Human Activity Recognition with Smartphones

Introduction

Human activity recognition (HAR) focuses on classifying sequences of accelerometer data recorded by specialized harnesses or smart phones into known well-defined movements [1]. These movements are often normal indoor activities such as standing, sitting, jumping, and going up stairs. With the recent progress in wearable technology, pervasive sensing and computing has become feasible. Thus, HAR has widely application into various fields, such as active and assisted living systems for smart homes, healthcare monitoring applications, monitoring and surveillance systems for indoor and outdoor activities, and tele-immersion applications [2].

Data Collection and Preparation

Source of Dataset

The data [3] was collected from an experiment carried out with 30 subjects aged between 19 and 48 years old performing one of 6 standard activities (Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing and Laying) while wearing a waist-mounted smartphone (Samsung Galaxy S II) that recorded the movement data. Video was recorded of each subject performing the activities and the movement data was labeled manually from these videos. These movements were recorded as 3-axial accelerometer data (linear acceleration) and 3-axial gyroscopic data (angular velocity) at a constant rate of 50Hz[4]. The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed width sliding windows of 2.56 sec and 50% overlap (128 readings/window). From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

Data Overview

The data consisted of a total number of 10299 entries and 563 variables, with 561 features, one label, *Activity* and one id number, subject_id. The distribution of the outcome variable *Activity* is shown in Figure 1. There is no missing data and no duplicates of entries. Each feature has been standardized to the value range [-1, 1].

Exploratory Analysis

The distribution of activity showed the relatively equal distribution of human activities captured in the experiment, which also indicated there is no class imbalance. Laying activity occurred in a relatively higher frequency than walking activities.

To check through which device the data is collected, I processed the column and count the feature of for three type of devices. Across 561 features, 345 features are collected using sensor accelerometer, and 213 features are collected by Gyroscope (Figure 2).

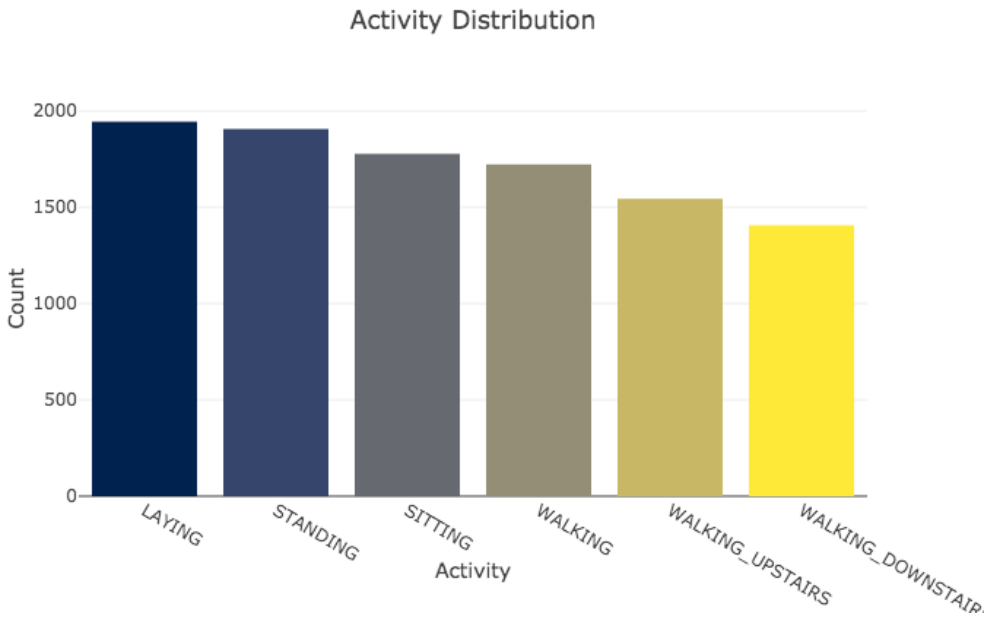

Figure 1: Smartphone activity label distribution in count.

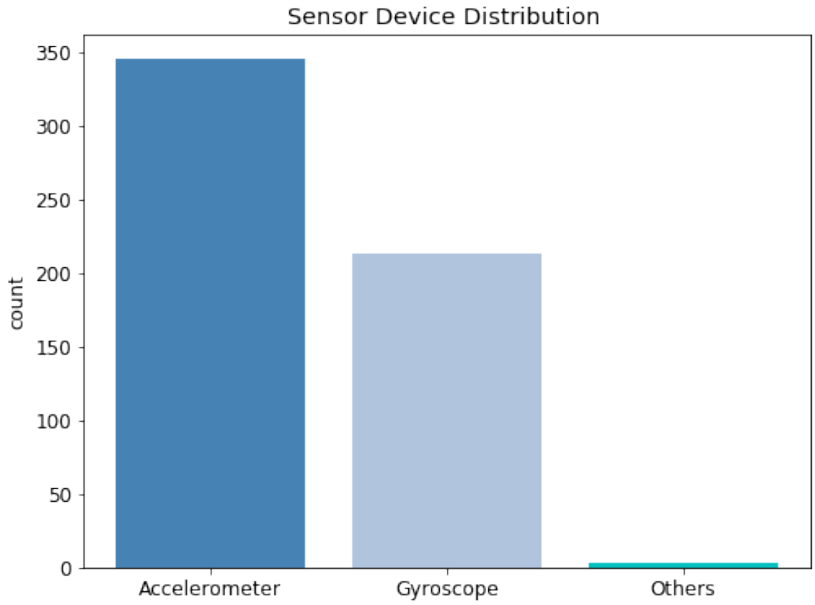

Figure 2: Sensor device distribution.

Given the large set of features included in the dataset, examining the bivariate feature correlation of all features does not provide us a clear understanding. Also, given most of features are generated from the main feature sets, my hypotheses are most movements will be highly correlated. Therefore, I randomly selected 12 features to test the bivariate correlation between features (See figure 3). The figure result indicated

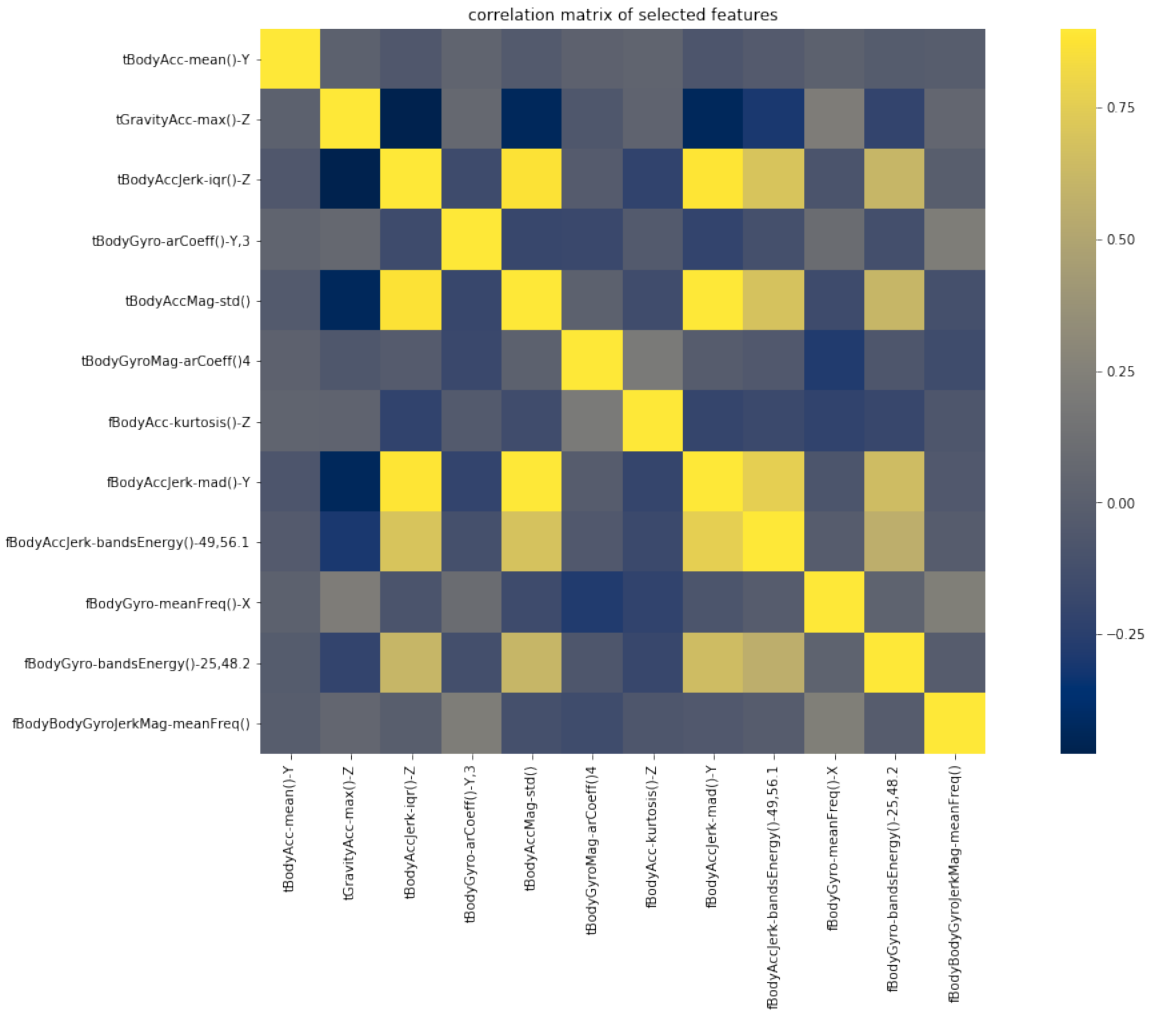

Figure 3: Sensor device distribution.

Data Preprocessing

Train and Test Data Split

The training and testing data are split by subject_id, to avoid the data leaking introduced by random splitting method. The datasets were split into training dataset with 80% participants (n = 24) and test dataset with 20% participants (n = 6). The 6 subject_id is chosen randomly with subject_id = 5,10,15,20,25, and 30. Such splitting resulted in 8229 entries in the training set and 2070 entries in the test set.

Label Encoding

The label Activity is a categorical feature consisted with 6 levels of activity. To prepare for the modeling, the label is encoded into numerical feature using label encoding method.

Modeling

Two models are used to predict the human activity: logistic regression and random forest.

Logistic Regression

Logistic regression serves as the baseline model for our analysis. Logistic regression analyzes the relationship between multiple features and a categorical label, to estimates the probability of occurrences of an event by fitting data to a logistic curve. Logistic regression acted as a probabilistic discriminative classifier. In our case, the label is transformed to multiclass, and thus logistic regression produces piece-wise linear decision boundary [5]. The advantage of using the logistic regression is to assume the independence between multiple features and ignore the covariance among features and are subject to confounding effects [6]. An initial model of logistic regression is applied and achieved an accuracy of 95.41%. The confusion result is shown below.

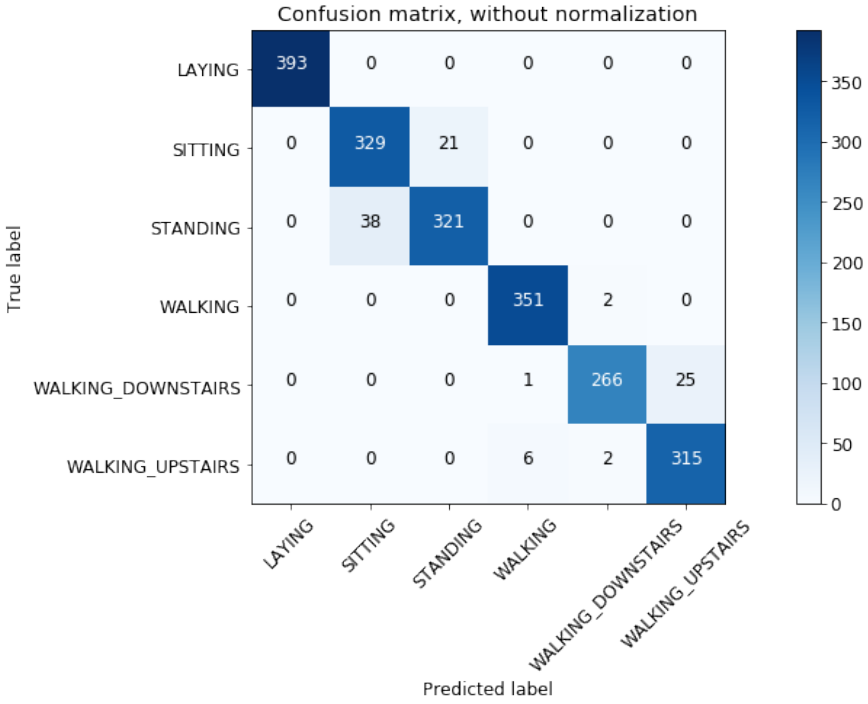

Figure 4: Confusion Matrix of Logistic Regression.

Random Forest

The second model I adopted is tree-based. Random Forest is an ensemble learning method, composed of a set of decision trees. Each decision tree acts as a weak classifier and pooling the responses from multiple decision trees leads to a strong classifier. A decision tree is trained independently and determines the class of input by evaluating a series of greedily learned binary questions [7]. Random Forest adopted a bagging method, which is the approach to decrease the variance of prediction by generating multiple different bootstrap training samples with replacement. Initially, I used the default feature of random forest, which yielded an accuracy of 89.37%. After tuning the parameter max_depth that controls the depth of nodes, and the n_estimator that controls the number of trees, the best accuracy score was achieved as 90.82%

when the random forest consisting of 40 trees, with the max_depth of 25 nodes was used. The confusion matrix is shown below.

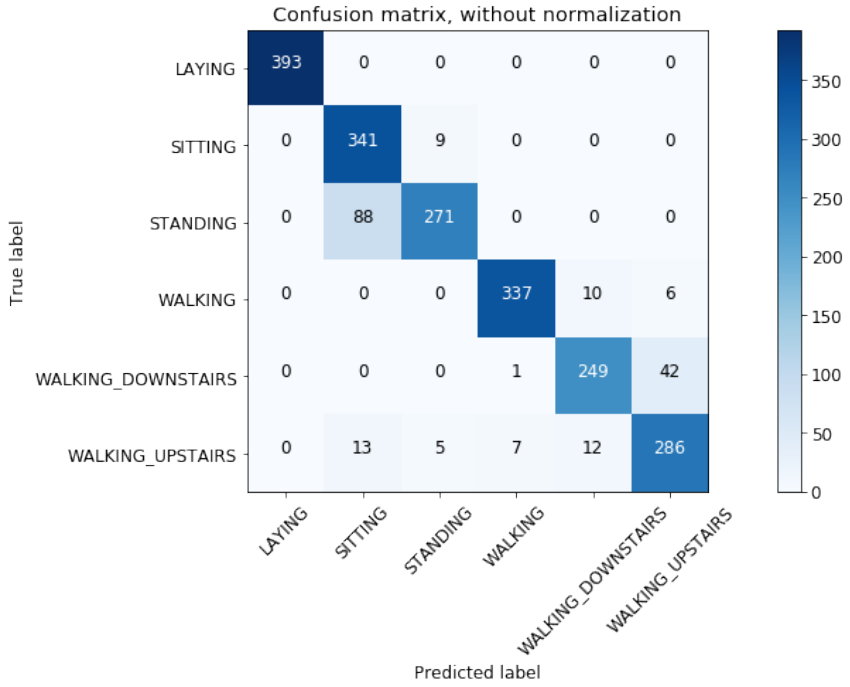

Figure 5: Confusion Matrix of Random Forest.

Model Comparison

When comparing the overall performance of the model, accuracy is used as the metric. The accuracy is defined as the ratio of several correct predictions in the number of predictions. The confusion matrix result presented the prediction result of predicted value and true label. Interestingly, the logistic regression resulted in higher accuracy of 95.41% than the random forest model 90.82%. In most of the cases, the tree-based model performs better than linear classifiers, but the interesting finding might be due to several reasons. Potentially, it could be our data is linearly separable, so the logistic regression performs better than the random forest. It is also possible that parameter tuning of random forest model does not work its best. To better understand this problem, I tested another tree-based model XGboost[8], which is a scalable and accurate implementation of gradient boosting machines designed for the speed and performance. It has been proven to push the limits of computing power for boosted trees algorithms, the accuracy achieved with XGboost is 93.09%, which is still lower than logistic regression. Therefore, it is hard to identify the causing factor in this single case, therefore, more experiment is needed to test the result.

Feature Selection

The purpose of feature selection is to reduce the number of features, to reduce overfitting and improve the generalization of models. I adopted the method of univariate feature selection techniques, which examines each feature individually to determine the strength of the relationship of the feature with the label [9]. I adopted the univariate feature selection, the SelectKBest class in scikit-learn to the best model-logistic regression. SelectKBest is a method to score the features using a function, and in this case, the classification result, and then removes all but the k highest scoring features. I plotted the accuracy change against the number of features (see figure 4), and identified the number of the features that achieve 80% accuracy is 11 and for 90% accuracy is 46. Figure 4 shows the accuracy increases dramatically when several features reach around 20 and reached a peak of accuracy as 90% when the number of features is approaching 30.

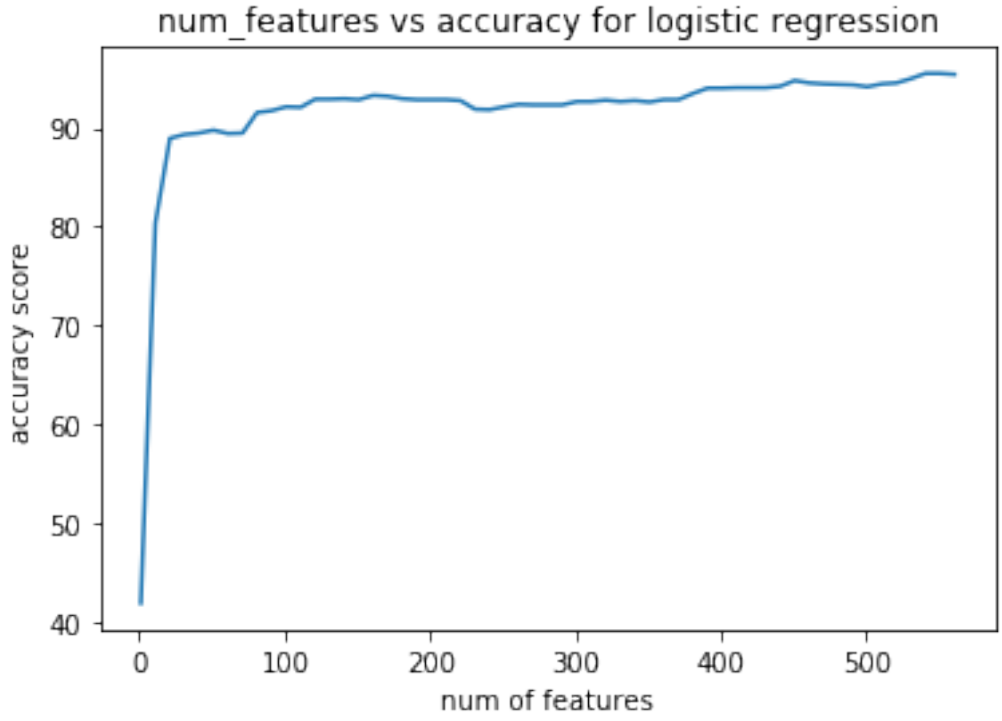

Figure 6: Number of Features vs Accuracy for Logistic Regression.

Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. Principal Component Analysis (PCA) is one of the main techniques for dimensionality reduction. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [10]. Using PCA method, the full features are deducted to 7 principal components to achieve 80% accuracy and 30 principal components to achieve 90% accuracy. The results are shown in figure 5.

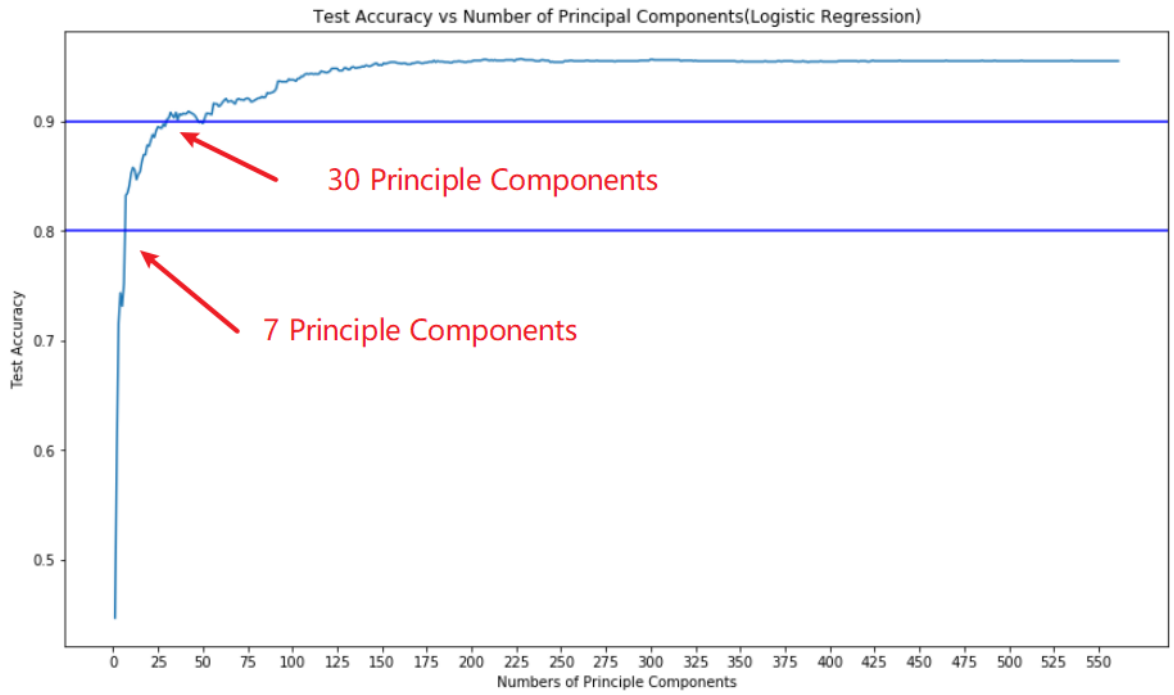

Figure 7: Principal Component Analysis (PCA) for dimensionality reduction to achieve 80% and 90% accuracy .

Conclusion

Two models, logistic regression and random forest are used for human activity recognition and achieved 95.41% and 90.82% accuracy respectively. Interestingly, the logistic regression performs better than the random forest model, some potential reasons are provided, but the in-depth exploration remains for the future. Using the feature selection methods, the univariate technique, I am able to detect the variation between number of features and accuracy and identified the feature that reaches 80% and 90% accuracy. Using the PCA method, the dimensionality is deducted to 7 and 30 principal components to reach the 80% and 90% accuracy. Other machine learning model is encouraged to contribute to studying the problem and improving the accuracy results for the future study.

Reference

[1] Rasekh, Amin & Chen, Chien-An & Lu, Yan. (2014). Human Activity Recognition using martphone.

[2] Ranasinghe, S., Al Machot, F., & Mayr, H. C. (2016). A review on applications of activity recognition systems with regard to performance and evaluation. International Journal of Distributed Sensor Networks. https://doi.org/10.1177/1550147716665520

[3] UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smart_phones

[4] Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013, April). A public domain dataset for human activity recognition using smartphones. In *21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN* 2013. Bruges, Belgium, 24-26, April 2013.

[5] Kim, E., Helal, S., & Cook, D. (2009). Human activity recognition and pattern discovery. IEEE pervasive computing, 9(1), 48-53.

[6] Sperandei S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12–18. doi:10.11613/BM.2014.003

[7] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[8] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

[9] https://blog.datadive.net/selecting-good-features-part-i-univariate-selection/

[10] Jolliffe, I. (2011). Principal component analysis (pp. 1094-1096). Springer Berlin Heidelberg.