ISSN No:-2456-2165

Crime Prediction Using K Mean and Customised Ml for Insurance Fraud

Darshna Parmar[1], Saurabh Rai[2], Yash Dwivedi[3], Rishabh Joshi[4], Shashank Chaudhary[5]
Assistant Prof[1]. , Dept. of Computer Science and Technology
Parul University

Abstract:- If an illegal act is performed to gain benefit or compensation by either the seller or buyer of an insurance contract is known as Insurance fraud. It may entail a fraudster filing a gasconade claim, post-dated policies, invalid medical history, viatical fraud, faked injuries or exaggerating their damages.

Keywords:- *Insurance Fraud, K Mean, Customized ML Approach, Hyper parameter Tuning.*

I. INTRODUCTION

The majority of insurance fraud cases involve exaggerated or false claims. Fraudsters try to make fraudulent claims by just showing up at the location and witnesses to get the amount which is counted as a section 420 Indian Penal Code.

We cannot ask human analysts to check every transaction one by one. We want accurate prediction, i.e., minimize missed frauds and false alarms. Recent report by Business Today paper shows a rapid growth in insurance frauds hence a proper system needs to be built for predicting fraudulent claims. 47% of companies became victims of financial fraud in the past two years. Recent surveys show that crimes in the dark web are increasing rapidly, hence a proper system needs to be designed in order to maintain the law.

II. OBJECTIVE AND SCOPE OF THE STUDY

The objective of our paper is to classify between a fraudulent claim and a genuine claim. As Many current models use traditional techniques that are traditional predictors. Less number of features used in classifying which leads to less accurate models. This model will be trained on 38 features with some of the most important features like “place_of_accident”, “type_of_vehicle”, “age_of_person”. Cyber crime departments can use this system to easily track the frauds with illicit records. Insurance companies can use this to detect and reduce their loss with less human intervention. By using standard methods of Data Ingestion, data pre-processing, customized machine learning, model selection, model tuning and lastly the deployment.

III. BACKGROUND KNOWLEDGE

We are going to classify fraudulent and genuine claims as following:

1. First we'll be taking file/batches of file with 37 features and 1 target column that is fraud or not from the client by some network share address, then validate it to check if it is in correct format if not then put it in bad file folder else pass it to be inserted in our training database that will be then exported as CSV file for training dataset.

2. Will do data pre-processing which involves handling missing values, categorical data to numerical, normalizing. After this next thing is to clusterize the data using K mean algorithm so that we can in later stage choose the best suited model for each and every cluster, this is also called customized machine learning approach.

3. Model selection for choosing best suited model for each cluster to get the best possible accuracy and tuning it by hyper parameter tuning (in specific GridSearch CV) to select best possible parameters for best suited models. Finally, we save the model as pickles

4. Deployment to cloud environments for making it available globally. After deployment this whole model will repeat the same process as 1-3 with the only difference that after data validation, the data will be stored to prediction DB from which csv file get imported and based on which cluster to which each data belongs model calls will be made to give the final prediction.

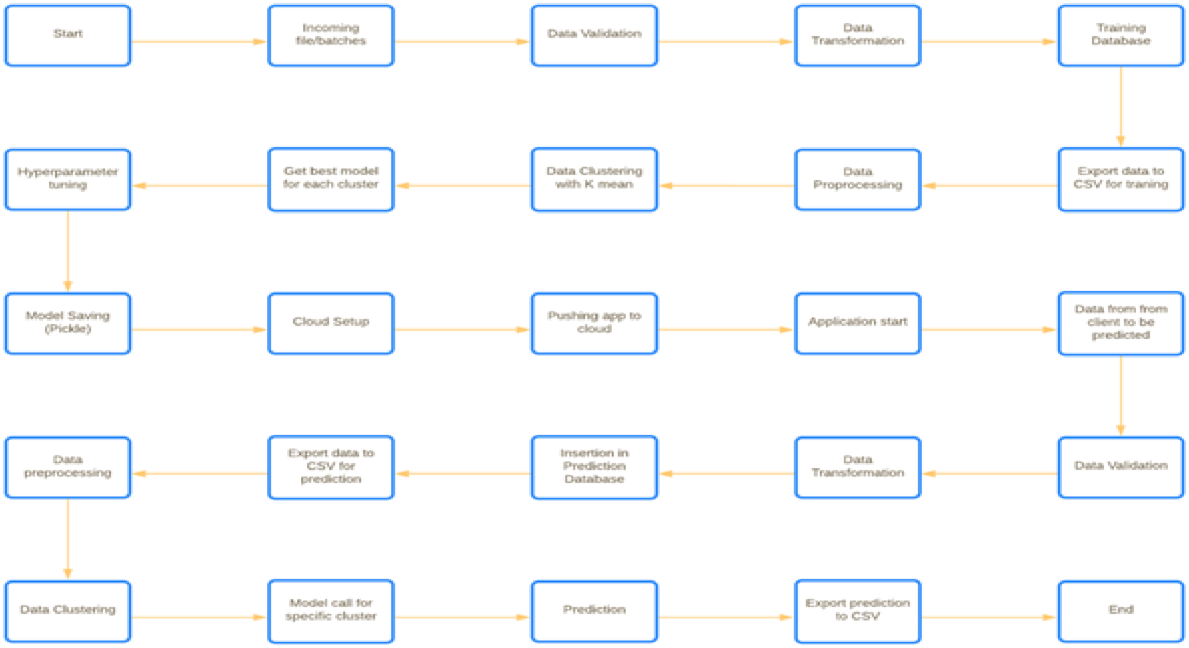

Fig. 1

IV. SURVEY OF LITERATURE

This section comprises some of the literature on various classification techniques that has been used to classify the various illicit content' s on the dark web, which was developed by various researchers and we have used them.

Insurance companies should rely on working with general data analytic techniques like methods having statistical touch, then classifying all the fraudulent claims by which business can be done. This paper also shows how having a few simple yet significant key variables can increase efficiency, reduce cost and time of the business. (For instance here just 20 variables are used instead of 31, reducing variables by 35%). As a result, we can give more time and focus on other pivotal variables.

The e-commerce transaction deceiving dataset is a database that has a class disparity. This study follows the synthetic minority over-sampling technique (SMOTE) method to deal with the class disparity in the e-commerce transaction deceiving dataset, the algorithm used is the decision tree, by naive bayes, neural network, and random forests. It was presumed that the application of decision trees, random forests and naive Bayes was able to manage the disparity of the e-commerce deceiving dataset by producing higher F-1 and G-mean results compared to the random forest, neural networks, and Naive Bayes.

In this survey of Machine learning techniques used in the Predicting Insurance Fraud has been inspected, which gives wide scope algorithms and methods for predicting fraudulent claims. The mix of supervised and unsupervised learning methods can be seen really often for achieving better accuracy. Many methods are blended together for getting better accuracy results. For instance, the Hybrid method, which demonstrates the blend of different algorithms resulting in flexibility and outperforming other algorithms as well. Another approach gaining prominence in recent times is ensemble learning, it fosters reliability and flexibility. Studies also show that ensemble learning tackles long-lasting machine learning problems like over-fitting, concept drift and class contrast. It's appealing because of its ability to generalize. Ensemble model and its application involves a good expenditure in resource and time which can be considered one-time investment, after all the aftermath is pretty much useful and efficient.

V. CONCLUSIONS

As the project is in the development phase, the only conclusion we can draw is that the given system will classify if a claim made by a person is genuine or fraud which will help the cyber crime agencies. In future we may have new unrecognised features that can be added in our model so as to keep our model updated with different fraudulent claims.

This model will be trained on 37 features and will be built using customized machine learning approach to first divide the valid data into cluster or groups by k mean then applying best suited model for each cluster, which will make this a highly accurate and stable model which will never overfit or underfit due to its hyper parameter tuning process.

VI. FUTURE WORK

In fraudulent claim detection higher the number of features available better the model accuracy. So, in future we should try to increase the number of features with the current trend of fraudsters.

REFERENCES

[1]. Stefan Furlan, Marko Bajec – Holistic approach to fraud management – University of Ljubljana, Journal of Informational and Organizational Science, Vol.32, 2018.

[2]. Carol Anne Hargreaves, Vidhyut Singhania- Analytics of Insurance Fraud Detection - American Journal of Mobile Systems, Applications and Services ,Vol.1, 2017.

[3]. Adi Saputra, Suharjito – Fruad Detection Using Machine Learning, IJACSA,Vol.10, 2019.

[4]. Nicola J Morley, Linden J Ball, Thomas C Ormerod- How Detection of Insurance Fraud Succeeds or Fails - Psycology, Crime & Law, Vol.12, 2016.

[5]. Clifton Phua, Vincent C.S Lee, Kate Smith-Miles, Ross W Gayler –A Comprehensive Survey of Data Mining- Monash University, Clayton campus, 2010.

[6]. Aggarwal, CC. 2015. Outlier analysis. In Data mining. Springer, pp. 237–263. DOI: https://doi.org/10.1007/978-3-319-14142-8_8

[7]. Ahuja, MS and Singh, L. 2017. Online fraud detection- a review. International Research Journal of Engineeringand Technology, 4(7): 2509–2515

[8]. Jun Lee, S and Siau, K. (2001), “A review of data mining techniques”,Industrial Management &Data Systems,URL: ttps://doi.org/10.1108/02635570110365989.

[9]. El Bachir Belhadji, Georges Dionne and Faouzi Tarkhani, “A Model for the Detection of Insurance Fraud ”, URL : https://www.researchgate.net/publication/233487 794.

[10]. Y.Sahin, S.Bulkan, E.Duman, A cost-sensitive decision tree approach for fraud detection, Elsevier, Expert Systems with Applications, Volume 40, Issue 15, p5916-5924 (2013).

[11]. D.Olszewski, Fraud detection using self-organizing map visualizing the user profiles, Elsevier, Knowledge-Based Systems, Volume 70, p324-333 (2014)

[12]. N.S.Halvaiee, M.K.Akbari, A novel model for credit card fraud detection using Artificial Immune Systems, Elsevier, Applied Soft Computing, Volume 24, p40-49 (2014)

[13]. Liu, W., Wang, S., Zhou, Y., Wang, L., Zhang, S., 2010. Analysis of forest potential fire environment based on gis and rs, in: The 18thInternational Conference on Geoinformatics: GIScience in Change, pp. 1–6. doi:10.1109/GEOINFORMATICS.2010.5567966.

[14]. Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. doi:10.1007/978-3-319-24574- 4_28.

[15]. Sahni, S., Kidwai, F., Sharma, P., Singhal, H., 2019. Implementation of iot to minimize post-harvest losses. Innovative Computing and Communication: An International Journal 1, 7 – 16.

[16]. Jörg Schiller, Universität Hamburg, Institut für Versicherungsbetriebslehre, Von-Melle-Park 5. 20146 Hamburg

[17]. Ke Nian Haofan Zhang, Aditya Tayal., Thomas Coleman B-2, Yuying Li.l. "Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3GI Combinatorics and Optimization.

[18]. European Journal of Economics, Finance and Administrative Sciences. ISSN 14502275 Issue 73 March,2015.©FRDN. http://wwweuropeanjournalofeconomicsfinanceandad ministrativescienc es.com

[19]. International Journal of Marketing, Financial Services & Management Research. Vol.2, No 5, May (2013) Online available at www.indianresearchjournals.com. ISSN 22773622.

[20]. Komal S. Patil, Prof.Anand Godbole, Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India.