Customer Behaviour Prediction Using Web Usage Mining

MrR N Devendra kumar[1], P Jayarajruban[2], R Madhan Kumar[3], M Harish Kumar[4]

[1]Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology,
[2,3,4] Student, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT: Customer behaviour prediction is really important in analysing what the customer needs and also to identify whether the customer is satisfied with the services. Retaining the current customers is as important as gaining new customers. So the behaviour pattern of the customers need to be analysed to see if they are satisfied with the services. Every telecom organizations use this technique to identify the customers who are most likely to unsubscribe from their service and provide them with better services to stay competitive. Our project aims at developing a system to identify the behaviour pattern for a bank dataset. This will help the banks identify the customer who will potentially leave their bank and try to retain them by offering better service. We have used Random Forest Classifier(RFC) for classification as it has proven to have better accuracy.

Key words:Customer Behaviour Prediction, Random Forest Classifier.

I. INTRODUCTION

The most important goal in building a behaviour prediction system is understanding the customer behaviour, needs and the similar patterns in order to provide good customer service. Nowadays there are lot of techniques and tools available in internet to analyse and predict the customer behaviour. Our system will identify and extract potentially useful patterns and hidden information from the user data by applying some classification algorithms. Therefore, our customer behaviour prediction system can be used to find those obvious data that have potential value to reduce competition and simultaneously increase business profit. Our project aims to develop a system which will be able to analyse and predict the user's behaviour pattern in a bank customer dataset. This will help the bank identify the customers who will potentially leave the bank and try to retain them.

As we know it is more difficult and expensive to get a new client than retaining an existing one. It is really important for any business to know what leads a client towards the decision to leave their service. Having the ability to predict that is necessary because it helps a business gain a better understanding of future expected revenue. In addition to that it also allows us to identify the particular customer who might discontinue the service and allows us to target that individual in attempt to prevent them from leaving their service. Also it helps the business identify and improve upon areas where customer service is lacking.

Analysing user's behaviour is one of the biggest fields of big data analysis. Several research has been done on this field of study and a lot of algorithms are developed to analyse the user's behaviour. Most of the research uses telecom service data to identify whether a customer will keep using their services or leave their service. This study is really important to retain the customers and to stay competitive. Every service sectors realise that retaining their existing customer base is as important and challenging as obtaining new customers. So they use prediction algorithms to see if their customers are satisfied with their service. In our project we are using bank dataset to predict whether the customer will leave the bank or not. This will help the bank to retain the customers and to provide better services.

II. LITERATURE SURVEY

Irfan Ullahet.al.,author focuses on developing an algorithm to identify if a customer will continue using a service or not for a telecom sector. Customer Relationship Management (CRM) analysers try to identify the reasons for a customer leaving a service and identify the pattern. RF algorithm is used because it has proven to have better accuracy than k-means clustering algorithm[1].

Sonia Sharma et.al., author explains about

data mining process and uses some random available tools online to analyze web log files and data is collected by using free web analyzer tools. They focus on education domain. Here the study is made between top most universities in Haryana by knowing the customer interaction between period December 2016 to February 2017 on the websites of the universities. The various data such as global rank, country rank, total no of views, loading time are collected as data. The given data fed into the tools such as page score and it provides the findings such as university which has good global rank and country ranking, the university which has less loading time of website compared to others and also the college website which has highest views. The main goal of this research is to understand the user behavior pattern in order to improve their organization's website[2].

Jitendra Net.alauthor focuses on data mining concepts and its tools and technologies to obtain proper result for customer behavior analysis. It explains about the stages of web usage mining and also detailed categories of web usage mining. This paper emphasize on user future next request prediction using web log record, click streams record and user information. Author classifies the scope of web mining based on the work done by each technique in respective fields. Those are global scope (web content mining and web structure mining) i.e. spanning an entire web and local scope(web usage mining)i.e. spanning an individual web.it shows the comparisons and summary of various methods of future request prediction with application[3].

AdebolaOrogunet.alauthor explains about association rule mining model to predict customer behavior. Here they using typical online retail store for data collection. The dataset contains all customer transaction occur between certain period of time and attributes such as unit price, invoice date, customer id, invoice number, stock code, description, quantity, country. Language used here is python. Using these dataset they implement Association rule mining and extract important trends from the customer behavior data.[4].

Tawfiq A. Al-asadi. author gives a table of information which explains about approaches in web mining taxonomy, data sources, representational models and techniques. Web Mining techniques are classified into three. Author clearly picturized an image which gives information about current techniques which used to analyse web contents. In this paper the author briefly explains about unstructured, structured, semi-structured and multimedia techniques. Here they gives a detailed information about client level log, server level log

and proxy level log also explains different web usage mining techniques such as classification, clustering, statistical analysis, Association rule which is applied to extract knowledge from log data[5].

Snehal Kulkarni et al.,author describes about awhich system records the behaviour and pattern of surfing of customers. The patterns are found out through web mining. Here the author has setup a dummy website to acquire the dataset. Various steps such as Data validation, data preprocessing applying back propagation neural network, prediction are carried out in the system. Data validation is done by using ANOVA and K Fold Cross Validation. It has been found that K Fold Cross validation is more effective than ANOVA. Errors are being evaluated using mean square error. Data preprocessing is done to transform raw data to require format. Back propagation neural network is the algorithm that is used in this system. This will significantly affect the predicted result. Using this output whether the customer will buy the product or not can be found out[6].

Dr.K.Maheswariet.al.,Author uses SVM algorithm is used to classify the customer and predict their behavior. Sales data set taken from the internet is used. Here they give a general information about data mining and explains about characteristics and applications of data mining. Here the attributes used for analysis is based on customer visit in the shopping site. If a customer visit every day, every week or three times in month they classified as regular customer. Based on non-availability the customer classified as customer on occasions. If a customer visits once in 3 months they classified as customer on festival seasons. Customer who never purchase but visit website are classified as Customer on window shopping. These data is preprocessed and the missing data and not supported data are identified and cleaned using data cleaning method and SVM model was created for this dataset. They give the results of frequency of product based on sales and purchase date and also percentage of online shoppers based on their age and years. Also give comparative analysis of customer classification with various parameters. With the help of linear kernel the SVM models are created. The Support Vector Machine with classification algorithm is used to predict the online purchase behaviour by data mining .The research has shown that in recent years kid with age below 7 is allured to buying things online. The purchasing amount of customers has been increased during the offer times[7].

R.Khanchanaet.al The author discusses about Fuzzy Possibilistic algorithm for clustering which is used for predicting customers' behavior. In this algorithm clustering is used as the classification technique. It uses two levels of prediction model which was suggested by Le and Fu. This model was created by merging Markov model and Bayesian theorem. Predicting categories and Predicting pages are the two levels in this model. They used a part of news data to perform web mining using the FPCM clustering. The results Show that the hit ratio of this algorithm is higher than other hierarchical clustering techniques.[8].

III. EXISTING SYSTEM

Customer behavior prediction models are widely used by many service providing organizations to predict their customers behavior. This is one of the fastest growing fields in big data analysis. Based on data that the model is dealing with several different algorithms are used to predict the customers behavior. The accuracy of the algorithm decides which algorithm works best. Some of the most popularly used algorithms to predict customer's behavior are decision trees, neural networks, SVM and many more. Based on the data at hand some algorithms perform better than other. Due to constant development of new algorithms the accuracy of the model increases when applying those algorithms.

IV. PROPOSED SYSTEM

Our proposed system aims at providing better accuracy than some of the most popularly used machine learning algorithms. Our model uses Random Forest Classifier algorithm to predict the customer's behavior for a bank user dataset. Random forest classifier are easier to implement than most algorithms. Random forest classifier is made up of decision trees. While decision trees use specific features to produce results random forest classifier uses its randomness to predict the result which provides better accuracy.

First the dataset is imported and preprocessing is done where the irrelevant fields from the data set are removed. Now the string variables in the data are converted to integer values and this process is known as label encoding. After pre-processingheatmap function is used which identifies the fields that highly influence the result.The fields identified using the heatmap function are age, geography and balance. the data is split for training and testing (x= 80% for training, y= 20% for testing) and then scaling process is done using standard scalarfunction.Finally the random forest classifier algorithm is applied and the

confusion matrix and the accuracy of the algorithm is obtained..

A. Flow Chart
A diagram which visually displays correlated information such as events, steps in a process, functions is knows as flowchart.

From the flow chart the process flow of the system is explained. First the dataset is imported .After that the data is pre-processed by removing the unwanted fields and by turning the string data into integer data. Then the pattern of the data set is analyzed and the important features are discovered. Finally the algorithm is applied and the results are obtained.
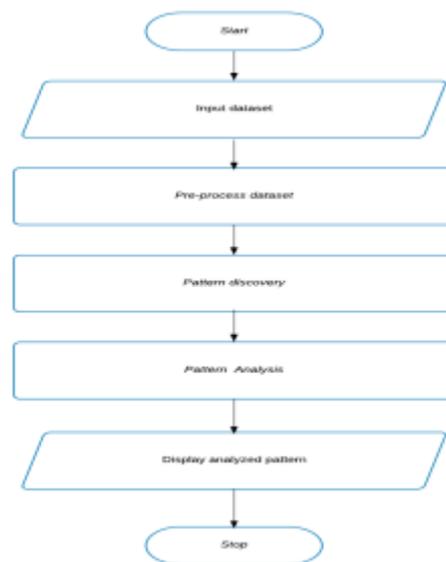

Fig 1.1. Flow chart

B. Use Case Diagram
Use case diagrams are used to display system functionality and requirements.

The use cases in this system are fetch dataset, pre-process data, discover pattern, analyze pattern, produce result, enter details, view results.The functionalities ofuser are: Input their details view the prediction results. The functionalities of Prediction algorithm are:import the dataset, pre-process the dataset, analyse and discover pattern, produce prediction result
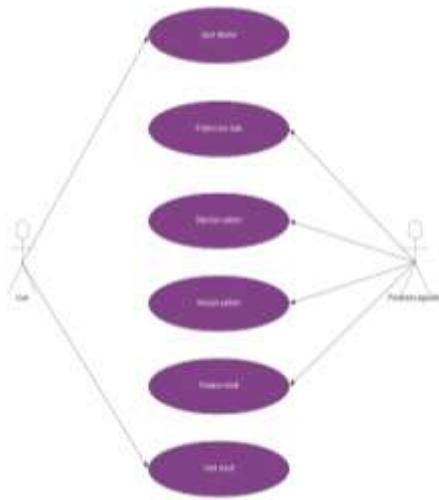
Fig 1: Use Case Diagram

C. Sequence Diagram
The flow of operation of the system is explained below:

- First the dataset is imported by the algorithm.
- Pre-processing is done where unwanted fields in the dataset is removed and the string variables are converted to integer variables.
- Now the factors which mostly influences the result is identified using heatmap.
- Finally the Random Forest Classifier algorithm is applied to produce the result and then the algorithm is saved.
- When we enter the customer details in the webpage the algorithm will process it and produce the result.
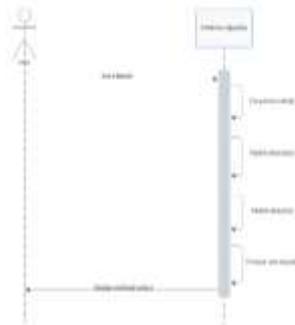
Fig 3: Sequence Diagram

D. Activity Diagram
The flow of activity starts by fetching the dataset using pandas library. Now pre-processing is done where certain operations are done on the data to retain only the useful information needed. Pattern discovery and analysis is done by applying the

Random Forest Classifier algorithm and then the confusion matrix and accuracy of the algorithm are displayed.

Fig 4: Activity Diagram

V. WORKING PRINCIPLE

Jupyter Notebook is launched using command prompt by using the command jupyter lab. Here numpy and pandas libraries are imported to fetch the bank user dataset and perform several operations on it. After that several process such as pre-processing, pattern analysis and pattern discovery are done. During pre-processing irrelevant fields in the data set are removed. In Pattern analysis stage the data is divided into two sets for testing and training purpose. Using Random Forest Classifier algorithm the pattern discovery stage is done and the accuracy of the algorithm is displayed. Using webpage we can predict whether the customer will stay or leave the bank.
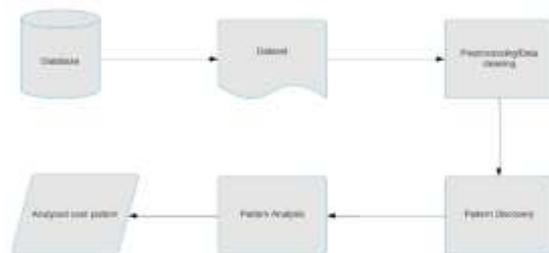

Fig 5: Block Diagram

A bank dataset with 10,000 unique customer data is used for this project. This data set consist of customer information such as row number, sur name, customer id, balance, geography, gender, age, active member, no of products, has credit card, expected salary, credit score and exited.

Using pandas library the dataset is imported. After that to remove unwanted fields pre-processing is done where the fields namely row number, customer id and sur name are removed. Now the string variables is converted into integer variable (label encoding process).

For pattern discovery the heatmap function is used which identifies the fields that highly influence the result. The heatmap function uses different colors such as green (high), red (low) and yellow (intermediate) to highlight the impact of different fields in the dataset. The fields identified using the heatmap function are age, geography and balance.

For pattern analysis first the fields are split into two sets i.e. training data and testing data. Now the scaling process in done on the data using standard scalar function. Finally Random Forest Classifier algorithm is applied to analyse the pattern in the dataset, the confusion matrix and accuracy is displayed.

VI. EXPERIMENTAL SETUP

For this experiment, Python IDE version 3.9 or higher is needed. Libraries and packages such as flask, matplotlib, numpy, pandas and sklearn should be installed. Also the Random Forest Classifier algorithm should be used for pattern identification.

VII. RESULTS

As a result, our smart wearable device can able to fetch all the values from the sensor and they can be viewed by the authorized person
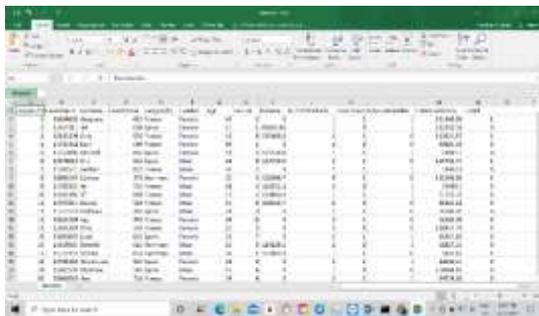
Fig 6: Data set
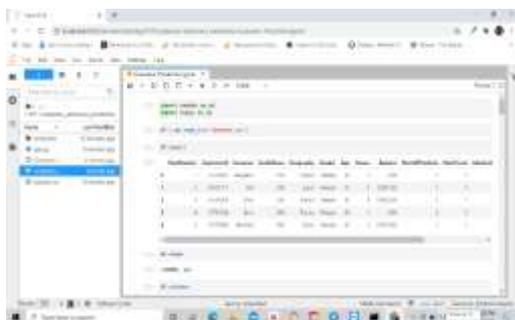
Fig 7: Importing the dataset

The data set is imported using pandas library and head function is used to display the first five rows of the data set.
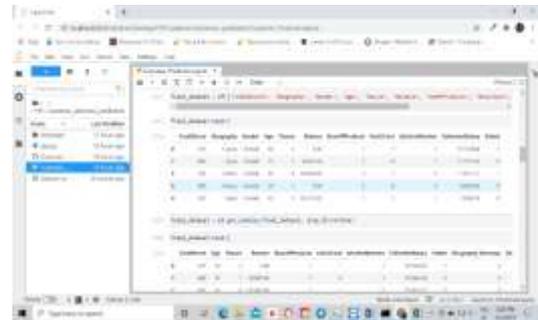

Fig 8: Pre-processing the data

From the dataset irrelevant fields such as sur name, customer id, row numbers are removed then the string variables are converted into integers variable using get_dummies.
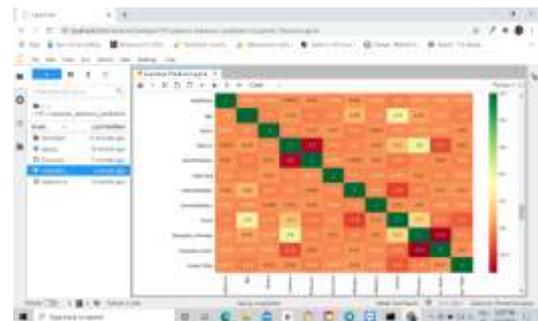

Fig 9: Heatmap

From the heatmap it is clearly visible that the fields age, balance and geography highly influence the result.
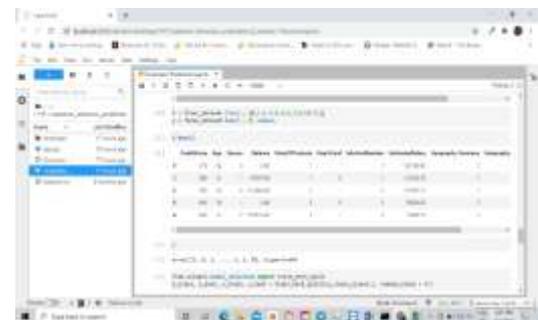
Fig 10: Splitting the dataset

The data set is split into x (input) and y (output).For training and testing purpose we have taken 20% of data for testing and 80% for training the data.
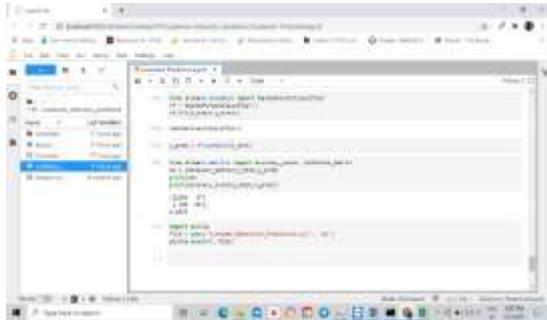
Fig 11: Result

The Random Forest Classifier(RFC) is applied to train and test the data and then the confusion matrix and the accuracy of the algorithm is displayed.

VIII. CONCLUSION

In this project we have designed a system to predict customer behaviour using random forest classifier for a bank user dataset. This algorithm takes a customer data as input and performs several operations like pre-processing, pattern discovery and pattern analysis on it and produces the predicted output. This algorithm will be very useful in identifying whether the customer will continue using the services provided by the bank or not. Thus the bank will be able to identify the users who might potentially leave the bank and identify the reason behind it and try to improve it.

IX. FUTURE WORKS

There are some ideas that we would like to implement in the future. As new algorithms are being developed frequently this project can be tested with those algorithms to see if the accuracy increases. As of now our system only predicts the results for one data at a time but in the future we would improve it to predict multiple results at once.

ACKNOWLEDGEMENT

REFERENCES

[1].   P. Prabhu, Umang, JyothiJayakumar, ChPhanindra Kumar (2018). IntelligentWearable Device for Coal Miners, International Journal of Engineering & Technology, pages – (677-680).

[2].   Jagadeesh R, Dr. R. Nagaraja (2017). IoT based Smart Helmet for unsafe event detection for mining industry, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume: 04 Issue: 01

[3].   Tarek Eldemerdash, Raed Abdulla, VikneswaryJayapal, ChandrasekharanNataraj, Maythem K. Abbas (2020). Iot Based Smart Helmet for Mining Industry Application, International Journal of Advanced Science and Technology, Vol. 29, No. 1pp. 373 – 387.

[4].   Pulishetty Prasad, Dr. K. HemaChandran, H. Raghupathi(2017).Zigbee Based Intelligent Helmet for Coal Mine Workers, InternationalJournal of Advance Research in Science and Engineering, ISSN:2319-8354, Volume no:6.

[5].   Ashish "Coalmine safety monitoring using Wireless sensor Networks", International Journal of Scientific Engineering and Technology (IJSET) Volume 2, Issue 10, October 2013

[6].   Wakode "Coalmine safety monitoring and alerting system", International research journal on Engineering and Technology (IRJET) Volume 4, Issue 3, March 2017

[7].   Dong "Coal Mine safety Monitoring system based on Zigbee and GPRS", Applied Mechanics and Material Volume 422,2013

[8].   Aarthi "Coal Mine safety Monitoring system using ARM 9", International Journal of Science and Research (IJSR), Volume3, Issue-11, November 2014

[9].   Dheeraj "IoT in mining for sensing, Monitoring and prediction of underground mines Roof support", conference on recent information and advancement technology 2018.

[10].   Madhu "Coalmine safety monitoring system", International journal of Mechanical Engineering and Technology (IJMET) Volume 8, Issue 12, December 2017

[11].   C. J. Behr, A. Kumar and G.P. Hancke (2016). A Smart Helmet for Air Qualityand Hazardous Event Detection for the Mining Industry, Conference Paper, DOI: 10.1109/ICIT.2016.7475079.

[12].   R.MUTHAIYAN, V.G.RAJARAMYA, A.BILOMIN RAMYA (2015). Implementation of Wireless Sensor in Coal Mine Safety System, IJCSEC-International Journal of Computer Science and Engineering Communications.Vol.3, Issue 1, Page.613-620, ISSN: 2347–8586.

IJAEM