Predictive Modeling

Prepared by
Kanksha Masrani (14IT058)
&
Kunal Mehta (14IT059)

Under the supervision of

Mr. Pritesh Prajapati

A Report Submitted to

Charotar University of Science and Technology

for Partial Fulfillment of the Requirements for the

Degree of Bachelor of Technology

in Information Technology

IT 414 Software Group Project-V (7^{th} Sem)

Submitted at

DEPARTMENT OF INFORMATION TECHNOLOGY

Chandubhai S. Patel Institute of Technology

At: Changa, Dist: Anand – 388425

November 2017

CERTIFICATE

This is to certify that the report entitled "**Predictive Modeling"** is a bonafied work carried out by **Kanksha Masrani (14IT058) and Kunal Mehta (14IT059)** under the guidance and supervision of **Mr. Pritesh Prajapati for** the subject **Software Group Project-V (IT 414)** of **7th** Semester of Bachelor of Technology in **Information Technology** at Faculty of Technology & Engineering – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate herself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Under supervision of,

Pritesh Prajapati
Assistant Professor
Dept. of Information Technology
CSPIT, Changa, Gujarat.

Prof. Parth Shah
Head & Associate Professor
Department of Information Technology
CSPIT, Changa, Gujarat.

Chandubhai S Patel Institute of Technology

At: Changa, Ta. Petlad, Dist. Anand, PIN: 38842. Gujarat

TABLE OF CONTENTS

List of Figures

List of Table

ABSTRACT

Analyzing data of insurance companies gives an important insight on how the customers are reacting to the offered insurance policies by the companies. This information can be used to predict the behavior of future policy holders. Insurance companies maintain a large database on their customers and policy related information. Data mining technique applied with proper preprocessing of data prove to be very efficient in extracting hidden information from data stored by life insurance companies. There are many data mining algorithms that can be applied to this huge set of data. The main focus of our work is to apply different classification techniques on the data provided by a insurance company. Attribute selection techniques are applied to properly classify the data. Classification techniques proved to be very useful in classifying customers according to their attributes. A comparative analysis of the performance of the classifiers is also reported in this Project.

Acknowledgement

The merits of our project depend only on the wide panorama of the people who have devoted their precious time, and provided valuable suggestion as well as guidance to our project. We are grateful to our project guide **Mr. Pritesh Prajapati** for his guidance throughout this project research and work

We also wish to thank all the faculty members of Information Technology and our respectable Head of Department **Prof. Parth Shah** for their constant help and efficient teaching procedures. I express my sincere gratitude to them for their constant support and valuable suggestion without which the successful completion of this project would not be possible.

CHAPTER 1
INTRODUCTION

1.1 PROJECT REVIEW

Data mining or Knowledge Discovery in Databases (KDD) aims at finding novel, interesting, and useful information in real-world data sets. Predictive modeling can be defined as the analysis of large data sets to make inferences or identity meaningful relationships, and the use of these relationships to better predict future events. It uses statistical tools to separate systematic patterns from random noise, and turns the information into business rules, which should lead to better decision making. Insurers have begun to turn to predictive models for scientific guidance of expert decisions in areas such as claim management, fraud detection, premium audit, target marketing, cross selling, and agency recruiting and placement.

Our analysis is aimed at explaining characteristics of people who are likely to or not to buy policy, based on sam pled product usage data, such as contribution and number of insurance policies, and socio- demographical data, such as average household size and average income. Our original dataset, consists of 86 variables and 2500 cust omer records, of which are originally 1500 training data records, 500 testing data records and 500 validation data records. The response variable is in the binary form whose value is either Buyer or Non-Buyer of a policy.

1.2 SCOPE

The modern paradigm of predictive modeling has made possible a broadening, as well as deepening, of actuarial work. Predictive modeling has been effective in domains traditionally thought to be in the sole purview of human experts.

1.3 OBJECTIVE

Our objective for the prediction method combines Machine Learning algorithms for prediction with evolutionary search for choosing the predictive features. The result is a predictive model that uses only a subset of the original features, thus simplifying the model and reducing the risk of overfitting while maintaining accuracy. The historical data of the customers like their age, income, various other policies taken, marital status etc. will be used in order to the analysis. Later on some analysis will also be done to find the most relevant attributes i.e. the factors that affect the prediction the most.

CHAPTER 2
SYSTEM ANALYSIS

2.1 TOOLS AND TECHNOLOGY USED

- **Front end:** R using ggplot2 and python using matplotlib.
- **Back end :** R and Python

2.2.1 SPECIFIC REQUIREMENTS

SOFTWARE:

- R 3.4.0
- Python 3.6.0
- Data Mining Software- XLMiner.

HARDWARE:

- Windows Vista
- 2 GB RAM +
- 32 Bit Operating System

2.2 LIBRARIES REQUIRED

- Ggplot2
- Plyr
- Dplyr
- SciKit
- Pandas
- Matplotlib
- NumPy

CHAPTER 3
SYSTEM DESIGN

3.1 FLOW OF PROJECT

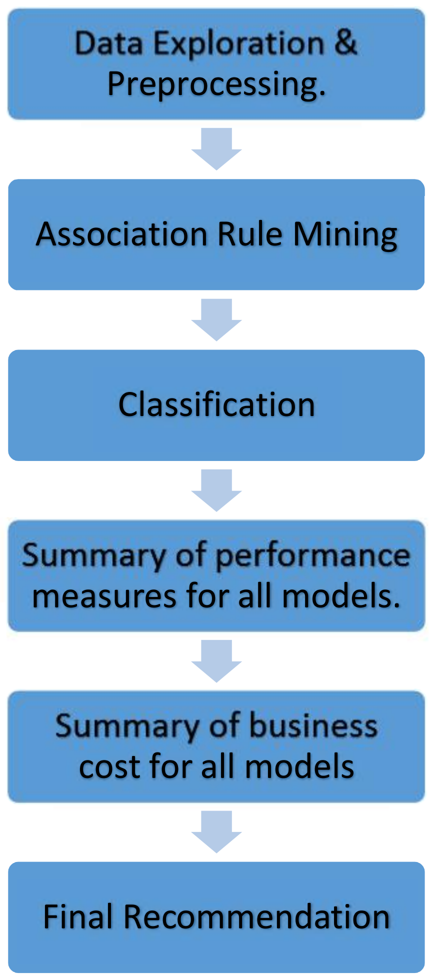

Fig: 3.1

3.2 DATA DICTIONARY

Attribute Number	Attribute Name
1	Customer Subtype
2	Number of houses
3	Avg size household
4	Avg age
5	Customer main type
6	Protestant
7	Other religion
8	No religion
9	Married
10	Living together
11	Other relation
12	Singles
13	Household without children
14	Household with children
15	High level education
16	Medium level education
17	Lower level education
18	High status
19	Entrepreneur
20	Farmer
21	Middle management
22	Skilled labourers
23	Unskilled labourers
24	Social class A
25	Social class B1
26	Social class B2
27	Social class C
28	Social class D
29	Rented house
30	Home owners
31	1 car
32	2 cars
33	No car
34	National Health Service
35	Private health insurance
36	Income < 30.000
37	Income 30-45.000
38	Income 45-75.000
39	Income 75-122.000
40	Income >123.000
41	Average income
42	Purchasing power class
43	Contribution private third party insurance
44	Contribution third party insurance (firms)
45	Contribution third party insurance (agriculture)
46	Contribution car policies
47	Contribution delivery van policies
48	Contribution motorcycle/scooter policies
49	Contribution lorry policies
50	Contribution trailer policies
51	Contribution tractor policies
52	Contribution agricultural machines policies
53	Contribution moped policies
54	Contribution life insurances
55	Contribution private accident insurance policies
56	Contribution family accidents insurance policies
57	Contribution disability insurance policies
58	Contribution fire policies
59	Contribution surfboard policies
60	Contribution boat policies
61	Contribution bicycle policies
62	Contribution property insurance policies
63	Contribution social security insurance policies
64	Number of private third party insurance 1 - 12
65	Number of third party insurance (firms)
66	Number of third party insurance (agriculture)
67	Number of car policies
68	Number of delivery van policies
69	Number of motorcycle/scooter policies
70	Number of lorry policies
71	Number of trailer policies
72	Number of tractor policies
73	Number of agricultural machines policies
74	Number of moped policies
75	Number of life insurances
76	Number of private accident insurance policies
77	Number of family accidents insurance policies
78	Number of disability insurance policies
79	Number of fire policies
80	Number of surfboard policies
81	Number of boat policies
82	Number of bicycle policies
83	Number of property insurance policies
84	Number of social security insurance policies
85	Whether the profile owns this insurance

Fig: 3.2

CHAPTER 4
IMPLEMENTATION

4.1 MODULE SPECIFICATION

Data Exploration and Pre-processing:

Explored the data set to see the validity of each variable as a useful predictor. Mainly used bar charts since almost all of the variables in our dataset are categorical. Based on our data analysis, we eliminated 78 variables that do not significantly distinguish people who are likely to buy the policy from those who are unlikely to. We also dropped variables which, by its nature, seem to be correlated and kept one in each group of correlated variables in our model.

We transformed some remaining variables to make our analysis more practical based on domain knowledge and insight from data visualization. For example, we created to binary variable, PRIV_3 RD, stating whether or not a person buy at least one private third party insurance while we dropped CON_PRIV_3 RD (Private third party insurance) even though the portion of success class in the records which have higher value of CON_PRIV_3rd tends to be higher. The model with binary variables is better than that with numerical variables because it is easy to interpret as long as the performance of each model does not show significant difference.

Analysis:

We constructed models using the following algorithms, Logistic Regression, Classification Tree, Naïve Bayes and SVM to analyse our data. We didn't use discriminant analysis because almost all of our variables are categorical and some of them are dummies which violates the assumption of discriminant analysis.

In our logistic regression analysis, we eliminated useless or insignificant variables based on our domain knowledge and p-values for each variable. After all, our final model had four explanatory variables;

- M_EDU_HIGH (people who have high education in a specific area)
- M_AVG_INCOME (Average income)
- PRIV_3 RD_INS
- CAR_INS

Of these four variables, CAR_INS had the lowest p-value, thus its significance contributed the most to the model. Based on our analysis, we found all these four variables were attributable to two main customer characteristics wealth and risk aversion. In our classification tree analysis, CAR_INS, M_EDU_HIGH and M_AVG_INCOME played an important role in explaining characteristics of insurance policy buyers. Although these two models yielded some similar results, that is, some variables in both the models had explanatory power, the logistic regression model seemed to perform better than the classification tree. In the logistic regression, the percentage errors of buyers in training and validation sets are (28.74% and 34.87%) lower than those of the classification tree (37.36% and 39.50%).

The following implications have been drawn:

1. The buyers are likely to live in a wealthy area.
2. Residents living in area which has high proportion of highly educated people are more likely to be high incomers, thus they have the policy.
3. A policy buyer is likely to own a car. In addition, private third party insurance policy is required for car owners. The ownership of car insurance and private third party insurance is a good indicator of car ownership.
4. If a person is risk averse, he/she is likely would buy insurance. The ownership of car insurance and private third-party insurance is a good indicator of risk averseness.

4.2 CODE IMPLEMENTATION.

Code of Main Modules:

Logistic Regression belongs to the family of generalized linear models. It is a binary classification algorithm used when the response variable is dichotomous (1 or 0).

We can evaluate Logistic regression model fit and accuracy by various metrics:

1. **AIC - Akaike Information Criteria (AIC)** : The smaller the better. Looking at the AIC metric of one model wouldn't help. It is more useful when you compare models and hence below we have three models with the third model with the least AIC value.

In R we use glm() function to apply Logistic Regression inclusive of all variable.

Model 1:

```
#logistic analysis
head(res[['train']])

logAnalysis <- glm(OUTCOME~.,data = res[['train']],family=binomial(link="logit"))
summary(logAnalysis)

anova(logAnalysis, test="Chisq")
```

Then we applied the ANOVA Chi-square test to check the overall effect of variables on the dependent variable and the variable which had p-value < 0.05 are listed below.

Variable	p-value
Customer Subtype	9.069e-05
Third Party Insurance	0.0007128
Car Policy	3.132e-06

We will remove those items which lessen the significance of the attributes above and create another model and compare their AIV values:

Model 2:

```
loganalysis2 <- glm(OUTCOME~ Customer.subtype+Contribution.third.party.insurance.f
summary(loganalysis2)

anova(loganalysis2, test="Chisq")
```

Variable	p-value
Customer Subtype	8.761e-05
Third Party Insurance	0.90461
Car Policy	0.0002246

Model 3:

```
loganalysis3<- glm(OUTCOME~ Customer.subtype+Number.of.car.policies, data = res[['train']], family = binomial(link = "logit"))
summary(loganalysis3)

anova(loganalysis3, test="Chisq")


anova(logAnalysis,loganalysis2,loganalysis3,test = "Chisq")
```

Variable	p-value
Customer Subtype	8.761e-05
Car Policy	0.0002254

Conclusion of AIC metric test:

Model	AIC Value
logAnalysis	546.68
Loganalysis2	502.95
Loganalysis3	501.01

Comparing the three models:

```
anova(logAnalysis,loganalysis2,loganalysis3,test = "Chisq")
```

```
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1      1068     306.88
2      1149     424.99 -81 -118.114   0.0045 **
3      1150     425.01  -1   -0.021   0.8847
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Third models p-value > 0.05 corroborates that the third model is better which has two attributes which are most significant with respect to the OUTCOME i.e. Customer Subtype and Number of car policies.

4.3 SNAPSHOTS

Below mentioned are the graphs of people who have the policy along with the other polices as mentioned in the title. Most of the people who have the policy do not have any other policy such as the life insurance policy or the social security policy. The only exception were the fire and car policy.

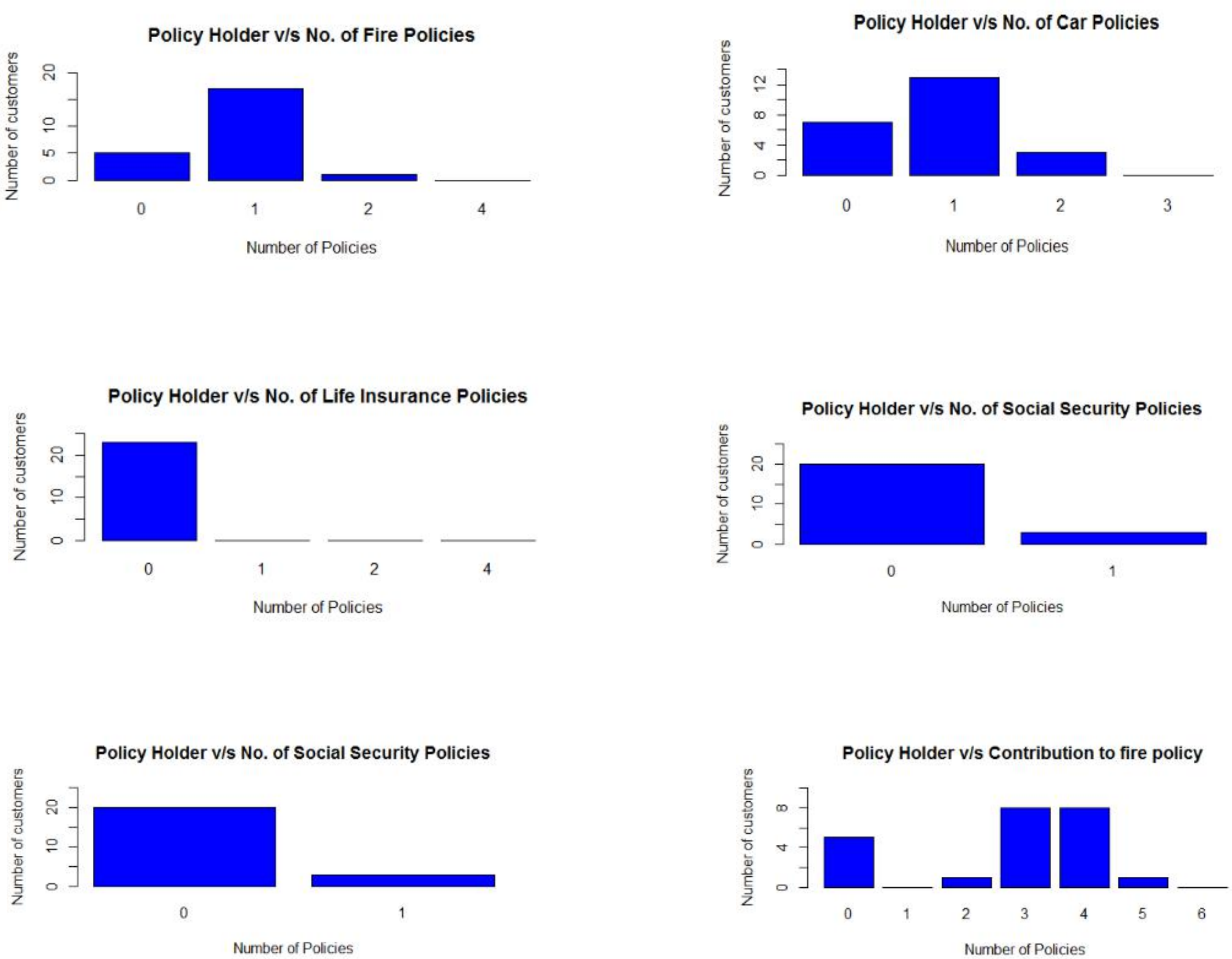

Fig: 4.3

The Distribution of the average age, Main Type of Customers and Subtype of Customers as been shown below:

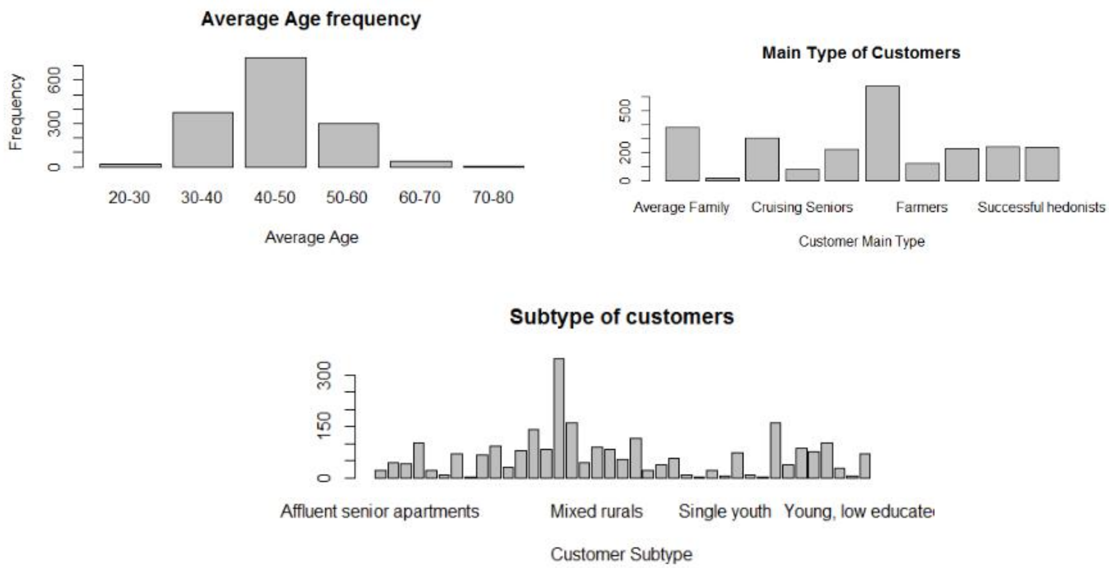

Fig: 4.4

The Regression Model:

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.05754209	0.31352952	0	*
M_EDU_HIGH	0.17362288	0.05885206	0.00317612	1.1896069
M_AVG_INCOME	0.17481972	0.07740599	0.02391586	1.19103146
PRIV_3RD_INS	0.4663696	0.17033134	0.00618115	1.59419608
CAR_INS	1.32716846	0.18058152	0	3.77035236

Residual df	691
Residual Dev.	847.1951904
% Success in training data	50
# Iterations used	8
Multiple R-squared	0.12195091

Table: 1

Training Data scoring – Summary Report:

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	248	100
Non-Buyer	131	217

Error Report			
Class	# Cases	# Errors	% Error
Buyer	348	100	28.74
Non-Buyer	348	131	37.64
Overall	696	231	33.19

Table: 2

Validation Data scoring – Summary Report:

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	155	83
Non-Buyer	1455	2307

Error Report			
Class	# Cases	# Errors	% Error
Buyer	238	83	34.87
Non-Buyer	3762	1455	38.68
Overall	4000	1538	38.45

Table: 3

Classification Tree:

Training Data Scoring – Summary Report (using full tree)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	218	130
Non-Buyer	96	252

Error Report			
Class	# Cases	# Errors	% Error
Buyer	348	130	37.36
Non-Buyer	348	96	27.59
Overall	696	226	32.47

Table: 4

Validation Data Scoring – Summary Report (using best pruned tree)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Buyer	Non-Buyer
Buyer	144	94
Non-Buyer	1191	2571

Error Report			
Class	# Cases	# Errors	% Error
Buyer	238	94	39.50
Non-Buyer	3762	1191	31.66
Overall	4000	1285	32.13

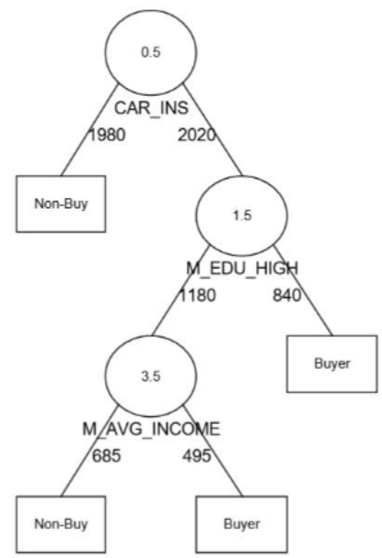

Table: 5

CHAPTER 5

CONSTRAINTS AND FUTURE ENHANCEMENT

1. PITFALLS:

- Significance & Sample Size: A key aspect of model construction is to select a good set of explanatory variables.
- Outliners: The modeller may throw out some unusually large input values or outcomes as “outliners”.
- Missing Data: Often there are cases when the dataset is incomplete. We know some of the attributes for these individuals but not all.
- Correlation, Causality and Hidden Variables: Statistical analysis on its own can only show whether an input is correlated to the output variable. This does not imply a casual relation.

2. FUTURE ENHANCEMENTS:

The following gives us hints for further action:

Area-Focused Marketing: Based on socio-demographic predictor such as M_EDU_HIGH and M_AVG_INCOME, marketers could place marketing campaigns in areas with high proportion of high educated people and high proportion of high average incomers.

Advertising Campaign: Based on customers’ past transactions, the insurance company could seek opportunities of cross selling by knowing who are likely to be potential customers.

Bundled Products: The Company can bundle the insurance with another one in order to attract more customers.

Joint Marketing: Marketers could join a few more people in launching promotion campaigns to focus target potential customers.

Data mining techniques are very useful to apply on insurance companies data. The regularity of a customer for instalment payment depends on certain important factors that the company stores which are obviously user specific and very sensitive. The source that provided us the data could not provide user specific information such as the actual income of the policy holder, health condition of the policy holder etc. which can be integrated in the attributes effecting classification of a customer. We intend to collect these user sensitive information which we believe will effect strongly in building a more specific and effective classifier in future.

CHAPTER 6

CONCLUSION

Our goal is to find a classifier that could effectively classify a non-regular customer from a regular customer of an insurance company. To do this initially we faced some problems in the pre-processing stage. To solve this we use first attribute selection methods to select the proper attributes that can have maximum effect on the classification. It proves to be very effective in action. We also use balancing algorithms on our data to balance the data. Without applying balancing techniques the classification is mostly favoured by the general class. But after balancing the results that we get are quite good. As the balancing is done maintaining the initial ratio, the result is equally applicable on original data set.

REFERENCES

1. Pujari, Arun K. *Data mining techniques*. Universities press, 2001.
2. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.
3. Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.
4. Elder, J., 2009. *Handbook of statistical analysis and data mining applications*. Academic Press.