Secure and Scalable Data Mining Technique over a Restful Web Services

*Francesco Solar[1]
Department of Computer Science
Complutense University of Madrid
Madrid, Spain

Smith Oliver[2]
Department of Computer Science
University of Western Australia.
Crawley WA 6009, Australia

***Abstract*— Scalability, efficiency, and security had been a persistent problem over the years in data mining, several techniques had been proposed and implemented but none had been able to solve the problem of scalability, efficiency and security from cloud computing. In this research, we solve the problem scalability, efficiency and security in data mining over cloud computing by using a restful web services and combination of different technologies and tools, our model was trained by using different machine learning algorithm, and finally using different validation and cross-validation methods to compare and validate our results. We build, train, test and deploy our model by restful API services which can be called through an endpoint, and with very high accuracy and huge success.**

***Index Terms*— Data mining, Restful API, Web services, cloud computing, machine learning.**

I. INTRODUCTION

Over the years, several methods and techniques had been proposed and implemented in an attempt to address issue of security, scalability and efficiency in data mining[17,18,19,20,21]. Despite all the methods and techniques, the issue of security, scalability and efficiency remains challenging problems still in need of solutions in the industrial. The reason being that those proposed technique fails to provide permanent solution to the problem

The fact that we are in an era of cloud computing cannot be overemphasize, and as more companies are shifting to cloud computing due to availability of Infrastructure As A Service (IAAS) [1,2,3,4,5,6], Platform As A Service(PAAS) and Software As A Service(SAAS) which all can be access over the internet. This bring about huge gap of challenge in security and scalability as we are in an era in which availability of data is paramount to making innovation and discovery in statistical artificial intelligence, inventing new things, and advancing some of the existing methods in both the research and industrial world.

As scientist and industries geared towards data mining on cloud computing, there is increasing demand on security due to increase in phishing through cyberspace whereby internet fraudster uses social engineering (a series of techniques commonly used by scammers to manipulate human psychologically) to trick people by sending malicious email to them. Most times, victims do end up sending sensitive, personal, and corporate information to them. While we have some in the mold of email, some are through malicious website, software, and social media platform. Between year 2013 and 2015, Facebook and Google were both tricked of a sum of $100 million through phishing attack by an extended campaign when Quanta was impersonated by a phisher to trick Facebook and Google in which both tech giants fell victims. While in 2016, Cretan bank in Belgium fell victim of business email compromise known as BEC in which the bank was ripped of $75.8 million. In the year 2016, an Austrian based aerospace part manufacturing industry. Austrian Aerospace Component Manufacturer lost $61 million to Business Email Compromise (BEC) scam which pushes the company to the wall as the company closed the financial fiscal year with a loss of $26.6 million.

While modern machine learning and data mining model should be able to learn and adjust itself overtime from based on existing data with very little or no human intervention, the problem of scalability coupled with performance issue which comes with this activities cannot be over emphasize. In order to resolve this issue, we implemented data mining over restful API (Application Programming Interface) services. We used different machine learning and data mining algorithm to build, train, and test our model through a restful API services with huge amount of success which completely solve the problem of security, scalability and performance. The model was also deployed in such a way that they could be called by using an endpoint from any device.

II. BACKGROUND STUDY

As scientists and industry shifted to cloud computing which is being delivered over the internet, scalability, efficiency, performance, and security becomes and issues which can never be overemphasize as it includes; unlimited storage, provisioning and updating, guaranteed privacy more secure (Ko et al., 2011). Also, it is possible for users of cloud services to optimize server utilization, dynamic scalability, and minimize the development of new application life cycles (Al-Ruithe et al., 2019). In addition to the numerous benefits in cloud computing, there are also problems as a result of cloud outage since data storage is centralize in the cloud

which can paralyze a company business (Gupta and Gupta, 2014), also attack on the integrated cloud environment can cause loss of data and finance for both the service providers and subscribers (T.K and B, 2016). There are also other risks associated with computing on cloud environment which includes the issues of threats to data security information confidentiality (Paquette et al., 2010, Wang, 2011), and the possibility of information leakage and vulnerability (Inuwa, 2015, Tchernykh et al., 2019).

Some of the problems associated with data mining on cloud computing are as a result of the existing method adopted in the industry and scientific world at large. Current cloud computation for data mining needed service provider to provide interface for user. The user does not need to bother about the infrastructure. It enables the retrieval of useful previously unknown data from integrated data warehouse, in such a way that users doesn't need to border about infrastructure, storage, or configuration and maintenance, the provider handles that. It is based on retrieving directly from the integrated data warehouse.

Ruxandra-Ștefania PETRE in his research work "Data mining in Cloud Computing", relies on extraction of previously unknown or meaningful pattern from unstructured or semi-structured data from the web sources. "the analysis steps in the Knowledge Discovery and Databases process" (Nodine, Ngu, Cassandra and Bohrer, 2003). It listed three stages of research involving data warehouse which are staging, Integration, accessibility for the purpose of reporting and analysis. In the Review of Data Mining Techniques in Cloud Computing Database by Astha Pareek etal.

In addition to the problems associated with data mining on cloud computing in which majority of the problems are due to existing methods of data mining on cloud computing. The existing methods also creates a wide gap between data mining on cloud computing and application Programming Interface (API) which are yet to be closed. In the existing method we are yet to see an instance in which we call we only need to call an API endpoint, and then have everything done for us.

III. RESEARCH METHODOLOGY

For this research, we used data obtained from social network from github. To begin this research, we did four basic things;

1. Subscription to IBM Cloud Object Storage facility to host our CSV file

2. Setting up a data warehouse on cloud from another channel different from IBM

3. Developed a background window service using .NET Technology

4. Wrote a web service to be consume in the cloud using python

5. Scheduler using python scheduler library to trigger the web API at interval.

In order to prevent the web service from being overwhelm due to multiple calling of the endpoint at regular interval, we developed a background window service to support the restful API service. Both the web API and the window service are picking records from the integrated data warehouse and pushing to the CSV file on the IBM Object Cloud Storage facility. They automatically pick new records to the CSV, and if a record is modified, it will be picked and modified on the CSV as well. The essence of the scheduler which was written in python is to be calling the web API at regular interval to check and push from the integrated data warehouse to the CSV file.

Having set up our cloud environment and the necessary programs written, we proceeded by adopting the following data cleansing and preparation techniques;

Data cleaning. We use preprocessing and cleaning methods to remove incomplete data that might cause system failure and also affect output prediction. Rows containing missing values where completely removed. We also use different methods to identify and remove noisy data, outliners, and other factors which can influence the output result

Data Reduction for Data Quality. In order to maintain data integrity, we needed to deal with all rows containing null value; hence we opted for python library tool called PyCaret. We have two options which is either to automatically fill all the null values or to remove any row(s) containing null values weighted. Having weighted the risk involves in both, we decided to remove any row with null or empty value from the data, and this was done using PyCaret python library

Data Transformation. For us to make our data to acceptable format for easy data mining and pattern recognition, it needed to undergo some data transformation. To ensure data is fully transformed to acceptable format, we used Discretization, normalization, and data aggregation technique.

Data Mining. Having successfully set up our apparatus which includes IBM cloud object storage facility, running background window service, active web service, scheduler, and with the data being thoroughly pre-processed and transformed.

Unlike current data mining in cloud computing process in

which, one needs to make direct connect to the data warehouse or directly call the csv file. We only called our restful API endpoint (Figure 1) which was developed and hosted on the cloud;

```
In [3]: data = pd.read_csv('http://cs4430.meritgbtech.com/customers/social')
        data.head()
```

Out[3]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

Fig.1. Preview of First five rows of dataset.

This ensures a higher level of security and control over the data, the only thing that needed to be called is the endpoint of our API which automatically displays the data as seen in Figure 1 displaying the first five (5) records in the data using python library in panda.

```
|--- feature_0 <= 0.63
|   |--- feature_1 <= 0.61
|   |   |--- feature_0 <= -0.16
|   |   |   |--- class: 0
|   |   |--- feature_0 >  -0.16
|   |   |   |--- feature_1 <= 0.40
|   |   |   |   |--- feature_1 <= -0.06
|   |   |   |   |   |--- class: 0
|   |--- feature_1 >  -0.06
|   |--- feature_1 <= 0.03
|   |   |--- class: 1
|   |   |   |   |   |   |--- feature_1 >  0.03
|   |   |   |   |   |   |   |--- feature_1 <= 0.26
|   |   |   |   |   |   |   |   |--- feature_1 <= 0.08
|   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.14
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |--- feature_0 >  0.14
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.24
|   |   |   |   |   |   |   |   |   |   |   |--- feature_1 <= 0.06
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |   |   |--- feature_1 >  0.06
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_0 >  0.24
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_1 >  0.08
|   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.28
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.14
|   |   |   |   |   |   |   |   |   |   |   |--- feature_1 <= 0.14
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |   |--- feature_1 >  0.14
|   |   |   |   |   |   |   |   |   |   |   |   |---  truncated
branch of depth 3
|   |   |   |   |   |   |--- feature_0 >  0.14
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_0 >  0.28
|   |   |   |   |   |   |   |   |   |   |   |--- feature_1 <= 0.13
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |   |   |--- feature_1 >  0.13
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_1 >  0.26
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |--- feature_1 >  0.40
|   |   |   |   |   |--- class: 1
|   |--- feature_1 >  0.61
|   |   |--- feature_0 <= -1.15
|   |   |   |--- class: 0
|   |   |--- feature_0 >  -1.15
|   |   |   |--- feature_1 <= 1.36
```

Fig.2. Text representation of the feature algorithm from the Decision Tree

Formatted Decision tree algorithm was used to train the model having slatted the data into two equal part, half of it to train the model and the remaining half for testing of the model;

```
feature_cols = ['Age','EstimatedSalary' ]
X = data.iloc[:,[2,3]].values y = data.iloc[:,4].values
#split the dataset into training and test
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.25, ran-
dom_state= 0) #perform feature scaling
from sklearn.preprocessing import StandardScaler sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train) X_test = sc_X.transform(X_test)
#fit the model in the decision tree classifier from sklearn.tree import Decision-
TreeClassifier classifier = DecisionTreeClassifier()
classifier = classifier.fit(X_train,y_train)
```

Fig.3. Code Snippet for important feature selection and training of model.

We are able to obtain 90% accuracy on testing our model, after which we decide to optimize for more accuracy and better performance as seen in Figure 4.

Fig.4. Code Snippet for our model accuracy and test performance

We are able to obtain accuracy of 94% after which we decided to visualize the performance of the model.

Fig.5. Validation, Cross validation and the Estimation of Mean Square Error (MSE)..

We needed to know the level of fitting, because having an accuracy of 94% can be as a result of over fitting of the model with the training set of data, hence we decided to do two additional things; Firstly, we implemented Random Forest Algorithm for training the data again to check

```
Mean Absolute Error: 0.09166666666666666
Mean Squared Error: 0.09166666666666666
Root Mean Squared Error: 0.30276503540974914
[[61  4]
 [ 7 48]]
              precision    recall  f1-score   support

           0       0.90      0.94      0.92        65
           1       0.92      0.87      0.90        55

    accuracy                           0.91       120
   macro avg       0.91      0.91      0.91       120
weighted avg       0.91      0.91      0.91       120
```

performance and then optimize, it gives an accuracy of 92%. Secondly, we fed another set of data in the format of the trained data but new to the model, the model performed very well with accuracy. So, we proceeded to measure the performance of the model using confusion matrix, classification report, and accuracy score which gives good indication of optimal performance. We are in a dilemma either to ensure about 100% accuracy or over fitting because in supervise machine learning, high accuracy of almost 100% can be as a result of over fitting of the model which we want to avoid by possible means. So, since we are able to feed the model with new set of data which had not been previously fed to the model for which it performed very well with high rate of accuracy. So, we gave priority to avoid over fitting of the model than the accuracy of the model since high accuracy for supervised learning can be as a result of over fitting of the model for which the model becomes less accurate or behaved weird when fed with unfamiliar data. The validation report can be seen in Figure 5

We cannot test for the scalability, performance, and security of our model through cloud computing without deploying it. So, we proceeded to deploying our model to Microsoft Azure Cloud Server through a CICD (Continuous Integration/Continuous delivery) pipeline. We also write and deploy another restful API service with parameterize endpoint through which the API will be called to access the model from any device. Series of testing such as User Acceptance Testing(UAT), end to end integration testing,

and beta testing was also conducted with huge amount of successful.

IV. CONCLUSION

In this applied research, we are able to achieve four goals;

1. Build, implement and deployed a highly secure, scalable, and efficient machine learning and data mining model in cloud computing through a restful web API services.

2. Data privacy and security was maintain as the whole process occurs through a restful web API services.

3. Bridging the gap between industry and academic research in data mining is imperative. This we are also able to achieve and test in real time over various devices.

4. Used some of the latest tool and technologies in the industry such as GITLAB code versioning, CICD pipeline, and Docker containerization to deploy our model. This ensures that our model is easy to maintain and update over a very long.

We successfully solve the problem of scalability, efficiency, and security of data mining in cloud computing through a restful web API service in a highly secure cloud computing environment with accuracy of ninety-four (94) percent after optimization in our decision tree algorithm over web API. All that an AI engineer, data scientist, or machine learning engineer needs is just the endpoint of the API. This removes complexity while at the same time simplifying the whole process, it also add additional layer of security, and also remove unnecessary bottleneck as scientist and engineers will be able to concentrate more on optimizing their algorithm and model for optimal result since only API endpoint is what is needed to be called.

We hope that this will be de-facto standard in the data mining, machine learning, data science and other similar industry at large.

IV. REFERENCE

[1] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus. Knowledge discovery in databases: an overview, 1992.

[2] W.Wu and L. Gruenwald. Research issues in mining multiple data streams. In Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques, Stream KDD‟10, pages 56–60, New York, NY, USA, 2010.ACM.

[3] David E.Y. Sarna, Implementing And Developing Cloud Computing Applications, CRC Press https://cwiki.apache.org/MAHOUT/kmeans-clustering.html.

[4] Weiss, A. (2007). Computing in the Clouds Networker.

[5] Wang, K., Xu, C. & Liu, B. (1999), Clustering transactions using large items, in„CIKM ‟99: Proceedings of the Eighth International Conference on Information and Knowledge Management‟, ACM Press, New York, NY, USA, pp. 483–490.

[6]Berson, Alex, Stephen Smith, and Kurt Thearling. Building data mining applications for CRM. New York: McGraw-Hill, 2000.

[7] Moving To The Cloud: Developing Apps in the new world of cloud computing , By Dinkar Sitaram, Geetha Manjunath.

[8] The Cloud Computing Handbook Everything You Need to Know about Cloud Computing, By Todd Arias.

[9]http://searchsqlserver.techtarget.com/definition/dataminin g.

[10]http://www.ijcaonline.org/volume15/number7/px c387 2623.pdf.

[11]http://www.waset.org/journals/waset/v39/v3972.pdf

[12]http://www.estard.com/data_mining_marketing/data_mi ning_campaign.asp.

[13]http://dssresources.com/books/contents/berry97.htm

[14] http://www.marketingprofs.com/articles/2010/35 67/the-nine-most-common-dataminingtechniques-usedin-predictive-analytics.

[16]Data mining in cloud computing by Ruxandra-Ștefania PETRE, Link: https://www.dbjournal.ro/archive/9/9_7.pdf.

[17] Review of Data Mining Techniques in Cloud Computing Database, by Astha Pareek1, Manish Gupta2

[18] Ige, T., & Adewale, S. (2022a). Implementation of data mining on a secure cloud computing over a web API using supervised machine learning algorithm. International Journal of Advanced Computer Science and Applications, 13(5), 1–4. https://doi.org/10.14569/IJACSA.2022.0130501

[19] Amos Okomayin, Tosin Ige, Abosede Kolade , ” Data Mining in the Context of Legality, Privacy, and Ethics ” International Journal of Research and Scientific Innovation (IJRSI) vol.10 issue 7, pp.10-15 July 2023 https://doi.org/10.51244/IJRSI.2023.10702

[20] Okomayin, A., & Ige, T. (2023). Ambient Technology & Intelligence. arXiv preprint arXiv:2305.10726.

[21]Ogaga, D., & Zhao, H. (2023). The Rise of Artificial Intelligence and Machine Learning in HealthCare Industry. International Journal of Research and Innovation in Applied Science.

[22] Ogaga, D., & Zhao, H. The Rise of Artificial Intelligence and Machine Learning in HealthCare Industry.

[23] Oliver, S., & Anderson, B. (2023). Comprehensive Review on Advanced Adversarial Attack and Defense Strategies in Deep Neural Network.