customer attrition prediction

**AProject reporton**

**CustomerAttritionPrediction**

ADissertationsubmitted inpartialfulfillment oftheacademicrequirementsforthe awardofthedegree.

**BachelorofTechnology**

**In**

**ComputerScienceandEngineering**

Submittedby

**M.Priyanka**

**(20H51A0568)**

**G.Nishith Reddy**

**(20H51A05K7)**

**G.Sai Karthik**

**(20H51A05K8)**

Undertheesteemedguidanceof

**Ms.T Adarana**

Assistant Professor 

### DepartmentofComputerScienceandEngineering

**CMRCOLLEGEOFENGINEERINGANDTECHNOLOGY**

(AnAutonomousInstitutionunderUGC&JNTUH,ApprovedbyAICTE,PermanentlyAffiliatedtoJNTUH,AccreditedbyNBA.)

KANDLAKOYA,MEDCHALROAD,HYDERABAD-501401.

##### 2020-2024

Customer attrition prediction

**CMRCOLLEGEOFENGINEERING&TECHNOLOGY**

KANDLAKOYA,MEDCHALROAD,HYDERABAD–501401

**DEPARTMENTOF COMPUTERSCIENCEANDENGINEERING**



## CERTIFICATE

This is to certify that the Major Project phase I report entitled **" CUSTOMER ATTRITION PREDICTION "** being submitted by **M.Priyanka(20H51A0568),G.Nishith Reddy(20H51A05K7)** ,**G.Sai Karthik(20H51A05K8)** in partialfulfillmentfortheawardof**BachelorofTechnologyinComputerScienceandEngineering** is a record of bonafide work carried out his/her under my guidanceandsupervision.

TheresultsembodiesinthisprojectreporthavenotbeensubmittedtoanyotherUniversityorInstitutefortheawardofanyDegree.

##### Ms.T.Adarana Dr.Siva Skandha Sanagala

**AssistantProfessor Associate ProfessorandHOD**

**Dept.ofCSE Dept.ofCSE**

Customer attrition prediction

## Acknowledgement

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to**Ms.T.Adarana, Assistant Professor** , Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

**M.priyanka (20H51A0568)**

**G.Nishith Reddy (20H51A05K7)**

**G.Sai Karthik(20H51A05K8)**

**TABLEOFCONTENTS**

|  |  |  |
| --- | --- | --- |
| **NO.** | **TITLE** | **PAGENO.** |
|  | **LISTOFFIGURES**  **LISTOFABBREVIATIONS** | ii |
|  | **ABSTRACT** | 3 |
| **1** | **INTRODUCTION** |  |
|  | 1.1 Motivation | 4 |
|  | 1.2 Objective | 5 |
| **2** | **BACKGROUNDWORK** |  |
|  | 2.1 DomainIntroduction | 6 |
|  | 2.2 Logistic Regression | 6 |
|  | 2.3 KNN classifier | 8 |
|  | 2.4 The random forests algorithm | 9-11 |
| **3** | **EXISTINGSYSTEM** |  |
|  | 3.1 VGG | 13-14 |
|  |  |  |
|  | 3.2 InceptionV3 | 15 |
|  |  |  |
|  | 3.3 Densenet | 16 |
|  |  |  |
| **4** | **RESULTSANDDISCUSSION** | 18-20 |
| **5** | **CONCLUSION** | 22 |
|  | **REFERENCES** | 23 |

**ABSTRACT**

Customer attrition, often referred to as customer churn or customer turnover, poses a significant challenge for businesses across various industries. Predicting and mitigating customer attrition is of paramount importance in today's competitive marketplaces, where retaining existing customers is often more cost-effective than acquiring new ones. This abstract provides a concise overview of the key elements in customer attrition prediction, highlighting the significance of data analytics and machine learning in this context.

Customer attrition prediction involves the application of advanced data analytics techniques to identify customers who are at risk of leaving a business, and it plays a crucial role in devising strategies for retaining them. Leveraging historical customer data, including transactional information, demographics, and customer interactions, predictive models are constructed to forecast the likelihood of customer attrition. Various algorithms, such as logistic regression, decision trees, random forests, and neural networks, have been successfully employed to address this challenge.

The benefits of accurate customer attrition prediction are multifold. First and foremost, it enables proactive customer retention efforts, allowing businesses to intervene and address the concerns of at-risk customers before they defect. Additionally, it aids in optimizing marketing strategies, as companies can allocate resources more efficiently by targeting high-risk customers with tailored retention campaigns. This, in turn, leads to improved customer satisfaction and long-term profitability.

Furthermore, data-driven customer attrition prediction is a dynamic field that continues to evolve. Innovations in artificial intelligence and machine learning have enabled the development of more sophisticated models that can capture subtle patterns in customer behavior. Moreover, the integration of real-time data sources and feedback loops has enhanced the accuracy and timeliness of predictions, making it possible to respond to customer churn in near real-time. In conclusion, customer attrition prediction is a vital tool for businesses seeking to maintain sustainable growth. By harnessing the power of data analytics and machine learning, organizations can not only identify at-risk customers but also take proactive steps to retain them. As the field continues to advance, businesses must remain agile and adapt to the changing landscape of customer attrition prediction to ensure their long-term success.

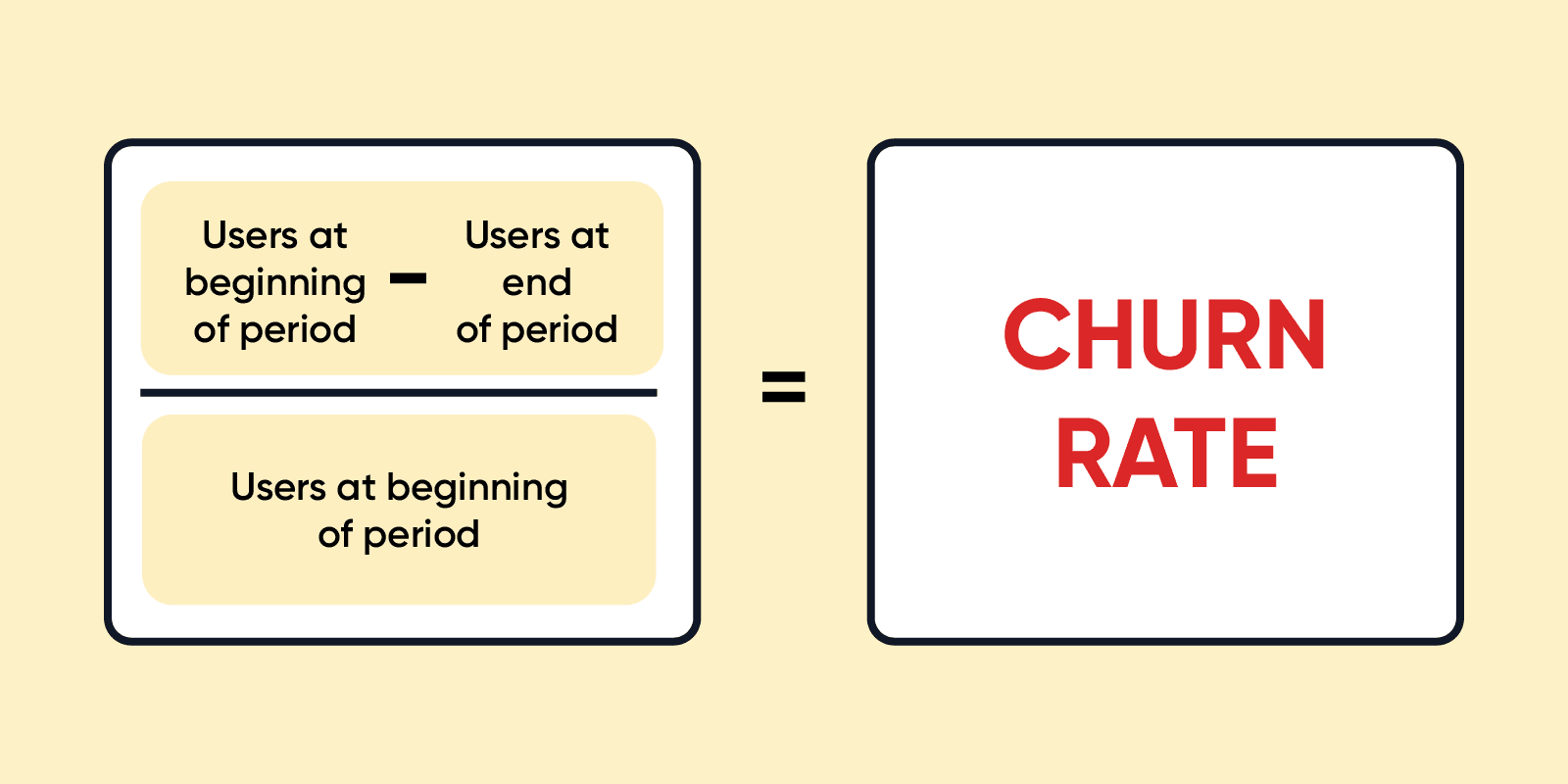
1

# CHAPTER1INTRODUCTION

#### INTRODUCTION

Customer attrition, the phenomenon of customers discontinuing their relationship with a business, is a pervasive challenge faced by companies across various industries. Also known as customer churn or turnover, it has significant financial implications and can erode long-term profitability. In today's competitive business landscape, retaining existing customers is often more cost-effective than acquiring new ones. To address this challenge, organizations are increasingly turning to data analytics and machine learning to predict customer attrition and implement proactive strategies for customer retention.

In the ever-evolving world of e-commerce, businesses are constantly seeking innovative strategies to thrive in an increasingly competitive marketplace. One of the most formidable challenges they face is customer attrition, often referred to as customer churn. In this context, customer attrition signifies the loss of valuable customers who cease engaging with an e-commerce platform. As e-commerce continues to gain prominence and reshape consumer habits, understanding and addressing the factors behind customer attrition has become paramount for companies aiming to ensure sustainable growth and profitability.



* 1. **Motivation:**

Intense Market Competition: The e-commerce industry is characterized by fierce competition, where consumers are just a few clicks away from exploring alternative products or brands. This competitive environment magnifies the importance of retaining existing customers, as it is often more cost-effective than acquiring new ones. Motivated by this competition, businesses are increasingly focusing on understanding and mitigating customer attrition.

Data-Driven Decision-Making: E-commerce platforms generate vast amounts of data, including customer browsing behavior, purchase history, and feedback. The motivation to analyze this data for customer attrition prediction arises from the desire to convert this information into actionable insights. Leveraging data analytics and machine learning, e-commerce businesses can harness this wealth of data to gain a competitive edge.

Customer Lifetime Value: An essential motivation for tackling customer attrition is the concept of Customer Lifetime Value (CLV). E-commerce companies recognize that each customer represents a potential stream of revenue over an extended period. The longer a customer stays loyal, the more valuable they become. Thus, predicting and addressing attrition is crucial for maximizing CLV and, consequently, overall profitability.

## Objective:

Early Identification of At-Risk Customers: The primary objective in the realm of e-commerce customer attrition is to identify customers at risk of churning at an early stage. By leveraging historical transaction data, browsing behavior, and other relevant factors, e-commerce platforms can develop predictive models to flag customers with a higher likelihood of attrition. Early identification empowers businesses to take preventive actions, such as personalized recommendations, loyalty rewards, or proactive customer support.

Tailored Customer Engagement: E-commerce businesses aim to tailor their engagement strategies based on the predicted attrition risk of individual customers. Instead of deploying generic marketing campaigns, companies can customize their approaches to re-engage and retain customers effectively. This personalization enhances customer satisfaction and loyalty, ultimately reducing churn.

In conclusion, the motivation behind studying e-commerce customer attrition is driven by the fierce market competition, the vast availability of data, and the desire to maximize Customer Lifetime Value. The objectives revolve around early identification of attrition risks and the implementation of personalized engagement strategies to enhance customer retention. As e-commerce continues to evolve, understanding and mitigating customer attrition remains an integral aspect of e-commerce business strategies, contributing to sustained growth and long-term success.

# CHAPTER 2BACKGROUNDWORK

#### DOMAININTRODUCTION:

E-commerce attrition prediction, within the domain of machine learning, represents a compelling intersection of technology and commerce. As the e-commerce industry continues to thrive and evolve, the ability to forecast and mitigate customer attrition using cutting-edge machine learning techniques has become a critical focus for businesses. This domain is a testament to the transformative power of data-driven insights and predictive models in enhancing the sustainability and profitability of e-commerce operations.

In the e-commerce landscape, customer attrition, or churn, signifies the departure of customers who once engaged with a platform but have gradually reduced their interaction or abandoned it altogether. Understanding and addressing this phenomenon is vital, as it directly influences a company's bottom line and market positioning. Machine learning brings a unique set of tools and methodologies to this challenge, enabling the prediction and proactive management of customer attrition in a highly dynamic and competitive environment.

E-commerce platforms are treasure troves of data, where every customer interaction generates valuable information. This data encompasses transaction histories, browsing behavior, demographics, and feedback. Machine learning algorithms are tailored to harness this data effectively, identifying hidden patterns and trends that would be difficult to discern through traditional analysis. By leveraging these techniques, e-commerce businesses can unlock the secrets of customer attrition, enabling data-driven decision-making and personalized customer engagement.

#### ****Logistic Regression:-****

Classification techniques are an essential part of machine learning and data mining applications. Approximately 70% of problems in Data Science are classification problems. There are lots of classification problems that are available, but the logistics regression is common and is a useful regression method for solving the binary classification problem. Another category of classification is Multinomial classification, which handles the issues where multiple classes are present in the target variable. For example, IRIS dataset a very famous example of multi-class classification. Other examples are classifying article/blog/document category.

Logistic Regression can be used for various classification problems such as spam detection. Diabetes prediction, if a given customer will purchase a particular product or will they churn another competitor, whether the user will click on a given advertisement link or not, and many more examples are in the bucket.

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem. Its basic fundamental concepts are also constructive in deep learning. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables.

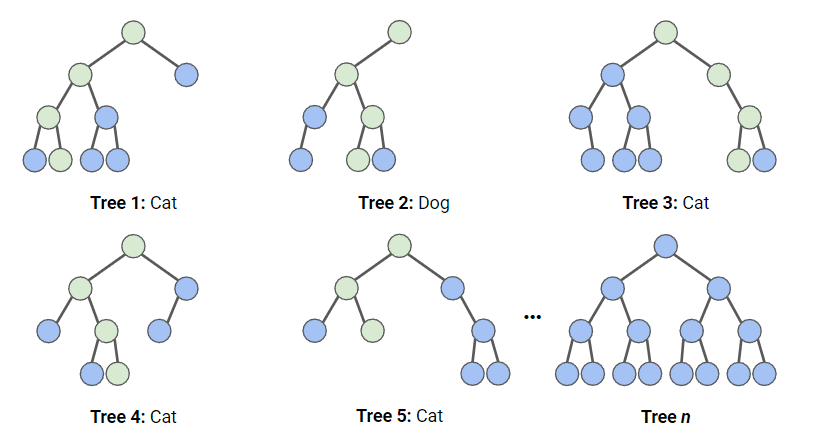
**2.3 KNN Classifier:**

K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won’t vote. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.

# 2.4 The Random Forests Algorithm

In a random forest classification, multiple decision trees are created using different random subsets of the data and features. Each decision tree is like an expert, providing its opinion on how to classify the data. Predictions are made by calculating the prediction for each decision tree, then taking the most popular result. (For regression, predictions use an averaging technique instead.)

In the diagram below, we have a random forest with n decision trees, and we’ve shown the first 5, along with their predictions (either “Dog” or “Cat”). Each tree is exposed to a different number of features and a different sample of the original dataset, and as such, every tree can be different. Each tree makes a prediction. Looking at the first 5 trees, we can see that 4/5 predicted the sample was a Cat. The green circles indicate a hypothetical path the tree took to reach its decision. The random forest would count the number of predictions from decision trees for Cat and for Dog, and choose the most popular prediction.



It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

# How does the algorithm work?

It works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.

# Random Forests vs Decision Trees

* Random forests is a set of multiple decision trees.
* Deep decision trees may suffer from overfitting, but random forests prevents overfitting by creating trees on random subsets.
* Decision trees are computationally faster.
* Random forests is difficult to interpret, while a decision tree is easily interpretable and can beconverted to rules.

# Advantages:

* Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
* It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
* The algorithm can be used in both classification and regression problems.
* Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
* You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

# Disadvantages:

* Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming.
* The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

# CHAPTER 3

# EXISTINGSYSTEM

# Logistic Regression

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:

Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:

Apply Sigmoid function on linear regression:

Properties of Logistic Regression:

* The dependent variable in logistic regression follows Bernoulli Distribution.
* Estimation is done through maximum likelihood.
* No R Square, Model fitness is calculated through Concordance, KS-Statistics.

# Linear Regression Vs. Logistic Regression

Linear regression gives you a continuous output, but logistic regression provides a constant output. An example of the continuous output is house price and stock price. Example's of the discrete output is predicting whether a patient has cancer or not, predicting whether the customer will churn. Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.

# K-Nearest Neighbors

KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure determined from the dataset. This will be very helpful in practice where most of the real world datasets do not follow mathematical theoretical assumptions. Lazy algorithm means it does not need any training data points for model generation. All training data used in the testing phase. This makes training faster and testing phase slower and costlier. Costly testing phase means time and memory. In the worst case, KNN needs more time to scan all data points and scanning all data points will require more memory for storing training data.

# How does the KNN algorithm work?

In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm. This is the simplest case. Suppose P1 is the point, for which label needs to predict. First, you find the one closest point to P1 and then the label of the nearest point assigned to P1.

Suppose P1 is the point, for which label needs to predict. First, you find the k closest point to P1 and then classify points by majority vote of its k neighbors. Each object votes for their class and the class with the most votes is taken as the prediction. For finding closest similar points, you find the distance between points using distance measures such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance. KNN has the following basic steps:

1. Calculate distance
2. Find closest neighbors
3. Vote for labels

# CHAPTER4

# RESULTSANDDISCUSSION

The main objective was to distinguish between churned and retained customers in addition to finding the associated attributes leading to churn. At first, it was observed that single male customers are having slightly higher probability of churn. In addition, Mobile preferred order category is related to customer churn as well. Furthermore, the churned customers are slightly higher in phone/Mobile phone preferred login device which might be caused by the E-commerce’s customer user experience phone version of the ecommerce. Also, it was found that churned customers are having higher mean in complain, city tier Number of addresses and number of registered devices. However, our study shows that satisfaction score is higher in churned customers which was not expected. On the other hand, Tenure, and count of number of orders is lower for churned customers which is reasonable.

# CHAPTER5

**CONCLUSION AND FUTUREWORK**

#### CONCLUSION:

E-commerce businesses are allocating huge amount of money to acquire new customers. However, customers lifetime depends on a lot of variables and this study was about building customer churn prediction model for e-commerce businesses. the dataset used for this project is foe leading e-commerce platform which was taken from Kaggle. The study started with exploratory analysis and data visualisations to increase our understanding to churned customers. It was noticed that churned customers associated with male gender, single marital status. Then, three different machine algorithms were applied to predict customer churn which are Decision tree, Logistic regression, and random forest. It was found that random forest has the best accuracy and kappa score at 93.5% and 0.75 respectively.

**REFERENCES:**

1. Zhang, D. (2015). Establishment and application of customer churn prediction model. Beijing Institute of Technology.

2. Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019, January). Churn prediction using neural network-based individual and ensemble models. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 634-639). IEEE.

3. Wu, X. J., & Meng, S. S. (2017). Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost. Industrial Engineering, 20(02), 99- 107.

4. Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia Cirp, 79, 403-408.

5. Shao, D. (2016). Analysis and prediction of insurance company’s customer loss based on BP neural network. Lanzhou University

6. Lu, N., Liu, X. W., & Lee, L. (2018). Research on customer value segmentation of online shop based on RFM. Computer Knowledge and Technology, 14(18), 275-276, 284.

7. Huang, J. (2018). A Comparative Study of Social E-Commerce and Traditional Ecommerce. Economic and Trade Practice, (23), 188-189.

8. Feng, X., Wang, C., Liu, Y., Yang, Y., & An, H. G. (2018). Research on customer churn prediction based on comment emotional tendency and neural network. Journal of China Academy of Electronics Science, 13(03), 340-345

9. Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Information Fusion, 54, 128-1