Machine Learning Engineer Nanodegree

Capstone Final Report

Michael Basca March 5th, 2017

I. Definition

Project Overview

The practice of predictive modeling defines the process of developing a model in a way that we can understand and quantify the model's prediction accuracy on future, yet-to-be-seen data. In short, predictive modeling is the process of developing a mathematical tool or model that generates an accurate prediction (Kuhn and Johnson 2013). Prediction models are used in a wide range of industries such as finance, healthcare, public policy and others.

Polling is an integral part of our political system. A poll is a survey or questionnaire used on a random sample of people in hopes that the random sample is representive of a large population. In assuming that the sample is representative, polling minimizes the resources required to gather data by only using a fraction of the population. Polling can be asset in the realm of politics because they gauge the views of the public on the current political issues. Politicians can use polling information along with predictive modeling to cater their message to the constituencies that they are trying to influence, in order to get elected or aid their decision making in governance.

The history of using machine learning in polling is limited. Some of the literature that was researched have modeled political affiliation based on text sentiment analysis of tweets or political speeches [1, 2]. The text can be processed into a bag of words vector that closely resembles binary features for this project. A variey of supervised linear and non-linear learning algorithms were used (Support Vector Machines, Logistic Regression, Linear Discriminant Analysis and Naïve Bayes) as well as unsupervised methods as well(Principle Component Analysis, K-Means Clustering). Another experiment details the use of Decision Trees to predict party affiliation of Congress members by analyzing how each member voted on particular legistlation [3, 4]. These yes or no votes also reflects the structure of the binary data that is used for this capstone.

Problem Statement

In this project I will be participating in a **Kaggle**[5] project that uses data from a mobie application called ***Show of Hands***[6]. In this competition ***Show of Hands***[TM] provides a questionnaire to thousands of data in a form of approximately one hundred **Yes** or **No** questions along with their political affiliation (**Democrat** or **Republican**). I will use this data to create a model that predicts political affiliation based unlabeled test data.

The problem is to achieve is to predict whether the person answering the questionnaire is either a **Democrat** or **Republican**. Kaggle has the actual test labels to compare your predictions with and will give you a score based on accuracy. The goal is to achieve an accuracy that acheives a score at 75 percentile (or more) of all participants.

This problem is a classification task. While there are many learning algorithms (linear and nonlinear) that can produce a classification model, the focus of this assignment will be the preprocessing of data that will enhance the signal while reducing the noise to aid in proper generalization. I will perform vizualizations of the data that will guide which processing will be required.

The preprocessing will include:

- Transformations
  - MinMax scaling
  - Box Cox Transforms
- Feature Extraction and Dimensionality Reduction
  - Tree Feature Importance
  - L1 Regularization
- Dealing with missing Values
  - Deletions of Samples via threshold
  - Imputation
- Create meta-features
  - additive/multiplicative features

There are a variety classification learning models that we can choose from:

- Linear Models
  - Logistic Regression
  - Linear Discriminant Analysis
  - Partial Least Squares Linear Discriminant Analysis
- Non-linear Models
  - Support Vector Machines

- Tree Models
  - Decision Trees
  - Random Forests
  - Boosting

Each model has their strengths and weaknesses. In general I will explore common models with the data minimal processing to provide a baseline score. I will then perform extra processing (feature reduction as well as feature creation) to see whether the score can be improved.

Metrics

While classification models can be evaluated in a variety of metrics (AUC, sensitivity, specificity, kappa, etc.), Kaggle will evaluate our predictions based on accuracy. We may use the previous metrics a guide to provide better accuracy. Accuracy is defined as:

$$\frac{TP + TN}{N}$$

Where:

- **TN** are Number of True Negatives
- **TP** are Number of True Positives
- **N** are Number of Samples

II. Analysis

Data Exploration

A data set of 6960 samples were gathered from ***Show of hands***. The data included the following features:

1. Date of birth - an Interval variable
2. Gender - a Nominal/Binary variable
3. Income Bracket - an Ordinal Variable
4. HouseHold Status - a Nominal Variable
5. Educational Level - an Ordinal variable
6. One hundred and one **Yes** or **No** questions - a Nominal/Binary Variable

The outcome is: 1. Party affiliation **Democrat** or **Republican** - a Nominal/Binary Variable

Kaggle has provided csv files for the training set with labels as well as the testing set (80/20 split) without labels to test your model against.

Here is a portion of the data frame:

	YOB	Income	EducationLevel	HouseholdStatus	Gender	Interact.with.someone.dislike.daily	Parents.Fight.i
USER_ID							
1	1938.0	NaN	NaN	Married (w/kids)	Male	No	NaN
4	1970.0	over $150,000	Bachelor's Degree	Domestic Partners (w/kids)	Female	NaN	Yes
5	1997.0	$75,000 - $100,000	High School Diploma	Single (no kids)	Male	NaN	Yes
8	1983.0	$100,001 - $150,000	Bachelor's Degree	Married (w/kids)	Male	No	Yes
9	1984.0	$50,000 - $74,999	High School Diploma	Married (w/kids)	Female	No	Yes

It was decided to remove samples where participants were born before 1933 (some samples were stated that their YOB was 1900 which indicated that the sample's data validity was questionable) as well as participants born after 2000 where the person might be too young to comply with some of the questions or answer them seriously. The result was a loss of 7% of the samples that were considered outliers.

Here are some example statistics for some selected feature columns after the removal of these outliers:

	YOB
mean	1979.63
std	14.95
min	1935
25%	1970
50%	1983
75%	1993
25%	1970
max	1999

Income	count
Income	count
under $25,000	729
$25,001 - $50,000	692
$50,000 - $74,999	805
$75,000 - $100,000	714
$100,001 - $150,000	744
over $150,000	701

EducationLevel	count
Current K-12	720
High School Diploma	662
Current Undergraduate	745
Associate's Degree	366
Bachelor's Degree	1162
Master's Degree	6150
Doctoral Degree	183

Gender	Count
Male	3112
Female	1984

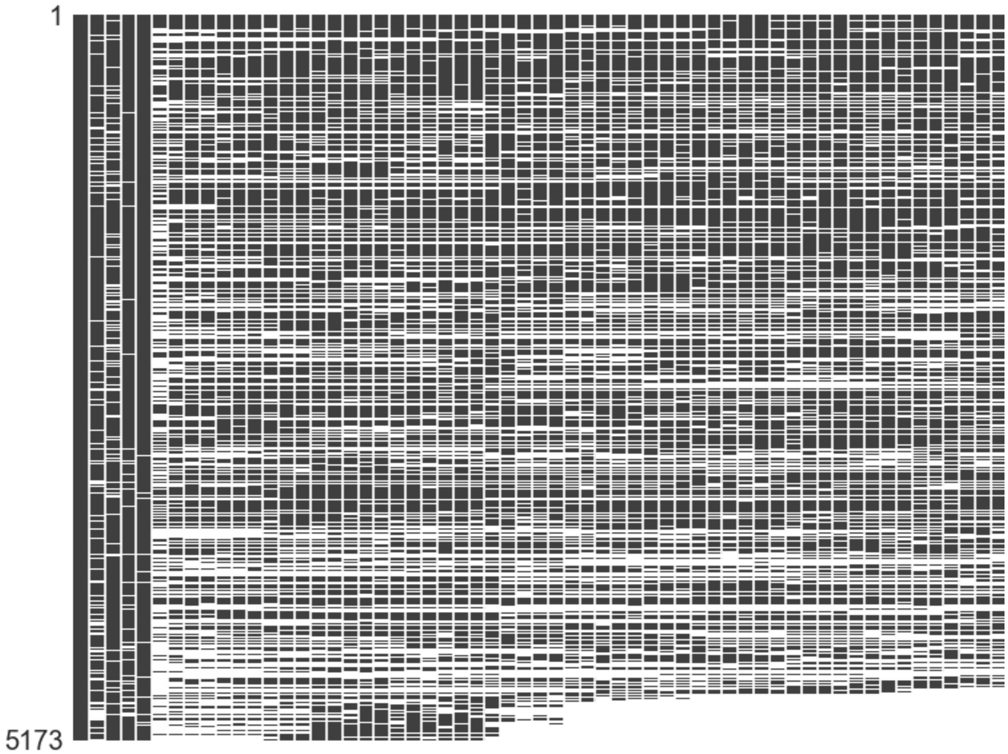
Work.Min.Wage	count
Yes	196
No	2906

The target breakdown is the following:

Party	count
Republican	2423
Democrat	2750

Which indicates relatively balanced classes.

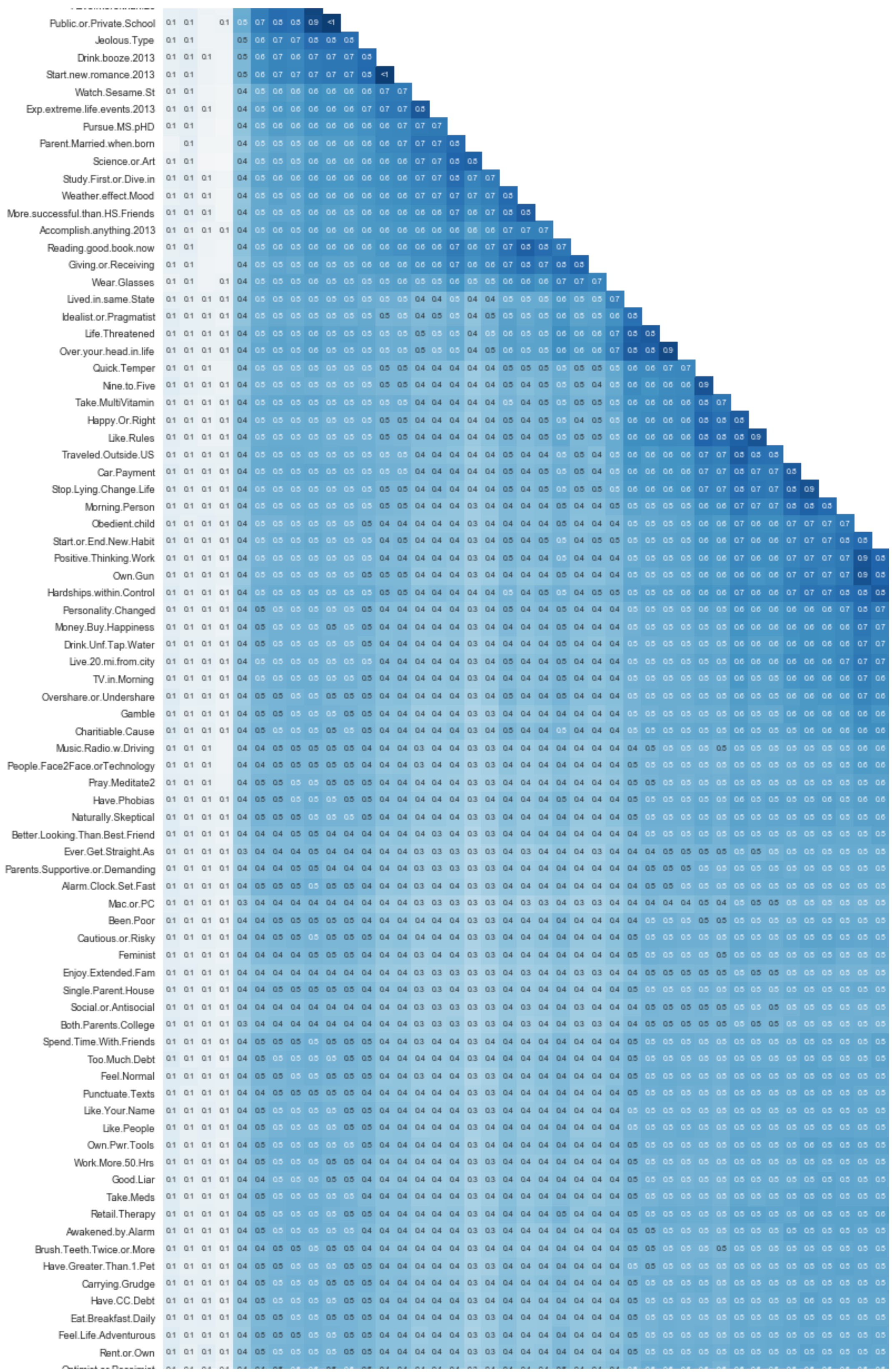
The data is somewhat sparse. Let's look at this graphically:

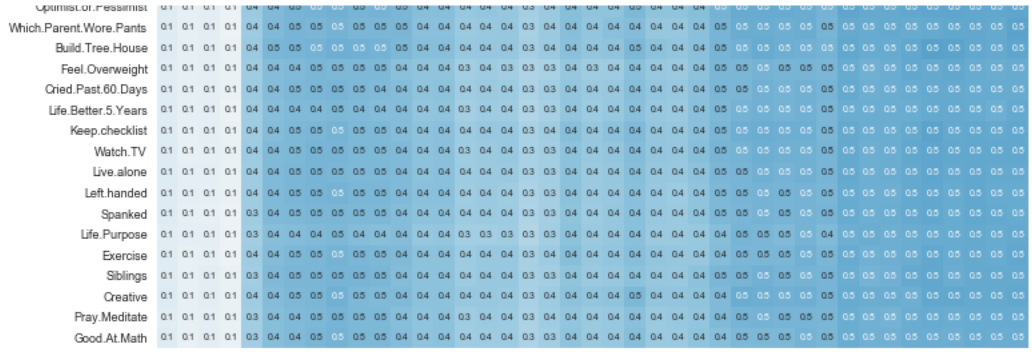


We will devise a method of deleting more samples that have a minimum sparsity threshold with imputation on the rest of the samples.

Exploratory Visualization

Let's take a look at the correlation plot.


Public.or.Private.School
Jeolous.Type
Drink.booze.2013
Start.new.romance.2013
Watch.Sesame.St
Exp.extreme.life.events.2013
Pursue.MS.pHD
Parent.Married.when.born
Science.or.Art
Study.First.or.Dive.in
Weather.effect.Mood
More.successful.than.HS.Friends
Accomplish.anything.2013
Reading.good.book.now
Giving.or.Receiving
Wear.Glasses
Lived.in.same.State
Idealist.or.Pragmatist
Life.Threatened
Over.your.head.in.life
Quick.Temper
Nine.to.Five
Take.MultiVitamin
Happy.Or.Right
Like.Rules
Traveled.Outside.US
Car.Payment
Stop.Lying.Change.Life
Morning.Person
Obedient.child
Start.or.End.New.Habit
Positive.Thinking.Work
Own.Gun
Hardships.within.Control
Personality.Changed
Money.Buy.Happiness
Drink.Unf.Tap.Water
Live.20.mi.from.city
TV.in.Morning
Overshare.or.Undershare
Gamble
Charitiable.Cause
Music.Radio.w.Driving
People.Face2Face.orTechnology
Pray.Meditate2
Have.Phobias
Naturally.Skeptical
Better.Looking.Than.Best.Friend
Ever.Get.Straight.As
Parents.Supportive.or.Demanding
Alarm.Clock.Set.Fast
Mac.or.PC
Been.Poor
Cautious.or.Risky
Feminist
Enjoy.Extended.Fam
Single.Parent.House
Social.or.Antisocial
Both.Parents.College
Spend.Time.With.Friends
Too.Much.Debt
Feel.Normal
Punctuate.Texts
Like.Your.Name
Like.People
Own.Pwr.Tools
Work.More.50.Hrs
Good.Liar
Take.Meds
Retail.Therapy
Awakened.by.Alarm
Brush.Teeth.Twice.or.More
Have.Greater.Than.1.Pet
Carrying.Grudge
Have.CC.Debt
Eat.Breakfast.Daily
Feel.Life.Adventurous
Rent.or.Own



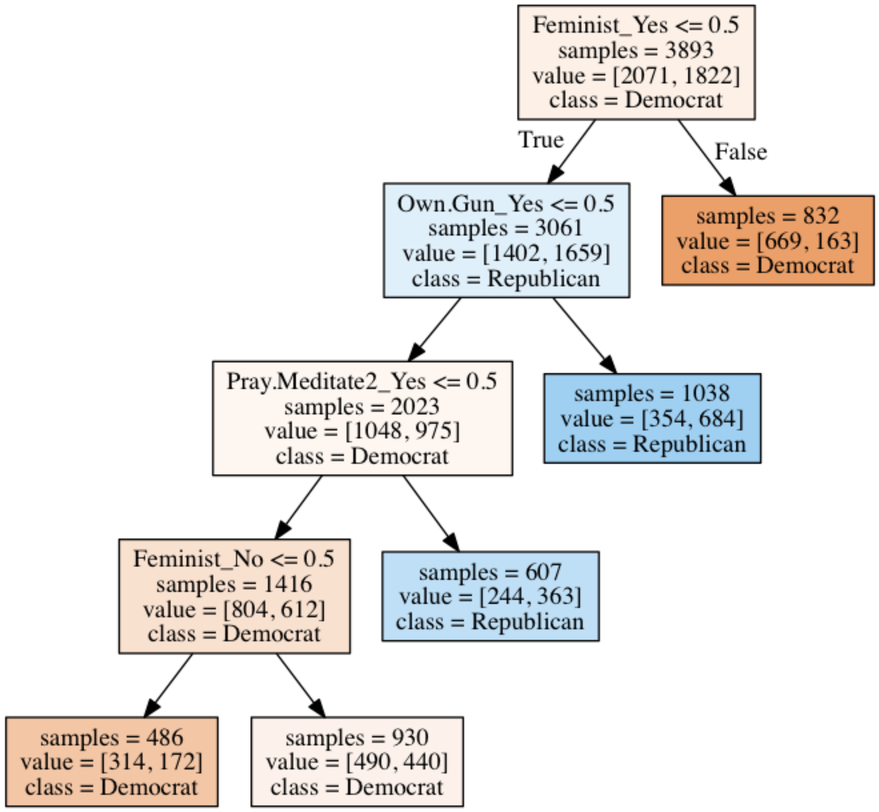
Using a threshold of +/- 0.7 we see that there are very few features that highly correlated.

Let's take a look univariate comparison between our categorical vs. our target. We will take the odds ratio for every binary categorical and list the highest or lowest (the lowest will be inverted in the table i.e. 1/score) odds ratios. These features are considered strong indicators of what the target value will be.

The target breakdown is the following:

Question/Answer Options	Odds Ratio	Party
Are you a Feminist?	5.43	Democrat
Own Gun?	2.37	Democrat
Pray/Meditate?	2.19	Republican
Pray/Meditate2?	2.11	Republican
Does life have a purpose?	1.90	Republican
Did your parents spank you?	1.61	Republican
Which parent wore the pants (Mom/Dad)?	1.55	Democrat

We can also create a Decision tree to have an idea of which features are important.



The tree reconfirms that features from the odds ratio analysis correspond to strong predictive features.

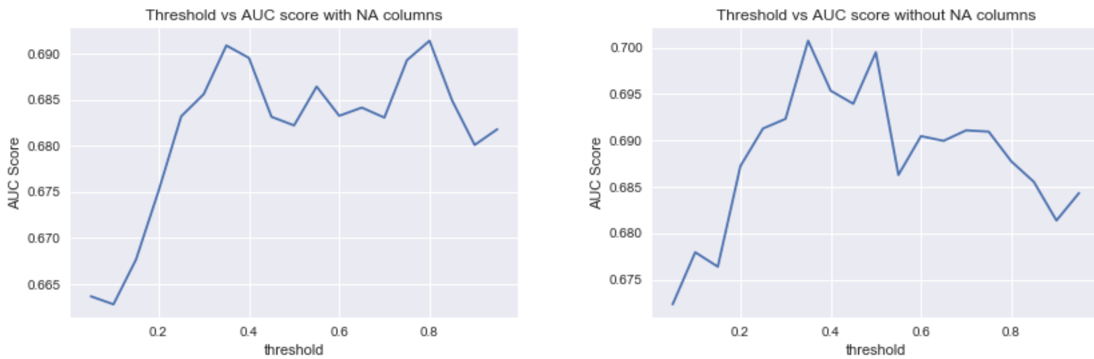
Based on the sparsity of the data set let's see if it's possible remove very sparse samples from the data set to ensure that the data to be trained is representative, but we need to ensure that not too many samples are removed so that the model has sufficient data to train on. Based on this, I applied the following strategy:

1. Remove samples from the data set based on a sparsity threshold, with threshold percentage defined as the percentage of features that are filled.
2. Impute using MICE [7] imputation scheme.
3. Cross validate using stratified shuffle split sampling, and area under ROC curve as scoring metric. This is too ensure that class imbalancing due to sampling will not lead to misleading score.

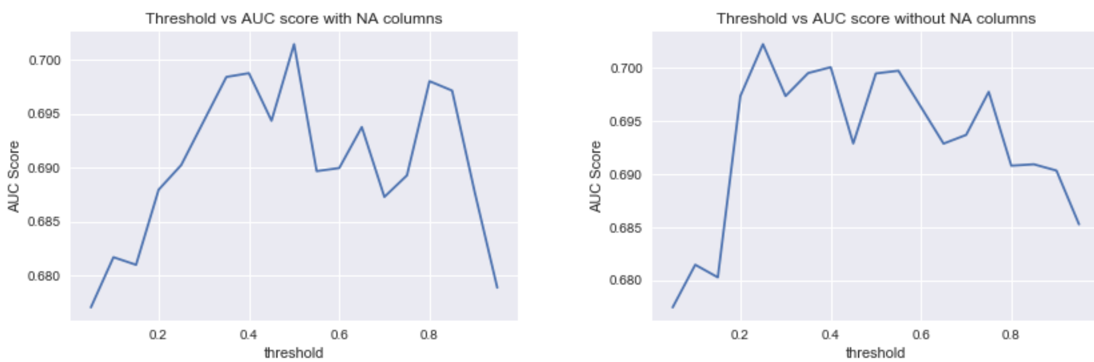
Plots of score vs. threshold value we generatdd for the following models using dummy variables that included whether the sample had an NA as feature and dummy variables that did not.

- Logistic Regression
  - L1, L2, elastic net regularization
- Suport Vector Machine using RBF kernel
- Random Forest (50 trees)
- Gradient Boosting (50 trees)

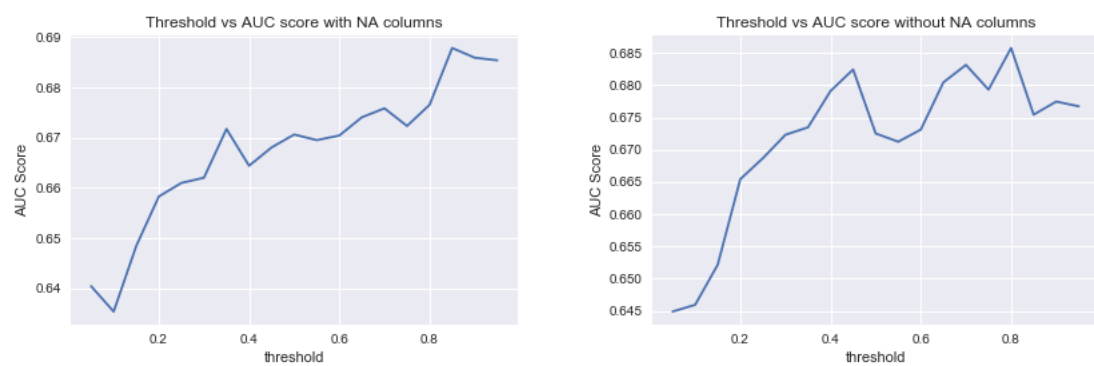
Support Vector Machine



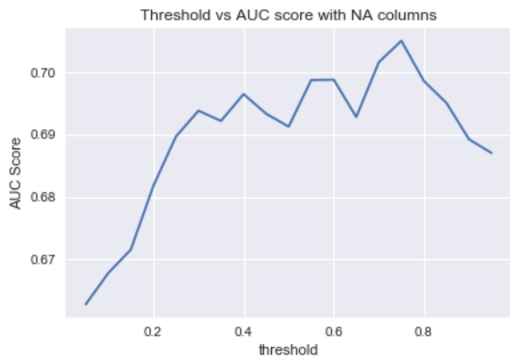
Gradient Boosting

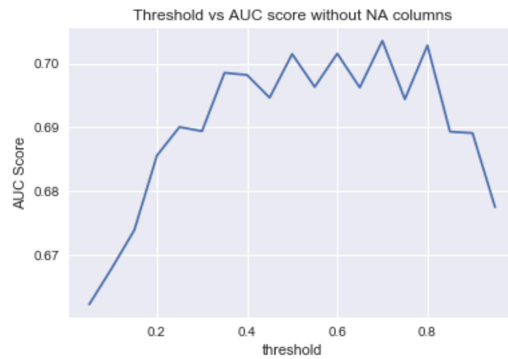


Random Forest

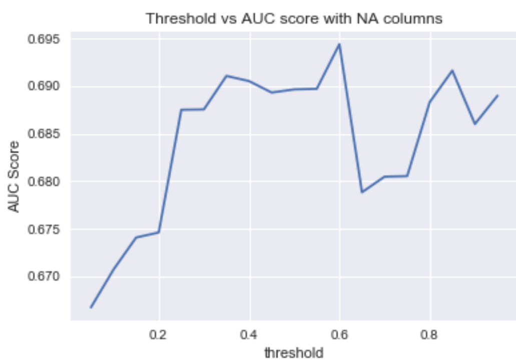


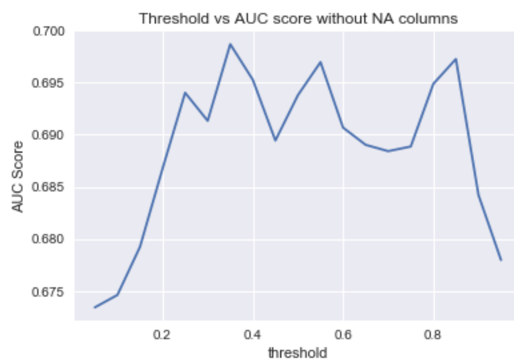
Logistic Regression L1 Regularization



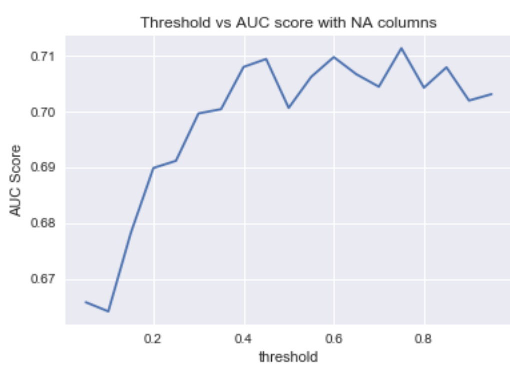


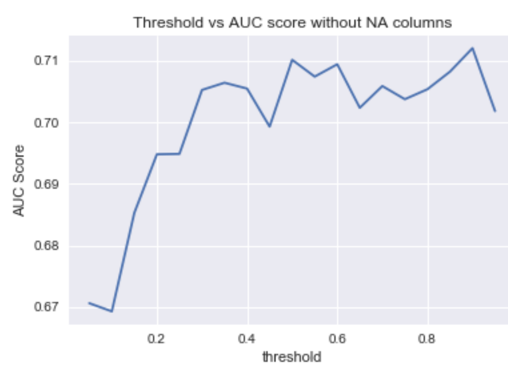
Logistic Regression L2 Regularization





Logistic Regression Elastic Net Regularization





From the analysis above I deduce the following:

- There is not a major difference in performance from:
- Model to Model
- Using dummy variables including NAs vs not inlcuding NAs.
- There performance is generally worse for thresholds of less than 0.3

Based on the following going forward we will the following processing for our training and tuning each model.

1. Utilize data set that includes dummy variables with NAs
2. Reduce samples to 0.3 threshold
3. Impute NAs via MICE algorithm
4. BoxCox transform continuous and ordinal features then rescale using MinMax to ensure that all values are between 0 and 1 (only for non tree models)

Algorithms and Techniques

Since the number of features p = 394 and the number of samples, for a threshold of 0.3, N = 3893, we have the ability of using linear models such as logistic regression. If the data is not linearly separable we can use non-linear models such as support vector machines and tree models (random forests and boosting). Tree models tend to do better if the data has a lot of noise. Gradient boosting tends outperform random forests since they have less bias (but may overfit), but random forests are simpler to tune in terms of hyperparameters.

Benchmark Model

The Kaggle competition creator provided code in the R language perform a simple logistic model to achieve a baseline score:

```
# KAGGLE COMPETITION - GETTING STARTED

# This script file is intended to help you get started on the Kaggle platform, and to show you how to make a submission to the competition.


# Let's start by reading the data into R
# Make sure you have downloaded these files from the Kaggle website, and have navigated to the directory where you saved the files on your computer

train = read.csv("train2016.csv")

test = read.csv("test2016.csv")

# We will just create a simple logistic regression model, to predict Party using all other variables in the dataset, except for the user ID:

SimpleMod = glm(Party ~ . -USER_ID, data=train, family=binomial)

# And then make predictions on the test set:

PredTest = predict(SimpleMod, newdata=test, type="response")

threshold = 0.5

PredTestLabels = as.factor(ifelse(PredTest < threshold, "Democrat", "Republican"))

# However, you can submit the file on Kaggle to see how well the model performs. You can make up to 5 submissions per day, so don't hesitate to just upload a solution to see how you did.

# Let's prepare a submission file for Kaggle (for more about this, see the "Evaluation" page on the competition site):

MySubmission = data.frame(USER_ID = test$USER_ID, Predictions= PredTestLabels)

write.csv(MySubmission, "SubmissionSimpleLog.csv", row.names=FALSE)

# You should upload the submission "SubmissionSimpleLog.csv" on the Kaggle website to use this as a submission to the competition

# This model was just designed to help you get started - to do well in the competition, you will need to build better models!
```

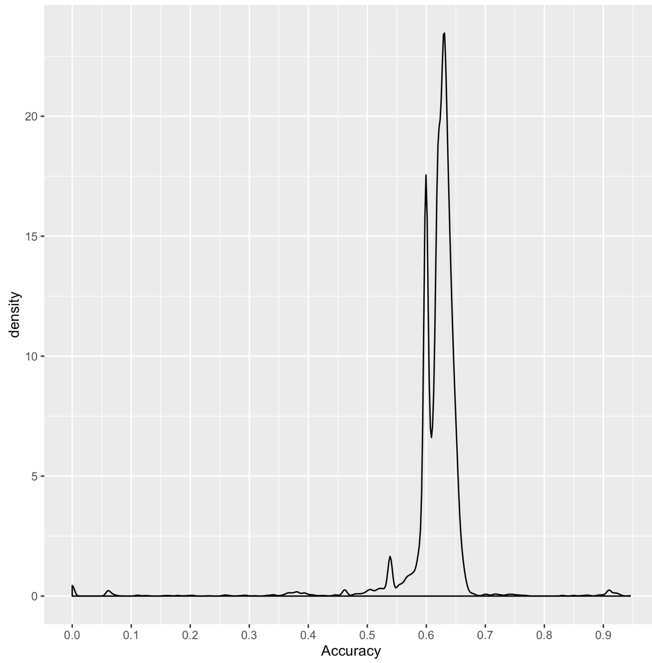
The accuracy on the test data was:

Public Score = **0.59914**

Private Score = **0.57902**

Contestants were expected achieve an accuracy better than this.

A density graph of the public scores was also created from the Kaggle scores generated from 8336 competitors:

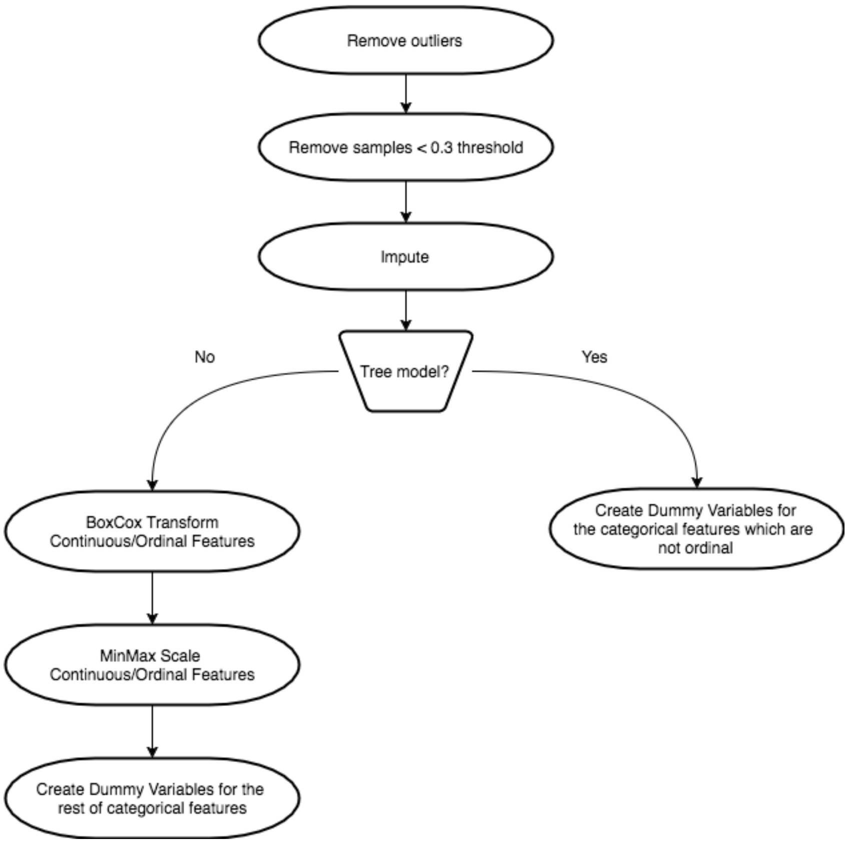


The first hump in the bimodal density graph represents the benchmark. While the second hump represents the students improvement on the data where the median is approximately **0.625**. The goal is to achieve an accuracy score in the neighborhood of the second hump.

III. Methodology

Data Preprocessing

As mentioned above the preprocessing steps are layed out in the diagram.



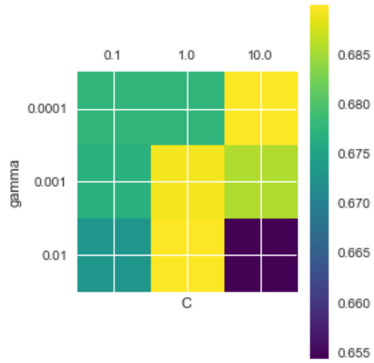
Note again that the MICE imputation will be utilized.

Implementation

The following algorithms and parameters were used:

- Logistic Regression, Specifically `ElasticNetCV` function.
- Perform a GridSearch on the following:
    - alpha values: ranging from $6 \text{ X } 10^{-5}$ and $6 \text{ X } 10^{-1}$
    - L1 ratios: .1, .5, .7, .9, .95, .99, 1. Where value of 0 corresponds to Ridge regression and value of 1 referring to Lasso regression
- Random Forest Classifier
- Perform a GridSearch on the following:
    - number of trees: 1000, 1500, 2000.
- Support Vector Machine
- Perform a GridSearch on the following:
    - Gamma: $1 \text{ X } 10^{-4}$, $1 \text{ X } 10^{-3}$, $1 \text{ X } 10^{-2}$
    - C: 0.1, 1, 10
- Gradient Boosting:
- Perform a GridSearch on the following separately in this order:
    - GridSearch 1
        - Minimum Samples per split: 1200, 1400, 1600, 1800, 2000
        - Minimum samples per leaf: 2, 4, 6, 8, 10
        - Apply best parameters to GridSearch 2
    - GridSearch 2
        - Subsample: 0.6, 0.7, 0.8, 0.9, 0.1
            - Apply best parameters to GridSearch 3
    - GridSearch 3
        - n_estimators: 50, 60, 70, 80, 90, 100, 110, 120, 130, 140

An example of picking the model with the highest prediction scores is shown below:



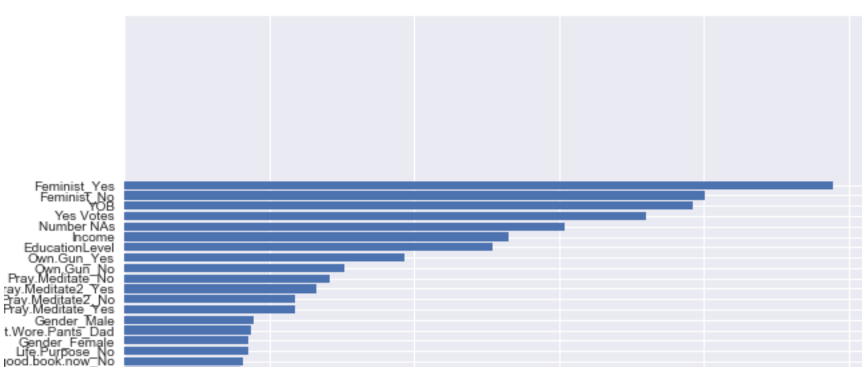
All of the code for processing the data and running the models are documented in the Capstone.ipynb notebook

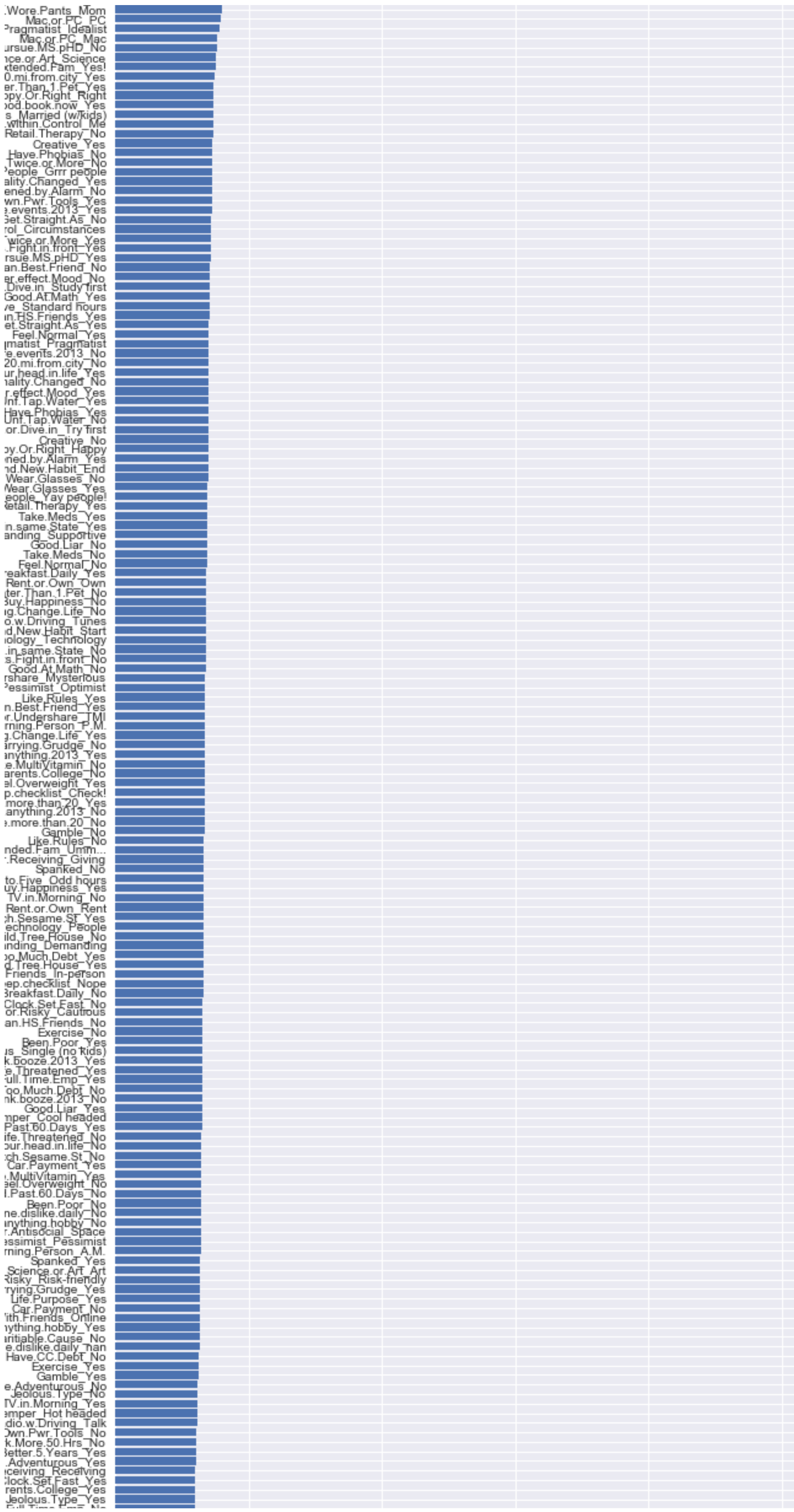
Refinement

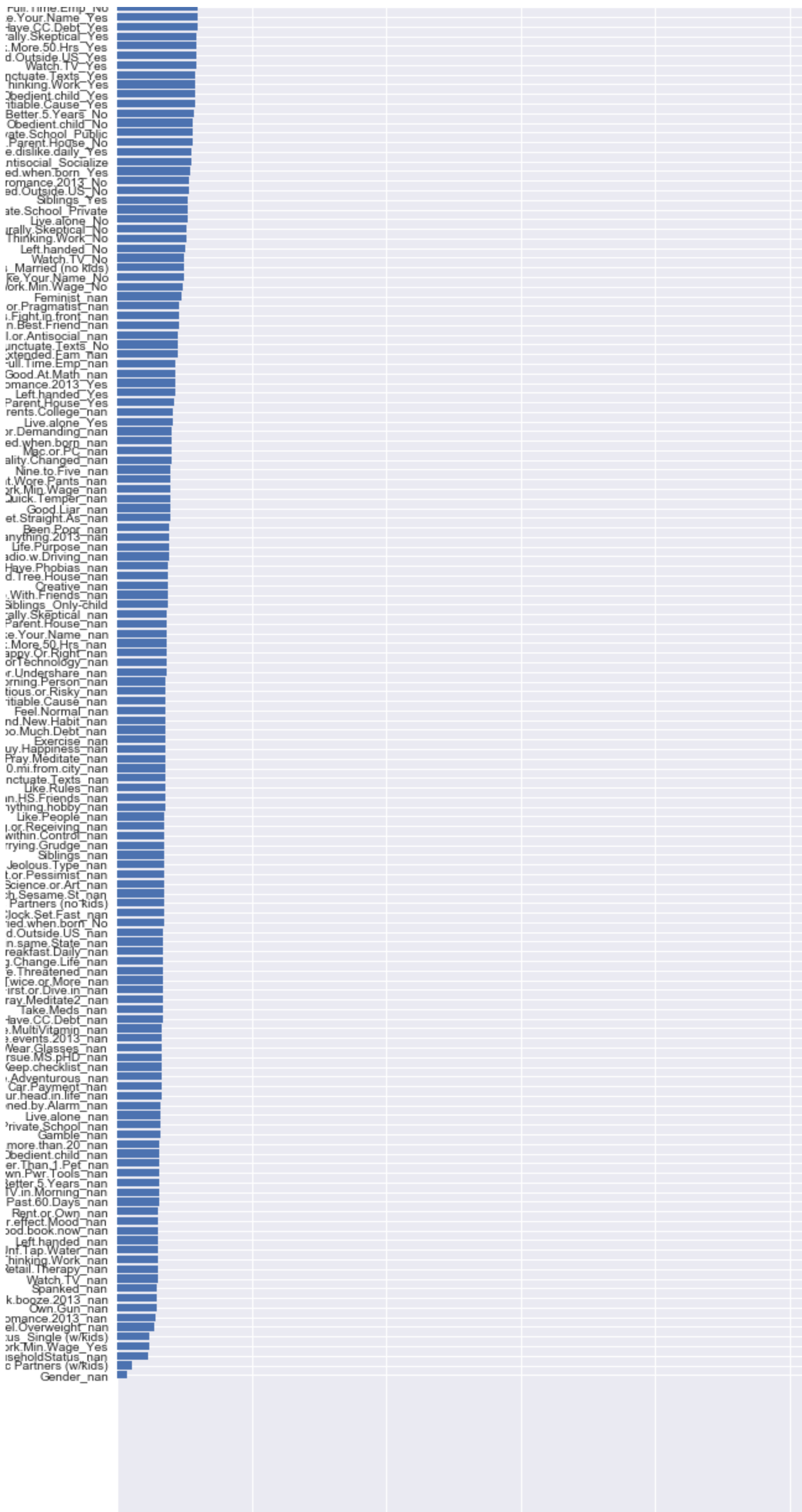
The initial models above were ran and achieved similar results (to be discussed in the evaluation section). In order to increase the prediction score, two metafeatures were created:

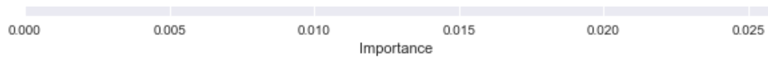
- Number of Questions answered 'Yes'
- Number of Questions left blank

The the new data set were ran again on the Random Forest and Logistic Regression models using the same hyperparameters derived from their grid searches. Here is the importances of the new data set with meta-features. The importance plot shows that Yes Votes and Number NAs are considered among the most important features for prediction.









IV. Results

Model Evaluation and Validation

Model	Parameters	Public Score
Logistic Regression (elastic net)	alpha = 0.0067; L1 ratio = 1	0.62500
Logistic Regression with meta-features(elastic net)	alpha = 0.0067; L1 ratio = 1	0.61638
Random Forest	2000 trees, random features selected = sqrt features	0.62644
Random Forest with meta features	2000 trees, random features selected = sqrt features	0.61925
Gradient Boosting Machine	max_depth = 3; min_samples_leaf = 6; min_samples_split = 1300; n_estimators = 80; subsample = 1.0	0.62213
Gradient Boosting Machine with meta features	max_depth = 3; min_samples_leaf = 6; min_samples_split = 1300; n_estimators = 80; subsample = 1.0	0.62213
Support Vector Machine	C = 10; gamma = 0.001	0.61063
Support Vector Machine with meta features	C = 10; gamma = 0.001	0.62069

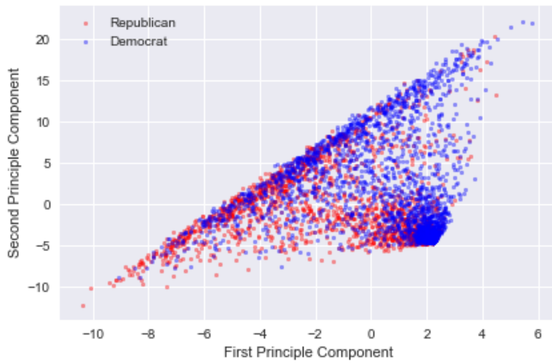
It seems that model performance is agnostic of model type and whether we add these two additional features. Since model parameters were cross validated via stratified k-fold (w shuffling), I have high confidence that model generalizes to data unseen from the model. The scores in the table justify this notion as all the scores are relatively consistent. Based on this I would pick the logistic model using L1 regularization (corresponding L1 ratio =1) with additional meta features as the winning model from a speed performance standpoint. The threshold test shown earlier that the model is robust to perturbations in the original set as long as not too much data is removed,hence the threshold of 0.3 value.

V. Conclusion

Reflection

We've shown that with straightforward preprocessing (removing samples based on thresholding and imputation) and gridsearch of parameters via cross-validation that the prediction score can be increased by approximately two extra percent. Given the Kaggle scores of previous submissions this score is on par of what other participants achieved on average. Given the sparsity of the data, there were numerous ways that the data could have been processed. Multiple imputation schemes as well as deleting entire features altogether. In terms of feature selection multiple types of subselection schemes could have been applied such as recursive feature elimination and univariate correlation with the target.

On sort of dimension reduction technique is an off shoot of PCA called Partial Least Squares. In short it's somewhat of a supervised version of PCA. Here is the plot below for two dimensions:



As you can see there are separation in certain areas but overlap in others. When a classifier was performed on it produced similar results to the previous models. When performed on higher dimensions the results were still in the same ballpark.

The elastic net L1_ratio set to 1 regularization (complete LASSO) is a method to select the best features. It's result still seemed to be on par with the rest of the models that included all of the features.

Improvement

Although the previous processing and gridsearch validation did increase the score from the baseline. It did not improve it drastically. I feel that in order to increase prediction score even further more features need to be created with use of **domain knowledge**. For example weighting certain features and combining them either additavely or multicatively based on domain knowledge to create features that the model hasn't previously seen would aid in separating Republican samples from Democrat samples further. Most likey the top Kagglers had used this strategy.

References

[1] *Party Predictor: Predicting Political Affiliation http://cs229.stanford.edu/proj2013/EwonusMcCannRoth-PartyPredictorPredictingPoliticalAffiliation.pdf*

[2] *Predicting the Political Alignment of Twitter Users https://pdfs.semanticscholar.org/ccaf/a80db5f4b19886d6bbe9a2a37e2048d52a28.pdf*

[3] *Decision Trees and Political Party Classification https://jeremykun.com/2012/10/08/decision-trees-and-political-party-classification*

[4] *http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records*

[5] *https://inclass.kaggle.com/c/can-we-predict-voting-outcomes*

[6] *https://www.showofhands.com*

[7] *https://stat.ethz.ch/education/semesters/ss2012/ams/paper/mice.pdf*