International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Application of Machine Learning Techniques to Predict the Impact of Health Insurance on the Wellbeing of an Individual

Poornima Taranath, Sweta Das, Gowrishankar S.

Abstract*: The healthcare domain in India has suffered considerably despite the advancement in technology. Several financing schemes are endorsed by the insurance companies to lessen the financial burden faced by the government and people. Nonetheless, Health Insurance segment in India remains underdeveloped due to various complexities that it faces. This paper exploits a heuristic sampling approach combined with the ensemble Machine Learning algorithms on the large-scale insurance business data to realize the current shape of the Health Insurance industry in India. Through the courtesy of Data Mining and Data Analytics, it is plausible to furnish insights that assist the common people in acquiring closure that helps in the process of decision making.

***Keywords*: Data Analysis, Machine Learning, Web Scraping, Health Insurance, Sentimental Analysis.**

I. INTRODUCTION

The dawn of the Internet and tools like mobile phones, PC's, Tablets and other devices that utilize Internet-based services has caused massive production of data that may be useful or worthless. Additionally, a tremendous proportion of the world's population use and are connected by the Internet that fuels the generation of data. Often these data have all the qualities of Big Data namely volume, velocity, variety, and veracity [16]. Another flourishing discipline is the Internet of Things (IoT) that connects physical devices to the Internet, has a major role in redefining machine learning in our lives. Securing information and knowledge through a process of Data Analytics on these large and complex datasets reveals patterns and trends that aid in delivering insights to mankind. However, the rampant growth, enormous size and the substantive complexity of Big Data rule out the usage of traditional data management, data storage and data processing tools and techniques.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Poornima Taranath*, Department of Computer Science & Engineering, Dr.Ambedkar Institute of Technology, Bengaluru 560056, Karnataka, India.
Sweta Das, Department of Computer Science & Engineering, Dr.Ambedkar Institute of Technology, Bengaluru 560056, Karnataka, India
Gowrishankar S, Department of Computer Science & Engineering, Dr.Ambedkar Institute of Technology, Bengaluru 560056, Karnataka, India.

With the help of real-time computing across distributed platforms, solutions like Hadoop and NoSQL databases are used to overcome some challenges encountered by Big Data. Data Analytics, a subset of Business Intelligence, demands the ordering and organizing of raw data in order to obtain insights and information. Mathematical and scientific methods employed by Data Mining, a part of Data Analytics, is used to discover patterns and extract information from datasets using software available for open source data mining, proprietary data mining, and marketplace surveys. These fields regularly use Online Analytical Processing (OLAP) that aids in Customer Relationship Management (CRM). Moreover, these fields also embody Machine Learning (ML) that empowers systems to learn from experience. Using the concept of ML, predictions and decisions are made by systems without the need for programming them. Various algorithms for supervised, unsupervised and reinforcement learning has been adopted. Classification algorithms, Regression algorithms, Clustering, Topic Modeling are commonly used to train the machines. The aforesaid concepts can be exploited massively with the intention of improving the present state of the Healthcare industry in numerous ways. This shall also provide fresh insights that can tackle the future market's trends and requirements.

The rest of the paper is organized as follows. Section two presents literature survey, common issues in health insurance are addressed in section three, system overview is discussed in section four, results are provided in the fifth section and we finally conclude the paper.

II. LITERATURE SURVEY

Big data has been a compelling name in the last decade. With a rush of variety of forms of digital data, this abyss region was the goldmine that gave rise and made all the new technologies like Artificial intelligence and deep learning, to name a few, happen [17] which are applicable to business and the economic structure. The health industry has always been crucial for any economic growth. Alongside health, insurance industry too surged being its pillion. The insurance industry especially of the health sector has not benefited the potential uses of big data for their growth. In here we show sample platform where insurance business data can be exploited for the betterment of both customer and insurance providers. Using random forest classifier, machine learning techniques, big data on insurance from the China Life Insurance Company has been analyzed,

to showcase how insurance platforms can be improved. The drawback of using this was the imbalanced datasets which lead to different opines for the basis of classification; techniques like, the use of random forest algorithm in parallel computing and spark for cache [7, 8].

Methods like boosting [14], reverse random under sampling [15] have been considered to solve the problem of classification i.e., oversampling and undersampling but recent MapReduce technique using Hadoop has been efficient in producing results [8] where each decision tree of different training sets are trained parallelly. Spark took the favor in the coming times with its better memory structure RDD, for real-time query processing scenarios. The other side of the story is the risk and governance of big data since we are dealing with a sensitive matter at hand. A breakthrough in big data under the above technology's influence would be a potential foreseeable future [3]. These findings shall help end users and potential insurers, lawmakers to get a gist about strategically using big data in the healthcare insurance field.

III. ADDRESSING COMMON ISSUES

The present-day Health Insurance market in India has a broad scope and is in need of attention. Health insurance schemes are provided by Government agencies, private sectors, and other agencies to the common people with the aim of providing them with financial relief from the expenses incurred during medical contingencies. Yet, this industry deals with a myriad of issues. Regardless of India being a country suffering from overpopulation, only a fraction of India's population effectively utilizes the Insurance policies. Aided by the lack of knowledge and low literacy rates in rural areas, the usage of Health Insurance schemes are further minimized. Additionally, sales representatives of such policies have little understanding of the product they are trying to sell. Inadequate healthcare infrastructures and distribution channels further add up to the challenges faced by the Health Insurance market. Fraudulent claims and claim settlement disputes are very common [9,2]. The most eminent challenge faced is data acquisition because of the above mentioned issues encountered by the Health Insurance sector. Privacy and security mechanisms embraced by Insurance companies in order to protect the integrity of their customers makes collecting data unsuitable. Unreliable data sources which are further characterized by the poor quality of data is a source of inconvenience. Establishing consistency among datasets obtained from several sources and ensuring the correctness of data is a difficult task. With years this has been improving thanks to the behemoth of digitalization [5] where many factors like risk, analysis and the listed above are tackled for a stable environment of use.

IV. SYSTEM OVERVIEW

Data assembling and data analytics are the two major tasks that give shape to our entire system. Primarily, familiarity and awareness of the present insurance sector in India is of paramount importance. The understanding of the various challenges faced during data acquisition and other non-technical challenges enables us to plan and execute our actions in an orderly fashion. With this knowledge, we are able to resolve the complication of choosing the source for our data. Data is then scraped using various tools and techniques which is then proceeded with data analytics. Various insights and outcomes are drawn from the results of the analysis. These processes are depicted in Fig 1. in the form of a flow chart.

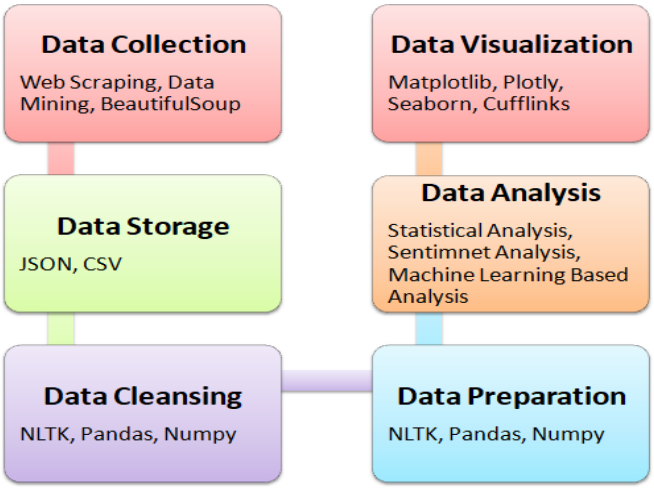

Fig. 1.Flow Chart representing the various stages of cleaning the data required for the analysis of Health Insurance plans.

Collecting data from valid and relevant sources is a demanding quest because of the characteristics big data possesses. Additionally, a multitude of challenges is faced by the insurance sector in India that is described in the above section. Even so, we succeeded in collecting data from the Web through a technique called Web Scraping. Python, a high-level programming language, is one of the most desirable languages that is used to carry out web scraping and data analytics. Data in the form of customer feedback about some of the most popular health insurance plans are scraped from websites like PolicyBazaar and MouthShut. Python's BeautifulSoup library aids in pulling out data from websites in XML and HTML format, through which preferred data is obtained and stored as a JSON or CSV file. The datasets stored includes customer opinions, ratings, and suggestions to the various insurance plans existing in India along with the customer information. Using the Jupyter Notebook that delivers an interactive execution environment, data gathered through web scraping is loaded into an IPython file using Python's Pandas DataFrame. DataFrame is a data structure provided by the Pandas library that empowers us to store tabulated data and modify them accordingly. Several statistical analysis is performed on the datasets and the same is displayed assisted by libraries like Numpy, Matplotlib, Seaborn, Cufflinks, Plotly, etc. Using a technique of Natural Language Processing (NLP) and the standard library NLTK provided by Python to implement NLP, we analyze and process the feedback of the customers. Furthermore, using NLTK tokenization, lemmatization, stemming, parsing and tagging is conducted on the textual data that is contained in the customer feedback column of the DataFrame.

A Word Cloud is then produced for the feedback column and it illustrates the words of utmost significance that are used by the customers to describe their experience on using the insurance plans. Words that portray pain points, grievances and satisfaction of the customers can be viewed using the Word Cloud. Using NLP, Sentiment Analysis is also performed on the textual data that examines the sentiments of the customers who provides the feedback. It assigns positive, negative and neutral scores to the opinions that help us in scanning their polarities. Methods and tools like Doc2Vec and TF-IDF are used for predicting similarities and for feature extraction. Machine Learning algorithms like Random Forest Classifier and Multinomial Naive Bayes are further used for sentiment analysis and predicting the features of utmost importance. A Receiver Operating Characteristic Curve and a Precision Recall Curve for the above Machine Learning algorithm is plotted to realize the accuracy of the models. Confusion Matrix is also generated for the same.

V. RESULTS AND DISCUSSION

Data Visualization is a technique where information and knowledge are presented to the common folks through charts, reports, pictures, graphs, etc. that entitles them to understand the various trends and patterns in an effort to draw out conclusions from them that is in one's best interests. This encourages individuals in making the right decisions that appear to be advantageous to their businesses and organizations or to different communities. Following the implementation of the concepts stated in the above section, we begin by visualizing the information obtained from these concepts. Libraries provided by Python like Matplotlib, Seaborn, Cufflinks, and Plotly are the most common libraries used by programmers to depict their observations with the help of various charts and plots. The results procured from carrying out analysis on the datasets are displayed in this section and are discussed further.

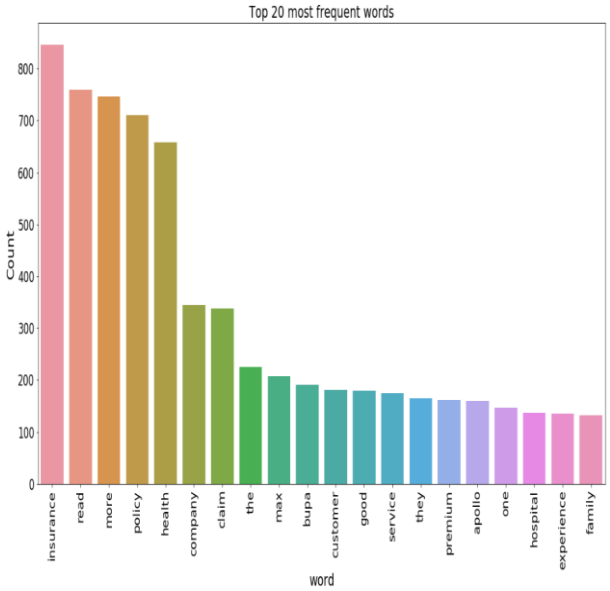

Fig. 2.Twenty most frequently used words to write feedback on an insurance plan.

The bar plot in Fig 2., shows the common English words used by a customer to present their viewpoint of an insurance plan. The customer opinions are condensed to prevent the appearance of words in the list called Stopwords which is provided by the NLTK library. This method assists us in excluding prepositions, conjunctions, and articles from the opinion.

Fig. 3.A Word Cloud for Health Insurance Sector.

The Word Cloud in Fig 3., illustrates the relative density of occurrence of the most used words to criticize or appreciate the insurance plan providers. Frequently observed words are bolder and they appear with a larger font than the others.

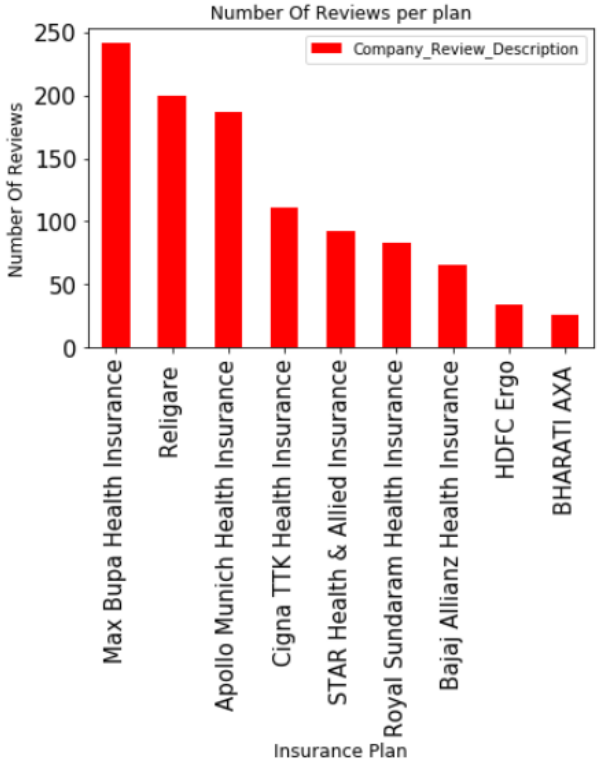

Fig. 4.Total number of reviews for each insurance plan in the dataset.

The bar plot as shown in Fig 4., identifies the most popular health insurance plans in India, based on the number of customer feedback for each insurance plan in the dataset.

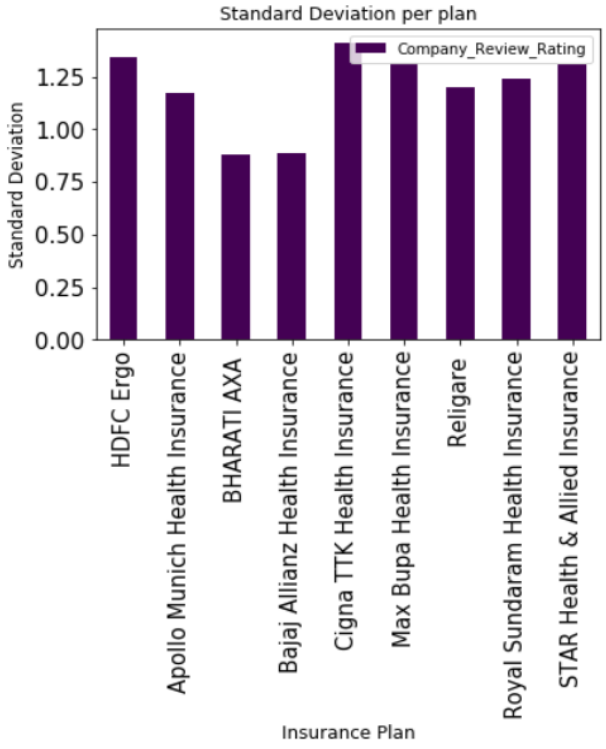

Fig. 5.Standard Deviation.

Fig 5., depicts the standard deviation among the ratings given to each insurance plan by customers thus identifying the variations in the ratings received by the health insurance plans.

Fig. 6.Relationship between text length and feedback rating.

A box plot in Fig 6., illustrates the correlation between the number of characters appearing in feedback and the rating given to the insurance plan by each customer.

In Fig 7., the heat map depicts the relation between the numeric fields in the dataset. It is observed that the ratings received by the company as a whole are related to the ratings given by each customer to the company by 79%. However, the text length is least related to the company rating as well as feedback ratings.

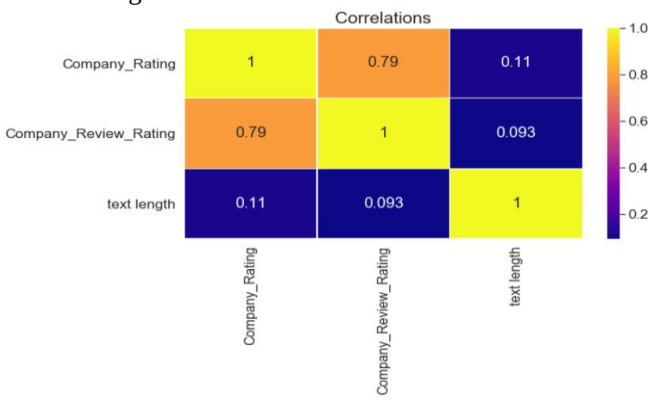

Fig. 7.Correlation Heatmap.

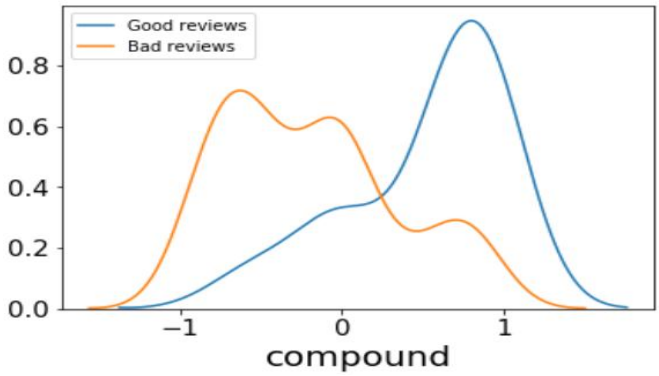

Fig. 8.Density Plot.

The Density Plot in Fig 8., portrays the densities of positive and negative reviews for the insurance sector in India. Using the polarity scores that deliberates the negative, positive and neutral response of customers, this density plot is achieved. In the following figure, the density plot for Max Bupa Health Insurance and Religare Health Insurance are compared.

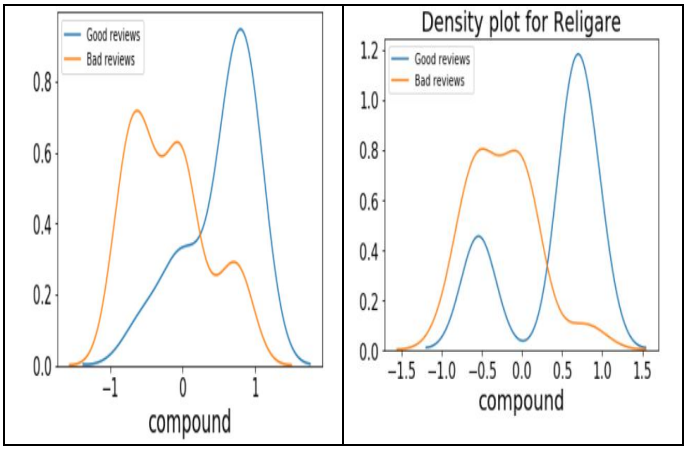

Fig. 9.Comparison between density plots.

Fig 9., outlines the difference between the density plots of the most reviewed health insurance plans. We can infer that Religare has more positive feedback in contrast to Max Bupa.

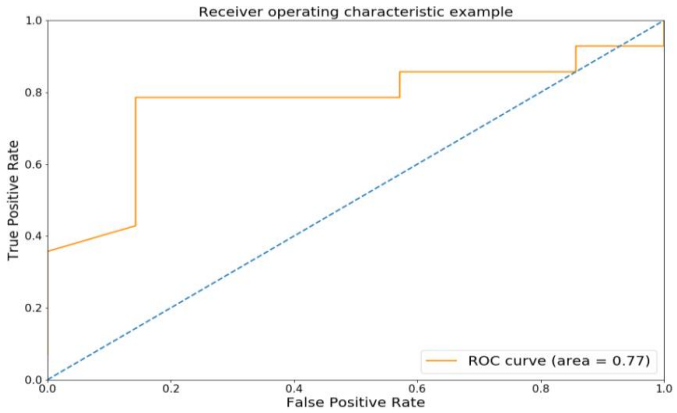

Fig. 10. A ROC Curve.

A Receiver Operating Characteristics Curve is plotted against the true positive and false positive rates predicted by Random Forest Classifier (Fig 10).

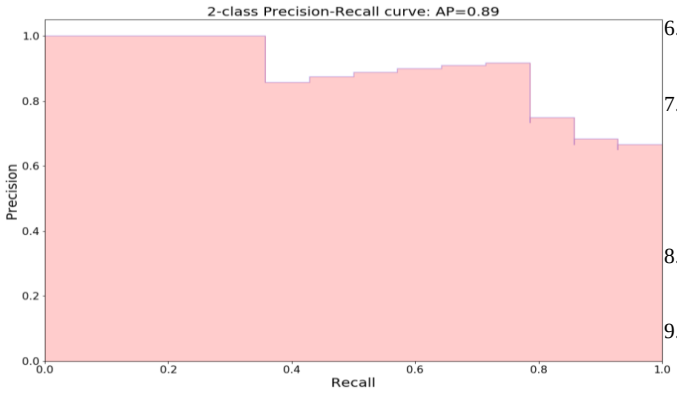

Fig. 11. PR Curve.

A Precision-Recall curve is plotted for the model implemented using Random Forest Classifier that portrays how the precision of the model varies with recall (Fig 11). It is observed that the higher the precision is, the lower is the recall and vice versa.

VI. CONCLUSION

Even with global connectivity access to most of the people, comparing and analyzing insurance claims would be a slight cumbersome task, with our concept of data analysis along with visualization, it gives a solution to a number of questions on India's health insurance market. Our work can be further developed by gathering information on every single insurance provider in the country, later in the world. With the growing financial and governmental policies, insurance's market is a promising field that would gather more big data in the future, joining this with Machine learning and other technologies we would be able to either create an application or a website with queries related to all the details and policies offered by the insurance providers. Many other indirect mechanisms can be developed on this research such as fraud prevention and cross-broker fraudulence in this sensitive domain.

In conclusion, having health insurance is vital for every citizen and our paper presents a way for a common man to understand and choose the right insurance plan that benefits him and others.

ACKNOWLEDGMENT

The third author would like to acknowledge that this research work was supported in part by the VGST grant of Govt. of Karnataka, India, under the RGS/F scheme.

REFERENCES

1. YiChuan Wang Raymond ,LeeAnn Kung, Chaochi Ting ,Terry Anthony Byrd Raymond, "Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care ",proceedings of 2015 48th Hawaii International Conference on System Sciences, Hawaii , January 05-08,2015.
2. S. Tennyson, "Insurance experience and consumers' attitudes toward insurance fraud," proceedings of Journal of Insurance Regulation, vol. 21, no. 2, pp. 35, 2002.
3. Wullianallur Raghupathi, Viju Raghupathi, "Big data analytics in healthcare: promise and potential" proceedings of Health Information Science and Systems, 2 (3) (, pp. 2– 10 ,2014.
4. M. Eling, H. Schmeiser, and J. T. Schmidt, "The solvency ii process: Overview and critical analysis," proceedings of Risk management and insurance review, vol. 10, no. 1, pp. 69–85, 2007.
5. Y. Shi, C. Sun, Q. Li, L. Cui, H. Yu, and C. Miao, "A fraud resilient medical insurance claim system." in proceedings of AAAI, pp. 4393–4394, Phoenix,Arizona , USA,February 12-17,2016
6. Uma Srinivasan , Bavani Arunasalam "Leveraging Big Data Analytics to Reduce Healthcare Costs" proceedings of computer.org/ITPro , 1520-9202/13 , IEEE ,2013.
7. Ziming Wu, Weiwei Lin, Zilong Zhang and Angzhan Wen,Longxin Lin "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis " in proceedings of IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) ,Guangzhou,China, July21-24, 2017.
8. Del Río S, López V, Benítez J M, et al." On the use of MapReduce for imbalanced big data using random forest", proceedings of Information Sciences, 285:112-137,2014.
9. S.-L. Wang, H.-T. Pai, M.-F. Wu, F. Wu, and C.-L. Li, "The evaluation of trustworthiness to identify health insurance fraud in dentistry," proceedings of Artificial Intelligence in Medicine, vol. 75, pp. 40–50, 2017.
10. Hido S, Kashima H, Takahashi Y" Roughly balanced bagging for imbalanced data" proceedings of Statistical Analysis and Data Mining vol 2(5,6), pp: 412-426, 2009.
11. "Oracle Health Insurance Analytics", in Oracle Health insurance back office applications: flexible solutions for complex healthcare systems.
12. Kuo M.H., Sahama T., Kushniruk A.W., Borycki E.M., Grunwell D "Health Big Data Analytics: Current Perspectives, Challenges and Potential Solutions.", Proceedings of Big Data Intelligence, 1(12): 114–126, 2014.
13. Garrison L.P. Jr "Universal Health Coverage-Big Thinking versus Big Data" proceedings of Value Health16(1 Suppl): S1-S3,2013.
14. Friedman J H, Hall P. "On bagging and nonlinear estimation" proceedings of Journal of statistical planning and inference, vol 137(3), pp: 669-683,2007.
15. Tahir M A, Kittler J, Yan F "Inverse random under sampling for class imbalance problem and its application to multi-label classification" proceedings of Pattern Recognition, vol 45(10), pp: 3738-3750,2012.
16. Kuo M.H., Sahama T., Kushniruk A.W., Borycki E.M., Grunwell D "Health Big Data Analytics: Current Perspectives, Challenges and Potential Solutions. Int J Big Data Intelligence "proceedings of vol 1(12) pp: 114–126,2014,
17. Chen H., Chiang H.L. Storey C,"Business intelligence and analytics: from Big Data to big impact" proceedings of MIS Quarterly ;vol 36(4) , pp:1-24,2012.

AUTHORS PROFILE

Poornima Taranath earned her B.E. in Computer Science & Engineering from Dr.Ambedkar Institute of Technology, Bengaluru, Karnataka in the year 2019. She is currently working as Software Developer in Metricstream, Bengaluru.

Sweta Das earned her B.E. in Computer Science & Engineering from Dr.Ambedkar Institute of Technology, Bengaluru, Karnataka in the year 2019. She is currently working as Software Developer in IBM-India, Bengaluru.

Gowrishankar S is currently working as Associate Professor in the Department of Computer Science and Engineering at Dr.Ambedkar Institute of Technology, Bengaluru, India. He earned his PhD in Engineering from Jadavpur University, Kolkata, India in 2010 and MTech in Software Engineering and BE in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belagavi, India, in the years 2005 and 2003, respectively. His current research interests are mainly focused on Data Science, including its technical aspects as well as its applications and implications. Specifically, he is interested in the applications of Machine Learning, Deep Learning and Big Data Analytics in Healthcare. He writes articles on his personal blog at https://www.gowrishankarnath.com. His Twitter handle is @g_s_nath.