Internet Image Ad Classification

A Predictive Model

Jiying Zou

Introduction

Can an internet image be automatically identified as an ad based on observations? Such a question is important in today's rapidly expanding marketing-related machine learning and algorithm development fields, for example in the creation of ad-blocker software. This report, through data investigation and regression methods, aims to build and compare various models, focusing on predictive power, classifying an image as an ad or not. Models examined are picked by principal component analysis (PCA), binomial LASSO regression, and combinations of the two. The model selected by LASSO surpasses the other models in performance, yielding about 95% cross validated prediction accuracy.

Data Description

This dataset consists of 1558 measured explanatory variables for 3279 internet images, of which 458 are ads and 2821 are not. The explanatory variables contain 3 continuous and 1555 Bernoulli variables that take on 0/1 values. The continuous variables contain the image height, width, and aspect ratio measurements, but no details are known about what the Bernoulli variables represent besides that "1" means presence of and "0" means absence of some characteristic. Only the continuous and first Bernoulli variable have any missing values.

In working with data, I treat the Bernoulli variables as numeric for simplicity. As for the continuous variables, all three are skewed, multimodal, and as seen from Figure 1(a), relatively correlated with one another with severe non-constant variance issues. Some problems are alleviated by transformations: logging the first variable, square root the second, and fourth root the third (Figure 1(b)).

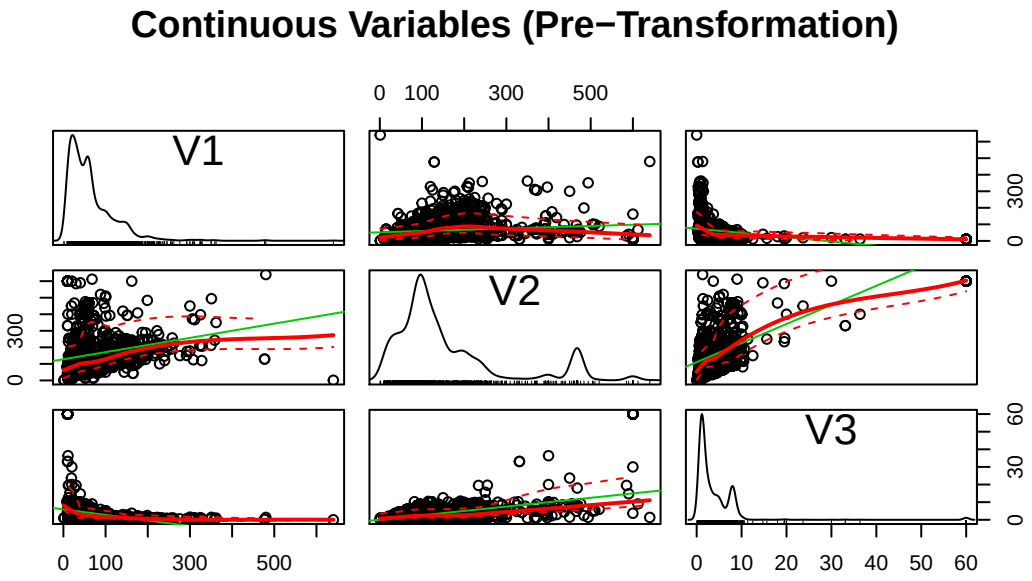

Figure 1(a). Three continuous variables show considerable skew, multimodality, and non-constant variance in their pairs plots.

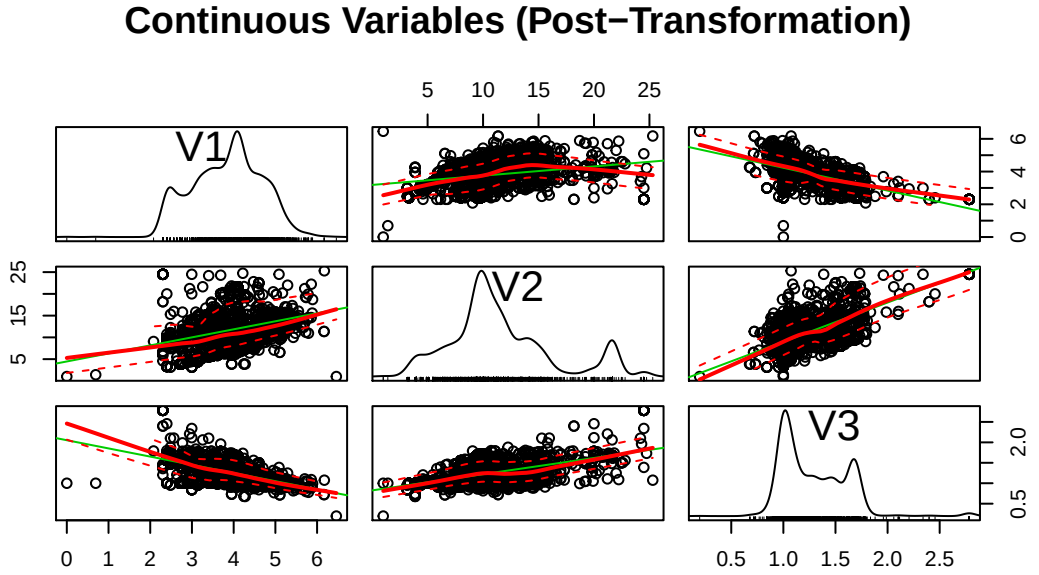

Figure 1(b). After transformations, the variables are much better behaved – some skew is fixed, and errors are a lot more consistent.

However, Figure 2 hints that non-ads are overrepresented in the missing data. Out of all cases, nearly 30% of non-ads yet only around 15% of ads lack all three continuous values. Thus, without information on missing value explanations, I remove the continuous variables for fear that they might skew the predictions later on.

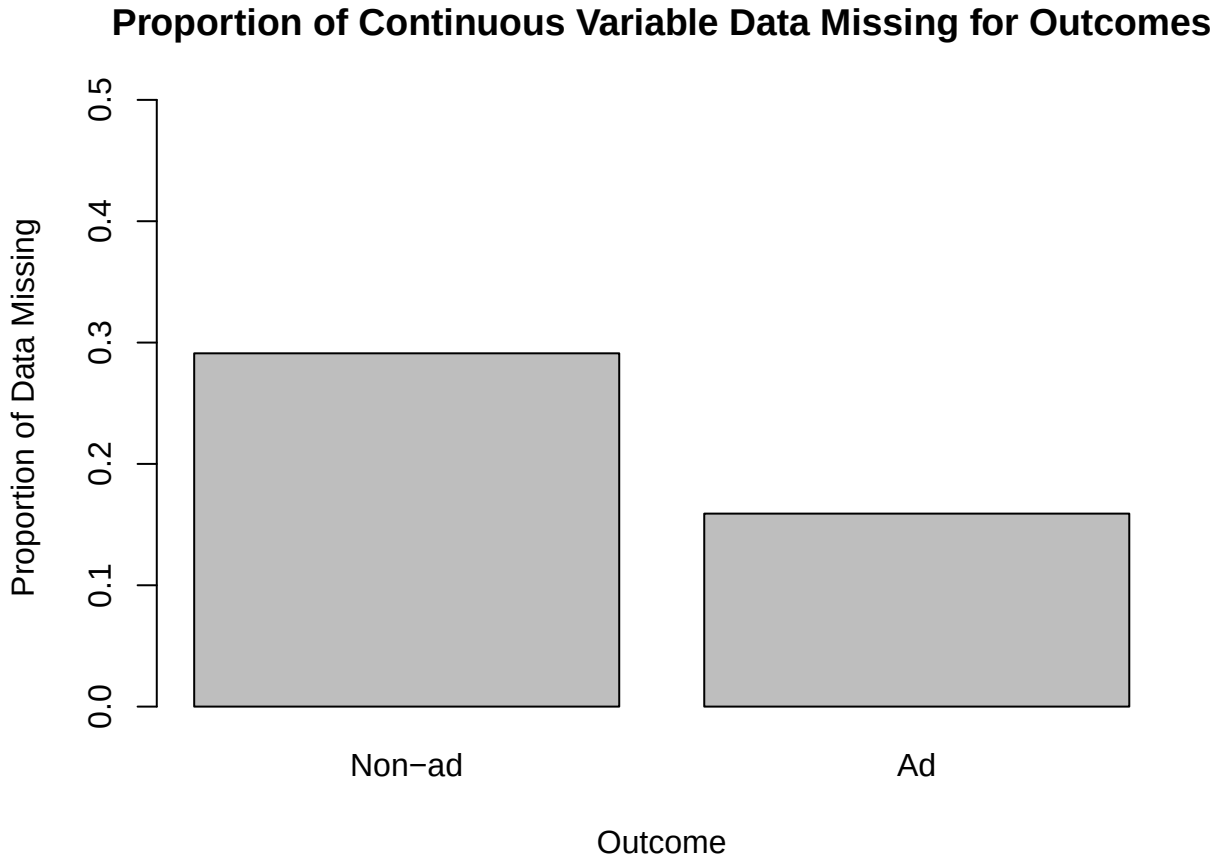

Figure 2. Around 30% of non-ad cases having missing continuous variable data, while only around 15% of ad cases are missing this information.

For the remaining Bernoulli variables, I remove 795 duplicated columns to avoid perfect collinearity, and weed out the leftover 15 cases containing missing values, which can be safely ignored in the grand scope of 3000+ other observations. Given that the remaining 760 variables consist mostly of zeroes, collinearity is still a crucial problem in model building and selection.

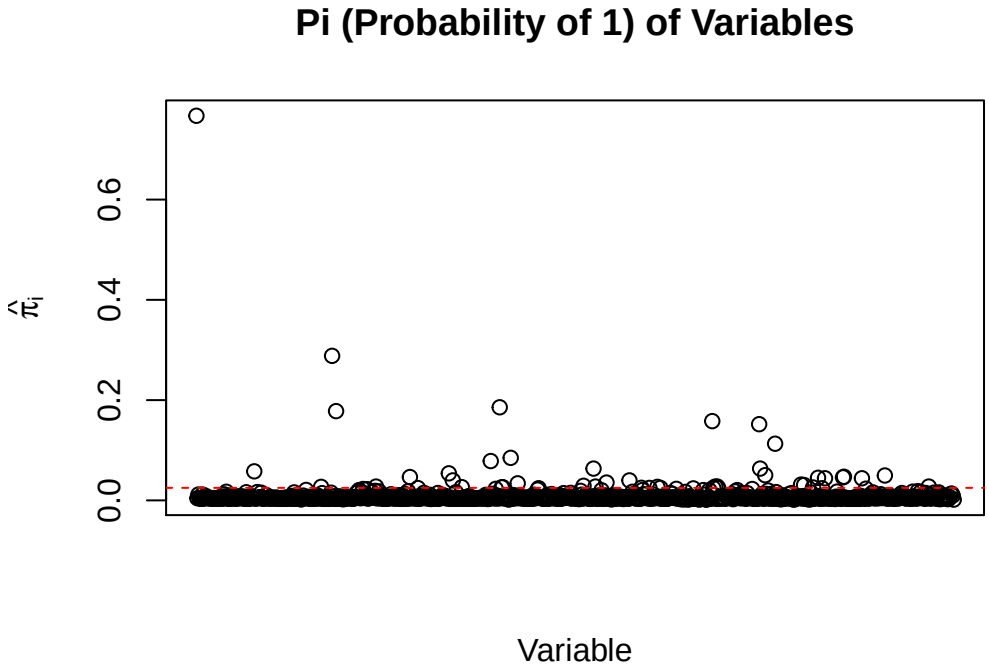

Figure 3. Only a couple Bernoulli (0/1) variables contain more than 2.5% of "1"'s; most explanatory variables contian mostly "0"'s.

Finally, the rows are shuffled randomly to eliminate any patterns, and the first 250 cases are set aside as the test set.

Main Results

Under the motivation of higher predictive power, I chose two main tools, PCA and binomial LASSO, to build four models with. The first model solely uses PCA to reduce collinearity, and the second follows up the results from the first with binomial LASSO regression to further reduce dimensionality. The order is then reversed, and the third model uses binomial LASSO by itself while the fourth model follows up the third with PCA. Binomial LASSO is implemented using the `glmnet` package, and the optimal penalty term is chosen by 5-fold cross validation based on AUC error.

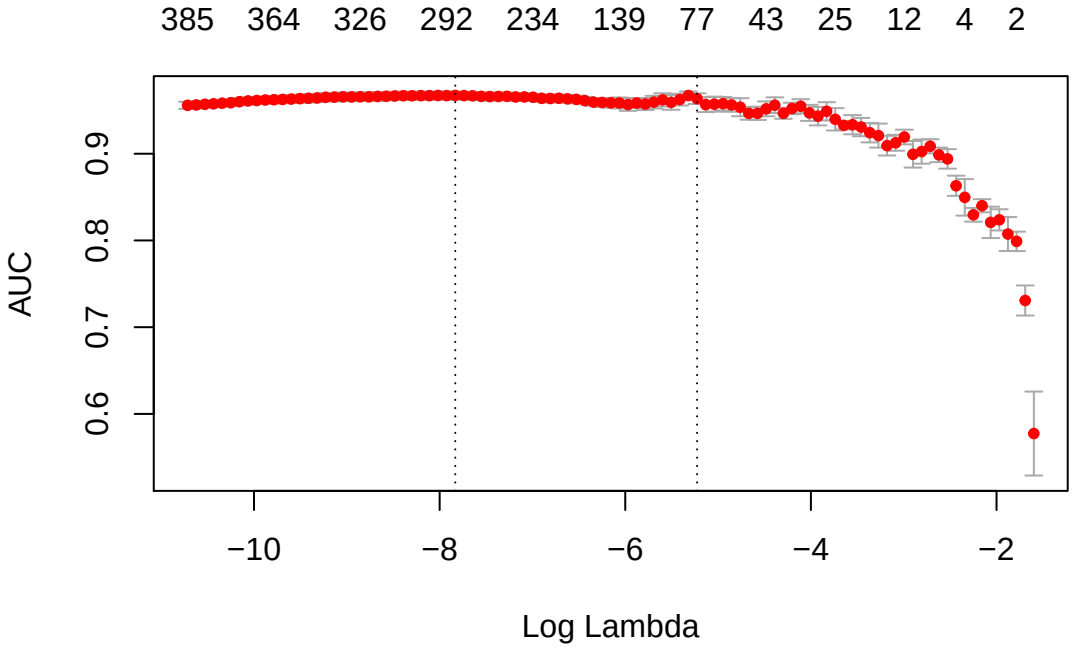

Figure 4. The optimal L1-penalty term λ is selected through AUC error, and resides at the highest point between these lines.

LASSO regression has the effect of driving some coefficients to zero:

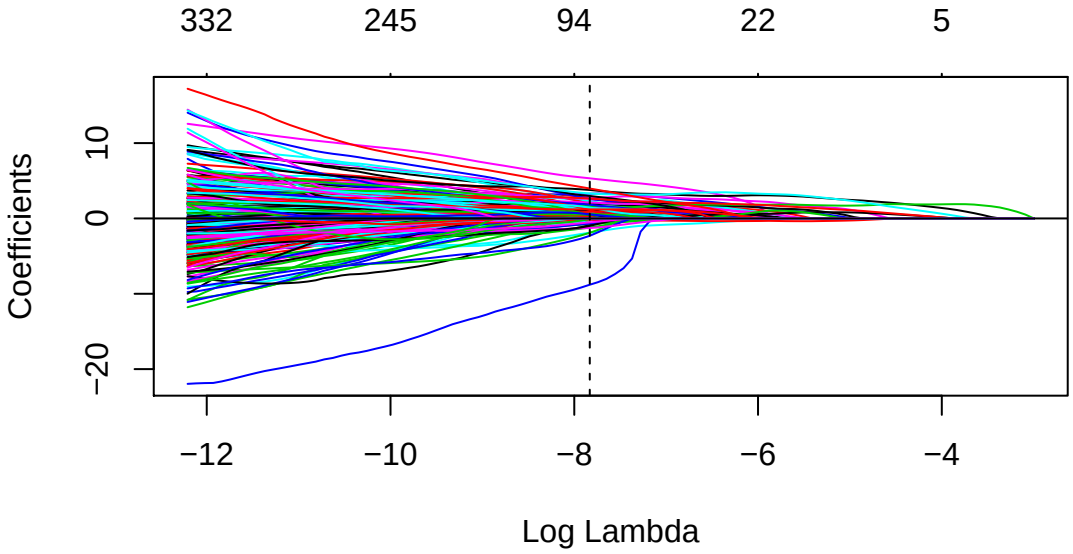

Figure 5. Each colored line represents a different coefficient's value. As the penalty increases, some coefficients are driven to zero, thus reducing dimensionality.

The resulting variables kept in each model are then extracted and fitted with coefficients in a general linear model (if no LASSO-provided coefficients are available), which is used to predict on the test set. The models

contain 43, 27, 316, and 13 variables, respectively. (These numbers may alter slightly during compilation, depending on the train-test arbitrary split.) Root-mean-squared prediciton error (RMSE) is the criteria used to assess model competency.

```
##        Method      RMSE
## 1    PCA only 0.1781704
## 2 PCA + LASSO 0.1996347
## 3  LASSO only 0.1699874
## 4 LASSO + PCA 0.2349575
```

Table 1. Four different types of models and their respective RMSE errors when predicting on the test set.

The best of the four models is the model with LASSO regression only, which yielded the lowest RMSE and a cross-validated prediction accuracy of around 95%! The variables kept are as follows:

```
## [1] "The variables present in the final model are: "
```

```
##   [1] "V4"    "V5"    "V6"    "V7"    "V9"    "V14"   "V19"   "V20"
##   [9] "V23"   "V29"   "V32"   "V33"   "V34"   "V35"   "V36"   "V37"
##  [17] "V38"   "V40"   "V43"   "V44"   "V48"   "V51"   "V52"   "V53"
##  [25] "V54"   "V57"   "V61"   "V65"   "V67"   "V68"   "V72"   "V75"
##  [33] "V76"   "V80"   "V82"   "V91"   "V95"   "V96"   "V104"  "V105"
##  [41] "V109"  "V114"  "V121"  "V126"  "V137"  "V146"  "V150"  "V156"
##  [49] "V167"  "V168"  "V172"  "V173"  "V175"  "V178"  "V181"  "V184"
##  [57] "V187"  "V188"  "V192"  "V195"  "V196"  "V200"  "V202"  "V203"
##  [65] "V208"  "V214"  "V218"  "V222"  "V229"  "V232"  "V236"  "V243"
##  [73] "V244"  "V247"  "V251"  "V265"  "V266"  "V283"  "V285"  "V291"
##  [81] "V296"  "V298"  "V303"  "V306"  "V318"  "V341"  "V349"  "V352"
##  [89] "V355"  "V358"  "V363"  "V366"  "V370"  "V371"  "V378"  "V379"
##  [97] "V380"  "V382"  "V388"  "V396"  "V399"  "V400"  "V407"  "V412"
## [105] "V418"  "V419"  "V421"  "V429"  "V436"  "V442"  "V451"  "V457"
## [113] "V458"  "V460"  "V462"  "V464"  "V465"  "V470"  "V476"  "V477"
## [121] "V479"  "V490"  "V497"  "V500"  "V507"  "V509"  "V510"  "V512"
## [129] "V521"  "V526"  "V533"  "V546"  "V549"  "V550"  "V568"  "V574"
## [137] "V581"  "V626"  "V628"  "V637"  "V641"  "V643"  "V646"  "V664"
## [145] "V692"  "V701"  "V702"  "V706"  "V771"  "V794"  "V810"  "V848"
## [153] "V861"  "V882"  "V885"  "V889"  "V893"  "V949"  "V951"  "V955"
## [161] "V959"  "V969"  "V975"  "V982"  "V983"  "V984"  "V986"  "V998"
## [169] "V999"  "V1001" "V1010" "V1012" "V1022" "V1023" "V1033" "V1036"
## [177] "V1044" "V1064" "V1075" "V1078" "V1082" "V1087" "V1088" "V1092"
## [185] "V1095" "V1110" "V1115" "V1117" "V1118" "V1126" "V1133" "V1135"
## [193] "V1141" "V1149" "V1159" "V1167" "V1168" "V1169" "V1173" "V1180"
## [201] "V1181" "V1192" "V1193" "V1195" "V1217" "V1229" "V1230" "V1248"
## [209] "V1254" "V1255" "V1263" "V1264" "V1265" "V1268" "V1272" "V1274"
## [217] "V1292" "V1293" "V1337" "V1342" "V1352" "V1360" "V1364" "V1368"
## [225] "V1377" "V1381" "V1383" "V1386" "V1400" "V1403" "V1414" "V1429"
## [233] "V1435" "V1437" "V1438" "V1440" "V1445" "V1446" "V1451" "V1454"
## [241] "V1455" "V1456" "V1459" "V1460" "V1463" "V1465" "V1468" "V1470"
## [249] "V1471" "V1474" "V1475" "V1478" "V1480" "V1481" "V1483" "V1484"
## [257] "V1485" "V1491" "V1493" "V1494" "V1496" "V1499" "V1503" "V1504"
## [265] "V1505" "V1508" "V1511" "V1515" "V1516" "V1519" "V1520" "V1523"
## [273] "V1529" "V1530" "V1531" "V1532" "V1533" "V1534" "V1537" "V1538"
## [281] "V1543" "V1545" "V1546" "V1548" "V1549" "V1550" "V1556" "V1557"
```

This result is surprising, since I hypothesized that the PCA model would perform the best due to its collinearity reducing properties. LASSO regression aims to minimize the size of the regressor coefficients and operates off of the L_1-penalty, but does not intrinsically regulate collinearity. A possible explanation is that

the PCA model has almost triple the number of variables as the LASSO model does, allowing for a much larger collinearity issue. Condition indexes generated from PCA confirm this idea – a good proportion of condition indexes are greater than the threshold (>10), indicating strong lingering collinearity. In addition, the PCA model yields separability issues causing the `glm()` fit to fail to converge. Separability issues cause estimated coefficients to be large and wildly inaccurate, which then interferes heavily with model stability.

Discussion

Subjectivity of model selection means that no model is "correct"; there are many possible methods beyond those I've implemented. I focus on PCA and LASSO specifically because these methods address collinearity and zero-out coefficients, respectively, effectively reducing dimensionality. Considering the large amount of variables, stepwise methods are computationally expensive for little gain, ridge regression will fail to zero out any coefficients at all, and without a doubt it is unrealistic to perform model selection by hand. Another model I attempted is motivated by the fact that only a handful of Bernoulli explanatory variables have a non-negligible proportion of "1"'s. Referencing back to Figure 3, this model selects variables with more than 2.5% of "1"'s, and uses BIC criteria to select a model from the five top models of each size up to 35 variables. While at first sight this method seems clever, in retrospect the method selects variables without guaranteed importance and fails to address collinearity. Another idea is to create the design matrix from the first few principal component vectors. This has the advantage of eliminating collinearity, since the design matrix will be orthogonal.

My final model is built upon assumptions of reasonable sampling methods, independently sampled images, and correct image judgement and data entry (implying no human error). This model has strong predictive power but weak interpretability, since no variable names are known, and there are a comparatively large number of variables. Although the original data is imbalanced in the amount of ad and non-ad outcomes, the predictive power is still fairly strong, indicating that the model performs well in predicting both outcomes. Predictive power may be improved by considering the three continuous variables, perhaps making up for the large amount of missing data with conditional imputation methods (i.e. fill in missing values from cases with a specific outcome with the mean of all cases with that outcome).

Regarding test set prediction error, RMSE is a more robust measure than percentage of correct predictions because it weighs in each model's probability of obtaining specific predictions and is not based on arbitrary cutoff values from which prediction outcomes are decided. A downside of my error comparisons is that I examined the RMSE of all models, while LASSO may be better compared through mean absolute error (MAE).

Summary

It is possible to predict whether or not an internet image is an advertisement with high accuracy through a model built from its characteristics. A 316-variable LASSO regression-only model yields the highest predictive power out of all the various approaches considered in this report. This model generates a cross-validated 95% prediction accuracy, but by no means is it the best model. Although this method gains legitimacy over others, many improvements can be made to enhance predictive power.

References

Packages used:

- `DataComputing` – for various methods: http://data-computing.org/accessing-data-computing-data-and-software/
- `stats` – for various methods: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
- `leaps` – for `leaps()` and `regsubsets()` exhaustive search: https://cran.r-project.org/web/packages/leaps/leaps.pdf
- `car` – for various methods: https://cran.r-project.org/web/packages/car/index.html
- `cvTools` – for cross-validation at the end: https://cran.r-project.org/web/packages/cvTools/cvTools.pdf
- `glmnet` – for binomial LASSO regression and cross-validation: ftp://debian.ustc.edu.cn/CRAN/web/packages/glmnet/glmnet.pdf

Lecture slides (from Professor Deborah Nolan) consulted:

- PCA.html – for PCA-related code
- ModelSelection.html – for Cp/AIC/BIC comparisons as criteria, and for `leaps()`, `regsubsets()`, and stepwise selection code

Lab examples (from Omid Solari) referenced:

- Lab 0419 – for LASSO graphs and decision on RMSE error model competency measure

Textbooks:

- John Fox's *Applied Regression Analysis & Generalized Linear Models, Third Edition*
- Section 14.5, on the topic of separability issues
- Section 20.2, on the topic of dealing with missing data

Notes taken in class:

- On the topics of collinearity, PCA, cross-validation, model selection, AIC/BIC, ridge and LASSO regression, logistic regression, and generalized linear models

Note: Some of this code was taken from my own implementation of HW5 about baseball data analysis.

A huge thank you to Professor Nolan and Omid Solari for helping me debug throughout the project!