ABSTRACT

Insurance market is growing day by day. Industries are facing a severe loss of profits due to increasing contests hence the loss of customers. They are trying to find the causes of losing customers by measuring customer loyalty to regain the lost customers. The customers leaving the current industry and moving to another insurance industry are called churn. The research paper is using application of data mining technology and R package to predict the results of churn customers on the insurance transaction customer dataset from LIC, AVIVA, MAX LIFE Insurance Company. The R tool has represented the large dataset churn in form of graphs which depicts the outcomes in various unique pattern visualizations. The Churn Factor is used in many functions to depict the various areas or scenarios where churners can be distinguished. The paper is logistic regression, decision tree, Random forest, Support vector machine model building to predict churn customer in insurance sector considering churn factor in account to depict various patterns for churners. R is a powerful statistical programming tool which can represent the dataset graphically with respect to different parameters and it also uses different packages available. Churns can be reduced by analyzing the past history of the potential customers systematically. In the past few years, the fast emerging requirements from both academia and industry has helped R programming language to emerge as one of the necessary tool for visualization, computational statistics and data science.

Keywords: Churn, Prediction, R Tool, Insurance Industry, Data mining

INTRODUCTION

Numerous insurance industries are present all over the world. Insurance market is facing a severe loss of revenue due to increasing competition among them and loss of potential customers. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up in the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning. Therefore, it has become important for the manager and marketing professional to predict churn customer. Operators to earn back the amount they invested along with at least the minimum profit within a very short period of time.

OBJECTIVE

1. To review of the Churn prediction model generation.
2. To find out customer transaction behaviour of various customers.
3. To demonstration of predictive model result for customer churn prediction Using R tool

LITERATURE REVIEW

Rao D Tripati,[1] In his PhD thesis the researcher has designed at perceptive the development of life insurance sector in India. The major findings of this study are states that it has been experimental that in spite of the creditable growth and presentation of life insurance sector, a huge potential still exists. Wadikar Ashok Laxaman [2] In his thesis on authenticate a general view that innovativeness in each action alone rules and control in the insurance Sector. Tesfaye Hintsay Atsmo [3] This research describes the development of predictive model, which determines the risk exposure of motor insurance policies. Patrick Amofah,Amer ijaz[4] This research has been undertaken for understanding Objectives, Strategies and benefits of CRM. Samuel Odei Danso[5] This dissertation explores the data mining methods in order to classify classification of consumers that will respond through mailings in the insurance industry. Latha M N[6] In his study researcher has created understanding about health insurance among the community, presented a clear picture about health insurance saturation and suggest the possible market potential offered in the District of Coimbatore.

John Hadden [7] A method is framed that can classify customers and aims to develop a consumer profiling methodology for predicting churn in advance. Jahromi(2009)[8] The purpose of this study was to develop predictive model or application to analyze and predict customers churn or attrition which support the customer retention and to help to recognize the customers with high probability of churn in close future and target them with incentives in order to convince them to stay. This models was developed for telecommunications service only for predict the churners form non-churners and help the Talia Co. to conduct a more efficient retention campaign and built up a dual model, which involves clustering and classification phase. Bagal Laxman Bhimrao,[9] In his research study on “Think about the market policies of this monopoly for select plans and the efforts of LIC to keep the monopoly intact, how the new plans are helpful in increasing the overall growth of society?” The significance of new product launching and recognition of new marketing strategies are crucial aspects in the purview of forthcoming introduction of private insurance companies in Indian market. Reddy Y. Nagendra[10] The study was an attempt to customer satisfaction in life insurance business, A case study of life insurance corporation of India Anantapur branches, The researcher has reviewed the attributes of the protection shopper markets, character the conceivable wellsprings of disappointment, work towards dealing with dissension from shoppers, which will offer assistance in upgrading consumer loyalty and counteract disappointment. This review goes for seeking setbacks in meeting the fulfillment levels as expected by the protection customers.

RESEARCH QUESTIONS

1. To prevent churning out of the customers, the insurance companies must know:-
2. What is the profile, tastes, preferences and purchasing behavior of the customer?
3. What is the transaction behavior of various customers?
4. What services and benefits would current customers likely desire?
5. Identifying the customers who are getting all types of services from your company?
6. Which are the satisfaction parameters and which is their impact to Consumer behavior?
7. How many Consumers are satisfied or dissatisfied?
8. Which is the average satisfaction level according to Consumers preference and expectations?
RESEARCH METHODOLOGY

Data Collection

Primary collection of data is collected through Interviews of development officer, Agent and customers, Questionnaire, Telephonic communication, E-mail, Group discussion and Observations.

The Sources for secondary data are as websites of IRDA and Insurance companies, IRDA Public disclosure of LIC, AVIVA, MAX LIFE INSURANCE, Annual reports of LIC, AVIVA, MAX LIFE INSURANCE, A presentation of company’s officials at various conferences, Reports and Publications by Insurance companies.

Data selection

By analysis of a customer’s usual usage behaviour and monitoring activity, By monitoring a customer’s spending behaviour; transaction history from a particular customer can be interpreted as a customer’s intention to churn.

By segmenting customers by demographic location a clear understanding of typical behaviour can be determined. This would be more beneficial to the insurance industry where poor coverage in a particular area could signify poor retention rates. By segmenting customers by age group, it is possible to determine the behaviour of certain age ranges. E.g. age group 18-24 could value technology most while a customer of 35-45 age groups could be more favourable towards service performance.

Prediction models generation on training set created vary in the relation between churners and non-churners. This is done to beat the imbalanced informational collection issue. Logistic Regression, Decision Trees, Support vector machine, Random Forest represents the training set 80 and test set 20 (non-churning: churning) with the original distribution. These training set distributions are always linked to the same model number for all techniques. The following Figure depicts an overview of how these training sets are generated.

Figure 1: Construction of the training and test from the insurance customer data base

Data sets of 10,000 cases are selected from the customer transaction database of insurance customers. One data set which represents the original distribution between churners and non-churners, a set with only churning customers and one set with non-churning customers are extracted from the insurance customer database.
DEMONSTRATION OF PREDICTIVE MODEL ANALYSIS

A data set for the year 2009 to 2015 of usage with 9084 customers’ records of an anonymous insurance sector in India is used for demonstrating and evaluating the proposed framework. In all 22 parameters are used in order to define five categories (behavioural, dissatisfaction, profitability, cross-selling, demographic) of churners, predicting the churners throughout comparative study between Random forest, Decision Tree, Support Vector Machine, and logistic regression of this data.

The results of the predictive models in table no indicates that outperforms Support Vector Machine but both outperform Decision Tree.

Table no 1: Predictive model performance

<table>
<thead>
<tr>
<th>Models</th>
<th>Logistic Regression</th>
<th>Decision Tree</th>
<th>Random Forest</th>
<th>Random Forest</th>
<th>Support vector machine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train data</td>
<td>Test data</td>
<td>Train data</td>
<td>Test data</td>
<td>Train data</td>
</tr>
<tr>
<td>Accuracy:</td>
<td>88%</td>
<td>87%</td>
<td>89%</td>
<td>91%</td>
<td>89%</td>
</tr>
<tr>
<td>Misclassification Rate:</td>
<td>12%</td>
<td>13%</td>
<td>11%</td>
<td>9%</td>
<td>11%</td>
</tr>
<tr>
<td>True Positive Rate:</td>
<td>75%</td>
<td>71%</td>
<td>68%</td>
<td>73%</td>
<td>73%</td>
</tr>
<tr>
<td>False Positive Rate:</td>
<td>8%</td>
<td>8%</td>
<td>4%</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>Specificity:</td>
<td>92%</td>
<td>92%</td>
<td>96%</td>
<td>97%</td>
<td>94%</td>
</tr>
<tr>
<td>Precision:</td>
<td>74%</td>
<td>75%</td>
<td>84%</td>
<td>89%</td>
<td>81%</td>
</tr>
<tr>
<td>Prevalence:</td>
<td>24%</td>
<td>25%</td>
<td>25%</td>
<td>24%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Graph No 1: Analysis of Predictive models

It is based on the insurance transaction data that combining multiple predictive models could outperform any of those models used separately. In the experiments, the researcher has used a combination of Random forests, logistic regression, decision trees and SVM. In graph No 1: Analysis of Predictive models represented predictive method is particularly useful for identifying churn customer in Insurance sector.
Table No 2: Predictive Model result

<table>
<thead>
<tr>
<th>Models</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
<th>Support vector machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gini Coefficient for train data:</td>
<td>0.76</td>
<td>0.78</td>
<td>1</td>
<td>0.86528</td>
</tr>
<tr>
<td>Gini Coefficient for test data:</td>
<td>0.74</td>
<td>0.8</td>
<td>0.86962</td>
<td>0.80616</td>
</tr>
</tbody>
</table>

As the observations are on Gini coefficient, if the coefficient above 0.6 it means model working good; it has been observed in table no 2: Predictive Model result. Random forests, logistic regression, decision trees and Support Vector Machine models gini coefficient is above 0.6 it means all predictive model working good; The abbreviation recorded henceforth i.e. from table no 2, the researcher opinion that Random forest model is best suitable for prediction methodology.

Graph No1: Key statistics of Model

Table No 3: Statistical result summary

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train data</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.88</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.89</td>
</tr>
<tr>
<td>Random Forest</td>
<td>1</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>0.93264</td>
</tr>
</tbody>
</table>

It has been observed Gini coefficient Random forest model is suitable for churn prediction

CONCLUSION

The proposed research has used application of data mining R tool and technique for developing predictive model in Insurance industry and using R package to predict the results of churn customers on the insurance churn dataset. It has evaluated, the number of churms using Random forests, logistic regression, decision trees and Support vector machine. The study predicts that there is a huge deviation in graph of churners when customer transaction histories are measured. The graphs are made taking churn factors as the deciding parameters. Graphs represent the different ways of observing the number of churners from the dataset. Once the root area is recognized the steps can be taken by insurance industry to improve their services and retain their old customers from churning.
REFERENCES

8. Jahromi , ”Predicting Customer Churn in Telecommunications Service Providers”, lulea University,2009
9. Bagal Laxman Bhimrao,” An analytical study of marketing strategies of life insurance corporation of india” Dr. Babasaheb Ambedkar Marathwada University, 2010