Loan Data Analysis Using Data Warehouse Techniques

Maduka Ashan Perera
Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology
Colombo, Sri Lanka
maduka.ashan@gmail.com

Chathura Madhuranga Boralugoda
Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology
Colombo, Sri Lanka
chathuraboralugoda@gmail.com

PPG Dinesh Asanka
Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology
Colombo, Sri Lanka
dineshasanka@gmail.com

***Abstract*— Most of the financial institutions are running their operations smoothly and profitable way without any interruptions with the help of data analytical techniques. This study will be able to enhance the business's ability to expand its market by providing meaningful and key analysis of consumer behavior. Financial institutions should have proper parameters to identify the right customer base with the capacity of their repayments. To identify those parameters, BI technologies and the data warehouse techniques such as inspecting, cleansing, transforming, and modeling were used to convert data to meaningful information. The star schema is used for this data warehouse design which includes one fact table surrounded by several dimensions. This study was mainly focused to identify the borrower's response to the calls taken by call center agents on a time basis per day. As a result, identified that several parameters such as age groups and gender-wise response times are different. Those factors will be evaluated by using a decision tree in future works. This will increase the loan collection efficiency.**

***Keywords*—Loan Repayment, Loan Arrears, Response Times, Decision Tree, Data Warehouse**

I. INTRODUCTION

Borrowing credit or loans from financial institutions by individuals, small organizations, or large organizations invest in self-employments. Individuals who borrow a certain portion of money at one time and repay that with certain steps with the realistic short or long time period. Most of the individuals borrow for education purposes or professional reasons with their repayment capacity calculating by themself with their monthly wages. Certain organizations in the country borrow the amount from the institution to invest in certain projects in which they are willing to expand their business or gain more profit on that. They are planning to repay that amount with the profit and income which they will achieve by the project they invest over the time period. Financial institutes such as banks, leasing companies, and other credit departments willing to earn their profits by funding or giving loans or credits to such organizations or individuals by taking risk of repayment over the period of time they agree to. This profit-earning cycle has risk factors and those factors should be calculate by the institution before they release the amount to borrowers.

Somehow over the repayment time period, some individuals or organizations show repayment rate deterioration and loan or credit recovery rate also reduced respectively [1]. The main problem is their criteria of accepting and rejecting borrowers with the small amount of analyzed data and mainly responsible for the bank contact. To identify, the legitimate borrowers have to change the screening criteria of the borrowers.

When the borrower applies for the loan or credit from a certain finance institute whether the loan or credit is accepted or rejected according to the screening criteria. After a certain time period, the accepted borrower receives his/her loan. This can be analyze further with the factors which we obtain by a decision tree. To accomplish the analysis by a decision tree, sample data was collected by certain financial institutions. To discover them, Microsoft BI tools were used to do certain operations such as inspecting, cleansing, transforming, and modeling to convert useful information.

A. Microsoft BI Tools.

SQL Server Integration Service (SSIS) tool which, Extract, Transform, and Load (ETL) data, which sample data were collected from the source systems to data warehouse designed to analyze data. This tool helps to do certain tasks such as cleansing, filtering, conversions, sorting, joining, validating, lookup, and aggregations. After these transform operations load data to the data warehouse, which includes Fact tables which contain measurers and Dimension tables which contain a set of detailed facts related to the business by surrounding the dimensions that elaborate the business details.

SQL Server Analytical Service (SSAS) tool processes the data. This tool is used by organizations to analyze data from multiple sources. SSAS helps to model multi-dimensional cubes which can measure sales amount, a quantity and so on. There are several methods such and drill down, roll up, slice, and dice to elaborate the date with the dimensions.

In order to illustrate such data, this study has utilized excel pivot tables and the dashboard for the managements using graphs. This is a framework. Framework in the sense, the user will be able to change and create their own reports or illustration visualization by themselves. This is called self-service. They will be able to change the parameters and analyze the data that we have provided under this framework.

This paper has used a decision tree to identify the categories which will affect the repayment, arrears amount, and the response of the borrowers using the data which gives output from the above tools.

II. Related Work

There are several studies conducted related to the fraud analysis of the financial institution by the borrowers or their customers. Loan or credit borrower's repayment capacity and recovery may or may not lead to the fraud.

SU-NAN WANG and JIAN-GANG YANG have conducted a study regarding the money laundering risk. Criminal activities, drug trafficking, bribing, smuggling are illegal activities that may lead to highly profitable areas. These will put through the cycle of a transaction before using them freely or otherwise those will become an illegitimate transactions. Putting through a cycle of transactions between banks and accounts that couldn't track back to the illegitimate sources. This shows as legitimate money to use them freely. That is called as money laundering.

China Anti-Money Laundering Monitoring and Analysis center received several reports of suspicious transactions. This includes more than one billion of foreign currency (US dollar) in 4926 accounts. The majority of the transactions reports came from the state-owned commercial bank in China. Therefore, Commercial Bank of China is facing money laundering.

The developed countries established advanced monitoring systems related to money laundering, such as the American Financial crime enforcement network Artificial Intelligent System (FAIS) [2]. Artificial intelligent systems greatly enhance the efficiency of the Anti-Money Laundering. However, the Anti-Money Laundering system was not developed in the china commercial bank. It is not recommended to apply directly developed countries AML systems because commercial banks faced many of their transactions reports captured regarding the foreign country. They need their very own pattern AML transaction system to be developed according to the Chinese financial market.

Numerous systems used decision tree [3] method to analyze the various studies due to the accuracy and the performance of it. To detect the Anti-Money Laundering they have been using the decision tree-based system. They have researched by selecting a midsize commercial bank in China with about 3 million customer accounts both individuals and companies. Their program is "Know your customer". This will detect suspicious transactions directly. When the customer opens an account, they have the risk of money laundering. The bank divide customers risk level as low, middle, and high. Initially they assign the risk rank for the customers as low when their transaction became high and with the amount of the money, they have begun to rate them as middle or high.

They have collected data of the 160,000 customers to the data warehouse to analyze them. They only selected 28 attributes of them due to the number of attributes. Among those attributes they only identified 4 of them as mostly related to the AML. And they change some values of the account numbers to show their result to cover the identity of the customers.

They have identified several risk factors, and they limited them to four inherited anti-money laundering risks of the customers as below.

1. Business and entity risk. What is the industry of the customer?
2. Location. Where the customer located?
3. Business size of the customer.
4. Products and transaction risk. Which kind of product or services is being offered to the customer?

Corresponding to the above attributes they have marked as low, middle, and high risk of the customer account regarding the Anti-Money Laundering.

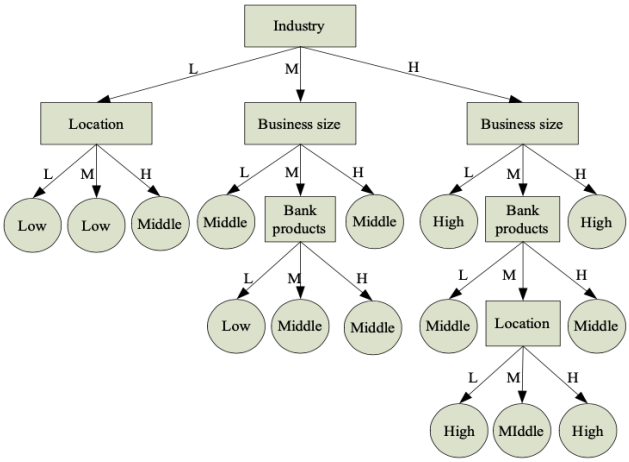

Figure 1. A decision tree for determination of customer's money laundering risk

JI Chengjun, WU Lijun, LI Jinping have conducted a study regarding the application of the decision tree analysis in the credit card. Since the 1990s of the last century, the Chines economy with the domestic needs they mostly person consumption for the credit option is very quickly developed. Commercial banks are the main source of taking credit risk by providing loan facility and credit to their customers, also their major source of income of the commercial bank is providing credit for the customers. Therefore, they have been looking for a method to evaluate their credit risk.

In china gradually opens several banks that will be major challenges for the banks that exist and open banks for their business due to the competition they face. They recognize with the studies credit card is the kind of way of non-cash transaction method to provide customers also they has been set up credit loans of the small amount from it as well. This method rapidly grows among the customers, it's spread over china rapidly. In the year 2003 according to the statistics they analyze 3,000,000 credit cards that have been issued. With the rapid issuing 50,000,000 cards had been issued at end of the 2006. Also, the 2008 china credit card issue quantity was 122,000,000. With the situation, the bank provided an overdraft facility over the credit cards that will cause fraud through the credit cards. That overdraft facility lends by the bank to increase their business and keep their business ongoing with the competitors. The risk that banks take with the fraud meantime some banks were bankrupt due to behaviors of escaping repaying their overdraft or credit amounts by their customers. The decision tree applies to identify the customer with past data they have been collected.

The decision tree starts with the root node which represents the sample data set. The second step or the second level is called a leaf, this level illuminates measure and the condition of the decision tree second level. All the attributes and discrete values are classified. The attributes of the values have to be discrete. In the third step branch and the sample, values will be divide into other branches. If the branch has no sample, That type contains the majority of the sample was created as a leaf. That defines the basic decision tree.

Select the index of the research paper of this. They have collected 21-word segments from the bank loan data set from the customers. Among those, they identify 20-word segments which are relevant to evaluate the customer financial credit information of the customers such as check account, credit history records, savings or deposit accounts, loan amounts, loan terms, loan purpose, controllable income, sex, working time, guarantors, address, age, property, telephone, registration, local or foreign. Finally, the bank divides them into two categories as "good customers" and "bad customers" [4].

The bank thinks the salary should be the root node for the decision tree according to their viewpoint [4].

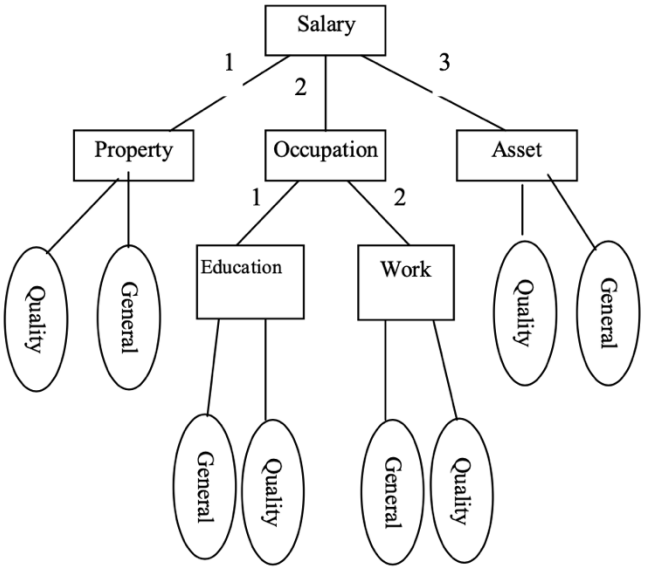

Figure 2. Annual salary of decision tree [4]

Statistical analysis on loan repayment and efficiency and its impacts on the borrowers, Yonas Shuke Kitawa, Nigatu Degu Terye conducted a study. Their major study is the factors affecting loan repayment efficiency and loan utilization in Ethiopia. They have mainly focused on the farmers and the individuals who carry microcredit as their loan facility to improve their small business or investments. In Ethiopia, they are provided a loan facility with a lesser interest rate. After some time they have faced some repayment rate deterioration. Their loan recovery rate was reduced by 64% to 31% from 1997 to 2000. Due to the issue, they have been facing they change in the screening criteria. When the client applies for a loan, the loan could be either accepted or rejected by the financial institute. The accepted loan will be received after a certain period of time to the applicant.

In this paper, they have used logical regression to predict the category of the loan borrowers that find the best fitting model for the individual case's relationship between response and explanatory variables. The Bayesian logistic regression method is used to predict the repayment efficiency of the borrowers.

As a problem, if there is high repayment efficiency, the relationship between the customer and the credit institution is high due to the help of the next higher amount obtained from the credit loan department or the financial institution. To change this, they considered some socio-economic factors. The main factors from the institution side are tight control, loan officer intensive, loan collection, the interest rate charged affect the repayment rate. The socio-economic factors from their customer side are marital state, gender, education, and income level. This conducted study has answered the following questions regarding the repayment factors at the end of the paper [6].

1. What are the main socio-economic factors?
2. What are the business and loan-related factors?
3. What are the major challenges faced by customers and the institute?

Their data collection method was a structured questionnaire. They have distributed the questioner among the people with the respective population. This includes social attributes, household characteristics, income, assets, financial characteristics such as credit and savings.

Logistic regression analysis is an extended technique of multiple regression analysis which is used to identify the categorical variables. They have considered the ratio of the probability of success, and this is the logistic model they used,

$$\frac{P(x_i)}{1-P(x_i)} = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}), i = 1,2,\cdots n \quad (1)$$

They have used Bayesian Logistic Regression to come up with the result. As a result, they came up with, out of 340 borrowers, 38.53% are efficient on repayment and 61.47% not efficient at the time of data collection on their study 11.8% for agricultural, 22.6% trades, 21% small enterprises, 10.9% general loans, 8.8% handcraft. With regards to the sex, they found out 38.2% female and 61.8% male borrowers [6].

III. Methodology

A. Data Warehouse Concept

A data warehouse is a comprehensive technology that helps key people to access any required level of information within the enterprise. It is an enterprise-wide framework that enables the management of all enterprise information.

Data warehouse consists of two different types of schemas, star and snowflake schema. This study has utilized star schema to develop analytical framework. The data warehouse consists of the Fact tables and dimension tables.

The fact table consists of measures which mainly address the business problem, process, and needs of the users, such as payment amounts, arrears amount, call count, inventory, sales. The fact table contains duplicate records for the analysis purpose.

A dimension table is a set of detailed business facts surrounded by many dimensions that describe those facts such as customer attributes, location attributes, product attributes, and so on.

B. Data Set

The sample data set was collected for the analysis from the financial institution of a particular country. Among the larger scale of data set factors were identified to analyze, such as profession, sex, age range, locations wise. The borrower's repayment amounts, arrears amount, arrears days for the particular month or quarter, response per call, and promise to pay amount particular customer range were derived from that source of data set.

Before loading data to the data warehouse, there was a workaround to accomplish the task such as cleansing data, conversion, filtering as per needs, sorting, joining, aggregating, and lookup.

C. Data Warehouse Design

As per the schemas mentioned above, star schema is used for this study data warehouse design which includes one fact table surrounded by several dimensions. Every fact points to one tuple of each dimension with additional attributes. Also, that does not capture hierarchies. In the star schema, dimension tables will not join each other and every dimension table joins with the special key called surrogate keys(SKs) with the fact table.

Surrogate keys are normally integer value, which is a unique value assigned to each row. SKs will do a very important role in the data warehouse, it helps to protect the data warehouse from unexpected administrative changes, and it helps to updates and inserts as well as tracking of the changes in the dimensions.

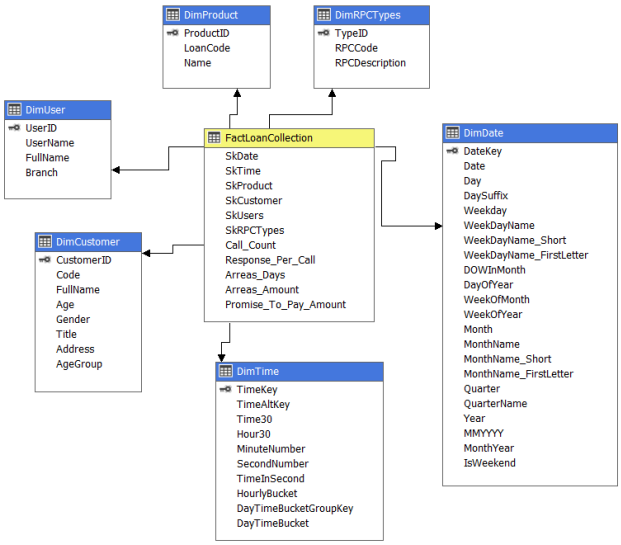

Figure 3. Loan Analyze Data Warehouse design

The data warehouse was designed and referred by this study which has one Fact table called "FactLoanCollection" with six dimension tables.

This data warehouse consists of dimensions DimCustomer, DimUser, DimRPCTypes, DimProduct, DimDate, and DimTime. Most of the data warehouse uses date dimension for the performance reasons and the functional reasons, that consist of the date hierarchy and other several attributes. This study contains two hierarchies for that with quarter and month name as well as numbers for the user's perspective. Apart from the date dimension, time dimension was added, which helps to analyze a data on an hourly basis, due to the sample data set that identified it is importance to analyze the data of the user's response time per the calls. This particular time dimension consists of the hierarchy call HourMinSec that contains hour, minute, and seconds. RPCType dimension contains two attributes that response per call type code and the description of that code. The product dimension consists of the borrower's loan types.

Furthermore, In Respect to the SCDs, This study has used Type 1 SCDs most of the dimensions

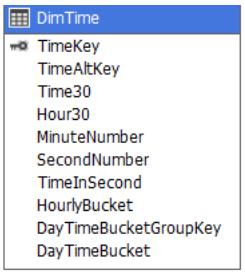

Figure 4. DimTime

The main focus of this analytical system is to check the actions of the borrowers and due to that FactLoanCollection fact table consists of five measures such as Call count, RPC (Response Per Call), PTP Amount (Promise To Pay), Arrears amount, and Arrears days.

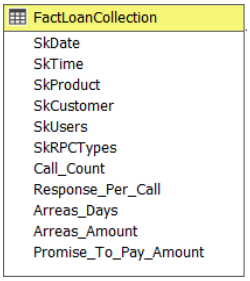

Figure 5. FactLoanCollection Fact Table

When considering this approach, it is similar to the Inmon approach which is known as "Father of data warehousing".

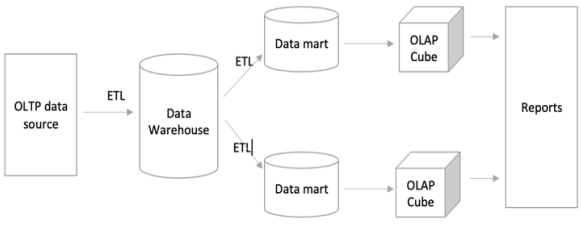

Figure 6. Inmon Approach

Similarly, this framework also consists of OLTP data source that loads data to the warehouse which is designed via ETL technology and process cube to browse dAata as multidimensional with several technologies that enable slicing, dicing, roll-up, drill-down, and pivot.

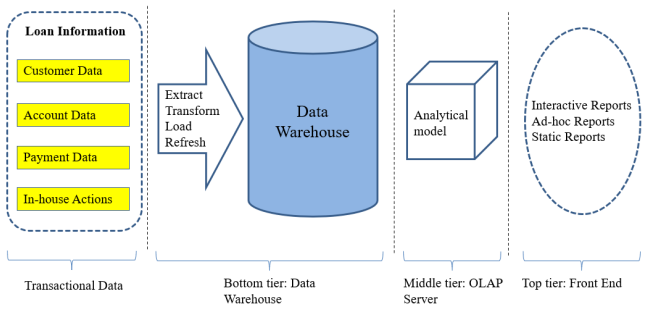

Figure 7. Architectural design diagram of Loan data analysis

D. ETL (Extract, Transform, Load)

ETL is a process of load data from source to destination. ETL defines Extract, Transform, and Load. Extract in the sense extracting data from the source. That source may not be an SQL environment that could be several types of sources. In this operation, we will face extracting data from many systems platforms such as flat files, web servers, emails, images. Due to this we are extracting several types of data into one database called staging area. You have to do cleaning and changes to the operational data that simplifies the building summaries that help to avoid slowness, security matters, impossibility, also it avoids the conflicts between source systems. That can be used as a backup option as well.

The second stage of the ETL is transform, it consists of several rules before load data to the destination. With this stage filtering, sorting, pivoting, validating, deriving columns, joining or merging, splitting, lookups operations can be performed. The above rules vary between the business need. The approach mostly uses derive functions in this stage to derive and convert data to certain types.

Third and final stage is loading. This stage will check the error logs send a notification upon that to the administration or the users regarding the availability of the data that can be performed after loading data into the destination.

This study has chosen Microsoft BI SQL Server Integration Service (SSIS) tool to perform the ETL process.

Figure 8. ETL for FactLoanCollection

The above figure illustrates the ETL process of this particular approach. With this ETL has done the full load for the fact table, the “SQL Execution” task was executed to perform truncation prior to load fact. Also, the cube dimension and the fact process are also done by this stage. This will be a very convenient process if the developer runs on the schedule of those tasks. No one needs to run every task manually. The administration just has to publish the ETL package to the package store and enable the SQL job for us. That process will run every time that you have scheduled previously. Apart from that, it will be able to add the task to send a notification to the administration or users.

E. Analysis

Data Analysis is a process to discover useful information for business decision making. This is used to extract useful information from the raw data to make a day-to-day decision based on that information. That may vary according to business needs or individual needs. Analysis is most of the time based on the future prediction by looking into past information or patterns that we have discovered, something like forecasting whether and the environmental factors.

There are several types of data analysis techniques such as Text analysis, statistical analysis, predictive analysis, diagnostic analysis, and prescriptive analysis.

There are several tools to accomplish the data analysis task, this study has used Microsoft SQL Server Analysis Services (SSAS) [11] to deliver our framework for the users.

F. Reporting

Reporting means illustrating information in a proper understandable manner to users. There are several technical tools such as excel, PowerBI reporting, SQL Server Reporting Service [9], Tableau so on. In this approach, we have used excel features to represent the dashboard with the pivot charts.

This study was conducted to identify the best time slots to make an effective customer call by a call center agent for financial institutes other than collecting a lot of information about loan borrowers.

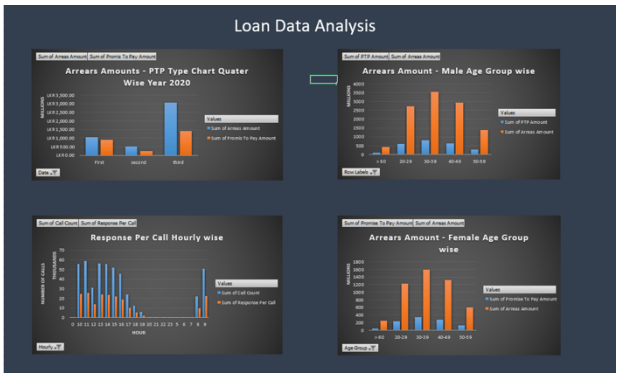

Figure 9. Dashboard of data analysis

This study was analyzed by certain areas such as arrears amount for the year 2020 quarter wise, Arrears amount gender and age group wise, and response per call hourly.

This study was conducted to identify the borrower’s response per the call on a time basis per day. To accomplish that, there is a special dimension call “DimTime”. This dimension contains an hour, minute, and second hierarchy to illustrate data which is the extension of Date Dimension [10]

TABLE I : Response Per Call Age Group Wise.

Hour	Age Groups				
	20-29	30-39	40-49	50-59	> 60
7	0.10%	0.15%	0.13%	0.29%	0.00%
8	4.14%	4.03%	3.28%	4.49%	4.43%
9	10.26%	10.45%	10.07%	10.49%	9.03%
10	10.63%	11.06%	11.17%	11.37%	11.12%
11	11.25%	11.55%	11.97%	12.03%	13.04%
12	5.87%	5.97%	6.17%	6.04%	7.86%
13	11.62%	11.06%	11.99%	11.21%	11.45%
14	11.72%	11.35%	11.82%	10.40%	11.71%
15	11.12%	10.81%	10.37%	10.91%	10.20%
16	9.23%	9.73%	10.15%	10.05%	8.36%
17	6.66%	6.86%	6.46%	6.32%	5.94%
18	5.05%	4.82%	4.57%	4.56%	4.52%
19	2.36%	2.15%	1.83%	1.85%	2.34%

G. Non Functional Requirements

Very important factors include the data analysis, that contains the non-functional requirements, which most of the parties do not consider and may be not known. This can be defined as quality attributes, such as usability, availability, security, and reliability. Those important factors must be involved with the data. If we are working with banking data sources those should be considered mainly due to security. Availability which means that data should available any time anywhere to access for the users with the credentials they provide, credential? Yes, when we get with the credentials first thing that comes up is security. The majority nowadays most organizations consider those factors. Analysis systems with nonfunctional requirements provide a better framework.

IV. Future Work

The financial institutions will reduce their risk factor if they have a proper evaluation of the loan or credit borrowers. Developing a proper classification algorithm with a decision tree will be able to avoid such a risk. To identify, the borrowers can develop a popular data mining technique. Decision tree is based on [8] algorithm with the factors such as gender, profession, and location of the borrowers according to the training data set.

A decision tree is a flowchart, it is like a tree. This is a top-down approach. This structure is made on training data set, those data are broken into a small set of subsets and illustrate in the nodes of that tree. The decision tree has a root node, branches, leaf nodes. The decision tree uses classification and regression which are used to predict and divide the nodes. The classification has been used for categorizing data set into certain parameters [7]. Regression is used for the prediction of the values, this is used to predict the model's continuous values.

V. Conclusion

In conclusion, some financial institutions are facing a huge risk due to the failure of recovering loans and collections on time. To avoid this and make decisions on the past data by analyzing them. This study was implemented in a framework. It enables to fetch the reports as per the user requirement with the available data set. For this, Microsoft BI tools were used in order to develop ETL and Cube. Excel enables access of this framework for the users, mainly for the management to make their decision, on that they will be able to reduce their risks by giving loans or credits to certain borrowers. This study was conducted to identify the borrower's response as per the call on a time basis per day. As a result, the researcher was able to identified different key response times of loan borrower's based on several parameters such as age groups and gender. This study varies from the other studies by recognizing the best time slots to make an effective customer call by a call center agent for financial institutes. For future work, we have decided to implement a decision tree-based algorithm with classification and regression techniques.

References

[1] AbrehamGebeyehu (2002). "Loan repayment and its Determinants in Small-Scale Enterprises Financing in Ethiopia: A Case of Private Borrowers AroundZeway Area", M. Sc. Thesis, Addis Abeba University.

[2] Ted E. Senator, Henry G. Goldberg, Jerry Wooton, etc., The financial crimes enforcement network AI system (FAIS) identifying potential money laundering from reports of large cash transactions[J], AI Magazine, Vol.16, No.4, pp. 21-39, Winter 1995.

[3] Safavin,S.R., Landgrebe,D. A survey of decision tree classifier methodology [J]. IEEE Transactions on Systems, Man and Cybernetics, Vol.21, No.3, pp.660-667, April 1991

[4] SU-NAN WANG, JIAN-GANG YANG. A MONEY LAUNDERING RISK EVALUATION METHOD BASED ON DECISION TREE, College of Computer Science and Engineering, Zhejiang University, Hangzhou 310027, China 2Shanghai Pudong Development Bank, Shanghai 200002, China 2007.

[5] JI Chengjun, WU Lijun, LI Jinping. The Application of the Decision Tree Analyzes in the Credit Card, Department of Management, Liaoning Technical University, Huludao, China

[6] Yonas Shuke Kitawa, Nigatu Degu Terye. (2020/10/30). Statistical Analysis on the Loan Repayment Efficiency and it's impact on the Borrowers: A case study of Hawassa city, Ethiopia Available :http://article.sciencepublishinggroup.com/html/10.11648.j.ajtas.20150406.28.html

[7] (202/09/13), Decision tree algorithm example in data mining. Available : https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/

[8] Wang, J. Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications. Hershey: Information Science Reference, 2008.

[9] Microsoft (2020/07/22). Analysis and reporting with Microsoft business intelligence (BI) tools. Available : https://docs.microsoft.com/en-us/sql/reporting-services/choosing-microsoft-business-intelligence-bi-tools-for-analysis-and-reporting?view=sql-server-ver15

[10] Create an Extended Date Dimension for a SQL Server Data Warehouse, Available : https://www.mssqltips.com/sqlservertip/5553/create-an-extended-date-dimension-for-a-sql-server-data-warehouse/. Dinesh Asanka, MSSQLTips.com.

[11] What is Analysis Services?, https://docs.microsoft.com/en-us/analysis-services/analysis-services-overview, Microsoft.