Risk Prediction for New Insurance Using Customer Information

***Durgesh Sharma*[1], *Harshala Gaikwad*[2], *Kartavya Verma*[3]**

[1]*Department of Computer Science Engineering, Acropolis Institute of Technology & Research, Indore, Madhya Pradesh*

[2]*Department of Computer Science Engineering, Acropolis Institute of Technology & Research, Indore, Madhya Pradesh*

[3]*Department of Computer Science Engineering, Acropolis Institute of Technology & Research, Indore, Madhya Pradesh*

Abstract:
Risk management is important for new authenticated customers to identify their correct insurance policy. Therefore, risk prediction of new insurance is a crucial factor in insurance business to classify the applicants. Analyzing the profile of individual applicant manually may take a very long time. By applying predictive modelling techniques may automatically classify risk level based on past data more quickly and accurately in less time and less labor.

Our project uses the data-set containing past customer information including age, height, BMI, family history, insurance history etc. along with the risk level. Hence here Supervised Machine Learning Algorithm is used to predict the risk level of different applicants. We apply Random Forest Algorithm to predict the risk level of new customer automatically.

Key Words: Predictive models, Supervised learning, Random forest algorithm.

I. INTRODUCTION

Insurance Industry plays an important role in the sustainable economic growth. Risk assessment in insurance industry is beneficial in determining fraudulent and genuine insurance buyers. Having detailed risk analysis has become an absolute necessity for an insurance company as the number of insurance buyers are increasing. It also helps during the underwriting and accept/reject stage of the insurance policy. Insurance Companies still rely on the conventional actuarial formulas to predict risk levels of policy holder. Analyze individual applicant data thoroughly and manually may take a very large amount of time.

Nowadays data analytics are gaining significance importance in risk assessment of insurance policy holders. Insurance companies have recently started carrying out predictive analytics to improve their business efficiency, but there is still a lack of extensive research on how predictive analytics can enrich the life insurance domain. Researchers concentrated on data analytics techniques to detect frauds among insurance holders, which is a crucial issue due to the companies facing great losses.

Today Insurance companies are focusing on the power of big data and machine learning to collect, process and manage huge amount of customer data in order to predict the risk level. With the help of these trending technology we try to apply predictive modelling approach that help to predict the risk level based on the customer information and previous historic insurance data. These predictions further help in the acceptance and rejection of the applied insurance policy.

Predictive analytics approach in life insurance mainly deals with modelling risk level of applicants to improve underwriting decisions and profitability of the business. Risk profiles of individual policy applicants are thoroughly analyzed by underwriters, especially in the life insurance business. The job of the underwriters is to make sure that the risks are evaluated, and premiums as accurately as possible to sustain the smooth running of the business.

Risk assessment is a common term used among insurance companies, which refers grouping customers according to their evaluated level of risks, determined from their historical data. Since decades, life insurance firms have been relying on the traditional methods and actuarial formulas to estimate life expectancy and devise underwriting rules. However, the traditional techniques are time-consuming, usually taking over a month and also costly. Hence, it is essential to find ways to make the underwriting process faster and more economical. However, broad research has not been conducted in this area.

The purpose of this research is to apply predictive modelling to classify the risk level based on the available past data in the insurance company and recommend the most appropriate model to identify risk and provide solutions to refine underwriting processes.

II. PROBLEM FORMULATION

Risk management is an important factor in any insurance industry. Insurers consider every available quantifiable factor to develop profiles of high and low insurance risk. Level of risk will be able to determine the expected insurance premiums.

Generally, insurance policies involving factors with greater risk of claims are charged at a higher rate. For this insurer collect a vast amount of information about policy holders. Calculating risk manually may take very long time. Therefore, calculating it using Machine Learning model may increase accuracy, performance and risk can be predicted automatically.

III. LITERATURE SURVEY

According to [1], Insuring involves gathering all the information about the customer, which can be a lengthy process. The customer usually undergo several physical tests and needs to submit all the proper documents to the insurance company. Then, the company calculates the risk profile of the applicant and evaluates if the application needs to be accepted or rejected. Afterwards, premiums are calculated [2]. On an average, at least 30 to 40 days are taken for an application of an insurance to be processed. However, in today's scenario, people are unwilling to buy the services that are slow. Due to the insuring processes being so lengthy and time & energy consuming, applicants are more liable to change to a competitor or they simply avoid buying this insurance policies. Unethical insuring practices can lead to applicants being displeased which leads to decrease in insurance policy sales.

The insuring company service standard is an important aspect in determining the fame of life insurance companies and helps to carry on an upper position in the competitive market [3]. Thus, it is an important part to constantly improve the insuring method to build up customer trust and their retention.

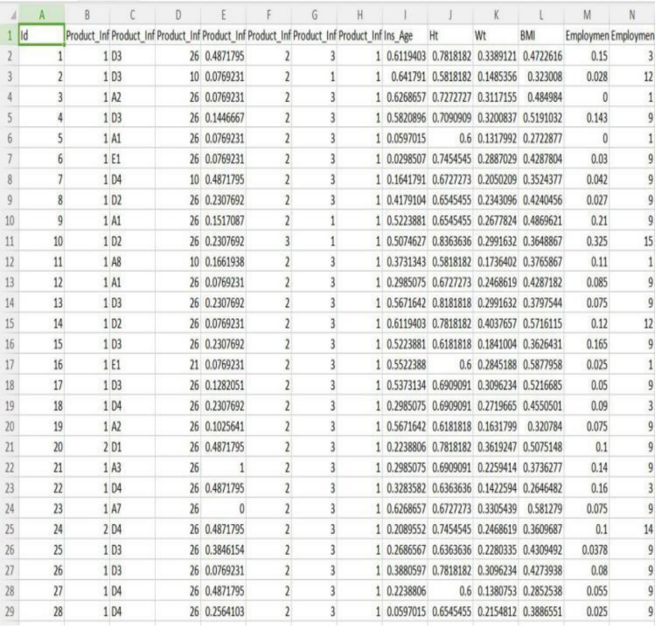
Likewise, insuring process and the plan of actions of medical methods are required by the insurance company to evaluate the risks of different applicants and customers which can be expensive [4]. Generally, all the costs to carry out the medical procedures are initially carried out by the firm. Costs of underwriting are fully paid from the contract and can last up to 10–20 years. And if, policy lapse happens, the insurer incurs great losses [5]. Therefore, it is important to automatize the insuring method using ordered processes. Forecasting the crucial factors impacting the risk management process can help to smoothen the procedures, making it well organized, efficient, economical and flexible for the applicants.

A research by [6] shows that cheap insuring proportions are a well-known problem among companies of insurance surveyed in some countries. One more ultimatum to the life insurance corporate is that they can bear adverse or unlucky selection. Unlucky selection refers to a situation where a high risky profile life insurance is given to the customer when the insurers do not have proper or all information on the customer [7]. Insurance companies always try to make least possible losses with certified insuring teams. Specially, the insurers attempt to avoid pitiful selection as it can have heavy impacts on the business of life insurance [8].

By accurately classifying the risk levels of independent applications through predictive analytics, adverse or pitiful selection can be avoided.

IV. METHODS AND TECHNIQUE

This research paper approach involves the collection of labelled insurance data from kaggle. This dataset consists of 128 different attributes of 59381 customers. It contains the customer information such as age, weight, bmi, height, family history etc. along the risk level between 1-8. Table 4.1 shows the insurance dataset taken from kaggle.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Id	Product_Inf	Product_Inf	Product_Inf	Product_Inf	Product_Inf	Product_Inf	Product_Inf	Ins_Age	Ht	Wt	BMI	Employmen	Employmen
2	1	1	D3	26	0.4871795	2	3	1	0.6119403	0.7818182	0.3389121	0.4722616	0.15	3
3	2	1	D3	10	0.0769231	2	1	1	0.641791	0.5818182	0.1485356	0.323008	0.028	12
4	3	1	A2	26	0.0769231	2	3	1	0.6268657	0.7272727	0.3117155	0.484984	0	1
5	4	1	D3	26	0.1446667	2	3	1	0.5820896	0.7090909	0.3200837	0.5191032	0.143	9
6	5	1	A1	26	0.0769231	2	3	1	0.0597015	0.6	0.1317992	0.2722877	0	1
7	6	1	E1	26	0.0769231	2	3	1	0.0298507	0.7454545	0.2887029	0.4287804	0.03	9
8	7	1	D4	10	0.4871795	2	3	1	0.1641791	0.6727273	0.2050209	0.3524377	0.042	9
9	8	1	D2	26	0.2307692	2	3	1	0.4179104	0.6545455	0.2343096	0.4240456	0.027	9
10	9	1	A1	26	0.1517087	2	1	1	0.5223881	0.6545455	0.2677824	0.4869621	0.21	9
11	10	1	D2	26	0.2307692	3	1	1	0.5074627	0.8363636	0.2991632	0.3648867	0.325	15
12	11	1	A8	10	0.1661938	2	3	1	0.3731343	0.5818182	0.1736402	0.3765867	0.11	1
13	12	1	A1	26	0.0769231	2	3	1	0.2985075	0.6727273	0.2468619	0.4287182	0.085	9
14	13	1	D3	26	0.2307692	2	3	1	0.5671642	0.8181818	0.2991632	0.3797544	0.075	9
15	14	1	D2	26	0.0769231	2	3	1	0.6119403	0.7818182	0.4037657	0.5716115	0.12	12
16	15	1	D3	26	0.2307692	2	3	1	0.5223881	0.6181818	0.1841004	0.3626431	0.165	9
17	16	1	E1	21	0.0769231	2	3	1	0.5522388	0.6	0.2845188	0.5877958	0.025	1
18	17	1	D3	26	0.1282051	2	3	1	0.5373134	0.6909091	0.3096234	0.5216685	0.05	9
19	18	1	D4	26	0.2307692	2	3	1	0.2985075	0.6909091	0.2719665	0.4550501	0.09	3
20	19	1	A2	26	0.1025641	2	3	1	0.5671642	0.6181818	0.1631799	0.320784	0.075	9
21	20	2	D1	26	0.4871795	2	3	1	0.2238806	0.7818182	0.3619247	0.5075148	0.1	9
22	21	1	A3	26	1	2	3	1	0.2985075	0.6909091	0.2259414	0.3736277	0.14	9
23	22	1	D4	26	0.4871795	2	3	1	0.3283582	0.6363636	0.1422594	0.2646482	0.16	3
24	23	1	A7	26	0	2	3	1	0.6268657	0.6727273	0.3305439	0.581279	0.075	9
25	24	2	D4	26	0.4871795	2	3	1	0.2089552	0.7454545	0.2468619	0.3609687	0.1	14
26	25	1	D3	26	0.3846154	2	3	1	0.2686567	0.6363636	0.2280335	0.4309492	0.0378	9
27	26	1	D3	26	0.0769231	2	3	1	0.3880597	0.7818182	0.3096234	0.4273938	0.08	9
28	27	1	D4	26	0.4871795	2	3	1	0.2238806	0.6	0.1380753	0.2852538	0.055	9
29	28	1	D4	26	0.2564103	2	3	1	0.0597015	0.6545455	0.2154812	0.3886551	0.025	9

Table 4.1: Dataset

The research paradigm focus in predictive study involving the use of machine learning algorithm to support the research objectives. Figure 4.1 shows the system flow chart. It gives an idea of the stages that have to flow to build the predictive models for classifying risk levels of the policy holders.

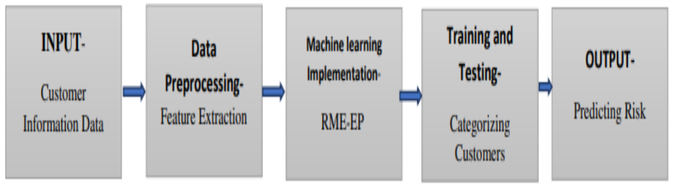
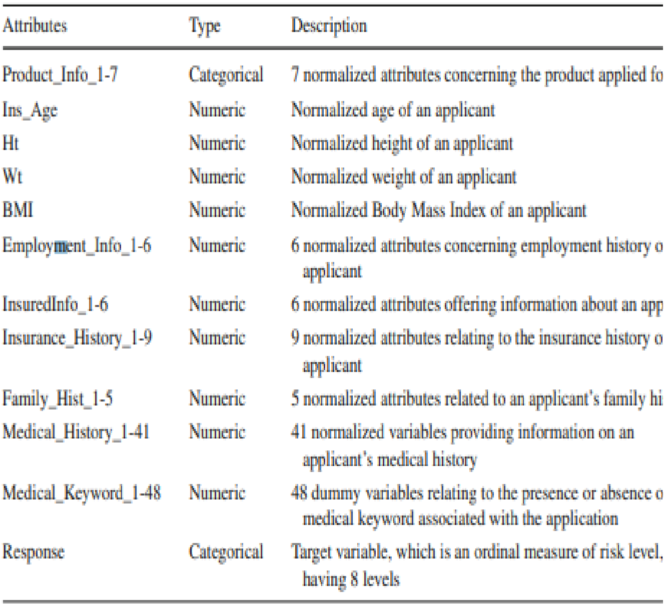


Figure 4.1: Flow Chart

Description of data set

In this dataset, we are provided over a various variables describing attributes of life insurance applicants. The customer attributes present in the dataset involves in the prediction of fraudulent or genuine customers. These attribute includes information of previous insurance customers such as their age, height, weight, bmi, family history etc.

The task of this research is to predict the "Response" variable for each customer having unique Id. "Response" is an attribute that measure risk that has 8 levels. The obtained dataset is labelled therefore we try to apply supervised predictive modelling approach to automate the process of analysis of insurance policy holders Table 4.2 describes the variables of attributes in dataset



Attributes	Type	Description
Product_Info_1-7	Categorical	7 normalized attributes concerning the product applied fo
Ins_Age	Numeric	Normalized age of an applicant
Ht	Numeric	Normalized height of an applicant
Wt	Numeric	Normalized weight of an applicant
BMI	Numeric	Normalized Body Mass Index of an applicant
Employment_Info_1-6	Numeric	6 normalized attributes concerning employment history o applicant
InsuredInfo_1-6	Numeric	6 normalized attributes offering information about an app
Insurance_History_1-9	Numeric	9 normalized attributes relating to the insurance history o applicant
Family_Hist_1-5	Numeric	5 normalized attributes related to an applicant's family hi
Medical_History_1-41	Numeric	41 normalized variables providing information on an applicant's medical history
Medical_Keyword_1-48	Numeric	48 dummy variables relating to the presence or absence o medical keyword associated with the application
Response	Categorical	Target variable, which is an ordinal measure of risk level, having 8 levels

Table 4.2: data description

Fig. 4.2 shows several graphs that are interactive with each other. This mainly presents the distribution of demographic variables in the data set. For instance, BMI, age, weight, and family history, response variable and how they vary. It provides insights into the customer data. Thus, the life insurance company knows its applicants better and has better assurance with them.

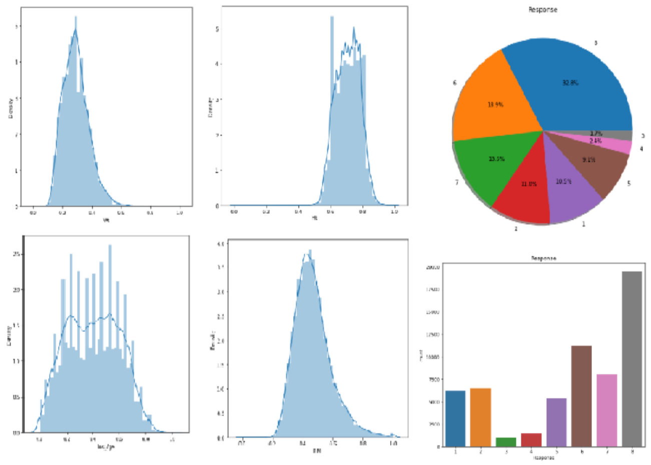
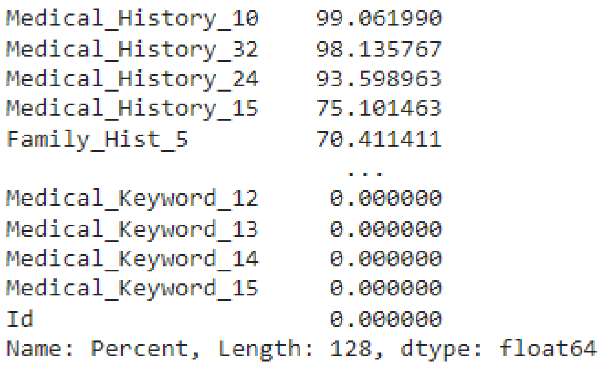


Figure 4.2: Graphs

Data Pre-processing

Data pre-processing is a technique in data mining used to transform the raw data into a useful and efficient format. It is also known as data cleaning. It helps to remove noisy data, inconsistent data, outliers and handles the missing data in the dataset. This step also grid the development of any strategies needed to deal with the inconsistencies in the target data. Specific attribute values will be transformed to ease analysis and interpretation.

In this the gathered data will be cleaned to treat missing values to make the data consistent with analysis. The data set has attributes with a remarkable amount of missing data which is show in Fig. 4.3



```
Medical_History_10    99.061990
Medical_History_32    98.135767
Medical_History_24    93.598963
Medical_History_15    75.101463
Family_Hist_5         70.411411
                         ...
Medical_Keyword_12     0.000000
Medical_Keyword_13     0.000000
Medical_Keyword_14     0.000000
Medical_Keyword_15     0.000000
Id                     0.000000
Name: Percent, Length: 128, dtype: float64
```

Figure 4.3: Missing data

In this research we have replace the missing values with the mean of the values in particular attribute. Also we have dropped the column having more than 70 percent missing values.

Encoding

Encoding is a Machine Learning technique used for converting categorical variables into numerical values so that it could be easily fitted to a machine learning model. There are different encoding techniques available: One hot encoding, mean encoding, label encoding, target guided ordinal encoding.

In this dataset, we have applied label encoding, which convert categorical values into numeric forms.

Supervised Machine Learning Algorithm

In this research we are using supervised machine learning techniques. The dataset collected from kaggle website is a labelled data that contains risk levels (1-8) as a response variable. This section will elaborate different Supervised Learning algorithms implemented on the insurance data set obtained from kaggle to build the predictive models. The techniques namely, logistic regression, random forest, K-Nearest Neighbors(KNN) and Naïve Bayes.

Random Forest

Random forest is a Supervised Machine Learning Algorithm. It is used for solving both Classification and Regression problems. It builds decision trees on different samples and takes their average for regression and majority vote in case of classification. Random Tree is a supervised machine learning algorithm which accounts for k randomly selected attributes at each node in the decision tree. Or we can say that, random tree classifier builds a decision tree based on random selection of data as well as by randomly choosing attributes in the data set. Random forest doesn't rely on one decision tree, rather it takes the prediction from each tree and on the basis of majority of votes for predictions, it predicts the final output as a result. The accuracy is directly proportional to the number of trees in the random forest. Greater number of tree in random forest prevents the problem of overfitting, and hence increases the accuracy.

Logistic Regression

Logistic regression is a Supervised Statistical Learning technique. Logistic Regression is a machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. It helps in the predictions of a categorical dependent variable. Therefore, the outcome of logistic regression must be a categorical or discrete value. The output of Logistic Regression is discrete (i.e. Yes or No, 0 or 1, true or False, etc.), but it gives the probabilistic values which lie between 0 and 1.

K Nearest Neighbors(KNN)

K-Nearest Neighbors is a Supervised Machine Learning algorithms. This algorithm is based on the assumption of the similarity between the new data and already available data. KNN tries to place the new case into the category that is most similar to the available categories with the help of different distance algorithm available. K-NN algorithm uses all the available data in train set (i.e. it does not split the data into training and testing set, rather it performs training on complete dataset) and classifies a new data point based on the similarity. Whenever the new data appears then it can be easily classified into a well suite category with the help of K-NN algorithm.

Naïve Bayes

Naive Bayes is a classification based Supervised Machine Learning algorithm. This algorithm is suitable for both binary and multiclass classification. It classifies future objects by assigning class labels to instances/records using conditional probability. Naive Bayes is a probabilistic classifier; it predicts the outcome on the basis of the probability of an object.

ALGORITHM	ACCURACY
Random Forest	82.94%
Logistic Regression	81.11%
KNN	68.47%
Naïve Bayes	64.97%

Table 4.3: Accuracy Comparison of different algorithm

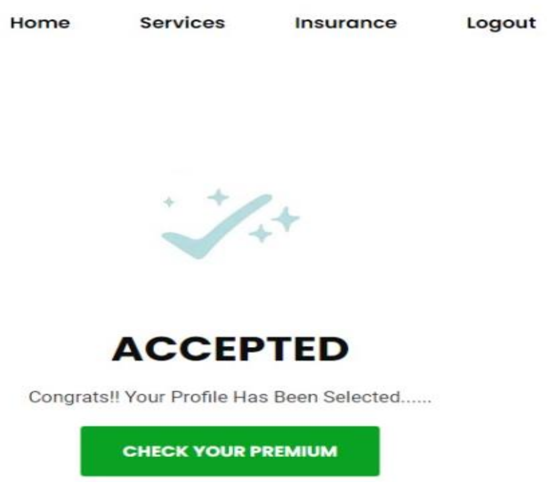
Table 4.3 shows the Comparison of accuracy of different predictive models. From the table it is clear that Random forest shows the best accuracy of 82.94%. Therefore, random forest is the best suited model for our research. By using random forest algorithm, we try to predict the risk levels of new customers which would be helpful of insurer to select the genuine customers for insurance policies.

Deployment

This research is deployed using Python Django Framework. We have used build-In Authentication forms provided by django for user signup/signin and Django forms to take inputs from new customers. Machine Learning model is integrated by converting the model into .sav file. This file helps to predict the risk level. Based on the acceptance of the policy it further display the premium to the customer.For frontend we had used HTML, CSS, Bootstrap and JS.

V. RESULT DISCUSSION

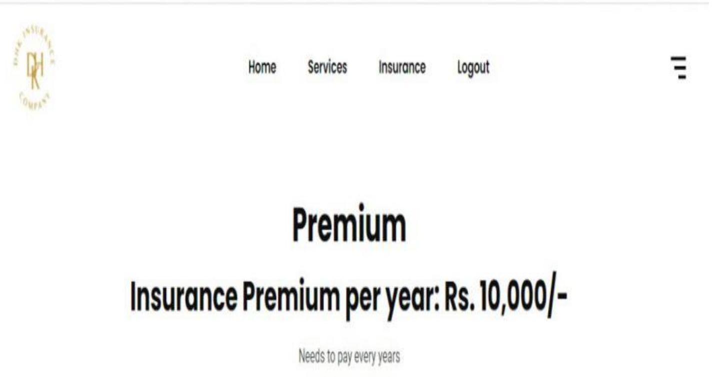
The deployed Machine learning model on Django framework will check the input values from the user and according to the values it will accept or reject the insurance policy as per the algorithm. For each user, if the insurance policy is suitable according to the parameters, the result will be accepted as shown in the figure5.1 and premium is display to the user as shown in fig. 5.2 and if the insurance policy is not suitable according to the parameters, the result will be rejected.



ACCEPTED

Congrats!! Your Profile Has Been Selected......

Figure 5.1: Accepting Policy

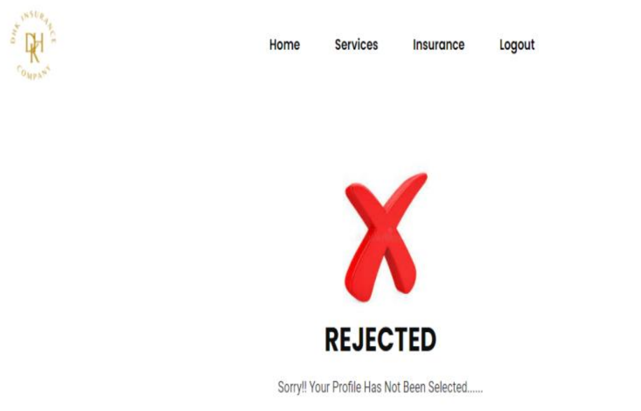


Premium

Insurance Premium per year: Rs. 10,000/-

Needs to pay every years

Figure 5.2: Premium



REJECTED

Sorry!! Your Profile Has Not Been Selected......

Figure 5.2: Rejected Policy

VI. CONCLUSION

Data analytics is gaining significance among the top IT companies worldwide. In the life insurance domain, predictive modelling using learning algorithm scan provide the notable difference in the way which business is done as compared to the traditional methods.

Previously, risk assessment for life endowing was conducted using complex actuarial formulas and usually was a very extensive process. Now, the work can be done faster and with better results with the help of data analytical solutions.

Therefore, it would enhance the business by allowing faster service to customer, thereby increasing fulfilment and loyalty.

VII. Acknowledgement:

We as the authors would like to extend a special thanks of vote to the Prof. Narendra Pal Singh Rathore, our guide for their valuable guidance and suggestions to improve this paper without which it would have been a very difficult task. The paper is supported by Acropolis Institute of technology and Research, Indore.

VIII. References:

[1] Wuppermann A (2016) Private information in life insurance, annuity and health insurance markets. Scand J Econ 119:1–45.

[2] Prince A (2016) Tantamount to fraud? Exploring non-disclosure of genetic information in life insurance applications as grounds for policy rescission. Health Matrix 26:255–307.

[3] Chen TJ (2016) Corporate reputation and financial performance of life insurers. Geneva Papers Risk Insur Issues Pract 41:378–397.

[4] Huang Y, Kamiya S, Schmit J (2016) A model of underwriting and post-loss Test without commitment in competitive insurance market. SSRN Electron J.

[5] Carson J, Ellis CM, Hoyt RE, Ostaszewski K (2017) Sunk costs and screening: two-part tariffs in life insurance. SSRN Electron J 1–26.

[6] Mamun DMZ, Ali K, Bhuiyan P, Khan S, Hossain S, Ibrahim M, Huda K (2016) Problems and prospects of insurance business in Bangladesh from the companies' perspective. Insur J Bangladesh Insurance Acad 62:5–164.

[7] Harri T, Yelowitz A (2014) Is there adverse selection in the life insurance market? Evidence from a representative sample of purchasers. Econ Lett 124:520–522.

[8] Hedengren D, Stratmann T (2016) Is there adverse selection in life insurance markets? Econ Inq 54:450–463.

IX. Authors:

1.Durgesh Sharma, B.Tech CSE,4th Year student, Acropolis Institute of Technology & Research.

2.Harshala Gaikwad, B.Tech CSE,4th Year student, Acropolis Institute of Technology & Research.

3.Kartavya Verma, B.Tech CSE,4th Year student, Acropolis Institute of Technology & Research.