

# Machine Learning in Advertising Technology

Vangie Shue

December 10, 2014

## Abstract

This project will explore the application of Machine Learning in Technology. Specifically, we will seek to develop a reliable process for developing a classification algorithm on given data.

## Introduction

Background on Advertising Technology use cases. Background on Cafemom and the specific problem: classify hispanic users

## Classifying Rare Events

(1, pg141) Rare events are more statistically informative than zeros is seen in the variance matrix (1, p142) When sampling, we must be careful not to select on X differently for the two samples. (2) The problem is that maximum likelihood estimation of the logistic method is well-known to suffer from small-sample bias. The penalized likelihood or Firth method are the general approach to reducing small-sample bias.

## Classification Algorithms

### Logistic Regression

(4) glmnet

## Methods

Demdex has Traits/Segments, uuid and traits/segments collected Table generating en/es language to determine hispanic/non-hispanic, 1 for es, 0 for en

- segment\_hispanic: 80324851 total uuids
- segment\_hispanic: 1698878 hispanic (2 percent)
- segment\_hispanic2: 9733751 total uuids
- segment\_hispanic2: 1868091 hispanic (19 percent)

## References

We used RStudio Sweave to build this L<sup>A</sup>T<sub>E</sub>X document

- 1 <http://gking.harvard.edu/files/gking/files/0s.pdf> (rare events)
- 2 <http://www.statisticalhorizons.com/logistic-regression-for-rare-events> (rare events)
- 4 <http://cran.r-project.org/web/packages/glmnet/glmnet.pdf>
- 3 <http://stats.stackexchange.com/questions/25389/obtaining-predicted-values-y-1-or-0-from-a-logistic-regression-model-fit>

## Logistic Regression

We used `glmnet` to generate logistic regression model. Then `predict.glm` is used to predict on the validation sample. (3) We then determine the how low of a predicted probability is needed to accurately classify the hispanic division.

```
> data(example)
```

The above is a snippet of code used.

## Results

The below is a sample graph of data.

Something like a plot centered.

## Conclusion

We demonstrated the application of Machine Learning in Advertising Technology, in particular for rare events.

Cannot accept comparisons without consideration to the implementation. Some may provide more "tuning" than other algorithms and therefore appear more accurate. However, we demonstrate which algorithm will work best for our usage.

## Acknowledgements

Patrick McCann & Cafemom, Professor Mohri