ViviSight: A Sophisticated, Data-driven Business Intelligence Tool for Churn and Loan Default Prediction

Barun Paudel[1], T.H. Gopaluwewa[1], M.R.De. Waas Gunawardena[1], W.C.H. Wijerathna[1], Rohan Samarasinghe[1]*, Hansa Perera[1]

[1]*Sri Lanka Institute of Information Technology,*
**Corresponding Author: rohan.s@sliit.lk*

Abstract — In today's world, we can see that data is being generated at a pace like never before. We have enough data which still don't have a suitable form of insightful information. For any business to turn into profitable ventures, top level management and work force highly rely on decisions - which in turn are dependent on the kind of information available. Commercially available BI tool possesses the significant limitations which are being highly expensive with limited features, insufficient customer profiling and need of expert knowledge to use them .This is creating a huge gap in catering the actual need of top level management in their decision making process. This research focuses on providing insightful information by overcoming existing limitations. If we consider Sri Lanka, telecommunication and finance industry are among the core driver of Sri Lankan economy. But they are facing problems in finding hidden patterns in their large chunk of data which have a big impact on their decision making process Most telecommunication companies are suffering from churn(customer leaving service). While in finance and banking sectors, number of loan defaults (customer not paying their loan back) is increasing day by day which is creating a huge financial loss. Through our product ViviSight, we are focused on problems faced by telecommunication and finance industries and provide the sophisticated, user-friendly, cost-effective and comprehensive data-driven business intelligence tool for efficient prediction.

Keywords— ***Business Intelligence, Data-mining, Predictive modelling, Statistical Analysis, Visual Analytics***

I. INTRODUCTION

This Research Project- ViviSight targets in developing set of tools, technologies and programmed products that are used to collect and make data available for better, faster and accurate decisions for telecommunication and finance cum banking sectors. Existing BI solution from Vendors like Microsoft, IBM, Oracle etc. are very expensive which cannot be afforded by small and medium companies for long run. Not only cost factor, these tools are very complex in nature that need an expert knowledge to use them which ultimately increase the cost. Significant limitations also exist in features which don't specifically cater the need of telecommunication and finance industry. When it comes to decision making process, customer profiling is a major component. Existing BI tools failed to provide effective way of analysing each customer behaviour effectively. When a company wish to buy this commercially available product, they need to buy ETL, predictive tool and Dashboard separately which is a huge drawback. This has resulted a huge gap between need of company and existing solution. This gap can be significantly reduced by carefully selecting appropriate tool and technologies focusing on specific problem. This Research adopts the same approach providing cost effective, ease to use BI tool with detailed customer profiling having ETL, predictive tool and Dashboard as a single product.

According to information carried out by Research [1] it has been stated that telecommunication companies has churn rate ranging from 10-67% which is highest among all other companies. There might be various reasons behind a customer leaving the service but most influencing is customer dissatisfaction. Customer shows certain churning behaviours which could be used as a parameter to identify likelihood of customer leaving a service. When it comes to finance and banking sectors, customer becoming a loan default has been a major issue. The research carried out by this article [2] shows a customer not being able to pay back loan to Bank creates a huge financial loss to Banking and Finance sector. Hence, considering all these problems faced by above mentioned companies, this research product –ViviSight in advance provides valuable, fully data-driven and fact-based information's and insights to top level management in a company so that they can make efficient, accurate and result-oriented strategic decision. ViviSight makes a prediction efficient by carrying out statistical analysis and data mining technique to calculate the probability of each customer leaving the service or being loan default.

II. METHODOLOGY

In order to carry out this research project, Agile methodology was followed. Before starting the Research, detailed investigation was carried out to identify major business problems faced by the companies. After listing down problem domains, literature survery was done based on the available business intelligence tools and their drawbacks and short-

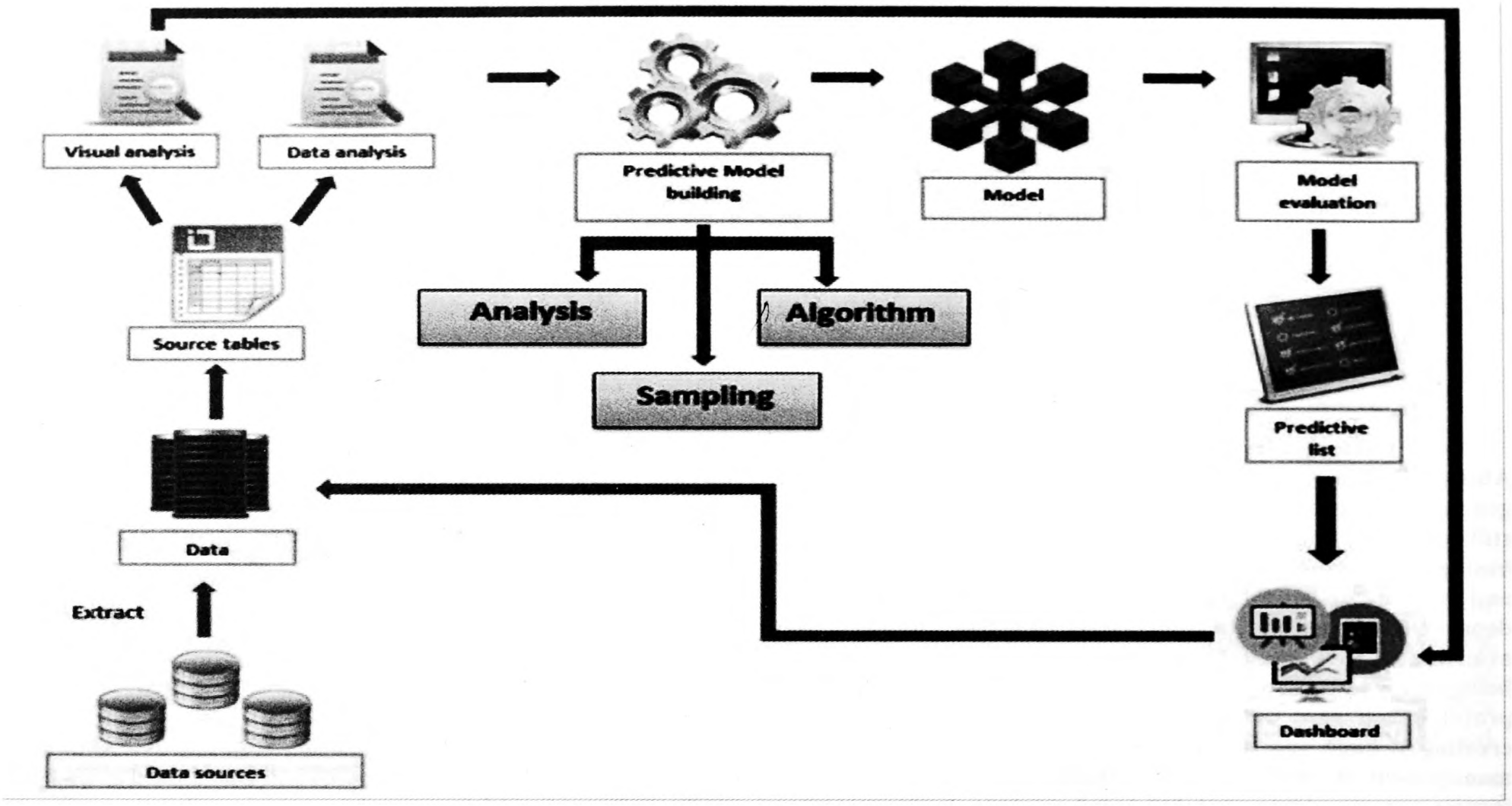

Figure 1. System Architecture of ViviSight

comes were listed down.Then Research team purposed solution-ViviSight which addresses major business problems in telecommunication and finance industry.

Research Team also carried out traditional feasibility study which proved that ViviSight is financially, technically and operationally feasible because it is built on open-source technologies and has no technological constraints and dependencies.Then focus was shifted on gathering fucntional and non-fucntional requirements of ViviSight.

During design phase, high-level architecture design was developed inorder the incorporate the gathered funtional and non-functional requirements of ViviSight.

Research Team also moduled the basic workflow of system along with architectural design.Basically ViviSight consists two arms: Desktop application and web-based application. Architecture is further categorized into three parts.1.ETL 2.Predictive model building and Evaluation 3. Dashboard .Figure 1 depicts high level architecture of ViviSight.

A. ViviSight Architecture

1) *ETL (Extract, Transform and Load):*

a) *Extract*: When it comes to practical scenario, there can be different data sources related to the system stored in different locations in different formats.

For example there can be several sources for Billing Support System, Operation Support System, Customer Relationship Management, Customer Experience Management… etc. Also these data can be available in different formats as xls, xlsx, csv, txt, xml... etc. The system should be able to identify these data sources as needed.

When it comes to data warehouse it could be implemented using Oracle, SQL, MySQL or other technology. The system should be able to extract data from whatever the format it's available. The required attributes for the prediction model should be selected and extracted from the DWH.

b) *Transform*

In the transform step set of rules to transform the data from the source to the target is applied. It includes joining data from various sources, aggretating them with suitable keys and apply advance validation rule to make the transformation process effective.

c) *Load*

The existing information from given data sources are then loaded and saved in a form of tables. Loading is done for 2 purposes:

i) Visual Analytics: It is used to show the transformed data in a graphical way (using bar charts, pie charts, table etc.) so that

the user can analyse them and get an idea about the behaviour of present and historical data.
ii) Data Analytics: In this component data is stored in a form of tables in order to use in predictive analytics. Predictive models use these tables as input for the analytics process.

3) *Predictive Model Building and Evaluation:*

In this process, research team have used most-effective approaches in predictive analytics to create a statistical model of future behaviour of given data set. In predictive analytics, classification approach of data mining concerned with forecasting probabilities and trends was used. Various algorithms using classification techniques were developed because research is dealing with binary classification problems: customer will be churn or not and customer will be loan default or not. Hence, in model building process, classification techniques of data mining is used.

For Model Evaluation process, confusion matrix techniques are used to check the accuracy of built algorithms.

4) Dashboard

Dashboard is the web-based component in ViviSight which provides content-rich visualization to data and gives the high-level perception of insights and information about data. ViviSight tool have generated graphical reports that describe data in most efficient manner. ViviSight Dashboard provides advanced data visualization and guided analysis through auto charting, and offers an array of visualization techniques to present data and results in the most insightful way.

B. Implementation

Before directly going into the implementation, research teams have designed the basic workflow of our system that helped them in an effective implementation of ViviSight.

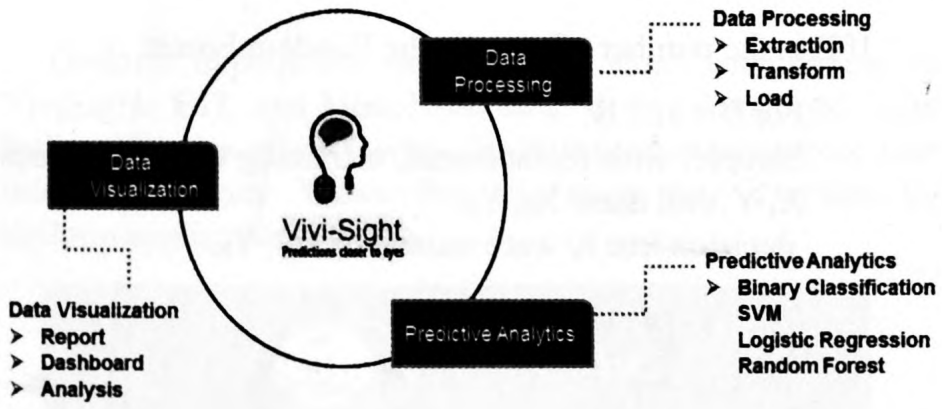

Figure 2. Basic Workflow of ViviSight

Considering functional and non-functional requirements for ViviSight, research team bundled their activities in following phases for the successful implementation of System.

- Preprocessing of data
- Finding most influential data attribute
- Data sampling into training and test set
- Model Building
- Benchmarking of each model where necessary and comparison to others to select the best one with highest accuracy.
- Dashboard Building
- ViviSight System integration

First and most important step was to obtain the dataset for research. Research Team have managed to obtain data set from a BI Company N-able which is based in Sri Lanka.Telco-data set have 10,000 data instances with different attribute included whereas Finance data-set has 20,000 data instances.

Research Team have carried out data imputation techniques in-order to cope up with missing data values. From the given dataset, it has been discovered that it is needed to apply data imputation techniques for those data instances which have errors during data entry by the users.

It is also found that some data fields are left empty and some values are with irrelevant information. For missing values, they substituted with median value. Data field which has same values across all data instances were removed so as to fit it to the models.

Table I presents the list of valuable data attribute with their description after preprocessing the financial data set.

TABLE I: DESCRIPTION OF IMPORTANT DATA ATTRIBUTE FOR FINANCE-DATASET

Parameter	Description
loan amount	Amount of loan requested by borrower
int_rate	Interest rate of the loan
annual_inc	Annual income of borrower
Purpose	Main intension of borrower for asking a loan
Installment	Monthly payment owed by borrower if he/she is granted a loan
sub-grade	More details about the loan
total_acc	Total account of the customer
emp_length	Employment details of borrower
Desc	Length of loan description

TABLE II: DESCRIPTION OF IMPORTANT DATA ATTRIBUTE FOR TELCO-DATASET

Parameter	Description
last_month_message	Message sent in and out last month
outgoing_call	Outside Calls made from the current subscribing sim
weekly_revenue	Total money spent by the customer in consuming the service
no_of_complains	Frequency of complains made by customer based on service dissatisfaction
no_of_short_call	Short duration calls made by customer
Incoming_call	Frequency of incoming call
Tower_id	Unique identification of tower
Contract type	Weather the customer is pre-paid or post-paid

Table II presents the list of valuable data attribute with their description after preprocessing the Telco data set.

Now focus was shifted and concerned to sampling of data into training and test set. Whole data was divided into training data (75%) for feeding to models and testing data (25%) in order to benchmark the models and find out the most accurate model.

Next approach was to choose between Python and R for statistical modeling and programming.

Python was selected as the best suitable programming tool for ViviSight due to following reasons:

- ViviSight includes ETL and Dashboard that are to be integrated with statistical model building component. Hence, Python is used since it is easy to integrate in a production workflow
- In programming point of view, python is easier to code than R.
- Python contain libraries like scikit-learn which are best for machine learning.
- Python's pandas package is now mature enough to outclass R data frame, especially where time-series indexing is concerned.

Once the decision for programming language was chosen, research team was focused in building predictive models using data-mining and machine learning techniques.

Since research is dealing with classification type problem- weather the customer will be churn or not and weather customer will be loan default or not, prediction model building was categorized into two parts:

TABLE III: COMPARISON OF R AND PYTHON

R	Python
High Complexity	Comparatively low complex than R
Already written packages for statistical modelling	Best if major concern is programming
Syntactically less cleaner than Python	Syntactically more cleaner than R
The libraries documentation isn't user friendly.	Rich documentation for libraries
It is harder to integrate to a production workflow.	It also integrates easily in a production workflow.
It has a huge set of libraries available for different statistical type analysis.	The scikit-learn library, panda, scipy, numpy are effective for machine-learning tasks.

1. Churn Prediction Model
2. Loan Default Prediction Model

Following classification algorithm were applied inorder to come up with ViviSight prediction model.

i. Random Forest

As the name suggest ,random forest of trees is grown during the learning process. Each tree has an internal node where the attribute are tested and classification is assign to every leaf node.

If B is the number of trees in the Random Forest

For b = 1,... B:
Sample, with replacement, n training examples from X, Y; call these X_b, Y_b
decision tree f_b were trained on X_b, Y_b.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x') \qquad (1.1)$$

ii. Support Vector Machine

Recently support vector machine has been used widely for solving classification problem. In this approach, we have reviewed the research work on SVM by (Vapnik, 1995;

Vapnik, 1998a; Vapnik, 1998b). In SVM, input Vector Class $\mathbf{X}$ can be decided by evaluating the sign of $y(\mathbf{x})$.

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{1.2}$$

Once the prediction models were built using above approaches, training set were used to train these models. Once these models were learned using the training set, following method were incorporated so as to benchmark the prediction models that have been developed so far.

i. Confusion Matrix as Performance Metrics

A classifier is typically evaluated by a confusion matrix:

	Predicted Negative	Predicted Positive
Actual Negative	**TN**	**FP**
Actual Positive	**FN**	**TP**

Where TN is total Negative x classified correctly by model, FN is total number of Positives classified incorrectly by model; FP is total number of Negative classified incorrectly by model, and TP is the number of positive that is classified correctly by model.

ii. Receiver Operating Curve (ROC)

ROC curve demonstrates the performance of a model that uses binary classifier. The X-axis is false positive rate (= FP/ (FP+TN)), and the Y-axis is true positive rate (=TP/(TP+FN)). The optimal threshold point can be found through locating the point on ROC that is closest to (0, 1).

After research team have selected above classification techniques in order to build churn prediction model and loan default prediction model, focus was shifted on implementing two arms of ViviSight application.

i. Desktop Application.

Desktop application includes two core components of ViviSight- ETL and Model Building. It was developed using Jython libraries which gives flexibility of creating nice user interfaces. Figure 3 describes the main user interface for desktop application.

Figure 3. Home screen for Desktop application

ii. Web - based application

It includes another core component of ViviSight-Dashboard. It is used for the visualization of data. Data can be visualized in two ways. One way is exploratory data and next one is insightful data. . It was developed using JavaScript, php and high-charts were used for graphical reports and visual analytics. Figure 4 shows the dashboard design of ViviSight.

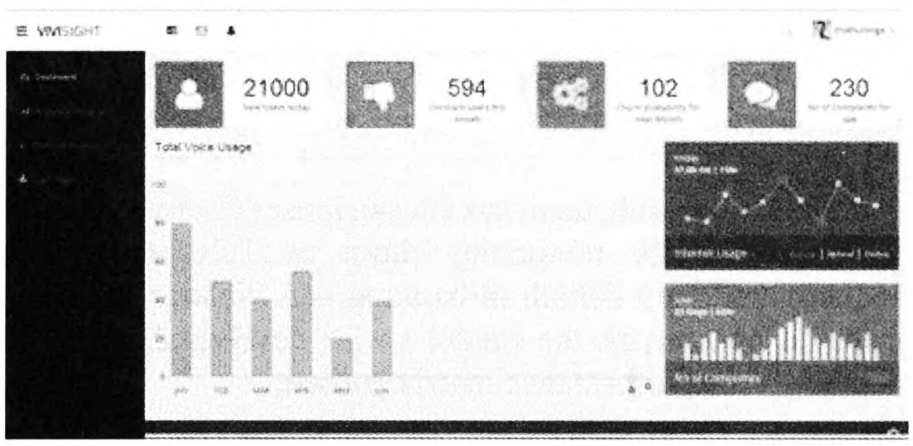

Figure 4. Home screen for Web application

III. RESEARCH FINDINGS, RESULTS & EVIDENCE

Final ViviSight system was benchmarked for its ETL Component, prediction result, model accuracy and Dashboard features. The results are shown in this section.
ViviSight ETL component has unique feature of mapping given dataset attribute to the database fields. It will help to proper transformation of dataset attribute to acceptable fields.

In this research, in order to build our models, scikit-learn a machine learning library in Python was used. They are: sklearn.ensemble.RandomForestClassifier for Random Forest model, and sklearn.linear model.LogisticRegression for Logistic Regression, sklearn.svm.SVC for Support Vector Machine, sklearn.neighbors.KNeighborsClassifier for K-Nearest Neighbors.

Research team were able to find churning probability of each customer in the given data set. Figure 5 and Figure 6 shows probability that is calculated using different algorithms.

ACCOUNT_ID	Gender	Age	Probability_rf	Probability_svm	Probability_knn
10617598	Male	65	0.5	0.4	0.6
10620056	Female	20	0.2	0.1	0
10620266	Female	74	0.1	0.2	0

Figure5.Churning probability calculated using different model

Figure 5 stored the churning probability that were calculated using different algorithm and stored them into the database and that database worked as the backend for data visualization for predictive analytics in Dashboard.
Probability_rf= probability calculated by Random Forest
Probability_svm= probability calculated by SVM
Probability_knn=probability calculated by K-NearestNeighbor

id	loan_amnt	int_rate	Probability_rf	Probability_svm	Probability_lr
631535	7000	13.80%	0.3	0.1	0.3
632417	35000	20.25%	0.4	0.6	0.7
633105	16000	19.74%	0.3	0.2	0.4

Figure6.loan-default probability calculated using different model

Figure 6 stored the loan default probability that was calculated using different algorithm.

Probability_lr= probability calculated by logistic regression

In this research, team has chosen most effective and widely used models for predicting churn in Telecommunication Industry and loan default in banking and Finance Industry. In order to benchmark the model so far developed, ROC-AUC Values as the performance matrix is used.

TABLE IV
ROC-AUC VALUES FOR TELCO-CHURN PREDICTION MODELS

Prediction Model	AUC Values
Random Forest	0.855
Support Vector Machine(SVM)	0.836
K-Nearest Neighbor	0.831

From table V, it is found the Random Forest is comparatively most accurate model among all for churn prediction having AUC score of 0.855.

TABLE VI
ROC-AUC VALUES FOR LOAN DEFAULT PREDICTION MODELS

Prediction Model	AUC Values
Logistic Regression	0.821
Support Vector Machine (SVM)	0.813
Random Forest	0.810

From table V, it is found the Logistic Regression is comparatively most accurate model among all for churn prediction having AUC score of 0.821.

Not only AUC curve, Confusion matrix is also employed in order to compare the result of different algorithms and calculated the matrix which significantly decides the best model to approach given classification problem.

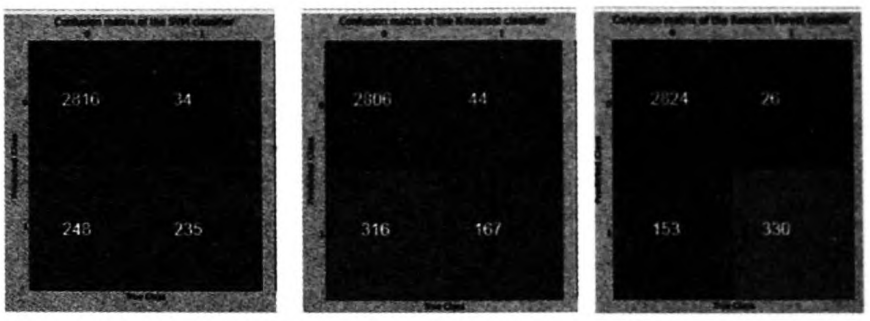

Figure 5. Confusion Matrix for each developed algorithms

Random forest again out preforms the other two at about 93% precision (330 out of 356) with support vector machines a little behind at about 87% (235 out of 269). K-nearest-neighbours lags at about 80%.

ViviSight possess the unique features in comparison to the existing BI tool which is customer profiling. Figure 6 talks about Customer profiling which is one of the best features of our tool-

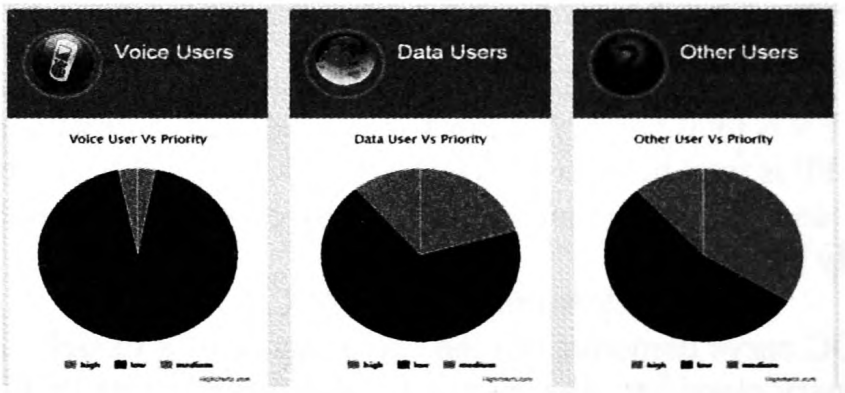

Figure 6. Customer Profiling in Telco Company

ViviSight. When it comes to company, customer are one of their biggest asset. They cannot afford to lose them. Research shows that it is always highly expensive to gain new customer rather than retaining existing customer.

IV. CONCLUSIONS & FUTURE WORK

This paper provides possible approaches that are best suitable to develop prediction model based on statistical model, data-mining and machine learning approaches. .ViviSight stands itself unique form existing product being as a comprehensive product that includes ETL. Prediction and Dashboard all in one.

ViviSight has developed now mainly for addressing the problem in Telco's and finances Industry. In next stage, in one hand, research team will be focusing on increasing the accuracy of these model, maintain the customer profiling mechanism and increase the performance of prediction process by introducing more effective data cleansing and transformation techniques. It is hoped that for any person who expects to build a similar system or any other real-time system, results of this research will be an aid and will provide insight on the performance, accuracy and reliability level that can be expected with the combination of tools, technologies, programming approach considered in this paper.

ACKNOWLEDGMENT

Our sincere gratitude goes to to the authority of Sri Lanka Institute of Information Technology (SLIIT) who helped us in so many ways in providing us with a good environment and facilities to complete this project. We would also express our thankful words to our families and friends for their understandings and supports on us in completing this project.

REFERENCES

[1] M. Flynn, "OpenGTS - Open Source GPS Tracking System," O'Reilli, 19 May 2009. [Online]. Available: http://whereconf.com/where2009/public/schedule/detail/7086. [Accessed 20 August 2014].

[2] ThreeHosts.com, "Fedora Vs. Ubuntu Vs. CentOS - Server Reviews From ThreeHosts.com," PRWEB, 28 October 2013. [Online]. Available: http://www.prweb.com/releases/compare-linux-centos-vs/ubuntu-vs-fedora/prweb11273960.htm. [Accessed 22 August 2014].

[3] Greystoke1337, "Big Data," Wikipedia, 22 August 2014. [Online]. Available: http://en.wikipedia.org/wiki/Big_data. [Accessed 23 March 2014].

[4] I. P. Singh, "JMeter Load Testing Beginner Tutorial," 28 October 2013. [Online]. Available: https://www.youtube.com/watch?v=4mfFSrxpl0Y. [Accessed 1 September 2014].

[5] S. Karlsson and E. Hansson, "Lossless Message Compression," 2013. [Online]. Available: http://www.diva-portal.org/smash/get/diva2:647904/FULLTEXT01.pdf. [Accessed 14 April 2014].

[6] P. Barry, "Which Programming Language?," Linux Journal, 7 February 2001. [Online]. Available: http://www.linuxjournal.com/article/4402. [Accessed 2 May 2014].

[7] R. Graham, "C10M," Atom, February 2013. [Online]. Available: http://c10m.robertgraham.com/p/manifesto.html. [Accessed 22 August 2014].

[8] "HAProxy," [Online]. Available: http://www.haproxy.org/. [Accessed 25 June 2014].