COMPARATIVE STUDY ON VARIOUS ALGORTHMS FOR DETECTION OF FAKE JOB POSTINGS

Dhanamma Jagli[1], Vishal Saroj Gupta[2]

[1]Assistant Professor, VESIT, Mumbai University, Mumbai, Maharashtra, India.
[2]MCA Final Year Student, VESIT, Mumbai University, Mumbai, Maharashtra, India.

ABSTRACT: We are living in such a situations where the unemployment rates is as high as it could be in anytime in near future. Thousands of employees have been removed on a daily basis creating a global effect on dependencies between the company and their employees.

In these frantic occasions, when thousands and a huge number of individuals are keeping watch for a vocation, it gives an ideal chance to online con artists to exploit their distress. The aim of this study is to compare of the most popular models available for detecting the accuracy, which helps to identify the jobs which are true or fake. Choosing an analysis model is necessary and can be difficult given the surplus of choices for this study, as it is used more than one model at a time to take advantage hide disadvantage of some models.

The selection of a good Analysis model will provide effortless performance of models in the system to deliver the best result. In this paper, we will see the various aspects of these seven models for analysing their ROC AUC and accuracy of the models. The comparison between the seven models will be done based on various parameters that can help analyse and decide which model will be better suited for different aspects.

Keywords: **Fake Job Postings, Logistic Regression, Support Vector Classifier, MultiLayer Perceptron Classifier, KNN Classifier, Decision Tree Classifier, XGBoost Classifier, Random Forest Classifier, ROC AUC and Accuracy.**

INTRODUCTION:

In order to have a high-quality model which predicts the highest accuracy, it's important that proper cleaning of the text have been done. The data I will be using for this analysis is a dataset of 18K job descriptions compiled by the University of the Aegean, Laboratory of Information & Communication Systems Security (http://emscad.samos.aegean.gr/). This dataset contains records which were manually annotated and classified into two categories. More specifically, the dataset contains 17,014 legitimate and 866 fake job descriptions.

THE PROBLEM STATEMENT:

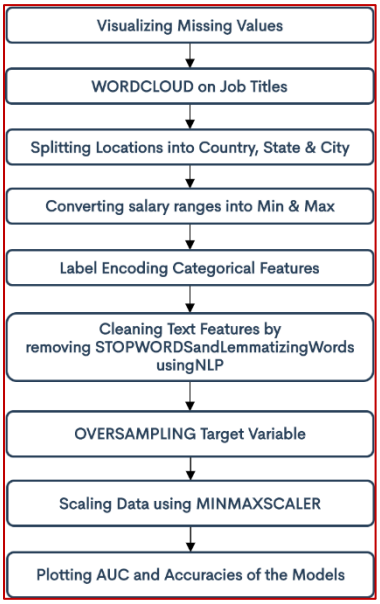
We see a day by day ascend in these phony employment postings where the posting appears to be really sensible, frequently these organizations will have a site also, and they will have an enlistment procedure that is like different organizations in the industry.

If one looks sufficiently hard, they can recognize the contrasts between these phony postings and real ones. More often than not these postings don't have an organization logo on these postings, the underlying reaction from the organization is from an informal email account, or during a meeting they may approach you for individual secret data, for example, your charge card subtleties by saying they need it for work force check.

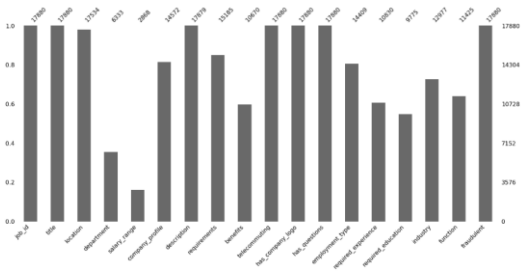
In typical financial conditions, all these are obvious indications that there something dubious about the organization, yet these are not ordinary monetary conditions. These are the most exceedingly awful occasions we as a whole have found in the course of our lives, and as of now, frantic people simply need an occupation, and by this, these people are straightforwardly giving way to the schemes of these tricksters.

PROPOSED MODEL ARCHITECTURE:

The architecture is shown as in the below:



Visualizing Missing Values:



Word Cloud for the dataset used:

IMPLEMENTATION:

ROC AUC:

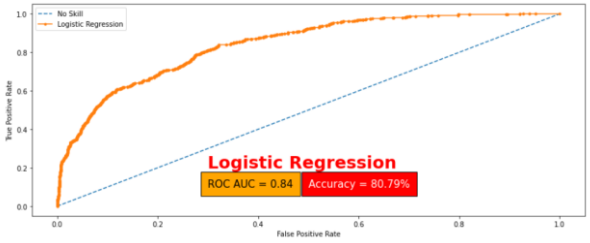
ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the "ideal" point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.The "steepness" of ROC curves is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate.

ROC curves are typically used in binary classification to study the output of a classifier. In order to extend ROC curve and ROC area to multi-label classification, it is necessary to binarize the output. One ROC curve can be drawn per label, but one can also draw a ROC curve by considering each element of the label indicator matrix as a binary prediction (micro-averaging).

Models used for the study:

1). Logistic Regression:

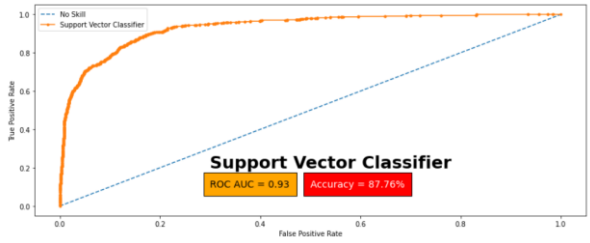
Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).



The Accuracy of this model is **80.79%.**

2). Support Vector Classifier:

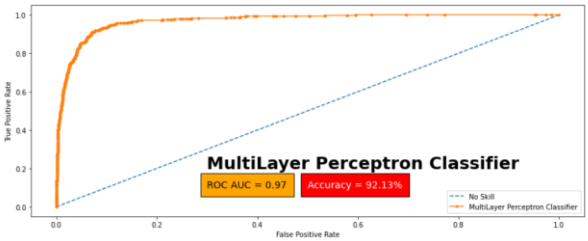
"Support Vector Classifier" is a supervised machine learning algorithm which can be used for both classification and/or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.



The Accuracy of this model is **87.76%.**

3). MultiLayer Perceptron Classifier:

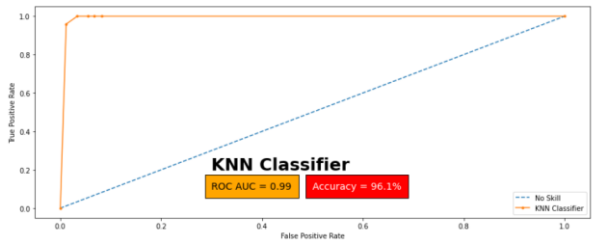
Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function \(f(\cdot): R^m \rightarrow R^o\) by training on a dataset, where \(m\) is the number of dimensions for input and \(o\) is the number of dimensions for output. Given a set of features \(X = {x_1, x_2, ..., x_m}\) and a target \(y\), it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers.



The Accuracy of this model is **93.30%.**

4). KNN Classifier:

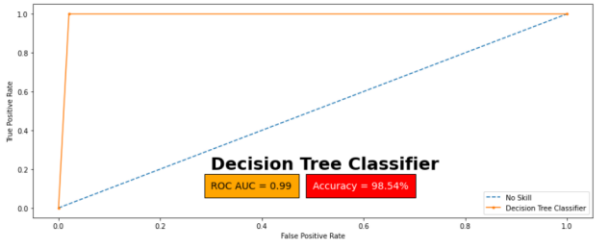
K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors.



The Accuracy of this model is **96.10%.**

5). Decision Tree Classifier:

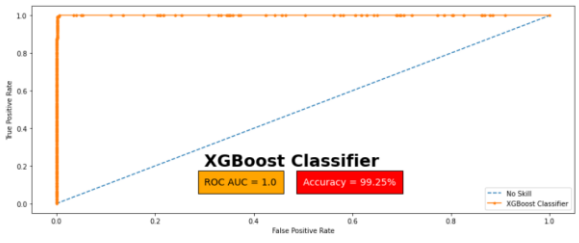
The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.



The Accuracy of this model is **94.83%.**

6). XGBoost Classifier:

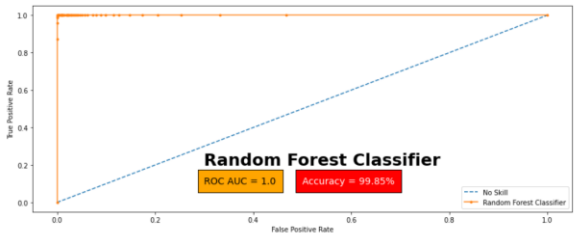
XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.



The Accuracy of this model is **99.25%.**

7). Random Forest Classifier:

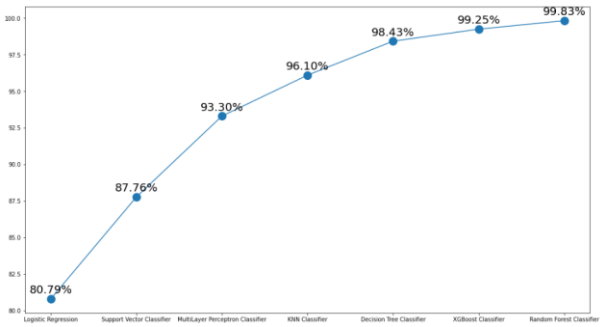
The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.



The Accuracy of this model is **99.93%.**

CONCLUSION

In this research, the results of the fake job posting dataset problem using the seven machine learning algorithms we came to conclusion that random forest classifier has the highest model to achieve the best accuracy.



REFERENCES

[1] V. Aswini and S. K. Lavanya, "Pattern discovery for text mining," 2014 Int. Conf. Comput. Power, Energy, Inf. Commun., pp. 412–416, 2014.

[2] L. Ge and T. S. Moh, "Improving text classification with word embedding," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-January, pp. 1796–1805, 2017.

[3] Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "Duplicate Questions Pair Detection Using Siamese MaLSTM," IEEE Access, vol. 8, pp. 21932–21942, 2020.

[4] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," 2016 1st IEEE Int. Conf. Comput. Commun. Internet, ICCCI 2016, pp. 471–475, 2016.

[5] C. Saedi, J. Rodrigues, J. Silva, A. Branco.