Results:

We used stratified 5-fold cross validation setup and random shuffling to perform our data splits and tested. With 3 different iterations and complete hyperparameter tuning on our random forest regressor model, we evaluated the model performance. our goal is to come up with the model to predict the candidate votes so choose ‘Pearson correlation coefficient (r)’ and ‘Mean square error (MSE)’ as our evaluation metrics of our model. As Pearson coefficient gives us the measure of the linear correlation between the variables and MSE gives us the value of averaged squared difference between the predicted values of our model. Hence, we optimized our model with high value of Pearson correlation coefficient and low value of mean squared error. The predicted ‘r’ and ‘MSE’ of our random forest regression model with different iterations is presented on table (1):

Table 1)

|  |  |  |
| --- | --- | --- |
| No. of iteration | Pearson Correlation coefficient (r) | Mean squared error (MSE) |
| Iteration 1 | 0.5493 | 0.0355 |
| Iteration 2 | 0.796 | 0.0204 |
| Iteration 3 | 0.709 | 0.019 |

From table 1), we can clearly see the prediction result gets better with no of iterations. On iteration 1) with our top identified features, we divide our data into 50/50 train/test sets which resulted the value of MSE= 0.0355 and ‘r’ = 0.5493.

On iteration 2) with added new features, **party affiliation** and **no of opponents and** splitting the data into CV 5 folds our model performs better by decreasing the MSE to 0.0204 and increasing the value of ‘r’ to 0.796.

On iteration 3) with Wikipedia NLP feature extraction and adding more features for candidates with Wikipedia page, our Random forest model was optimized on CV splits=5 and no of estimator=150.with this feature set and optimized hyperparameters tuning our model accuracy increases by decreasing the MSE to 0.019.

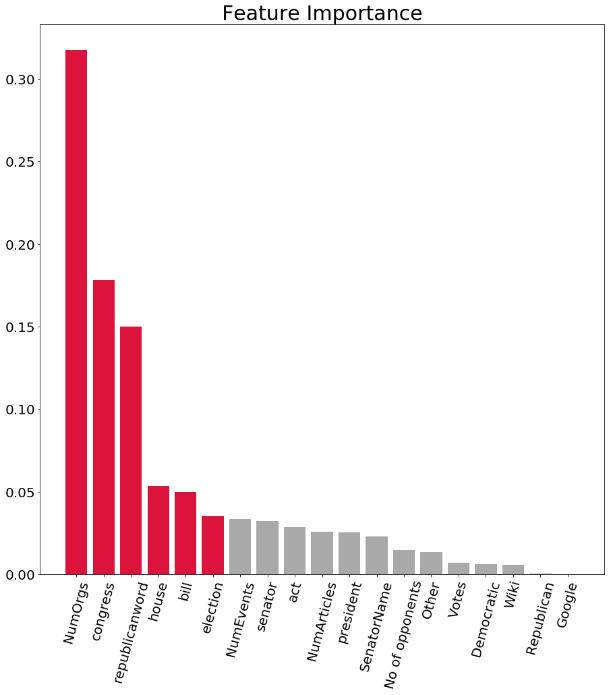
We can also conclude that our final model is a whole combination of 2 feature sets, one with politician with wiki page plus NLP and another without wikipage plus no NLP. RF model predictions between both models clearly indicates that having Wikipedia page with our top features listed in table 2) increases the chances of candidates to win.

Table 2) Top features Sets for candidates

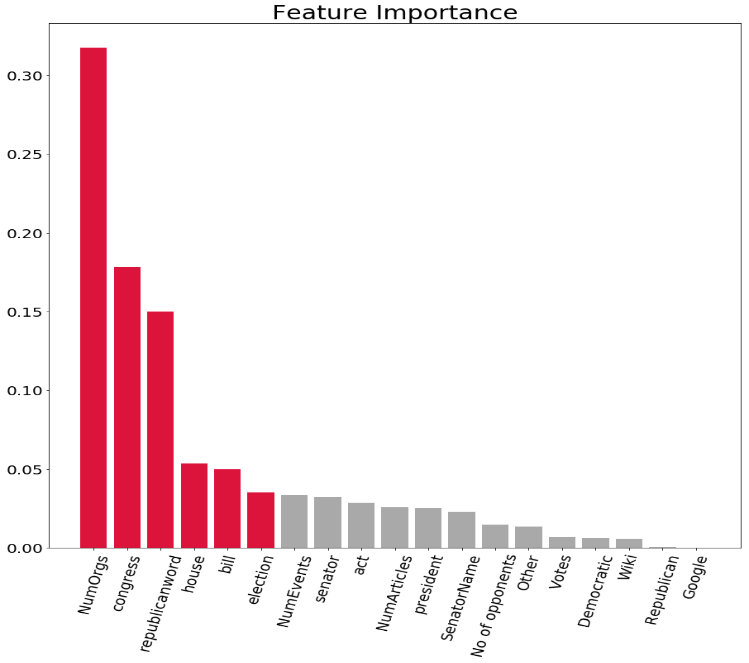
|  |  |
| --- | --- |
| Congress | president |
| Republican word | bill |
| Election | act |
| House | committee |
| Senator | senate |

Additional features for candidates with Wikipedia

|  |  |
| --- | --- |
| Google | No of Articles |
| Wiki | No of Opponents |
| No of events | Party |
| No of Orgs |  |

 Features for all candidates

we visualized the top features and their weights in feature importance chart as shown in fig 1). we can see top 6 features are: number of organizations the candidates are associated with, congress, republican word, house, bill and election. Our largest weight was just above 0.30 with features number of organizations the candidate is associated with.

 Fig 1) Feature Importance of top features