Prediction of Responsive Companies Based On Consumer Data Complaints

Rishi Sai Reddy Sudireddy (810968740)

Sankar Nadendla (810938098)

Venkata Sai Kiran Kuchipudi (810986594)

Table Of Contents

1. Introduction:

1.1 Explosion of Data

More than 90% of the data in the world has been generated in last 5 years alone. There has been a lot of financial and consumer data made available in many of the data sets. Data scientists have been using this data for mining and coming out with productive results which enhance the current solutions to the existing problems. Generating useful prediction and rules from these data sets is called mining. Financial and consumer reports data sets have been on a rise of late . Because the market is so competitive that every company wants an in depth analysis of their sales and product structure. Not all data sets have useful information. However, selectively choosing the attributes and by applying the algorithms one can derive useful information out of the data. The major steps in any data mining project are:

i) Data preprocessing
ii) Algorithm selection
iii) Model evaluation
iv) Results and rules interpretation

The project used Weka software to analyze the data and apply algorithms. The tool has over 50 data preprocessing and regression algorithms respectively. All the process methodology works on GUI based interactive methods for all the cleansing and filtering of data. It also includes 8 clustering and 3 association algorithms built in. Some of the popular algorithms are KNN and Apriori. Data preprocessing refers to the cleansing of data .It includes removing missing values, labeling of data if need, Removing the outliers. The missing values can either be deleted or imputed. Deletion of records may lose the information gain of the attributes of the project. The data can become biased or random which is completely unpredictable. However imputation is the concept where the missing fields are replaced with the mean of the attribute or mapped to the nearest neighbor using K nearest neighbour algorithm. Algorithm selection is the next part where the appropriate algorithms like Naive Bayes/Apriori are applied depending on the data set available. Not all algorithms apply to any data sets. Depending on the type of attribute, and its nature, algorithms have to be applied. In the case of finding the relations between the attributes, Association algorithms like Apriori generate the rules with support and confidence. These rules can be helpful in shaping up the conclusions for financial services companies. Finally, the results that are generated have to visualized in order to figure out what all can be derived from the data. Not every result/outcome gives us an understanding of the problem stance. As said earlier attributes and records have to be fine tuned in order to achieve the results with highest information gain.

2. Problem Description

With the advent of a large number of financial institutions offering all types of financial products like Credit cards, Mortgage, Car loans, Student loans, Medical insurances, etc. Consumer reports can be highly misleading if prediction correctness is low because companies spend lots of money in the belief of improving the consumer confidence. So, banking on the right things to improve upon in crucial any data mining project. Reliability is an issue of high degree when

maintaining a product line since, there are lots of factors associated with customer service like company's 'Brand Value'. Improving the timely response for existing products can greatly enhance the customer satisfaction and improve the sales and revenue of a company. Also, It has become a difficult task for an average consumer to approach the right company for the right service and get a timely response in turn. It is also highly complex for some companies to analyze the products in which most consumers have raised disputes. So solving such kinds of issues requires the results to be highly accurate .By having in depth information about companies, response type and time, the company could provide good customer service and improve overall performance of system on both ends of customer and company.

3. Data And Preprocessing

Dataset for this project has been taken from *data.gov*. It is consumer complaint dataset from Bureau of Consumer Financial Protection. Initial data had lots of missing values and was available in CSV format. Therefore, some of the data cleansing was done on the excel. Then after, using explorer command in the Weka the dataset can be loaded into the program. In order to work with Weka the CSV shouldn't have any end of the line attachments . As lot of values of class attribute was same therefore, to avoid , building of biased classifier. The data set is preprocessed using filter "resample" which generate random sample of data with different values of attributes. so, that random sample could be choose for the training of algorithm. Secondly, the "add values" filter is used for adding the "N/A" for the missing value is the class. The original dataset has 19 attributes like date sent, Product, Sub-product, Issue, Sub-issue, Zip code, Tags, Consumer consent provided, Submitted Via, Consumer complaint narrative, Company public response, Company, State, Timely response, Consumer disputed, Complaint ID. However, in order to come up with precise rules some attributes are removed. Some of the work had to be done on MS-Excel even as the initial dataset was unable to get recognized by Weka as a valid CSV file. However after all the preprocessing , the final data set the project used has 7000 records with 5 attributes. which are Product, Company, Consumer Disputed, Timely response, Company response to consumer. As, removing every missing data field would not be a viable option and majority of the rows has some or the other field either missing or a duplicate. So, it retained some of missing values. Attribute information:

A) Product: It is a normal attribute type with 16 distinct values. It can be seen by selecting the attribute on the preprocess tab of Weka. It houses values like mortgage, credit reporting, Consumer loan, Credit card , etc.

B) Company: It has all the banks and financial institutions combined with 767 distinct values. Examples include Bank of America, Wells Fargo, Jp Morgan. All the values nominal in type.

C) Consumer disputed: This is an attribute where the values are only either Yes or No. Some not accountable records are as well present because there is difference between value being N/A and missing . Some people might not want to disclose their status despite having an issue.

D) Timely response: It is the quiet essential attribute for the project , as the research revolves around it. It has two stances as Yes /No. In further report the topic deals with an in depth analysis of how the attribute's value can be used for interpretation of results and make informed decisions for both incoming and existing customers and companies.

4. Proposed model

4.1 Implementation Of Classification Algorithms

4.1.1 Naive Bayes Classifier

Naive Bayes Classifier is derived from the concept of bayes theorem with random assumptions. In the mathematical perspective, Naïve Bayes is all about finding the probability of a class for a given instance. Therefore, this algorithm is going to help in finding the overall responsive companies in the dataset. In order to achieve this result "Timely Response" is considered as a class for Naive Bayes classifier. The formula for finding the probability of a class for particular instance is as follows:

$$Pr[Event/Evidence] = \frac{Pr[Evidence/Event]\ Pr[Event]}{Pr[Evidence]}$$

Where,
Pr[Event/Evidence] is Probability of a particular Event after seeing the evidence.

Pr[Evidence/Event] is Probability of a particular Event before seeing the evidence.

Pr[Event] is Probability of particular Event.

Pr[Evidence] is Probability of particular Evidence

After performing Naive Bayes on given dataset in Weka by taking class as "Timely response" the following results were perceived:

Correctly Classified instances are 96.84%
Incorrectly Classified Instances are 3.15%
Kappa Statistic generated is 0.43
Mean Absolute Error generated is 0.06

Table I. Confusion Matrix for Timely Response Class

Classified as ==>	**a**	**b**
a	2274	18
b	58	31

Here in Table I, a= Yes and b= No.
As per the above results, most of the companies have timely responsive to different consumer complaints on various products. However, sometimes the response might not be satisfactory for

the consumer. So, this may lead to company being disputed by the consumer. Therefore,this paper finds the probability of consumer disputes from the dataset by considering the "Consumer Disputed" as Class for the algorithm.
The Results generated in weka are as follows:

Correctly Classified instances are 80.67%
Incorrectly Classified Instances are 19.32%
Kappa Statistic generated is 0.03.
Mean Absolute Error generated is 0.19

Table II. Confusion Matrix for "Consumer disputed" class

Classified as =>	**a**	**b**	**c**
a	1906	15	0
b	427	4	0
c	18	0	10

Where a= No, b=Yes, c= N/A

As per above results, Some of the consumers have disputed with the response given by the company for the particular product. This helped us to evaluate more on the class consumer disputed and find out the companies which have received disputes. To achieve this let project used the visualization tool in Weka. It is useful in visualizing a particular class and also provides a better understanding of the various instances of the class .And, If analyzed the product which has got more consumer disputes for the response given the company using the visualization tool in weka.

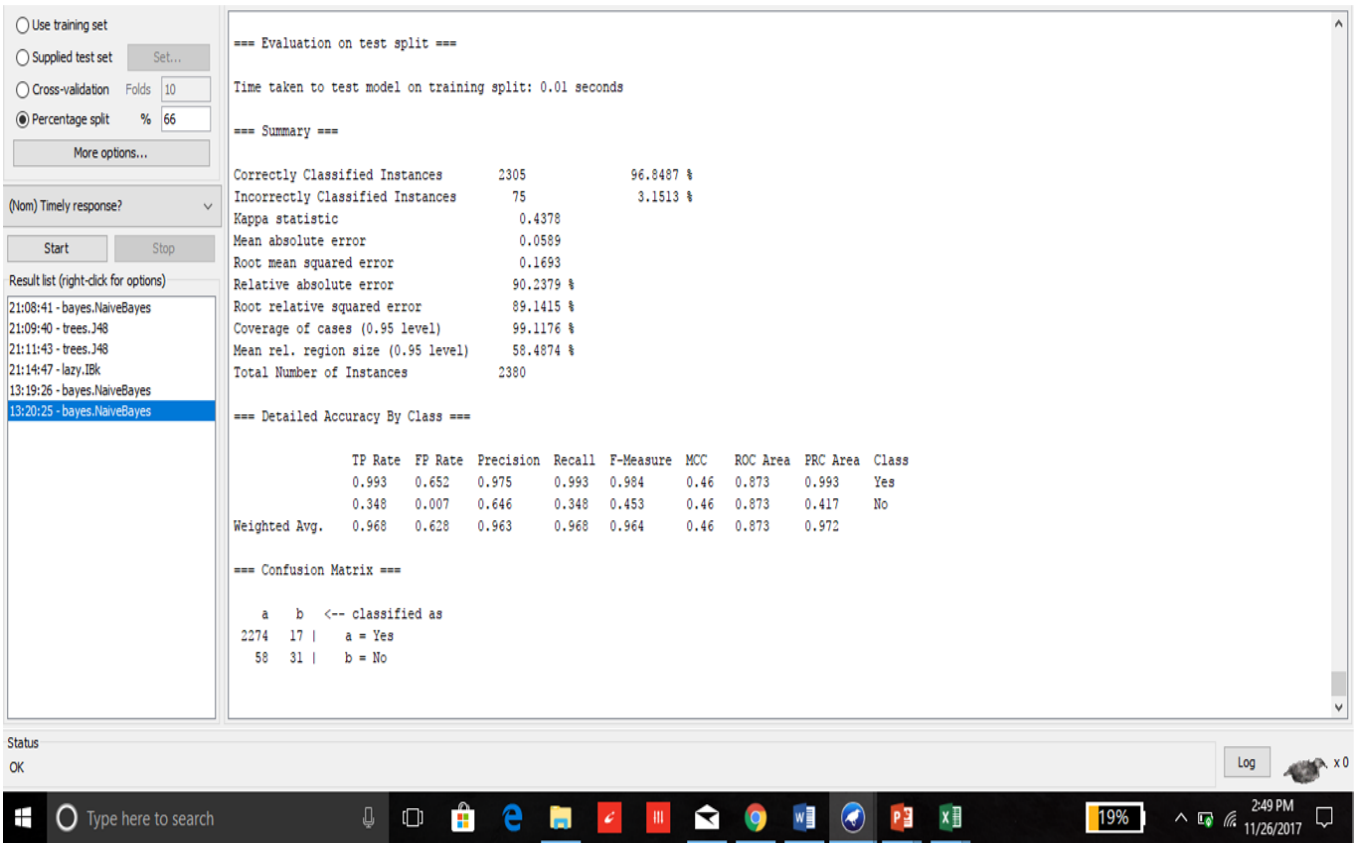

Fig 4(i)

5. Implementation:

5.1 Visualization:

In visualizing the consumer disputed class, The following analysis is made.By plotting a graph by considering product on x-axis and timely response on y-axis.Product attribute has different attribute values and timely response attribute has two values

i.e. yes and no. After observing the graph in visualization tool it was difficult to analyze the instances of consumer disputed as they are so small. Therefore, this research concentrated on increasing the jitter value so that we can get clear vision of the graph and makes it easier to get good results out of the tool.

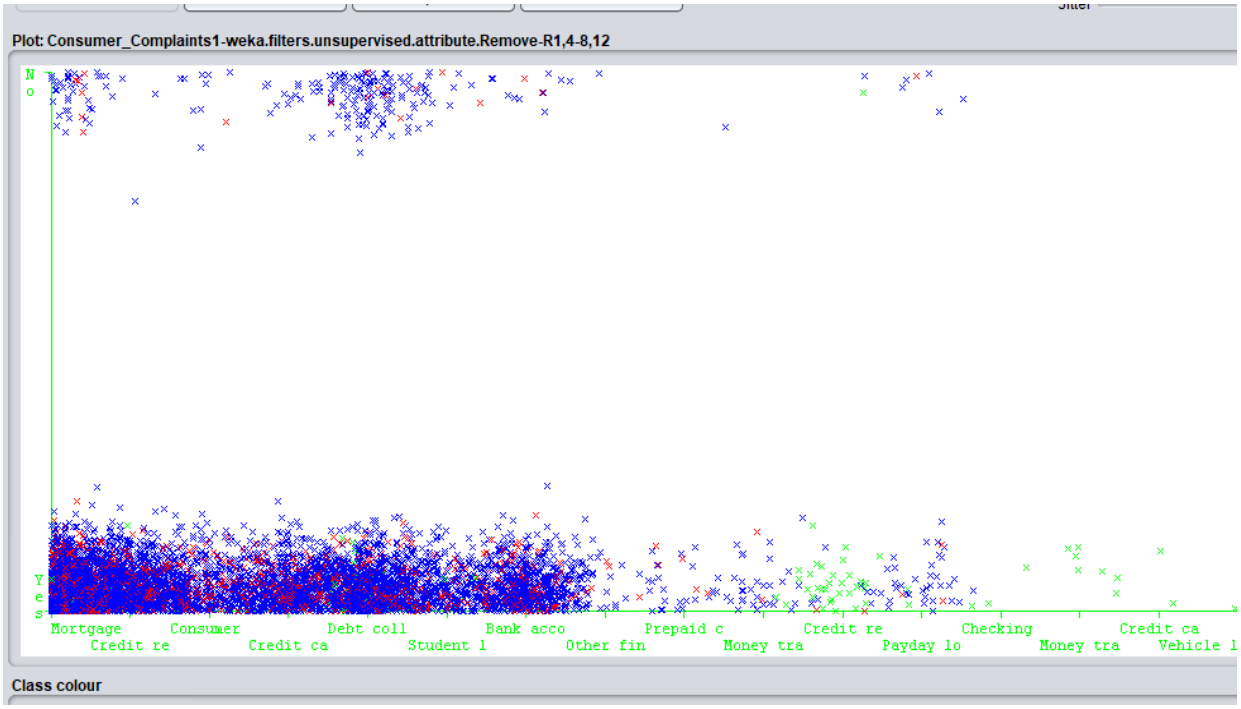

Fig 5(a) Visualization graph for Product vs Timely Response

After observing the graph in Fig5(a), the consumer disputed class has three values: Yes, No and N/A. They are represented with different coloured points which are as follows:
Red points are for the value “Yes”
Blue Points are for the value “No”
Green Points are for the value “N/A”
In Most of the cases the consumer disputed has either a yes or a no. This research concentrated more on the value yes.By observing that particular value we found that for the product mortgage has got more accumulated red dots. Therefore, the product mortgage has got more disputes for the responses given by the company. In generally, the consumers are disputed because they do not want to pay back there money for this particular product to the company.Later, in this paper the association rule mining will help us to learn more about the consumer disputes.

Additionally, let us visualize class consumer disputed by considering Company and Timely response

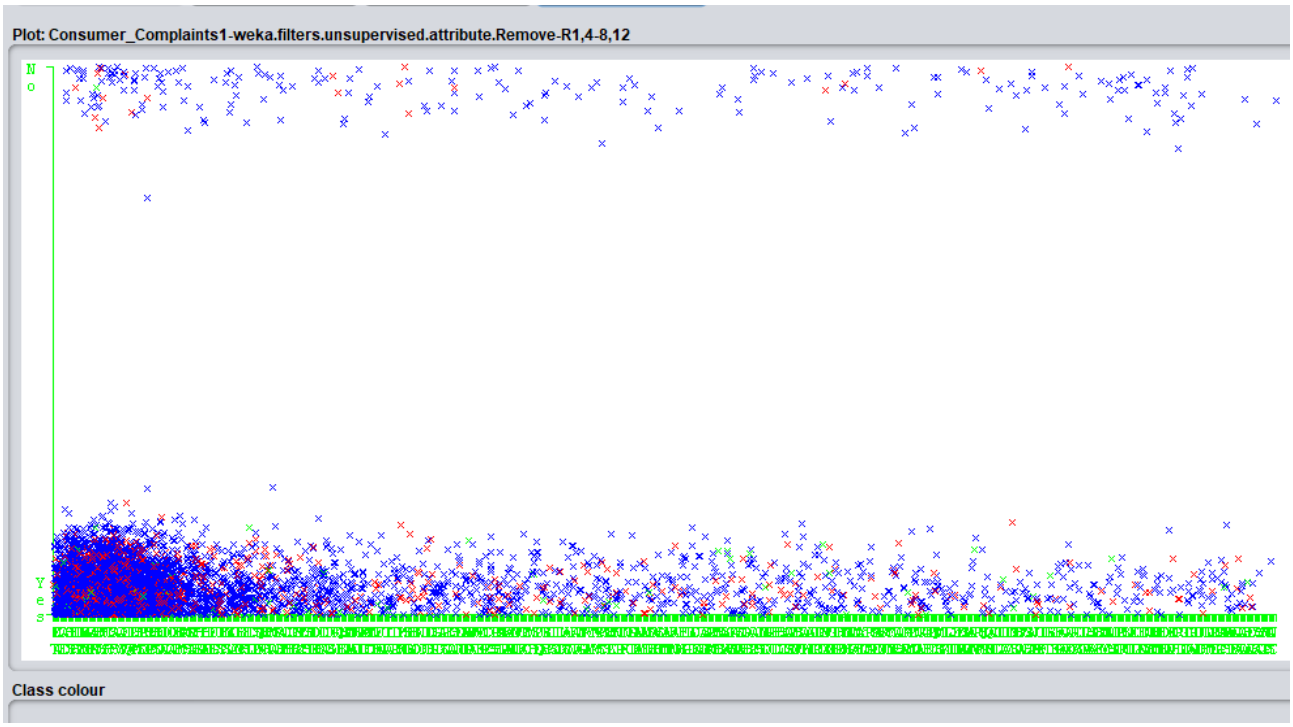

Fig2. Visualization graph for Company versus Timely response

As per Graph Fig 5(b), There are some many red instances that we can see from the graph for various companies. Therefore, the companies have received some disputes from the consumers for various products. However, in the graph it is difficult to find which company is mostly disputed by the consumer because there are some attribute values for the companies it is creating ambiguity to choose the company. Later, in this paper association rule mining will clearly explain and solve this issue and finds us a solution.

Accuracy evaluation:
Case 1:
Class: timely response?
Accuracy is evaluated by sum of diagonals/ total no. of classification.
2305/2381= 0.96%
If response is “yes” and classified as “yes” then result is true positive.
If response is “no” and classified as “yes” then result is false positive
If response is “yes” and classified as “no” then result is false negative
If response is “no” and classified as “no” then result is true negative
Precision = TP/TP+FP = 0.97
Recall = TP/TP+FN = 0.99

Case 2:
Class: consumer disputed?
Accuracy is evaluated by sum of diagonals/ total no. of classification.
1920/2390=80.33%

If consumer is not disputed and classified as “not disputed” result is true positive for class A..
If consumer is not disputed and classified as “disputed” or “ N/A” result is false negative for class A.
If consumer is disputed or “N/A” and classified as “not disputed” then result is false positive for class A.
If consumer is disputed and classified as “disputed” result is true positive for class B.
If consumer is disputed and classified as “NO” or “N/A” result is false negative for class B.
If consumer is not disputed or “N/A” and classified as “yes” then result is false positive for class B.
If consumer record is not available and classified as “N/A” result is true positive for class C.
If consumer record is not available and classified as “disputed” or “not disputed” result is false negative for class C.
If consumer is not disputed or disputed and classified as “N/A” then result is false positive for class C.

Precision = TP/TP+FP=0.81
Recall = TP/TP+FN = 0.99

5.2 Bagging

Ensemble learning improves the rate of correctly classified instances. Ensemble learning is performed with the well-known method of bagging. It is used for classifying two attributes here namely:
(1) timely response?
(2) consumer disputed?

It is also called bootstrap aggregation of different training sets. Aggregation of the results are done with voting in the case of bagging. The naive bayes is used as the classifier in the bagging. Bagging is explained as following:

Step 1:
For every iteration select, n instances with replacement from training set
Step 2:
Build model on each of n training set and bag the result by voting/average

In dataset, more than 90% consumers got timely response from the financial service companies making the model quite stable. Therefore, Bagging doesn’t decrease the overall error of the model in the case of attribute ‘timely response’. However, in the class ‘consumer disputed’, it decreases the overall error of model by 0.0010.

Table III. Difference of ensemble learning and naïve bayes

	% Correctly classified	Mean absolute error
Naive bayes	80.67	0.1973
Bagging	80.98	0.1962

Table III shows the difference in the results of ensemble learning and naive bayes for class consumer disputed.

5.3 Implementation of association rule mining

This learning model is to find relation between five attributes (1) company (2) product (3) company response to customer (4) consumer disputed (5) timely response. Apriori algorithm is used to find the relation between these five attributes.

5.3.1 Apriori algorithm:

This generates different attribute-values pairs which is called itemsets. Algorithm count the support and the confidence for the rules. Where support and confidence are as follows:

(a) support: proportion of instances which satisfy a rule. Where rule can be between two or more fields of dataset.
(b) confidence: proportion of the instances, where two or more items are always together.

The strategy of finding the best relation between the different attributes is, to iteratively reduce support to get the desired rules with selected confidence. This is one of the basic algorithm for association rule mining. Apriori algorithm works with two factors i.e. support and confidence. However, the problem with support factor is, in case there are many attributes and each have posses many distinct values. Looking for high support with given confidence provides very less rules. On the other hand looking for less support provides high number of rules. Weka has default parameters to find the rules. In consumer complaints there is more than 7000 instances. Therefore, there is a need for tuning of the parameters to find the rules inorder to find the relation between these attributes. This algorithm works in cycles with the user-defined parameters i.e. upper-bound support= 0.1, lower- bound= 0.06 and confidence=1.0 respectively. Algorithm starts with looking for maximum support for the given confidence and decreases by delta 0.05 default (can also be user-defined) after every cycle. It chose the top 10 rules with max support for the given confidence. With these parameters, Largest itemset L(4) found. The most informative rules are: .

[product=credit-reporting=430==>Company=Experia Information Solutions Inc. 430 ==> Timely response=Yes 430 ⇒ consumer disputed=No 430 <conf:(1)> lift:(1.03) lev:(0) [13] conv:(13.95).]

As, dataset of consumer complaints is very large with 5 attributes and every attribute have lot of distinct values. The attribute (company) has 767 distinct values. Therefore, frequent item sets are generated for rule mining. With the parameters used above it was quite difficult to find the best rule out of dataset. In the second step, the algorithm looks for lesser support i.e. upper-bound=0.009, lower-bound=0.001 with same confidence. Largest itemset is L(5) with 167 rules however, top 10 rules are selected e.g. Product=Mortgage Company=WELLS FARGO & COMPANY Consumer disputed?=Yes 57 ==> Timely response?=Yes 57 <conf:(1)> lift:(1.03) lev:(0) [1] conv:(1.85).

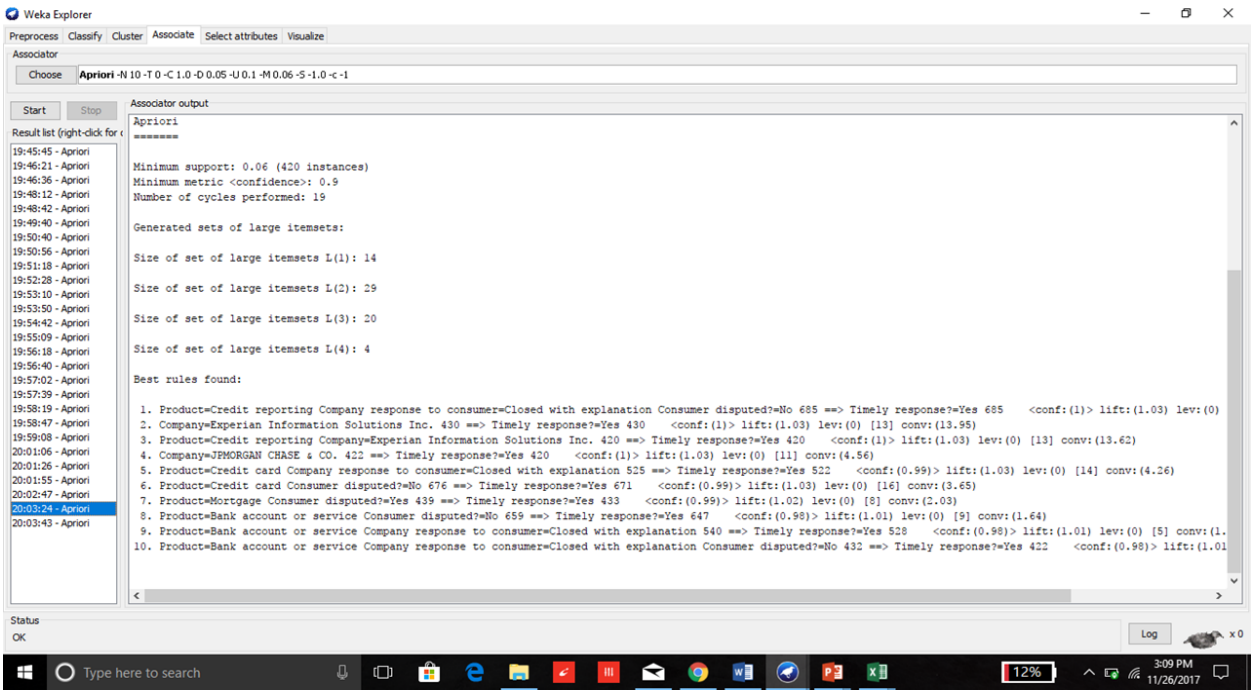

Fig 5(i)

TABLE IV. Results of Association Rule Mining

Support	Confidence	Rule Associated
0.06	0.9	Product= Credit-Reporting Company= Experia Information Solutions Inc. ==> Timely response=Yes
0	1	Product= Mortgage Company=WELLS FARGO & COMPANY Consumer disputed?=Yes 57 ==> Timely response?=Yes 57

6. Read me (Instructions)

(i) We have the data file which was collected in the form of Microsoft Excel.

(ii) By using the weka tool we have loaded our 2/3 part of the entire data to recognize the instructions how they were analyzed and the system or tool will understand it based on their classification to test the remaining data.

(iii) In the Next step, we have loaded our testing data (i.e. 1/3 part of the entire data) to test the accurate results.

(iv) In the same they can check for every algorithm which are inbuilt in the weka tool.

7. Conclusion

After observation of various results generated by different algorithms and methods, This project shows that most of the companies are timely responsive to the consumer complaints. However, there are some companies with which consumers have disputed. As per the experimental results, the company "Wells fargo & company" is has the most disputes for the product "mortgage" On the other hand, the company "Experian information solution inc" has the most timely response to the product "credit reporting" without creating any disputes.

8. References

[1] Torben Hansen, Ricky Wilke, Judith Zaichkowsky, (2010) "Managing consumer complaints: differences and similarities among heterogeneous retailers", International Journal of Retail & Distribution Management, Vol. 38 Issue: 1, pp.6-23.

[2] Eric W.T. Ngai, Vincent C.S. Heung, Y.H. Wong, Fanny K.Y. Chan, (2007) "Consumer complaint behaviour of Asians and non-Asians about hotel services: An empirical analysis", European Journal of Marketing, Vol. 41 Issue: 11/12, pp.1375-1391

[3] W. R. A. Fonseka *et al*., "Use of data warehousing to analyze customer complaint data of Consumer Financial Protection Bureau of United States of America," *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, Galle, 2016, pp. 1-6.
.

[4] Prashanth U. Nyer, (2000) "An investigation into whether complaining can cause increased consumer satisfaction", Journal of Consumer Marketing, Vol. 17 Issue: 1, pp.9-19.

[5] Chulmin Kim, Sounghie Kim, Subin Im, Changhoon Shin, (2003) "The effect of attitude and perception on consumer complaint intentions", Journal of Consumer Marketing, Vol. 20 Issue: 4, pp.352-371.

[6] Barlow, Janelle and Claus Moller (1996), A Complaint Is a Gift: Using Customer Feedback as a Strategic Tool. San Francisco: Berrett-Koehler.

[7] Ronald D. Anderson (2000), "A Bayesian Network Model of the Consumer Complaint Process," Journal of Service Research, 2 (4), 321-38.

[8] Andreasen, Alan R. (1988), "Consumer Complaints and Redress: What We Know and What We Don't Know," in The Frontier of Research in the Consumer Interest, E. Scott Maynes and the

ACCI Research Committee, eds. Columbia, MO: American Council on Consumer Interests, 675-722.