Bank Note Validation Using Machine Learning

Abdul Wahid, *Artificial Intelligence and it's Applications, University of Essex,*

Abstract—The advancement in technology has not only made things easier for us but it has also given rise to some unintended issues for example in case of bank notes it has made it easier to make counterfeiters by using modern tech available and they are so realistic that most of the people can not distinguish them from fake which is a huge concern for banks. Way out of this problem would be an automated system that can distinguish real from fake with good accuracy. In this paper we will be looking into different Machine Learning methods to resolve this problem [6]. We are using classification methods in order to solve this problem [3]. The best method was selected on the basis of confusion matrix [13] with best feature selection using Grid search [16]. After implementation of different models, SVM and K Nearest Neighbour came on top as they successfully predicted authenticity of the bank notes with full accuracy.

Index Terms—Confusion Matrix, Grid Search, ROC and AUC curve, ASM(Attritube Selection Measure), Desission Tree, Random Forest, Logistic Regression, KNN(K nearest Neughbour) SVM(support Vector Machine), TP(True Positive), TN(True Negative, FP(False Positive), FN(False Negative).

1 Introduction

THE counter measures against "forgery" remains the consequential matter for the banks to this day. Banks are really important for the economical stability. They provide assistance in inflation state by increasing flow of money in economy [2]. They are a Financial Institution that circulates money by the process of borrowing and lending money.

Banks borrow money from people in the form of deposit and provide interest on that money to the person who deposited. On the other hand they lend money to people who need it as a loan and take a certain amount of interest on that money. This all needs to be in balance and banks are really considerate on whom they lend the money [7] which goes to show the importance of money for the economy which is the exact reason why they need to make sure that the money being deposited is not forged as it will have to be removed from the circulation [10].

There have been many different approaches used in order to solve the problem at hand for example use of Neural Networks and other machine Learning algorithm [14]. We will be explicitly using different Machine Learning models to make a comparison between supervised and unsupervised algorithm in bank note validation.

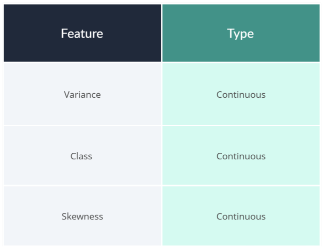
2 Literature Review

There have been different projects done similar to this which focus on the difference between classification and regression method [14] [3] [13]. Mostly image classification has been performed on the data set and then classifying the bank note. We are working with the numeric data set of results and want to train our model to get us best accuracy in terms of right classification of bank note. We will be using the classification methods to solve the problem as the final outcome needs to be Boolean.

3 Dataset Description

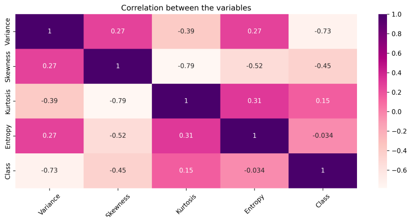
Dataset used in this report is from the UCI Machine Learning respository [17]. The dataset is in the text format and consists of 1372 rows and 5 coulmns. In Figure[1] the overall shape of the dataset is defined. The target class is divided into bool values 0 and 1. The relationship between different Varibles is also represenetd using Heatmap is Figure[2]

Fig. 1. Dataset Description



Feature	Type
Variance	Continuous
Class	Continuous
Skewness	Continuous

Fig. 2. Dataset Description



4 Methodologies

classification is a two step problem, learning and prediction. In learning we train our model on the dataset and when it comes to prediction we test our model on the given data and check the accuracy which is our measure of how good the model is performing. The duplictaes and null values were removed from the dataset. We are dividing the datasets into training and validation dataset in the ratio of 70:30. We will be implementing five different Machine Learning methods to find the best one possible. We will also apply confusion Matrix, Classification Report, ROC and AOC curve to determine the best method on the basis of the results. The attributes for all of the methods were optimized by performing Grid Search to get best accuracy.

4.1 Decision Tree Classifier

Decision tree classifiers are good at efficiency and accuracy when performing on large datsets [15].It is a Supervised Machine Learning model. Decision tree works as a node and branch framework like a tree. Node represents the start and branches are the splits or partitions. Decision Tree uses ASM(Attritube Selection Measure) to split the dataset by creating one parent or root node and rest of the dataset into smaller subsets or child nodes and keeps on repeating the process until it satisfies the requirements [1].

Decision Tree performed really well on the dataset in terms of both accuracy and efficiency. The Confusion Matrix of decision tree had 4 wrongly categorized exmaples 2 True positive and 2 True Negative as shown in the Figure[3]. The decision Tree received accuracy of 99 percent. The Roc and AUC graph projects the outcome of the decision Tree in Figure[4]

Fig. 3. Decision Tree Classifier Confusion Matrix

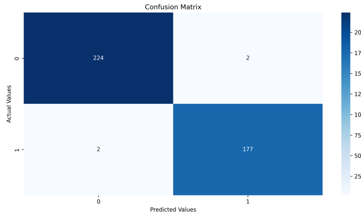
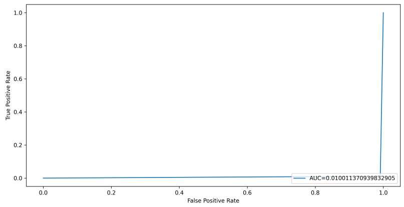


Fig. 4. Decision Tree Classifier ROC and AUC Curve



4.2 Logistic Regression

Logistic Regression is a Supervised machine Learning model for classification problems [9]. It is mostly used for predicting boolean outcomes where the variables are independent of each other and are linearly related to log of odds. [11].

Even though logistic regression had accuracy of 99 percent it was made 5 wrong classifications as shown in confusion matrix in the Figure[5]. The Roc and AUC graph projects the outcome of the Logistic Regression in Figure[6]

Fig. 5. Logistic Regression Confusion Matrix

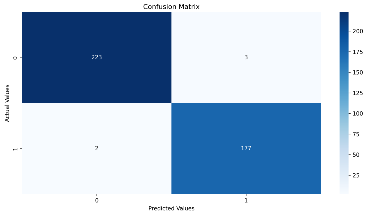
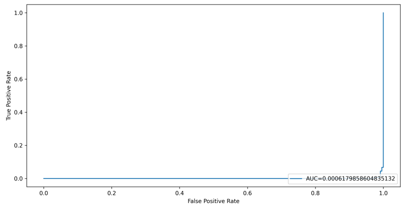


Fig. 6. Logistic Regression ROC and AUC Curve



4.3 Random Forest Classifier

Random Forest Classifier is a Supervised machine Learning model for classification problems [12]. It creates multiple tress like a forest of all possible combinations on basis of categorial variables in dataset.

Random Forest Classifier had accuracy of 100 percent and it only made one wrong classifications as shown in confusion matrix in the Figure[7]. The Roc and AUC graph projects the outcome of the Random Forest Classifier in Figure[8]

Fig. 7. Random Forest Classifier Confusion Matrix

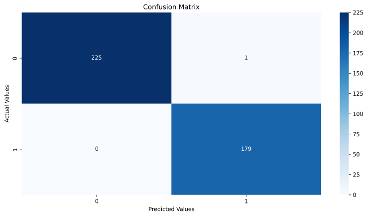
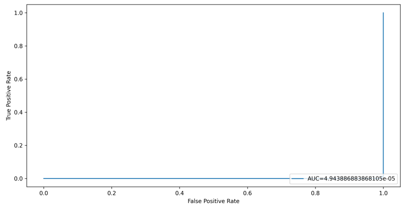


Fig. 8. Random Forest Classifier ROC and AUC Curve



4.4 K Nearest Neighbour Classifier(KNN)

KNN is a Supervised machine Learning model for classification problems and is one of the miost simplest algorithm to work with [8]. It is no linear algorithm which uses labeled parameters for classification. The value of K determines the number of the classification labels. [5].

KNN was correctly able to calculate all the outcomes correctly with 100 percent accuracy and no false categorization as shown in confusion matrix in the Figure[9]. The Roc and AUC graph projects the outcome of the KNN in Figure[10]

Fig. 9. KNN Classifier Confusion Matrix

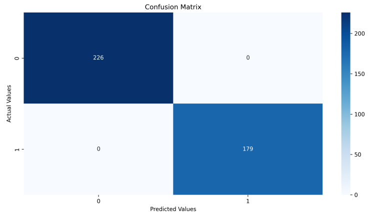
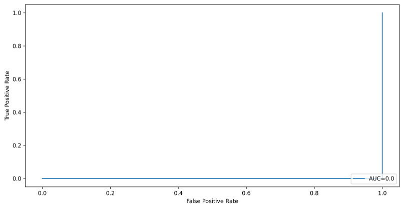


Fig. 10. KNN Classifier ROC and AUC Curve



4.5 Support Vector Machine (SVM)

SVM is a Supervised machine Learning model for classification problems that is really good and accurate on small datasets [4]. It is fast and reliable and good to work with textual datasets in the values of thousands for classification problems. Support Vector Machine divides the data points in the hyperplane using a line which is called the decision boundry.
SVM was the fastet alogorithm in terms of calculation and had 100 percent accuracy as shown in confusion matrix in the Figure[11]. The Roc and AUC graph projects the outcome of the SVM in Figure[12]

Fig. 11. SVM Confusion Matrix

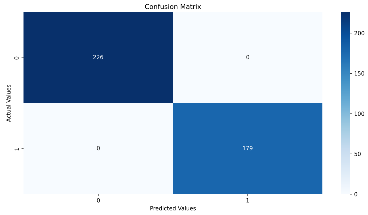
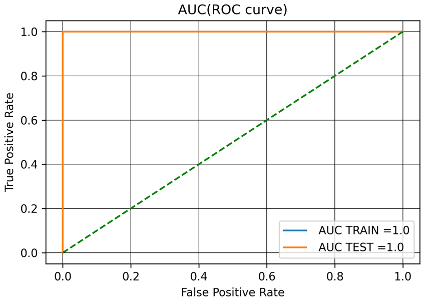


Fig. 12. SVM ROC and AUC Curve



5 Results

The performance and Evaluation of the models was done on the basis of correlation Report as shown in Table[1]. The prameters for evaluation are as following

- Accuracy - Ability to correctly predict outcome on given data

  Formula for Accuracy = **(TP + TN)/(TP + TN + FP + FN)**

- Precision - It is Measure of Quality

  Formula for Precision = **TP/(TP + FP)**

- Recall - It is Measure of Quantity

  Formula for Recall = **TP/(TP + FP)**

- F1-Score - Gives us good Accuracy Measures as compared to Accuracy

  Formula for Recall = **2*[Precision * Recall/(Precision + Recall)]**

Model	Accuracy	Dataset	Precission	Recall	F1-score
Decission Tree Classifier	0.99	Actual Trained	0.99 0.99	0.99 0.99	0.99 0.99
Logistic Regression	0.99	Actual Trained	0.99 0.98	0.99 0.99	0.99 0.99
Random Forest Classifier	1.00	Actual Trained	1.00 0.99	1.00 1.00	1.00 1.00
K Nearest Neighbour	1.00	Actual Trained	1.00 1.00	1.00 1.00	1.00 1.00
Support Vector Machine	1.00	Actual Trained	1.00 1.00	1.00 1.00	1.00 1.00

TABLE 1
Classification Report

6 Discussion

The dataset was trained and implemented on total five Machine Learning algorithm with optimized parameters using Grid Search. Some algorithms were more efficient in terms of speed but some were effective in terms of accuracy.

Field of data Science is very waste and there is still room for improvement. Overall SVM and KNN are the most best suitable algorithms. In the future, I would also like to experiment with MLP and Deep Learning models as well as Neural Network.

7 Conclusion

Almost Every method we used had it's on Strong points and draw back as well.Parameters for all the models were optimized by using Grid Search in order to get best outcome. Different methods are viable for different type of Data set and Problem. As for this sort of dataset and Problem, no doubt that SVM and KNN came on top. KNN is a less effective as it is a bit time consuming but it is accurate and is not restrained by the length of the dataset. But SVM can only work with limited amounts of datasets for best results. Both of them were equally good in terms of accuracy and were able to correctly predict real and fake Note without any Failure with 100 percent Accuracy. Random Forest classifier also had an accuracy of 100 pdercent but it as much more time consuming as compared to any other model we used in this report.

Acknowledgments

I would like to thank our module supervisor 'Haider Raza' as for a student who didn't had any experience with data science and in small time period actually enabling us to properly do a research. It had created an interest for me to opt Data Science as my profession. The course itself was really informative and all the material was well thought and organized for even a beginner.

References

[1] Data Camp. Origins — datacamp. https://www.datacamp.com/tutorial/decision-tree-classification-python.

[2] Asli Demirguc-Kunt, Erik Feyen, and Ross Levine. The evolving importance of banks and securities markets. Technical report, National Bureau of Economic Research, 2012.

[3] Bahareh Ghasemain, Dawod Talebpoor Asl, Binh Thai Pham, Mohammadtghi Avand, Huu Duy Nguyen, and SJVJOES Janizadeh. Shallow landslide susceptibility mapping: A comparison between classification and regression tree and reduced error pruning tree algorithms. *Vietnam Journal of Earth Sciences*, 42(3):208–227, 2020.

[4] Vikramaditya Jakkula. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5):3, 2006.

[5] Sero Kell. Origins — sero kell. https://serokell.io/blog/knn-algorithm-in-ml.

[6] Chhotu Kumar and Anil Kumar Dudyala. Bank note authentication using decision tree rules and machine learning techniques. In *2015 International Conference on Advances in Computer Engineering and Applications*, pages 310–314. IEEE, 2015.

[7] Michael Manove and A Jorge Padilla. Banking (conservatively) with optimists. *The RAND Journal of Economics*, pages 324–350, 1999.

[8] Antonio Mucherino, Petraq J Papajorgji, and Panos M Pardalos. K-nearest neighbor classification. In *Data mining in agriculture*, pages 83–106. Springer, 2009.

[9] Todd G Nick and Kathleen M Campbell. Logistic regression. *Topics in biostatistics*, pages 273–301, 2007.

[10] Bank of England. Origins — bankofengland. https://www.bankofengland.co.uk/banknotes/counterfeit-banknotes.

[11] Capital One. Origins — capital one. https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/.

[12] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.

[13] E Pantaleoni, RH Wynne, JM Galbraith, and JB Campbell. Mapping wetlands using aster data: a comparison between classification trees and logistic regression. *International Journal of Remote Sensing*, 30(13):3423–3440, 2009.

[14] Sumeet Shahani, Aysha Jagiasi, and RL Priya. Analysis of banknote authentication system using machine learning techniques. *International Journal of Computer Applications*, 975:8887, 2018.

[15] Philip H Swain and Hans Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, 1977.

[16] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4):1502–1509, 2016.

[17] UCI. Origins — uci.