Data Analytics Project Proposal

Priyanka Reddy

Southern New Hampshire university

This Data Analytics Project based on a online retailer company where they were launching new products every term. This project describes how Richard who works in the company analyze data of the sales and increase efficiency of getting profit to the company.

**Background:**

Richard works for ecommerce company. This organization is planning to launch a next generation eReader and the organization is also planning to increase the effectiveness of their marketing. This organization have many customers. Richard noticed that there are categories of people who are interested in old products, who are interested in coming products. He wants to know why people were motivated to buy as soon as the gadgets are out and why some people are prefer to buy the old products.

The organization additionally offers a great many different sorts of media, for example, customary printed books and hardware of each kind. Richard trusts that by mining the clients' information in regards to general purchaser practices on the site, he'll have the capacity to make sense of which clients will purchase the new gadgets early, which ones will purchase next, and

which ones will purchase later on? Richard needs to have the capacity to foresee the planning of purchasing practices, yet he likewise needs to see how his clients' practices on his organization's

site demonstrate the planning of their buy of the new eReader. So, based on diffusion theories by roger he came to know about the adopting the new technology tends to follow the “S” shaped curve with customers who purchase products at early stage followed by smaller groups of late adopters.

Comprehension Rogers' hypothesis, Richard trusts that he can order his organization's clients into one of four gatherings that will inevitably purchase the new eReader: Innovators, Early Adopters, Early Majority or Late Majority. Based on analysis using roger’s theory, Richard figured out to use data mining which helps him to predict which category of customers falls into what kind of adopters.

**Data Sources, Needs & Analytics Initiative:**

Decision tree model was preferred to find good estimators of what customers want to purchase. The Ecommerce website will have huge set of data on their own and no need to depend on third parties to gather data. So, every customer who login on their website have their data, with what they are looking, what they have purchased, when they have purchased. So now they made two sets of data:

The first data is training data which contains the activates of the users who bought the previous edition of gadgets and the timings of their purchases. The another set of data which consists of attributes of the present users who are going to purchase the new gadgets as soon as they are into the market.

He plans to make sense of which class of adopter every individual in the scoring informational collection will fall into in light of the profiles and purchasing timing of those individuals in the preparing informational collection. In investigating his informational index, Richard has found that clients' action in the regions of computerized media furthermore, books, and their general movement with gadgets available to be purchased on his organization's site,

appear to have a part just the same as when a man purchases an eReader.

The attributes that were gathered by the set of data that have been gathered are as follows

User-ID, gender, Age, Marital-status, user-activity, products purchased based on months, years. Classifying the types of products, Payment method, whether they are using credit card, PayPal, EMI, checks. With the gathered data now the organization is going to prepare the data analysis using decision tree.

Using the gathered data now organization will increase their sales fore casting by targeting the customers based on the data they gathered. Optimize the marketing strategies with the data to target various clients based on their targeted market with more effective and personified messages. This data analysis can also help to retain customers in order to attract the new customers so that they can convert the initial product sales in to profits. With the decision modelling the organizations can identify which workers are well on the way to leave, permitting them to proactively guarantee that the individuals who are most profitable are fulfilled and suitably boosted. HR offices as of now have a great part of the information required to apply prescient examination in this specific situation, including level of pay, residency, execution level, participation records, and history of advancement. The data analysis can increase capacity utilization and can reduce the time in maintaining the equipment’s or machines and their maintenance. The analysis can help the organization to make better decision in order to when to release product which helps in identifying the growth opportunities of the organization business.

Choice trees are phenomenal prescient models when the objective trait is clear cut in

nature, and at the point when the informational collection is of blended sorts. choice trees are superior to anything more measurements based methodologies at dealing with qualities that

have absent or conflicting qualities that are not dealt with—choice trees will work around such

information and still create usable outcomes.

Decision trees consists of parent node, child node and leaves (associated by marked branch bolts), speaking to the best indicator qualities in an informational collection. These hubs and leaves prompt to certainty rates in view of the real qualities in the preparation informational collection, and can then be connected to correspondingly organized scoring information keeping in mind the end goal to create forecasts for the scoring perceptions.

**Goals:**

Decision Tree helps in creating model based on enormous data for a target variable. In this scenario, the goal is to make a decision tree based on the data that has been gathered by the group of customers who bought the previous version of e-reader and who are waiting for the new generation. The people behavior of buying products can be determined as the branches and the product buying time can determined as the leaves of the decision tree. As organization, main goal is to determine the time of buying the products by new customers, So the proposed decision tree model can depict and analyze the required goal. Depending on the buying time of old customers the decision tree can analyze the buying time of new customers. So, for any organization to be successful, customer satisfaction plays a key role. To get customer satisfaction, organization must follow the set of rules like always getting feedback from stake holders, customers, conduct meetings. Company must set timeline, Company objective, company goal and company budget goals. So, customer satisfaction and good relations with stakeholders help to achieve success. Decision tree analysis, in this scenario helps organization to follow marketing strategies like providing gift vouchers and coupon for certain group of people based on the analysis made by decision tree based on the buying time of the customers. Through this process and analysis if organization can achieve some sales i.e. if they achieve more sales than previous quarter then we can say that the organization is successful based on the strategies followed by decision tree analysis.

**DATA ANLYTICS LIFE CYCLE**

Life cycle allows organizations to follow a procedure to place themselves in the success track. Life cycle helps in analyzing what to do and when to do. And it also helps in completing the project successfully considering with customer satisfaction

Here I am discussing the 6 phases of life cycle which helps in best practices for project completion.

**Discovery Phase:**

In Discovery Phase first we need to get familiar with e-commerce business domain. Initially we need know the nature of data we are using, and look for similar kind of products and their marketing strategies. Make sure we have all available resources available for the people who work in the team. Resources in the sense, technology, hardware, project time and data. In this phase, we make sure the available set of data is appropriate and also look for any additional information available in order to make the project more better which helps company understand the organization performance.

**Data Preparation Phase:**

For data preparation, we need all the required software’s in order to work with data and perform statistical analysis and also for preparing the time duration of the project. “The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to

get data into the sandbox (DocPlayer, 2015) ”. So, to analyze the data and to work with data the data should be transformed to ETLT. The data like browsed history, purchased history should be kept in the columns which can be loaded into the XML file, so that this can be stored in the database and can extract whenever it is necessary.

**Modelling phase:**

In modelling phase the team describes the data flow and relation between the variables and targeted variables. The team decides the strategies, procedures, and work process it plans to take after for the consequent model building stage. The group investigates the information to find out about the relationships between factors and along these lines chooses key factors and the most appropriate models. Here we make conclusions from the browsing activity, and purchased activity. But we can’t analyze anything from the web activity login.

**Model Building phase:**

In model building phase, the project is in testing phase. In this phase the team execute the models that were proposed in the modelling phase. In this phase the team work on with work environment and determines if the work environment needs any other extra packages in order to work with data. Work environment includes hard ware and software.

**Communicate Results Phase:**

In this communicate results phase the team discuss with stake holders and partners regarding the information they gathered by analyzing the data from customers purchasing time, customers purchasing previous e-readers and people waiting for the next generation of e-readers.

**Operationalize Phase:**

In this operationalize phase the team discuss about the marketing strategies and training for the sales persons on how to pitch the product and deployment of the project happens in this stage.

**VALUE OF LIFE CYCLE:**

If the organization follow the life cycle, then the company is going to achieve its goals as it is going to process the work in step by step procedure where even if we do any mistakes in the project then it is easy to estimate the budget and also save the time. In projects if they move forward to the next step or phase and later if they find new requirements, they can go the intial phase where this process can help to build models. The modelling phase helps in getting the work environment set up and also allows to improve the work environment if necessary. Data Analytics life cycle helps in achieving good strategies and help in achieving good models and project can have good performance. Business user, project sponsor, project manager, database administrator, data engineer, data scientist plays a key role in the data analytics life cycle. Data Security is the main issue where every individual role worry about. The data project is held in every team member hand. So, teams should also consider the data security as main issue and they must worry about from malware threats or to steal data by others.

**Data Evaluation:**

With the gathered data in the case study, the data can be analyzed and can work with the data easily. Apart from browsing data, purchasing time, still it is not clear about the browsed media, and books that was searched by the customers. So, when the customer login with their account its not predictable of what exactly they are going to browse or buy. Instead of their searched goods it better to look for their browsed media so that we can find patterns or algorithms during buying time. So, data collected like gender, address, will not help in analyzing what exactly the customer is looking for. The payment method will also not help to analyze the purchasing data. So, getting the credit card details and analyzing the credit reports where we can get the sensitive information about the customer help analyze the strategies. Hence the Data security play a key role in analyzing the sensitive information. Any loss of data will lead to loss of trust and reputational damage for the organization. So, while gathering data, especially sensitive data, organizations have to make sure the data they are transferring or sending should be always encrypted and make sure the data is secured at every stage of life cycle and at every process.

**TOOLS APPLICABILITY TO INITIATIVE**

The tools used by Richard’s company to analyze the data that has been gathered from customers and information that customers have provided are been saved in database. So, this data has been taken out rom database and has been imported into the grid layout where only store manager has permissions to access this grid layout in the site. Now the data presented in the grid layout have been converted to excel sheet and has been sorted the data according. The excel sheet has many tools for organizing the data and sort the data and convert them in to statistical graphs. So, the excel has been a good analytical tool for analyzing the data. But when we want to add another set of data or any information in the excel sheet then all the sorted data has to be change and the formulas need to change in order to get the new analyzed data. Or lese we need to create another excel sheet to analyze the data. This may lead to bring errors into the analyzed data which again added budget for the company or organization. There is even chance of getting spyware or malware with the excel sheet. So, excel sheet is not preferred every time. When we want to transfer an excel file and if the excel file is big enough then we can’t send it through our mails, so we need to break the data into smaller pieces to transfer the data which is again a drawback where there can data loss while transferring the data files.

**TOOLS APPLICABILITY TO DATA**

Richard’s goals were to predict the purchasing time of products by new customers. So, for predicting the purchasing time Richard need to understand the old customers purchasing time and must analyze when did old customers bought the e-reader earliest as soon as they hit the market. And having data of previously collected data he must match with the new customer’s interest, preference of their purchasing time. The excel spread sheets has the feature of performing the what-if analysis. The changes in the values to check the changes that effect the whole values of data is the process of what-if analysis. There are few what-if analysis tools. Those are Scenarios, Goal seek and data tables. The scenarios and data tables are similar which get the data sets and provide the possible outcomes. Whereas the Goal seek analyze the output and provide the possible input values which leads to the analyzed output. In what-if analysis there is Solver add-in which is like goal-seek and we will be using this add-in for the present organization. Solver add-in can accommodate more variables. The Solver add-in utility can dissect the situations in basic leadership circumstances that include thought of qualities and imperatives for a few factors at the same time. This capable capacity utilizes different changing factors and imperatives to locate the ideal answer for take care of an issue. Let us take Richard’s scenario. We consider variable like no of customers bought the electronic media, no of customers bought electronic gadgets, and no of customers bought other products and using this data we can find the relation between person’s interest and when they are interested of buying products and what kind of products. But these variables aren’t enough to analyze the future buying time of new customers. So, excel enough is not a good option to analyze data in this case. And as the large set of data is loaded then excel sheets takes lot of time in order to function and produce the output and this slow can also produce errors in the output which is not recommended. So considering the previously analyzed on spread sheets Richard has to look into efficient data analytic tool which helps Richard to analyze the purchasing time of new customers with more accurately.

**RECOMMENDATION OF TOOLS**

After analyzing all scenarios i would like to suggest Richard to use Rapid Miner with R for analyzing the data. Using these tools Richard can analyze the data by using techniques like decision tree.

Decision tree is a graph model which has root node which we can consider as the new customers purchasing time, and for branches and leaf nodes we can consider as other variables which helps in analyzing the final purchasing time. The decision tress helps to analyze the data easily and quickly. When it comes to analyzing with analytical tool with the technique, here is what it goes, the root node which a label in rapid miner. Each node represents the input values of the scenario. The edges from tree which is equivalent to the values of inputs of the scenario. In analytic tool, we can provide the all the available data as input variables until all the instances have a same root value. we can use recursive order based on various inputs of clients until every one of the clients who purchased the past tablet in a similar period are gathered together which would help decide the quality estimations of the client's which impacts their purchasing time and advancement of the new tablet can be contrived in view of that.

Now we will analyze the analytic tool Rapid Miner and R for analyzing the data.

R has lot of tools which access the data base, undergo the process of all kinds of files, also saves the files from other statistical sources. To perform all these functions in R one has to get familiar with programming language in order to write commands to process the data. Whereas the Rapid Miner has tools like Excels, CSV, databases where we can use wizards for processing the data and establishing the statistical environment.

R and Rapid Miner are open sources with licensed packages and R supports multi-clusters for individual package and Rapid Miner also supports parallel execute add-in. Rapid Miner executes R scripts for transforming data into Statistical format so we can merger the Rapid Miner and R together.

So, for getting the desired output for Richards scenario I would recommend him to use both R and Rapid Miner tools for Analyzing the data to get the purchasing time of new customers.

**Data Analytics Value**

Data Analytics add value to the existing industries and help in getting more profits for the organization. Big data is practiced in many sectors like below:

**Banking and Securities**: In banking sectors the data analytics is used to monitor the financial market activities daily. So, this data analytics helps this industry to keep track of illegal activity happening in the financial market.

**Communications, Media & Entertainment:** In this sector data analytics is used to analyze the mobile and web users in real time.

**Healthcare Providers:** Data analysis is implemented in the spread of chronic disease and also to have a eye on patients and their problems.

**Education:** In this se tor data analysis is used to track the login time and logout time and over all progress of the students.

**Manufacturing & natural resources:** In this sector data analysis is used to enhance the supply chain capabilities which later helpful to increase the productivity.

**Government:** In this sector, big data is helpful for analyzing to detect and the patterns they use for the process to fasten the process.

**Insurance:** In this sector data analysis is used for customer insights and the products they are looking for.

**Retail & Wholesale trade:** In this sector data analysis is used for reducing fraud, optimizing the staff through data from customers and their shopping patterns. It is also useful to analyze the inventory on time.

**Transportation:** Data analysis is useful in traffic control, the route planning and to keep track of all transportation system.

**Energy & Utilities:** To track all meter reading using data analysis.

In this project, which I am currently working on works with data analysis for analyzing the marketing strategies to increase the sales of new e Reader. Since This is an e-commerce site so the organization might have gathered lot of data from customers. The data like customer behavior preferences, their age, their income, their purchase history e.t.c. By analyzing the obtained data, the organization can use the analysis to use them for improving their business. The ways where company can benefit from analyzing the data are as follows:

* By breaking down the informational index of clients who purchased past e Reader organization can get to know their inclinations and the variables that affected them to purchase. As in this informational index, there are different variables, decision tree would help examine all alternatives and think of a correct arrangement of special offers for each arrangement of clients.
* Churn and Loyalty models: utilizing information like Bought Electronics and Bought Digital media, organization can make sense of the more beneficial and more standard clients and energize them with dependability focuses or cashback offers.
* Discover the privilege new clients: Decision Tree when laid out with the inclinations and estimations of the prior tablet clients as hubs, their purchasing time and the favored items at clears out. The new arrangement of client’s information when mapped on to this structure of decision tree, their purchasing time and the items favored can be predicted.

Analyzing all these will benefit the organization to achieve its goal to fulfill and which in turn will leads to run a smart business with more efficient and higher profits and making their customers happy forever.

**Insights gained on the Data:**

With the choices of information accessibility and the potential information that could be brought from the Database, data model can be intended to improve the efficiency of marketing strategy and sales of the new e Reader. This Data model would particularly help the offers of the new e Reader by confining better advertising techniques. The methodologies would be based after anticipating when the new arrangement of clients will probably purchase the new e Reader, then encircling our advancements and rebates on the e Reader just to those arrangements of clients who will probably purchase in that period. If the client is searching for any kind of products of advanced media or computerized books frequently, then we could offer rebates on that only if they are purchasing the new e Reader. By implementing this process the execution is comprehend the past e Reader client's inclination and purchasing examples and mark their purchasing time and after that attempt to coordinate the new arrangement of client's purchasing design and the inclinations with the prior tablet one's and anticipate their purchasing time.

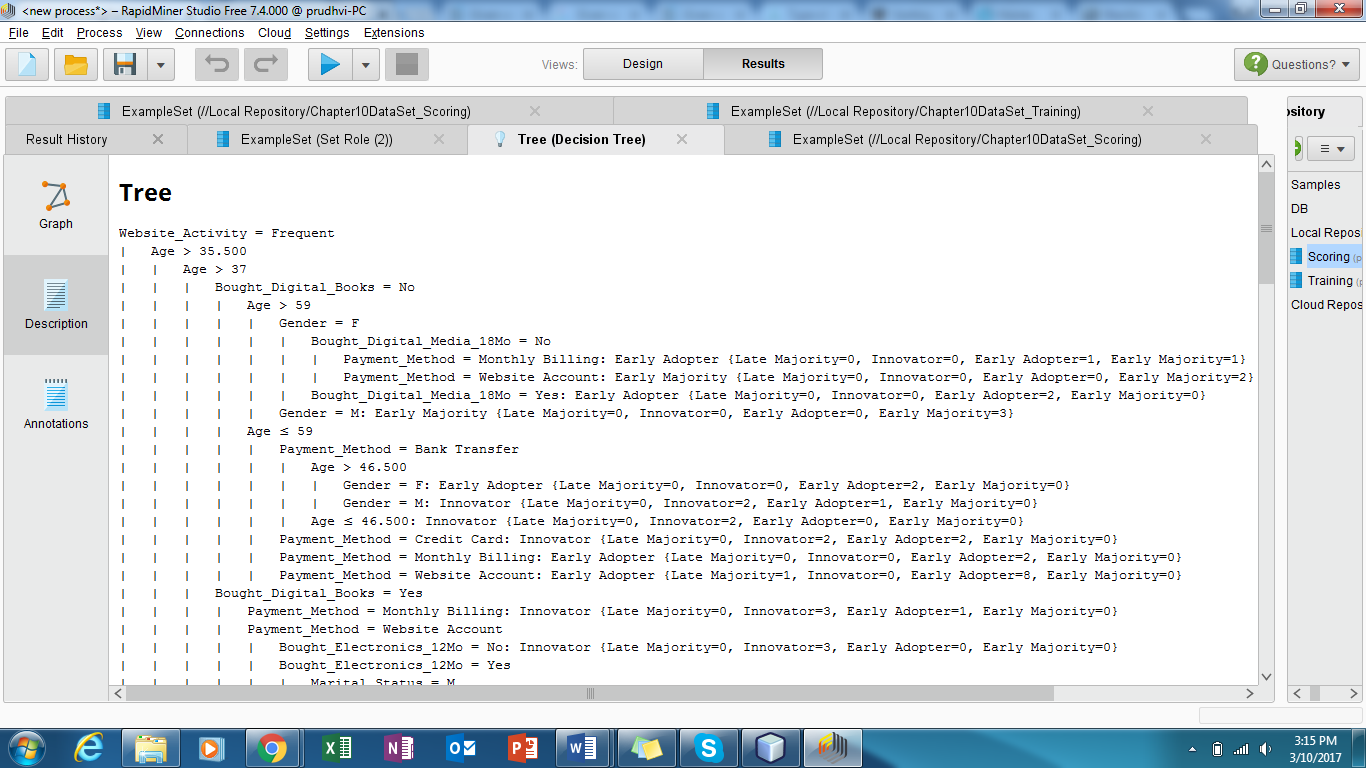
After investigation on different strategies like decision tree, Random Forest and Navies' Bayes hypothesis I found this would be effective in anticipating the purchasing time since this is the ideal informational index to execute decision tree as we have information on past possibilities on which we can make thorough evaluations of the probabilities. The Tools which we considered are R, Rapid Miner. After examination, we concluded to use Rapid Miner over different devices because of the complexity, multivariate dataset and inspired by getting and anticipating patterns as opposed to making point gauges Rapid Miner would be correct choice. Business can likewise break down information utilizing spreadsheet however for quicker and better basic leadership with enormous measure of information Rapid Miner would be ideal.

The experiences got from Decision tree would be certainly valuable to the organization in increase in the sales of e Reader, as organization would have the capacity to elevate e Reader to the clients when they will probably purchase with rebates and offers which would draw in their consideration. With these advancements and rebates it would build client maintenance complimenting the eventual fate of the organization.

Understanding the information and investigating the different strategies and devices as a feature of this project has given me tremendous introduction to data Science and Big data. Encourage a few hands-on in building Decision Tree utilizing Rapid Miner would help me to set up myself better in Big data and analyzing the data sets using various tools.

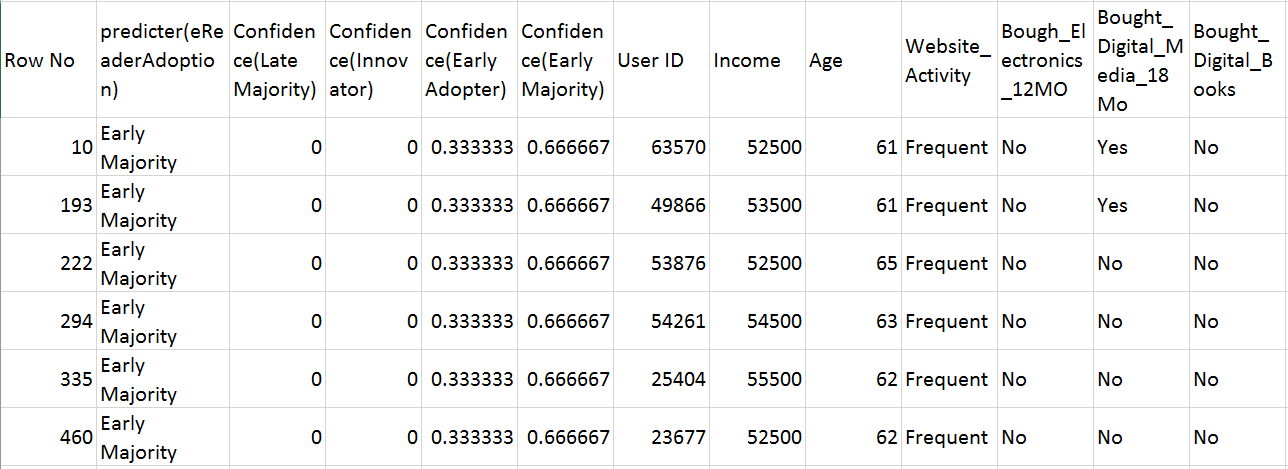
**Communication:**

The attributes which are considered for creating the decision tree are Age, Website activity, Income, bought digital books, bought digital media, which is provided in the data set which was given. So, our target is to get analyzed how many will be ready to buy e-reader, so we make e-reader adoption as the labels which means a leaf which is presented in the decision tree. Based on the training set data and the scoring data the e reader adoption label is allocated to the individual user-ID. The figure 1 shows the decision tree splitting with branches and leaves to get the desires e-reader adoption.



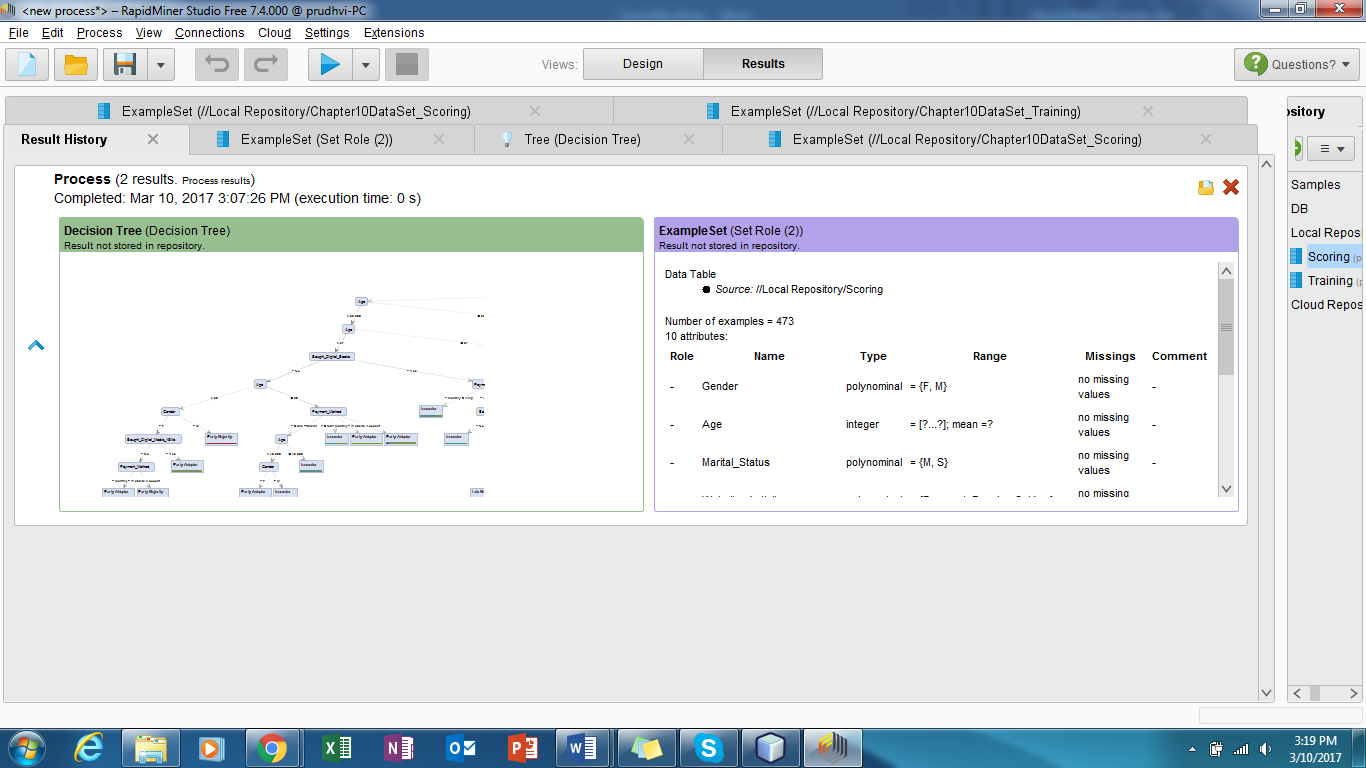
**Figure-1**

The below diagram shows how the decision tree has let us to make decisions based on the decision tree. Here is the final data of predictors of e-reader adoption, and confidence level. All this analysis is made on what the data sets we have provided. So, based on certain income level and certain age all these analysis has been made.

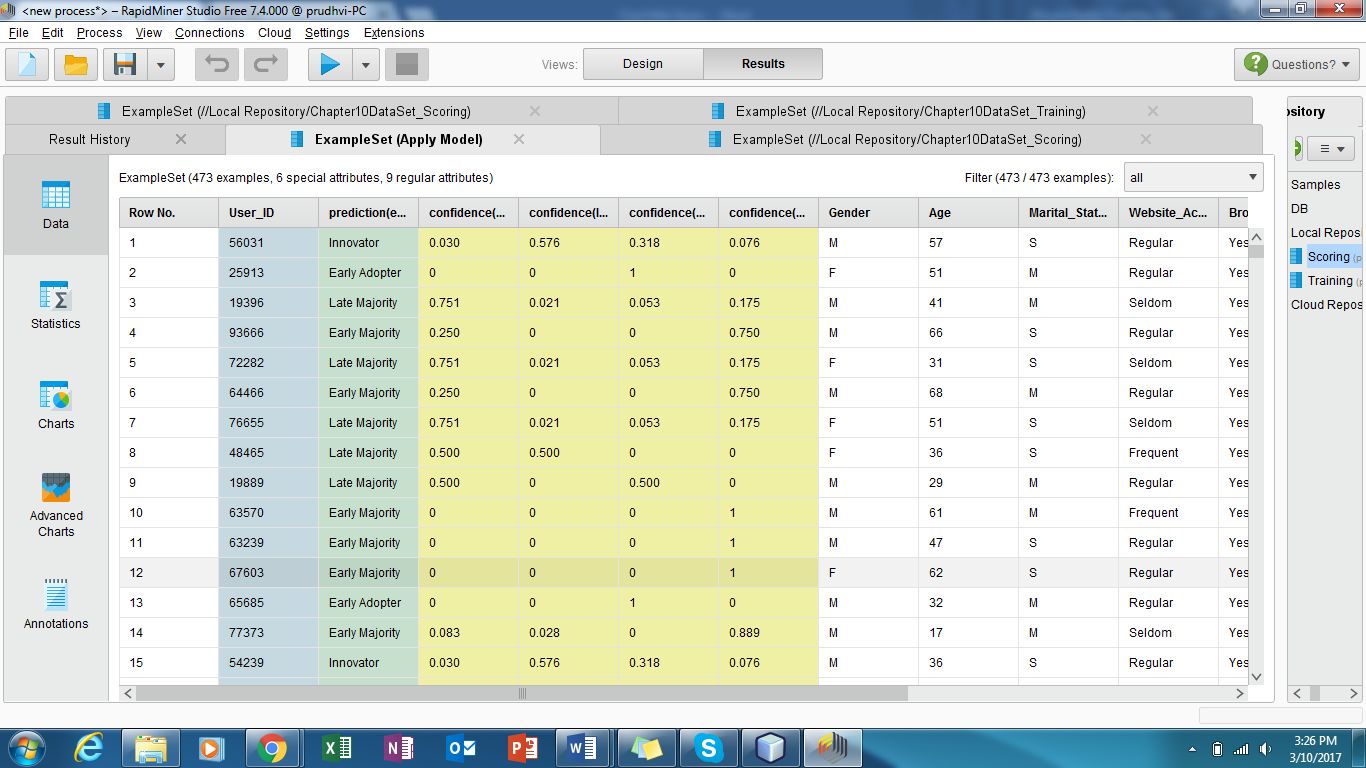


**Figure - 2**

The Whole decision tree model is represented in the below figure.



**Figure – 3**

Below figure represents the confidence values that has been predicted with e reader adoption value for every individual USER\_ID.

**Figure – 4**

**Conclusion:**

So, now based on all the analysis Richard and his organization can definitely benefit if they follow the same analysis by analyzing more data as they get and update the reference set by following the data analytics life cycle. In this case the decision tree method help Richard and his organization to analyze and achieved their targets and could predict when the buyers want to purchase the products. Rapid miner and R tools helped them too. In the same way, they can make many analysis and make sorting of data based on their customer preferences.

**REFERENCES:**

North, M. (no date) *Data mining for the masses*. Available at: https://dl.dropboxusercontent.com/u/31779972/DataMiningForTheMasses.pdf (Accessed: 6 February 2017).

Arseniev, V., Babenko, D., Technologies, H., Gaul, P. and Mayatt, A. (2011) *Big data: Big opportunities to create business value*. Available at: https://www.emc.com/microsites/cio/articles/big-data-big-opportunities/LCIA-BigData-Opportunities-Value.pdf (Accessed: 6 February 2017).

Allied (2014) *How to increase your company’s value with predictive Analytics*. Available at: http://www.alliedbizgroup.com/resources/publications/predictive-analytics-to-increase-company-value.html (Accessed: 6 February 2017).

DocPlayer (2015) *Chapter 2: Data Analytics life cycle - PDF*. Available at: http://docplayer.net/6882511-Chapter-2-data-analytics-life-cycle.html (Accessed: 20 February 2017).

Erl, T., Buhler, P. and Khattak, W. (2016) *Informit*. Available at: http://www.informit.com/articles/article.aspx?p=2473128&seqNum=11 (Accessed: 20 February 2017).

Smith, S. (2010) *The advantages of using spreadsheets*. Available at: http://yourbusiness.azcentral.com/advantages-using-spreadsheets-3647.html (Accessed: 26 February 2017).

Microsoft (2017) *Introduction to what-if analysis - excel*. Available at: https://support.office.com/en-us/article/Introduction-to-What-If-Analysis-22bffa5f-e891-4acc-bf7a-e4645c446fb4 (Accessed: 26 February 2017).

AnalyticBridge (2017) *Comparison on RapidMiner, SAS enterprise miner, R and orange*. Available at: http://www.analyticbridge.com/group/miningterabytesofdata/forum/topics/comparison-on-rapidminer-sas-enterprise-miner-r-and-orange (Accessed: 26 February 2017).

*How top 10 industries use big data applications* (2015) Available at: http://www.datascienceassn.org/content/how-top-10-industries-use-big-data-applications (Accessed: 5 March 2017).

McKinsey (2016) *How companies are using big data and analytics*. Available at: http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/how-companies-are-using-big-data-and-analytics (Accessed: 5 March 2017).