Optimizing profit by retaining customers using machine learning techniques

M. Robinson Joel[1] and M.V. Srinath[2*]

Article History

Received: 04-01-2021

Revised and Accepted : 05-06-2021

Published: 26-06-2021

https://doi.org/10.56343/STET.116.014.004.007

http://stetjournals.com

Abstract

In recent years, report points out that the customer churn in banking sector have seen a spike. Churn can be found in various categories. It's a known fact that the cost of customer acquisition is far greater than cost of customer retention. The aim of this paper is to investigate machine learning based techniques for churn by predicting the results in best accuracy. To capture the various pieces of information like data cleaning, different variable identification, missing value, analyse the data validation and data visualization on a given set of data set using dataset analyzer supervised machine learning technique denoted as SMLT. Here we propose the best accuracy, by comparing the various machine learning algorithm on a specific dataset like the credit card dataset we have taken as a sample. We do the evaluation classification report, identify the confusion matrix and do the categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique.

Key words: Artificial Intelligence, customer churn, machine learning, optimization profit, SMLT

INTRODUCTION

Machine learning (Fig.1) is a concept used to predict the upcoming future from past data. In the field of artificial intelligence (AI), Machine learning (ML) (Shai and Shai, 2014) enables to learn the system. New developments have happened in computer programmes using machine learning when new data is exposed. Python is used to develop a new machine-learning algorithm. Machine Learning is the process of training the given data and predicting the given data by feeding them into the algorithm and in turn they produce the new test data. Supervised learning concept, unsupervised learning method and reinforcement learning concept are the different categories of machine learning.

Supervised machine learning is mainly used to train data and also predict the outcomes by providing labelled datasets as input to train algorithms. This supervised learning process mainly helps in many practical issues like in our mail account it segregate the incoming mails into the spam folder (Dada *et al.* 2019; Divya and Kumaresan 2014) and in primary as good mails.

Like supervised machine learning technique, this unsupervised learning is also a machine learning technique where these are not supervised by using training dataset. To find useful insights from data and like a human perform activities using their own experience like artificial intelligence. This unsupervised machine learning applies to uncategorized and unlabelled data to get a good result.

Like supervised and unsupervised machine learning, the next one is reinforcement learning noted as RL. People try to ought an action on the environment to maximize the result. Nowadays Data scientists implement many algorithms to find out the patterns using languages like python.

For the prediction of the given set of data into different classes, we are using classification. We denote classes as targets for some cases and labels or categories for some cases. For mapping function from the set of inputs denoted as X and Y respectively, we can form classification predictive modeling (Oleksandr et al 2020). Classification is a supervised learning method under machine learning and statistical analysis. Through this classification, new observations are

M. Robinson Joel

email: *sri_induja@rediffmail.com*

Department of Computer Science and Engineering, Ponnaiyah Ramajayam Institute of Science and Technology (PRIST), Chennai, Tamil Nadu, India.

[1]Department of Computer Applications, Sengamala Thayaar Educational Trust Women's College (Autonomous), Sundarakkottai, Mannargudi - 614 016, Tamil Nadu, India.

P - ISSN 0973 - 9157

E- ISSN 2393 - 9249

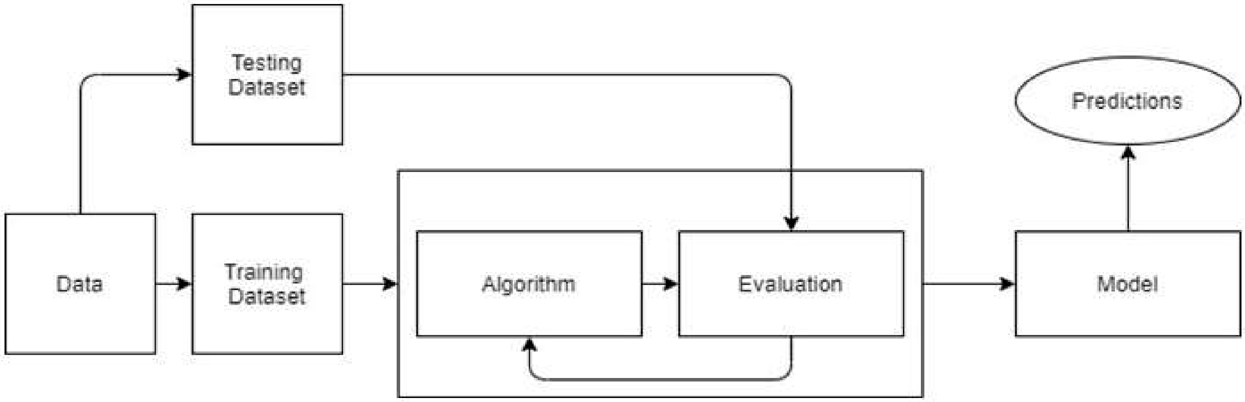

Fig. 1. Process of Machine Learning

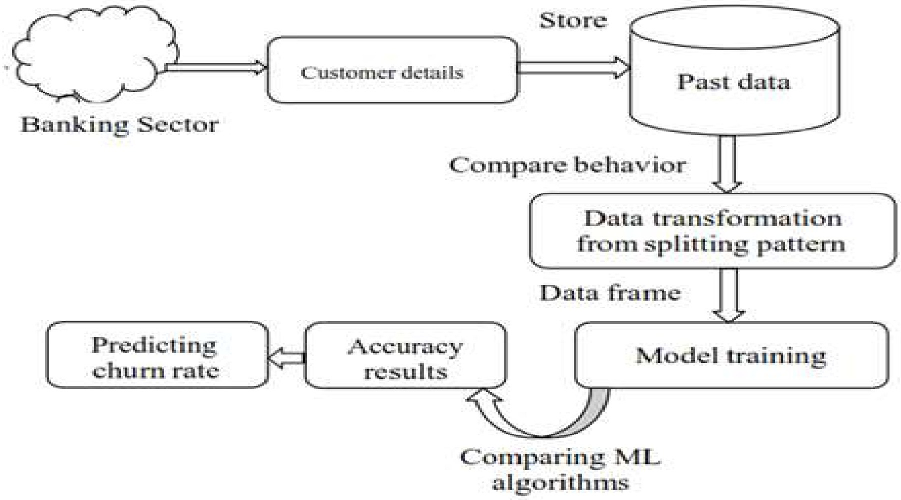

Fig. 2. Proposed architecture

derived from the data given as input to the computer programs. The test data may contain bi-class like female or male differentiation and sometimes spam or not. Similarly, they may be multi-class variables also available. The techniques like human handwriting recognition (Preetha *et al*, 2020), document classification (Kamran *et al*, 2019), human biometric identification (Besbes *et al*, 2008) (Kazi *et al*, 2012) speech recognition are good examples of classification techniques.

Churn is the significant problem in the business sector. The churn rate have been increased at a faster rate and it is the responsibility of banking department to control and reduce the churn rate. Churn prediction and customer identification are the major problems to the banking sector as there are tremendous amount of churn data that exist. There is a need of technology through which the case solving could be faster. The problem made us to go for a research is how can solve a churn case made easier. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster.

RELATED WORK

In the present competitive market of telecom domain, churn prediction is a significant issue of the Customer Relationship Management (CRM) to retain valuable customers by identifying a similar groups of customers and providing competitive offers/services to the respective groups. Irfan Ullah *et al*.(2019) proposed customer churn model for data analysis. In this study, a customer churn model (Irfan *et al*, 2019) is provided for data analytics and validated through standard evaluation metrics. Finally, they provided guidelines on customer retention for decision-makers of the telecom companies. The study can be further extended to explore the changing behavior patterns of churn customers by applying Artificial Intelligence techniques for predictions and trend analysis.

Zhang *et al*.(2020) mainly focused on a new interesting intra operator customer churn problem that some customers switch their telecommunications services from 4G to 3G/2G. In order to achieve profit maximizing classification, in which the effect of individual variances among customers in terms of their monthly fee is considered, they set classification threshold for each customer and make optimal decision each time. Then their assign 4G customers switching scores to reflect their current switching likelihood. Through establishing a GBDT-Gradient Boosting Decision Tree- based regression model, they demonstrate the predictability of switching behaviors of 4G customers, which lay a foundation to design evaluation test for 4G service plans. Their proposed framework solve this intra-operator customer churn problem well, and provides insight into the reasonable design of 4G service plans.

Eunjo Lee *et al.*(2020) proposed a churn analysis. The purpose of churn analysis is to prevent losses caused by user churn. Consequently, churn prediction is required to not only improve prediction accuracy but also maximize expected benefits. There are three main features of their proposed method (Divya and Kumaresan, 2014). First, they define churn *via* analyzing the access patterns of users. Second, long term loyal customers with a high benefit are identified and used for churn prediction. After that they calculate the expected profit per user *via* cost-benefit analysis and optimize the prediction model. Finally, they consider that the profit estimation method they used will be necessary for other researchers to analyze user churn in online game services. However, there is still the limitation that sufficient verification has not been achieved in practice. They plan to verify and improve the proposed method in subsequent studies rigorously.

In their paper (Eunjo *et al*, 2019) they propose a competition framework. For this, game data was collected to do mining using commercial game log data. Between the prediction window and the training data, there was a long time span provided by the researchers which result in the minimum time required to execute the churn prevention methods. Second, test sets include a change of business model to drive concept drift issue. Third, they provided only log data of loyal users for the competition. According to our experiments, they tend to be more difficult to predict churn than others, however they are more valuable in business.

Ahni *et al.*,(2020) proposed comparing the churn prediction analysis techniques using log data. In the field of insurance, games, and management similarly, like in the fields of Internet services Churn analysis is used. Churn Prediction Models of this paper used deep learning for churn prediction with data timestamps in the order of seconds or with vast amounts of customer log data in total. In such case, feature engineering techniques for processing logs have a significant effect on model performance enhancement. Also the deep learning model can learn customer's behavioral patterns from vast amount of data by layerwise stacked neurons structure. Therefore, given minute timestamps and abundant observations, applying this data to deep learning algorithms for the generation of latent features is expected to produce better performance than conventional churn prediction models.

Prasad Babu Gowd and Robinson Joel (2018) proposed a system that converts the buyer's reviews in the form of voice format into text format. This conversion happened to base on the speech recognition module. The customer's reviews are stored in the cloud and fed into the sentimental analysis domain to get neutral reviews, positive reviews and negative reviews. These all reviews help the customers to find and take decisions to go for the product. The author proposed mainly a voice or speech-based model to collect the speech reviews or feedback and its data are fed into a mining algorithm. The author assigned machine learning to find out the value or weightage of each and every word.

Sasikala (2020) proposed a new methodology named DLMNN abbreviated as a Deep Learning Modified Neural Network, a sentiment analysis of online products customer's reviews. Also, the author proposed the Improved Adaptive Neuro-Fuzzy Inference System denoted as IANFIS methodology on the online products to be predicted. The author analyzed performances of both methodologies and got a good result. These Deep Learning Modified Neural Networks are proposed on the scenarios such as Content-based, Grade based, and Collaboration based of the customer's review analysis.

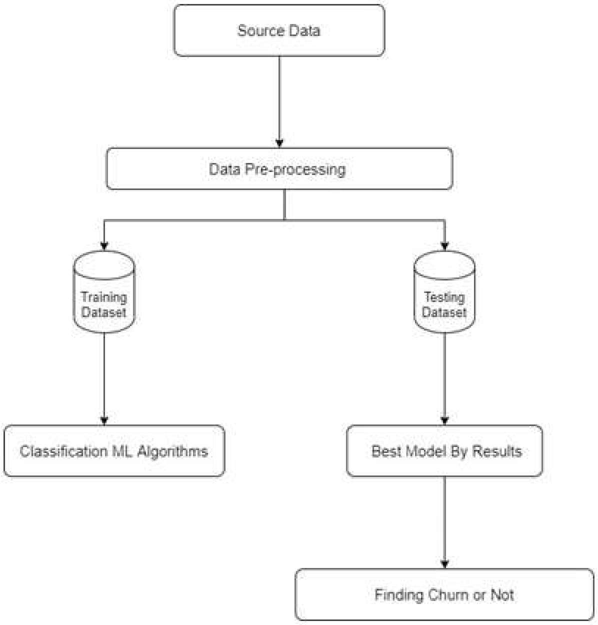

Fig. 3. Data Preprocessing model

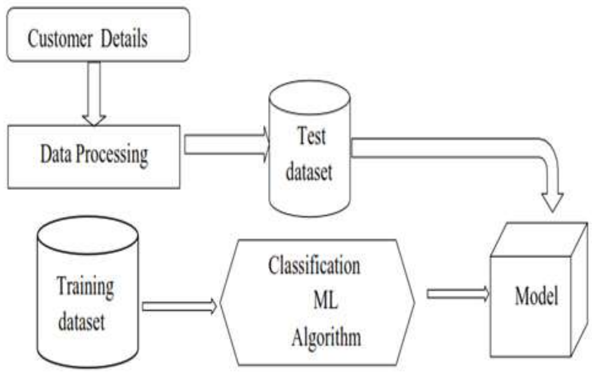

Fig. 4. Architecture of Proposed Model

Singla *et al.* (2017) mainly focus on the different features of the mobile phones available on e-commerce websites and their reviews. The authors collect both the customer's review data and the manufacture's view on the product as the dataset. The authors conduct the sentiment analysis on the three mobile brands namely Samsung, Apple and BLU.

Proposed Work (Fig. 2)

The proposed model have the capability to identify the churn customers early. It finds the reasons behind churn to avoid loss of customers and provides measures to retain them. To handle this limitation, multiple algorithms are used and the best classifier model is selected for retention. It can support companies to identify, predict and retain churning customers, help in decision making and CRM. We used a number of machine learning algorithms for churn and non-churn classification on a large dataset of the banking sector. We observed that the Random Forest (RF) algorithm produced better accuracy of 96.4% as compared to other machine learning algorithms. It can work efficiently on huge data. Improves profit by pruning the reason for churn. Helps in decision making with the help of useful insights.

Machine learning is a computer system's method of learning by way of examples. There are many machine learning algorithms available to users that can be implemented on datasets. Churn Prediction ways: To utilize the resources, identify the hotspots of churns and allocate vigilante resources such as customer satisfaction, requests, wants etc. reschedule patrols according to the vulnerability of a churn. Through that we avoid churns and ensure better visualization through avoiding happening churns such as customer loss, attrition etc.

The data set collected for predicting churn is split into a Training set and a Test set (Figs. 3 and 4). Generally, 8:2 ratios are applied to split the Training set and Test set. K-Nearest Neighbor (KNN), Random Forest, and Decision tree algorithms are applied to the training set *via* data model to get the accurate result, which leads to perfect prediction.

It creates cells to freely to explore the given data so that it should not perform too many operations in each cell. One option is, that it can take with this is to do a lot of explorations in an initial notebook. These don't have to be organized, but make sure you use enough comments to understand the purpose of each code cell. Then, after done with your analysis, create a duplicate notebook where it will trim the excess and organize steps so that you have a flowing, cohesive report and make sure that informed on the steps that are taking in your investigation. Follow every code cell, or every set of related code cell, with a markdown cell to describe to the reader what was found in the preceding cell.

Churn forecasting or predictive policing is based on large amounts of data collected from previous customers. It uses algorithms and other methods to help banking sectors to handle and share observations so that better early warning systems can be created to ensure more satisfaction to customers.

Churns such as customer dissatisfaction, loss, and much more can be predicted through churn patterns within a neighborhood or community to better prevent churn in the future or locate resources in much-needed sectors to handle churn it occurs.

The well known supervised machine learning algorithm is K-Nearest Neighbor. It does the closest K number and gives the common class as prediction.

One of the best supervised machine learning algorithms is Logistic Regression used to predict the target variable. In this algorithm, there are only two possible classes known as dichotomous variables. That the dependent variable is denoted as a binary value. That binary value will be denoted as 1 when it is a success or yes and denoted as 0 when it is a failure or no.

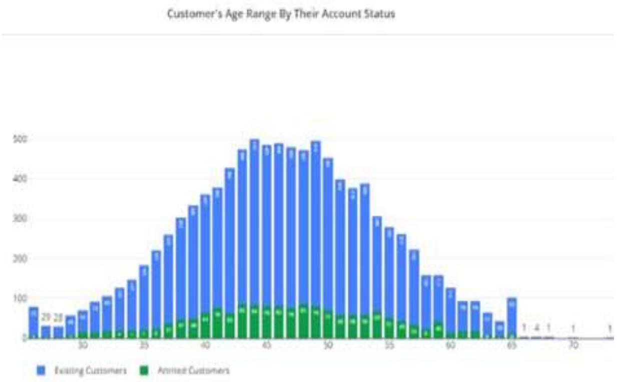

Fig. 5. Customer's age range by their account status

One of the famous machine learning algorithms is Random Forest which also belongs to the supervised learning algorithm. These random forest machine learning algorithms are used as classification processes and regression problems in machine learning. It is also termed as the concept of ensemble learning since it uses multiple classifiers to solve a problem and output the good result.

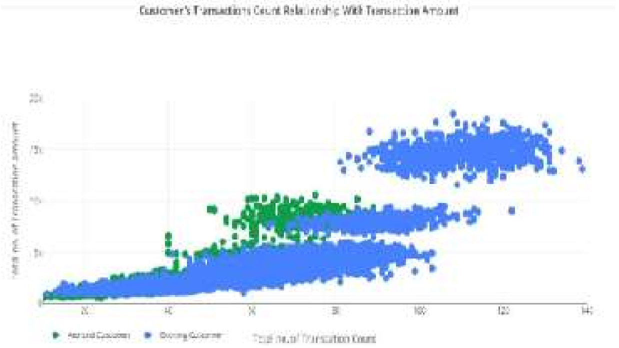

Fig. 6. Customer's credit limit relationship with average utilization ratio

Table 1. Algorithms and their accuracy

Method	Accuracy
Decision tree algorithms	92%
K-Nearest Neighbor (KNN)	93%
Random Forest	96.40%

Figure 5 shows the Customer's age range by their account status from the dataset of the e-commerce. Here the age-wise processing and cleaning happened.

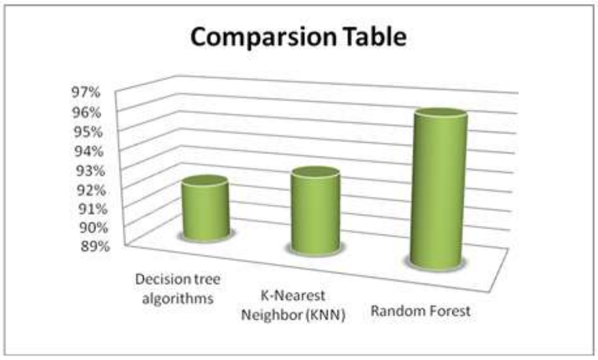

Fig. 7. Comparison table of the algorithms

Figure 6 shows the Customer's credit limit relationship with average utilization ratio from the dataset of the e-commerce. Here the Customer's credit limit relationship with average utilization are processed and cleaning happened.

Table 1 gives the accuracy of the machine learning algorithms like decision tree algorithm, K-Nearest Neighbor (KNN) and Random forest.

Figure 7 shows the accuracy of the machine learning algorithms like decision tree algorithm, K-Nearest Neighbor (KNN) and Random forest.

CONCLUSION AND FUTURE WORK

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is found out. This brings some of the following insights about churn rate. It has become easy to find out relation and patterns among various data. It mainly revolves around predicting the type of churn which may happen if we know the location of where it has occurred in the credit card department. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation. Data visualization generated many graphs and found interesting statistics that helped in understanding customer churns datasets that can help in capturing the factors that can help in keeping customers safe. Credit card department wants to automate the detecting of the churn from eligibility process (real time) based on the churn rate of areas. In future, we can develop some web applications and applications related to desktop to broadcast the predicted results. In Future, this work can be modified by using Machine Learning models for Forecasting churn, as the data points will sufficiently increase to apply ML models and increase the accuracy of the predictions.

REFERENCES

Ahn .J, J. Hwang, D. Kim, H. Choi and S. Kang, A Survey on Churn Analysis in Various Business Domains. *IEEE Access*, 8:220816-220839, 2020, doi: 10.1109/ ACCESS.2020.3042657.
https://doi.org/10.1109/ACCESS.2020.3042657

Baker, J. L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shgughnessy, 2009. Research developments and directions in speech recognition and understanding, part i, *IEEE Signal Process. Mag.*, 26(3):75–80.
https://doi.org/10.1109/MSP.2009.932166

Besbes, F.,Trichili, H., and Solaiman, B. 2008. Multimodal biometric system based on Fingerprint identification and Iris recognition In: Proc. 3rd Int. *IEEE Conf. Inf. Commun. Technol.: From Theory to Applications (ICTTA 2008)*, 1-5. DOI:10.1109/ICTTA.2008.4530129. https://doi.org/10.1109/ICTTA.2008.4530129
https://doi.org/10.1109/ICTTA.2008.4530129

Dada EG, Bassi JS, Chiroma H, Abdulhamid SM, Adetunmbi AO and Ajibuwa OE, 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*. 10;5(6):e01802. doi: 10.1016/ j.heliyon.2019.e01802. PMID: 31211254; PMCID: PMC6562150.
https://doi.org/10.1016/j.heliyon.2019.e01802

Divya .S, And T. Kumaresan, 2014. Email Spam Classification Using Machine Learning Algorithm. *Int. j. innov. res. comput. commun.*, 2(1).
https://doi.org/10.1109/TG.2018.2871215

Eunjo Lee, Boram Kim, Sungwook Kang, Byungsoo Kang, Yoonjae Jang and Huy Kang Kim. 2020. Profit Optimizing Churn Prediction for Long-term Loyal Customer in Online games
https://doi.org/10.1109/TG.2018.2888863
IEEE Trans. Games (IF1.851), Pub Date : 2020-03-01, *DOI: 10.1109/tg.2018.2871215*

Eunjo Lee, Yoonjae Jang, Du-Mim Yoon, Jihoon Jeon, Seong-il Yang, Sang-Kwang Lee, Dae-Wook Kim, Pei Pei Chen, Anna Guitart, Paul Bertens, Africa Perianez, Fabian Hadiji, Marc Muller, Youngjun Joo, Jiyeon Lee, Inchon Hwang and Kyung-Joong Kim,Game Data Mining Competition on Churn Prediction and Survival Analysis using Commercial Game Log Data, *IEEE Trans Games* (IF1.851), Pub Date : 2019-09-01, *DOI: 10.1109/tg.2018.2888863*
https://doi.org/10.1109/ACCESS.2019.2914999

Huang, X., Deng, L.: An overview of modern speech recognition. In: Indurkhya, N., Damerau,. F.J. (eds.) Handbook of Natural Language Processing, 2nd edn. Boca Raton, FL, USA: CRC, Taylor and Francis

Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, SailUl Islam, Sung Won Kim, A Churn Prediction Model Using RandomForest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification I Telecom Sector. *IEEE access*.7(1):60134-60149.
https://doi.org/10.1109/ACCESS.2019.2914999

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu Laura Barnes and Donald Brown. 2020. Text Classification Algorithms: A Survey, *arxiv.org*,Received: 22 March 2019; Accepted: 17 April 2019;

Kazi M. M., Rode Y.S. ,Dabhade S.B., Al-dawla N.N. H.,Mane A.V., Manza R.R., Kake K.V. 2012. Multimodal Biometric System Using Face And Signature : A Score Level Fusion Approach, *ACR*. 4(1):99-103

Oleksandr Dorokhov, Liudmyla Dorokhova, Lyudmylamalyarets and Iryna Ushakova. 2020. Customer Churn Predictive Modeling By classification Methods. *Bull. Transilv. Univ. Bras. III: Math. Inform. Phys.* 13(62):347-362https://doi.org/10.31926/but.mif.2020.13.62.1.26
https://doi.org/10.31926/but.mif.2020.13.62.1.26

Preetha S, Afrid I M, Karthik Hebbar P and Nishchay S K. 2020. Machine Learning for Handwriting Recognition, *IJC-GSSRR*. 38(1): 93-101

Prasad Babu Gowd. P.N.V.V. and M.Robinson Joel. 2018. User Satisfaction Assessment Extraction From Vocal Reviews Using Speech Recognition. *IJICS*, 5(3):89-98.

Sasikala, P. and Mary Immaculate Sheela, L. 2020. Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS. *J. Big Data* 7:33. https://doi.org/10.1186/s40537-020-00308-7

Shai Shalev-Shwartz and Shai Ben-David 2014. Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, U.K. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning
https://doi.org/10.1017/CBO9781107298019

Singla Z, Randhawa S, and Jain S. 2017. Statistical and sentiment analysis of consumer product reviews. *In:* 8th *International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.7(33):1–20.

Zhang, Y., He, S., Li, S., & Chen, J. 2020. Intra-Operator Customer Churn in Telecommunications: A Systematic Perspective. *IEEE T. on Veh. Technol.*, 69:948-957.
https://doi.org/10.1109/TVT.2019.2953605