IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol. 8, Issue.12, December 2020 E-ISSN: 2347-2693

Insurance Approval Analysis using Machine Learning Algorithms

CH. Lakshman Vinay[1], G. Vijay Sagar[2], M. Ajay[3], SK. Hussain[4], Bh Padma[5*]

[1,2,3,4,5]Dept. of Computer Science and Engineering, Gayatri Vidya Parishad College for Degree and PG Courses, Visakhapatnam, India

[*]*Corresponding Author: padma.bhogaraju@gmail.com, Tel.: 0891-2718053*

DOI: https://doi.org/10.26438/ijcse/v8i12.4650 | Available online at: www.ijcseonline.org

Received: 11/Oct/2020, Accepted: 10/Dec/2020, Published: 31/Dec/2020

Abstract— Risk Management is important for insurance industry to ensure the eligibility of a new customer for approval. Insurance companies need to analyze the existing customer's information such as income, assets, occupation, premium payment records to decide whether a new customer is qualified for an insurance policy. This paper focuses on forecasting the eligibility of the new customers for insurance approval by performing classification on a real time insurance company dataset using three Machine Learning algorithms such as Decision Tree Induction, Naive Bayes Classification and K-Nearest Neighbor algorithms. These algorithms are examined against their classifier accuracy after implementation and the algorithm that demonstrates the best accuracy is chosen for predicting the new customers.

Keywords—Insurance, Machine Learning, Decision Tree Induction, Naive Bayes Classification and K-Nearest Neighbor, Classifier

I. INTRODUCTION

Insurance is a contract, signified by a policy, in which an entity receives financial security or reimbursement against financial losses from an insurance company. The company collects clients' risks to make payments further affordable for the insured. Insurance policies are employed to evade against the risk of financial losses, even either big or small, that may occur from damage to the insured or her property or from accountability for damage or injury caused to a third party. There are a large number of different types of insurance policies that are available, and virtually any business or individual can come across an insurance company willing to insure them for a certain price. The most widespread types of personal insurance policies are health, auto, Home, life and owners. Most individuals in the United States have at least one of these types of insurance, and car insurance. Over the past years, insurance industry poses many challenges; one of them is maintaining the data either in legacy systems or in paper files for underwriting transactions.

Most of the insurance companies are automating their data collection. Complexity of underwriting process has been increased with the increase of data [1]. Therefore, this paper performs an analysis on a real time insurance company dataset for making a decision that a new member can be promoted for the insurance approval or not [2]. An ML based decision support system is taken into the consideration which can act as underwriter and it is capable enough to take decisions based on the trained data. This paper uses a 'decision-tree' classification model, Bayesian Classification model and KNN model by training the insurance company dataset for predicting new customers.

The new customers are asked to enter the required insurance details which are used to predict whether the new customer is eligible for insurance or not. Admin can check the records of new and old customers from the dataset. Based on the input dataset dynamically a model is set by calculating accuracy of each every model.

II. PROBLEM STATEMENT

Over the past years, insurance industry poses many challenges one of them is maintaining the data either in legacy systems or in paper files for underwriting transactions. Most of the insurance companies are automating their data collection. Complexity of underwriting process has been increased with the increase of data. So we need a system, which analyses whether the new customer is eligible for an insurance or not using machine learning algorithms for an insurance provider and also for the user to enquire of his insurance approval [3].

This research therefore performs an analysis on a real time insurance company dataset for making a decision that a new member can be promoted for the insurance approval or not [4]. An ML based decision support system is taken into the consideration which can act as underwriter and it is capable enough to take decisions based on the trained data. This project derives a 'decision-tree' classification model and Bayesian Classification model by training the insurance company dataset for predicting new customers. The main objective of the research is to build a reliable system that is feasible to use by partially and completely by people.

III. LITERATURE SURVEY

Aman Dubey, TejismanParida, AkshayBirajdar, Ajay Kumar Prajapati, published a research paper named after "Smart Underwriting System: An Intelligent Decision Support System for Insurance Approval & Risk Assessment" which describes its abstract as "insurance industry poses many challenges, one of them is maintaining the data either in legacy systems or in paper files for underwriting transaction. Most of the insurance companies are automating their data collection process.

Traditionally, information of the client (such as personal details, medical records etc.) who needs insurance is sent to the underwriter through an email and after proper analysis, underwriter sends the quick Quote back to the agent based on his intuition and experience. Generally, quick Quote consists of insurance approval conditions and insurance plan name [5]. Due to enormous amount of diseases and medicines, complexity in underwriting process has been increased. In a nutshell, an improved and optimized way of underwriting process is required. Introducing Artificial Intelligence can help to transform the traditional underwriting process to smart one. Usually data given to the underwriter is in unstructured format.

Using Natural Language Processing and by training numerous statistical machine learning classifiers over the unstructured texts, important features were extracted out from unstructured emails. Main challenge is to exploit the information embedded in emails using automated tools, because of noisiness, un-cleaned and unstructured data. Based on the features extracted, a model was trained and tested for unseen mails to get the proper insurance plan name and advice. This data was drafted to a template and sent back to the agent through an automated email reply. Main Objective of the paper is to handle dynamic situations efficiently and to automate the underwriting task.".

Narander Kumar, Vishal Verma, Vipin Saxena, published a research paper named after "Construction of Decision Tree for Insurance Policy System through Entropy and GINI Index "which describes its abstract as "In the modern era of computing, sparse and irregularity in a data sample is needed for rebuilding of the huge and different types of dimensions of data. But still, there is a challenge to make an analysis on this type of data. One issue is the vigorous selection of data.

There are a number of analytical tools and weapons are available but vital part is the decomposition, and deciding the clusters as well as gradient estimation of data. When wee extract the attributes of sample data then most common attributes may guide to inaccurate results. So, the present paper offers solutions through entropy calculation and GINI Index computation of an insurance company. After calculation of entropy and GINI Index, a decision tree of sample data of an insurance company is presented.

There are many different types of classification tasks that and specialized approaches to modeling that may be used for predicting the class labels.

III.I Decision Tree Algorithm:

Decision Tree algorithm belongs to the family of supervised learning algorithms. As compared to other supervised learning algorithms, the decision tree technique can be utilized for solving classification and regression and problems too [6]. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We evaluate the values of the root attribute in comparison with the record's attribute. Depending on this comparison, we pick up the branch corresponding to that value and then jump to the subsequent node. This method comes under supervised learning algorithm (though it can be used for unsupervised learning).It means we have to provide the dataset with labels already there which is the correct choice. Then we train the Decision tree model with it. In the end, we predict future input with unlabeled data. Steps involved are: Label the dataset, generate a vocabulary and Create document term matrix. For Example: Our goal is to predict if the constraints given by user are meeting the requirements or not and finally predicting the output.

III.II Naïve Bayes:

This classification technique of machine learning uses Bayes theorem with assumption that there is independence among predictors [7]. In nutshell, it is assumed that in a class, feature is independent to presence of any other feature. It is simpler to build and useful for big data set with random features. Its performance is also fast. It calculates posterior probability P(c|x) from P(c). P(c|x) is the posterior probability of class, given predictor (x, attributes). P(c) is the prior probability of class. P(x|c) is the likelihood i.e. probability of predictor given class. P(x) is the prior probability of predictor.

Steps involved in Naive Bayes:

- Frequency table is converted to data set
- Find probabilities such as overcast probability and create likelihood table
- Calculate probability for each category by using Bayesian equation.

III.III KNN (K- Nearest Neighbor):

K-nearest algorithm is an efficient algorithm if we wanted to see the characteristics of an object matching with the other-object nearby it that means if you know about the characteristic of one object then you can easily determine the characteristic of the object near it. In the research, basically the credibility of the client/ applicant was seen and how much asset can it provides if his insurance is approved. Hence to evaluate this data mining techniques

were used here; the given data is pre-processed so that all metadata can be extracted will be used to find the closest neighbour of the given data with all the training data. The proposed classifier used to recognize the new object follows these steps: Testing data, Data Abstraction, KNN Classifier, Insurance plan evaluation, and finally output the result.

IV. SYSTEM DESIGN

The aim of this research is to predict whether a new customer is eligible for insurance or not. So user will interact through user interface and enters the required details which are passed to the prediction system and this will generates the output. Whereas admin will interact with admin interface and from there he can manage the dataset and also checks accuracy for each and every model. Dataset is the input for the model used for training and testing the records.

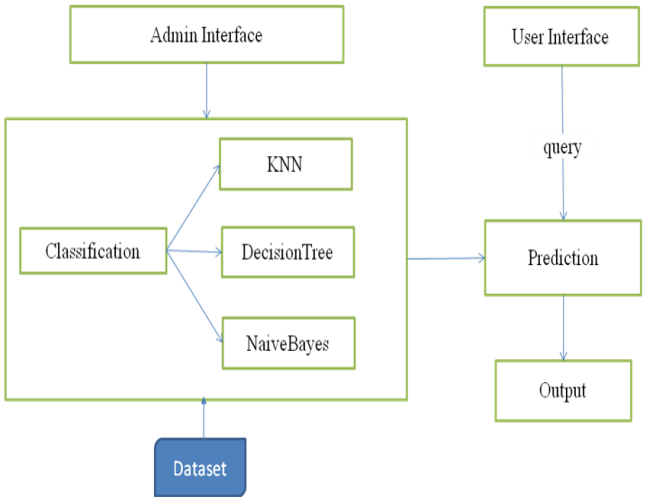

Fig.1 Proposed System Design

IV.I System functions:

- Insurance Policy Holder's dataset is input to the Machine Learning techniques.
- Perform feature selection for selecting the attributes which are relevant and contribute to the machine learning process.
- Derive a best classification model by training the Policy holder's dataset.
- These models are used to predict whether an individual is approved for an insurance policy or not.
- A comparative study with respect to 'model accuracy' is also performed.

V. IMPELEMENTATION

This system is implemented using Django Web Framework and Python 3. Python has a design philosophy that gives emphasis to readability of code, particularly using significant whitespace. It presents programming constructs that offers clear coding on both small and large scales. Python possesses a dynamic type system and automatic memory management. It maintains multiple programming paradigms, together with imperative, object-oriented, procedural and functional. It also has a complete standard library. Django is a high-level Python web framework that facilitates rapid development of maintainable and secure websites. Built by skilled developers, Django takes care of much of the problems of web development, so you can spotlight on writing your app without needing to spin the wheel. It is free and open source, and has an active and thriving community, great documentation, and other options for free and paid-for support.

V.I Dataset:

This paper uses a 'decision-tree' classification model, Bayesian Classification model and KNN model by training the insurance company dataset for predicting new customers. The new customers are asked to enter the required insurance details which are used to predict whether the new customer is eligible for insurance or not.

	sno	Age	Gender	Married	Bmi	Children	Smoker	Region	Income	Education	InsuranceApproval
0	1	19	0	0	27.900	0	1	3	0	1	1
1	2	18	1	1	33.770	1	0	2	0	1	1
2	3	28	1	1	33.000	3	0	2	1	1	0
3	4	33	1	1	22.705	0	0	1	1	0	0
4	5	32	1	0	28.880	0	0	1	0	1	1
5	6	31	0	1	25.740	0	0	2	0	1	0
6	7	46	0	1	33.440	1	0	2	0	0	1
7	8	37	0	1	27.740	3	0	1	1	1	0
8	9	37	1	1	29.830	2	0	0	0	1	0
9	10	60	0	1	25.840	0	0	1	0	1	0
10	11	25	1	1	26.220	0	0	0	1	1	1
11	12	23	1	1	34.400	0	0	3	0	1	1
12	13	56	0	0	39.820	0	0	2	0	1	1
13	14	27	1	1	42.130	0	1	2	1	1	1
14	15	19	1	0	24.600	1	0	3	1	1	0
15	16	52	0	0	30.780	1	0	0	0	0	1
16	17	23	1	0	23.845	0	0	0	0	1	0
17	18	56	1	1	40.300	0	0	3	0	0	1
18	19	30	1	1	35.300	0	1	3	1	1	1
19	20	60	0	1	36.005	0	0	0	1	0	1
20	21	30	0	1	32.400	1	0	3	0	1	1
21	22	18	1	1	34.100	0	0	2	0	0	1
22	23	34	0	1	31.920	1	1	0	0	0	1
23	24	37	1	1	28.025	2	0	1	0	1	0
24	25	59	0	1	27.720	3	0	2	1	1	1
25	26	63	0	1	23.085	0	0	0	0	1	0

Fig 2. Insurance Dataset

VI. OBSERVATIONS AND RESULTS

A comparative analysis is done among the three algorithms with respect to the accuracy of the algorithm when tested for prediction. Decision tree and KNN both are non-parametric methods. Decision tree supports automatic feature interaction, whereas KNN can't. Decision tree is faster due to KNN's expensive real time execution. A much better method for evaluating the performance of a classifier is to consider the confusion matrix. A confusion matrix is a way of summarizing the performance of a classification technique. The confusion matrix shows the way in which your classification model is confused when it makes predictions. It offers us an insight not only into the errors being made by the classifier but also several types of errors that are being made. Each row in the confusion matrix

stand for an actual class, where each column is corresponding to a predicted class label in the classification process [8][9].

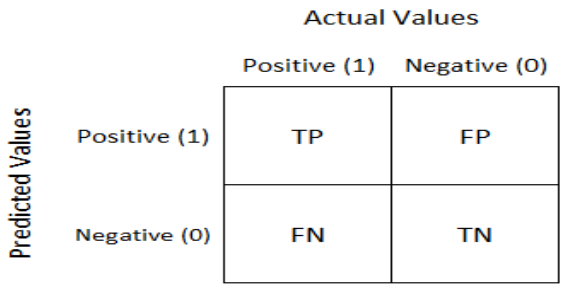

Fig 3.Confusion Matrix

Following table 1. shows confusion matrix after the implementation of different classifiers using Django Web Framework tool and Python implementation. The results are taken on a computer with 4GB RAM with Win8 Operating System.

Table 1.

Classification Alg.	*Confusion Matrix*
Decision Tree Induction	Predicted Yes / Predicted No; Actual Yes: 149, 2; Actual No: 3, 249
Naïve Bayesian Classification	Predicted Yes / Predicted No; Actual Yes: 144, 24; Actual No: 67, 168
K- Nearest Neighbor	Predicted Yes / Predicted No; Actual Yes: 126, 42; Actual No: 36, 199

After we observe the confusion matrix of all the three classifiers, it is clear that Decision Tree Classifier predicts insurance grant with greater probability as compared to the other two classifiers .It exhibits less number of false positives and true negatives.

Classifier Accuracy: Classifier Accuracy simply measures the number of correct decisions the classifier makes out .This will achieved by dividing the total number of test examples, by the number of correct decisions your classifier makes as shown fig 4. Accuracy is a simplistic measure that is misleading on many real-world problems.

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{False positives} + \text{True negatives} + \text{False negatives}}$$

Fig. 4. Formula for Classifier Accuracy

We calculated and compared the accuracy of all the classifiers [10]. The results are shown in the Table 2. And Fig. 5. We can clearly observe that Decision Tree Classifier shows more accuracy compared to the rest of the two. KNN classifier shows better performance in comparison with Naïve Bayesian classifier.

Table 2. Accuracy of the 3 classifiers

Classification Algorithm	*Accuracy*
Decision Tree Induction	0.9875930521091811
Naïve Bayesian Classification	0.7741935483870968
K- Nearest Neighbor	0.8064516129032258

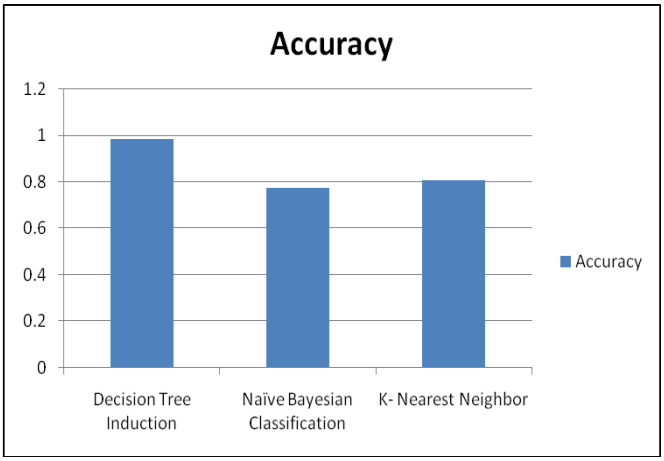

Fig 5. Accuracy Comparison of the 3 classifiers

VII. CONCLUSIONS AND FUTURE WORK

This paper performs an analysis on insurance customers' information by classifying the dataset of existing customers using three Machine Learning algorithms Decision Tree Induction, Naïve Bayes classification and K-Nearest Neighbour algorithm [11]. By employing this project, Insurance Company and customers both get benefited where an insurance company can predict the eligibility of the new customer for issuing a policy and customer confirms whether he is sanctioned against the policy. After generating the classification models, testing the accuracy of the classifiers and analyzing the confusion matrix, it is proved that the results are accurate for prediction [12]. This research can be further enhanced in future to compare accuracy of the existing methods against other classifiers and machine learning techniques such as Multiple Regression Analysis and Support Vector Machines [13].

REFERENCES

[1] Bhalla A. Enhancement in predictive model for insurance underwriting. Int J Comput Sci Eng Technol **3:160–165, 2012.**
[2] Sagar S. Nikam,2015. "A Comparative Study of Classification Techniques in Data Mining Algorithms". Oriental Journal of Computer Science & Technology, **Vol. 8, April 2015.**
[3] Mamun DMZ, Ali K, Bhuiyan P, Khan S, Hossain S, Ibrahim M, Huda K. Problems and prospects of insurance business in

Bangladesh from the companies' perspective. Insur J Bangladesh Insurance Acad **62:5–164, 2016.**

[4] Fang K, Jiang Y, Song M. Customer profitability forecasting using Big Data analytics: a case study of the insurance industry. Comput Ind Eng **101:554–564, 2016.**

[5] Cummins J, Smith B, Vance R, Vanderhel J. "Risk classificaition in Life Insurance". 1st edn. Springer, New York, **2013.**

[6] S.Archana and Dr. K. Elangovan, 2014. "Survey of Classification Techniques in Data Mining". International Journal of Computer Science and Mobile Applications, **Vol. 2 Issue. 2, February 2014.**

[7] Bhavesh Patankar and Dr. Vijay Chavda, 2014. "A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining". International Journal of Advanced Research in Computer Science and Software Engineering, **Vol. 4, Issue 12, December 2014.**

[8] K. P. Soman, 2006 . "Insight into Data Mining Theory and Practice", New Delhi: PHI, **2006**.

[9] S. B. Kotsiantis, 2007. "Supervised Machine Learning: A Review of Classification Techniques". Informatica, **vol. 31, pp. 249-268, 2007.**

[10] H. Bhavsar and A. Ganatra, 2012. "A Comparative Study of Training Algorithms for Supervised Machine Learning". International Journal of Soft Computing and Engineering (IJSCE), **Vol. 2, Issue. 4, September 2012.**

[11] Brijain R. Patel and Kushik K.Rana, 2014. "A Survey on Decision Tree Algorithm for Classification". International Journal of Engineering Development and Research, **2014**.

[12] Matthew N. Anyanwu and Sajjan G. Shiva, 2009. "Comparative Analysis of Serial Decision Tree Classification Algorithms". Researchgate, **January 2009.**

[13] Saurav Singla , Vikash Kumar, 2020. Multi-Class Sentiment Classification using Machine Learning and Deep Learning Techniques. International Journal of Computer Sciences and Engineering (IJCSE). **Vol. 8, Issue.11, November 2020** E-ISSN: 2347-2693. DOI: https://doi.org/10.26438/ijcse/v8i11.1420.

AUTHORS PROFILE

Ch. Lakshman Vinay is an engineering graudate in computer science from Gayatri Vidya Parishad College for Degree and PG Courses (A) . He areas of interest are Web Technologies and Machine Learning. He believes in hardworking and dreams to lead becom a projet lead in software industry.

G. Vijay Sagar is an engineering graudate in computer science from Gayatri Vidya Parishad College for Degree and PG Courses (A) . He areas of interest are Web Development and Data Science. He is interested in contributing to develop open source projects, pursuiing higher studies engage in research.

Merapureddy Ajay is an engineering graudate in computer science from Gayatri Vidya Parishad College for Degree and PG Courses (A). He got the Merit Scholarship Awardee thrice in a row under Educational Scholarship Scheme from Director General Naval Projects (2014 – 2016) and stood in the 3rd place in "SWISH Sunrise Indian Innovative Student Hackathon" conducted by Japanese Company The Denso Group at GITAM University in 2019.

SK. Hussain is an engineering graudate in computer science from Gayatri Vidya Parishad College for Degree and PG Courses (A). He wants to become a Software Engineer and is passionate in working in Web Development and he dreams to work as a software engineer in MNCs.

Dr. Smt Bh Padma is working as an Associate Professsor in the Department of Computer Science and Engineering, GVPPG,Visakhapatnam, India. She has got her doctorate from GITAM University under the supervision of Dr GVS Raj Kumar, Dept of IT, GITAM. Her area of specialization is Cryptography and Network Security and Machine Learning. Till now she has published 16 research articles in various prominent peer reviewed journals and presented papers in various conferences in this research area.

