Applied Machine Learning

Assignment 1

Sai Vishnu Kanisetty SXK175300

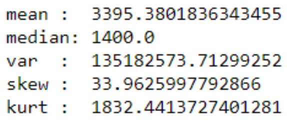
Objective of this exercise:

- To predict the number of shares for articles published by Mashable
- To classify the articles as high shared or low shared

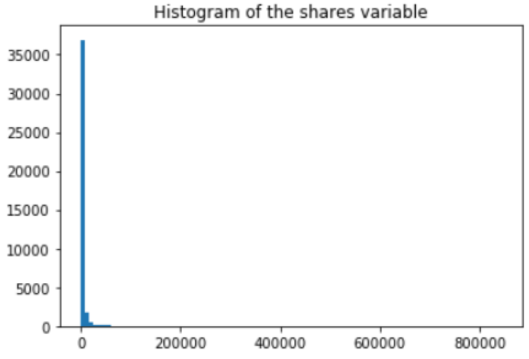
Index: This exercise has been implemented in python and different topics discussed in this article are:

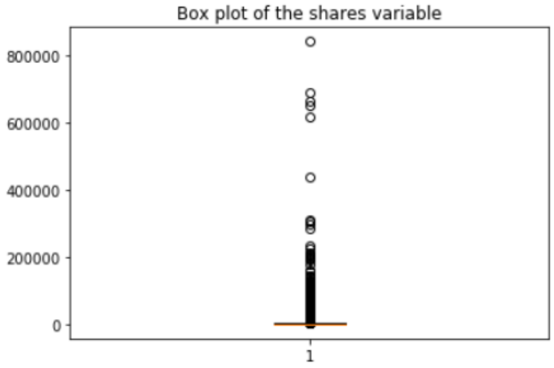
Step 1: Understanding data and construction of classes for the classification

- Data has 39,644 records of 61 variables and there are no missing values
- Two non-predictive features (URL and Timedelta) have been dropped from further analysis
- All the numerical predictor features have been normalized
- Some information about the shares variable:



```
mean :  3395.3801836343455
median: 1400.0
var  :  135182573.71299252
skew :  33.9625997792866
kurt :  1832.4413727401281
```





- Mean being more than the median and having values above the 2 standard deviations (box plot) shows that the data is positively skewed

- Median of the shares (1400) has been considered cutoff for the classification. All the articles which have more than 1400 shares have been labelled to “high share”(1) articles and all the articles below the median have been labelled to “low share” (0) articles. The split of the articles to high shares (20,082) and low shares (19,562) will be ~50% because of considering median as the split point.

Step2: Definition for different algorithms used and their hyper parameters

- For predicting number of shares, two versions of linear regression have been implemented. Error metric for both the methods is R-Square (percentage of the response variable variation that is explained by the predictor variables).

  Method 1: Linear Regression with Normal Equations

  Method 2: Linear Regression with batch gradient descent. Cost function that is minimized is the sum of squares of the actual and predicted shares.

  Hyper parameters are:

  - ♦ Learning rate for gradient descent
  - ♦ Convergence threshold

- For classifying the articles as high shares and low shares, two version of Logistic Regression have been implemented. Error metric for both the methods is Accuracy (Overall how often the classifier is correct).

  Method 1: Logistic Regression with Normal Equations

  Method 2: Logistic Regression with batch gradient descent. Cost function that is minimized is **cost (hθ, (x),y) = - ylog(hθ(x)) - (1-y)log(1- hθ(x))**

  Multiple random initialization of the estimates and intercept have been implemented in Logistic Regression because there can be a possibility of multiple local minima for the defined cost function.

  Hyper parameters are:

  - ♦ Learning rate for gradient descent
  - ♦ Convergence threshold

  Cutoff threshold probability for both logistic regression with normal equations and batch gradient descent has been defined as a probability at which True positive rate is high and False positive rate is low.

  True positive rate: Proportion of events that are correctly predicted as event

  False positive rate: Proportion of non-events that are wrongly predicted as an event.

- For the entire exercise data has been randomly divided into 70% train and 30% test.

Step 3: Results of experimentation using all the heterogenous features

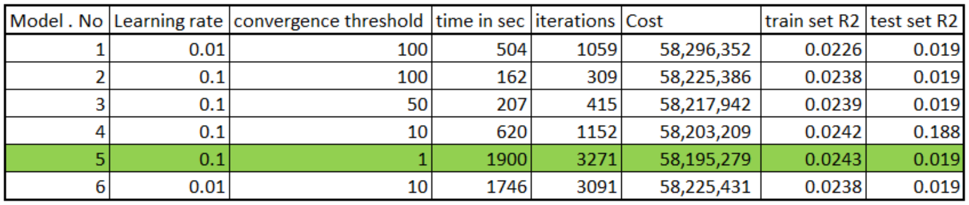
In both the gradient descent approaches, experimentation has been performed by changing the convergence threshold and learning rate to obtain the best value of the defined error metric for repetitive algorithms (Linear Regression R-Square of 1 and Logistic Regression Accuracy of 100%).

Linear Regression with Normal Equations

Train data R2	0.0239
Test data R2	0.0176

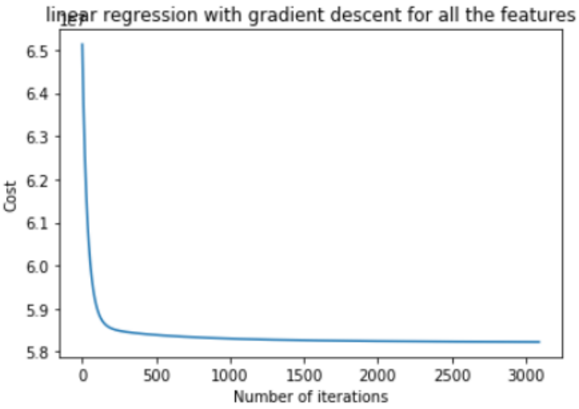
Linear Regression with Gradient Descent

Execution time and Number of iterations taken to reach the convergence, Cost, train set R2 and Test set R2 associated with different values of convergence threshold and learning rate (alpha).



Model . No	Learning rate	convergence threshold	time in sec	iterations	Cost	train set R2	test set R2
1	0.01	100	504	1059	58,296,352	0.0226	0.019
2	0.1	100	162	309	58,225,386	0.0238	0.019
3	0.1	50	207	415	58,217,942	0.0239	0.019
4	0.1	10	620	1152	58,203,209	0.0242	0.188
5	0.1	1	1900	3271	58,195,279	0.0243	0.019
6	0.01	10	1746	3091	58,225,431	0.0238	0.019

Visual display cost for the best mode (6) from the above graph.



Findings:

- Model with alpha of 0.1 and convergence of 1 has the least cost and highest R2 for both test set and train set.
- Train and test R2 of gradient descent approach are high compared to Normal Equations method

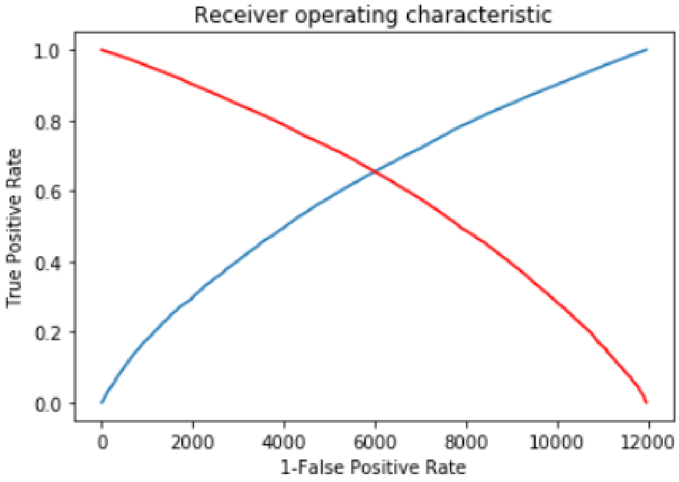
Logistic Regression Baseline

Because of considering median as the split point, the baseline accuracy of the model is ~50%.

Logistic Regression with Normal Equations

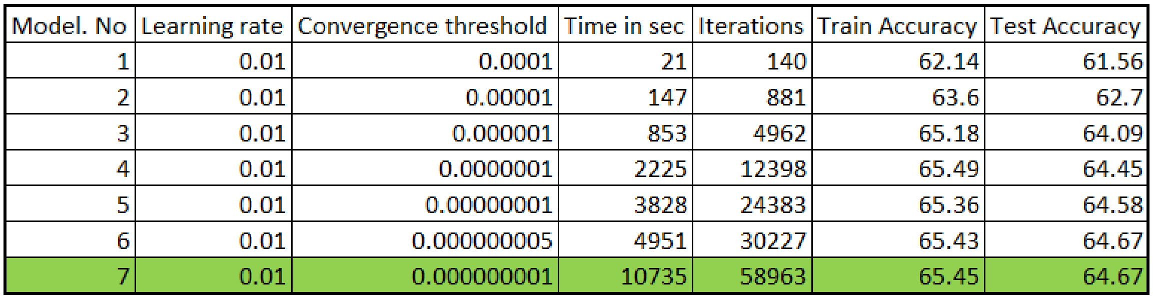
Cutoff Probability Threshold	0.48
Train data R2	65.40
Test data R2	64.6

ROC curve to determine the cutoff probability



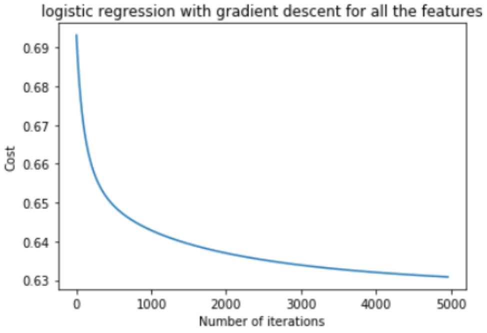
Logistic Regression with gradient Descent

Execution time and Number of iterations taken to reach the convergence, train set R2 and Test set R2 associated with different values of convergence threshold and learning rate (alpha).



Model. No	Learning rate	Convergence threshold	Time in sec	Iterations	Train Accuracy	Test Accuracy
1	0.01	0.0001	21	140	62.14	61.56
2	0.01	0.00001	147	881	63.6	62.7
3	0.01	0.000001	853	4962	65.18	64.09
4	0.01	0.0000001	2225	12398	65.49	64.45
5	0.01	0.00000001	3828	24383	65.36	64.58
6	0.01	0.000000005	4951	30227	65.43	64.67
7	0.01	0.000000001	10735	58963	65.45	64.67

Plot of the cost for Model 3



Findings:

- Model with alpha of 0.01 and convergence of 0.000000001 has the highest accuracy for both test set and train set
- Train and test R2 of gradient descent approach are high compared to Normal Equations method
- Cutoff Probability Threshold for the best model (Model 7) in logistic regression is 0.49

Step 4: Results using ten random features

Random selected ten columns are:

```
data_channel_is_socmed, kw_min_avg, kw_max_avg, num_videos,
data_channel_is_entertainment, global_rate_negative_words,
n_unique_tokens, num_hrefs, title_subjectivity, data_channel_is_tech
```

The hyper parameters learnt from respective experimentation of gradient descent in step 3 have been used. Results are present in Image 1 (Page 6)

Finding:

- For both linear regression and logistic regression, test and train error metric are very nearby for Normal Equations method and using gradient descent

Step 5: Results using ten selected features

Hypothesis: An article about business, published on a weekend, with highest number of non-stop words and having extreme positive polarity or negative polarity, covering images and title conveying a sentiment can have good chance of getting highest number of shares.

Based in the above analogy ten features have been selected from the data:

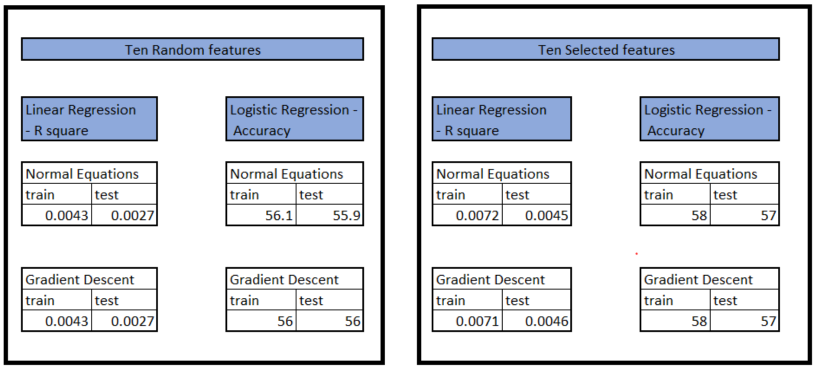
n_unique_tokens, n_non_stop_unique_tokens, self_reference_max_shares, is_weekend, max_positive_polarity, title_sentiment_polarity, average_token_length, max_negative_polarity, num_imgs, data_channel_is_bus

The hyper parameters learnt from respective experimentation of gradient descent in step 3 have been used. Results are present in Image 1 (Page 6)

Finding:

- For both linear regression and logistic regression, test and train error metric are very nearby for Normal Equations and using gradient descent

Image 1



Ten Random features

Linear Regression - R square

Normal Equations	
train	test
0.0043	0.0027

Gradient Descent	
train	test
0.0043	0.0027

Logistic Regression - Accuracy

Normal Equations	
train	test
56.1	55.9

Gradient Descent	
train	test
56	56

Ten Selected features

Linear Regression - R square

Normal Equations	
train	test
0.0072	0.0045

Gradient Descent	
train	test
0.0071	0.0046

Logistic Regression - Accuracy

Normal Equations	
train	test
58	57

Gradient Descent	
train	test
58	57

Step 6: Comparison of the results from Step 3, 4 and 5

Findings:

- Model with all the features has the highest R2 and highest accuracy for both train and test datasets
- Model with selected features has high R2 and high accuracy compared to random selection.

Business intuition and understanding of data has helped in selecting the features that can impact the shares of any article. The selected variables are vey less related with each other but cover a holistic view of several reasons that can lead to sharing an article.

For example, randomly selected features has only “unique tokens” but not the “nonstop unique tokens”, which is an important metric to consider the amount of sentiment contributing content in a article.

Step 7: Comparison of the results from Step 3, 4 and 5

- Low R2 suggests data is not linear. Below mentioned methods can be tried to draw a better fit line
  a) polynomial regression with regularization
  b) Non -linear like decision trees and random forest
- Gradient Descent: Execution time taken for the model with highest accuracy in logistic regression is ~ 3 hours. To decrease the running time, methods mentioned below can be considered.
  a) momentum
  b) Xavier Initialization

Reference:

1. Andrew NG, Deep Learning Course in Coursera
2. Gradient descent http://ruder.io/optimizing-gradient-descent/index.html#momentum
3. Cut off threshold in Logistic regression https://stackoverflow.com/questions/28719067/roc-curve-and-cut-off-point-python