**IE5202 Project 1**

Dong Zheng

A0119545B

**Step 1: Primitive Regression Model**

In this step we are only allowed to include 5 predictors and their power transformations into our model.

**1.1 Methodology**

The method for building Primitive Regression Model can be broken down into three sub-steps: The first is to identify which 5 predictors can explain most of the response value ‘total votes change percentage’. Then we need to analyze each predictor and its residuals to identify the area of improvement. Lastly, we try different power transformations on original variables, to reduce the root mean residual sum of squares and improve the model prediction power.

To find the top 5 predictors, we standardize all 52 quantitative predictors and rank them by their feature importance. The result (in Appendix 1) shows that the top 5 important predictors are 'RHI225214', 'RHI125214', 'RHI325214', 'PST045214' and 'POP010210'. We should not use all of them in our primitive model, because they could be highly correlated. Here we only pick the most important feature 'RHI225214'. The categorical covariate ‘state\_abbr’ in the first column should also be included, then we have three more predictors to choose. The best way is to enumerate all possible 3 predictors combinations and compare their AIC. Thus, we enumerated and compared 20825 models of 5 predictors. The ‘getAll’ function result exhibits that (RHI225214 , state\_abbr, RHI325214, LFE305213, HSG495213) are the best 5 predictors in terms of AIC value, and (RHI225214 , state\_abbr, POP645213, LFE305213, HSG495213) are the second best model predictors.

**1.2 Result Interpretations and Main Findings**

Now let us build up our primitive regression model using formula: **Total\_votes\_change\_percentage ~ RHI225214 + C(State\_abbr) + RHI325214 + LFE305213 + HSG495213 + 1**

Then plug it into statsmodels.formula.api and get OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | total\_votes\_change\_percentage | **R-squared:** | 0.385 |
| **Model:** | OLS | **Adj. R-squared:** | 0.372 |
| **Method:** | Least Squares | **F-statistic:** | 29.43 |
| **Date:** | Tue, 21 Sep 2021 | **Prob (F-statistic):** | 4.30e-217 |
| **Time:** | 14:15:57 | **Log-Likelihood:** | 3388.1 |
| **No. Observations:** | 2500 | **AIC:** | -6670. |
| **Df Residuals:** | 2447 | **BIC:** | -6361. |
| **Df Model:** | 52 |  |  |
| **Covariance Type:** | nonrobust |  |  |

The Adj. R-squared value (0.372) is not very high, suggesting that there is room for improvement. F-statistic (29.43) determines that at least one of the covariate’s coefficient is not zero. Look into P-value list, RHI225214, RHI325214, LFE305213, HSG495213’s P-values are all very small, indicates that they are all statistically significant variables. Categorical variable ‘State\_abbr’has different P-values for different states. States with small P-values indicates that these states have significant impact on ‘total votes change percentage’. While in other states with large P-values, they are not significantly different from the average response value.

Let us then look at what are these 5 variables mean and what insight we can get from them.

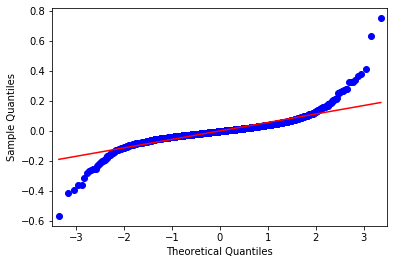
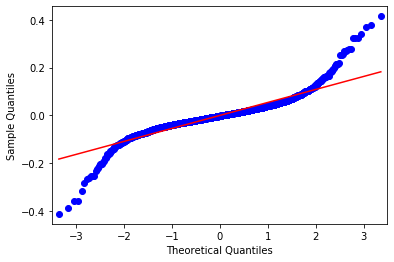
*Table 1.1: Top 5 Important Predictors*

|  |  |
| --- | --- |
| Predictors | Description |
| RHI225214 | Black or African American alone, percent, 2014 |
| RHI325214 | American Indian and Alaska Native alone, percent, 2014 |
| HSG495213 | Median value of owner-occupied housing units, 2009-2013 |
| LFE305213 | Mean travel time to work (minutes), workers age 16+, 2009-2013 |

RHI225214 and RHI325214 are two variables that related to ethnic composition of a state. Since in 2012 most Black and American Indian people votes for Barack Obama, while in 2016 they are less likely to vote for Donald Trump, it is reasonable that RHI225214 and RHI225214 are two of the most relevant predictors that are included in our primitive model.

LFE305213, HSG495213 are two variables that relate to income and living standard in a state. LFE305213 implies whether people in the state own a car and also indicates that whether they are living in city or rural area. HSG495213 indicates people’s income level (whether they are poor or middle class). Income level could affect whether people vote for Donald Trump. Thus there is no surprise that these two variables are included in the model.

**4. Model Diagnostics**

In this section we diagnose our primitive regression model and find ways to improve it. The three assumptions made for error terms in linear regression are constant variance, independence of variables and normality of the distribution. These assumptions need to be checked and see if our model violates any of the them. First, we plot the normal qq-plot to check the normality of the data.

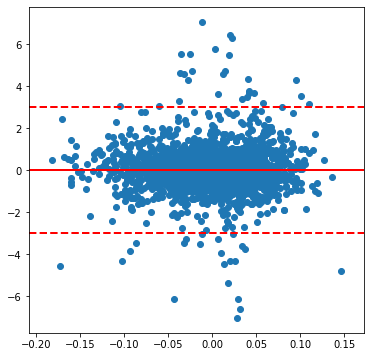
*Figure 1.1 Normal QQ-plot before and after removing 3 Outliers*

From the plot we can see that within the middle part, our data is quite normally distributed. However, at two extremes our model data deviate and do not follow normal distribution. we can easily identify two outliers at top right corner and one at lower left corner. After checking original data set, we found that these outliers are all from Texas. In the future model, we can create a new dummy variable called ‘isTexasOutlier’ to consider these special data, or probably remove them from our training set because there could be leverage and influence points.

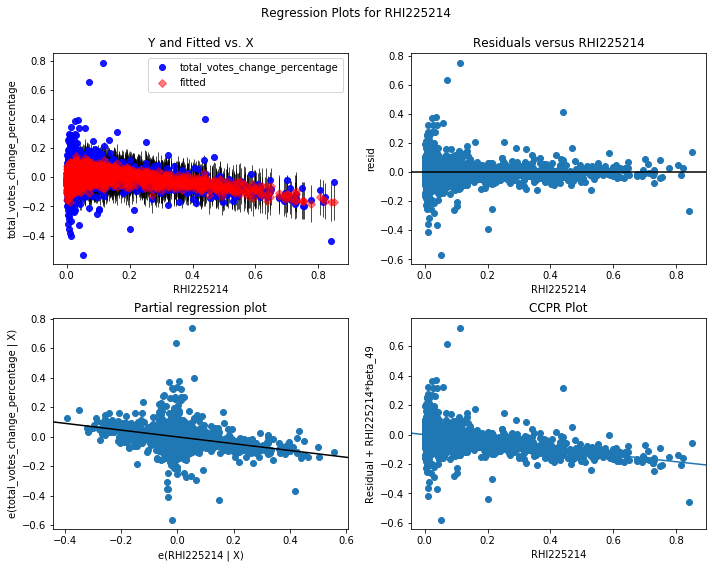
*Table 4.1: Extreme Outlier ID*

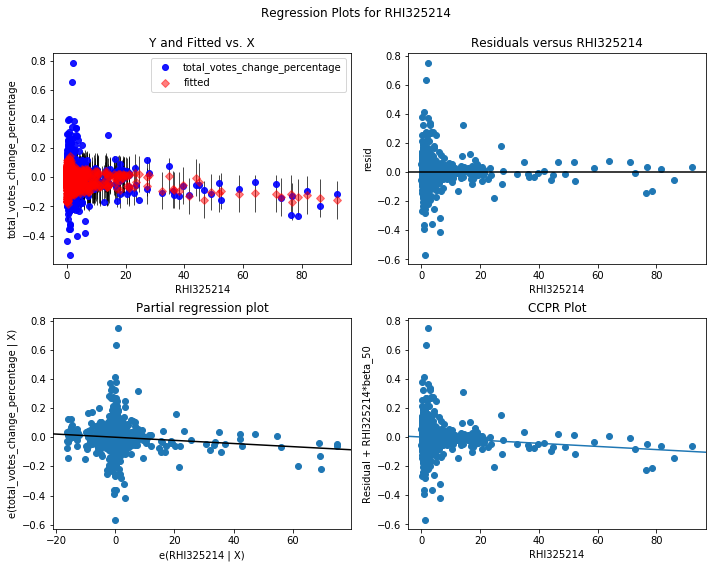
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Row ID | State\_abbr | |  | | --- | | Total\_votes\_change\_percentage | | Dem\_change\_percentage |
| 578 | TX | 0.78785468 | -0.045181407 |
| 2274 | TX | 0.65608987 | 0.005040385 |
| 785 | TX | -0.53045616 | -0.042610196 |

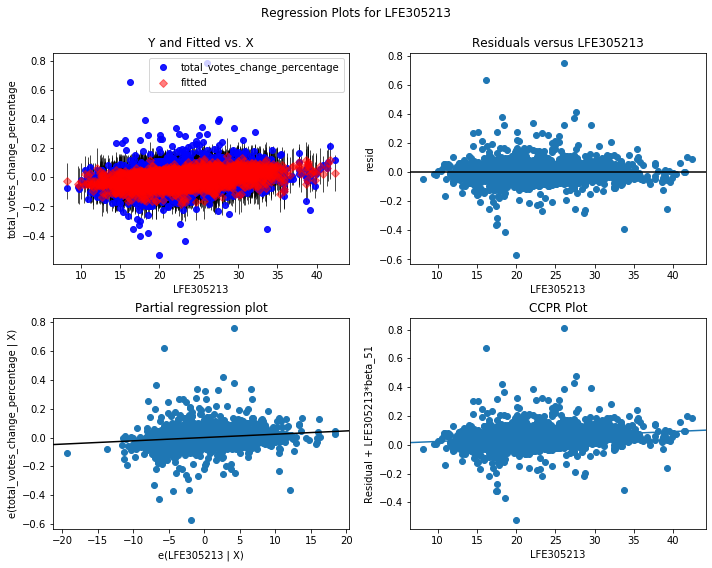
Secondly, the residuals against fitted values plot is displayed as below to check the constant variance assumption. We can see that despite there are a few outliers in the middle part, the overall residuals variance are quite constant along the x-axis.



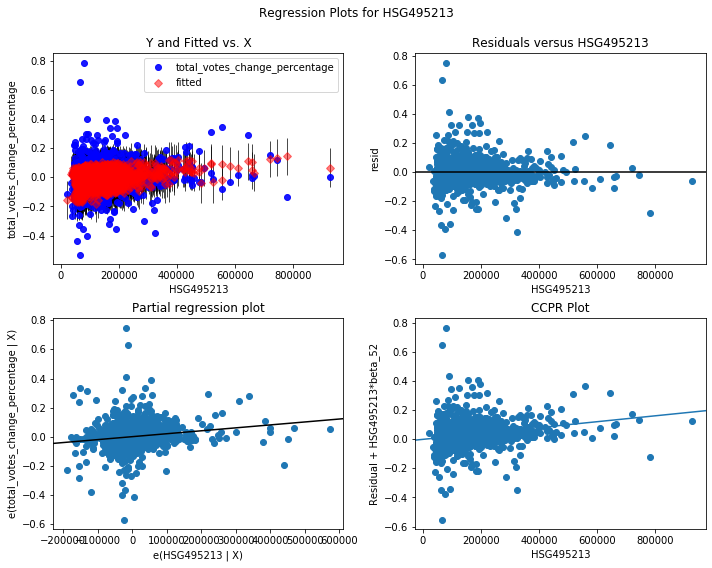
*Figure 1.2 residuals against fitted values*

Lastly we also need to check model residuals against each predictor to see if they are independent of each variable. Regression plots for RHI225214, RHI325214, LFE305213, HSG495213 are listed below in sequence:





From RHI225214 and RHI325214 Regression plots, we can see that there is a decreasing trend of variance along x-axis, and there is a little non-linearity in the RHI225214 scatter plot. Thus, we may consider applying a log transformation for RHI225214 and RHI225214 data to see if that can make residual terms more constant.



In terms of the LFE305213, the residual versus LFE305213 plots shows that variance are generally constant. Therefore, we do not need to apply any transformation to LFE305213 data. Lastly, residuals versus HSG495213 plot shows that variances are not so normally distributed and we may consider applying Box-Cox transformation to the data.

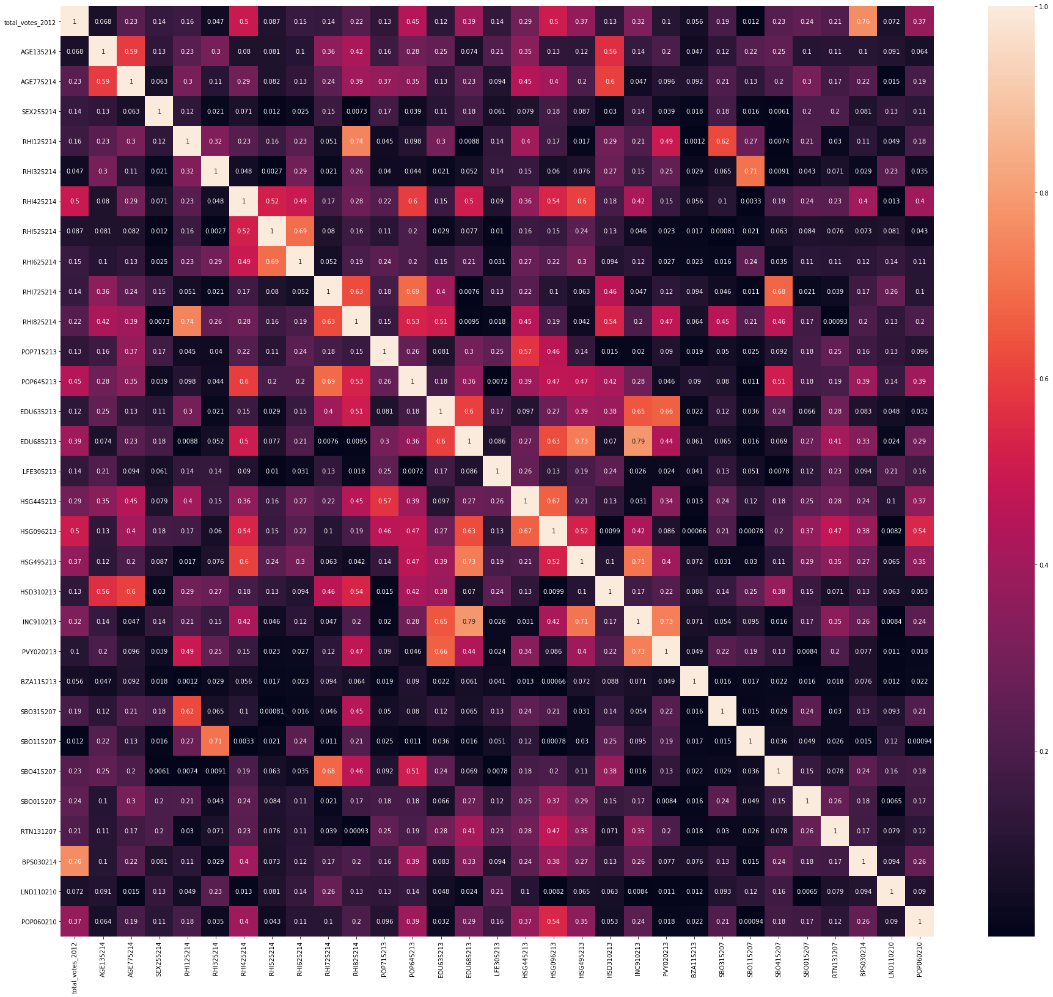
In conclusion, assumptions made for linear regression model are generally satisfied, except that some extreme data points may not follow normal distribution. The partial regression plots also imply that some systematic variance in our preliminary regression model could be reduced by including higher-order terms, such as interactions and transformations. We will explore that in our Step 2 Advanced Regression Model.

**Step 2 Advanced Regression Model**

In step two, we can include interaction and any new variables in order to build the best regression model for predicting the responses.

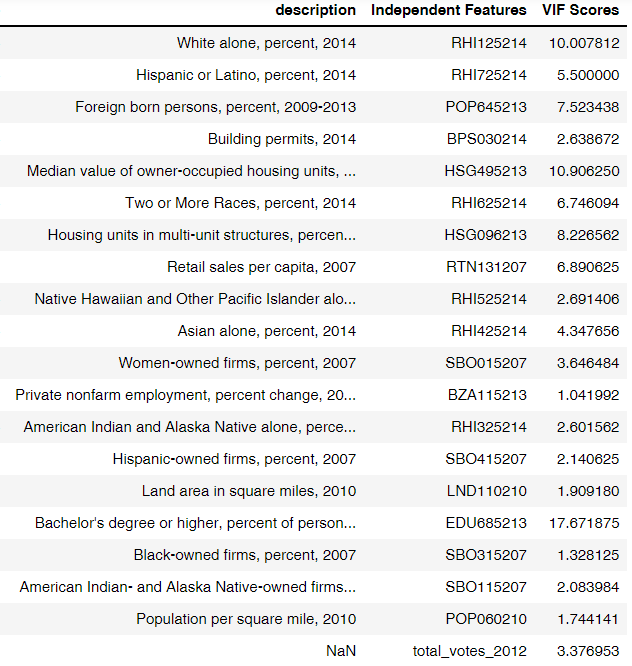
**2.1 Methodology**

From the model diagnostic part in step 1, we observed that there are some non-linearly in the data. Therefore, including interaction terms into our advanced model would be a good improvement. Since we have 53 predictors, which means that there will be more than thousands of two-factor interactions for us to consider. Thousands of interactions are a little bit redundant for the forward selection algorithm and it will take very long computation time.

In order to efficiently find the best interaction for the model, we need to reduce the number of predictors considered for interaction effects. The correlation matrix of 52 variables exhibits there is some collinearity in our data set. As you can observe from the correlation heatmap (Appendix 2), white color girds indicate highly correlated covariate pairs. To reduce highly correlated predictors in the model, we use SelectNonCollinear function in ‘Collinearity’ package to remove them. The resulted correlation heat map displays that there are no more highly correlated covariates in the data set.

*Figure 2.1* Correlation Heat Map

There are around 30 predictors remained which is still too much for interaction terms. Then sorting the remaining predictors by their VIF (Variance Inflation Factor) score and filtering out predictors, that have a VIF Score higher than 10, will give us final 20 predictors list as below:

*Table 2.1 List of Predictors to be Considered for Interaction Effects*

Lastly, we use Forward Selection Algorithm to find out interaction terms that will benefit our regression model. By hierarchy principle, we should also include their first-order terms into our regression model.

**2.2 Justifications and Findings**

By including second-order interaction terms and their main effect, our advance-regression model for predicting ‘total\_votes\_change\_percentage’ results are shown in Table 2.2 below. The Adj. R-squared increase from 0.372 to 0.536 and the AIC value decrease from -6670 to -7400. If it is also true in the test data set, our advanced model’s prediction power has increased around 50% when comparing with our primitive regression model.

*Table 2.2 OLS Regression Results for Advanced Regression Model – R1*

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | total\_votes\_change\_percentage | **R-squared:** | 0.593 |
| **Model:** | OLS | **Adj. R-squared:** | 0.536 |
| **Method:** | Least Squares | **F-statistic:** | 10.41 |
| **Date:** | Thu, 23 Sep 2021 | **Prob (F-statistic):** | 7.23e-263 |
| **Time:** | 02:10:23 | **Log-Likelihood:** | 4007.1 |
| **No. Observations:** | 2497 | **AIC:** | -7400. |
| **Df Residuals:** | 2190 | **BIC:** | -5613. |
| **Df Model:** | 306 |  |  |
| **Covariance Type:** | nonrobust |  |  |

Our advanced regression model for predicting ‘dem\_change\_percentage ‘ results are shown in Table 2.3 below. The Adj. R-squared value is 0.849 indicates that the statistical patterns in ‘dem\_change\_percentage’ are more obvious and it could be easier for us to predict ‘dem\_change\_percentage’ than ‘total\_votes\_change\_percentage’.

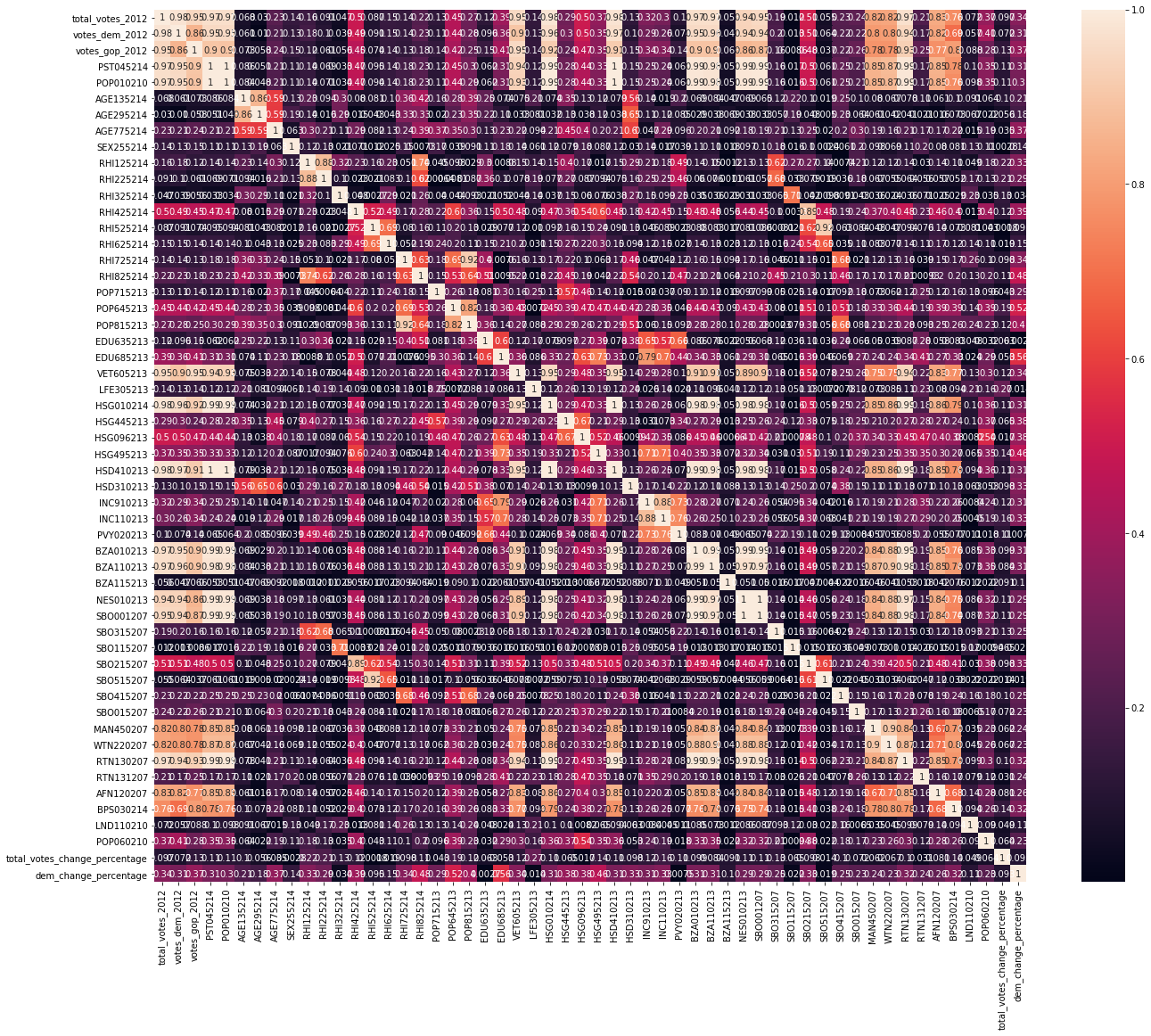
*Table 2.3 OLS Regression Results for Advanced Regression Model – R2*

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | dem\_change\_percentage | **R-squared:** | 0.878 |
| **Model:** | OLS | **Adj. R-squared:** | 0.849 |
| **Method:** | Least Squares | **F-statistic:** | 30.80 |
| **Date:** | Sun, 26 Sep 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:19:35 | **Log-Likelihood:** | 6450.7 |
| **No. Observations:** | 2500 | **AIC:** | -1.196e+04 |
| **Df Residuals:** | 2027 | **BIC:** | -9201. |
| **Df Model:** | 472 |  |  |
| **Covariance Type:** | nonrobust |  |  |

**Appendix**

Append 1. List of Most Important Features by Descending Order

|  |  |  |
| --- | --- | --- |
| ['RHI225214', | 'total\_votes\_2012', | 'AFN120207', |
| 'RHI125214', | 'BPS030214', | 'BZA115213', |
| 'RHI325214', | 'INC110213', | 'SBO015207', |
| 'PST045214', | 'POP645213', | 'POP815213', |
| 'POP010210', | 'WTN220207', | 'SBO215207', |
| 'HSD410213', | 'LFE305213', | 'POP715213', |
| 'HSG010214', | 'HSG495213', | 'PVY020213', |
| 'RHI425214', | 'RHI525214', | 'POP060210', |
| 'BZA010213', | 'INC910213', | 'HSD310213', |
| 'RHI725214', | 'EDU685213', | 'SBO315207', |
| 'SBO001207', | 'AGE135214', | 'AGE775214', |
| 'RHI825214', | 'MAN450207', | 'AGE295214', |
| 'RTN130207', | 'EDU635213', | 'SEX255214', |
| 'RHI625214', | 'LND110210', | 'SBO115207', |
| 'votes\_dem\_2012', | 'RTN131207', | 'HSG445213', |
| 'BZA110213', | 'NES010213', | 'HSG096213', |
| 'VET605213', | 'SBO415207', | 'SBO515207'] |

Append 2. Correlation Heatmap for All 52 Predictors in the data set.