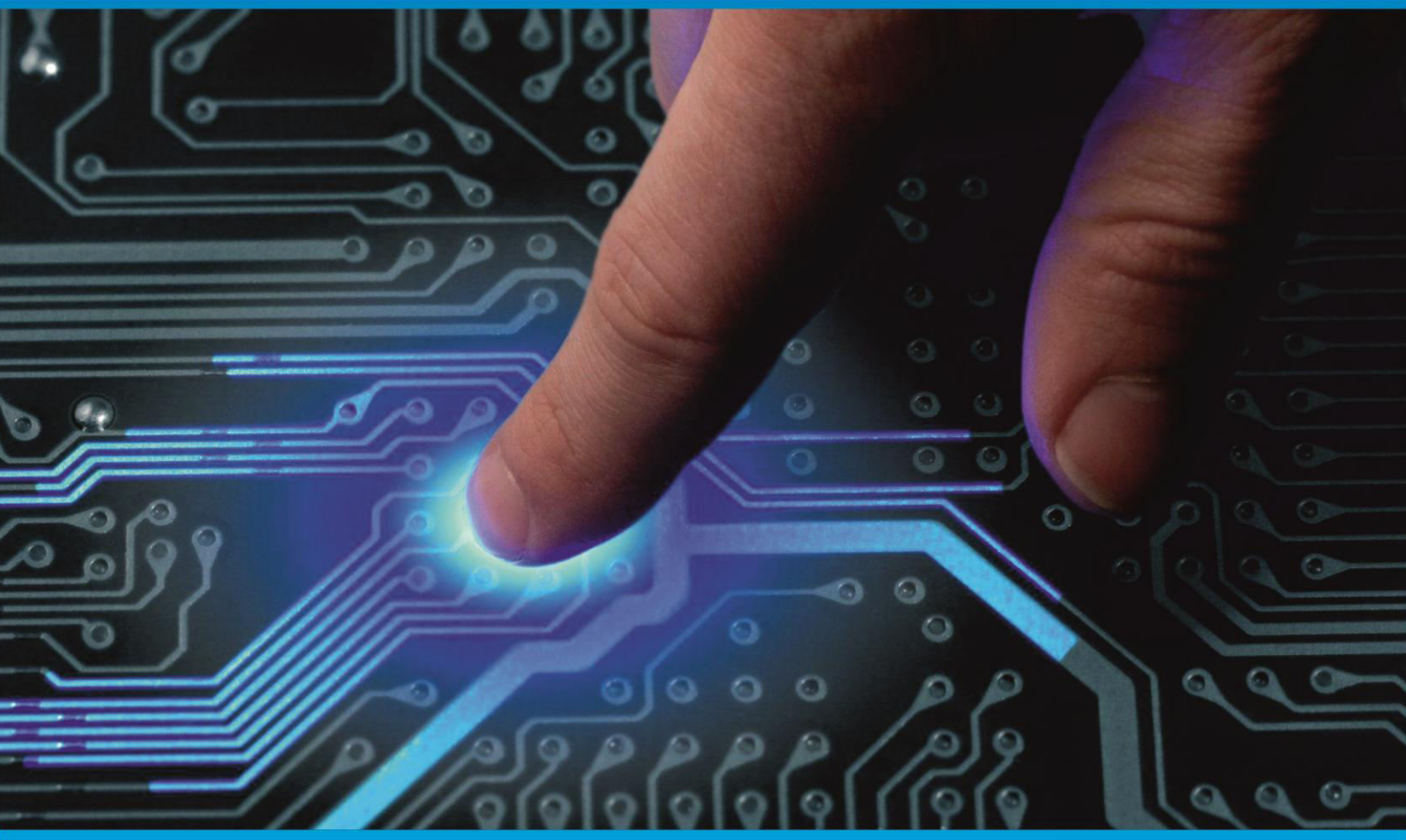
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

Medical Cost Price Prediction with Random Forest(RFA)

Latikarani Deepchand, Mausmi Devalekar, Nikita Vishwakarma, Prof. Asmita Deshmukh

Dept. of Computer Engineering, K.C College of Engineering and Management Studies and Research Thane, Kopri Thane(E), Mith Bandar Road,Thane, India

ABSTRACT: In the medical insurance system, their must be decision made about the treatments and what insurance should cover treatments one way to make this decision is to rank treatments by their ratios of sex, children, age, region and mainly smoker which affects the treatment costs..

KEYWORDS: Medical insurance costs, Random Forest Algorithm, Big data analytics, Machine learning, Kaggle.

I. INTRODUCTION

Big data technology was widely applied. Big data technology does not only create significant value but also promote the change and progress of traditional industries. The traditional marketing method of selling insurance is mainly based on off-line sales business. Insurance salesman sells the company's products by calling or visiting the customers. The blind marketing way has achieved good results in the past, which maintained the company sales performance for a long time through widespread sales with the gradual opening of the insurance industry, a large number of private insurance companies enter the market, which forms a healthy competitive environment and constantly promote the reform of the insurance industry. Due to the lack of purpose and innovation of traditional marketing methods, the poorly organized insurance business data and obscure customers' purchasing characteristics directly lead to a serious imbalance in the category of product data, which bring difficulties to user classification and recommendation of insurance product.

Classification of imbalanced data sets has puzzled many researchers. In real life, we could not get the expected distribution of data because of various reasons, especially in some cost sensitive business scenarios. For the unbalanced distribution of data in the same sample space, we usually choose some re-sampling methods which sacrifice some features to construct relatively balanced training data sets. Information has been the key to a better organization and new developments. Te more information we have, the more optimally we can organize ourselves to deliver the best outcomes. That is why data collection is an important part for every organization. We can also use this data for the prediction of current trends of certain parameters and future events. As we are becoming more and more aware of this, we have started producing and collecting more data about almost everything by introducing technological developments in this direction. Today, we are facing a situation wherein we are flooded with tons of data from every aspect of our life such as social activities, science, work, health, etc. In a way, we can compare the present situation to a data deluge. Te technological advances have helped us in generating more and more data, even to a level where it has become unmanageable with currently available technologies. This has led to the creation of the term 'big data' to describe data that is large and unmanageable. In order to meet our present and future social needs, we need to develop new strategies to organize this data and derive meaningful information. One such special social need is healthcare. Like every other industry, healthcare organizations are producing data at a tremendous rate that presents many advantages and challenges at the same time. In this review, we discuss about the basics of big data including its management, analysis and future prospects especially in healthcare sector.

II. LITERATURE REVIEW

Machine learning is a technology where machines can learn from the previous data and predict new samples. Machine learning models are applicable in all fields. Medical files also not having any exclusion to machine learning. Medical field using ML models in different situation from last several year. Many of the researchers applied machine learning techniques to medical related cost prediction. B.Nithya [1] et.al applied machine learning models in predictive Analytics in health care. They applied various supervised and unsupervised models for predictive analysis. They also

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

suggested machine learning tools and technique are decisive in health care province and exclusively used in the diagnosis and prediction of various types of diseases.

III. MACHINE LEARNING APPROACHES

Price prediction history shows that authors have used machine learning techniques in this domain extensively. Health domain is no exception for this where medical prices are being predicted using health related data. Broadly machine learning techniques are categorized into two types of learnings, supervised and unsupervised learning.

A. *Supervised Learning*

Supervised learning is more widely used among the mentioned types. It is called supervised learning because in this approach, the algorithm is trained on the input data to get the desired result. Another way to define it is an algorithm is trained under a supervision of training data. In this case, the data used to train the algorithm is a labeled data. When the target or the result is known, the data is called labeled data. In a typical supervised machine learning problem, the data has two parts. The first part consists of input variables which are called features. The second part is actual target variable or label. Features help to find out the target. The mapping of features and the label would look like,

$$Y= f(X)$$

Where Y is the label and X is the input variable. There are two phases in supervised learning. The first is the training phase and the second is the testing phase. In the training phase, data consists of input variables and a label associated with those. When the algorithm is sufficiently trained on this data, new data is given for testing. This new data is not labeled. Therefore, based on the training algorithm has gained in the training phase; it has to predict the label for new set of data. Examples of supervised learning are 8 classification and regression techniques. These techniques train your model/machine with the chosen dataset and when the new input comes classifies or predicts the output.

IV. ARCHITECTURE

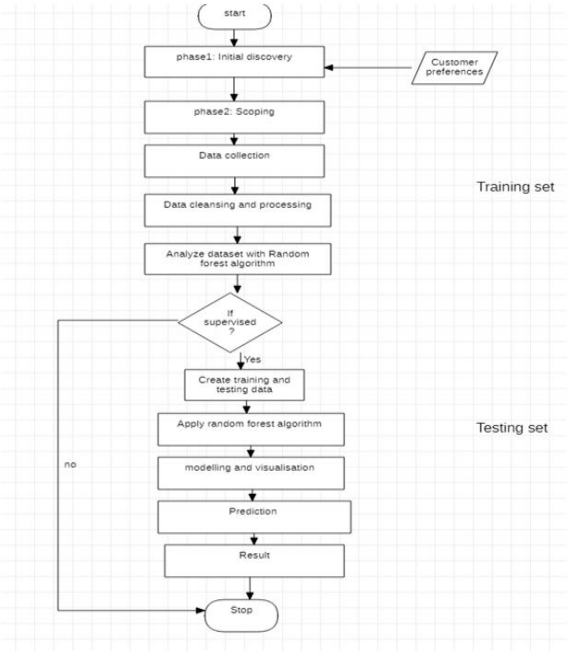


Fig. 1. This shows figure of the architecture of the system.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

V. DESCRIPTION OF ARCHITECTURE

The initial discovery is the first face to face meeting between the consultancy team client and will take place after background information has been exchanged by telephone or email . The information obtained during the initial discovery, additional information gathered through research will be analysed and evaluated against stated challenges and objectives . Data will be collected securely in accordance with an agreed methodology . Once collected the data needs to be 'cleaned' to prepare it for processing. Based on patterns and features, models will be created to answer questions set during the scoping phase.

After creating the answers questions set during the scoping phase the results will be presented to the client in a way that answers the challenges set for the project allowing the client to implement the findings.

A. *RFA(Random Forest Algorithm)*

Random forests or decision forests are an ensembled learning method for classification,regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.The random forest is made up of several decisions trees, each decision will be full growth,it do not need to cut processing,the more tree it has the more accurate the result will be,the overall estimate,and it has the advantage of automatic feature selection etc.We have used RFA it is supervised learning algorithm .The "forest" it builds, is an ensemble of decision trees,usually trained with the "bagging" method . Random forest builds multiple decision trees and merges them together to get a more accurate and stabble predicition.Random forest is a collection of decision trees,there are some differences.If you put input a training dataset with features and labels into a decision tree,it will formulate some set of rules ,which will be used to ,make the predictions.The random forest algorithm is used in a lot of different fields,like banking,the stock market ,e-commerce.It overcomes the problem of over fitting by combining the result of different decision trees.It works well on large range of data items than a single decision tree does.It has less variance than single decision tree.Random forest are very flexible and possess very high accuracy.

- **ADVANTAGES OF RANDOM FOREST**

- Effectively capture the non-linearity between the response and explanatory variables as compared to multiple linear regression models.
- They can effectively handle many types of predictors (categorical, continuous, skewed, etc.) without the need to pre-process them.
- It does not require the user to explicitly specify the form of the predictor's relationship to the response as opposed to multiple linear regression

- **DISADVANTAGES OF RANDOM FOREST**

- Model instability – minor changes in data can significantly alter the tree's structure resulting in inaccurate interpretations.
- They become highly computational as the number of trees increases.
- They can sometimes lead to over fitting.

B. *Classification*

The target variable or the label in a classification problem holds categorical values. That means the target variable is discrete. The categorical values of this target variable are usually finite. The mapping function given in the previous section then would try to predict one of the categorical values, the target variable has with the help of input variables. For example, given a set of input variables an algorithm would try to predict whether a person will buy a house or not. In this example, Buying a house is a label and it has two values Yes or No. When the label has only two values such as either "Yes" or "No", or "1" or "0" then the problem is called as a binary classification problem. The label can have multiple class values as well then it is called as multi-class classification problem. For example, given a set of input variables an algorithm would try to predict whether a given fruit is "Mango" or "Apple" or "Banana". There are many algorithms used to solve classification problems. Some of the algorithms are Decision Trees, Random Forest.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

C. *Decision Tree*

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any boolean function on discrete attributes using the decision tree. Random forest or Random decision forest is a method that operates by conducting multiple decision trees. The Decision of the majority of the trees is chosen by the random forest as the final decision. Decision tree is a tree shaped diagram used to determine a course of actions. Each branch of the three represents a possible decision.

VI. METHODOLOGY

A. *Phase One: Initial Discovery*- The initial discovery is the first face to face meeting between the consultancy team and the client and willtake place after the background information has been exchanged by telephone and email. The consultants will will looking to gather additional information for the project scoping phase. The client will have the opportunity to speak in depth about the business and what they are hoping to achieve. The client will also be provided with the information about how data analytics pojects work, how the Data Science manages these projects and the commitments the client is expected to make.

B. *Phase Two: Scoping*- The information obtained during the initial discovery, additional information requested from the client and supplementary information gathered through research will be analysed and evaluated against stated challenges and
.

C. Data Collection- Data will be collected securely in accordance with an agreed methodology. This process varies from client to client and is dependant on the type of quatity of data available and how it is stored.

D. Data Cleaning and Processing- Once collected the data need to 'cleaned' to prepare it for processing. This involves identifying gaps in the data, making data compatible and fixing errors in storage systems.

E. *Modelling*- Based on patterns and features, models will be created to answer questions set during the scoping phase.

F. *Visualisation*- The results will be presented to the will be created to answer questions set during the scoping phase.

G. Prediction- Machine learning AI models will be trained and evaluated using historical data. These will the be applied to fresh data.

VII. IMPLEMENTAION

A. *Learn About Your Actual And New Customers:*
Currently, many companies have much information about their customers: who they are, where they live, what their behaviour is online, how is their shopping history, what are the propducts that season after season after season buy (or not). The next step is using that data into information, information that can be used to understand the different profiles, preferences are, and also, the possible actions and associated outcomes. Also, it is entirely possible to detect new customers and new oppourtinities by combining different data sources.

B. Forecasting of Future Scenarios:
The last level of any implementation of data science project into a company is forecasting: once that your data is reliable and consolidated it is possible to analye future patterns and predict behaviors.

C. Improving Business Decision- Making Process:
Usually, the process of business decision making relies on the expertise of management and data analysis of the business. Nevertheless, this process can change and improve through the simulation of a variety of potential scenarios using the in-depth knowledge of customers and the forecasting of the next tendencies to conduct us to the best business outcomes.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

VIII. TOOLKIT:- JUPYTER

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

IX. CONCLUSION

Based on the findings of the project, it can be said that predictive modeling has tremendous benefits for the health insurance industry in determining how much the premium should be charged to the insured based upon his/her behaviors and health habits. Health insurance companies can then accurately charge the premium based upon a specific individual's attributes. This will not only help the individuals in getting charged the right amount of premium for their health insurance but will also help in forgiving better relationships and a level of trust between the insurance company and the insured. However, there are certain limitations which are the scope of further studies. The data did not include any information on an individual's medical costs, the real-time data i.e. data collected from the sensors in the wearable health devices such as fit bits etcetera. If we take all these types of different data sources into account then we can have a better picture of an individual's behavior and can more accurately predict the insurance premium charge and the associated risk.

REFERENCES

1. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 2002: 321-357.
2. Ramentol E, Caballero Y, Bello R, et al. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory [J]. Knowledge and information systems, 2012, 33(2): 245-265.
3. Sáez J A, Luengo J, Stefanowski J, et al. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291: 184-203.
4. Ramentol E, Verbiest N, Bello R, et al. SMOTE-FRST: a new resampling method using fuzzy rough set theory[C]//10th International FLINS conference on uncertainty modelling in knowledge engineering and decision making (to appear). 2012.
5. GU Ping, OU YANG Yuan-you. Classification research for unbalanced data based on mixed-sampling[J]. Application Research of Computers, 2015, 32(2): 379-381.
6. Tomek I. Two modifications of CNN[J]. IEEE Trans. Syst. Man Cybern.1976, 6: 769-772.
7. Tahir M A, Kittler J, Yan F. Inverse random under sampling for class imbalance problem and its application to multi-label classification[J]. Pattern Recognition, 2012, 45(10): 3738-3750.
8. Angiulli F. Fast condensed nearest neighbor rule[C]//Proceedings of the 22nd international conference on Machine learning. ACM, 2005: 25-32.
9. Laurikkala J. Improving identification of difficult small classes by balancing class distribution[M]. Springer Berlin Heidelberg, 2001.
10. Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[M]//Advances in intelligent computing. Springer Berlin Heidelberg, 2005: 878-887.
11. Friedman J H, Hall P. On bagging and nonlinear estimation [J]. Journal of statistical planning and inference, 2007, 137(3): 669-683.
12. Hido S, Kashima H, Takahashi Y. Roughly balanced bagging for imbalanced data [J]. Statistical Analysis and Data Mining, 2009, 2(5-6): 412-426.
13. Del Río S, López V, Benítez J M, et al. On the use of MapReduce for imbalanced big data using random forest [J]. Information Sciences, 2014, 285: 112-137.
14. Bhagat R C, Patil S S. Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest[C]//Advance Computing Conference (IACC), 2015 IEEE International. IEEE, 2015: 403-408.
15. Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

16. Goldston D. Big data: Data wrangling [J]. Nature, 2008, 455(15).
17. Reichman O J, Jones M B, Schildhauer M P. Challenges and opportunities of open data in ecology [J]. Science, 2011, 331(6018).
18. TU Xin-li, LIU Bo, LIN Wei-wei. Survey of big data[J]. Application Research of Computers, 2014, 31(6): 161.
19. Beibei Li, Bo Liu, Weiwei Lin, and Ying Zhang. Performance Analysis of Clustering Algorithm under Two kinds of Big Data Architecture. Journal of High Speed Networks, 2017, 23(1): 49-57.
20. Ping Li, Jin Li, Zhengan Huang, Tong Li, Chong-ZhiGao, Siu-Ming Yiu, Kai Chen. Multi-key privacy-preserving deep learning in cloud computing. Future Generation Computer Systems, 2017. DOI: 10.1016/j.future.2017.02.006.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904153|

The Jupyter Notebook is an open source web

application that you can use to create and

share documents that contain live code, equations, visualizations, and text. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details