**Problem Understanding**

With the attention paid to the forecasting of all aspects of elections continuing to increase, the importance of implementing sound practices to identify the strength, or lack thereof, regarding a model, continues to grow alongside. While Presidential year election cycles will always garner a higher focus, modeling a mid-term cycle presents its own unique set of challenges.

Often the first step in understanding the specifics of what may occur in an election cycle is understanding what the voter turnout may look like. Turnout itself, if properly predicted, can then be used as a variable to help forecast specific candidate races and legislation ballot initiatives. Stratifying predicted turnout by location and/or demographic groups can also help to provide a clearer look at the way results may ultimately fall.

Providing an assist to the turnout modeling process is the benefit of data regarding past turnout at the individual level. This, of course, is not the case regarding the specifics of who or what was actually cast on an individual’s ballot. Add to that, the fact that while candidates may still impact a person’s inclination to vote or not from cycle to cycle, that inclination is not generally as volatile as what *type of candidate* or legislation a person votes for is. With these two concepts in hand, and coupled with a robust data set detailing other demographic components, properly modeling turnout, even in a mid-term cycle begins to become more possible. The goal of the ensuing writing is to attempt to do so as well as to provide insight into the process.

**Data Understanding**

**Dataset Specifics**

The original dataset used for conducting this project consisted of 50,000 individual observations, each representing a unique voter, and 39 variables. The data was compiled in the state of Nevada following the primary elections of the 2014 mid-term election cycle. As was stated above, the ultimate goal of the project was to predict turnout for the general election as accurately as possible. The original dataset consisted of 17 demographic or characteristic variables, while the remaining 21 variables consisted of voter history on precinct turnout history. A list of the original variables within the dataset can be seen below.

Table

Description automatically generated

**Selection of Target**

With all predictive modeling practices rendered impossible sans the presence of a target variable, the first step of the project was to identify the proper one. The decision ultimately became binary—should turnout from the 2012 general election be used or turnout from the 2010 general election? The answer to this question centered around whether or not recency or consistency should be valued more. If 2012 turnout were chosen, the benefit of having the most recent possible data point for each individual would be present. However, 2012 represented a Presidential election cycle. As was briefly mentioned above, turnout in mid-term cycles can greatly differ from Presidential cycles for a variety of reasons. To quantify that statement consider the following: within the dataset the turnout for the 2012 general election (Presidential cycle) was 13,195 amongst the 50,000 individuals. However, in the proceeding 2010 mid-term cycle, the turnout was 21,580. These figures represent 63.55% decrease in turnout from the mid-term cycle to the following Presidential year. To ensure that this wasn’t an outlier the same comparison was performed on the 2008 and 2006 general election turnout figures, and similarly, a 65.08% decrease in turnout was found from mid-term to Presidential cycles.

The stark contrast between the two types of cycles led to the ultimate determination that the voting history of individuals in the 2010 general election should be used as the target variable in the trained model to ensure mid-term consistency. The chosen target was a binary variable simply indicating whether or not an individual voted in the 2010 general election or not. An added benefit of selecting the 2010 turnout as the target rather than the 2012 turnout (although not one that drove the decision making process) is that the target class is more evenly represented in 2010, with 43.16% of the individuals having voted as opposed to just 26.39% in 2012. A more even distribution amongst the two classes allows for more model stability on unseen data, and eliminates the need to oversample the minority class while preparing the data for modeling.

**Nature of the Variables**

Most of the variables in the dataset were either categorical or binary in nature. In fact, in the original dataset there existed only seven numeric variables—age, and the six variables quantifying previous elections’ turnout at the precinct level. A handful of categorical variables possessed ten or more levels in the original dataset as well. The distributions of the numeric variables were reasonably normal, which is to be especially expected with age. There was a slight right-tail due to the fact that the voting age is 18, but again, that is to be expected. Many of the binary inputs possessed extreme counts in one class. The counts of the variables and their levels can be seen in the charts below.

A picture containing text, receipt

Description automatically generated

Table

Description automatically generated

**Nan vs. Missing**

The dataset possessed no true missing values, but many values coded as “Nan” were present across numerous variables. “Nan” was simply meant to represent the absence of information. Intuitively, this could mean that the information was truly missing, or in the case of a binary variable, it could be meant to represented “No”, or “0”, or an absence of yes. While it may seem trivial on the surface, this distinction was a key to ensuring the structure of the modeling process. The approach chosen in this project was to inspect each variable one at a time and implement logic and reason to justify any changes or lack thereof. Below are the findings of that inspection.

* To begin, the variables “Party”, “CD” (Congressional District), and “DMA” (Metropolitan Area) possessed zero “Nan” values and thus were left untreated.
* The variables “Donates to Liberal Causes”, “Donates to Conservative Causes”, “Owns a Pet”, “Interested in Musical Instruments”, and “Interested in NASCAR” were all binary variables containing values of “Yes” or “Nan”. Since “Nan” represents the absence of information it also represents the absence of a “Yes” value in a binary sense. As a result, the “Nan” values in these variables were simply recoded as “No”. While there may be traces of “Nan” values that were coded as “No” inaccurately, the assumption is made that these occurrences are few and far between.
* “Marital Status” and “Home Owner or Renter” were treated very similarly to the above binary variables. While “Marital Status” actually possessed three levels—“Married”, “Non-Traditional”, and “Nan”—the “Non-Traditional” designation still indicates the presence of a marriage. Therefore the “Nan” values could be safely coded as “Not Married” and the variable could possess three levels. The same presumption was made in “Home Owner or Renter” that the absence of information could indicate that a home was not owned by the individual and that they must therefore be a renter. As a result, “Nan” was recoded as “Likely Renter”.
* The remaining categorical variables in the dataset all possessed multiple levels and could not reasonably see their “Nan” values recoded. Thus, they each possessed a certain percentage of missing values. The variables are their “Nan” (missing) percentages are:
  + Occupation (83.62%)
  + Income (39.20%)
  + Ethnicity (10.38%)
  + Education (44.82%)
  + Dwelling Type (52.17%)
  + Net Worth (51.90%)

It should be noted that none of the numeric variables in the dataset possessed any missing values.

**Data Preprocessing**

**Variable Derivation**

When modeling any dataset this large, with a fairly high number of input variables, one of the most important tasks revolves around identifying the proper set of final inputs to be fed into the model. While there are supervised ways to conduct this task that will be discussed further in the ensuing sections, there are also unsupervised methods that can help to reduce the collection of inputs that serve as candidates. In this particular dataset, voter history for each individual encompasses 15 of the 39 total variables. While the information in these 15 variables varies from cycle to cycle, both the scale as well as the essence behind what is being measured does not. As a result, a decision was made by the modeler to derive a combination of these variables. To expound, four new variables were created as derivations. They were:

* Total Number of Times Voted in the Past 2 Midterm Primaries
* Total Number of Times Voted in the Past 2 Midterm Generals
* Total Number of Times Voted in the Past 2 Presidential Primaries
* Total Number of Times Voted in the Past 2 Presidential Generals

The 2010 primary election was NOT included in this count, as it remained its own variable. A key element to this decision once again revolved around creating a separation between Presidential cycles and mid-term cycles, with the reasoning remaining the same as was outlined above. Another important aspect to the decision regarded how many previous cycles to include in the derived variables. Given that there were three Presidential cycles available and only two mid-term cycles, the first consideration was whether or not this would place a greater weight on one versus the other. While intuition may lead one to answer yes to this question, the fact that both variables would exist in their own exclusive scales should make this a non-issue.

However, of greater concern was the fact that as the number of past cycles looked back upon increases, a bias is introduced into the dataset. If an individual were found to have voted in three of the previous three mid-term general elections for example, yet another individual was found to not have voted in any, that discrepancy could be due to the fact that the latter individual was not of voting age, and was therefore not eligible to vote in the previous three cycles. This would inherently give a higher weight to older individuals.

While this weight would be present in the dataset whether the variables were derived or not, it is exponentiated in these particular derivations with each previous cycle added. Thus, the decision was made to limit the number of previous cycles in the derived variables to two. A final, albeit brief, consideration was given to removing the voter history variables as input candidates because of their inherent age bias, but intuition held that they would be too important as potential predictors to eliminate altogether. The four derived variables listed above were numeric in nature.

**Consolidating Categorical Variables**

As was mentioned above, a number of categorical variables within the dataset possessed as many as six levels or more. The best practice for modeling these variables is to try at all costs to reduce the number of levels, as each level becomes its own dummy variable in some modeling techniques. Unpruned decision trees were trained on the target variable using single categorical inputs at a time in order to identify the purest splits within said input. For example, the variable “Party” produced four root nodes that consolidated the eight original levels of the variable in half. Levels were combined that possessed similar primary outcome proportions in the tree. A full chart detailing the consolidated levels of each variable can be found below. It should be noted that the variable “DMA” was consolidated using general analyst domain knowledge rather than a decision tree, but that it was the only such one. Furthermore, the variable “Income” was not consolidated.

Table

Description automatically generated

**Removing Variables**

Continuing with the process of reducing the candidate pool for potential inputs to the model, a handful of variables were able to be removed via logical reasoning. “2014 Primary Voter History”, “2012 General Election Precinct Turnout”, and “2012 General Election Precinct Turnout” were all removed given that the target variable occurred in a proceeding cycle. However, all three were added back into the score dataset and renamed.

**Renaming Variables**

Expounding on the final sentence in the previous section, some preservation of recency is still desired throughout the model. As a result, when the score dataset is used to predict the 2014 general election turnout, using the model trained on the 2010 general election turnout, some of the variables have to shift, and thus, must be recoded or renamed. For example, as was mentioned above, the “2010 Primary Turnout” was left alone in the trained model as it was used as its own input to illustrate the primary turnout for the year of the same target variable. However, in the scored final model, where the target variable shifts to the 2014 general election turnout, the role of the “2010 Primary Turnout” variable is instead served by the “2014 Primary Turnout” variable.

**Data Partitioning**

The dataset as a whole was split into two partitions—a training partition and a validation partition. The former contains 70 percent of the original data, while the latter contains the remaining 30 percent. Both samples were drawn at random. The training dataset, as implied, is used to train the model, while the validation model is used to predict and ensure that the model is not overfitted. Overfitted models will be subject to erosion when introduced to unseen data.

**Missing Values**

As was mentioned above, a number of variables in the dataset possessed missing values, even after logic based recoding. Critical assessment was necessary when approaching the missing values within the categorical variables. It should be noted that imputing missing values involves making up a value, which of course, results in unreal data points. Because of this, although it is permissible to impute values within numeric variables by using mean imputation, using mode imputation for categorical variable missing values was poor practice in this particular dataset for two reasons. First, the number of levels in each of the categorical variables was low (never greater than 6), and thus too much weight would be given to the majority class in a fashion that would significantly alter the distribution of the variable. Secondly, the sheer number of categorical variables in the final dataset was cause for concern as it pertains to using mode imputation. If there were one or two, the impact would not be felt as greatly. Ultimately, with these points in mind, and after consulting prior research on the topic, it was resolved that the only way to fully treat missing values would be to remove all the rows/observations inside of a categorical variable that possessed them.

However, this decision does not come without a potential cost. Consideration had to be put into what potential information loss would be incurred by removing a significant number of observations. Conversely, the same question applies to removing any of the variables in the dataset altogether that possessed missing values, which would be the alternative option. In order to decide on the trade off a sensitivity analysis was ultimately performed. While this will be detailed further in later sections, the crux of the analysis involved first using various variable selection techniques to assess whether an input would even be a candidate for the model in the first place. Following that, models were trained with and without certain variables, keeping in mind that if the variables were inside the model, the sample size was reduced due to the removal of observations containing missing values.

**Variable Selection**

As was previously mentioned, numerous variable selection techniques were employed as a means for beginning the sensitivity analysis. Beyond that, variable selection also had utility in that it served as the final step in paring down the original collection of variables in the dataset. This final step served as the supervised learning step in the process.

***Random Forest Models***

Random Forest models serve many purposes, but one of their most useful is actually variable selection. Three different Random Forest models were trained in this step. The models contained 100, 500, and 750 trees, respectively. This was done to reduce the inherent randomness of the models as they pertain to selecting inputs as candidates for node splitting. The randomness of the selection of inputs within a Forest model helps to make the variable selection process less greedy, and thus, less likely to overfit when the model is finally applied. The ranked results of the candidate splits are displayed below.

***LASSO***

Least Absolute Shrinkage and Selection Operator, or LASSO, works as a constrained form of the Ordinary Least Squares (OLS) process. In LASSO, the sum of the absolute values of the regression coefficients must be smaller than a predetermined value. In this particular case, that value was 1. LASSO integrates variable selection and model fitting all at once, thus providing a clear picture of variable importance. The ranked results of input importance from LASSO are displayed below.

***PLS***

Partial Least Squares Regression (PLS) is a flexible approach to modeling and variable selection because it can deal well with wide data, tall data, collinear data, and noisy data. The collinear aspect proved to be useful for this particular dataset, given that, intuitively, many of the categorical inputs were likely to be correlated. Examples of this include “Age” and “Net Worth” or “Net Worth” and “Home Owner or Renter”. PLS is very similar to Principal Component Analysis (PCA), but it combines multiple regression techniques of the principal components to ensure that the original variables are ultimately selected rather than just a linear combination of them as is the case in PCA. The variables chosen as inputs with the PLS technique are displayed below.

***Logistic Regression with Stepwise Selection***

The stepwise method of selecting input variables in logistic regression involves a blending of both forward and backward selection. As is implied, all coefficients begin equal to 0. The predictor most correlated with the target variable is then added to the model and the model steps in both directions until all relevant predictors are in the model. It should be noted that the level of significance for selection (alpha) was 0.05. The variables chosen as inputs with the stepwise selection method are displayed below.

***Chi-Square***

Variable selection using chi-square selection criteria was the final method employed. The variables chosen as inputs with the stepwise selection method are displayed below.

Table

Description automatically generated

The above table displays the ranked order in each of the variable selection techniques for the various inputs. The far right two columns indicate the average ranking for each variable and the number of time each was selected to pass on to the modeling phase. The variables highlighted were the ones selected in the final dataset for modeling. This will be detailed further in the ensuing sections.

**Sensitivity Analysis**

After completing the variable selection process and attaining the results displayed in the above chart, the next step in the process involved comparing the results of models both with and without the variables containing missing values. The variables that served as the primary focuses of the investigation were “Education” and “Net Worth”. The general rule of thumb regarding missing values within a variable is that if the count of missing values exceeds 30 percent, extra caution and investigation needs to be given. For this reason, “Ethnicity” was not included as part of the sensitivity analysis. The percentage of missing values in the “Occupation” variable was over 80%, which is too high for consideration inside of any model, and the variables “Income” and “Dwelling Type” were not significant enough in the variable importance rankings (as shown in the chart above) to merit further investigation.

To conduct the analysis, a logistic regression model, with ALL of the above highlighted variables included as inputs was trained on the target variable. This model included both “Net Worth” and “Education”. In order to do so while also complying with regression model rules, the missing values in either variable were removed from the sample. Another logistic regression model was then trained using all of the above highlighted variables and NOT including “Education” nor “Net Worth”. This allowed for a much fuller sample.

The comparison ultimately boiled down to whether more predictive power would be lost by removing observations or by removing inputs. The results were fairly drastic. The model that removed the variables yielded a misclassification rate of 16.35% while the model that kept the variables but removed observations yielded a misclassification rate of 18.58%. Thus, the decision was ultimately made to remove the variables.

It should be noted that the only variable containing any missing values that remained in the final model was “Ethnicity”. It contained a mere 10.38% missing value count, and was shown to be a strong candidate across the various variable selection techniques. As a result, two final datasets were produced—one with no missing data that contained all 50,000 observations, and one with missing data that contained 44,808 observations.

**Variable Removal**

With all of the above information in tow, the final dataset was absent the following variables:

* Education
* Occupation
* Net Worth
* Marital Status
* Pet Owner
* Dwelling Type
* DMA
* Income
* NASCAR Interest
* Musical Instrument Interest
* Donates to Conservative Causes
* Donates to Liberal Causes
* VH12g
* VH12p
* VH08g
* VH08p
* VH06g
* VH06p
* VH04g
* VH04p
* VH02g
* VH02p
* VH00g
* VH00p
* Most Recent Primary Precinct Turnout

**Modeling**

**Random Forest Champion Model**

After performing all of the necessary data preparation steps, multiple model types and techniques were employed in order to identify a champion model. The models were compared using overall misclassification rate as the criteria since the cost of incurring false positives or negatives in the context of this particular problem is not high. The types of models trained included Decision Trees, Logistic Regressions, and Gradient Boosted Random Forest Models. Each of these techniques utilized different parameters within as well. Both the Decision Tree models as well as the Random Forest models were training using the dataset containing missing values given their robust nature and ability to handle missing values without model deconstruction. However, the logistic regression models were trained using the reduced sample dataset of 44,808 observations.

The final champion model was a Random Forest Model trained using 500 trees. 60 percent of the observations in the training data were used for each tree in the forest. The number of input variables considered for splitting for each node of the Random Forest was the square root of the total number of inputs, which in this particular case, was rounded up to four.

The misclassification rate yielded from the champion model was 15.44%. When compared to the Random Forest model trained using 700 trees, the misclassification rate was reduced by only .09%. Given the potential for computational costs incurred by having models with larger parameters, the 500 tree model was chosen instead. As was mentioned above, the champion logistic regression model yielded a misclassification rate of 16.35%. The best champion decision tree model yielded a misclassification rate well over 20%.

The variable importance rankings from the champion Random Forest model are shown in the images below. The rankings reflect the decrease in accuracy and the decrease in Gini when each variable is systematically left out of the model.

Chart

Description automatically generated

**Model Application to Unseen Data**

Earlier it was mentioned that some variables needed replaced and/or recoded in order to comply with the symmetrical structure of the train and score datasets. After this process was complete, the champion Random Forest model trained using the 2010 general election turnout as the target variable, was applied to the score dataset to yield predictions as to whether or not each individual in the dataset would be expected to vote. The anticipated turnout on the 44,808 individuals in the dataset was 11,467, or 25.59%.

**Potential Fixes, Adjustments, Concerns**

To begin, the champion model yielded a misclassification rate that was likely too high to render the model useful. While this is a less than ideal outcome, the practices employed throughout the process were sound, and more importantly, the reasonings behind them were rooted in thoughtfulness, domain knowledge, and logic. However, that is not to suggest that the model was perfect, because obviously, it was not.

Two core issues come to mind when investigating the cause of the lower accuracy rate. First, as was mentioned above, there is the possibility that the input variables were too highly correlated with one another. While this is certainly true, the issues caused by this problem would have more to do with the long-term integrity of the model than with the accuracy of its predictions. The second issue that again comes to mind involves the inherent bias introduced by deriving variables that could possess higher or lower weights that are too dependent upon age—a variable that penalizes potential younger voters. Again, this issue is certainly possible, however as was iterated above, this bias would be present on some scale whether the derived variables were present or not. “Age” could be removed if it were of greater concern. However, the variable importance chart and selection table above both indicating it is too important a driving factor in predicting turnout to be ignored entirely.