Machine Learning Report

I. Introduction :

Many webpage usually show images that are sort of advertisements for some other webpages or products. Advertisements are always very annoying because they tend to dominate the webpage and result in very slow loading and somehow redirect user to other undesired webpages. In our project, we try to solve ads detection problem that given an image extracted from webpage with certain set of features classifies it into 2 sets: non ads and ads.

In reality, the source of annotated instances is very limited where there are large number of images extracted from webpages. We could use the large source of extracted image to improve the performance of classifier. A research paper 'Learning to remove advertisements' by Nicholas Kushmerick was studied to gain proper insights on the features of data. Additionally, to reduce the dimensional of features contain in dataset, paper "Experiments with Random Projections in Machine Learning" will be presented to observe the comparison of performances between PCA and Random Projections on some standard machine learning tools.

II. Describe the pipeline :

Overall, the pipeline of our works contains 4 steps: Pre-processing, Feature Engineering and Reduce Dimensional, Evaluation. All the steps will be presented as below:

Pre-processing: The goal of this task is to make the other steps become easier. Dataset combines 3279 samples (458 advertisements and 2821 non advertisements) which is described by 1559 attributes:

- Height, width, aspect ratio, local.
- 19 caption features.
- 111 features for alternate word in image's tag.
- 495 features for base URL.
- 472 features for destination URL.
- 457 features for image's URL.

This task will include:

- Handing with missing values (28% of dataset)

	Missing value 'ads'	Missing value 'nonads'	Total
Height	830	73	903
Width	828	73	901
Aspect Ratio	837	73	910
Local	10	5	15

The missing values in height and width will be replaced by the mean of height and width values with respect to 'ads' and 'nonads'.

The missing values of ratio aspect will be calculated by the division of width on height.

The missing local cells are replaced by the value which is in majority in the available aspect ratio of all the sample belonging to the same class 'ad' and 'nonads'.

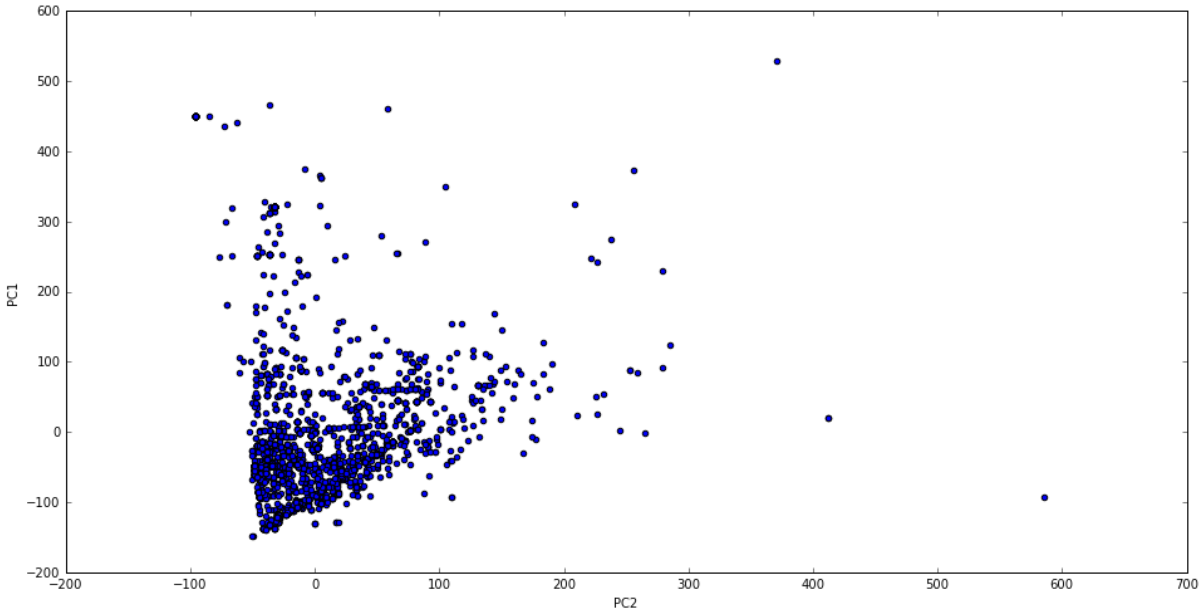
- Converting string target values to integer form.
- The target value at column 1558 will be converted to binary value to make the process become easier.

Feature Engineering and Reduce Dimensional: as we have discussed in the introduction part, we will use PCA and Random Projects (RP) as the method to reduce dimensions of features which affects on the dataset. Then we will compare 2 methods to have an overview in the paper "Experiments with Random Projections for Machine Learning" using a number of standard machine learning tools, such as K-nearest neighbor or SVM… The purpose of paper is not to compare the performances of these methods to each other, but to see the differences in their performances using PCA and RP based on the pseudo code in paper. Due to this code, the dataset will be split into 70% training and 30% testing set.

- PCA (Principal Component Analysis):

  In this technique, variables are transformed into a new set of variables, which are linear combination of original variables. These new set of variables are known as principle components. They are obtained in such a way that first principle component accounts for most of the possible variation of original data after which each succeeding component has the highest possible variance. The second principal component must be orthogonal to the first principal component. In other words, it does its best to capture the variance in the data that is not captured by the first principal component.

  In our case, we tried to test on 2-dimensional and received the result the first PC already explains almost 85,62% of the variance, while the second one accounts for another 14,2% for a total of almost 99,82% between the two of them.



- Random Projection: A much easier way to map training points to lower dimension space in preserving distance. A theorem due to Johnson and Lindenstrauss states that for a set of points of size n in p-dimension al Euclidean space, there exists a linear transformation of the data into a q dimensional space, q ≥ O(ε^{-2}log(n)) that preserves total distance up to a factor 1±ε. Other theorem with a construction of projection matrix states that: Given n point in R^p , choose ε, β > 0 and q > $\frac{4+2\beta}{\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}}$ and let E = $\frac{1}{\sqrt{q}}$XP for projection matrix P. Then mapping X to E preserves distances up to factor 1 ± ε for all rows in X with probability (1 - $n^{-\beta}$). The projection matrix P, p x q, is constructed as following:
    - With probability 1/2: p_{ij} = 1
    - With probability 1/2: p_{ij} = -1

  Or

    - p_{ij} = $\sqrt{3}$ * ±1 (with probability 1/6 for each)
    - p_{ij} = 0 (with probability 2/3)

Experiment Pseudocode:

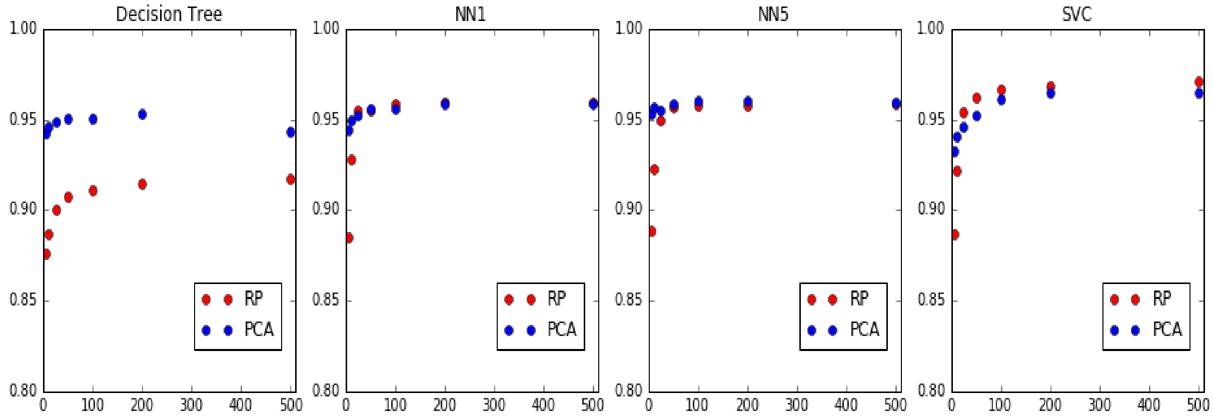
```
Data set D, set of projection dimensions {5, 10, 20, 25, 50, 100, 200, 500},
30 random splits (training and testing set: 70%, 30%).
    For each random splits s:
        ▪ Normalize data
        ▪ For each dimension q in set of dimensions:
            • Do PCA on training set and project both training set and
              testing set into $R^q$
```

- Create random projection matrix as described in the theorem and project both set into R^q

- Train learning methods on projected training set and apply them to projected testing set to evaluate the performance.

Evaluation:

The figure below shows the results of experiments. The x-axis represents the different dimensions of the projection space and the y-axis represents the accuracies.



III. Discuss the results:

We apply PCA and random projection to SVM, K-nearest neighbors (k=1, k=5) and Decision Tree.

For 3 over 4 classifiers, PCA gives better results than random projection. However, when the number of components increase, accuracies resulted by random projection are closed to those resulted by PCA. Moreover, the complexity of random projection is really low O(pq) while PCA needs O(p^2n) + O(p^3) .

4 Classifiers have same performances in using PCA features. However, Decision tree does not work well with random projection features. Moreover, SVM works well in using random projection features and results high accuracy.

IV. Conclusion:

In this report, we compared the comportement between two methods PCA and Random Projection. In most of the cases, PCA gives good results but for large, high-dimensional data like ad data, it is very expensive in computation.
Random Projection has an advantage in computational cost, but also keeps some desired properties. Although Random Projection shows some poor results with small projection dimensions, its predictive performance improves if we increase the projection dimension, especially when used with appropriate learning methods.

V. Reference:

N. Kushmerick (1999). "Learning to remove Internet advertisements"
Dmitriy Fradkin, David Madigan “Experiments with Random Projections for Machine Learning”