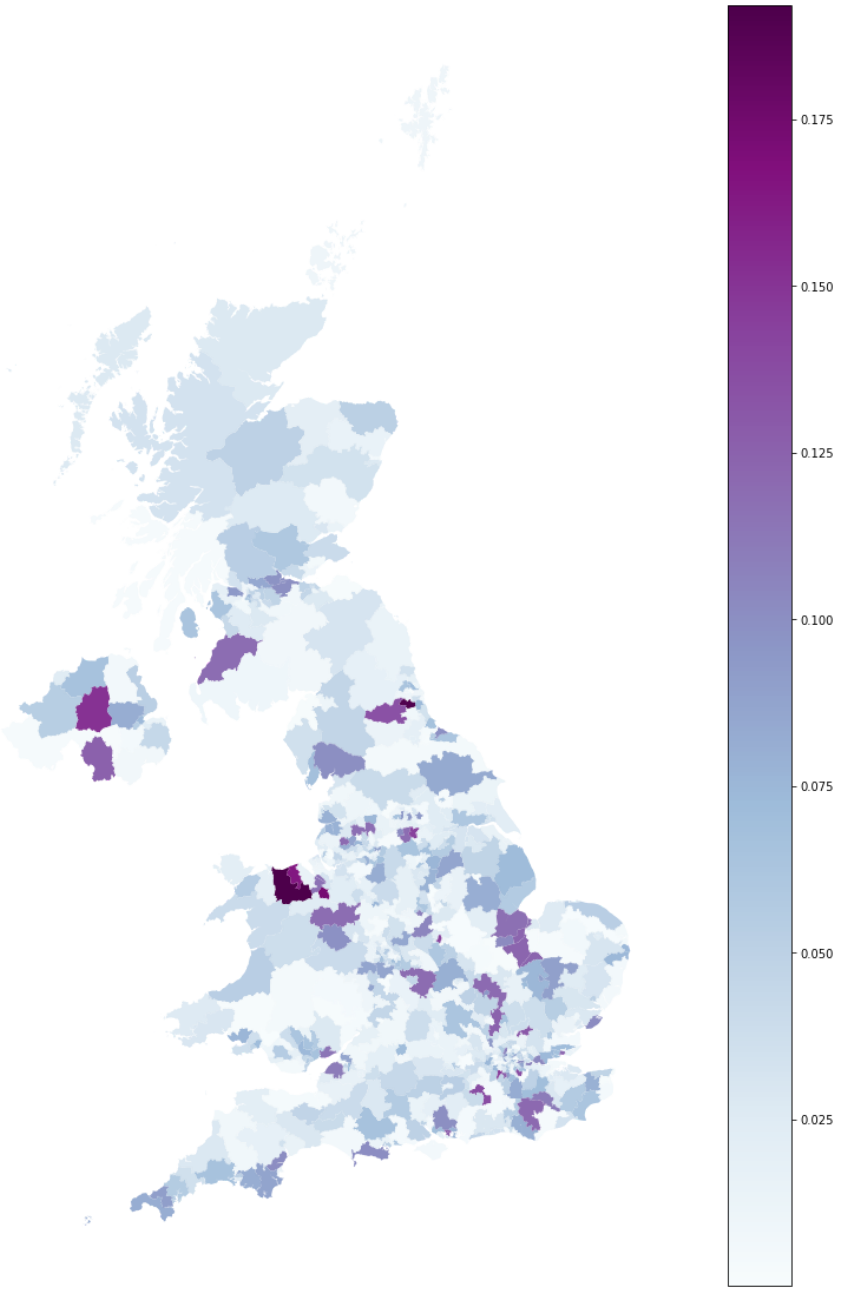

Figure 1 - Visual Representation on the Accuracy of the Artificial Neural Network Predictions on the 2019 Election

Figure 1 shows the wide distribution and variability of how accurate the models predictions were. In general observation. The darker the blue of the constituency, the less accurately the turnout predictions between the predicted value and the true value match. There are no specific grouped geographical areas in which predictions were specifically poorer than other areas. Constituencies where the Artificial Neural Network model was the poorest at predicting seems to be randomly distributed throughout the UK rather than particularly localised.

ii. Feature Importance

One method that can help further the understanding of which data features were beneficial and had the most positive impact on the model. The method used to devise feature importance is permutations. Permutation was used as it surpasses the limitation of impurity-based feature importance and can be used on a holdout test set, which is useful as due to the nature of the results, it is more useful to learn of the feature importance on the 2019 election data. As shown in Figure 2, we can see the most important features in regards to the Gradient Boosted Regression model.

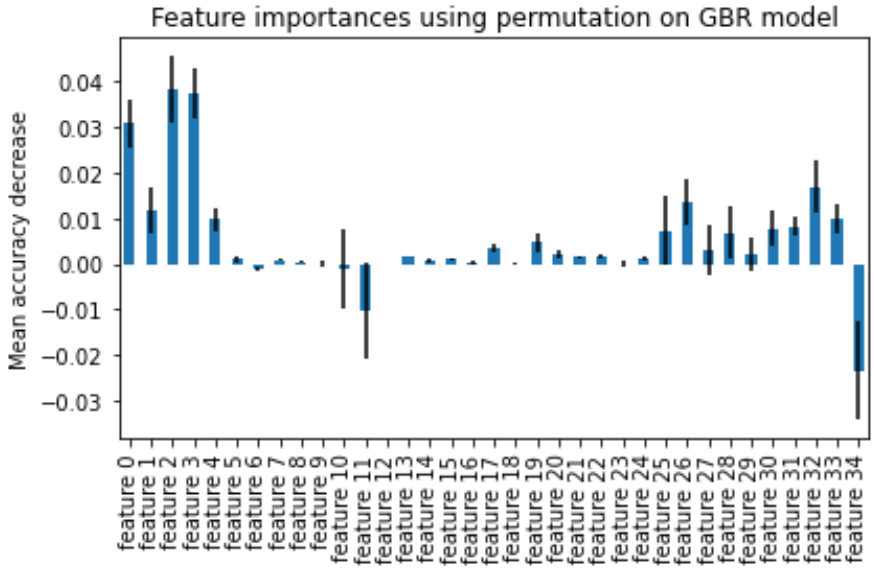

Figure 2 - Feature Importance Permutations of the Gradient Boosted Regression Model (GBR)

Gradient Boosted regression was the model chosen to calculate feature importance through permutation, as of the time of writing this paper Keras does not have support for calculating feature importance. Therefore, the second-best Gradient Boosted Regression model was chosen as it is the best performing model that supports SciKit-learn implementation of calculating feature importance via permutation. Results in Figure 2 show that the first few features have a relevant positive mean accuracy decrease value. Features 0-4 include the electorate, Conservative vote share, Liberal Democrat vote share, Labour vote share and SNP vote share are all features that are useful to the model. Interestingly, Liberal Democratic vote share and Labour vote share have higher feature importance than the Conservative vote share according to permutation. Features added additionally such as age groups also have significant relevant feature importance, especially the 10-19 (feature 26) age group and 70-80 (feature 33) age group both have high permutation mean accuracy scores. Constituency location dummy encoded features have relatively little feature importance compared with the already outlined. The features with negative permutation values include the prior turnout feature (feature 11) as well as the unemployment feature (feature 34). The negative permutation values suggest these features are not useful to the prediction model as randomly initiated values for these features would have

better prediction relevance. Therefore, in future iterations of similar models, it would be prudent to remove these data features as they offer no benefit to the prediction model. Feature 12, the feature associated with the terror attack feature also has no relevance to the model predictive capability as it has a negligible value. It may be worth noting this could be the case due to the universality of the values contained in the feature.

6. Discussion

Unfortunately, even with the data utilised with the methodology, the overall prediction results on the 2019 election turnout was disappointing. Although the main issue of poor prediction likely stems from the inability to obtain and use enough quality data to build an accurate model. This is one of the largest challenges in building an accurate prediction model using machine learning algorithms that needs to be discussed. In an ideal environment, a surplus of all the data discussed in the literature would have been available to build a conclusive model. Even data that seems more obscure may impact voter turnout rates in a given constituency. An example of this is the number of polling stations there are at the time of a given election (Orford et al, 2011. pg 149-150). Data like this is still rather difficult to obtain in a practical non-time consuming manner for a researcher, especially if they do not have enough evidence to support the potential effectiveness of introducing it as a feature. Therefore, the practicality of obtaining useful data indicates that it can be useful data features is one of the most prominent challenges in applying machine learning algorithms to create accurate prediction models in social science.

In future experiments constructing a voter turnout model of this kind, it would likely be beneficial to increase the number of elections used in the training data. The reason why more prior elections were not added to the training dataset was that the further back the election the fewer data could be found regarding constituencies. The methodology used in this paper was to construct a model using higher quality, more complete datasets rather than high quantity data that contains many significant missing values. Although the results suggest that a much higher quantity of data is needed to build a better prediction model. Another large problem with building a voter turnout model to predict constituency turnout includes that the number of constituencies changes over time. Meaning, depending on the election there may be a significantly different number of constituencies in different locations which may make the individual training instances of the model more unreliable the more historical election is used.

A methodological step that can be taken to potentially improve the results of the predictions is to expand the scope of the number of different machine learning algorithms used. The reason the 6 models were chosen and used for this paper was mainly due to their established effectiveness and varied parameters and mathematical components. There are a much larger range of algorithms out there that may boast a better performance if trialled on these datasets. However, due to the consistently weak prediction variables on the holdout set for all the machine

learning algorithms, the overarching dominant issue seems to be the lack of informative data features as already outlined, rather than the selection of the machine learning algorithm

Another area of improvement that should be investigated in further research on voter turnout models is more precise feature engineering. Most of the features utilised in this model were features directly obtained from the datasets used and created through the methodology. These features may not have been precise enough in terms of informative data to get the best use of the data. As an example, rather than just including the results of the prior elections as the training DataFrame, it could be more prudent to simplify these features into a general electoral competitiveness feature that consists of the difference between the party that obtained the highest number of votes and the rest of the parties.

A further element to address that is likely a potential reason why the models did not perform well was due to extraneous elements and contexts that can dominate an election. The elephant in the room for the 2019 election was, of course, Brexit. Much of the 2019 election was enamoured over different strategies and responses to the EU referendum results in 2016, in which there was a majority who voted to leave the European Union. The reason why Brexit has not been addressed in the methodology behind building the turnout model was that it is an incredibly difficult political phenomenon to model. Furthermore, for the model trained to be somewhat generalisable, it was decided in this methodology not to focus resources on trying to incorporate Brexit regarding data features, as Brexit only became the forefront issue for the electorate specifically in 2017 and 2019 elections. Although, after analysing the results it would have perhaps been very useful for the prediction model to have involved data regarding Brexit for the 2019 election, as it can be argued the 1.5 percentage points decrease in turnout for the 2017 election seen was specifically due to the drawn-out unclarity on where the country was heading in regards to its relationship with the EU (Cutts et al, 2020, pg 9-10).

Voter fatigue is another aspect of the voter-turnout topic worth discussing, as between 2015 and 2019, there were 3 general elections as well as the EU referendum. Due to the relatively high frequency of nation-wide votes that would otherwise only occur once every 4-5 years, it is reasonable to postulate that the overall decline in turnout observed in the 2019 election could be due to disenfranchisement and discouragement of the electorate due to how many elections have happened in an untypical timeframe (Garman, 2017, 33-34). To involve this as a feature in the machine learning model it would be useful to add a data feature stating the amount of time it has been since the prior election. However, the feature would only likely be useful if a similar pattern before elections in the training set has been observed.

7. Conclusion

While there are hints of potential in using machine learning algorithms to predict voter turnout, the results of this paper suggest there is still much work to be done to precisely narrow down useful data features that can be used to predict voter turnout. As observed through permutation of

feature importance, many of the features added and supported by the literature show a utility for model prediction on unseen elections. Data features of age groups, as well as electorate and prior electoral results, show promise for further utilization in a predictive model involving electoral turnout. This is in contrast to data features such as unemployment and prior electoral turnout which are features that have no use as far as the models constructed for this investigation are concerned. Despite this, however, much larger availability and access to relevant data are also to be needed to be able to build better quality models that utilise machine learning. Out of all the models trialled in this investigation, Artificial Neural Networks was the best suited for the task despite its relatively weak performance on the 2019 election turnout predictions. Artificial Neural Networks would especially work more effectively when provided a larger quantity of data to train on, as it seems the most effective machine learning algorithm for finding complex relationships between data features in regards to predicting electoral turnout. Ultimately, while choosing the right algorithm, increasing the quantity of data and improving the quality of features may unlock the full potential of machine learning to create useful prediction models in space of electoral turnout, there may always be extraneous variables and unforeseen contextual elements that machine learning orientated turnout models may fail to capture, such as the Brexit in the case of UK, that keeps electoral turnout an elusive problem to precisely capture and predict.

8. Bibliography

Azzollini, L. (2021). The scar effects of unemployment on Electoral PARTICIPATION: Withdrawal and Mobilization across European societies. European Sociological Review. https://doi.org/10.1093/esr/jcab016

Bansal, B., & Srivastava, S. (2018). On predicting elections with hybrid topic-based sentiment analysis of tweets. Procedia Computer Science, 135, 346-353. doi:10.1016/j.procs.2018.08.183

Chen, N., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science. ACM Transactions on Interactive Intelligent Systems, 8(2), 1-20. doi:10.1145/3185515

Childs, S. (2004). A British gender Gap? Gender and political participation. The Political Quarterly, 75(4), 422–424. https://doi.org/10.1111/j.1467-923x.2004.00646.x

Cravens, M. D. (2020). Measuring the strength of voter turnout habits. Electoral Studies, 64, 102117. doi:10.1016/j.electstud.2020.102117

Cutts, D., Goodwin, M., Heath, O., Surridge, P. (2020). Brexit, the 2019 general election and the realignment of British politics. The Political Quarterly, 91(1), 7–23. https://doi.org/10.1111/1467-923x.12815

Denny, K., Doyle, O. (2008). Political interest, Cognitive Ability And Personality: Determinants of voter turnout in Britain. British Journal of Political Science, 38(2), 291-310. doi:10.1017/s000712340800015x

Dhillon, A., & Peralta, S. (2002). Economic theories of voter turnout. The Economic Journal, 112(480). https://doi.org/10.1111/1468-0297.00049

Di Franco, G., Santurro, M. (2021) Machine learning, artificial neural networks and social research. Qual Quant 55, 1007–1025 . https://doi.org/10.1007/s11135-020-01037-y

Fieldhouse, E., &; Cutts, D. (2008). Diversity, density and turnout: The effect of neighbourhood ETHNO-RELIGIOUS composition on voter turnout in Britain. Political Geography, 27(5), 530-548. doi:10.1016/j.polgeo.2008.04.002

Garmann, S. (2017). Election frequency, choice fatigue, and voter turnout. European Journal of Political Economy, 47, 19–35. https://doi.org/10.1016/j.ejpoleco.2016.12.003

Géron, A. (2020). Hands-on machine learning with Scikit-Learn, Keras, and Tensorflow: Concepts, tools, and techniques to build intelligent systems. O'Reilly.

Geys, B. (2006). Explaining voter turnout: A review of aggregate-level research. Electoral Studies, 25(4), 637-663. doi:10.1016/j.electstud.2005.09.002

Grimmer, J., Roberts, M. E., Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. Annual Review of Political Science, 24(1), 395-419. doi:10.1146/annurev-polisci-053119-015921

House of Commons Library, Constituency data: Population, by age. House of Commons Library. (2021, June 14). https://commonslibrary.parliament.uk/constituency-statistics-population-by-age/.

House of Commons Library, Constituency data: People claiming unemployment benefits. House of Commons Library. (2021, July 15). https://commonslibrary.parliament.uk/constituency-data-people-claiming-unemployment-benefits/.

Lauderdale, B. E., Bailey, D., Blumenau, J., & Rivers, D. (2020). Model-based pre-election polling for national AND sub-national outcomes in the US and UK. International Journal of Forecasting, 36(2), 399-413. doi:10.1016/j.ijforecast.2019.05.012

McCartney, A., Harris, B., Rojanasakul, M., Burgess, J., Murray, P., & Vann, A. (2020). Explaining the Bloomberg News 2020 Election Turnout Model. Retrieved August 17, 2021, from https://www.bloomberg.com/graphics/2020-us-election-results/methodology

McInnes, R. (2021, June 28). General election 2019: Turnout. House of Commons Library. https://commonslibrary.parliament.uk/general-election-2019-turnout/.

Hindman, M. (2015). Building better models. The ANNALS of the American Academy of Political and Social Science, 659(1), 48-62. doi:10.1177/0002716215570279

Hoffman, R. &, Lazaridis, D. (2013). The limits OF Compulsion: Demographic influences on voter turnout in Australian state elections. Australian Journal of Political Science, 48(1), 28-43. doi:10.1080/10361146.2012.755670

Kim, S. S., Alvarez, R. M., & Ramirez, C. M. (2020). Who voted in 2016? Using Fuzzy forests to Understand voter turnout. Social Science Quarterly, 101(2), 978-988. doi:10.1111/ssqu.12777

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 14th international joint conference on Artificial intelligence - Volume 2. 1137–1143.

Orford, S., Railings, C., Thrasher, M., & Borisyuk, G. (2011). Changes in the probability of voter turnout WHEN Resiting Polling Stations: A case study in Brent, UK. Environment and Planning C: Government and Policy, 29(1), 149-169. doi:10.1068/c1013r

Pattie, C., Hartman, T., & Johnston, R. (2018). A close-run THING? Accounting for changing Overall turnout in UK general elections. Representation, 55(1), 101-116. doi:10.1080/00344893.2018.1555676

Park, N. (2020, September 9). Parliamentary constituency population estimates (Experimental Statistics) - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/parliamentaryconstituencymidyearpopulationestimates.

Pelkonen, P. (2010). Length of compulsory education and voter turnout—evidence from a staged reform. Public Choice, 150(1-2), 51–75. https://doi.org/10.1007/s11127-010-9689-3

Richardson, B., & Hougen, D. F. (2020). Districts by Demographics: Predicting U.S. House of representative elections using machine learning and demographic data. 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). doi:10.1109/icmla51294.2020.00136

Robbins, J., Hunter, L., & Murray, G. R. (2013). Voters versus terrorists: Analyzing the effect of terrorist events on voter turnout. Journal of Peace Research, 50(4), 495-508. doi:10.1177/0022343313479814

Smets, K. (2012). A widening generational divide? The age gap in voter turnout through time and space. Journal of Elections, Public Opinion & Parties, 22(4), 407-430. doi:10.1080/17457289.2012.728221

Srivastava, N., Hinton, G., Krizhevsky , A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15, 1929–1958.

Stockemer, D., & Scruggs, L. (2012). Income inequality, development and electoral turnout – new evidence on a burgeoning debate. Electoral Studies, 31(4), 764–773. https://doi.org/10.1016/j.electstud.2012.06.006

The National Consortium for the Study of Terrorism and Responses to Terrorism. Global terrorism database. GTD. https://www.start.umd.edu/gtd/. Retrieved August 17, 2021.

Tsai, M., Wang, Y., Kwak, M., & Rigole, N. (2019). A machine learning based strategy for election result prediction. 2019 International Conference on Computational Science and Computational Intelligence (CSCI). doi:10.1109/csci49370.2019.00263

Uberoi, E., & Johnston, N. (2021, August 20). Political disengagement in the UK: Who Is disengaged? House of Commons Library. https://commonslibrary.parliament.uk/research-briefings/cbp-7501/.

Watson, C., Uberoi, E., Loft, P. (2020, April 17). General election results from 1918 to 2019. House of Commons Library. https://commonslibrary.parliament.uk/research-briefings/cbp-8647/.

Yegnanarayana, B. (2006). Artificial neural networks. Prentice-Hall of India.

8. Appendix

A. Training Data with all the features

	Electorate	con_vote_share	lib_vote_share	lab_vote_share	snp_vote_share	plc_vote_share	dup_vote_share	sf_vote_share	sdlp_vote_share	uup_vote_share	oth_vote_share	turnout	prior_turnout_1	terror_attack	East Midlands	Eastern	London	North East	I
count	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
mean	71428.389511	0.346735	0.166756	0.324561	0.024465	0.007606	0.005538	0.005109	0.003161	0.003262	0.112807	0.663226	0.645620	1.532115	0.076016	0.094873	0.121391	0.048321	
std	8055.795066	0.156398	0.121493	0.167307	0.100028	0.038064	0.044606	0.044371	0.029048	0.027178	0.086962	0.056501	0.060869	1.945074	0.265103	0.293126	0.326677	0.214506	
min	21301.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.443100	0.372100	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	66846.000000	0.218500	0.059400	0.175800	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.051200	0.627300	0.607700	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	71943.000000	0.367200	0.154700	0.336700	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.084700	0.668200	0.650600	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	76460.000000	0.482300	0.233000	0.453800	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.169300	0.704300	0.691100	4.000000	0.000000	0.000000	0.000000	0.000000	
max	110683.000000	0.658800	0.619700	0.813000	0.619100	0.443300	0.498100	0.710800	0.484800	0.464100	1.000000	0.819400	0.819400	4.000000	1.000000	1.000000	1.000000	1.000000	

North West	Northern Ireland	Scotland	South East	South West	Wales	West Midlands	Yorkshire and The Humber	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80+	UnempConstRate
1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697.000000	1697
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1697
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.047420032
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1
0.120801	0.021214	0.069534	0.137890	0.086623	0.047142	0.090748	0.085445	0.119225	0.113112	0.130290	0.128403	0.134207	0.132232	0.109394	0.080947	0.052185	NaN
0.325992	0.144139	0.254436	0.344887	0.281365	0.212005	0.287336	0.279625	0.018066	0.010940	0.047387	0.027539	0.013873	0.017866	0.023494	0.021736	0.014795	NaN
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.072197	0.055559	0.049835	0.075587	0.087200	0.071026	0.043043	0.023900	0.013900	NaN
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.106500	0.106800	0.101900	0.111300	0.125400	0.122100	0.094800	0.067246	0.042800	NaN
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.117400	0.112300	0.117700	0.124501	0.133200	0.135457	0.111359	0.081300	0.051800	NaN
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.130200	0.118800	0.138400	0.138700	0.142714	0.145400	0.126075	0.094800	0.061100	NaN
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.196939	0.176500	0.386200	0.270894	0.176730	0.167300	0.175874	0.152500	0.121813	NaN

B. Artificial Neural Network Architecture

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 256)               9216
_________________________________________________________________
dropout (Dropout)            (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_2 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_3 (Dense)              (None, 32)                2080
_________________________________________________________________
dropout_3 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_4 (Dense)              (None, 16)                528
_________________________________________________________________
dropout_4 (Dropout)          (None, 16)                0
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 17
=================================================================
Total params: 52,993
Trainable params: 52,993
Non-trainable params: 0
_________________________________________________________________
```

C. Tuned Hyperparameters

Random Forest Regression = {'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100}

Decision Tree Regression = {'max_depth': 3}

Gradient Boosting Regression = {'learning_rate': 0.02, 'max_depth': 10, 'max_features': 0.3, 'min_samples_leaf': 2, 'n_estimators': 1000}

Support Vector Machine Regression = {'epsilon': 0.1, 'kernel': 'poly'}

Multiple Linear Regression = {'fit_intercept': False}