|  |
| --- |
| university of arizona |
| Telecom Market Advertising |
| Data Mining for Business Intelligence |
|  |
|  |
|  |

|  |
| --- |
|  |

The document is a report on the final project for the course Data Mining for Business Intelligence - MIS545. The document describes the project in detail and explains the various data mining techniques and methodologies used for the project. It also provides details about the dataset which was used for the project. Also, mining algorithm recommendations have been made based on the results.

Table of Contents

1. EXECUTIVE SUMMARY2
2. PREMISE2
3. PRoblem statement3
4. dataset description4
5. dataset cleansing & preprocessing4
6. attribute selection5
7. descriptive statistics6
8. algorithms used7
9. final algorithm recommendation11
10. interesting facts11
11. successful or not11
12. presentation comments12
13. conclusion12
14. Appendix……………………………………………………………………………………………………..14
15. Executive Summary

Telecom sector is one of the most competitive sectors in the world. To stay ahead of their competitors, companies try to take every step possible. One such step is to understand the needs of its customers. Our project aims to do that through data mining techniques.

We collected the telecom sector customers’ data. It had a number of attributes regarding the customers. Based upon this attribute we tried to understand the nature of the customers. With this understanding in place, we helped devise a robust and effective marketing strategy.

Our aim was to find out customers who use or do not use internet on their phones and classify them. We came up with a number of interesting facts like income, age and employment status are one of the most important attributes in classifying the customers. Income beinggreater than $45,000 annually is the first classifying criteria in classification. Thus these and many more interesting facts could be of use for the company to advertise and reach out to customers in a better manner.

1. Premise

In today’s world of cut throat competition, staying ahead of one’s competitor is a key to success. This is more so in the telecom industry because of the reasons mentioned below:

* **Low switching cost:** It is very easy and cheap for a consumer to change from one service to another without having to pay much. If the service is not up to the person’s expectation, he would easily switch to another.
* **Low barrier for new entrants:** New players can easily enter the telecom market as there is liberalization and entry barriers are low. This leads to exhaustive competition among the operating companies.
* **Existing economy:** The current economy is going through doldrums. It adds to the difficulty of operations.

# Keys to Success

In the adverse market situation mentioned above, there are certain things which a company can do to. Some of them are:

* **Understand the customer base:** Understanding the customers’ needs is the key to success. What problems they are facing with the service, what improvements they want and so on.
* **Develop “customer intimacy”:** This is somewhat similar to the first point with the only difference that here the company may want to customize its offerings in order to get closer to their customers
* **Advertise to the right people, at the right time and with the right offers:** There are three key features in this point: firstly, figuring out the right people, secondly, catch them when the demand is there and finally cater to their direct needs. If thesecan be achieved then one can lead a successful advertising campaign.
* **Be proactive:** Instead of waiting for the right opportunity, through data mining one can be proactive and create opportunities.

**DATA MINING:** A game changer. Our team believes that all the above and much more can be achieved through data mining.

1. Problem Statement

Through this project we would try to devise a marketing campaign by classifying the customers based upon their usage of internet on their phone. We try to help the telecom company find potential customers who do not use internet on their phone and thereby can be sold internet plans.

1. Dataset Description

The dataset is about the telecom sector customers. It contains a number of attributes about the customers. Some of the main points of the dataset are given below:

* There were 887 records in total
* There were 23 attributes in all
* Some of the important attributes were:
  + **Age:** Age of the customers
  + **Income:** Annual Income
  + **Marital Status:** Married of Single
  + **Sex:** Male or Female
  + **Bank\_Account:** Has a bank account or not
  + **Ethnicity:** Race
  + **Education:** Education level
  + **Cell\_Count:** No of cells owned by the customer
  + **Cell\_Type:** Uses smartphone or not
  + **Use\_Phone\_for:** Uses phone for just calls or more
  + **Home\_Phone:** Has a home phone or not
  + **Chat\_Cell:** Does chatting on phone or not

1. Data Cleansing and Preprocessing

This is the first and foremost step in data mining process. We need to cleanse the data and pre-process it so that it be used easily for various mining and analysis tools. We did the following:

**Dealing with missing values:** Attributes like income and age had some missing values. Since these were continuous variables, we replaced them with the mean value of the entire dataset. Also Employment status and ethnicity had some missing values. They were also replaced by mode of the entire dataset as they were categorical variable. For example ethnicity was replaced by white as that was the mode value for the entire dataset.

**Dealing with Skewness:** The int\_use attribute was highly skewed in favor of people using skewness. There were 661 records of people who used internet on their phone while 226 who did not use internet on their phone. To tackle this we took 250 records of value 1. This was an incorrect approach as was pointed out during the presentation by our professor. We touch more upon this in the “Presentation Comments” section.

We also removed columns which did not have aid in our analysis like transaction id, billing details, family\_size. After this we carried forward with the attribute selection step. Also we converted the .xls format to .csv so that it could be loaded into Weka.

1. Attribute Selection

Once we cleansed and pre-processed data, we performed Attribute Selection so as to select those attributes which play the most significant role in our analysis and the ones which are the key contributors to the results that we desire.

## Independent variable

**Int\_Use:** This variable told whether the customer used internet from his phone or not. Its value was Yes and No. For the purpose or our analysis we had converted them to 1 and 2. 1 meant people used internet on their phone while 2 meant they did not use internet on their phone.

## Dependent variable

We usedCfsSubsetEvalfor the purpose of attribute selection.Using this method, we chose the three most dependent attributes. They are:

* **Age:** This told us about the age of the customer. It was a continuous number. For the purpose of our analysis we had created age\_bins variable. For all those people who were aged between 10-19 were coded as 1, 20 – 29 were coded as 2 and so on.
* **Income:** Just as age, income\_bin was created. So people with annual income between $0 and $10,000 were coded as 1, $10,001-$20,000 were coded as 2 and so on.
* **Employment Status:** It was a categorical variable coded as Yes and No for employed and not employed respectively. We converted these to 1 and 2 for the purpose of our analysis.

1. Descriptive Statistics

We performed descriptive analysis on the dataset and following are a few observations that we made once we completed the descriptive analysis:

* **Age:** Most people were in the age group of 50-59 who were customers of the company. Next biggest chunk came in 40-49 range.
* **Ethnicity:** White group dominated the dataset followed by Hispanics
* **Gender:** There was not much difference as to the number of people with different genders. They were almost equal in number.
* **Int\_Use:** Among the internet users, females dominated males. Also people in the age group of 40-45 were extensive users of internet.

1. Algorithms Used

For the purpose of classification, we used three different classification methods.

Decision rules, Decision Trees and Logistic Regression. Let us take a look at each one of them, the reason for choosing them and finally how they fared.

## Decision Rule

We used decision rule classifier. It creates rules by first building a decision tree, then converting the “best” leaves into rules.

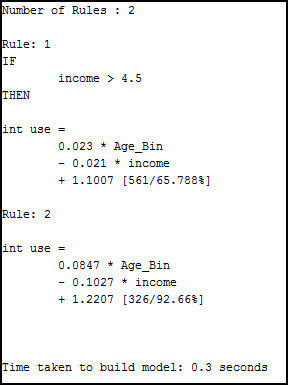
### Reasons for Choosing:

* **Interpretability:** Decision rules output are easy to interpret. One can easily put in the values and of the variables to check what the final classification is.
* **Speed:** Speed refers to the time it takes to build the model. For the purpose of our analysis, decision rules fared well in terms of building the model.
* **Scalability:** If the dataset is scaled up, decisions rule can still be used.

### Algorithm Performance:

The algorithm fared well. The model took 0.3 seconds to build. The results were also easily interpretable and accuracy was good at 64%.

### Output Screenshot:



**Explanation of the Output:** Two rules were generated. Rule 1 means that when the income is above 4.5 or $45,000, then whether the person uses internet or not can be known by feeding in the Age\_Bin and income value for a particular record. For example age\_bin= 4 and income is 6. Then

int\_use= 0.023\*4 – 0.021\*6 + 1.1007 = 0.092 – 0.126 + 1.1007 = 1.0667. This value is as it is closer to 1 means that this customer uses internet on his phone as 1 is coded as people using internet.

## Decision Tree

The second classification method used was decision tree. In this each leaf node predicts the target variable. It bifurcate the variables on certain value and creates branches.

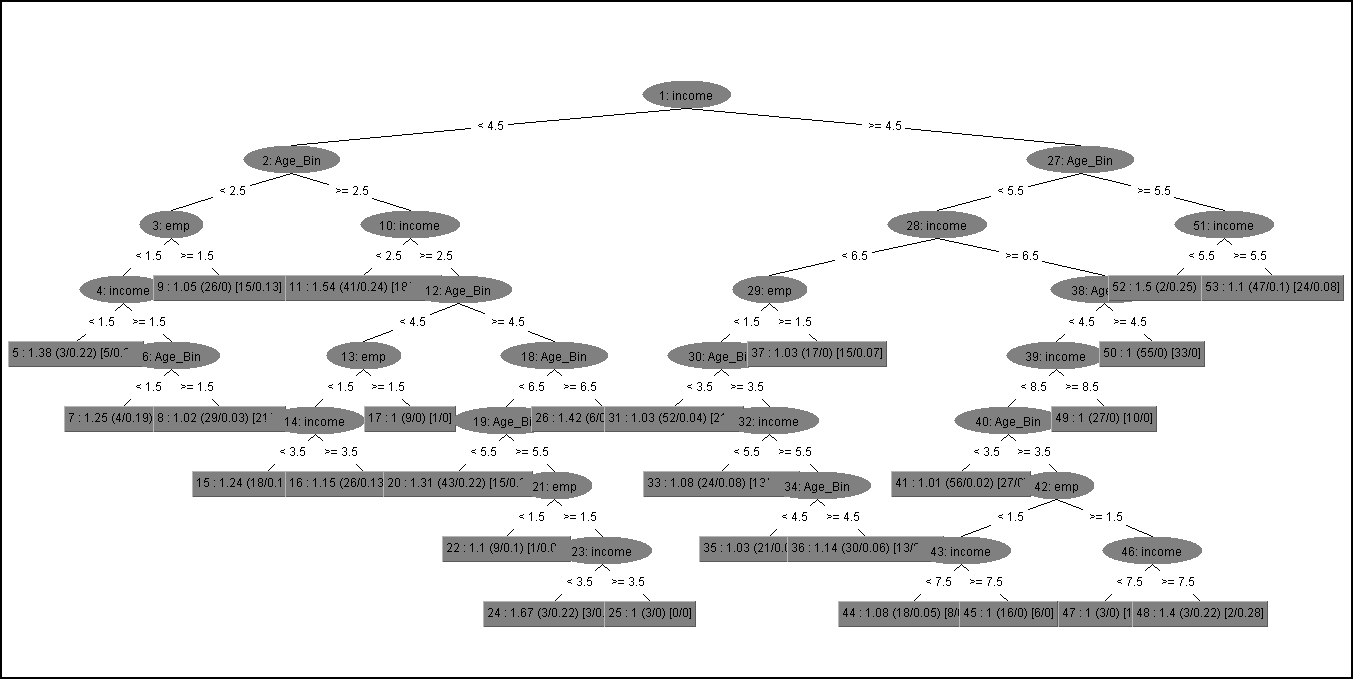
### Reasons for Choosing:

* **Interpretability:** Decision tree output is easy to interpret. However if the depth increases, interpreting the result becomes difficult.
* **Speed:** Speed refers to the time it takes to build the model. For the purpose of our analysis, decision tree fared well in terms of building the model.
* **Scalability:** If the dataset is scaled up, decisions treecannot be used effectively.

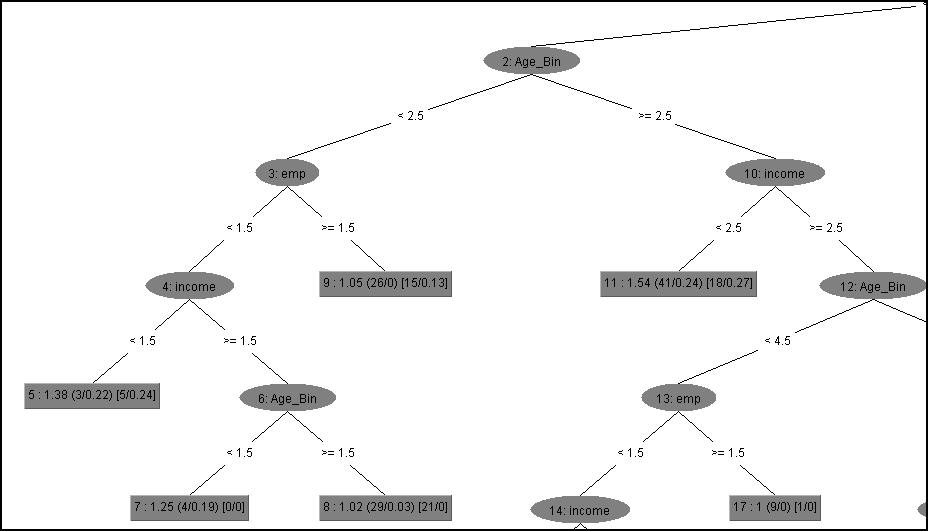
### Algorithm Performance:

The algorithm fared well. The model took 0.8 seconds to build. The results were also easily interpretable and accuracy was good at 46%.

### Output Screenshot:



### Zoomed in Version:



**Explanation of the Output:** We can decipher the tree by nothing a few important things. Firstly, the leaf nodes determine the class attributes or the target attributes. The nodes are variables and branches are created on the bases of some value of the variable. So, if we were to find out whether a person uses internet or not, we would look at their income and if it is greater than or less than $45,000, we move towards a branch. We keep repeating this till we reach the leaf node.

## Logistic Regression

Logistic Regression is the third classification model used. As we have binary categorical target variable, hence we are able to use logistic regression.

**Reasons for choosing:**

* **Interpretability:** Logistic regression output can be easy to interpret. One can easily put in the values and of the variables to check what the final classification is.
* **Speed:** Speed refers to the time it takes to build the model.
* **Scalability:** If the dataset is scaled up, logistic regression can still be used.

### Algorithm Performance:

The algorithm did not fare well. The results were however easily interpretable but accuracy was not good and it could manage only 43%.

### Output Screenshot:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | Age\_Bin | .382 | .052 | 53.118 | 1 | .000 | 1.466 |
| emp | .324 | .065 | 24.849 | 1 | .000 | 1.383 |
| income | .001 | .002 | .290 | 1 | .590 | 1.001 |
| Constant | -3.955 | .364 | 117.831 | 1 | .000 |  |
| 1. Final Algorithm Recommendation   **Our final recommendation is Decision rules based on four yardsticks**:   * **Interpretable:** Decision rules output was most easily interpretable * **Speed:** Decision rules model was built very quickly, 0.3 seconds to be precise, as can be seen in the output above. This is particularly of value as we had only some 800 odd records. In a real world, this number runs into millions, and it is then that speed adds value. * **Scalability:** Decision rules are most scalable of the three algorithms used. * **Accuracy:**We achieved the best accuracy of 64% using decision rules  1. Interesting Facts   Income is the most important factor which determines whether the person uses internet on their phone or not. Also $45,000 is the differentiating value for both decision tree and decision rules.  Another interesting factor is that people of age group between 40 and 49 and having income between $70,000 and $80,000 make the most number of people using the internet.  People aged above 30 and having income between $60,000 and $90,000 are the potential candidates who if not using internet can be sold the plans.   1. Successful or Not   Yes, the analysis helped in answering the problem statement. We were successful in classifying the customer population into internet and not internet users. This is what was required. Based upon this the company can go ahead with their advertising campaign. | | | | | | | |

1. Presentation Comments

Mainly three comments were made:

1. **Tackling Skewness:** We had incorrectly tackled skewness. We had randomly selected 250 records for 1 type if int\_use attribute. The correct way is to under sample and over sample. We performed the same afterwards. However, we did not find any major change in the output results.
2. **Logistic Regression:** It was performed on the dataset but not included in the presentation. However we are including it here in our final report.
3. **Interesting Facts:** All such facts which have been obtained from classification have been incorporated in the final report as asked.
4. Conclusion

We would conclude by saying that in today’s competitive times, data mining has become very important in terms of providing organizations with proper knowledge and understanding of the market. It empowers them with the information they need to constantly work towards improvement. It provides them with the capability to stay proactive and stay ahead of the competition. We realized the importance and the impact that data mining can have on an organization’s success more and more as we progressed through various stages of the project and that organizations should try their best to capitalize on whatever data mining has to offer.

1. Appendix

23:09:03: Weka Explorer

23:09:03: (c) 1999-2013 The University of Waikato, Hamilton, New Zealand

23:09:03: web: http://www.cs.waikato.ac.nz/~ml/weka/

23:09:03: Started on Sunday, 5 May 2013

23:09:33: Started weka.attributeSelection.CfsSubsetEval

23:09:33: Command: weka.attributeSelection.CfsSubsetEval -s "weka.attributeSelection.BestFirst -D 1 -N 5"

23:09:33: Filter command: weka.filters.supervised.attribute.AttributeSelection -E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5"

23:09:33: Meta-classifier command: weka.classifiers.meta.AttributeSelectedClassifier -E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

23:09:33: Finished weka.attributeSelection.CfsSubsetEval weka.attributeSelection.BestFirst

23:15:37: Started weka.classifiers.trees.REPTree

23:15:37: Command:weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1

23:15:38: Finished weka.classifiers.trees.REPTree

23:16:10: Started weka.classifiers.rules.M5Rules

23:16:10: Command:weka.classifiers.rules.M5Rules -M 4.0

23:16:15: Finished weka.classifiers.rules.M5Rules