Decision of Advertising Strategy Using Machine Learning Methods

by

M.S., Haoxi Ma

University of Connecticut, 2020

Project Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Statistics

in the

Department of Statistics

Advisor: Emiliano A. Valdez

Autumn 2020

Content:

ABSTRACT

To magazine companies, forecasting the probability of purchasing specific magazines after looking at ads can significantly help identify the advertising strategy. In this report, we will focus on developing "best" machine learning models for "Kid Creative" magazine to forecast whether the customers will buy this magazine after looking at the ads in email.

First of all, after drawing learning curves on a training dataset, we do model selections to identify significant features. Then, we fit logistic regression and do some diagnoses. After that, we construct SVM, Naïve Bayes, Random Forest, Artificial Neural Networks models combining with grid search. Finally, because of the skewness of response, we compare all models by F_1 score.

In conclusion, we choose to use SVM with a Gaussian kernel which achieves 93.62% prediction accuracy after 10-fold cross validation.

Key words: Learning curves, Logistic regression, SVM, Naïve Bayes, Random Forest, Artificial Neural Networks

BACKGROUND

A magazine reseller is trying to decide what magazines to market to customers. In the "old days," this might have involved trying to decide which customers to send advertisements to via regular mail. In the context of today and the "web," this might involve deciding what recommendations to make to a customer viewing a web page about other items that the customer might be interested in and therefore want to buy. The two problem are essentially the same.

In recent years, forecasting the probability of purchase after looking at ads becomes more and more important. In order to be able to develop an equation that predicts the probability that a customer will buy a particular magazine, the company will need to run an experiment in order to collect data on customer purchase behavior. One way to do this is to randomly select some customers from the customer database and then send them emails with randomly selected ads. Whether or not these customers buy the advertised magazines can provide the data necessary to estimate the equations that will be used to predict the probability that a customer purchases a particular magazine.

As the development of machine learning methods, magazine companies can build more and more powerful machine learning models to help identify the advertising strategy. Advertising people to what they may be willing to buy is a Win-win method.

MOTIVATION

In this paper, we will focus on the issue of developing "best" machine learning models for one magazine called "Kid Creative" whose target audience are children between the ages of 9 and 12m to forecast whether the customers will buy this magazine after looking at the ads in email. First of all, we will do model selection using LASSO regression to delete some redundant features. Then, fit the Logistic regression and do some diagnoses. What's more, we are going to use other machine learning methods like SVM, Naïve Bayes, Random Forest and Artificial Neural Networks to fit the data and compare the prediction accuracy among these models.

Main goals:

(i) Plot learning curves to decide if more features (i.e. interaction or polynomial terms) are likely to help

(ii) Determine which features influences purchase significantly

(iii) Construct Logistic regression model and check its adequacy

(iv) Fit SVM, Naïve Bayes, Random Forest, Artificial Neural Networks models and compare these models by F_1 score

(v) Apply K-Fold cross validation to give the prediction accuracy

DATA DESCRIPTION

The dataset is available online on LogisticRegressionAnalysis.com and the link is:

http://logisticregressionanalysis.com/303-what-a-logistic-regression-data-set-looks-like-an-example/

This Dataset has 673 observations and 17 variables corresponding to 1 response and 16 features:

Purchased "Kid Creative" (Buy = 1 if purchased "Kid Creative", 0 otherwise)

Household Income (Income; rounded to the nearest $1000.00)

Gender (Is.Female = 1 if the person is female, 0 otherwise)

Marital Status (Is.Married = 1 if married, 0 otherwise)

College Educated (Has.College = 1 if has one or more years of college education, 0 otherwise)

Employed in a Profession (Is.Professional = 1 if employed in a profession, 0 otherwise)

Retired (Is.Retired = 1 if retired, 0 otherwise)

Not employed (Unemployed = 1 if not employed, 0 otherwise)

Length of Residency in Current City (Residence.Length; in years)

Dual Income if Married (Dual.Income = 1 if dual income, 0 otherwise)

Children (Minors = 1 if children under 18 are in the household, 0 otherwise)

Home ownership (Own = 1 if own residence, 0 otherwise)

Resident type (House = 1 if residence is a single family house, 0 otherwise)

Race (White = 1 if race is white, 0 otherwise)

Language (English = 1 is the primary language in the household is English, 0 otherwise)

Previously purchased a parenting magazine (Prev.Parent.Mag = 1 if previously purchased a parenting magazine, 0 otherwise)

Previously purchased a children's magazine (Prev.Child.Mag = 1 if previously purchased a children's magazine, 0 otherwise)

Continuous features summary:

$Table\ 1 - Continuous\ \ features\ summary$

Features	Range	Mean	Median
Income	0~75000	35079	32000
Residence.Length	0~72	17.62	16.00

$Note: There\ is\ no\ missing\ value$

Categorical features summary:

$Table\ 2 - Categorical\ \ features\ summary$

Features	Value=1	Value=0
Buy	125	548
Is.Female	371	302
Is.Married	235	438
Has.College	195	478
Is.Professional	230	443
Is.Retired	39	634
Unemployed	21	652
Dual.Income	156	517
Minors	245	428
Own	244	429
House	449	224
White	466	207
English	612	61
Prev.Parent.Mag	48	625
Prev.Child.Mag	57	616

$Note: There\ is\ no\ missing\ value$

EXPLORATORY DATA ANALYSIS

1 Data preprocessing:

After checking the missing value, we encode the categorical variables and split the dataset by randomly selecting 80% data into the train dataset and leaving 20% into the test dataset. Then, we do feature scaling which will be helpful to improve the performance of machine learning models, especially important when using Gaussian kernel in SVM.

2 Learning curves

For convenience, we use python to draw the learning curves:

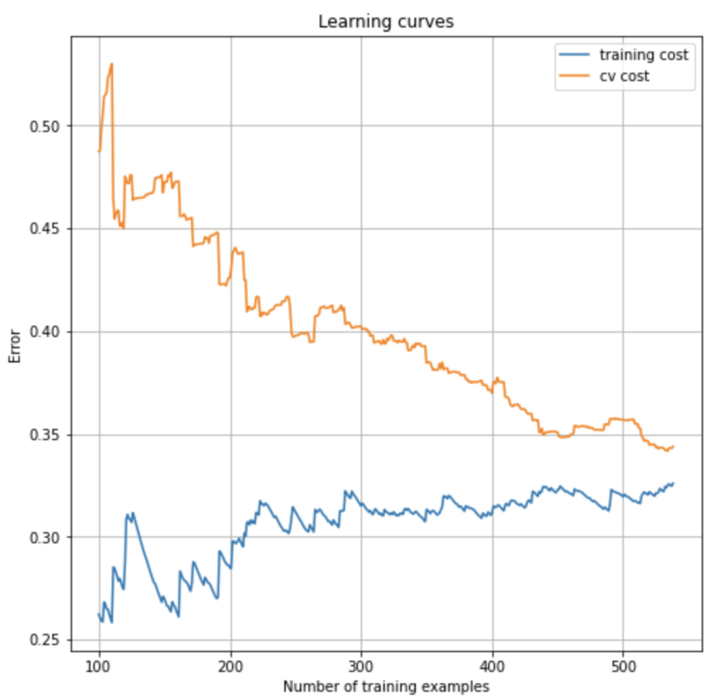


Figure 2 − 1: Learning curves for training dataset

According to the figure above, we can argue that a learning algorithm is not suffering from high bias or high variance which means we do not need to add polynomial terms or give more data to training dataset.

3 Check the skewness of response

In machine learning, we need to decide which criteria to use for comparing different models. From the table 2, we know that response “Buy” has #125 equal to 1 while #548 equal to 0, so it is obviously skewed.

In this two-group classification, if the response is skewed, it will not be appropriate to use accuracy as criteria. However, $F1\ score$ is a good choice, defined as

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

4 Check the multicollinearity

We calculate the VIF for each feature and the output is shown in Appendix F-1. From the output, all VIF values are less than 5 indicating no multicollinearity but we still need model selection to delete redundant features.

5 Model selection

5.1 Lasso regression

When we do Lasso regression to select a model, we have two choices: one is the simplest model with larger deviance, another is the more complicated model with minimum deviance. Here, we choose the second one and the output is listed in Appendix F-2.

According to the output, model 1 selected by Lasso regression is

$$Buy \sim Income + Is.Female + Dual.Income + Minors + Own + House + White + Prev.Child.Mag + Prev.Parent.Mag$$

5.2 Stepwise AIC

We can also use stepwise AIC method to get the model having the smallest AIC. The output is shown in Appendix F-3 so the model 2 is

$$Buy \sim Income + Is.Female + Minors + Own + White + English + Prev.Child.Mag + Prev.Parent.Mag$$

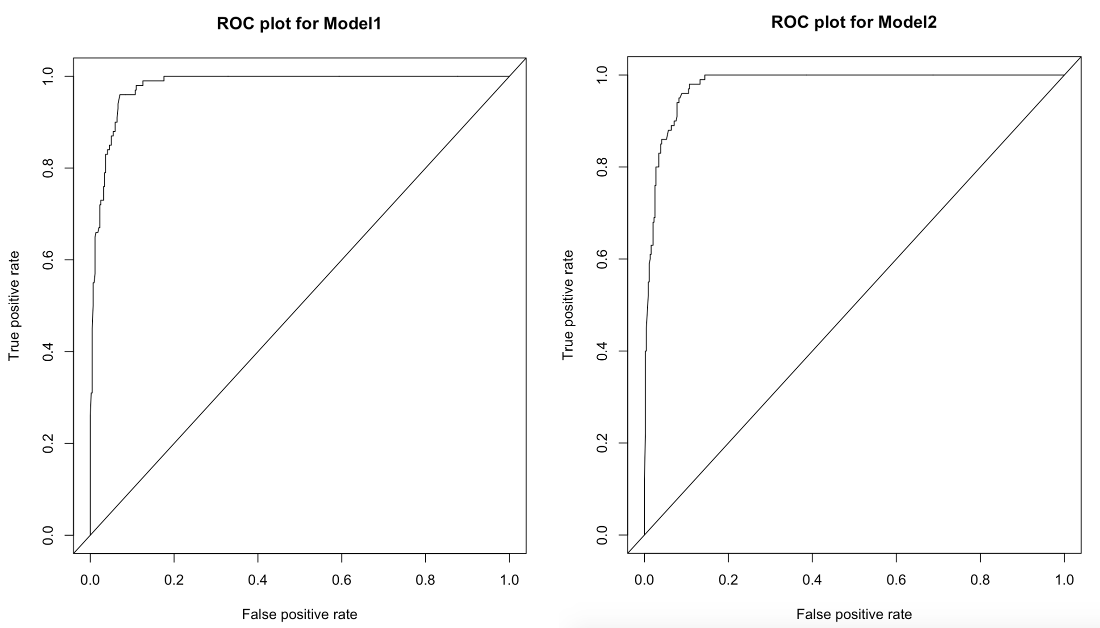
5.3 Choose the better model

After two parts above, we get two "best" models under different standards. Now, we need to choose the better one. Summarize the AIC and area under the ROC plot:

$Table\ 3 - \ Summary\ AIC\ and\ area\ for\ two\ models$

Model	k	AIC	Area
Model 1	9	173.2	0.9796347
Model 2	8	172.58	0.9790183

And the ROC plots are



$Figure\ 5 - 1\!: ROC\ plot\ for\ two\ models$

From the table 3 and two ROC plots, model 1 performs as well as model 2. Then, considering Principle of Parsimony, we choose model 2. Therefore, the useful or significant features are Household Income, Gender, Children, Home ownership, Race, Language, Previously purchased a parenting magazine and Previously purchased a children's magazine.

6 Logistic regression

6.1. Fit model

We are going to fit the logistic regression using the selected features above. Because all categorical features have 2 group, we set the dummy variables like

$$Dummy_i = \begin{cases} 1 & value = 1 \\ 0 & value = 0 \end{cases}$$

Then we get the logistic regression as (output shown in Appendix F-4):

$$\begin{aligned} log\left(\frac{p}{1-p}\right) &= -16.4968 + 13.822 Income + 1.4461 Is.Female + 0.8993 Minors \\ &+ 1.4081 Own + 1.9043 White + 1.3392 English \\ &+ 1.2166 Prev.Child.Mag + 1.169 Pre.Parent.Mag \end{aligned}$$

Interpretation:

For continuous features, Income, the log-odds of buying "Kid Creative" magazine will increase 13.822 from one dollar increase in income when all other features are held fixed. For categorical features, Is.Female, the odds of buying "Kid Creative" magazine for female is $e^{1.4461}$ times the odds for male when all other features are held fixed. Others are similar.

6.2. Diagnose

In this part, we draw Pearson residuals and deviance residuals plot first:

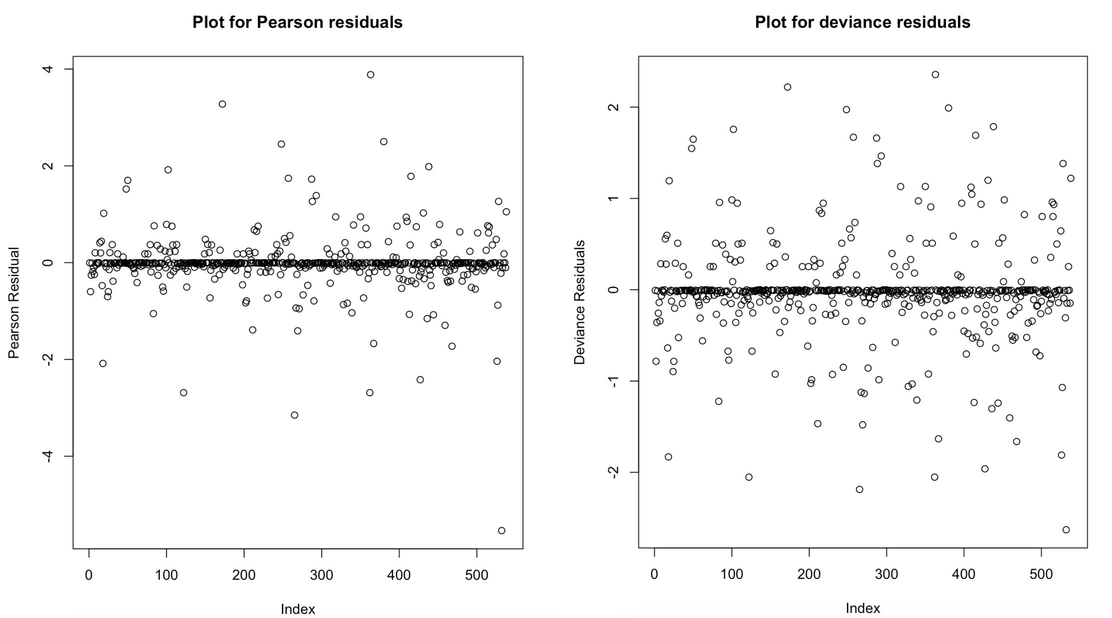


Figure 6 − 1: Pearson residuals and deviance residuals plot

According to the Figure above, there is no specific pattern but some modulus of Pearson residuals are larger than 2 indicating necessity of checking lack of fit. Therefore, we apply Hosmer-Lemeshow Test to the response and fitted value (Output shown in Appendix F-5) and P-value is 0.998 super larger than 0.05. Then, we can't reject the null hypothesis and conclude that a logistic classifier is a good fit.

6.3. Select threshold

In logistic regression, choosing the best cut-off value will improve the performance of the classifier. We suppose some thresholds and plot the prediction accuracy:

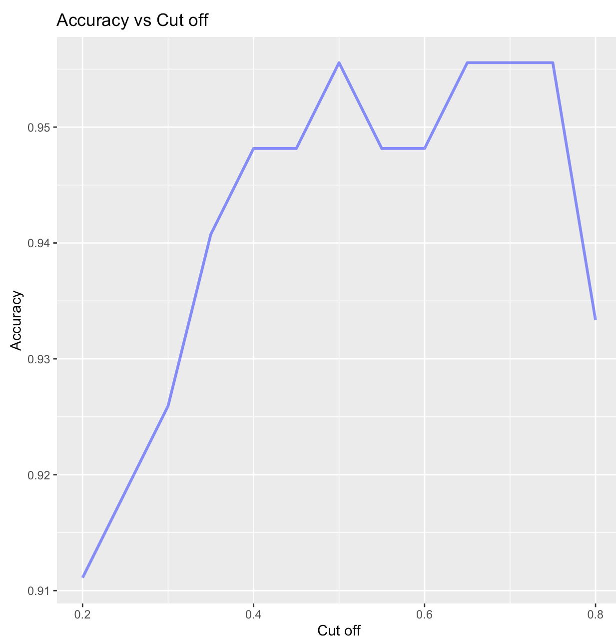


Figure 6 – 2: Different thresholds vs accuracy

From the plot, thresholds 0.5, 0.65, 0.7 and 0.75 have the same accuracy. However, the people who will buy the magazine are extremely important to a magazine seller so we tolerate more type II error. Therefore, we pick 0.5 as the cut-off.

6.4. Confusion matrix and F1 score

Construct the confusion matrix within test dataset:

Table 4: Confusion matrix in logistic regression

actual / Predict	1	0
1	23	4
0	2	106

Finally, the F1 score of logistic regression is 0.9724771.

7 Support vector machine

Because #features are small and #observations are intermediate, we choose to use Gaussian kernel which is

$$f_1 = exp\,(-gamma * ||x - l^{(1)}\,||^2)$$

7.1. Grid search

In order to fit the "best" SVM model, we need to decide which regularized weight C and which dispersion parameter $gamma$ to use. Then, we use grid search method and the result is

Table 5: Grid search results

C	$gamma$
1	0.04986254

7.2. Confusion matrix and F1 score

Construct the confusion matrix within test dataset:

Table 6: Confusion matrix in SVM

actual / Predict	1	0
1	23	2
0	2	108

Finally, the F1 score of logistic regression is 0.9818182.

8 Naïve Bayes

This method is a simple method just based on the Bayes formula. We can directly fit it and make predictions in the test dataset. Then, construct the confusion matrix:

Table 7: Confusion matrix in Naive Bayes

Predict \ actual	1	0
1	23	6
0	2	104

Finally, the F1 score of logistic regression is 0.962963.

9 Random forest

In order to fit a better random forest classifier, we need to figure out how many variables randomly sampled as candidates at each split. Again, we use grid search method and find the best #variables are 2. After that, we construct the random forest classifier and draw a plot to show the importance of features:

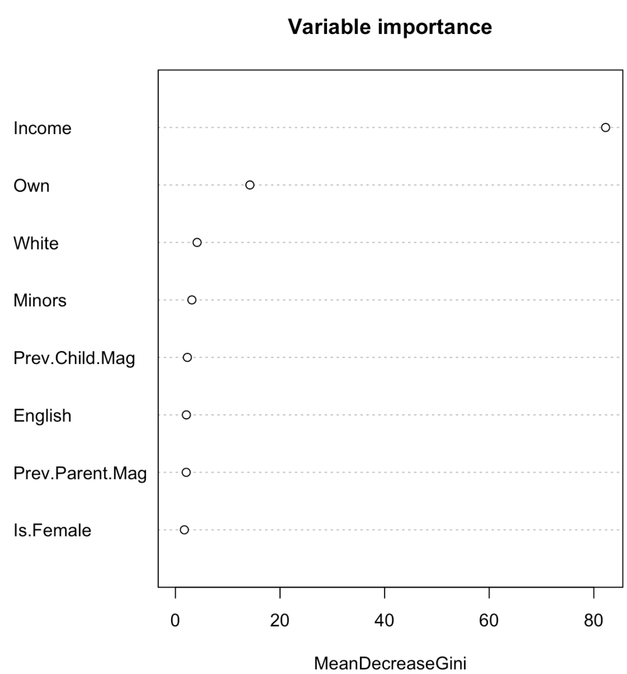


Figure 9 − 1: Features importance plot in random forest

According to the figure, we can conclude that income plays the most important role in buying magazines. This is a really logical conclusion.

Construct the confusion matrix:

$Table$ 8: $Confusion\ matrix\ in\ random\ forest$

actual / Predict	1	0
1	20	3
0	5	107

Finally, the F1 score of logistic regression is 0.963964.

10 Artificial neural network

Because neural networks will fit very complicated new features, it usually takes a long time to finish. In this time, we are going to use the parallel computing in h2o platform which can save a lot of time. Here, we will use 2 hidden layers with 8 neurons in the first layer and 4 in the second layer.

Construct the confusion matrix:

$Table$ 9: $Confusion\ matrix\ in\ ANN$

actual / Predict	1	0
1	20	5
0	2	105

Finally, the F1 score of logistic regression is 0.9677419.

DISCUSSION

Above all, we fit 5 machine learning models to do this two-group classification. As mentioned before, because of the skewness of the response, we are going to compare different models by F1 score.

Table 510: Grid search results

Logistic regression	SVM	Naïve Bayes	Random forest	ANN
0.9724771	0.9818182	0.9629630	0.9639640	0.9677419

According to the table above, the SVM model has the largest F1 score. Besides, its false negative cases are smallest. In conclusion, we prefer the SVM with the Gaussian kernel method to deal with this forecasting problem.

After selecting the "best" model, we do 10-fold cross validation to evaluate the overall prediction accuracy and the result is that SVM with Gaussian kernel has 93.63% prediction accuracy which is really high.

At last, we want to argue that there is an insufficiency left. That is, the dataset might be too ideal to distinguish these models significantly. Next time, try some more nonperfect datasets and apply more complicated models.

Appendix

F-1:

```
> vif(check_model)
         Income       Is.Female       Is.Married      Has.College  Is.Professional       Is.Retired
       2.162220        1.427314        2.318978        1.343797         1.553143         1.580532
     Unemployed Residence.Length     Dual.Income           Minors              Own            House
       1.021111        1.275572        1.883748        1.379817         2.217700         1.634060
          White         English  Prev.Child.Mag  Prev.Parent.Mag
       1.362022        1.213629        1.092942        1.098589
```

F-2:

```
> coef(model_select)
17 x 1 sparse Matrix of class "dgCMatrix"
                          s0
(Intercept)      -0.27271486
Income            0.73348396
Is.Female1        0.04501878
Is.Married1       .
Has.College1      .
Is.Professional1  .
Is.Retired1       .
Unemployed1       .
Residence.Length  .
Dual.Income1      0.06524198
Minors1           0.02722275
Own1              0.01803923
House1            0.01412153
White1            0.06346235
English1          .
Prev.Child.Mag1   0.06542867
Prev.Parent.Mag1  0.01803936
```

F-3:

```
Coefficients:
    (Intercept)           Income       Is.Female1          Minors1             Own1           White1
       -16.4968          13.8220           1.4461           0.8993           1.4081           1.9043
       English1  Prev.Child.Mag1 Prev.Parent.Mag1
         1.3392           1.2166           1.1690
```

F-4:

```
Call:
glm(formula = final_formula, family = "binomial", data = train_set)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.62870  -0.10747  -0.01373  -0.00273   2.35721

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -16.4968     2.1641  -7.623 2.48e-14 ***
Income            13.8220     1.6913   8.172 3.03e-16 ***
Is.Female1         1.4461     0.4726   3.060 0.002212 **
Minors1            0.8993     0.4419   2.035 0.041837 *
Own1               1.4081     0.4589   3.068 0.002154 **
White1             1.9043     0.5578   3.414 0.000641 ***
English1           1.3392     0.8652   1.548 0.121646
Prev.Child.Mag1    1.2166     0.7490   1.624 0.104304
Prev.Parent.Mag1   1.1690     0.6836   1.710 0.087257 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 516.68  on 537  degrees of freedom
Residual deviance: 154.58  on 529  degrees of freedom
AIC: 172.58

Number of Fisher Scoring iterations: 8
```

F-5:

```
> hoslem.test(y_goodness,fitted(logistic_classifier),g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  y_goodness, fitted(logistic_classifier)
X-squared = 1.0253, df = 8, p-value = 0.9981
```