MACS 301
Homework 8

Reid McIlroy-Young

March 6, 2017

1 Biden

Part 1

To start with the data were loaded and a testing set of 30% of the data was created.

Part 2

The initial tree using just *dem* and *rep* is show in figure 1. The MSE from the model, against the testing set is 394.97.

Part 3

When we go to the full tree with all variables considered the lowest MSE of the prunings considered is 342.33, better than before, but we will see if we can improve it. The optimal tree is shown in figure 2 and the plot of minimum samples per leaf is shown in figure 3 this show hows each of MSE's for a specific number of samples, along with their mean as the line. We obtained this plot by using 10-fold cross validation and varying the minimum number of samples per node in the tree, once generated trees cannot be pruned under *sklearn*. This is still tree pruning though, just not of a pre-existing tree. The plot indicates that from about 400 to 900 samples per leaf the MSE is lowest so we will be using 500 were appropriate in later models.

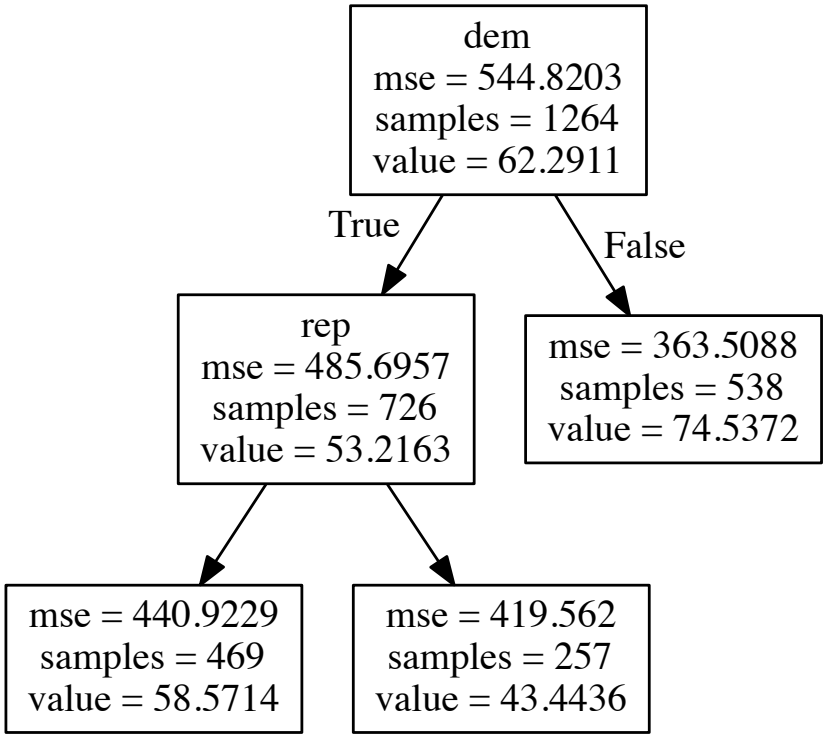

Figure 1: *Question 1, Part 2 tree*

Part 4

Creating a series of models with bagging and varying the number of hypothesis (estimators), produced a model with 78 hypothesis and an MSE of 506.86. The plot of hypothesis vs MSE is shown in figure 4. *sklearn* does not support variable importances for bagging so, I have computed them for boosting as well as random forest.

Part 5

Creating another series of models with a random forest and varying the number of considered features gives a minimum MSE of 411.95 with 4 features considered. The MSE vs features plot is shown in figure 5. The table of

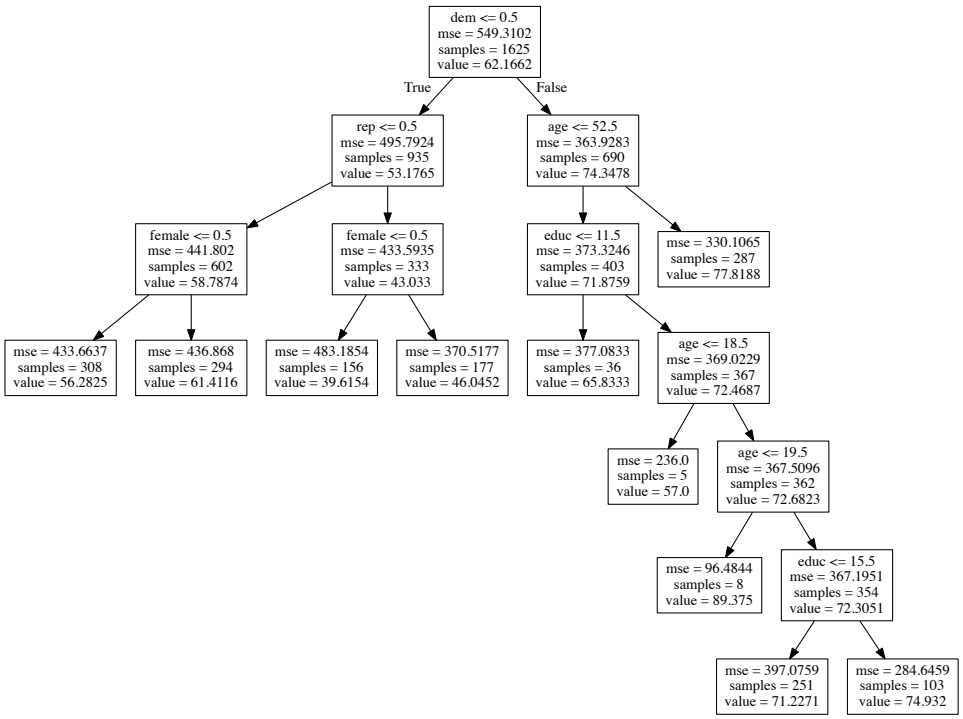

Figure 2: *Question 1, Part 3 tree*

variable importances is shown in figure 6 and shows that the model puts almost all of its weight on the affiliation of the respondent and unlike some past models virtually no weight on their gender.

Part 6

Creating another series of models with boosting and varying the learning rate gives a minimum MSE of 394.33 at a value of 0.1. The MSE vs learning rate plot is shown in figure 7. The table of variable importances is shown in figure 8 and shows that the model puts most all of its weight on the age of the respondent and spreads out the rest, a more diffuse set of importances shows it is using more of the information provided and thus the significant decrease in error is not surprising.

When comparing to all the models we can see that simply pruning the decision tree leads to the best outcome this tells us that it is likely the more complex models were over fitting. That boosting with a reduced learning rate came in second also suggests this since a higher learning rate would allow

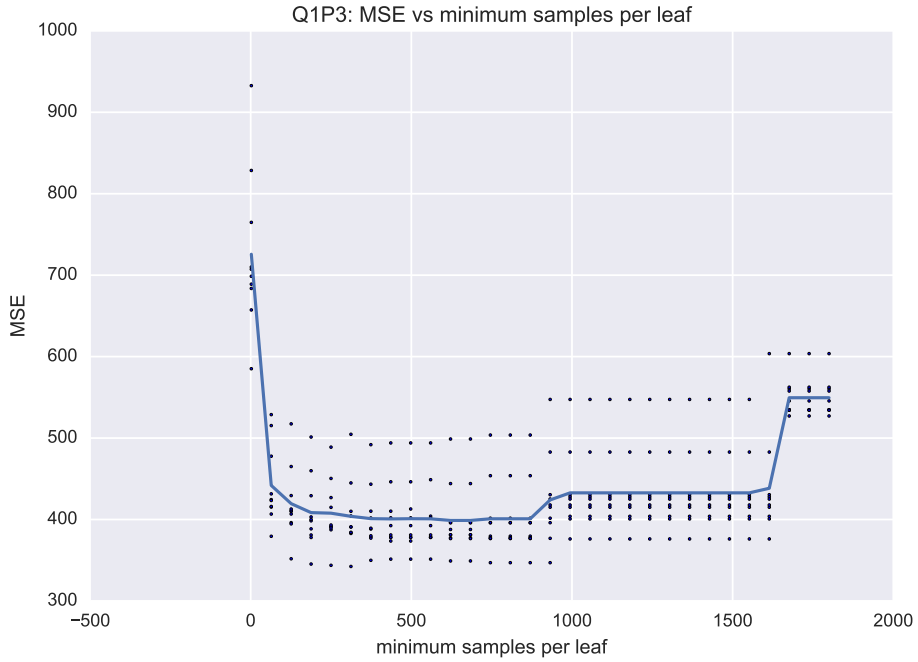

Figure 3: *Question 1, Part 3 plot*

the model to become more complex faster, although as we can see from the plot making it lower does not help past 10^{-2}.

2 Voter Turnout

Part 1

The models we choose to use are:

DecisionTreeReduced a decision tree using only $mhealth_sum$ and age

DecisionTree a full decision tree

Bagging a classifier derived from a bagging approach

RandomForest a classifier derived from a random forest approach

AdaBoost a classifier derived from application of adaboost (the best boost)

These gave the statistics found in figure 9 and the ROC curves in 10.

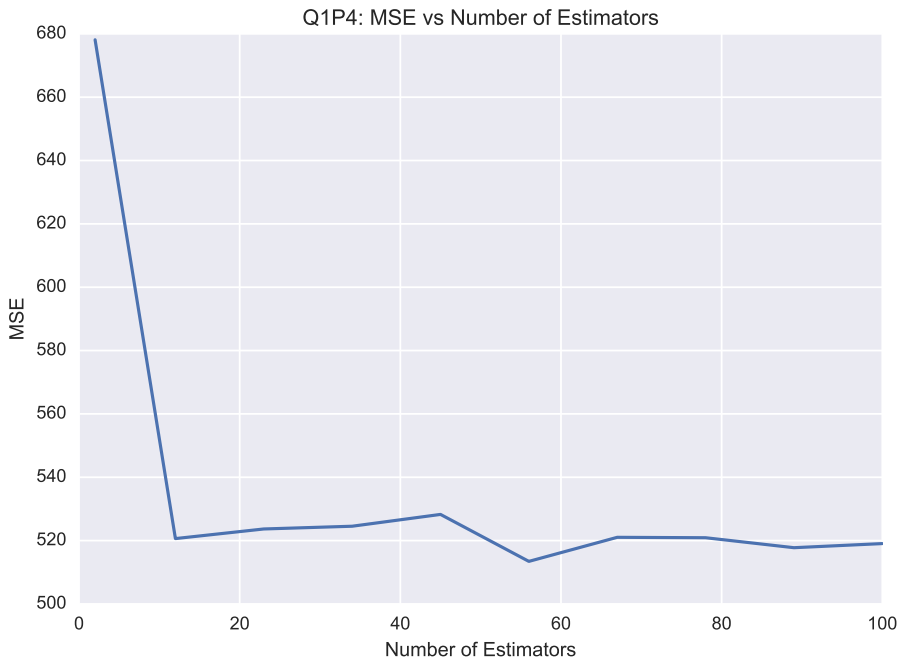

Figure 4: *Question 1, Part 4 plot, going beyond 100 does not improve things but does take much longer to compute*

From looking at the table we can see that *adaboost* is the best on all measures but recall, where *RandomForest* is sightly better. The gap between these two is narrow on all metrics though so either would be a valid choice for the best, that said we choose *adaboost* as the general best fit. We can look at the variable importances for it and see what it considers important. Figure 11 shows them. Interetsingly enough *age* is by far the most important factor, with the remainder of the influence spread between *educ*, *inc*10 and $mhealth_sum$. These variables being important helps explain why the decision tree with full range of variables was worse than that with only *age* and $mhealth_sum$, those variables are likely more important and thus a simple model is sufficient.

None of the models we examined were significantly better than any of the others, the range of the AUC values is only about .15. This tells us that our best model is not much better than out worst and that likely none of the ones considered will allow us more than 70% accuracy.

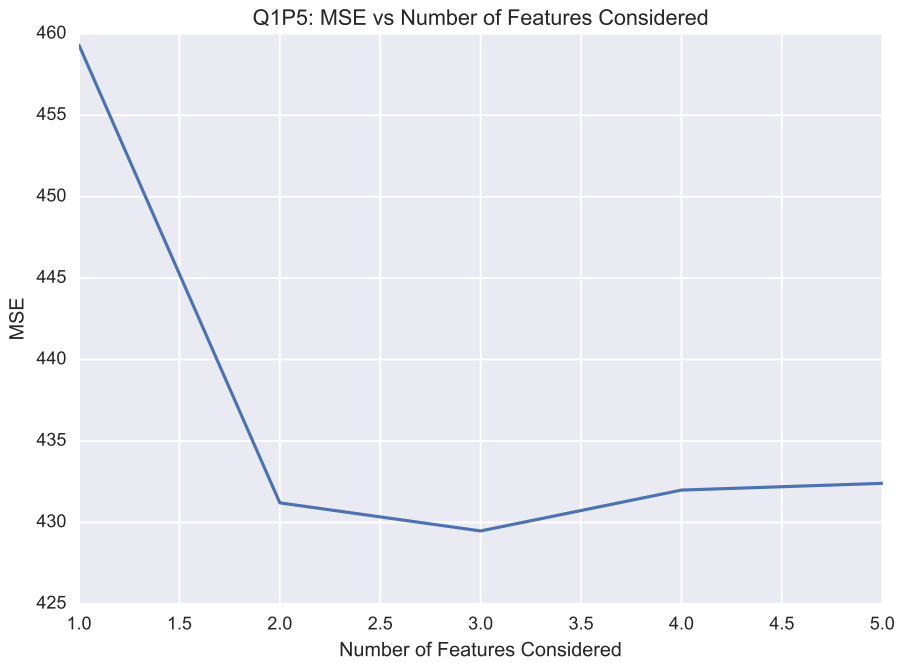

Figure 5: *Question 1, Part 5*

Part 2

The models we choose to use are:

linearSVM reduced a linear SVM trained with only *age* and $mhealth_sum$

linearSVM a linear SVM trained with the full data set

radialSVM a radial kernel SVM trained with the full data set

nuRadialSVM a radial kernel SVM trained with the full data set, changes number of support vectors

Figure 6: *Question 1, Part 5: Variable importances*

```
female :  0.0000
age    :  0.1000
educ   :  0.0000
dem    :  0.8349
rep    :  0.0651
```

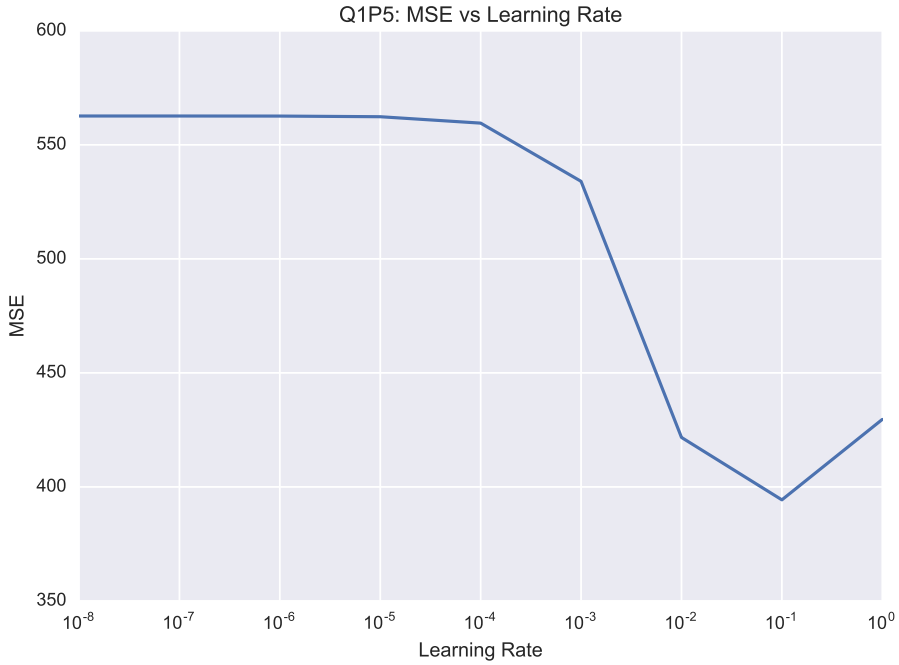

Figure 7: *Question 1, Part 6*

poly2SVM a 2nd order polynomial kernel SVM trained with the full data set (3rd order took too long to run)

These gave the statistics found in figure 12 and the ROC curves in 13.

From looking at the table we can see that *poly2SVM* and *radialSVM* are in close contention for the best values. *poly2SVM* has a slighly better AUC and recall so we will say it is the best, although either would be acceptable. What is notable is how similar this model is than the best form part 1, *poly2SVM* has an AUC of 0.669227 while *adaboost* has an AUC of 0.676059 the difference is tiny. The variance amongst models though is much larger for

Figure 8: *Question 1, Part 5: Variable importances*

```
female :  0.0466
age    :  0.5756
educ   :  0.1679
dem    :  0.1265
rep    :  0.0834
```

Figure 9: *Question 2, Part 1: table of statistics*

	AP	PRE	RE	auc
DecisionTreeReduced	0.834994	0.752137	0.752137	0.626068
Bagging	0.850517	0.777778	0.748971	0.631495
AdaBoost	0.913532	0.880342	0.741007	0.676059
RandomForest	0.876923	0.820513	0.761905	0.666667
DecisionTree	0.834164	0.752137	0.733333	0.603030

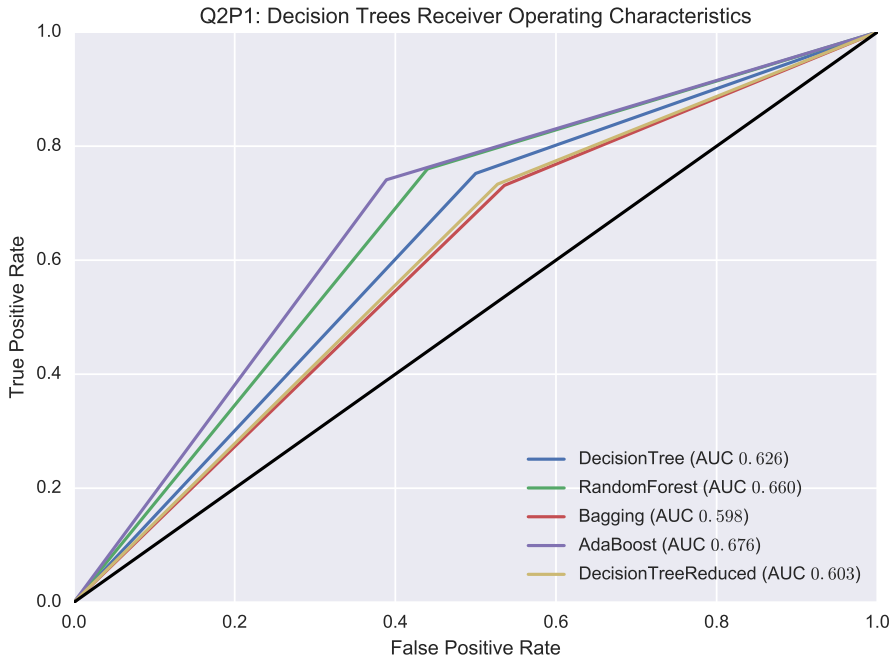

Figure 10: *Question 2, Part 1*

the SVMs *linearSVM reduced* is by far the worst of all the models examined. The *nuRadialSVM* was only included because of how difference it is from the other *radialSVM*, but neither is good.

3 Voter Turnout

To use the data provided for this question the the *income* and *educ* variables had to be converted to numerical values. To do this we converted the *income*s to the mean value of that stated and for *educ* the responses were ranked from

Figure 11: *Question 2, Part 1: adaboost Variable importances*

```
mhealth_sum:  0.1200
age       :  0.5000
educ      :  0.1600
black     :  0.0200
female    :  0.0000
married   :  0.0200
inc10     :  0.1800
```

Figure 12: *Question 2, Part 2: table of statistics*

	AP	PRE	RE	auc
linearSVM	0.502137	0.004274	1.000000	0.666189
poly2SVM	0.919213	0.888889	0.732394	0.669227
linearSVM_reduced	0.632290	0.423077	0.744361	0.561121
radialSVM	0.942516	0.923077	0.710526	0.659611
nuRadialSVM	0.890041	0.841880	0.729630	0.633565

lowest to highest and assigned a number corresponding to their rank, NAs were assigned the middle category.

Part 1

The models we choose to use are:

linear	linear fit, round at .5
linear highThreshold	linear fit, round at highest category value
linear LowThreshold	linear fit, round at lowest category value
DecisionTree	decision tree
logit	logistical regression
RandomForest	a classifier derived from a random forest approach
linearSVM	a classifier derived from a linear kernel SVM

Figure 13: *Question 2, Part 2*

When we looked at just race and belief there was a quite starting result, all the models made the same predictions. Figure 14 shows the resulting statistics, and figure 15 the ROC curves.

Further examination showed why, figure 16 shows the results of the Random Forest on all 4 possible inputs. If the respondent is black not guilty otherwise, guilty. This is also supported by looking at the model's variable importances figure 17.

All the models initially had this effect, we changed the rounding on the linear fit to see if it would affect the results positively. As the table shows, rounding down or rounding up both result in worse predictive power. Thus we can say that race is significant predictor of belief.

Part 2

The models we choose to use are:

ridge Tikhonov regularization OLS model

linear linear fit

Figure 14: *Question 3, Part 1: table of statistics*

	AP	PRE	RE	auc
linear	0.965577	0.958065	0.836620	0.825453
linear_highThreshold	0.921619	0.880645	0.842593	0.738128
DecisionTree	0.965577	0.958065	0.836620	0.825453
logit	0.965577	0.958065	0.836620	0.825453
linear_LowThreshold	0.994904	0.993548	0.735084	0.700875
RandomForest	0.965577	0.958065	0.836620	0.825453
linearSVM	0.965577	0.958065	0.836620	0.825453

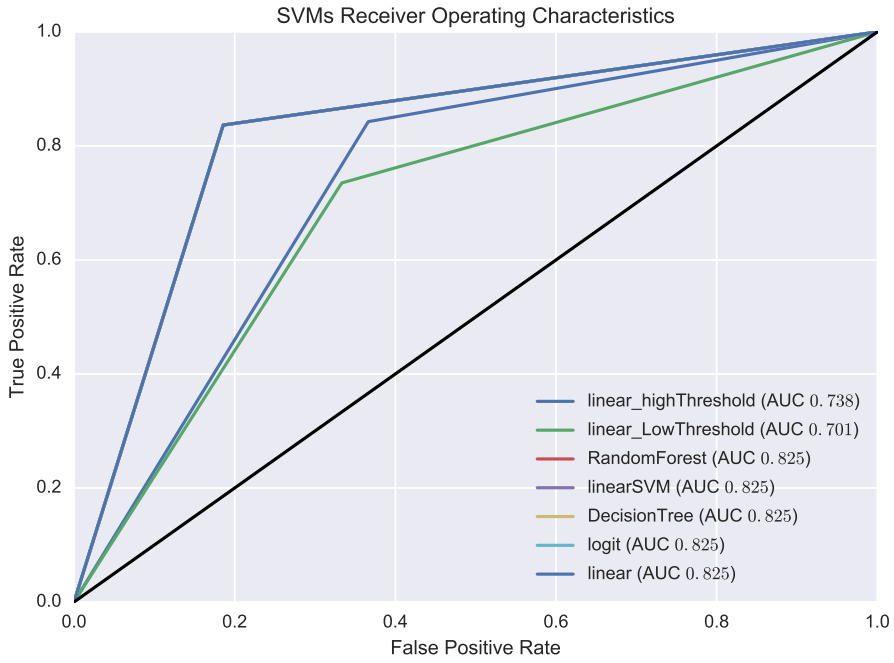

Figure 15: *Question 3, Part 1*

logit	logistical regression
DecisionTree	a classifier derived from a decision tree approach
RandomForest	a classifier derived from a random forest approach
linearSVM	a classifier derived from a linear kernel SVM
poly	2nd order polynomial fit

Figure 16: *Question 3, Part 1: table of results for RandomForest*

	hispanic	not hispanic
black	0	0
not black	1	1

Figure 17: *Question 3, Part 1: RandomForest Variable importances*

```
black  :  0.9913
hispanic:  0.0087
```

radialSVM a classifier derived from a radial kernel SVM

These gave the statistics found in figure 18 and the ROC curves in 19.

Looking at all these models there are four in close contention for being the best which when ranked by AUC are $linearSVM$ followed by, *logit*, *ridge* and *linear*. The last three are all very similar models, all belong to the class of generalized linear models so their closeness is not surprising. That the SVM is best is interesting, it would be the one we choose as being the best, although the linear model is tempting as it is much easier to interpret. Luckily the SVM is linear so we can interpret the coefficients as linearly relating to the classification. Figure 20 has the coefficients, these show that *black* is by far the most important factor, with *educ*, *hispanic* and $female$ all being significant.

Figure 18: *Question 3, Part 2: table of statistics*

	AP	PRE	RE	auc
ridge	0.965577	0.958065	0.836620	0.825453
linear	0.965577	0.958065	0.836620	0.825453
logit	0.965583	0.958065	0.838983	0.827942
DecisionTree	0.850412	0.758065	0.827465	0.647775
RandomForest	0.876824	0.803226	0.832776	0.674324
linearSVM	0.979006	0.980645	0.815013	0.849814
poly	0.943561	0.919355	0.840708	0.775005
radialSVM	0.945399	0.919355	0.762032	0.635918

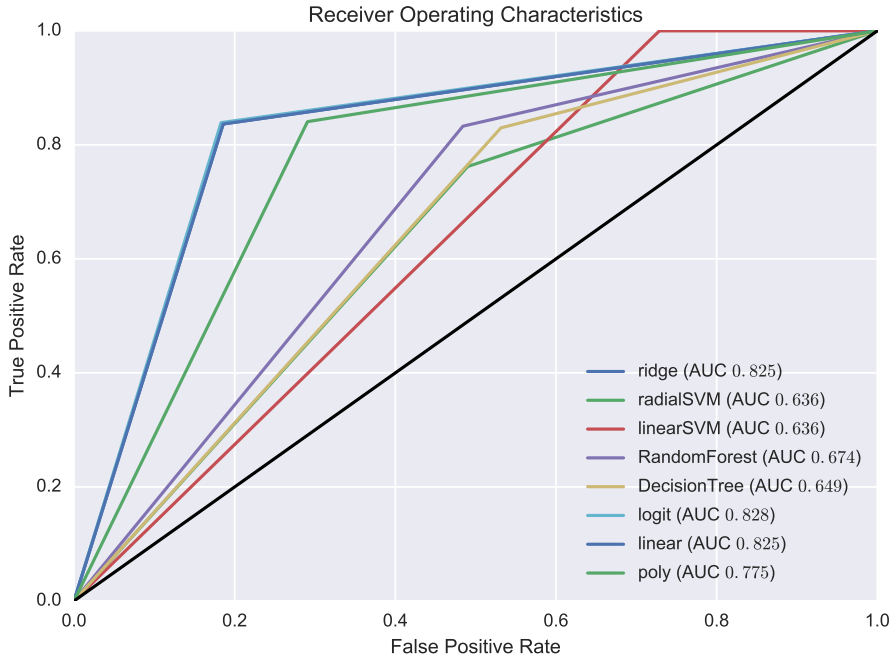

Figure 19: *Question 3, Part 2*

One interesting note about the top models is they tend to assume guilt, figure 21 shows the full table for *logit*. Notice how the precision for only 0 is much lower than for 1, this is in part due to the sample having more 1's but obtaining a model with a more even spread would be better for making unbiased predictions.

Figure 20: *Question 3, Part 2: coefficients of linearSVM*

```
dem      :  -0.0366
rep      :  0.0204
ind      :  0.0000
age      :  -0.0057
educ     :  0.1271
female   :  -0.1251
black    :  -1.1353
hispanic :  -0.1030
income   :  0.0117
```

Figure 21: *Question 3, Part 2: table of statistics for logit*

	precision	recall	f1-score	support
0	0.50	0.82	0.62	71
1	0.96	0.84	0.89	354
avg / total	0.88	0.84	0.85	425