

Predicting Alumni Major Donors

Rong, Luke, n' Brenton

12/7/18

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

The Data

We have data on all Umass Amherst alumni from the past century with information detailing a past giving history, their school spirit or participation in extracurricular activities, involvement with UMass as alumni, and many socio-economic attributes, $n = 264837$.

Question of Interest

A major donation is a considered a single donation of \$25,000 or more.

- **We would like to determine candidates who are likely to make a major donation in the next 5 years.** Model determines characteristics of likely major donors based on major donors in the data set.
- These false positives, or “candidates” are people who are not major donors but are similar enough to major donors that the model falsely labels them that way. We consider them the most likely people to become major donors in the near future.

This has a new trendy name called
Lookalike Modeling!



Variable Selection

- Only variables which could be determined to have occurred previous to major donation were considered
 - Random forest was used to aid in variable selection
 - Permutation Feature Importance to decide which variables to include in GAM fit
1. Fit a random forest to the data
 2. Predict on the data, measure prediction accuracy
 3. Permute the values for predictor, refit the random forest
 4. Measure the mean decrease in prediction accuracy with that one variable permuted
 5. Positive score: large mean decrease in accuracy indicates predictor which adds value to the model
 6. Negative score: indicate the permuted values are predicting better than the true values for the variable

Variables

Response

- Major Donor Indicator = 0,1
 - Has a donor made a one time donation of \$25,000 or more
 - 99.7% Non major donors vs 0.3% Major donors

Binary Variables

- Undergraduate student club member indicator
- Greek life indicator
- Student Athlete indicator
- Graduate school school indicator

Variables

Factors [0,1,2]

Date recorded in time span of interest, where this span of interest is the five years prior to major donation for major donors, or the five years prior to the last time data were updated for non-major donors. Three level factor where we code "0" for those who do not have a date prior to date of interest, date added within our five year span of interest as "2", and recorded prior to span of interest as "1".

- Child indicator
- Job Indicator
- C-suite indicator (CEO, CFO, etc)

Variables

Other Factors

- School indicator (CNS, Engineering College etc.) - Multi-level Factor
- Years of consecutive giving, Multi-Level Factor [0,1,2,3,4,5]
 - This is evaluated within the time span of interest

Spatial Statistics

- Latitude & Longitude of home (when available) or business
- Binary indicator when Latitude and Longitude were missing

Variables

Continuous Variables

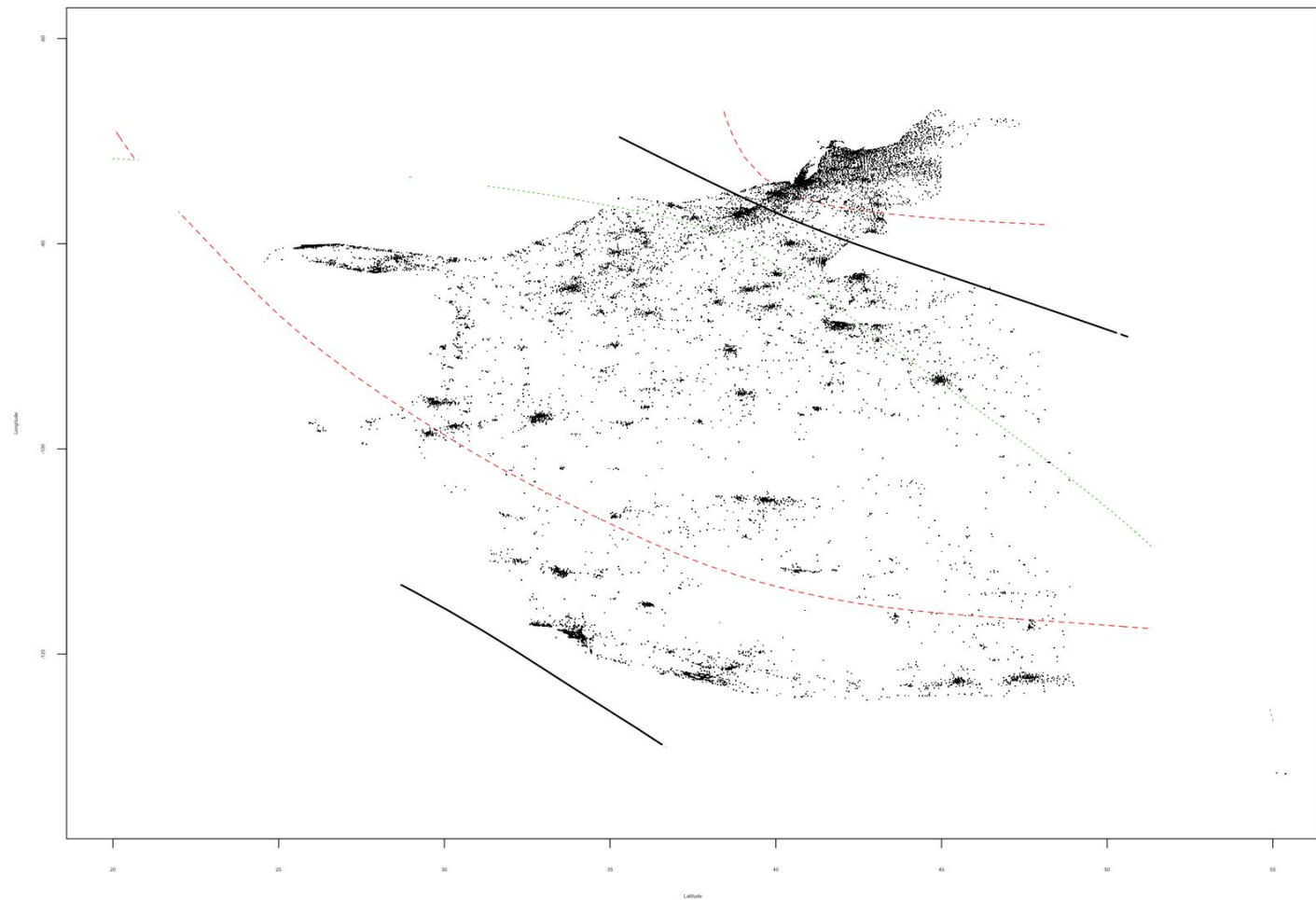
- Giving ratio
 - Calculated in five year span of interest
 - Ratio > 1 meaning someone's giving increased
- Average Donation in five year span of interest
- Age at first major donation for major donors or current age for non-major donors
- Count of Umass Events Attended as alumni previous to major donation, current total for non major donors
- Count of Alumni Clubs or groups someone were a part of previous to major donation, current total for non major donors

Model

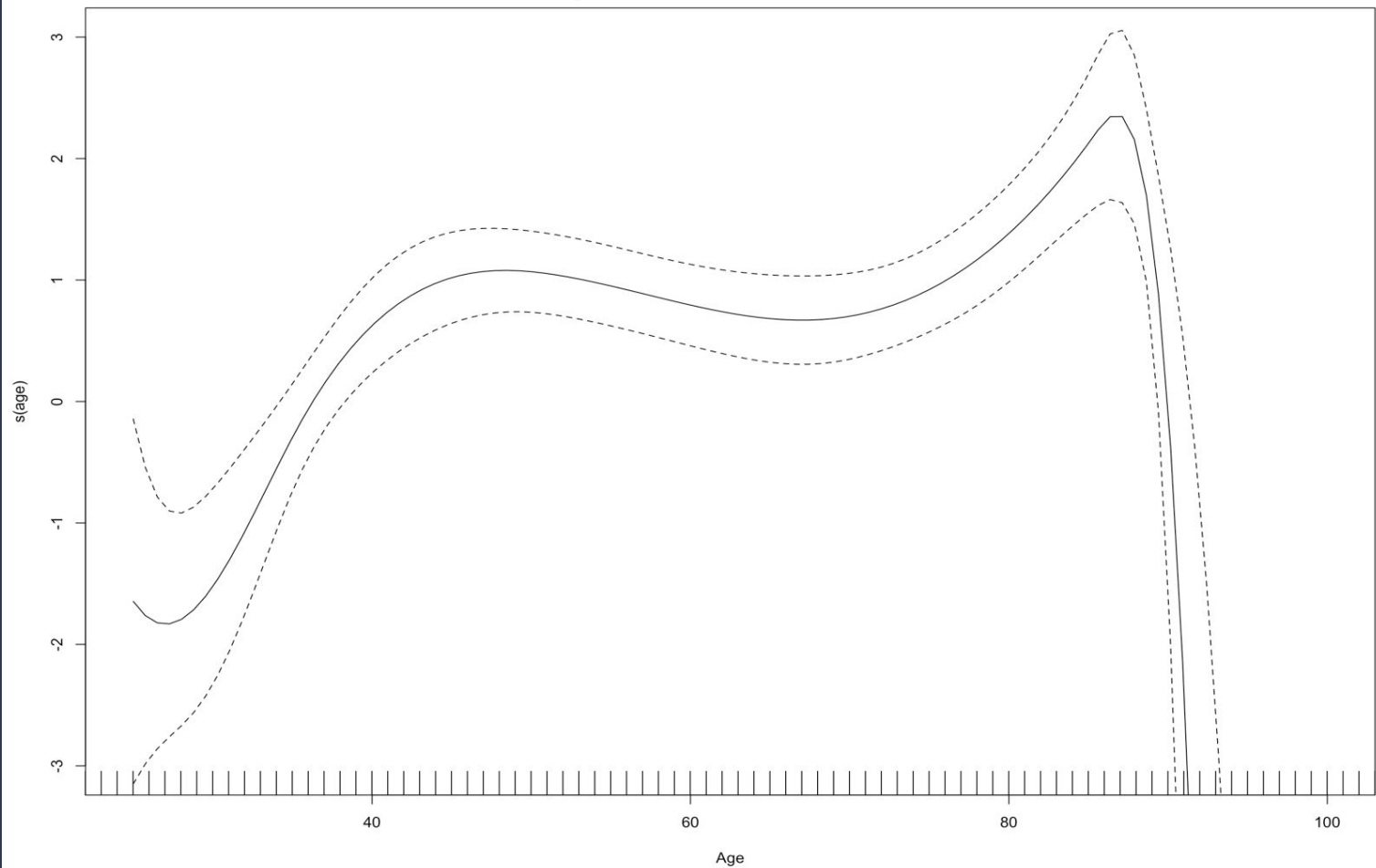
$$\text{Logit}(\text{Pr}(Y = 1)) = \beta_o + \beta_1 X_{club} + \beta_2 X_{Greek} + \beta_3 X_{athlete} + \beta_4 X_{grad} + \beta_5 X_{location_{NA}} + \\ factor_{child} + factor_{job} + factor_{executive} + factor_{school} + factor_{consecutive_giving} + \\ f_1(X_{age}) + f_2(X_{ratio}) + f_3(X_{events}) + f_4(X_{clubs}) + f_5(X_{Avgdonation}) + f_6(X_{LAT}, X_{LONG})$$

Fit in BAM

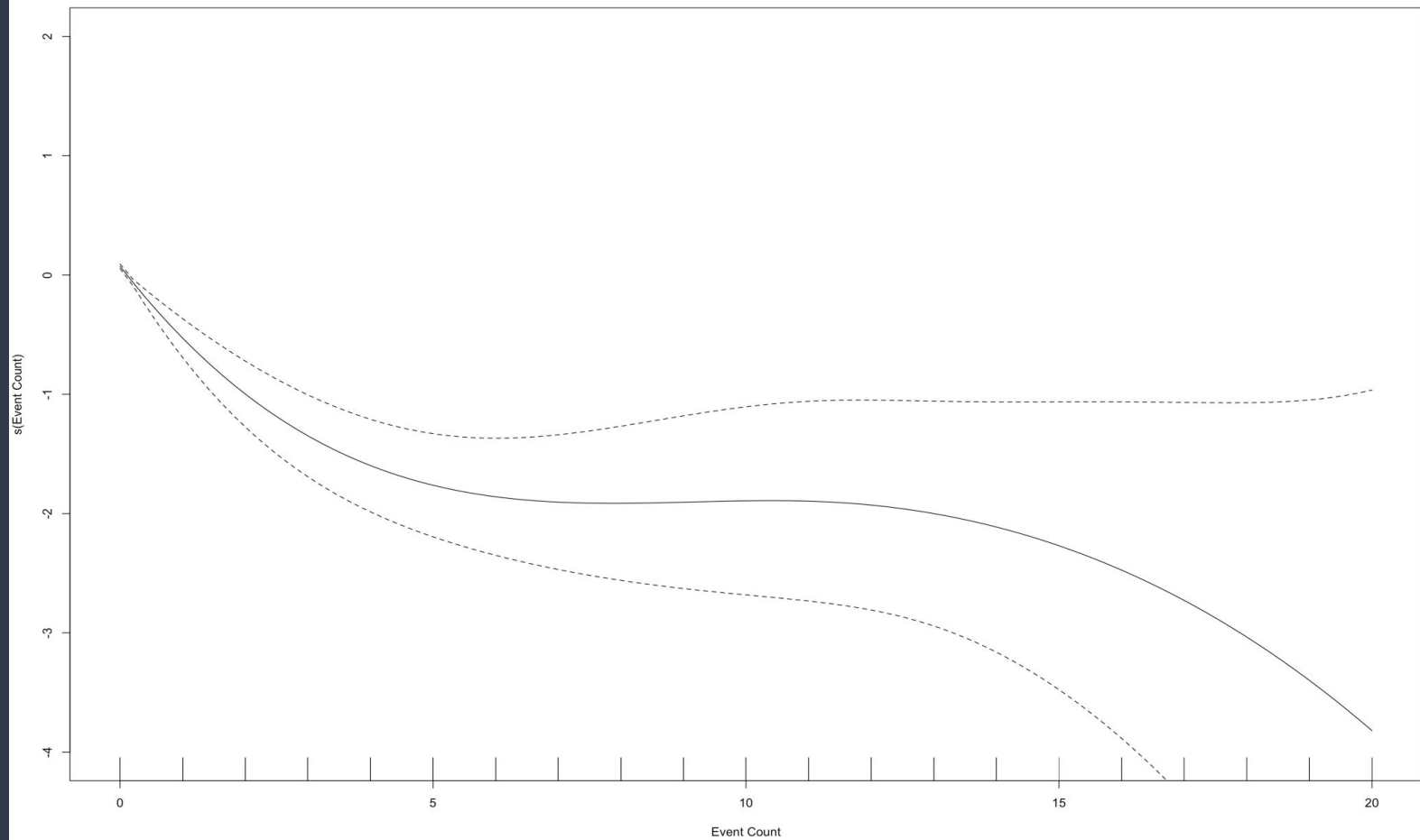
- Penalized p-splines for continuous variables combine a B-spline basis with a discrete penalty on the basis coefficients
- Thin plate regression splines for the spatial statistics: Longitude and Latitude
 - “by” was used to only include this term when we had valid coordinates
 - When missing this term contributes zero to the linear predictor from its smooth
 - When missing a binary indicator was implemented
 - 14% of major donors were missing coordinates, vs only 7% of non major donors
- “REML” to estimate parametric portions of the model



Age Smooth Term Model Contribution




Event Count Term Model Contribution




Evaluating Model Performance

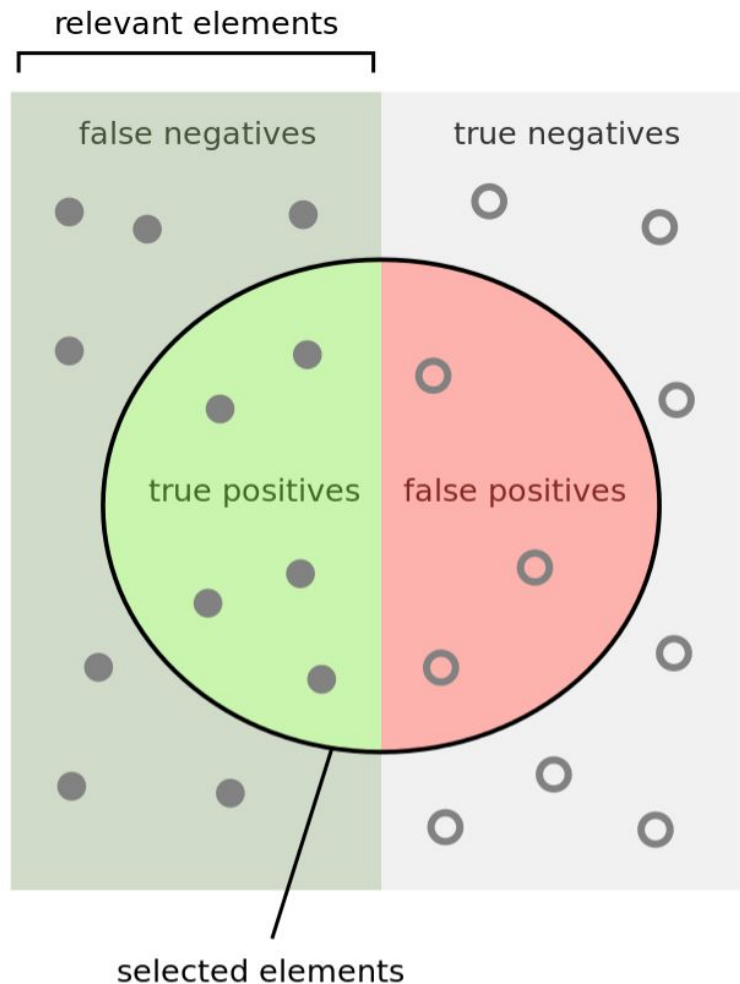
The F1 Score is a measure of accuracy for binary classification that accounts for both precision p and recall r . It is the harmonic mean of p and r .

How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$


How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$




GAM vs Logistic Regression

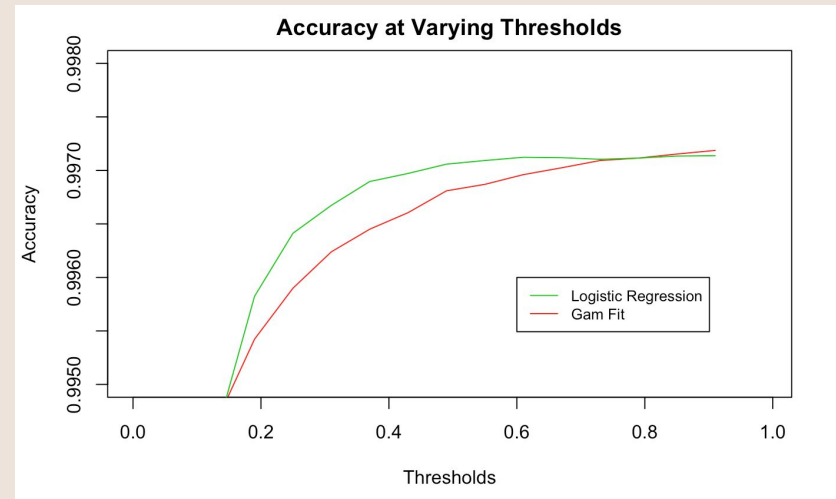
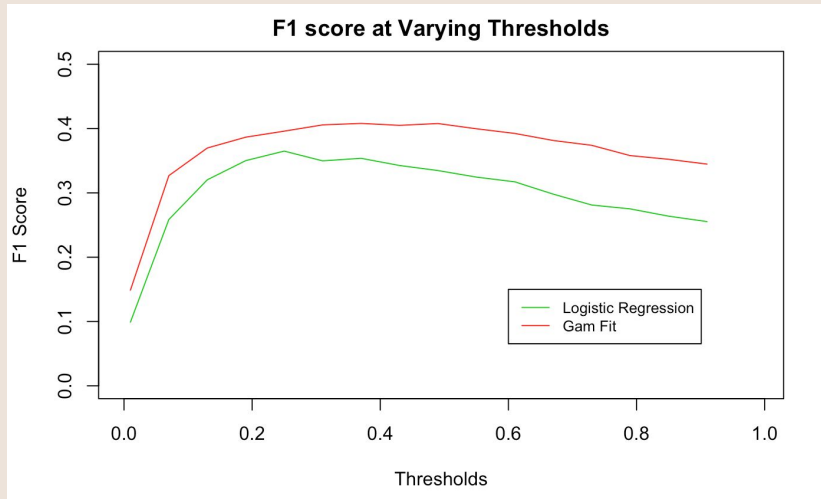
Average 5-Fold Cross Validated Metrics

GAM Model:

Overall Deviance Explained: 42.8%

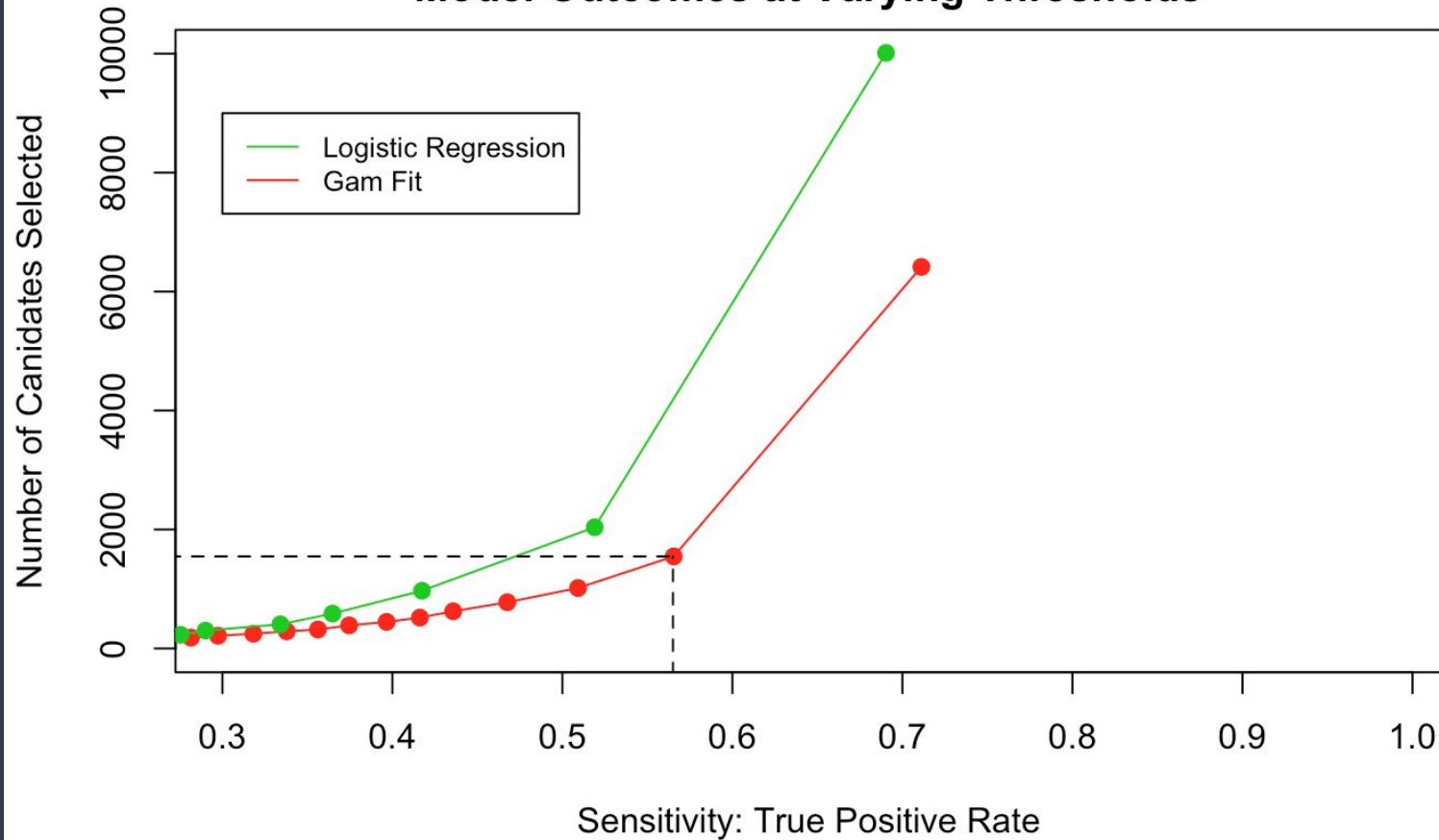
Logistic Regression:

Overall Deviance Explained: 36.5%



Evaluating on accuracy does not indicate how well our model performs due to class imbalance, while F1 score gives more weighting to true positives, while still penalizing for false positives.

Model Outcomes at Varying Thresholds



GAM Confusion Matrix

Truth \ Model	0	1
	0	1
0	262470 (TN)	1548 (FP) Candidates!
1	355 (FN)	462 (TP)

Threshold = .1

F1 Score = .328

Accuracy = .993

MC Simulation

```
func_sphere <- function(la,lo) { ##  
  result<-sin(lo)*cos(la-.3)  
  result<-result-mean(result)  
  return (result)  
}  
n=10000  
lo <- runif(n)*2*pi-pi ## longitude  
la <- runif(3*n)*pi-pi/2  
ind <- runif(3*n)<=cos(la)  
la <- la[ind];  
la <- la[1:n]  
la = la *180/pi  
lo = lo *180/pi  
sphere_dat <- func_sphere(la,lo)
```

- Simulation was used to determine how well/if the gam model we've built can detect encoded relationships between various predictors and Y.
- Latitudes and longitudes were simulated on 3-D surface
- Over 100 runs we simulate data as well as several other covariates imitating age, number of school events attended, number of clubs.
- We also add two factors to the model, imitating child indicator and job indicator.

Generate Data

```
```{r}
int = -6
Z = sphere_dat+age_dat+events_dat+
 clubs_dat+est_child+est_job+int
pr = exp(Z)/(1+exp(Z))
y = rbinom(n,1,pr)
table(y)
```
```

First we randomly generate x 's and input those into the “true” functions to generate Z 's.

The true functions are centered so their overall mean is 0, removing the effects of the intercept.

We then apply the sigmoid function to the Z 's to transform the Z values into probabilities p .

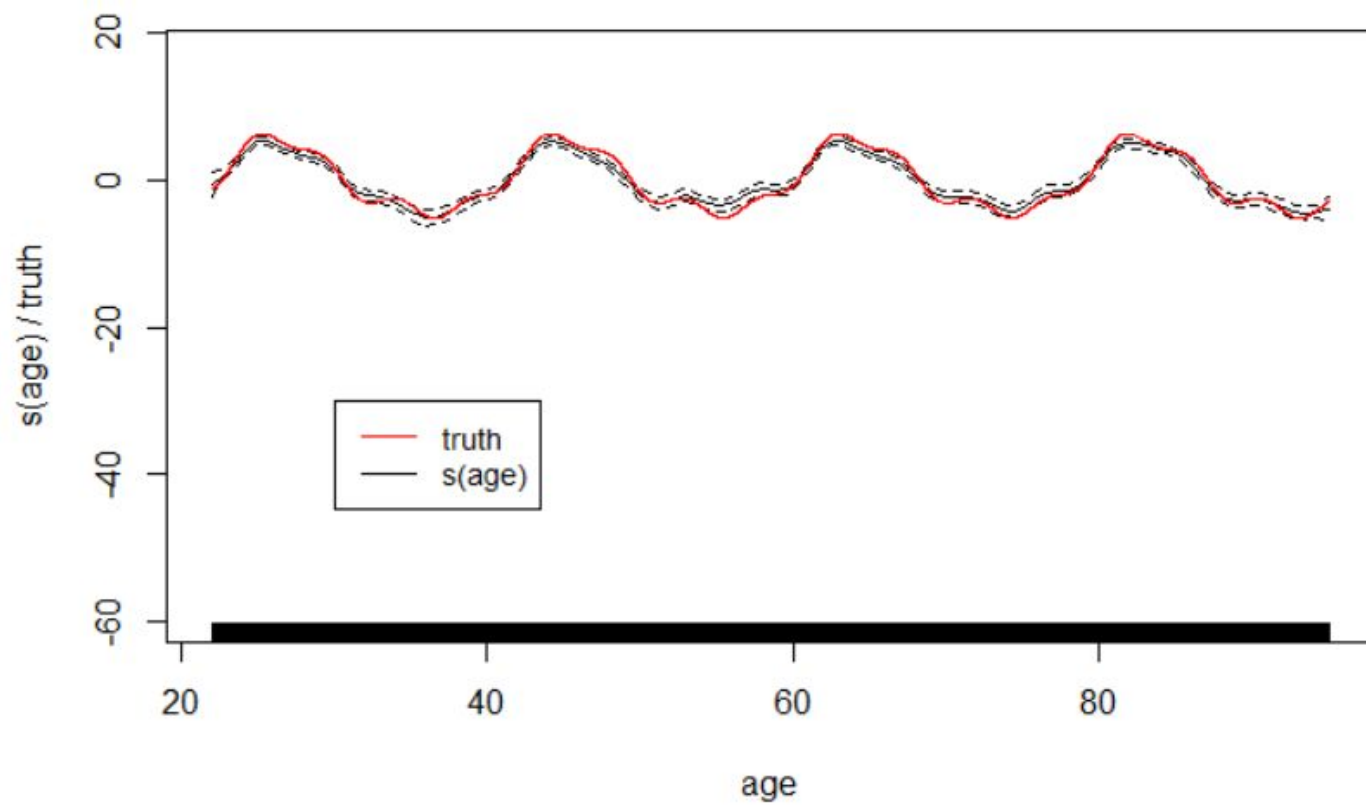
← Here is our way to randomize y 's.

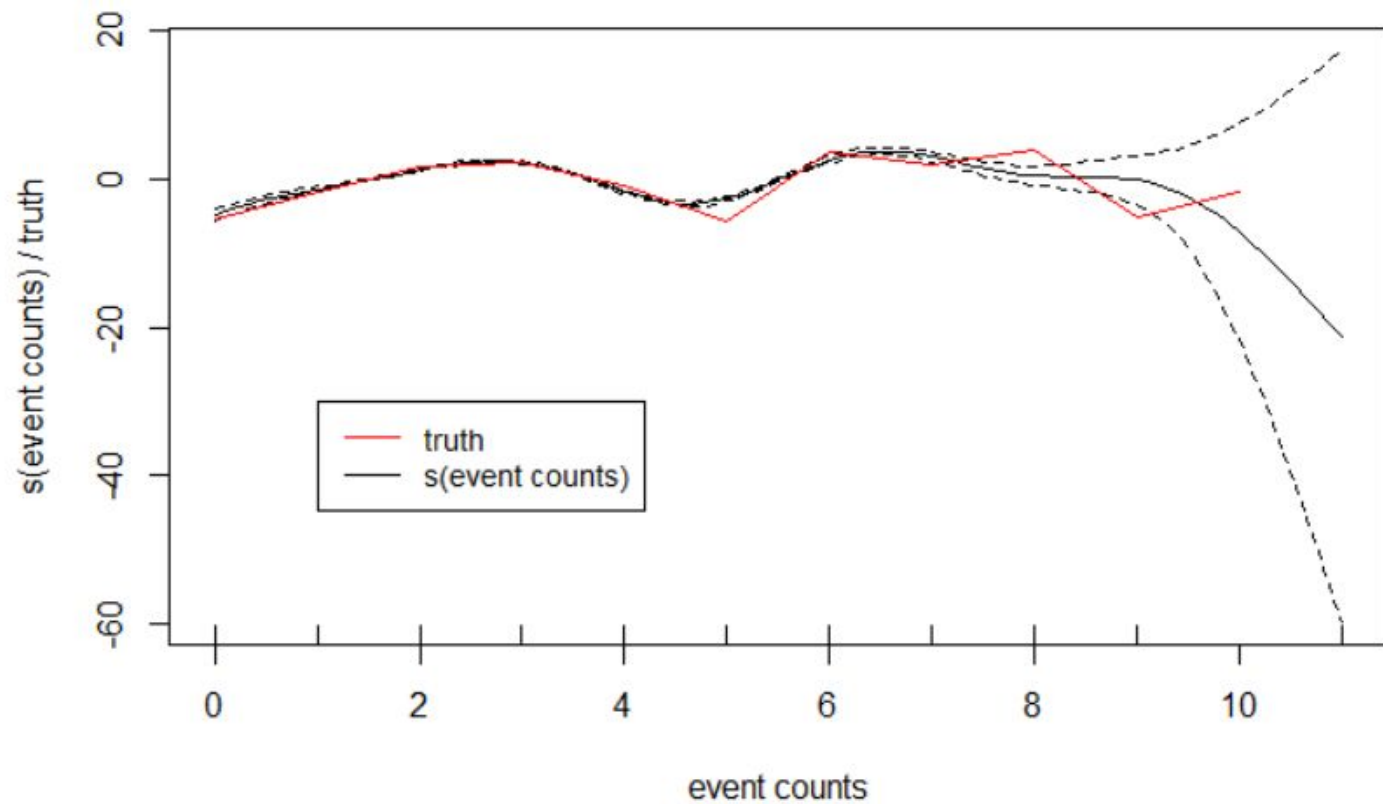
We generate y 's in binomial distributions with probability $P(Y=1)=pr$

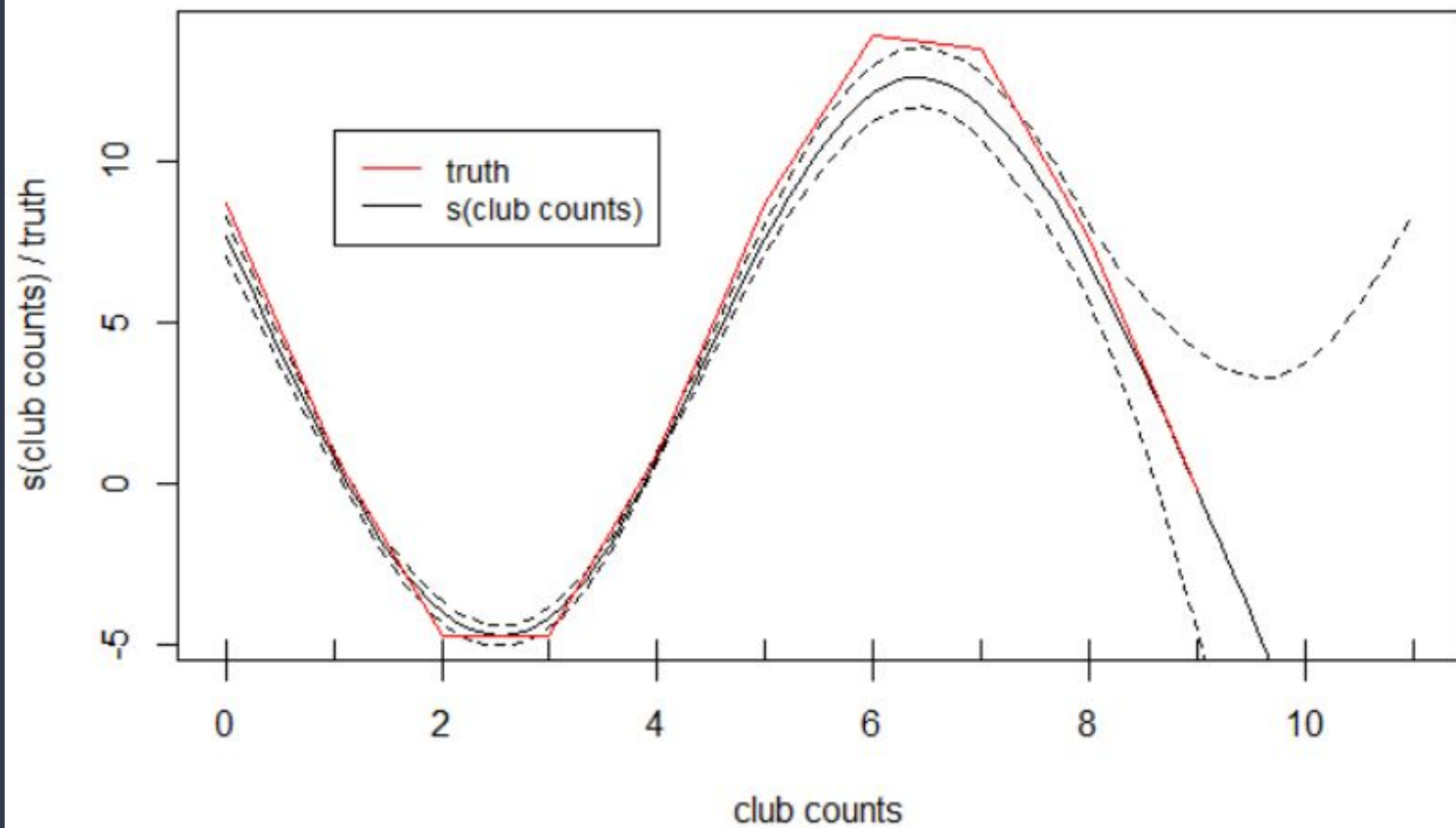
In this way, we add noises to our true functions.

Simulation Accuracy

- We split the dataset into training set and test set with 10% of the simulated data in the test set, leaving 90% of the data for each iteration to be fit on ($n=10,000$).
- We fit the training data with GAM and predict on the test data.
- For the bivariate function, we use Mean-Squared Error to check the accuracy of the estimate from gam.
- For the univariate functions, we plot the graphs of the true functions and compare with the plots produced by GAM.
- In order to evaluate the overall performance of the model, we compute the accuracy and the F1-score of the model.







Example Simulated Confusion Matrix

| <div>Model
Truth</div> | 0 | 1 |
|----------------------------|-------------|-------------|
| 0 | 767
(TN) | 22
(FP) |
| 1 | 34
(FN) | 178
(TP) |

Simulation Results

- Mean MSE for thin plate spline estimates = 0.256
 - The results of the true function of latitude and longitude typically range from -1 to 1.
- Average accuracy of model = 93.6%
- Average F1 score = 84.6%

THE END

