The goal of our preprocessing will be to ensure that we identify missing values for each of the variables and ensure that we treat ordered and non-ordered categorical variables properly. To accomplish this, we read the through the CSES4 Questionnaire document on the CSES website to see if a ordered relation exists between the labels and identify the corresponding value for a missing variable (typically 9, 99, or 999). We then mark the missing values in our dataset and do listwise deletion to remove records that include a missing value. However, before doing this, we exclude from our analysis any independent variable that has more than 10% of values missing.

Of our remaining sample, we have 1,561 true non-voters and 7,699 true voters (16.9% of respondents left after dropping duplicates did not vote; this also informs us that our model needs to outperform the naïve model of predicting everyone to vote with accuracy of 83.1%). Because of this imbalanced ratio, we will elect to use an under-sampler on the majority class. Given that we have almost 10,000 samples, there should be a sufficient number of samples left to extract information from the majority class even after under-sampling. We will empirically test to find the optimal majority-minority ratio to use in this sampler.

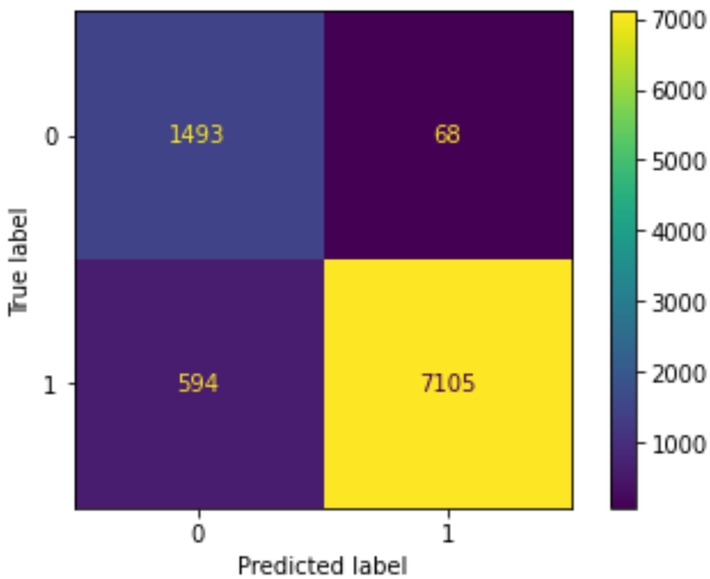
After applying the under-sampler, we will use a standard scaler to scale our input variables. Although this is not strictly necessary for the Random Forest, it will be useful in our Logistic Regression as we apply a both L1 and L2 penalties, so we need our variables to be on the same scale.

We will not select features on the basis of theory and will instead use a feature selection technique to maximize predictive accuracy empirically. We choose to use a select-k-best feature selector using mutual information as our selection criteria. We optimize empirically to find the number of features to select.

We will try a number of different model constructions, including a k-nearest neighbor classifier, a random forest classifier, and a logistic regression. We vary a number of hyperparameters for the construction of each of these models to adjust for overfitting and introduce regularization.

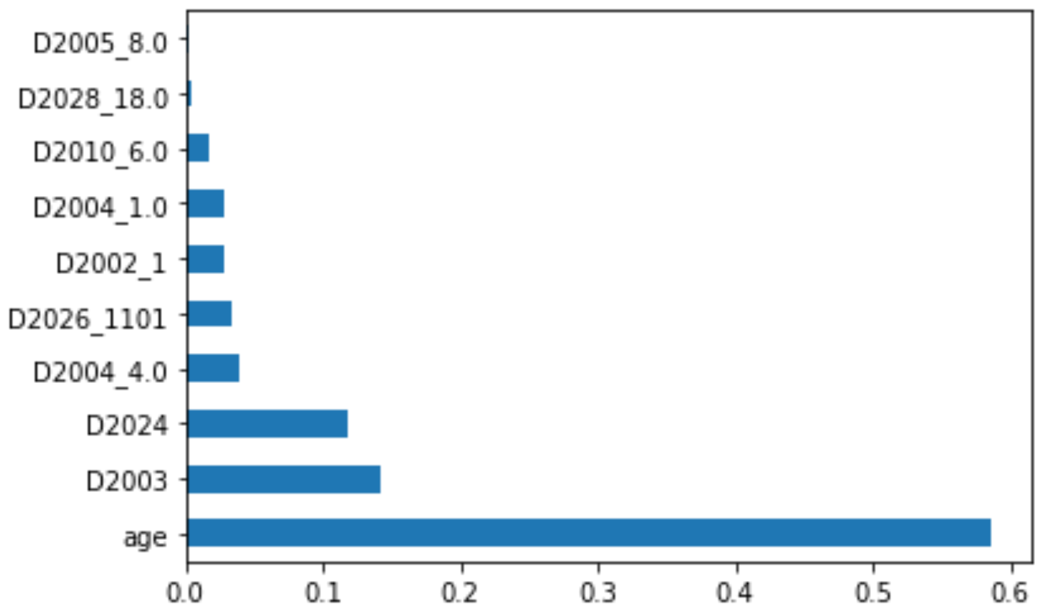
To identify the best set of hyperparameters, we use a cross-validated grid search. We judge our models on the basis of the F-1 score. Since we do not know the specific application of this ML-model, we think an equal tradeoff between precision and recall is fair as the cost of a false negative and false positive are somewhat equal with no use case to guide us. As we are using a 10-fold cross validation in our hyperparameter selection, we will elect not to hold out a test set and will instead judge our models on the basis of the averaged cross-validated score. Our final model has an average F-1 score of 0.887 from the cross-validation. We find that this best model is a Random Forest Classifier with and uses the 10 features shown in the feature importance chart below.

Confusion Matrix:



- Note that our model has very few False Positives. If we were using this model to create a campaign where we will target people who are likely to not vote, then this is a good model as most of the people who will not vote are predicted that way.

Feature Importance:



- After feature selection, we find that age, education, religious services attendance, marital status as single/not single, religious denomination, gender, employment status as student or not are the key features.