Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/54/20241420

Overview of potential customer prediction for products based on machine learning

Qihang Wang

School of Education, Johns Hopkins University, Baltimore, United States

qwang142@jh.edu

Abstract. With the development of the Internet, e-commerce, as a new way of online shopping, has provided great convenience to people's lives. Due to the complex potential relationship between consumers and products, it is difficult to recommend products according to consumers' needs, which increases the difficulty of online shopping. Therefore, how to predict the potential consumers of commodities has attracted the wide attention of researchers. Fortunately, machine learning-based approaches have made text and images that can model complex underlying relationships between data. Therefore, researchers have introduced machine learning into the product potential consumer prediction field and achieved good results. This paper first introduces the relevant data set and the product potential users' forecast evaluation index. Then, it summarizes the relevant product potential consumer prediction methods based on machine learning. Finally, the paper summarizes the whole article and looks forward to future research methods.

Keywords: Potential customer prediction for product, Machine Learning, product recommendation.

1. Introduction

Nowadays, through the development of technology, many new technologies have emerged, like artificial intelligence and self-learning programs like machine learning. For example, AI was a recent development like ChatGPT, a Large Language Model [1]. So many researchers have started researching how technology, like machine learning, can exert its power in business, like possible customer predictions for products in the business world. In the business world, most small and large companies always need to face one question: what customer wants and how to make long-term relationship with their customer. So, possible customer prediction for products is a strategy that helps companies identify which kind of customer would be most likely to buy the company's product. Thus, possible customer prediction for products is significant in the business field since in the business field, to compete with other companies, information and time become the essential or main advantage. If a company could accurately find customers who would most likely buy their product, their advertising efficiency would increase a lot, and companies' advertising budgets could also decrease. Plus, since advertising to all customers may have a negative response, finding the correct set of customers becomes crucial.

So, many companies have tried different methods to achieve the prediction goal and make predictions more accurate. And the first thing that forecasting needs might be data. For example, in retail, fashionable product companies would use their databases to extract product information from

DOI: 10.54254/2755-2721/54/20241420

companies' ERP systems or seek frontline staff for feedback for qualitative data. To make forecasting, they would use classical statistical methods such as the Bayesian approach, auto-regression, and exponential smoothing [2]. This kind of strategy is straightforward to implement. It has an acceptable performance for forecasting, but over time, when people have more and more information sources, this strategy also has its' limitations. This strategy is suitable for products with stable demand [2]. But with the significant data era coming, customers' requirement changes quickly with an irregular pattern, so this strategy becomes ineffective because it may not find the irregular pattern. Another example of possible customer prediction for products is that an article shows that in Bangladesh, most people are related to business like many people opened a shop. Hence, they need target customers because losing orders and consumers is an important problem that retailers do not wish to encounter. Considering the competition and limitations of finance in the retail industry, it is tough to make precise predictions [3]. So, these two examples show that possible customer prediction is crucial for most companies to gain profit or to manage their budget, and an accurate prediction method is hard to find. In addition, a classical statistical method used to make demand forecasting have an acceptable performance but also has many limitations.

Thus, many researchers have started researching how advanced technology like machine learning models can integrate with the concept of possible customer prediction to help the prediction become more efficient and increase its accuracy. So, the main objective of this article is to provide a basic introduction to different kinds of machine learning algorithms and how these different algorithms can be applied to recommendations or possible customer prediction. And potential future research direction for machine learning model. Plus, the hope of machine learning can be more efficient and solve more problems people and companies could have.

Machine learning is a technique or tool that has become increasingly popular in recent years, and many fields have used it to make upgrades. Machine learning's definition is a technique that enables one to learn without explicitly being programmed to do so, with its performance to be optimized as they are exposed to more data samples [4]. In other words, machine learning is based on algorithms that can learn data without requiring programmers to ask them to do it. And machine learning has three significant sections and seven steps to achieve the goal. The three major sections are supervised learning, like Linear Regression, K-nearest neighbor, and unsupervised learning, like K-means clustering and reinforcement learning [4]. The application of some algorithms in business and a more detailed explanation of algorithms will be discussed in the machine learning method section later in this paper. In addition, the seven steps that must be taken for a machine learning model to achieve its goal are data collection, data cleaning and data preprocessing, feature engineering, the definition of the machine learning model, training, performance evaluation, and the final step prediction. When data are first collected, they are raw data that cannot be used for machine learning models, and only after the data have gone through the process can they be used for prediction. In feature engineering, it creates new features, but its goal is to select suitable parts for the model. For example, there might be 100 models, and to do feature engineering, these features will be ranked by how relevant the feature is to the problem. And the top rank, like 20 or 30, would be chosen. Selecting good features is vital because good features can help to improve model performance. So, integrating machine learning models into possible customer prediction problems can allow an increase in the accuracy of the prediction because after training for a model has been completed with a good performance, machine learning models can recognize patterns that may not recognized by a simple statistical review of history data which is crucial for most companies. In addition, after the prediction model accuracy has been increased, the budget and business decisions would be more precise and accessible. The model could improve efficiency significantly because new data from the machine learning model can be updated without further training so that the prediction will be based on the latest pattern.

DOI: 10.54254/2755-2721/54/20241420

2. Dataset

The dataset being used for the article is from a Portuguese banking institution. The Portuguese banking institution wants to predict whether customers would subscribe to their term deposit. Different types of user information were recorded because multiple contact with the same client was often required [5].

2.1. Data description

In the dataset, users' information was recorded by seventeenth distinct attributes. These attributes included age, job, marital, education, balance, loan, y, and other elven attributes. Some attributes are easy to understand, but some still need to be. For example, the y attribute is the target attribute for prediction results, or balance means the average yearly balance in the bank. These attributes were also categorized by different types, including integer, binary, categorical, and date [5]. For example, the attribute age is in the integer category; martial is in the categorical category.

2.2. Data Processing

Several steps need to be done before this dataset can be used to train a model to predict whether customers would subscribe to banks' term deposits, as the introduction part discussed. Clean or fill in the missing value, but there are no missing values or outliers in this dataset, so this step can be skipped. The next step is to create features; in this dataset, different attributes that might have a connection with each other would connect and develop new features. And to develop features, different kinds of correlation should be considered to create more features for future selection.

2.3. Feature selection

After different features were created, the time has come to select features for the dataset and make further analysis. In feature selection, the purpose is to choose features with high quality. So, to choose high-quality features, there are different methods. For example, in this dataset, to determine high-quality features, the correlation quality should be examined by using scores to represent the importance of different features to the target variable. And rank these features top-down and choose the top features for future analysis. The reason for doing this is that training a machine learning model is very time-consuming, so if the feature selection part is high quality, it can either increase model performance or help decrease the model complexity. High complexity in models can cause overfitting and decrease model performance.

In conclusion, choosing a model is essential for companies, but data preparation is necessary before selecting models. Making good data preparation, like feature selection, could help to increase machine learning models' performance, which helps companies better exert the power of machine learning models and improve their work efficiency.

3. Method

3.1. From questionnaire to recommendation system

3.1.1. Classic method of product recommendation

Before advanced technology like machine learning emerged, many companies used classical ways to analyze and collect data. They usually need to send out many questionnaires for people to answer, which is very time-consuming. Also, although classical data analysis methods are acceptable, they still need help finding an irregular product pattern to predict the product's target customers. Also, finding hidden relationships between different data points is challenging for these methods. So, after machine learning emerged, many companies wanted to integrate machine learning to help them increase the possibility of finding the correct set of customers to recommend or help companies' websites become more user-friendly.

DOI: 10.54254/2755-2721/54/20241420

3.1.2. Product Recommendation based on early machine learning

An example of applying machine learning models could be using k-means clustering algorithms and k-nearest neighbor to improve the movie recommendation system's work process and root mean squared error value. In the article on movie recommendation system, three authors, Rishabh Ahuja, Arun Solanki, and Anand Nayyar, propose to implement a k-means clustering algorithm to cluster different kinds of movies and use k-nearest neighbor to replace the work of collaborative filtering to optimize the work process [6]. The result of implementing two algorithms is that the root mean squared error value is better than any existing technique [6].

In this example, multiple algorithms have been introduced to apply to the movie recommendation system. One is the k-means clustering algorithm, and another is the k-nearest neighbor algorithm. How k-near neighbor work is that there will be k points representing data and calculating the similarity between each point to compare whether or not two data points are from one cluster and how close they are for future actions [7]. This definition has been applied to this movie recommendation system because the dataset has different users for different movie ratings. The k-near neighbor algorithm would suggest movies based on calculating the similarity of two users' movie ratings. If the similarity is high, which means these two users might have the same interest in movies, the algorithm will introduce movies one user has watched to another. If the similarity is low, the algorithm will pass to other users to continue. The k-means clustering algorithm works because the K-means algorithm uses random number selection to select cluster centers from the dataset, so cluster number is required. [8]. So, this algorithm can be applied to the example of a movie recommendation system because it would randomly select k movies in the dataset and calculate the distance between these k movies to split all movies into k clusters and continue iterating to find the optimal cluster result. The authors of this article want to apply this algorithm because they want it to help them reduce the number of clusters that need to be used by the recommendation system to optimize the work process of the system.

3.2. Problems related to the system which predicts target customers and solutions

Through the development of advanced technology, new machine learning models like more efficient XGBoost emerged, and some problems related to the recommendation system have not been solved—for example, the cold start problem. This problem often occurs when the recommender system cannot connect with existing data to make recommendations [9]. When a new user comes to the system but does not have data for this customer, they cannot predict whether this customer is a potential customer for any product, which is a problem that needs to be solved. A possible solution was presented with an example of integrating different machine learning models into a music recommendation system. In the article, Haoye Tian and other authors want to solve the problem by enhancing the LR model and adding the XGBoost model to solve the cold start problem and improve the system [10]. Their work shows that their enhanced LR model is slightly better than the traditional method used in the system. If they just used the LR model and the XG boost model, they all had problems, so finally, they combined two models to create a new model, which is the LX model. Eventually, their experiment proved that the LX model had the optimal result.

In this example, different algorithms were introduced, like logistic regression, the XGBoost model, and a new LX model that combined logistic regression and XGBoost. First, how logistic regression works is that logistic regression models are statistical models that evaluate relationships between two or more variables. One is a dependent qualitative variable, and another is one or more independent explanatory variables [11]. This model is the most popular one used in recommendation systems. However, the performance of this model is not that good in the music recommendation system because this model is not good at dealing with nonlinear features. So Haoye Tian and other authors want to support the LR with the XG boost model. XGBoost model is an abbreviation of the eXtreme Gradient Boosting package. It is an efficient implementation of a gradient-boosting framework [12]. So, in the article, XGBoost was used to overcome the weakness in linear models and help improve the model's speed [5]. Finally, the LX model, which Haoye Tian and other authors created, combined the logistic regression and the XGBoost models. They wanted to make this model because they thought it could take

DOI: 10.54254/2755-2721/54/20241420

advantage of both models to better deal with nonlinear features [10]. The result shows that the created model LX is genuinely better than the other models, the logistic regression model and the XGBoost model. The LX model has the lowest error rate in the music recommendation system.

3.3. Application

In the previous section, different kinds of machine learning models, like accessible models for logistic regression or k-near neighbor and tree-based models for XGBoost, were introduced. Various examples of how machine learning models can be applied to different circumstances, such as upgrading systems or creating a new one, have been shown to both achieve their purpose. However, improvement can be made only if companies know their question and choose machine learning models according to that question. After companies are transparent about their question, they can evaluate different aspect models, such as how big the training data is, how long it would take to train a model, and so on. [13] Thus, when companies want to make improvements like the examples above, they need first to identify the purpose of what they wish their machine learning models to do for them. For the movie recommendation system example, three authors want their algorithm to help optimize their working process. And after they identify the purpose, they can start to evaluate the performance of different models. Like in the music system example, Haoye Tian and other authors first use logistic regression, the XGBoost model to support logistic regression, and finally, the LX model. Eventually, they compare three models and find the model with the highest performance and the model that suits their purpose.

When a company wants to sell a new product, it always wants to predict what kind of people would most likely buy it. Nevertheless, it took much work in the past to connect the product and the people who buy it. In the past, to predict which kind of customer is interested in their development, many companies usually needed to send out many questionnaires for people to answer, which is very time-consuming and might cost much money. In addition, even after technology is developed and some ways can be used to predict target groups of customers, there are also some limitations like the irregular pattern of customers or some hidden relationship between different customer information. But nowadays, with information technology and computer science development, machine learning algorithms have grown and gradually become the mainstream to improve prediction efficiency. Also, it can help to improve the accuracy of prediction results. We can see in the example result with the help of machine learning models like Logistic regression or XGBoost, or models people create in particular situations that serve special needs. The music or movie recommendation systems are upgraded as in the previous section. These systems become more powerful or user-friendly than the last or other versions. This is how the machine learning model can be applied to product lead forecasting.

4. Conclusion

In conclusion, this article provides an overview of how machine models can be applied to possible customer prediction. The article starts by introducing the concept of possible customer prediction. Possible customer prediction for products is a strategy that helps companies identify which kind of customer would be most likely to buy the company's product. And then this introduces what machine learning is. Machine learning is a technique that enables one to learn without explicitly being programmed to do so, with its performance to be optimized as they are exposed to more data samples. And why customer prediction may want to integrate with machine learning models. Then, a specific dataset relating to possible customer prediction has been analyzed in the dataset section. In the method section, different examples of how machine learning has been applied to possible customer prediction. These various examples have introduced many models, like logistic regression, k-near neighbor, k-means clustering, etc. These models all solve problems that people may have and make an improvement. The k-means clustering algorithm helps the movie system to reduce the number of clusters used in the work process, and the k-near neighbor algorithm replaces an old way of finding similarities between customers to optimize the working process. The logistic regression algorithm and XGBoost model combined to create a new LX model, which helped the system increase its accuracy. However, these models also have problems like logistic regression, which needs to handle nonlinear features better. And

DOI: 10.54254/2755-2721/54/20241420

XGBoost can go overfitting. Overfitting means making the model too complex, so the performance of the training part looks excellent, but when it comes to the actual testing part, the performance decreases significantly. Also, like the k-means clustering model, processing a large amount of data might cost much time. So, in the future, machine learning models should enhance the ability to process big data sets because time flies; so many companies now have massive databases that store data, so processing huge datasets is a vital ability. Also, future machine learning directions should help people better understand why these machine learning models work and why some effective models work still need to be clarified to people. Companies can better achieve their purpose by understanding how machine learning models work. In addition, future machine learning should be more people-centric, meaning the model's answer and action can be explained to people rather than current days; it is hard to explain the answers from models. By reaching this goal, it is easier for companies to adjust the model they are using and produce the better result they want because people now understand the results that models give better. In conclusion, machine learning models in the future should continue to make progress to be more powerful, easier to understand, and help people solve more and more different kinds of problems that people might have.

References

[1] Rahman , M., Terano, H. J. R., Rahman, N., Salamzadeh, A., Rahaman, S. (2023). ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. Journal of Education, Management and Development Studies. 3(1). 1-12. Doi: 10.52631/jemds.v3i1.175

[2] Ren, S., Chan, HL. & Siqin, T. Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. Ann Oper Res 291, 761–777 (2020). https://doi.org/10.1007/s10479-019-03148-8

[3] M. A. I. Arif, S. I. Sany, F. I. Nahin and A. S. A. Rabby, "Comparison Study: Product Demand Forecasting with Machine Learning for Shop," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 171-176, doi: 10.1109/SMART46866.2019.9117395.

[4] H. M. E. Misilmani and T. Naous, "Machine Learning in Antenna Design: An Overview on Machine Learning Concept and Algorithms," 2019 International Conference on High Performance Computing & Simulation (HPCS), Dublin, Ireland, 2019, pp. 600-607, doi: 10.1109/HPCS48598.2019.9188224.

[5] Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.

[6] R. Ahuja, A. Solanki, and A. Nayyar, "Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 263-268, doi: 10.1109/CONFLUENCE.2019.8776969.

[7] R. A. Jarvis and E. A. Patrick, "Clustering Using a Similarity Measure Based on Shared Near Neighbors," in IEEE Transactions on Computers, vol. C-22, no. 11, pp. 1025-1034, Nov. 1973, doi: 10.1109/T-C.1973.223640.

[8] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming, K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, Information Sciences, Volume 622, 2023, Pages 178-210, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2022.11.139.

[9] Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 9, 59 (2022). https://doi.org/10.1186/s40537-022-00592-5

[10] H. Tian, H. Cai, J. Wen, S. Li, and Y. Li, "A Music Recommendation System Based on logistic regression and eXtreme Gradient Boosting," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-6, doi: 10.1109/IJCNN.2019.8852094.

DOI: 10.54254/2755-2721/54/20241420

[11] S. Domínguez-Almendros, N. Benítez-Parejo, A.R. Gonzalez-Ramirez, Logistic regression models, Allergologia et Immunopathologia,Volume 39, Issue 5, 2011,Pages 295-305, ISSN 0301-0546, https://doi.org/10.1016/j.aller.2011.05.002.

[12] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

[13] M. D. Tamang, V. Kumar Shukla, S. Anwar and R. Punhani, "Improving Business Intelligence through Machine Learning Algorithms," 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2021, pp. 63-68, doi: 10.1109/ICIEM51511.2021.9445344.