A Prediction of Customer Behavior using Logistic Regression, Naivesbayes Algorithm

Ruchita Atre
Research Scholar
Department of Computer Science Engineering
IES IPS Academy, India

Namrata Tapaswi
Head of Department
Department of Computer Science Engineering
IES IPS Academy, India

ABSTRACT
In past two decades e-commerce platform developed exponentially, and with this advent, there came several challenges due to a vast amount of information. Customers not only buy products online but also get valuable information about a product they intend to buy through an online platform. Customers share their experiences by providing feedback which creates a pool of textual information and this process continuously generates data every day. You can analyze the content in the form of comments, ratings and reviews. Consumers decide to buy a given product by looking at these reviews and reviews rating. Such content may be positive or negative reviews made by consumers who have used the product before. Our data analysis and multi-agent simulation demonstrate the feasibility of this framework. Perform behavioral analysis on data retrieved from Amazon reviews. These comments are divided into four categories: happy, up, down and rejection. When we analyze data to calculate the sense of user reviews, our goal is to use data-driven marketing tools such as data visualization, natural language processing, and machine learning models to help understand the organization's demographics. The system is developed based on classification algorithms includes Naïve Bayes, Logistic Regression. For each topic, the existing problems are analyzed, and then, current solutions to these problems are presented and discussed. The experimental results show that the proposed sentiment analysis method has higher precision, recall and F1 score. The method is proved to be effective with high accuracy on comments.

Keywords
Costumer Behavior, Logistic Regression, NaivesBayes , Customer Reviews, Data Mining, machine learning

1. INTRODUCTION
Consumers decide to buy a given product by looking at these reviews and reviews. Such content may be positive or negative reviews made by consumers who have used the product before. Machine learning algorithms can help us visualize and victories data. Natural language processing is a sub-field of machine learning that is used to analyze text and identify positive or negative reviews given by consumers. This is also called sensitivity analysis. The influence of traditional media can be estimated using established methods such as ratings. At the same time, the effect of the new strategy is difficult to quantify. Sentiment analysis can be used to extract meaningful insights from customer reviews and ratings. First, it explains the research background for analytical methods for text sentiment and expression of word vectors. Second, the details of the proposed method of analysis for comment sentiment are described. Third, experiments were performed and the results of the experiments were analyzed and discussed. Finally, the proposed method is summarized and the next research direction is introduced. Based on multiple dimensions such as emotional traits, negative traits, and emoticons, we use classification algorithms such as naive Bayes and logistic regression to predict the data. Naive Bayes algorithm is a classification technique based on Bayes' theorem, provided independence between predictors. Simply put, the Naive Bayes classification assumes that the existence of a particular function in a class has nothing to do with the existence of other functions. Logistic regression assumes a linear relationship between input variables and output. Data conversion of input variables that better reveal this linear relationship can produce a more accurate model. Data mining can be used to guide decision making and predict the impact of decisions aimed at discovering and consistently using beneficial knowledge in organizational data. Each CRM element can be sustained by dissimilar data mining models; these models usually include classification, association, clustering, regression, sequence detection or visualization.

1.1 Data Mining
One of the key features of CRM is its ability to distinguish between customer groups. Treating customers to the level of service they like forms the basis of segmentation. The main goal of customer segmentation is to identify and realize profitable areas and provide products or services that meet customers' common needs .The key areas where data mining can generate new knowledge include segmentation of customer databases based on demographics, purchasing patterns, geography, attitudes, and other variables. Data mining can be used to develop segmentation schemes based on current or expected / estimated value of customer. In order to prioritize customer care or market intervention based on the importance of each customer, these subdivisions are important. Segmentation insights are properly designed to develop a strategic roadmap so you can leverage the key profit-driven opportunities in each unique customer base. This can shorten the customer's purchasing cycle, increase costs, build higher customer loyalty, deepen cross-product penetration, or reduce service and support costs Data mining can provide customer insight, which is critical to setting up an effective CRM strategy. This can lead to more personalized customer relationships, improving customer satisfaction and beneficial relationships through data analysis. It can support “customized” and updated customer management at all stages of a customer’s life, from acquiring and establishing strong relationships to preventing staff turnover and gaining repeat customers. Complex customer segmentation enables organizations to target profitable customers, understand their needs, allocate resources and compete with competitors. Data mining can also be used to develop segmentation schemes based on customers' current or expected / estimated value. In order to prioritize customer care and market intervention

based on the importance of each customer, these segments are important. Data mining models can identify patterns that even the most experienced business people might miss. They can help fine-tune existing business rules, enrich, automate and standardize judgmental practices based on personal opinions and opinions. They include an objective, data-driven approach that minimizes subjective decision-making and simplifies time-consuming processes.

1.2 Objectives

Focusing on scientific way to estimate how a new service is accepted in society, we developed consumer behavior modeling framework. An accurate analysis of this user-generated content can be helpful to e-commerce organizations to gain insights and understand their consumers' intentions and requirements. Machine Learning Algorithms can help us plot accurate illustration representations of such consumer behavior. Machine learning classifiers include Naïve Bayes, Logistic Regression are used in the designing of the system.

• To analyze the consumer behavior(i.e) positive or negative

• To implement the machine learning algorithms

• To enhance the performance analysis.

• To explore the customers' insights from the data provided by Amazon, including features from text analysis.

• Classifying the customers according to their level of satisfaction (satisfied dissat-isfied).

1.3 Problem Statement

Analysis of customer feedback using text data focuses more on models of prediction. The text information contains not only the customers' sentiments, but also other important aspects of information. The focus on most of the researches is whether the text is subjective or objective, and if it is objective then it is positive or negative. Where there was work mainly focusing on sentiment analysis, the researchers opted the subjective approach to deal with the analysis and model building and predictions. However, there are still challenges to deal with the text data.

2. LITERATURE SURVEY

Soumi Ghosh et.al (2020) Predictive analysis of customers' buying behavior is an interesting and challenging task in modern life. Our goal is to commence concept of machine learning in depth using arbitrary shoe algorithms. In this article, a model has been projected to predict which cloud services have been purchased based on a number of factors. Use various parameters (such as advertising keywords, previously purchased cloud services, etc.) to create a random forest model and train our model. In order to implement the proposed model, the advertising log dataset has been obtained and necessary changes have been made to it. As a result, the proposed model provides customers with very accurate predictions. Five factors have been considered that will influence the customer's purchasing decision within cloud services, such as previous purchasing habits, seen advertising processes, customer placement, etc.

BoraBardüket.al (2020) is more advantageous for companies to preserve the interests of customers than to win new customers. Identifying customers who are at risk of churn and churn can enable companies to change their strategy in advance to achieve the best results. This study proposes a new method of non-contractual business. One of the most widely used methods in the literature for these types of companies is to use Beta Geometric Negative Binomial Distribution (BG / NBD) to model customer behavior. Although this method performs well in predicting the overall churn rate, it does not perform well in marking individual customers. This research aims to improve the BG / NBD model by incorporating machine learning into the decision-making process. In the proposed method, the mathematical definition of the BG / NBD model is used for feature of features and is used through decision trees. When testing two data sets with different characteristics, it can be observed that the proposed method can significantly improve the performance.

3. PROPOSED SYSTEM

Online reviews have become an important source of information for users before making an informed purchase decision. The main aspect of analyzing customer reviews on e-commerce sites. The main purpose of the project is to characterize and predict "early reviewers" to increase product sales. With the rapid progress of online social platforms or the availability of large amounts of data from social networks, research into innovative communication has been extensively researched in social networks.

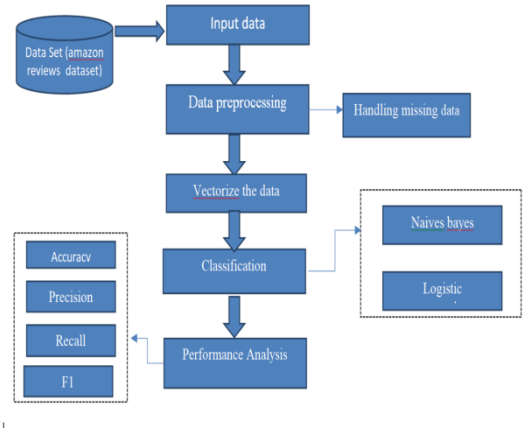

Figure 1 Flow chart

MODULES

- Data Selection and Loading
- Data Preprocessing
- Vectorize the data
- Splitting Dataset into Train and Test Data
- Classification
- Prediction
- Result Generation.

3.1 Data Selection and Loading

• Data selection is the process of selecting data for the detection of behavioral analysis.

• In this project, the dataset from Amazon is used to detect diseases.

• A dataset containing information about ID, name, comment, rating, etc.

• It provides useful information about the product, such as name, description, image URL, sales ranking, category, price, brand and the product most relevant to it.

3.2 Data Preprocessing

• Data preprocessing is the process of removing unnecessary data from the dataset.

• Failure to delete data

• Coding classification data Input data Data preprocessing Classification Performance Analysis Naivesbayes Data Set (amazon reviews dataset) Accuracy Precision Recall Handling missing data Logistic regression Vectorize the data F1 measure

• Failure to remove data: In this process, null values such as missing values and Nan values are replaced by 0.

• Removed missing and duplicate values and cleared any irregularities in the data. Coding of categorical data: categorical data is defined as a variable with a limited set of label values.

• Most machine learning algorithms require digital input and output variables.

3.3 Splitting Dataset into Train & Test Data

• Data sharing is the act of dividing available data into two parts that are normally used for crossvalidation purposes.

• Part of the data is used to develop predictable models, and the other part is used to evaluate the performance of the model.

• Dividing data into training and test sets is an important part of the evaluation of data mining models.

• Generally, when you share a data set in a training set and a test set, most of the data is used for training and a small portion of the data is used for testing.

3.4 Splitting Dataset into Train & Test Data

Naive Bayes is a classification technique based on Bayes' theorem, assuming independence between predictors. Simply put, the Naive Bayes classifier assumes that the existence of a particular function in a class has nothing to do with the existence of other functions.

Logistic regression is a linear algorithm (non-linear transformation of output). It is assumed that there is a linear relationship between input variables and output. Data conversion of input variables that better reveal this linear relationship can produce a more accurate model. \This is a process of predictable behavioral analysis from a data set.

• The project effectively predicts the data in the dataset by improving the performance of the overall prediction results.

3.5 Result Generation

The final result is generated based on the overall classification and prediction. Use some measures to evaluate the performance of this proposed method, e.g.

- Accuracy
- Precision
- Recall
- F-Measure.

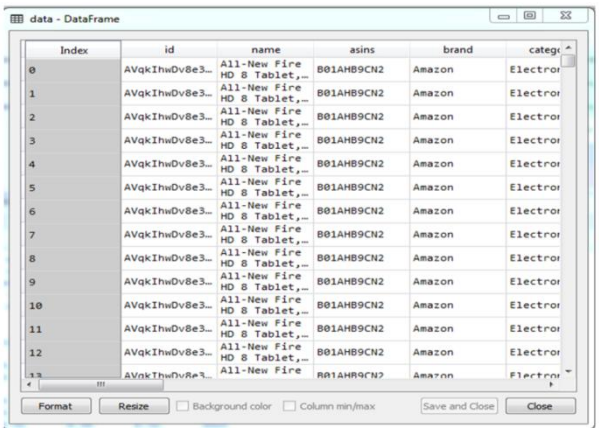

Figure 2 Database

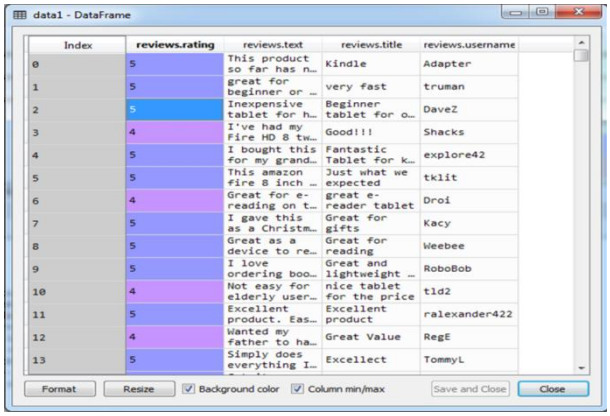

Figure 3 Dataframes

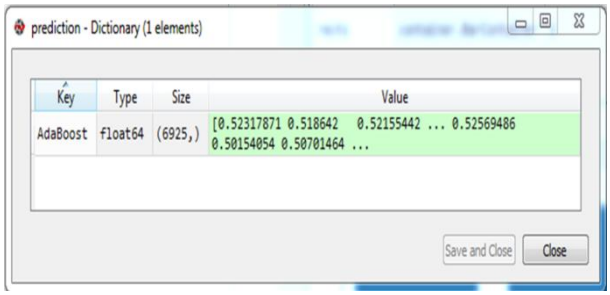

Figure 4 Prediction

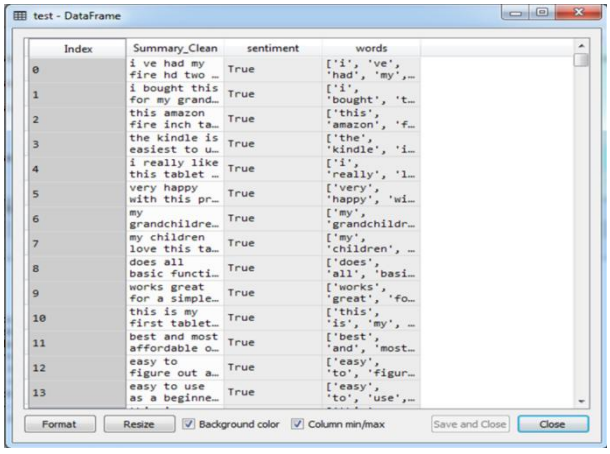

Figure 5 Test

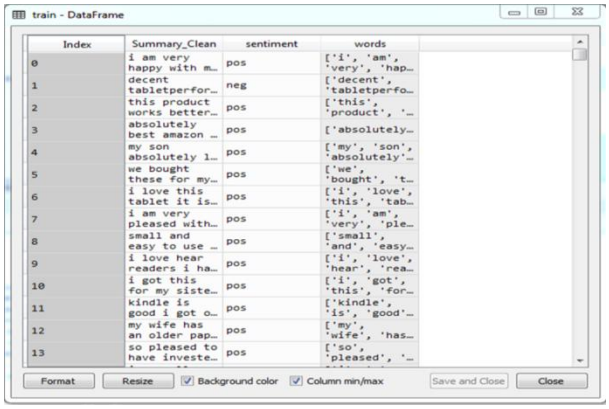

Figure 6 Train

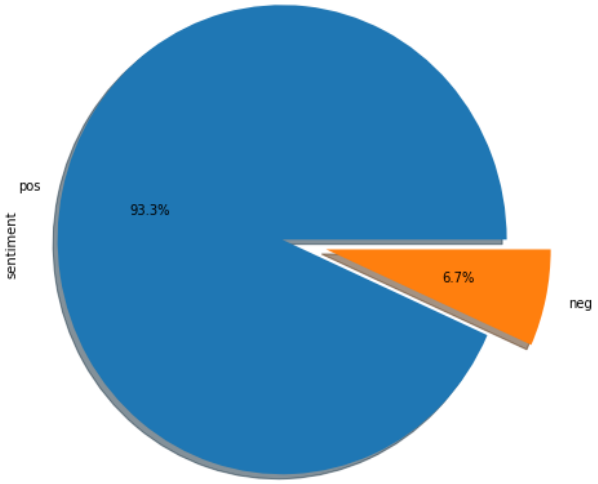

Figure 7 Sentiment analysis

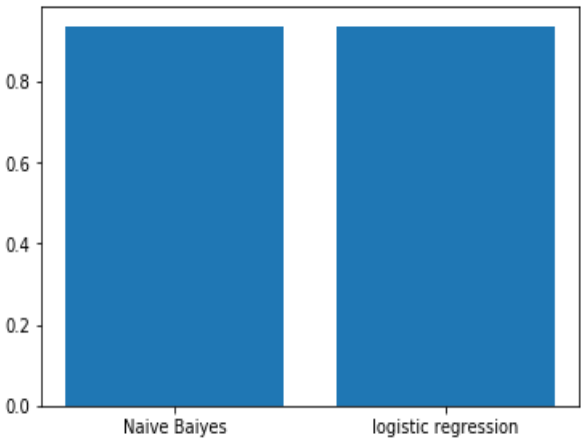

Figure 8 Classification analysis

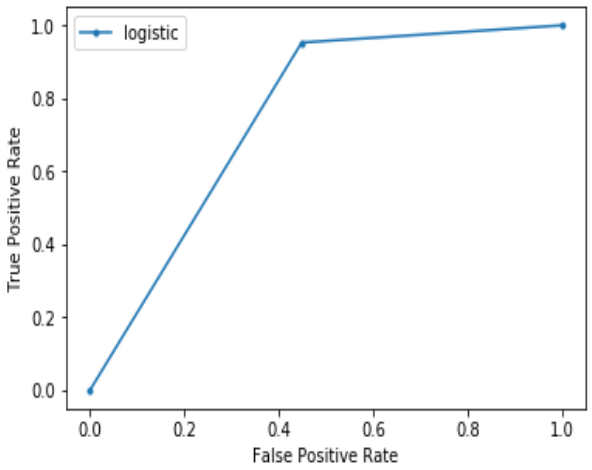

Figure 9 TPR and FPR Performance

Table 1 Comparison table with existing work

	Techniques	Accuracy (%)	Precision (%)	Recall(%)	F1score (%)
Existing work	Naivesbayes	93.41	92%	93%	92%
Proposed work	Logistic Regression	96.62%	0.96%	0.94 %	0.94%
	Naivesbayes	93.85 %	0.92 %	0.94%	0.92 %

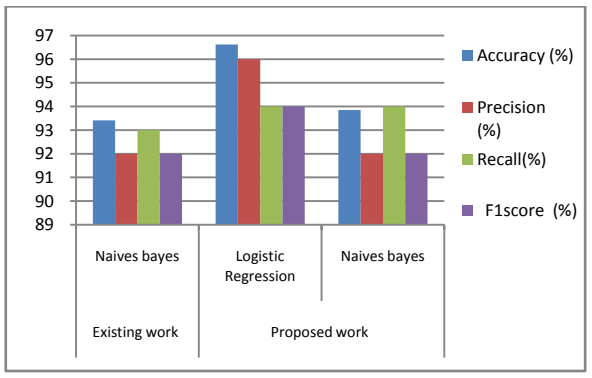

Figure 10 Comparison with existing work

4. CONCLUSION

The goal of this work is to help organizations understand their customers and combine targeted marketing techniques to increase their customer base and profits. Sentiment analysis helps us to evaluate consumers 'attitudes towards different products, which in turn helps us to analyze the products' performance in the market. We use two classification models to perform sentiment analyzes on reviews using natural language processing and achieve the best results in the naive Bayesian classifier and logistic regression. Logistic regression works well when the number of datasets is small and its output can be interpreted as a probability. Naive Bayes do

well on small datasets. Experimental results show that the proposed method is better than machine learning algorithms and achieves the highest performance in terms of accuracy, precision and F1 score. The results show that the proposed algorithm is better than other classifiers. The empirical analysis results provide evidence that our proposed method also has significant benefits in consumer behavior analysis functions. In the future, intelligent agents can be used to extend or modify the proposed optimization and classification algorithms to further improve performance. We will explore the application of more advanced methods of deep learning and possible combinations of machine learning Equal number of cases can be sampled for each rating and analysis can be made to classify customer satisfaction.

- Implementation of multinomial techniques to classify customer satisfaction by expanding classes. 89 90 91 92 93 94 95 96 97 Naivesbayes Logistic Regression Naivesbayes Existing work Proposed work Accuracy (%) Precision (%) Recall(%) F1score (%)

- Some objective terms are express enough about customer behavior when assessed with the emoticons.

- Analyze the difference when features are extracted using TD-IDF, Count Victories and Chi-square. Moreover, it can be implemented with multinomial classification approaches.

- Some unsupervised machine learning approaches to cluster features of each class and compare techniques.

5. REFERENCES

[1] SerhatPeker;AltanKocyigit;P. ErhanEren An empirical comparison of customer behavior modeling approaches for shopping list prediction 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) Year: 2018 DOI: 10.23919/ IEEE Opatija.

[2] QiongWu;Wen-LingHsu;TanXu;ZhenmingLiu;GeorgeMa;GuyJacobson; Shuai Zhao Speaking with Actions - Learning Customer Journey Behavior 2019 IEEE 13th International Conference on Semantic Computing (ICSC) Year: 2019

[3] SoumiGhosh;Chandan Banerjee A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE) Year: 2020

[4] Bora Bardük Modelling Time Statistics for Customer Churn Prediction 2020 28th Signal Processing and Communications Applications Conference (SIU) Year: 2020

[5] Chiaki Doi;MasajiKatagiri;TakashiAraki;DaizoIkeda;HiroshiShigeno Is he Becoming an Excellent Customer for us? A Customer Level Prediction Method for a Customer Relationship Management System 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA) Year: 2018

[6] Dehua Kong;XingLi;Yongxia Zhao Research on Product Recommendation Based on Web Space-Time Customer Behavior Trajectory 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) Year: 2019

[7] Harsh Valecha;AparnaVarma;IshitaKhare;AakashSachdeva;Mukta Goyal Prediction of Consumer Behaviour using Random Forest Algorithm 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) Year: 2018

[8] Asniar;KridantoSurendro Predictive Analytics for Predicting Customer Behavior 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT) Year: 2019

[9] SumitChavan;AvantiDorle;SiddhivinayakKulkarni;SitalakshmiVenkatraman Prediction Model Development using Neural Network Approach 2019 IEEE Pune Section International Conference (PuneCon) Year: 2019

[10] HaoKang;HailongZhao;Ting Ai The Description of Optimal Decision Tree Algorithm and Its Application in Customer Consumption Behavior 2020 IEEE International Conference on Information Technology,

[11] JinyoungYeo;Seung-wonHwang;sungchulkim;EunyeeKoh;NedimLipka Big Data and Artificial Intelligence (ICIBA) Year: 2020 Conversion Prediction from Clickstream: Modeling Market Prediction and Customer Predictability IEEE Transactions on Knowledge and Data Engineerin Year: 2020

[12] T. Yoshida, M. Hasegawa, T. Gotoh, H. Iguchi, K. Sugioka and K. Ikeda, "Consumer behavior modeling based on social psychology and complex networks," The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007), Tokyo, 2007, pp. 493-494.

[13] . R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016

[14] J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015.

[15] Ben Yedder, Hanene&Zakia, Umme& Ahmed, Aly &Trajkovic, Ljiljana. (2017). Modeling prediction in recommender systems using restricted boltzmann machines. 2063-2068. 10.1109/SMC.2017.812292.