Automation Tool Development to Improve Machine Results using Data Analysis

Tanushka Malhotra[1]

[1]*Department of Computer Science and Engineering, SRM University, Chennai, India*

Abstract - *The volume of data generated by different organizations has increased consistently in the past few years. Hence, it is necessary to analyze this voluminous data to produce profitable results. This paper presents a way of analyzing merchant datasets using the confusion matrix concept to attain the maximum precision, recall and accuracy for the given merchants. Moreover, various python libraries have led to a better automation tool algorithm for analysis.*

Key Words: **Data Analysis, Pattern Analysis, Big Data, Automation Tool, Confusion Matrix, Precision, Recall, Accuracy**

1. INTRODUCTION

Data Analytics[1] is the science of examining the data contained in datasets and drawing out meaningful patterns and correlations with the help of various technologies and software's. Big data analysis provides an organization with efficient operations and higher profits. Figure 1 shows the advantages of big data analysis performed by an organization.

Cost Reduction: The technologies used for big data analysis such as cloud technologies help organizations to increase their revenue and reduce the costs.

New Product and Services: Big Data Analytics has provided the companies the ability to produce new products and services to offer to their customers. These new products suffice the customers' needs and help the companies to gain profits.

Better and faster decision making*:* The availability of new technologies and data analytics tools have provided companies the ability to quickly analyze data and produce accurate and meaningful results.

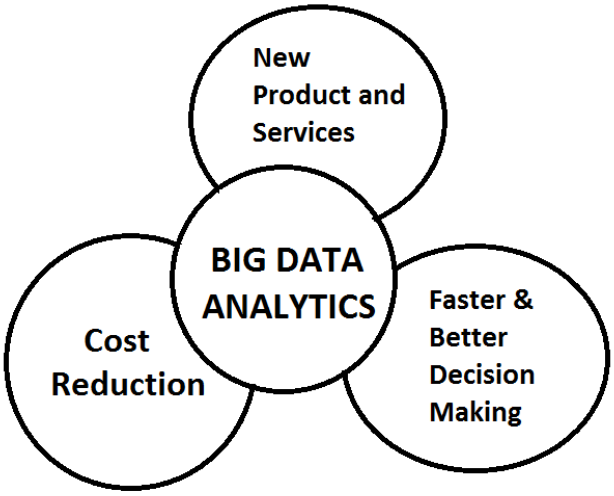


Figure 1: Advantages of big data analysis

1.1 Big Data

Big data[2] is a term that is most often used today to denote the amount of data that is generated in almost every sector due to the advent of technology and communication means. It is in the recent years that maximum amount of data is generated and the term "Big Data" has been coined for it. There are a lot of challenges like capturing data, duration, storage, searching, sharing, analysis and presentation, which are linked to big data but the organizations are finding different methods and tools to deal with these challenges.

1.2 Python and its libraries used for Analysis

There are many in-built libraries in python that help to efficiently analyze the big data. The most common and resourceful ones are manifested in the following sections.
Pandas

Pandas is one of the most popular data science library. It is able to read data from different sources and handles them in the form of data frames which can then be used to perform different kinds of analysis. It also possesses the feature of plotting charts and graphs as well as export functions. Export functions lets you generate an excel file from the results of your analysis.

Numpy/Scipy

Numpy stands for Numerical Python and Scipy is a library built upon Numpy. Both of these are open source libraries used for scientific computing and mathematical functions like integration, Fourier Transform, linear algebra, etc.

Matplotlib

Matplotlib is a library for creating simple 2D graphs and plots.

Re

This is a regex (regular expression) module which is used for easy pattern searching in data analytics.

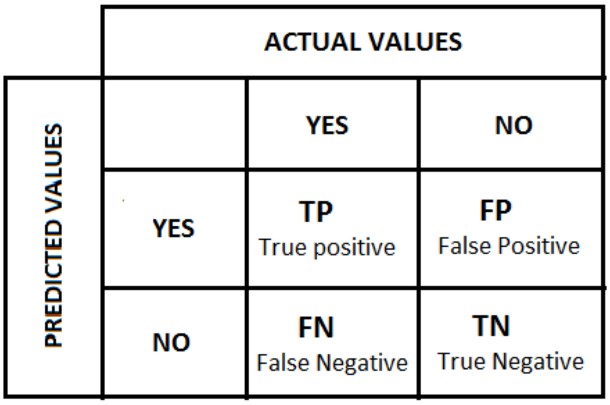
Fuzzywuzzy[4]

It is used for fuzzy string matching, i.e., find the string that matches approximately with the pattern.

2. CONFUSION MATRIX

Confusion matrix[5], as the name suggests, is a matrix which contains the actual and the predicted values of the classification done by a classification algorithm. This matrix is used to measure the performance of a machine. Figure 2 shows the confusion matrix.

TP: This stands for true positive. It means that the predicted value matches the actual value and our result is correct.
FP: This stands for false positive. It means that the predicted value is incorrect with respect to the actual value.
FN: This stands for false negative. It means that there is no value predicted when the actual value says that there should be some value predicted.
TN: This stands for true negative. It means that there is no actual value and the predicted value correctly predicts the no-event value.



		ACTUAL VALUES	
PREDICTED VALUES		YES	NO
	YES	TP True positive	FP False Positive
	NO	FN False Negative	TN True Negative

Figure 2: Confusion Matrix

3. PARAMETERS FOR PERFORMANCE MEASURE

Precision

Precision calculates the percentage ratio of the number of observations that were predicted correct with respect to the actual value and the number of observations predicted true irrespective of the actual value.

Precision = TP/ (TP+FP)

Recall

Recall describes the percentage ratio of the number of observations that were predicted correct with respect to the actual value and the actual cases predicted as yes.

Recall = TP/ (TP+FN)

Accuracy

Accuracy is defined as the percentage ratio of the number of true cases and the sum of all the cases.

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

4. EXPERIMENT AND OBSERVATIONS

There are two merchant datasets. One is the factual data, which contains the merchant city, state, address, phone number, store id and many more details. The other one is a machine generated result which contains the merchant description, city, state and store id. The results generated by the machine are not accurate enough and hence various patterns have to be deduced as well as analysis is to be performed to find out the flaws in the machine results. These kinds of deductions and pattern analysis[7] leads to an increase in the accuracy of the overall results. The main aim of this experiment is to create an automation tool for the aforementioned analysis.

The analytical automation tool is developed using python and its various libraries. The tool handles various parameters for comparison and finally judges whether the predicted result by the machine was a TN case, FN, case, TP case or an FP case using the confusion matrix technique. These parameters are stated below.

- The merchant description contains partial/complete city name and that city has only one store id in the factual.
- The store id is a part of merchant description and is also present in factual.
- There is partial address match in description and there is only one id for that address in factual.
- Wrong store id is predicted by the machine.
- If there is more than 1 store present for a given city in factual and there is nothing mentioned in the merchant description, then the machine cannot predict the store id.
- If the phone number of a merchant in factual matches with the merchant description.

The automation tool was run on 5 different merchants and it was observed that the precision, recall and accuracy were improved by the analysis of the automation tool. Figure 3 shows the percentage of precision and Figure 4 shows the percentage of recall for the machine generated results and the analysis done by the automation tool on the machine results. Furthermore, Figure 5 shows the percentage of accuracy for the same analysis.

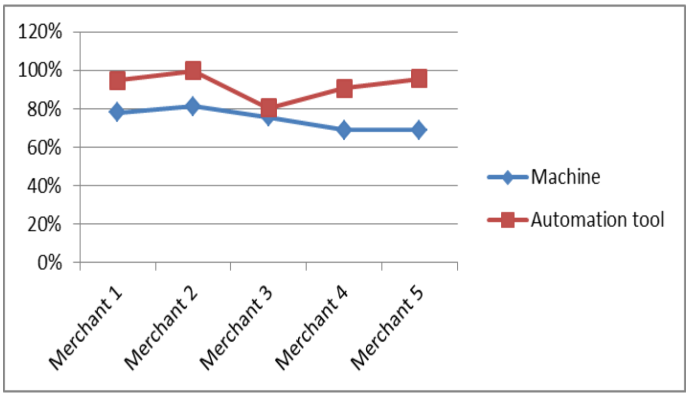


Figure 3: Precision of machine results and automation tool results

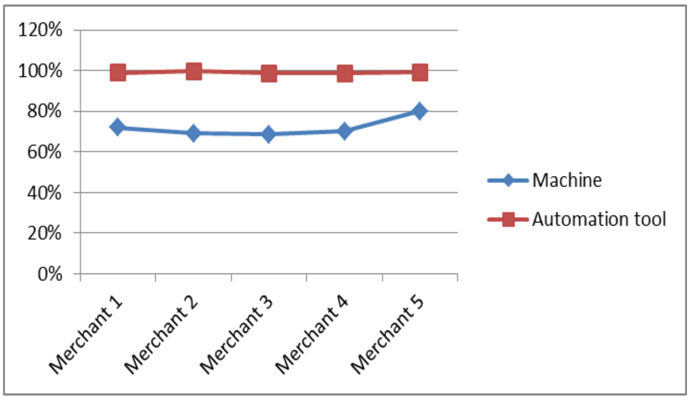


Figure 4: Recall of machine results and automation tool results

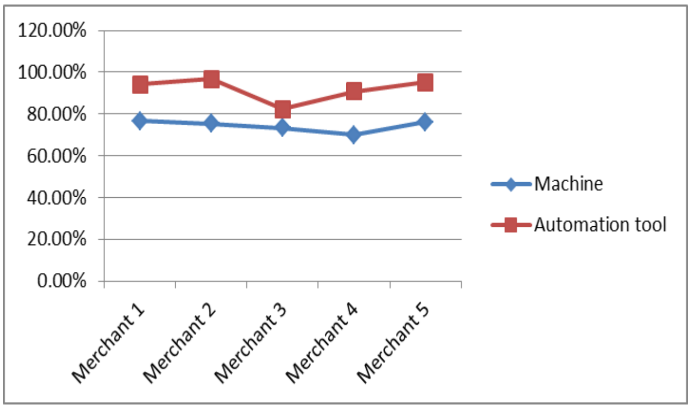


Figure 5: Accuracy of machine results and automation tool results

5. CONCLUSIONS

In this world of increasing technology and social media, where vast amount of data is being spawned, the advent of data analytics has helped organizations to manage their data and use it in a resourceful manner. This paper discusses the development of an automation tool using python. This automation tool can be run on datasets having thousands of entries and can produce meaningful results in a few minutes. Furthermore, the accuracy of results generated is quite high and these results can further be used to improve the machine. This kind of analysis can be used by various organizations to gain profit from merchants. The data from different merchants is collected and analyzed by profit-making organizations who then sell their predictions and analysis back to the merchants.

The confusion matrix is a perfect method to calculate various performance parameters and python, with its inbuilt libraries, offer a convenient platform to develop the analytical tool. Based on the analysis, it is observed that the machine results are less accurate and the patterns & deductions made by the automata result in an improved accuracy.

REFERENCES

[1] https://www.sas.com/en_us/insights/analytics/big-data-analytics.html.

[2] Significance and Challenges of Big data Research, Volume 2, Issue 2, June 2015.

[3] Python for Scientists and Engineers, Volume 13, Issue 2, March-April 2011.

[4] https://pypi.python.org/pypi/fuzzywuzzy.

[5] Genetic Algorithm and Confusion Matrix for Document Clustering, International Journal of Computer Science Issues, Vol. 9, issue 1, No. 2, January 2012.

[6] Performance Evaluation of Predictive Classifiers for Knowledge Discovery From Engineering Materials Data Sets, Doreswamy, Hemanth K.S., 2003.

[7] Jain, A., and Zongker, D. 1997. Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence.