Homework 7

Jeff Tilton

10/6/2018

Question 10.1

Using the same crime data set uscrime.txt as in Questions 8.2 and 9.1, find the best model you can using

- a regression tree model, and
- a random forest model.

In R, you can use the tree package or the rpart package, and the randomForest package. For each model, describe one or two qualitative takeaways you get from analyzing the results (i.e., don't just stop when you have a good model, but interpret it too).

Goals

1. Create a regression tree model for the crime data
    - Create a decision tree
    - Use decision tree to create a regression for the leaf nodes
2. Create a random forest model for crime data
3. Choose the best model
4. Interpret model

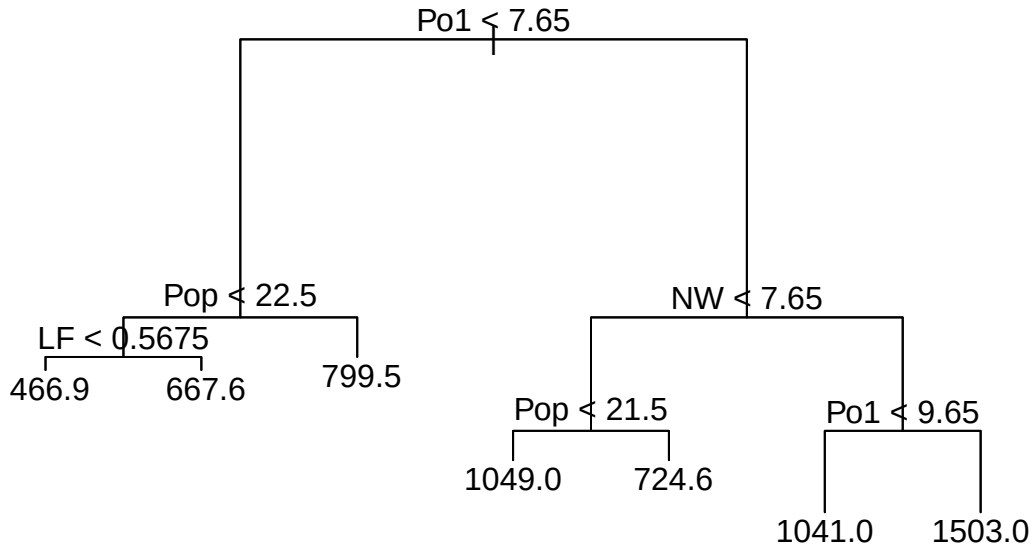
Regression tree model

Method

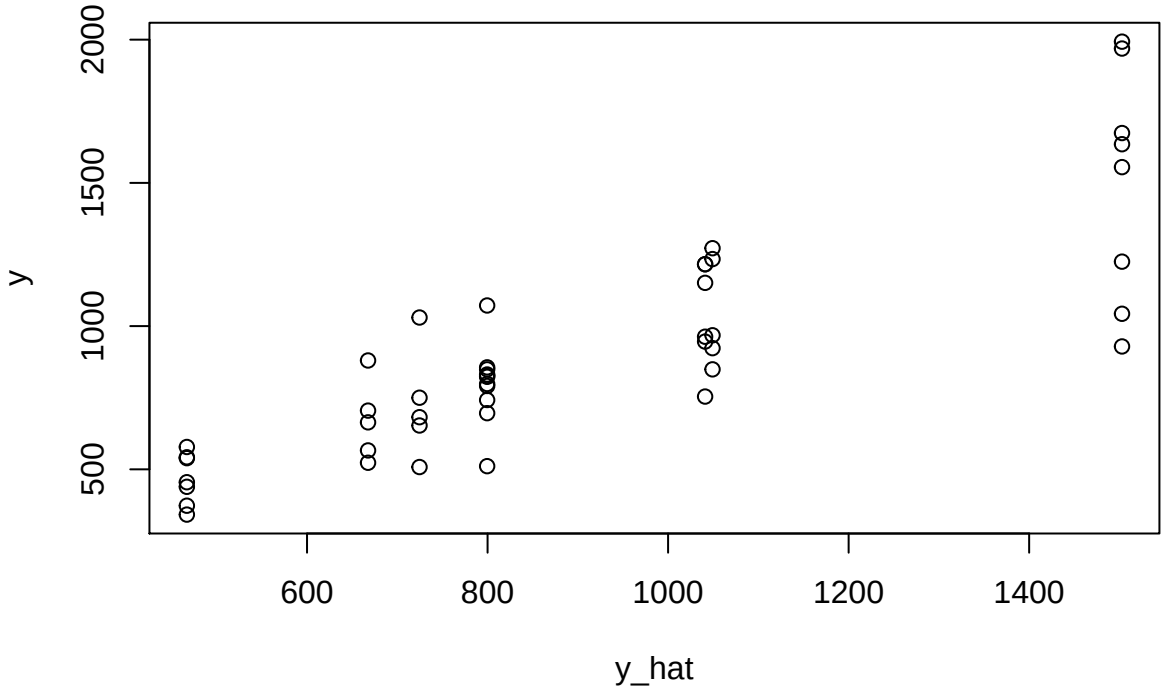
1. Create a decision tree using the `tree` package‘
2. Use cross-validation to determine an appropriate size of the tree
3. Create new pruned tree
4. Create regression models for the new tree leaves

Create a decision tree

Initial Tree

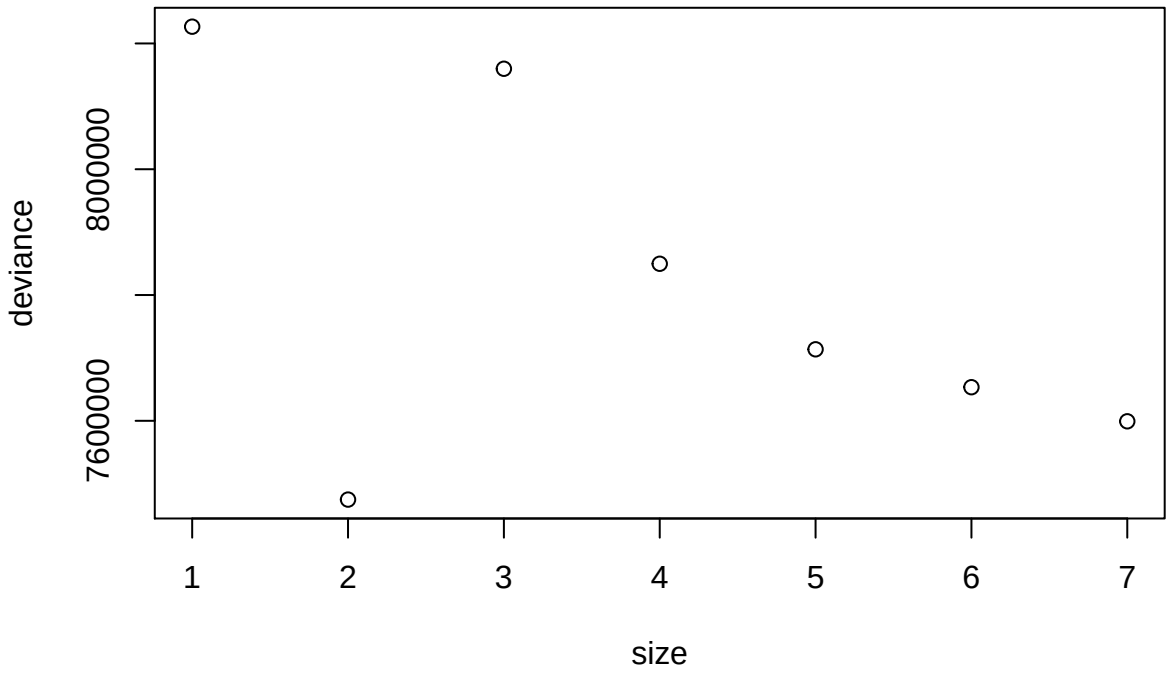


The above tree is the initial tree created by the `tree` library with default settings. It split 6 times and has 7 leaves. I will call this model `initial`.



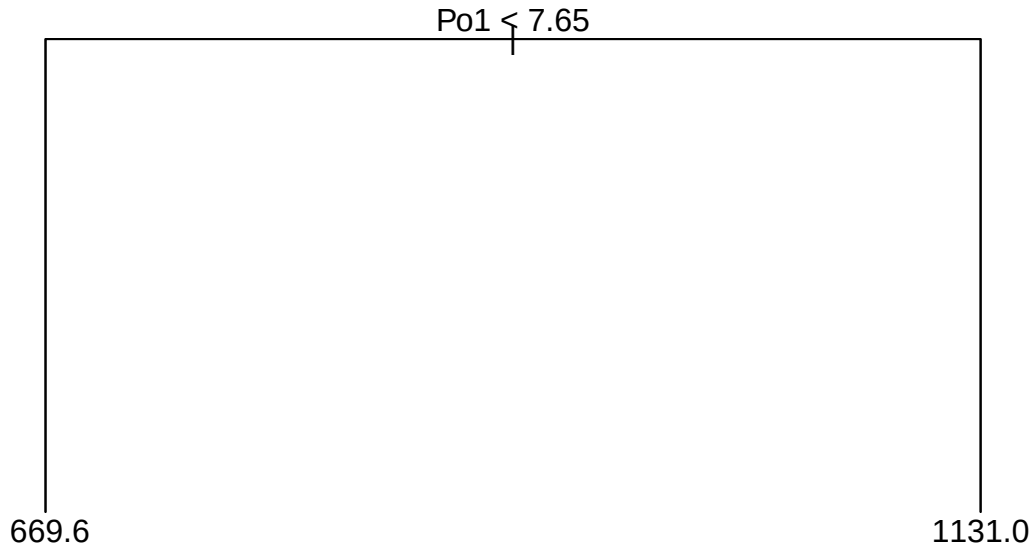
A plot of the `initial` predictions vs actual crime.

Cross Validate



A plot of of the tree size vs deviance after a 10,000 fold cross-validation. A size 2 tree had the lowest deviance.

Pruned Tree



A plot of the newly pruned size 2 tree chosen by cross-validation.

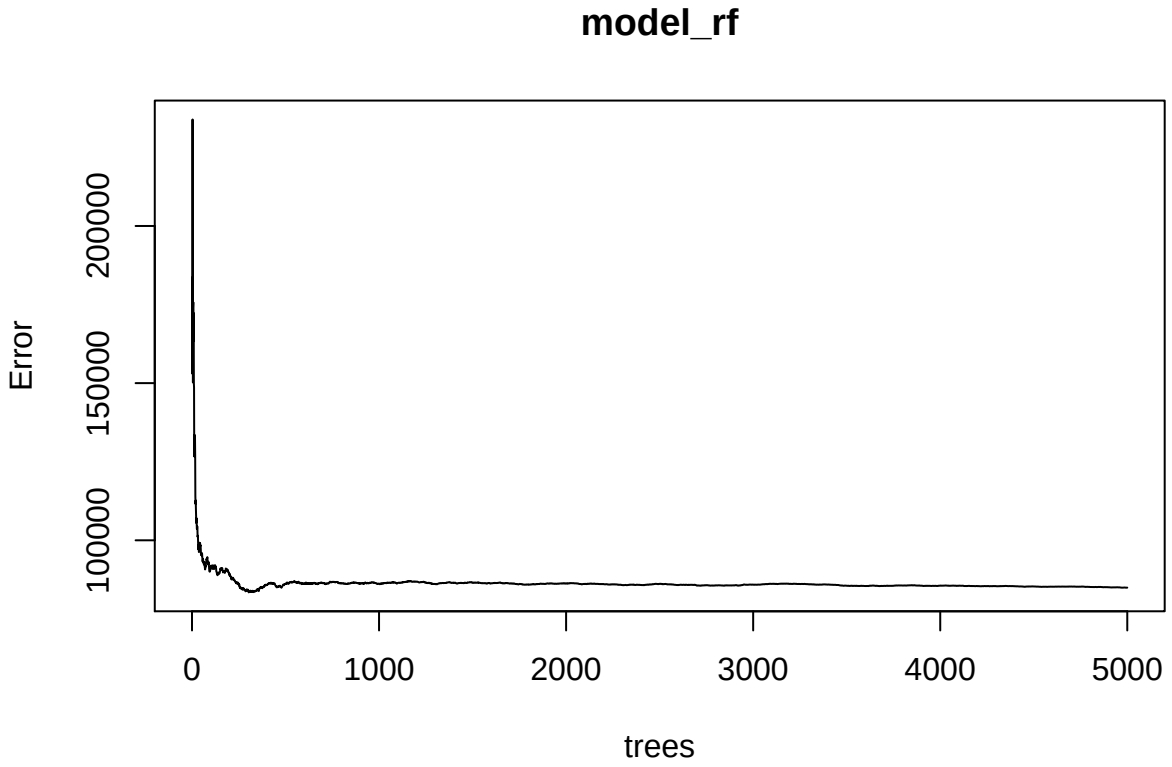
Regression Leaves

I created 4 regression tree models, 2 models for each leaf. The first models regressed on all data, then I chose to reduce the predictors to predictors with a p-value $< .075$ and created two more models. I ran 5-fold cross-validation on each model. I will call these models `leaf 1`, `leaf 1 significant`, `leaf 2`, `leaf 2 significant`.

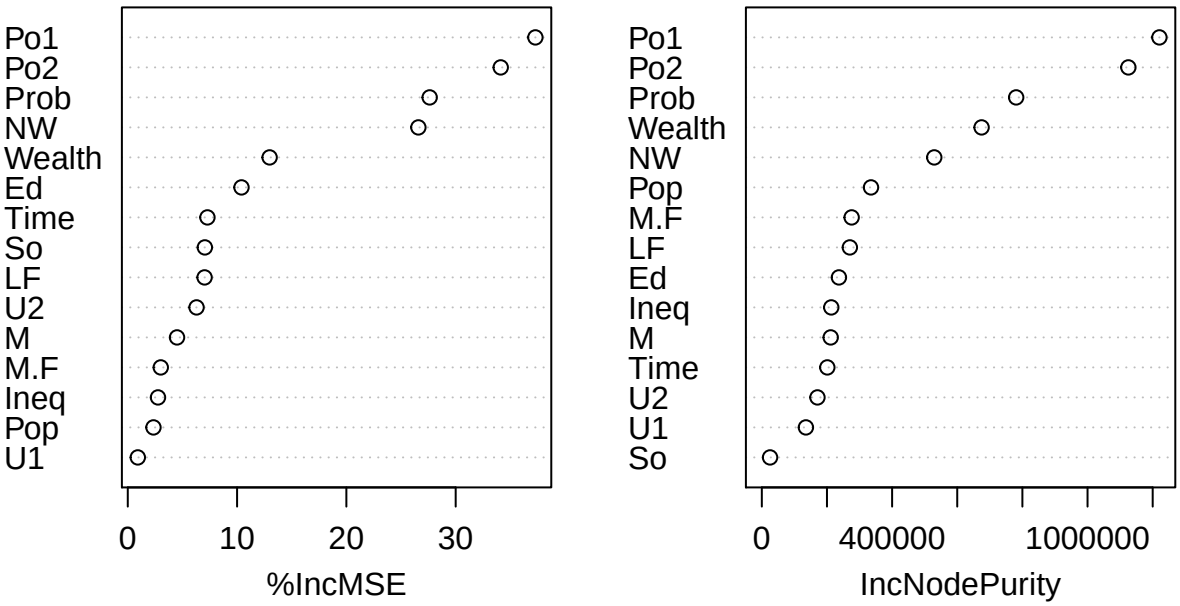
Random Forest model

Method

1. Create a random forest model with package `randomForest`



A plot of the number of trees vs error on a random forest model.



Above is a dotchart of variable importance as measured by a Random Forest.

Choose the best model

Method

1. Gather all R^2 values
2. Gather all Mean Squared Error (mse) values
3. Choose model based on results

Model results

	Initial	leaf 1	leaf 1 significant	leaf 2	leaf 2 significant	Random Forest
R^2	0.72	0.88	0.45	0.88	0.03	0.42
mse	4.03e+04	1.25e+05	3.09e+04	6.10e+05	1.75e+05	1.54e+04

I would choose the Random Forest model as the best model based on the results above because it has the lowest mean squared error.

Interpret Model

Although the leaf 1 and leaf 2 models have higher R^2 values they are demonstrably overfit as seen in the large mse values. It is interesting that the leaf 2 significant R^2 value is 0.03, but still outperforms the leaf 2 model in cross-validation.

Regression Tree

These models do not have a significant amount of data and I found it interesting that after cross-validation the tree size with the smallest deviance was 2. Although it did not "feel right" to use a tree of size 2, I decided I would go where the data leads me and it offered me the chance to have same data to work with in the regression tree.

The only significant split in my size 2 model is Po1, which is the per capita expenditure on police protection in 1960. Po1 is an interesting predictor, at first I thought that crime and Po1 were correlated, with crime being the dependent variable. I thought that the most likely scenario was a spike in crime precipitated a city's increase in police expenditure. I then contemplated that it is probably more nuanced. I can imagine a situation when a city mayor ran on increasing police expenditure, won and without any significant increase in actual crime the crime rate went up, because now there are more officers looking to make arrests. An example of trying to make yourself useful.

Random Forest

Po1 is also shown the most important factor for the Random Forest as seen in the dotchart. The predictors look to be grouped in 3 clusters of importance. It would be interesting to limit the random forest to 1, or 2 clusters and compare the results to the original model.

Question 10.2

Describe a situation or problem from your job, everyday life, current events, etc., for which a logistic regression model would be appropriate. List some (up to 5) predictors that you might use.

Response

Logistic regression can be used to determine how systems work and to predict future outcomes. There are many situations when this type of model would be appropriate. A company could use it to determine if there is work place bias in how they promote staff. Often times companies keep data on their employees these data could be used to create a logistic model to determine a person's probability of promotoion. Some possible predictors are:

1. Race
2. Sex
3. Hours worked (including overtime)
4. Educational achievement
5. Training events attended

Question 10.3

1. Using the GermanCredit data set germancredit.txt from http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german / (description at http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29), use logistic regression to find a good predictive model for whether credit applicants are good credit risks or not. Show your model (factors used and their coefficients), the software output, and the quality of fit. You can use the glm function in R. To get a logistic regression (logit) model on data where the response is either zero or one, use family=binomial(link="logit") in your glm function call.
2. Because the model gives a result between 0 and 1, it requires setting a threshold probability to separate between "good" and "bad" answers. In this data set, they estimate that incorrectly identifying a bad

customer as good, is 5 times worse than incorrectly classifying a good customer as bad. Determine a good threshold probability based on your model.

Goals

1. Use logistic regression to find a good predictive model for whether credit applicants are good credit risks or not
2. Show your model (factors used and their coefficients), the software output, and the quality of fit
3. Determine a good threshold probability based on your model.

Logistic Regression

Method

1. Use `glm` package to create an initial model
2. Make predictor limited model based on p-values<=0.05

Initial model Summary

```
##
## Call:
## glm(formula = V21 ~ ., family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.526  -0.676   0.337   0.663   2.615
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.69e+00   1.22e+00   -1.38  0.16810
## V1A12        3.68e-01   2.43e-01    1.52  0.12964
## V1A13        9.31e-01   4.15e-01    2.25  0.02473 *
## V1A14        1.86e+00   2.65e-01    7.02  2.2e-12 ***
## V2          -3.21e-02   1.00e-02   -3.20  0.00137 **
## V3A31       -1.04e-01   5.91e-01   -0.18  0.86013
## V3A32        7.09e-01   4.64e-01    1.53  0.12652
## V3A33        1.04e+00   5.09e-01    2.04  0.04090 *
## V3A34        1.34e+00   4.67e-01    2.86  0.00422 **
## V4A41        1.88e+00   4.21e-01    4.48  7.6e-06 ***
## V4A410       1.97e+00   8.68e-01    2.27  0.02321 *
## V4A42        1.10e+00   2.97e-01    3.71  0.00021 ***
## V4A43        9.72e-01   2.72e-01    3.58  0.00035 ***
## V4A44        7.66e-01   8.32e-01    0.92  0.35732
## V4A45        7.43e-01   6.15e-01    1.21  0.22694
## V4A46       -3.72e-02   4.29e-01   -0.09  0.93080
## V4A48        1.55e+01   4.27e+02    0.04  0.97101
## V4A49        9.02e-01   3.67e-01    2.45  0.01411 *
## V5          -1.06e-04   4.70e-05   -2.26  0.02370 *
## V6A62        5.02e-01   3.25e-01    1.54  0.12242
## V6A63        5.40e-01   4.62e-01    1.17  0.24214
## V6A64        2.37e+00   7.18e-01    3.30  0.00098 ***
## V6A65        1.02e+00   2.86e-01    3.59  0.00034 ***
## V7A72        4.52e-01   4.92e-01    0.92  0.35903
```

```
## V7A73        6.54e-01   4.73e-01    1.38  0.16654
## V7A74        1.18e+00   5.12e-01    2.31  0.02099 *
## V7A75        3.93e-01   4.68e-01    0.84  0.39996
## V8          -3.01e-01   9.77e-02   -3.08  0.00204 **
## V9A92        5.39e-01   4.24e-01    1.27  0.20363
## V9A93        1.07e+00   4.18e-01    2.56  0.01049 *
## V9A94        7.44e-01   5.06e-01    1.47  0.14136
## V10A102     -5.56e-01   4.39e-01   -1.27  0.20581
## V10A103      1.14e+00   4.72e-01    2.41  0.01592 *
## V11          5.97e-02   9.62e-02    0.62  0.53485
## V12A122     -1.12e-01   2.81e-01   -0.40  0.68922
## V12A123     -1.57e-01   2.62e-01   -0.60  0.54853
## V12A124     -7.82e-01   4.62e-01   -1.69  0.09050 .
## V13          2.91e-02   1.06e-02    2.75  0.00594 **
## V14A142      5.72e-02   4.48e-01    0.13  0.89834
## V14A143      7.32e-01   2.62e-01    2.79  0.00525 **
## V15A152      4.88e-01   2.56e-01    1.91  0.05666 .
## V15A153      6.01e-01   5.22e-01    1.15  0.24941
## V16         -1.23e-01   2.14e-01   -0.57  0.56593
## V17A172     -9.16e-01   7.58e-01   -1.21  0.22681
## V17A173     -8.30e-01   7.33e-01   -1.13  0.25764
## V17A174     -6.57e-01   7.31e-01   -0.90  0.36840
## V18         -4.63e-01   2.76e-01   -1.68  0.09308 .
## V19A192     -5.59e-02   2.22e-01   -0.25  0.80146
## V20A202      1.54e+00   7.10e-01    2.18  0.02955 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1046.80  on 849  degrees of freedom
## Residual deviance:  733.48  on 801  degrees of freedom
## AIC: 831.5
##
## Number of Fisher Scoring iterations: 14
```

P-value limited model Summary

```
##
## Call:
## glm(formula = V21 ~ ., family = binomial(link = "logit"), data = new_train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.793  -0.865   0.407   0.769   2.067
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  9.96e-01   3.50e-01    2.84  0.00448 **
## A13TRUE      7.07e-01   3.68e-01    1.92  0.05479 .
## A14TRUE      1.76e+00   2.15e-01    8.18  2.9e-16 ***
## A34TRUE      7.52e-01   2.11e-01    3.57  0.00036 ***
## A41TRUE      1.52e+00   3.71e-01    4.08  4.4e-05 ***
## A42TRUE      5.49e-01   2.48e-01    2.21  0.02711 *
```

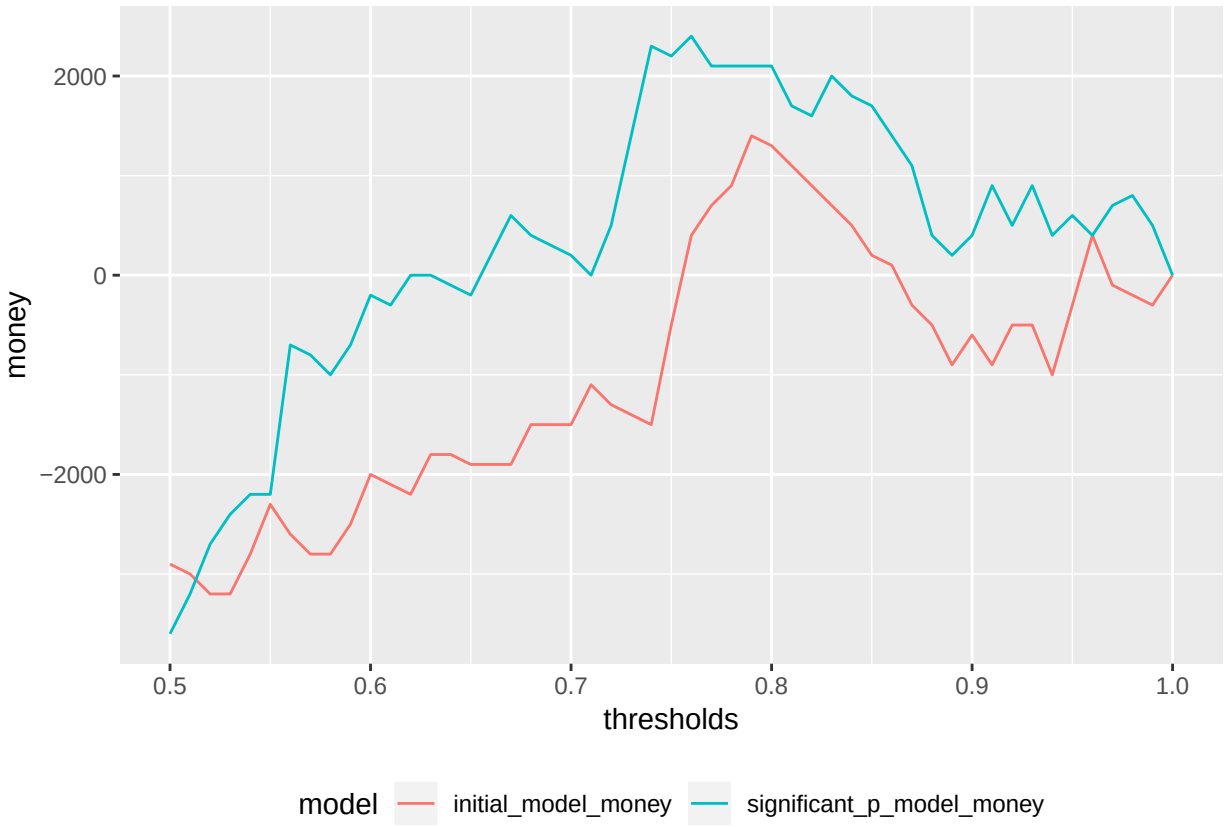
```
## A43TRUE      7.38e-01   2.30e-01    3.21  0.00131 **
## A49TRUE      4.54e-01   3.11e-01    1.46  0.14366
## A64TRUE      1.87e+00   6.33e-01    2.96  0.00311 **
## A65TRUE      1.04e+00   2.60e-01    3.98  6.9e-05 ***
## A74TRUE      6.19e-01   2.51e-01    2.47  0.01356 *
## A103TRUE     1.02e+00   4.32e-01    2.35  0.01861 *
## A124TRUE    -4.04e-01   2.46e-01   -1.65  0.09996 .
## V2          -3.25e-02   9.37e-03   -3.47  0.00052 ***
## V5          -7.47e-05   4.18e-05   -1.79  0.07375 .
## V8          -2.38e-01   8.80e-02   -2.70  0.00690 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1046.80  on 849  degrees of freedom
## Residual deviance:  808.96  on 834  degrees of freedom
## AIC: 841
##
## Number of Fisher Scoring iterations: 5
```

The initial logistic regression model has an AIC of 831.481 and used all predictors. The extra coefficients are dummy variables created automatically by the glm package, which creates n-1 dummy variables for factors of length n. The p-value limited predictor model has an AIC of 840.962 and used predictors of the previous model with p-value<=0.05.

Determine a good threshold probability based on your model

Method

1. Create a function that determines money made/lost for a given threshold
2. Compare both models
3. Choose model that makes the most money



Although the initial model had a lower AIC, the p-value limited model performed better as seen in the plot above and the table below.

	initial model	p-value limited model
AIC	831.481	840.962
Max Money	1400	2400
Threshold	0.79	0.76

I choose the p-value limited model with a threshold of 0.76 because:

1. It makes more money from the given dataset
2. It uses a lower threshold for a decision surface

It is significant to note that the second model made more money at a lower threshold. This is a second positive because you are not turning down clients that have the potential to bring future business. It would be good to cross validate this dataset, but I do not have time.