https://nmrj.ui.ac.ir/?lang=en
New Marketing Reserch Journal
E-ISSN: 2228- 7744
Vol. 13, Issue 2, No.49, Summer 2023
Document Type: Research Paper
Received: 16/11/2022 Accepted: 11/07/2023

Customer Churn Analysis Based on the Data-mining Approach: Hybrid Algorithm Incorporates Decision Tree and Bayesian Network

Atefeh Aghakhani Bezdi Langari
MA Student, Department of Industrial Engineering and Management, Shahrood University of Technology, Shahrood, Iran
aqakhani396@gmail.com

Aliakbar Hasani *
Associate Professor, Department of Industrial Engineering and Management, Shahrood University of TechnologyShahrood, Iran
aa.hasani@shahroodut.ac.ir

Abstract

Today, companies and organizations are aware of the fact that customer retention leads to greater profitability. Increasing competition causes the rate of customer churn to grow. Therefore, studying the features influencing the tendency of customer churn is important. In the present study, a hybrid model based on the data mining approach is presented to analyze the features of churn customers. In the first step, the feature selection node has been used to identify the features with higher importance and remove redundant items. Then, the C5.0 Decision Tree and Bayesian network were used to classify the customers into two groups, turning customers, and non-turning customers. These are data mining techniques and terms that can help in forecasting. Finally, the proposed model has been implemented in the chain store industry as a case study. Key findings indicate that both the decision tree model and the Bayesian network can predict churn customers with different accuracies, the area under the

*Corresponding author

10.22108/NMRJ.2023.135756.2797

receiver operating curve in the decision tree model is greater than the Bayesian network model and it has better performance. The results indicate the optimal efficiency of the proposed method. In addition, the results show that three features of gender, marital status, and age from the set of demographic characteristics and five factors of average monthly income level, number of purchases per month, the share of online shopping, how to get to know the store, type of market-related to customer transaction records are among the most effective factors.

Introduction

Customers are among the most important assets for businesses. Customer relationship management has been introduced as a comprehensive key strategy to stay focused on customers' needs and integrate methods of dealing with customers in the organization. In the key area of customer relationship management, the importance of customer orientation is very evident. This term means leaving the organization on behalf of customers and turning to competing organizations to receive services. Customers may leave the organization for various obvious or hidden reasons. The goal of organizations is to maintain existing customers by using customer retention methods. Customer loss creates a situation for competing organizations to attract customers of an organization. This situation has made the importance of predicting the churn of customers to double. Researchers have concluded that a small change in customer retention rate can have a huge impact on overall business improvement. Predicting customer churn is a suitable tool for describing the customer retention process of an organization, and the purpose of using it is to identify a group of customers who are prone to churn. Knowing this group of customers and taking preventive measures can play a very important role in preventing customers from turning away. In such a situation, predicting customer turnover has attracted a lot of attention in management and marketing studies. To efficiently manage customer churn forecasting within an organization, it is of great importance to provide an effective and highly accurate customer churn prediction model.

Methodology

The research is practical according to the purpose. Because the results of this research can be practically used, the nature of this research is post-event. Considering the existence of three types of research methods: quantitative, qualitative, and hybrid, the method of this research is quantitative. Since a specific methodology must be used to perform data mining operations, standard Crisp methodology has been used in this regard. Also, various data mining techniques including feature selection and classification have been used in this research. Finally, a case study is the method used in this research. In this research, based on the real data of the customers of the chain store industry, customer churn is predicted and the

effective factors of churn are identified. In this research, a hybrid model based on a data mining approach is presented to analyze the factors of customer churn. In the first step, the Bayesian network algorithm was used to identify factors with higher importance. Bayesian network is a non-circular directed graph where each node represents a variable and arcs represent direct causal relationships between connected nodes and conditional probability tables are assigned to nodes that have conditional dependence. In the second step, the C5.0 decision tree technique was used to classify customers based on their churning status. The C5.0 decision tree is an improvement over the C4.5 and ID3 decision trees. The division of each node is calculated based on information gain. This index is used to select the fragile variable in the process of tree growth.

Findings

The results of the research indicate that the demographic characteristics and purchase records of customers are effective in the behavior of customers. Based on the results of the classification in this research and the rules provided by the decision tree, the eight key factors identified in turning away customers with a significant effect on the studied case have been analyzed in the following. The classification results show that people with a monthly income of 50 million Rials and above are among those customers who are likely to buy again. People with a monthly income of less than 50 million Rials are not customers. People whose number of purchases was more than 2 times a month are among those loyal customers who are likely to buy again, and those customers who are likely to buy less than 2 times would not repurchase and they are part of the churning customers. The obtained results show that people with an internet shopping share of more than 20% are loyal customers and people with a lower shopping share are turned away (churn) customers. The percentage of women's purchases is higher than that of men, and women are more loyal than men. Married people have a higher percentage of the store's customers and are more loyal than single people. The overall amount of people's purchases from a chain store is much higher than from a traditional supermarket.

Conclusions

Analyzing and predicting customer behavior is very important because the cost of losing a customer is very high for an organization. In this regard, in the current research, the method of combining the Bayesian network and a C5.0 decision tree has been developed for predicting the analysis of customer churn. For this purpose, the data of customers of a chain store in Mashhad City has been examined as a case study. The variable of customer repurchase probability is considered as a dependent variable and then the most important independent variables for the implementation of the C5.0 decision tree and Bayesian network are identified by the feature selection node. The results of the present research show that the application of

the feature selection algorithm can help the decision makers to accurately classify the model and choose the best model and focus on the variables with the highest importance in predicting the turning of customers. The results also indicate that the eight factors of age, marital status, average monthly income, number of purchases per month, familiarity with the store, type of market, share of online shopping, and special sales are among the most important factors affecting diversion. According to the comparison of two Bayesian network algorithms and the C5.0 decision tree based on ROC diagram results, it is emphasized that the C5.0 decision tree with the highest accuracy has a better performance in identifying returning customers. Finally, a set of managerial insights for formulating marketing plans and facing all kinds of customers has been presented.

Keywords: Customer Churn, C5.0 Decision Tree, Bayesian Network, Data Mining, Machine Learning.

تحقیقات بازاریابی نوین
سال سیزدهم، شماره دوم، شماره پیاپی (۴۹) تابستان ۱۴۰۲
تاریخ دریافت: ۱۴۰۱/۸/۲۵ تاریخ پذیرش: ۱۴۰۲/۴/۲۰
ص: ۲۲- ۱

مقاله پژوهشی

تحلیل روی‌گردانی مشتریان مبتنی بر رویکرد داده‌کاوی: الگوریتم ترکیبی درخت تصمیم و شبکهٔ بیزین (مورد مطالعه: فروشگاه‌های زنجیره‌ای)

عاطفه آقاخانی بزدی لنگری [۱]، علی‌اکبر حسنی [۲]*

۱- دانشجوی کارشناسی ارشد مدیریت صنعتی، دانشکدهٔ مهندسی صنایع و مدیریت، دانشگاه صنعتی شاهرود، شاهرود، ایران
۲- دانشیار گروه مهندسی صنایع، گروه مهندسی صنایع، دانشکدهٔ مهندسی صنایع و مدیریت، دانشگاه صنعتی شاهرود، شاهرود، ایران
aa.hasani@shahroodut.ac.ir

چکیده

امروزه سازمان‌ها به این آگاهی رسیده‌اند که حفظ مشتریان باعث سودآوری بیشتر می‌شود. همچنین، افزایش رقابت نیز باعث می‌شود تا میزان روی‌گردانی مشتریان افزایش یابد؛ از این رو مطالعهٔ عوامل مؤثر بر تمایل به روی‌گردانی یا عدم روی‌گردانی مشتریان برای پژوهشگران و فعّالان کسب‌وکارها اهمیت دارد. در پژوهش حاضر یک مدل ترکیبی مبتنی بر رویکرد داده‌کاوی برای تحلیل عوامل روی‌گردانی مشتریان ارائه شده است. در گام نخست برای شناسایی عوامل با درجهٔ اهمیت زیادتر و حذف موارد زائد از گره انتخاب ویژگی استفاده و در گام دوم نیز برای طبقه‌بندی و پیش‌بینی مشتریان به دو دسته مشتریان روی‌گردان و مشتریان غیر روی‌گردان از درخت تصمیم C5.0 و شبکهٔ بیزین استفاده شده است. درنهایت، مدل پیشنهادی در صنعت فروشگاه‌های زنجیره‌ای به‌عنوان مطالعهٔ موردی پیاده‌سازی شده است. یافته‌های پژوهش حاکی از آن است که هر دو مدل درخت تصمیم و شبکهٔ بیزین توانایی مناسب را برای پیش‌بینی روی‌گردانی مشتریان دارد و سطح زیر منحنی عملیاتی گیرنده در مدل درخت تصمیم بیشتر از مدل شبکهٔ بیزین بوده است؛ درنتیجه مدل درخت تصمیم عملکرد بهتری دارد. همچنین، نتایج نشان می‌دهد که سه عامل جنسیت، وضعیت تأهل و سن از مجموعه مشخصه‌های دموگرافیک و پنج عامل متوسط سطح درآمد ماهیانه، تعداد خرید در ماه، سهم خرید اینترنتی، نحوهٔ آشنایی با فروشگاه و نوع بازار مربوط به سوابق تراکنش‌های مشتریان از مهم‌ترین عوامل مؤثر بر روی‌گردانی مشتریان است.

کلیدواژه‌ها: روی‌گردانی مشتری، داده‌کاوی، الگوریتم درخت تصمیم C5.0، الگوریتم شبکهٔ بیزین، یادگیری ماشین.

* نویسنده مسؤول

10.22108/NMRJ.2023.135756.2797

۱. مقدمه

مشتریان ازجمله مهم‌ترین دارایی‌ها برای کسب‌وکارها هستند. در دهه‌های اخیر، به‌دلیل اهمیت مشتریان در کسب‌وکار، تمرکز بسیاری از سازمان‌ها از محصول‌گرایی به مشتری‌مداری بوده است (Ye et al., 2013). شناسایی گروه‌های مختلفی از مشتریان و ایجاد طرح‌های بازاریابی، فروش و خدمات مناسب با نیازها و ویژگی‌های هر گروه، ازجمله هدف‌های مهم مدیریت ارتباط با مشتری است (Kotler, 2000). مدیریت ارتباط با مشتری به‌عنوان یک استراتژی کلیدی جامع برای تمرکز بر نیازهای مشتریان و یکپارچه‌سازی روش‌های برخورد با مشتریان در سازمان معرفی شده است (Brown & Cooper, 1999). درصورتی که سازمان‌ها قادر باشند میزان نگهداری مشتریان خود را به ۵ درصد افزایش دهند، سود کسب‌شده برای سازمان‌ها بین ۲۵ تا ۸۵ درصد افزایش پیدا می‌کند. در این حالت، مشتریان فعلی خرید بیشتری را در زمان کمتری انجام می‌دهند. همچنین، اگر مشتریان فعلی به نوسان‌های قیمت حساسیت کمتری را نشان دهند در این صورت می توانند تبلیغی برای جذب مشتریان جدید باشند (Ganesh et al., 2000). درحوزهٔ کلیدی مدیریت ارتباط با مشتری، اهمیت موضوع روی‌گردانی مشتریان بسیار نمایان است. درواقع، این اصطلاح به‌معنای ترک سازمان از جانب مشتریان و مراجعه به سازمان های رقیب برای دریافت خدمات است. مشتریان ممکن است به دلایل گوناگونِ آشکار یا پنهان سازمان ارائه‌دهندهٔ خدمت یا محصول را ترک کنند. درواقع، روی‌گردانی مشتری نقش بسیار مهمی در بازارهای اشباع‌شدهٔ امروزی دارد. هر سازمان برای پیشتازی خود در بازار، کسب سود و مدیریت هزینه‌های خود تلاش می‌کند. این درحالی است که جذب مشتریان جدید پنج تا شش برابر گران‌تر از حفظ مشتریان موجود است (Chung et al., 2016)؛ بنابراین هدف سازمان‌ها بر نگهداری مشتریان موجود با استفاده از روش‌های علمی و کارآمد حفظ مشتری متمرکز است. درواقع، از دست دادن مشتری فعلی منجر به ایجاد جذب مشتریان برای رقبا می‌شود (Lin et al., 2011). این وضعیت باعث شده است تا اهمیت پیش‌بینی روی‌گردانی مشتریان سازمان دوچندان شود. محققان به این نتیجه دست یافته‌اند که اندکی تغییر در میزان نگهداری مشتری می‌تواند بر بهبود کلی کسب‌وکار تأثیر بسیار زیادی بگذارد. پیش‌بینی روی‌گردانی مشتریان ابزاری مناسب برای توصیف فرآیند نگهداری مشتریان یک سازمان است که هدف از به‌کارگیری آن شناسایی گروهی از مشتریان است که مستعد روی‌گردانی هستند. با شناسایی این گروه از مشتریان و انجام‌دادن اقدام‌های پیشگیرانه می‌توان نقش بسیار مهمی در جلوگیری از روی‌گردانی مشتریان ایفا کرد. در چنین شرایطی، پیش‌بینی روی‌گردانی مشتریان توجه زیادی را در مطالعات مدیریت و بازاریابی به خود جلب کرده است. برای مدیریتِ کارآمدِ پیش‌بینیِ روی‌گردانیِ مشتری در درون یک سازمان، ارائهٔ مدل پیش بینی اثربخش با صحت زیاد روی‌گردانی مشتری اهمیت فراوانی دارد. مدل‌های پیش‌بینی‌کننده متعدّدی برای این مسئله ارائه شده است. در این میان، تکنیک‌های داده‌کاوی به‌طور مؤثری، قادر به شناسایی مشتریان متمایل به روی‌گردانی است. این تکنیک‌ها قابلیت کشف الگوها و ارتباط‌های درونی داده‌ها را دارند. همچنین، می‌توانند به دسته‌بندی و پیش‌بینی رفتار مدل براساس داده‌های دردسترس بپردازند. به عبارت دیگر، داده‌کاوی یک مبحث میان‌رشته‌ای است که با به‌کارگیری الگوریتم‌های مختلف به کشف الگوهای پنهانی مجموعه داده‌های گسترده می‌پردازد (Tsai & Lu, 2009). در صنعت فروشگاه‌های زنجیره‌ای نیز به‌دلیل افزایش رقابت و توجه

ویژهٔ مشتریان به کیفیت و قیمت کالا، مسئلهٔ تحلیل روی‌گردانی مشتریان اهمیت فراوانی پیدا کرده است. بدیهی است که در این فضای رقابتی و هزینهٔ اندک تغییر فروشگاه، مشتریان به‌راحتی فروشگاه مدنظر خود را به یک فروشگاه دیگر با خدمات و محصولات بهتر تغییر می‌دهند. روی‌گردانی مشتریان از یک فروشگاه به‌معنای کاهش سود عملیاتی است. به همین دلیل، فروشگاه‌ها باید تلاش کنند تا مشتریان باارزش فعلی خود را حفظ کنند؛ بنابراین هدف از پژوهش حاضر ارائهٔ مدلی کارآمد برای پیش‌بینی احتمال روی‌گردانی مشتریان فروشگاه زنجیره‌ای و همچنین، ارائهٔ پیشنهادهایی اثربخش برای کاهش روی‌گردانی مشتریان است.

۲. پیشینهٔ پژوهش

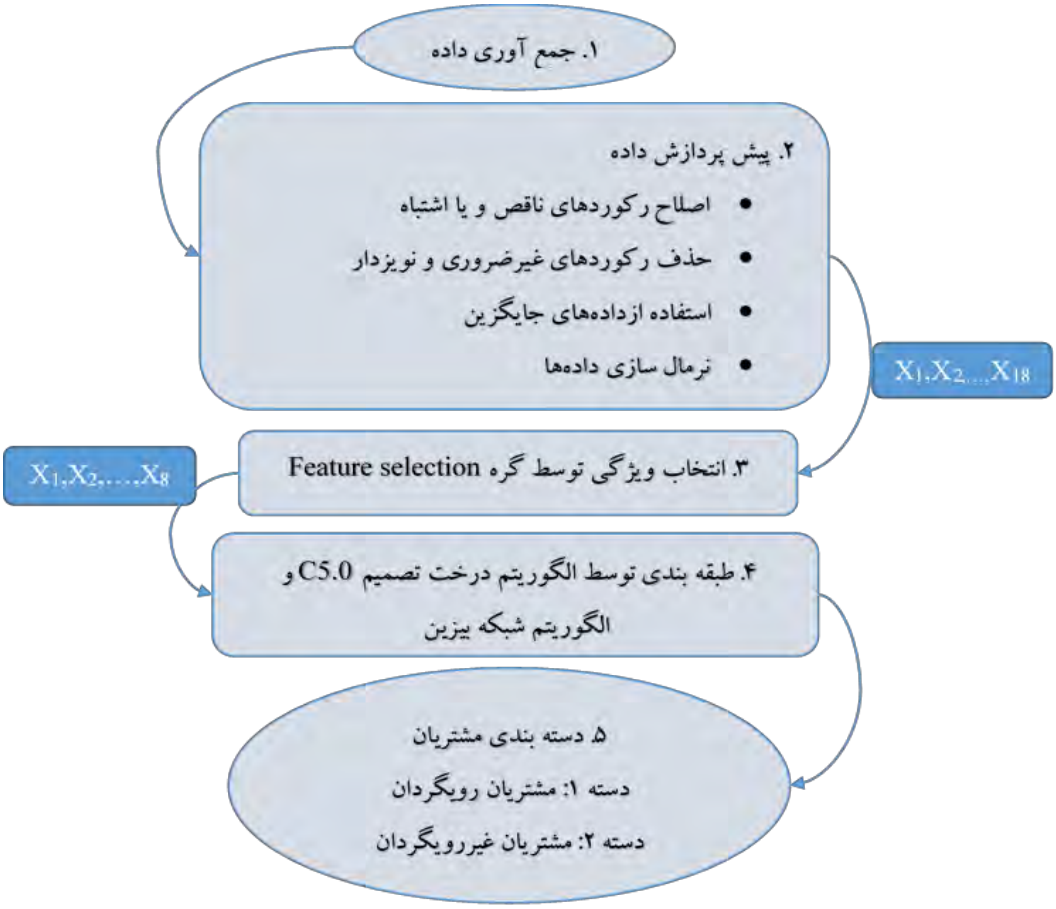
در این بخش مطالعات انجام‌شده درحوزهٔ تحلیل روی‌گردانی مشتریان ازمنظر معیارها و روش تحلیل مرور می‌شود. در پیشینهٔ موضوع، روی‌گردانی مشتری به فرآیندی اطلاق می‌شود که مشتریان از یک ارائه‌دهنده خدمات به ارائه‌دهندهٔ دیگری تغییر جهت می‌دهند؛ از این رو تکرارنشدن خرید به‌منزلهٔ روی‌گردانی از خدمت‌دهنده خواهد بود (Shobana et al., 2023). روی‌گردانی مشتری منجر به از دست دادن مشتریان بالقوه و کاهش درآمد سازمان خواهد شد. در این راستا، شناسایی و تحلیل علت‌های روی‌گردانی مشتریان اهمیت دارد و سازمان را قادر می‌کند تا خدمات خود را برای پاسخگویی بهتر به نیازهای مشتریان بهبود دهد. به‌طور کلی، روی‌گردانی مشتری به دو دستهٔ اصلی اختیاری و اجباری تقسیم‌بندی می‌شود. شناسایی وضعیت روی‌گردانی اجباری راحت‌تر از وضعیت اختیاری است؛ زیرا مشتریانی را دربرمی‌گیرد که به‌دلیل پاسخگونبودن به‌موقع و مناسب سازمان، تصمیم به قطع ارتباط گرفته‌اند. این درحالی است که شناسایی روی‌گردانی اختیاری در دو نوع عمدی و تصادفی و زمان رخداد آن برای سازمان بسیار دشوار است. روی‌گردانی اختیاری تصادفی می‌تواند به دلایل پیش‌بینی‌نشده همچون تغییرات شرایط مالی یا مکانی مشتری رخ دهد. روی‌گردانی اختیاری عمدی نیز تحت‌تأثیر عواملی همچون کیفیت سطح خدمات دریافت شده درمقابل رقبا، تحولات فناورانه و مسائل مشابه خواهد بود. شناسایی و تحلیل عوامل مؤثر بر روی‌گردانی یک مسئلهٔ پیچیده و تأثیرگذار بر مدیریت عملکرد سازمان‌هاست. به‌طور کلی، عوامل مؤثر بر روی‌گردانی به دو دستهٔ درونی و بیرونی تقسیم‌بندی شده است (امامی و همکاران، ۱۳۹۴). برخی از مهم‌ترین عوامل درونی عبارت است از: قوانین، عوامل مدیریتی، تبلیغات، خدمات، ساختار هزینه، کارکنان، تعامل و عوامل محیطی. برخی از مهم‌ترین عوامل بیرونی نیز عبارت است از: شهرت و اعتبار، روی‌گردانی اجباری، شلوغی و ازدحام، عوامل اجتماعی، تجربه‌های گذشته، ملاحظه‌های قومی و هزینهٔ روی‌گردانی. برای نمونه، در یک پژوهش برای تحلیل عوامل مؤثر بر روی‌گردانی مشتریان در صنعت بانکداری از داده‌های تراکنش مشتریان بانک، اطلاعات دموگرافیک، طول مدت ارتباط مشتریان با بانک و شکایت مشتریان استفاده شده است (Keramati et al., 2016). نتایج به‌دست‌آمده با درخت تصمیم نشان‌دهندهٔ ویژگی‌های مشتریان روی‌گردان است. در پژوهشی دیگر از مدل یادگیری ماشین افراطی برای پیش‌بینی روی‌گردانی مشتریان در صنعت بانکداری استفاده شده است. نتایج کسب‌شده حاکی از دقت بیشتر مدل ارائه‌شده درمقایسه با تکنیک‌های ماشین‌های بردار پشتیبان و شبکه‌های عصبی پس انتشار است (Mohanty & Sree, 2018). محققان در یک پژوهش به این نتیجه دست یافته‌اند که تأثیر وفاداری و رضایت مشتریان بر

قصد جابه‌جایی معنادار است و موانع جابه‌جایی این ارتباط را تقویت می‌کند (Mosavi et al., 2018). محققان در یک مطالعهٔ قیاسی برای پیش‌بینی روی‌گردانی مشتریان، Apache Sark ML و MLlib را بررسی کرده‌اند (Sayed et al., 2018). در یک پژوهش دیگر از تکنیک شبکهٔ عصبی برای تحلیل روی‌گردانی مشتریان استفاده شده است (Rosa, 2019). نویسنده در این پژوهش از داده های یک دورهٔ پنج‌ماهه برای تعریف مشتریان وفادار باتوجه به متغیرهای RFM استفاده کرده است. سپس از روش یادشده برای پنج ماه بعد به جهت تشخیص اینکه کدام یک از مشتریان وفادار مشخص‌شده در دورهٔ قبلی در این دوره روی‌گردان شده‌اند، استفاده می‌شود. در پژوهشی دیگر با استفاده از شبکه‌های عصبی به‌پیش‌بینی مشتریان روی‌گردان در صنعت مخابراتی پرداخته شده است (Khan et al., 2019). در ابتدا مجموعه داده از شرکت‌های مخابراتی پاکستان جمع‌آوری و سپس برای حذف اطلاعات نامناسب، مراحل پیش‌پردازش برروی‌داده‌ها پیاده‌سازی و اجرا و درادامه، ۲۶ ویژگی مهم که برای تشخیص روی‌گردانی مشتریان تأثیرگذار بوده انتخاب شده است. درنهایت، مدلی با به‌کارگیری الگوریتم شبکه‌های عصبی ایجاد شد که دقتی برابر با ۷۹ درصد را برای پیش‌بینی مشتریان روی‌گردان نشان می‌دهد. در یک پژوهش دیگر با استفاده از تکنیک‌های یادگیری ماشین، الگوریتم‌های درخت تصمیم، جنگل تصادفی گرادیان بوستینگ و XGBoost به‌پیش‌بینی مشتریان روی‌گردان درحوزهٔ مخابرات با استفاده از محیط Hadoop پرداخته شده است (Ahmad et al., 2019). در ابتدا مجموعه دادهٔ به‌کاررفته با استفاده از مراحل پیش‌پردازش و فرآیند تحلیل ویژگی‌ها بررسی شد. درادامه، برای مدل‌سازی پیش‌بینی و طبقه‌بندی مشتریان روی‌گردان، ویژگی‌های مناسب انتخاب و در اختیار الگوریتم‌ها قرار داده شد. نتایج نشان می‌دهد که مدل ایجادشده با به‌کارگیری XGBoost نسبت به مدل‌های ایجادشدهٔ دیگر دقت بیشتر و عملکرد بهتری را برای طبقه‌بندی مشتریان روی‌گردان دارد. در یک پژوهش دیگر با به‌کارگیری رویکرد متن‌کاوی، ابعاد رضایت مشتری در خطوط هواپیمایی با استفاده از نظرهای برخط مشتریان بررسی شده است (Lucini et al., 2020). محققان نظرهای مشتریان را از خطوط هواپیمایی و مسافران جمع‌آوری و ابعاد رضایت مشتریان را در ۲۷ بُعد شناسایی و همچنین، نظرهای مشتریان را با به‌کارگیری روش طبقه‌بندی بیزینِ ساده به مثبت و منفی تقسیم کرده‌اند و راهکارهایی را برای بهبود رضایت مشتریان ارائه داده‌اند؛ بنابراین حفظ مشتریان فعلی و جذب مشتریان جدید اهمیت زیادی دارد و انجام‌دادن یک پژوهش جامع را می‌طلبد. تحلیل مناسب رفتار روی‌گردانی و پیش‌بینی دقیق آن می‌تواند منجر به کسب سود بیشتر شود. در اغلب مطالعات رفتار روی‌گردانی مشتریان به‌صورت مستقل از رفتار دیگر مشتریان درنظر گرفته می‌شود. در یک مطالعه محققان عوامل مؤثر بر روی‌گردانی مشتریان و پیش‌بینی آن را با درنظرگرفتن وابستگی رفتار آنها با استفاده از ابزار تحلیل شبکه و الگوریتم یادگیری ماشین تحلیل کرده‌اند (Ljubičić et al., 2023). همچنین، پژوهشگران در مطالعه‌ای دیگر یک مدل پیش‌بینی را برای تحلیل روی‌گردانی مشتریان در فضای تجارت بنگاه‌به‌بنگاه با استفاده از ابزار داده‌کاوی مبتنی بر الگوریتم جنگل تصادفی ارائه کرده‌اند (Gattermann-Itschert & Thonemann, 2022). در یک مطالعهٔ دیگر، پژوهشگران یک مدل پیش‌بینی برای تحلیل عوامل مؤثر بر روی‌گردانی مبتنی بر الگوریتم تحولی و شبکهٔ بیزین در صنعت مخابرات ارائه کرده‌اند (Amin et al., 2023). در یک مطالعهٔ

دیگر محققان روش‌های مختلف مبتنی بر یادگیری ماشین را برای پیش‌بینی عوامل مؤثر بر روی‌گردانی مشتریان بررسی کرده‌اند (Geiler et al., 2022). همچنین، محققان در یک مطالعهٔ دیگر روش‌های مختلف طبقه‌بندی را برای پیش‌بینی روی‌گردانی مشتریان در فضای تجارت الکترونیک بررسی کرده‌اند (Baghla & Gupta, 2022). همچنین، محققان در یک مطالعه‌ای دیگر روش‌های مختلف طبقه‌بندی را برای پیش‌بینی روی‌گردانی مشتریان در فضای تجارت الکترونیک بررسی کرده‌اند (Kim & Lee, 2022). محققان با استفاده از رویکرد یادگیری ماشین مبتنی بر پیاده‌سازی و با تقویت گرادیان درخت تصمیم‌گیری اقدام به پیش‌بینی عوامل مؤثر بر روی‌گردانی مشتریان در صنعت مخابرات کشور نپال کرده‌اند (Shrestha & Shakya, 2022).

۳. روش پژوهش

پژوهش حاضر ازنظر هدف کاربردی و ازنظر رویکرد تجزیه‌وتحلیل داده‌ها کمّی است. شناسایی و تعریف شاخص‌های روی‌گردانی و ارزیابی مشتریان مبتنی بر پیشینهٔ موضوع و نظر خبرگان انجام شده است. برای این منظور در ابتدا مطالعهٔ معیارها براساس مطالعات کتابخانه‌ای و یافته‌های حاصل از پژوهش‌های پیشین جمع‌آوری شده است. سپس مجموعه ویژگی‌ها با روش دفی فازی دو مرحله‌ای به بررسی و تأیید خبرگان پژوهش رسیده است. گام‌های کلی پژوهش حاضر در شکل ۱ نشان داده شده است.



شکل ۱: گام‌های الگوریتم ترکیبی مبتنی بر درخت تصمیم C5.0 و شبکهٔ بیزین (منبع: یافته‌های پژوهش)

Figure 1: Steps of Hybrid Algorithm Based on C5.0 Decision and Bayesian Network

۳-۱. جمع‌آوری داده‌ها

داده‌های بسیار زیادی در بانک اطلاعاتی فروشگاه‌های زنجیره‌ای ثبت شده است. ویژگی‌های لازم پژوهش حاضر با نظر خبرگان و پژوهش‌های پیشین انتخاب شده است. با استفاده از پژوهش‌های گذشته و مبانی نظری دربارهٔ موضوع مدنظر، عوامل اصلی روی‌گردانی شناسایی و استخراج شده است. سپس با کمک تکنیک دلفی فازی در دو مرحله که مبتنی بر نظرهای خبرگان بود، ویژگی‌های اصلی در دو مرحله شناسایی شد تا در طراحی مدل از آنها استفاده شود.

۳-۲. پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها به‌عنوان اولین مرحله در ساخت مدل اهمیت فراوانی دارد. به‌کارگیری روش‌های مناسب پیش‌پردازش برروی داده‌ها قبل از ورود آنها به مدل‌های هوش مصنوعی ازجمله مواردی است که می‌تواند نتایج حاصل از شبیه‌سازی‌ها را به واقعیت نزدیک کند (حبیبی‌پور و همکاران، ۱۳۹۶). در پژوهش حاضر داده‌ها در ابتدا با استفاده از گره انتخاب ویژگی حاضر در نرم‌افزار مدلر، بررسی و محدودسازی شده است؛ به‌گونه‌ای که ویژگی‌هایی که رکودهای ناقص و یا اشتباه داشتند، اصلاح شدند. درادامه، ویژگی‌هایی که اطلاعات نویزدار داشتند و یا از رکودهای غیرضروری تشکیل شده بودند با به‌کارگیری داده‌های جایگزین اصلاح شدند. درنهایت، داده‌های موجود برای ارتقا کیفیت نتایج حاصل، طبق معادلهٔ ۱ نرمال‌سازی شد. به این ترتیب، داده‌ها در محدودهٔ بین ۰ و ۱ مقدار می‌گیرند.

$$\text{کردن نرمال} = \frac{\text{ارزش واقعی} - \text{مقدار پیشینه داده ها}}{\text{مقدار پیشینه داده ها} - \text{مقدار کمینه داده ها}} \quad (۱)$$

۳-۳. انتخاب ویژگی

در روش‌های انتخاب ویژگی با انتخاب زیرمجموعه‌ای بهینه و با حداقل اندازهٔ ممکن از ویژگی‌های اولیه، ابعاد داده‌ها کاهش می‌یابد. روش‌های گوناگون انتخاب ویژگی سعی بر این دارند که از میان 2^N زیرمجموعه کاندید برای یک مجموعهٔ N عضوی، بهترین زیرمجموعه را بر اساس یک تابعِ ارزیابی پیدا کنند. در تمام این روش‌ها براساس کاربرد، زیرمجموعه‌ای به‌عنوان جواب انتخاب می‌شود که بتواند مقدار تابع ارزیابی را بهینه کند. اگرچه هر روش تلاش می‌کند تا بهترین ویژگی‌ها انتخاب شود، باتوجه به وسعت جواب‌های ممکن و افزایش تعداد آنها به‌صورت توانی از N، پیداکردن جواب بهینه مشکل است و در Nهای متوسط و بزرگ هزینهٔ زیادی دارد. به‌طور کلی، روش‌های انتخاب ویژگی به سه دستهٔ اصلی بسته‌بندی، فیلترینگ و ترکیبی تقسیم‌بندی می‌شود (Kumar & Elavarasan, 2014). در پژوهش حاضر از روش فیلترینگ برای انتخاب ویژگی‌ها استفاده شده است. در روش فیلترینگ باتوجه به ویژگی‌های کلی داده‌ها، ویژگی‌های بهینه ازمیان کل ویژگی‌های موجود انتخاب می‌شود. این روش هزینهٔ محاسباتی کم و سرعت بالایی دارد و به‌راحتی می‌توان آن را پیاده‌سازی کرد (Tomar & Agarwal, 2014). در پژوهش حاضر از گرهِ انتخابِ ویژگی برای حذف داده‌های زائد استفاده شده است. این گره با استفاده از ضریب همبستگی به انتخاب فیلدهایی که می‌پردازد که ضریب همبستگی آنها به عدد یک نزدیک است. در طبیعت حوادث متعدّدی اتفاق می‌افتد که بین آنها همبستگی وجود دارد؛ به‌طوری که افزایش در مقدار یک متغیر منجر به افزایش یا کاهش در مقدار یک متغیر دیگر می‌شود که نشان‌دهندهٔ همبستگی است.

وجود همبستگی بین متغیرهای تحت ارزیابی، به این معنا نیست که یک متغیر، علت متغیر دیگر است. ضریب همبستگی، شاخص آماری است که میزان و حدود ارتباط بین متغیرها را نشان می‌دهد. در این گره پنج آیتم وجود دارد که درآیتم اول می‌توان براساس مقدارهای مفقوده انتخاب کرد؛ یعنی در این آیتم ٪۷۰ رکوردها و فیلدهایی که مقدارهای مفقودشده را دارند، به‌ترتیب حذف و غربال می‌شوند. در آیتم دوم فیلدهایی غربال می‌شود که رکوردهای زیادی را براساس فیلدهایی که در دسته‌های مشابهی قرارگرفته، داشته باشد؛ برای مثال، ٪۹۰ مشتریان فروشگاه اگر یک نوع کالا را خریداری کرده باشند این ویژگی نمی‌تواند تمایز دهندهٔ یک مشتری از مشتری دیگر باشد؛ درنتیجه این متغیر به‌عنوان یک متغیر بی‌اهمیت می‌تواند از مدل تحلیل حذف شود. در آیتم سوم فیلدهایی غربال می‌شود که تعداد زیادی دسته‌های جزئی و متفاوت با سایر دسته‌ها وجود داشته باشد؛ یعنی اگر هر مشتری کالای متفاوتی را خریداری کند این اطلاعات در مدل‌سازی مشتری بعدی به‌طور طبیعی، نقشی ندارد و مفید نخواهد بود و به همین صورت دو آیتم بعدی براساس ضریب اطمینان، واریانس و انحراف معیار انتخاب ویژگی را انجام می‌دهند. برای انجام‌دادن انتخاب ویژگی، احتمال خرید مجدد مشتری به‌عنوان متغیر وابستهٔ پژوهش درنظر گرفته می‌شود که مقدارهای ۰ و ۱ آن به‌ترتیب شامل عدم خرید مجدد و خرید مجدد می‌شود. متغیرهای پیش‌بینی‌کننده‌های پژوهش حاضر موارد ۱ تا ۱۷ جدول ۱ است. در این پژوهش از گره انتخاب ویژگی واقع در نرم‌افزار مدلر برای متغیر وابسته و پیش‌بینی‌کننده استفاده و عملکرد رویکرد انتخاب ویژگی نیز با دقت حاصل از فرآیند طبقه‌بندیِ سنجیده شده است. به این ترتیب که یکبار از الگوریتم انتخاب ویژگی برای انجام‌دادن طبقه‌بندی استفاده و بار دیگر بدون استفاده از الگوریتم انتخاب ویژگی سعی در طبقه‌بندی داده‌ها می‌شود. درنهایت، با مقایسهٔ میزان دقت هرکدام، می‌توان به اهمیت به‌کارگیری الگوریتم انتخاب ویژگی پی برد. در پایان نیز با استفاده از شبکهٔ بیزین به انتخاب ویژگی‌ها (اولویت‌بندی ویژگی‌ها) پرداخته شده و سپس درخت تصمیم مبتنی بر خروجی بیزین ترسیم شده است.

٤-۳. درخت تصمیم

تکنیک درخت تصمیم یکی از قدیمی‌ترین تکنیک‌ها برای ایجاد مدل دسته‌بندی است. الگوریتم‌های مبتنی بر درخت تصمیم، نتیجهٔ کار را به‌صورت درختی از کلیهٔ حالت‌های مختلف مقدارهای ویژگی‌ها ارائه می‌دهند (صنیعی آباده و همکاران، ۱۳۹۱). از این رو درخت تصمیم قادر به ایجاد توصیف‌های درک‌پذیر برای انسان از مجموعه روابط موجود در یک مجموعه داده است و همچنین، می‌تواند برای وظایف دسته‌بندی و پیش‌بینی از درخت تصمیم استفاده شود (Wu et al., 2022). این ساختار تصمیم‌گیری به شکل‌های گوناگون مانند تکنیک‌های ریاضی و محاسباتی که به توصیف، دسته‌بندی و عام‌سازی یک مجموعه از داده‌ها کمک می‌کند نیز معرفی می‌شود (Tsai & Chiou, 2009). ازجمله نقاط قوت درخت تصمیم عبارت است از: ۱- قابل فهم و درک‌بودن مدل ایجادشده؛ ۲- ارائهٔ پیش‌بینی‌ها درقالب یک رشته قوانین؛ ۳- بی‌نیازبودن به محاسبات طولانی و پیچیده برای دسته‌بندی داده‌ها؛ ۴- قابلیت به‌کارگیری برای انواع مختلف داده‌ها؛ ۵- شناسایی متغیرهای تأثیرگذار در پیش‌بینی و دسته‌بندی.

در پژوهش حاضر از درخت تصمیم C5.0 استفاده

شده است که ازجمله مزایای الگوریتم C5.0 درخت تصمیم، سرعت زیاد، کارایی حافظه، ایجاد درخت‌های تصمیم کوچک‌تر، وزن دهی و غربال‌سازی است (بحرینی‌نژاد و سروش، ۱۳۸۸).

الگوریتم درخت تصمیم C5.0 نسخهٔ بهبودیافته برپایهٔ درخت‌های تصمیم C4.5 و ID3 است که در آن شکسته‌شدن هر گره برپایهٔ بهرهٔ اطلاعاتی خواهد بود. از شاخص بهرهٔ اطلاعاتی برای گزینش متغیر شکننده در فرآیند رشد درخت به کار گرفته می‌شود. همچنین، میزان همگنی نمونه‌ها در یک گره با شاخص آنتروپی ارزیابی و از آن در تعیین بهرهٔ اطلاعاتی نیز استفاده می‌شود. درصورتی که متغیر هدف، C مقدار متفاوت داشته باشد، شاخص آنتروپی S وابسته به C کلاس از رابطهٔ ۲ محاسبه خواهد شد.

$$Entropy(s) = -\sum_{i=1}^{c} P_i . log_2^{P_i} \qquad (۲)$$

شاخص P_i نسبتی از S است که به کلاس i تعلق می‌گیرد. درواقع، شاخص بهرهٔ اطلاعاتی میزان کاهش انتظار را در شاخص آنتروپی ارائه می‌کند. به عبارت دیگر، بهرهٔ اطلاعاتی تعیین‌کنندهٔ اثرگذاری یک متغیر در فرآیند کلاس‌بندی است. بهره اطلاعاتی برای متغیر A وابسته به داده‌های S با استفاده از رابطهٔ ۳ محاسبه خواهد شد.

$$\begin{aligned} &Entropy(S,A) \\ &= Entropy(S) \\ &- \sum_{v \in value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \end{aligned} \qquad (۳)$$

دربخش نخست رابطهٔ (۳) میزان شاخص آنتروپی S درحالت اولیه و دربخش دوم آن میزان شاخص آنتروپی انتظار پس از تقسیم براساس متغیر A نمایش داده شده است.

۵-۳. شبکهٔ بیزین

شبکهٔ بیزین یک روش گرافیکی است که محققان در مطالعات تحلیل ریسک و ایمنی برپایهٔ دانش احتمال و عدم قطعیت به آن توجه کرده‌اند. شبکهٔ بیزین یک گراف جهت‌دار غیرمدور است که هر گره نشان‌دهندهٔ یک متغیر و کمان‌ها نیز نشان‌دهندهٔ روابط علّی مستقیم بین گره‌های بهم متصل‌شده، است. جدول‌های احتمال شرطی به گره‌های با وابستگی شرطی اختصاص داده می‌شود. براساس استقلال شرطی به‌دست‌آمده از مفهوم d-separation و قاعدهٔ زنجیره ای، شبکهٔ بیزین توزیع احتمال مشترک $p(x)$ از متغیرهای $X = \{U_1, \ldots, U_n\}$ را طبق رابطهٔ ۴ نشان می‌دهد (Nielsen & Jensen, 2009; Kjaerulff & Madsen, 2008).

$$P(X) = \Pi_{i=1}^{n} p\left(U_i | P_a(U_i)\right) \qquad (۴)$$

$P_a(U_i)$ نشان دهندهٔ احتمال والد متغیر U_i است؛ از این رو احتمال متغیر U_i به‌صورت رابطهٔ ۵ محاسبه می‌شود.

$$P(U_i) = \Sigma_{x \setminus U_i} P(X) P(U_i) = \Sigma_{x \setminus U_i} P(X) \qquad (۵)$$

یکی دیگر از ویژگی‌های بی‌نظیر شبکهٔ بیزین درمقایسه با سایر روش ها، قابلیت بهره گیری از تئوری بیز برای به‌روزرسانی احتمال وقوع رویدادهای اولیه به‌محض دریافت شواهد جدید (مانند آمار وقوع حوادث، داده‌های فرآیندی لحظه‌ای برای محاسبهٔ مقدارهای احتمالات به‌روزشده) است. (رابطهٔ ۶) (Nielsen & Jensen, 2009; Kjaerulff & Madsen, 2008).

$$P\left((X|E) = \frac{P(X.E)}{P(E)} = \frac{P(X.E)}{\Sigma_X P(X.E)}\right) \qquad (۶)$$

در رابطهٔ (۶) X نشان‌دهندهٔ متغیر تحت ارزیابی و E شواهد دریافتی است.

٤. یافته‌ها و بحث

در این مطالعه، اطلاعات مرتبط با خرید محصولات مصرفی از نوع تندمصرف از شعبه‌های منتخب یک فروشگاه زنجیره‌ای در شهرستان مشهد با استفاده از ابزار داده‌کاوی تحلیل شده است. مجموعه دادهٔ استفاده‌شده در این پژوهش، شامل اطلاعات ۱۰۰۰ مشتری با ۱۸ ویژگی است که ۷ ویژگی آن مربوط به اطلاعات دموگرافیک مشتریان مانند سن، جنسیت، وضعیت تأهل، مدرک تحصیلی، تعداد خانوار، وضعیت اشتغال و متوسط سطح درآمد ماهیانه و سایر ویژگی‌ها دربارهٔ سوابق خرید هر مشتری است. محققان اطلاعات را با استفاده از پرسشنامهٔ طراحی‌شده جمع‌آوری کردند که دربرگیرندهٔ سؤال‌های مرتبط با ویژگی‌های یادشده در فوق بوده است. ویژگی‌های استفاده‌شده برای انجام‌دادن این پژوهش (جدول ۱) باتوجه به پیشینهٔ موضوع، نظرهای خبرگان با روش دلفی دو مرحله‌ای جمع‌آوری شده است. جامعهٔ خبرگان پژوهش دربرگیرندهٔ ۲۰ نفر از مدیران ارشد فروشگاه زنجیره‌ای شهرستان مشهد و استادان دانشگاه در حوزهٔ بازاریابی بوده است.

جدول ۱: ویژگی‌های انتخاب‌شده

Table 1: Selected Features

شمارهٔ ویژگی	نام ویژگی	منبع	شمارهٔ ویژگی	نام ویژگی	منبع
۱	جنسیت	Keramati et al. (2016)	۱۰	تعداد خرید در ماه	Xue & Lu (2011) Baghla & Gupta (2022) Kim & Lee (2022)
۲	سن		۱۱	ارزش خرید در ماه	
۳	وضعیت تأهل	Gattermann & Thonemann (2022) Amin et al. (2023)	۱۲	تعداد خرید در سال	
۴	مدرک تحصیلی		۱۳	تأثیر فروش ویژه در خرید	
۵	تعداد خانواده		۱۴	ارزش آخرین خرید در زمان فروش ویژه	
۶	وضعیت اشتغال		۱۵	نحوهٔ آشنایی با فروشگاه	
۷	متوسط سطح درآمد ماهیانه		۱۶	سهم خرید اینترنتی در طی سال	
۸	روش آخرین خرید	Xue & Lu (2011)	۱۷	نوع بازار	
۹	ارزش آخرین خرید		۱۸	احتمال خرید مجدد	Shobana et al. (2023)

منبع: یافته‌های پژوهش

برای انجام‌دادن و تحلیل مدل ارائه‌شده در این پژوهش از نرم‌افزار IBM SPSS MODELER 18.0 استفاده شده است. بعد از انجام‌دادن پیش‌پردازش از گره انتخاب ویژگی برای شناسایی ویژگی‌های نامرتبط و تکراری که مسئله را با کمترین کاهش درجهٔ کارایی تشریح می کند، استفاده و سپس از الگوریتم درخت تصمیم C5.0 و الگوریتم شبکهٔ بیزین برای طبقه‌بندی داده‌ها استفاده می‌شود.

٤-١. بررسی اثربخشی رویکرد انتخاب ویژگی

در پـژوهش حاضـر داده‌هـا بـرای ارزیـابی مـدل ایجادشده به دو بخش آموزش و آزمایشی تقسیم‌بنـدی شده است. دربارهٔ تعداد رکوردهای آمـوزش و آزمـون می‌توان گفت که اگر تعداد دادهٔ آموزشـی زیـاد باشـد، مـدل سـاخته‌شـده بـه واقعیـت نزدیـک‌تـر خواهـد شـد. درواقـع، از ۷۵ درصـد داده‌هـا بـرای آمـوزش و از ۲۵ درصد نیز برای آزمون و اعتبار مدل استفاده شده اسـت. داده‌های بخش آموزش باعث می‌شود که مـدل بـه‌طـور هوشـمند، خـود را بـه‌روز و مـدل بهینـه را ایجـاد کنـد. داده‌های بخش آزمـون نیـز مـدل ایجادشـده را ارزیـابی می‌کند. برای بررسی اثربخشی رویکرد انتخاب ویژگـی از معیار دقت طبقه‌بندی استفاده شده است. بدین‌جهـت، بـرای بررسـی اثربخشـی ایـن رویکـرد در میـزان دقـت طبقه‌بنـدی یکبـار از گـره انتخـاب ویژگـی و بـار دیگـر بدون استفاده از گره انتخاب ویژگی سعی در طبقه بندی داده‌ها شده است که نتایج الگـوریتم درخـت تصـمیم و شبکهٔ بیـزین بـه‌صـورت جداگانـه در جـدول‌هـای۲ و ۳ آمده است.

جدول ۲: بررسی اثربخشی رویکرد انتخاب ویژگی درخت تصمیم C5.0

Table 2: Evaluation of the Effectiveness of the C5.0 Decision Tree Feature Selection Approach

الگوریتم درخت تصمیم C5.0	مجموعه دادهٔ یادگیری		مجموعه دادهٔ آزمون	
	درصد پیش‌بینی درست	درصد پیش‌بینی نادرست	درصد پیش‌بینی درست	درصد پیش‌بینی نادرست
دقت بدون انتخاب ویژگی	۹۸/۶۲	۱/۳۸	۹۳/۹۸	۶/۰۲
دقت با انتخاب ویژگی	۱۰۰	۰	۹۷/۵۹	۲/۴۱

منبع: یافته‌های پژوهش

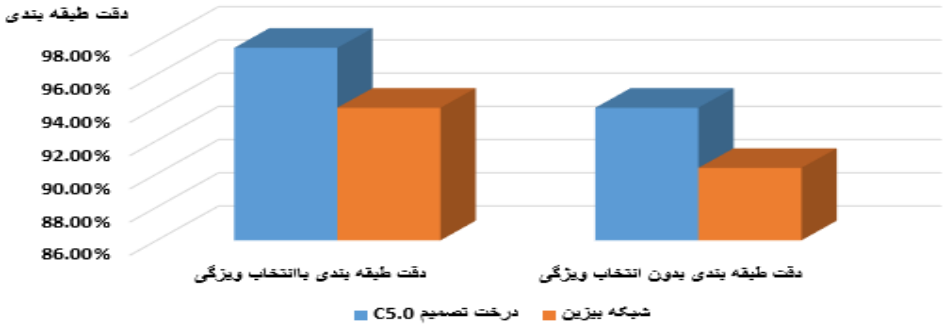
جدول ۳: بررسی اثربخشی رویکرد انتخاب ویژگی شبکهٔ بیزین

Table 3: Evaluation of the Effectiveness of the Bayesian Network Feature Selection Approach

الگوریتم شبکهٔ بیزین	مجموعه دادهٔ یادگیری		مجموعه دادهٔ آزمون	
	درصد پیش‌بینی درست	درصد پیش‌بینی نادرست	درصد پیش‌بینی درست	درصد پیش‌بینی نادرست
دقت بدون انتخاب ویژگی	۹۹/۰۸	۰/۹۲	۹۰/۳۶	۹/۶۴
دقت با انتخاب ویژگی	۹۶/۷۷	۳/۲۳	۹۳/۹۸	۶/۰۲

منبع: یافته‌های پژوهش

همـان‌گونـه کـه در جـدول‌هـای ۲ و ۳ و شـکل ۲ مشاهده می‌شود، دقت طبقه‌بنـدی دو الگـوریتم درخـت تصمیم C5.0 و شـبکهٔ بیـزین بـدون اسـتفاده از انتخـاب ویژگی در مجموعه دادهٔ آزمون به‌ترتیب برابر بـا ۹۳/۹۸ و ۹۰/۳۶ است و دقت طبقه‌بنـدی دو الگـوریتم درخـت تصمیم C5.0 و شبکهٔ بیزین با استفاده از انتخاب ویژگی در مجموعـه دادهٔ آزمـون بـه‌ترتیـب برابـر بـا ۹۷/۵۹ و ۹۳/۹۸ محاسبه شده است. با مقایسهٔ نتایج به‌دست‌آمـده می‌توان نتیجه گرفت که استفاده از گره انتخاب ویژگـی (به‌عنوان تکنیـک اسـتفاده‌شـده بـرای رویکـرد انتخـاب ویژگی) تأثیر بسزایی در دقت طبقه‌بندی دارد.

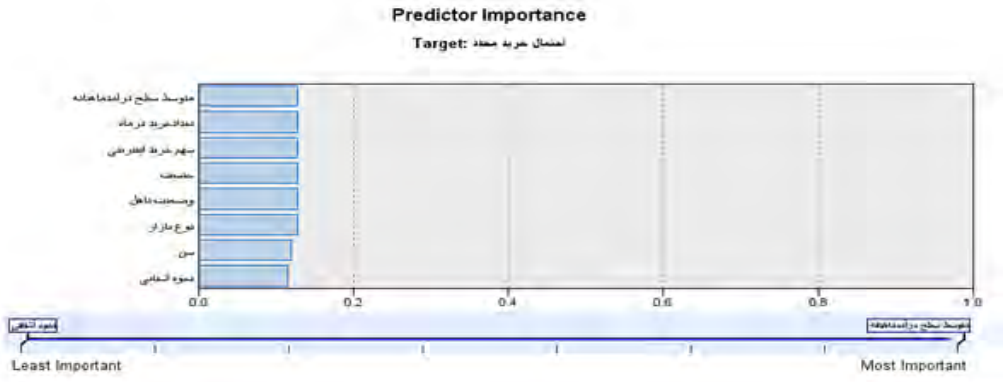


شکل ۲: نمودار میله‌ای بررسی اثربخشی رویکرد انتخاب ویژگی (منبع: یافته‌های پژوهش)

Figure 2: Bar Graph of the Effectiveness of the Feature Selection Approach

۲-٤. بررسی اثربخشی رویکرد طبقه بندی با استفاده از الگوریتم درخت تصمیم C5.0

در پژوهش حاضر برای ایجاد درخت تصمیم، داده‌ها به دو بخش آموزش و آزمون با نسبت ۳ به ۱ تقسیم و از الگوریتم درخت تصمیم انتخابی C5.0 استفاده شده است. همچنین، از متغیر احتمال خرید مجدد به‌عنوان متغیر هدف (متغیر وابسته) و از متغیرهای شناسایی‌شده با گره انتخاب ویژگی درقالب متغیرهای بااهمیت به‌عنوان متغیرهای ورودی (متغیرهای مستقل) برای ساخت درخت تصمیم C5.0 استفاده شده است. درواقع، درخت تصمیم C5.0 با بررسی تمامی فیلدهای پایگاه داده قصد دارد تا به فیلدی برسد که بهترین دسته‌بندی و پیش‌بینی را با تقسیم داده‌ها به زیرگروه‌ها انجام دهد. این فرآیند به‌صورت بازگشتی تکرار می‌شود تا بازهم زیرگروه‌ها به زیرگروه‌های دیگری شکسته شود. در این فرآیند زیرنمونه‌ها مجدد شاخه زده می‌شود و تا زمانی ادامه می‌یابد که نتوان زیرنمونه‌ها را شاخه زد. در پایان، پایین‌ترین سطح شاخه‌ها آزموده می‌شود و با حذف یا هرس شاخه‌هایی که نقش مهمی در مدل ندارد، عملیات پایان می‌یابد (علیزاده و ملک محمدی، ۱۳۹۳). باتوجه به نتایج به‌دست‌آمده در شکل ۳ بااهمیت‌ترین متغیرها با گره انتخاب ویژگی انتخاب و به‌ترتیب میزان اهمیت آنها شناسایی شده است که بااهمیت ترین آنها متوسط سطح درآمد ماهیانهٔ مشتریان را نشان می‌دهد.

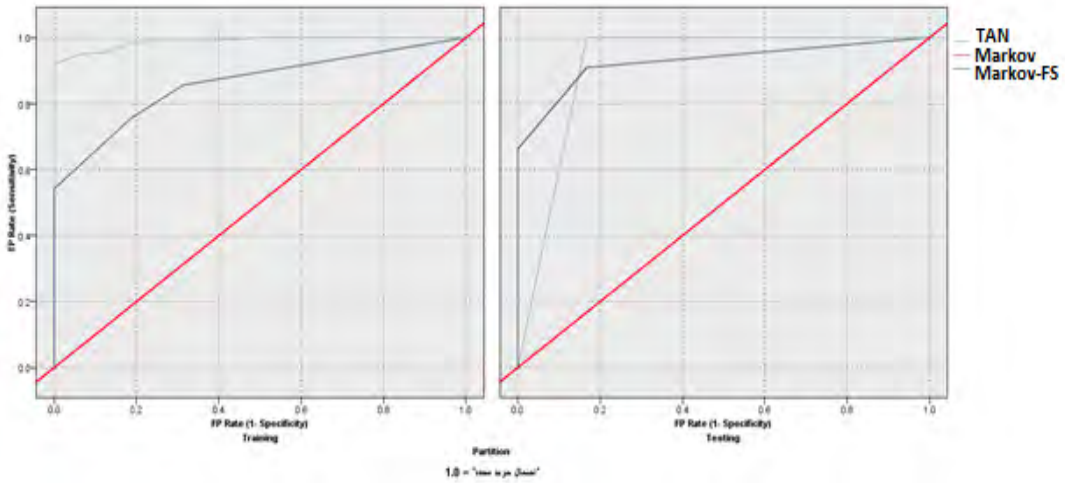


شکل ۳: نتایج طبقه‌بندی با الگوریتم درخت تصمیم C5.0 (منبع: یافته‌های پژوهش)

Figure 3: Classification Results with C5.0 Decision Tree Algorithm

۳-٤. بررسی اثربخشی رویکرد طبقه بندی با استفاده از شبکهٔ بیزین

در پژوهش حاضر با استفاده از الگوریتم شبکهٔ بیزین و درخت تصمیم، مدل و قوانین رویکرد طبقه‌بندی در راستای احتمال خرید مجدد مشتریان ارائه و استخراج شده است. پس از جمع‌آوری داده‌های مربوط به متغیرها به ساخت مجموعهٔ آموزش و آزمون پرداخته شده است که ۷۵ درصد داده‌ها به‌عنوان مجموعه آموزش و ۲۵ درصد به عنوان مجموعه آزمون به‌صورت تصادفی انتخاب شده است. داده‌های بخش آموزش مدل را می‌سازند و داده‌های بخش آزمون نیز به ارزیابی مدل ایجادشده می‌پردازند. درواقع، یکی از نیازمندی های شبکهٔ بیزین و درخت تصمیم، فیلد هدف طبقه‌ای است. فیلد هدف (متغیر وابسته) این پژوهش احتمال خرید مجدد و فیلدهای ورودی ۱۷ متغیر مستقل است. به‌طور معمول، در پژوهش‌های مرتبط با مدل‌سازی برای بررسی میزان دقت مدل و میزان پیش‌بینی از نمودار receiver operating characteristics (ROC) استفاده می‌شود. سطح زیر منحنی، میزان توانایی مدل را در تفاوت قائل شدن بین چند نتیجه نشان می‌دهد که میزان تمایز نامیده می‌شود. هرچقدر عدد سطح زیرین منحنی ROC به ۱ نزدیک‌تر باشد، می‌توان نتیجه گرفت که دقت مدل در معیار خوب است و برعکس. همچنین، هرچه این عدد به ۰/۵ نزدیک‌تر باشد، می‌توان نتیجه گرفت که دقت مدل پایین است و پیش‌بینی نامناسب دارد. درادامه، طبقه‌بندی داده‌ها با سه الگوریتم شبکهٔ بیزین به نام‌های TAN MARKOV و MARKOV-FS بررسی شده است و باتوجه به مقایسهٔ دقت به‌دست‌آمده در نمودار ROC (شکل ۴) می‌توان نتیجه گرفت که تابع TANاز دو تابع MARKOV و MARKOV-FS عملکرد بهتری دارد.

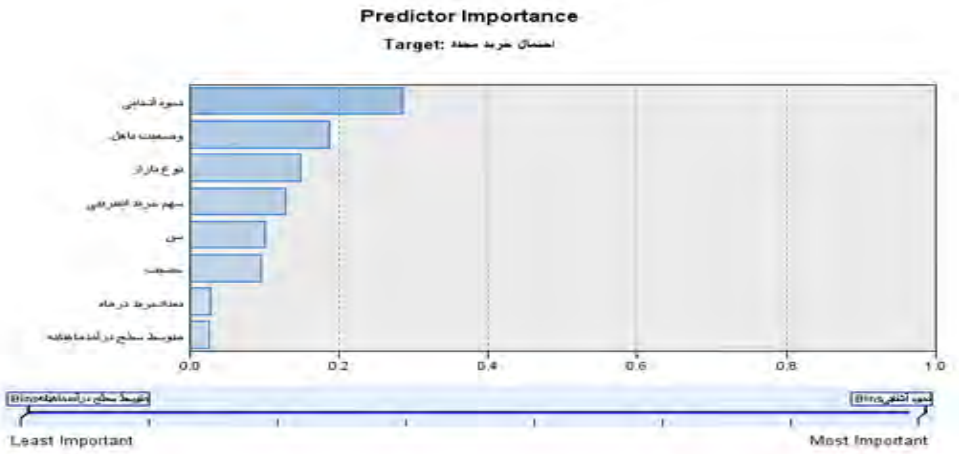


شکل ٤: مقایسهٔ دقت سه الگوریتم شبکهٔ بیزین (منبع: یافته‌های پژوهش)

Figure 4: Comparison of the Accuracy of Three Bayesian Network Algorithms

باتوجه به نتایج به‌دست‌آمده در نمودار ROC برای اینکه بتوانیم شبکه‌ای با کیفیت بالا داشته باشیم باید از الگوریتم جست‌وجوی شبکهٔ بیزین TAN استفاده شود. عملکرد الگوریتم جست‌وجوی TAN به‌گونه‌ای است که با ایجاد یک شبکهٔ بیزین ساده این امکان را فراهم می‌کند که هر پیش‌بینی‌کننده علاوه‌بر وابستگی به متغیر

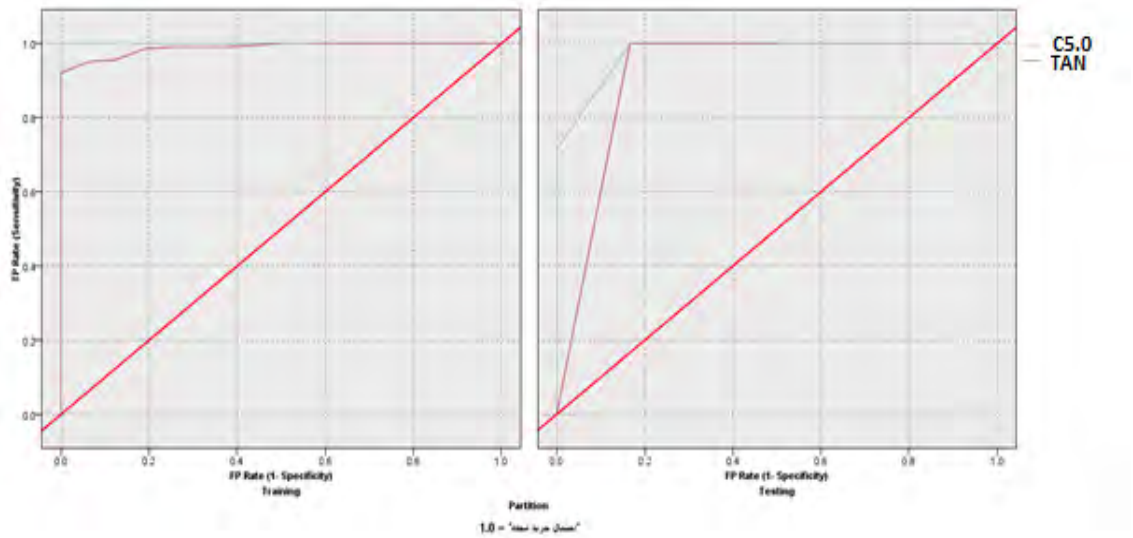
هدف بتواند به دیگر پیش‌بینی کننده‌ها نیز وابستگی داشته باشد؛ بنابراین ساختار TAN باعث می‌شود تا دقت طبقه‌بندی افزایش یابد (علیزاده و ملک محمدی، ۱۳۹۳). شکل ۵، نمودار اهمیت طبقه‌بندی متغیرهای به‌دست‌آمده را با گره انتخاب ویژگی نشان می‌دهد.



نمودار ۵: طبقه‌بندی متغیرها باتوجه به میزان اهمیت آنها با شبکهٔ بیزین (منبع: یافته‌های پژوهش)

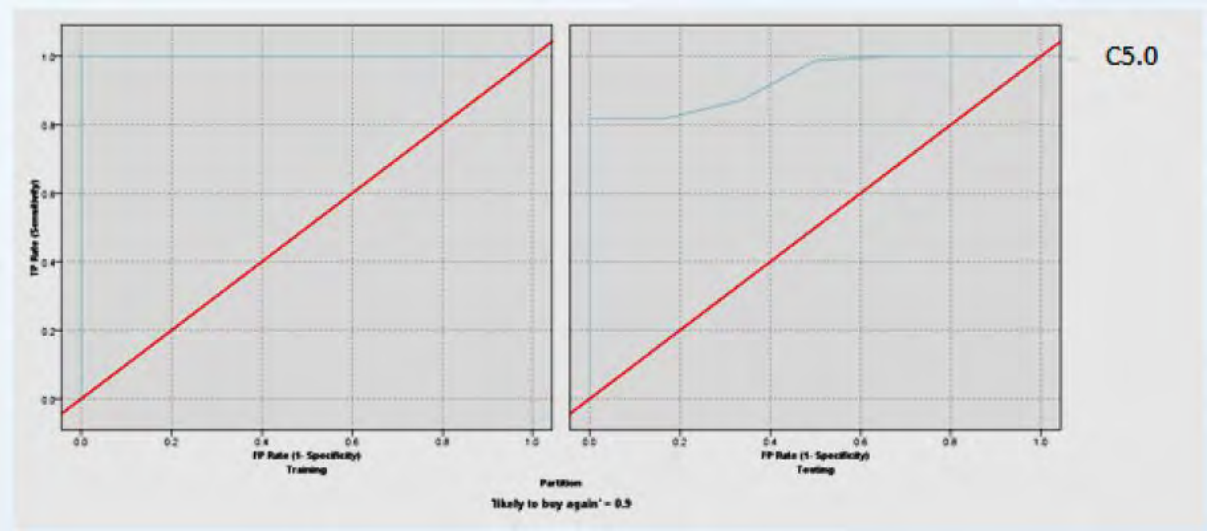
Diagram 5: Classification of Variables According to Their Importance with Bayesian Network

در شکل‌های ۶ و ۷ مشخصهٔ عملکرد سیستم ROC برای الگوریتم‌ها آورده شده است. این نتایج بیانگر آن است که ازنظر سطح زیر منحنی ROC مدل درخت تصمیم C5.0 عملکرد بهتری در پیش‌بینی روی‌گردانی مشتریان دارد.



شکل ٦: نمودار ROC مقایسهٔ دو الگوریتم درخت تصمیم C5.0 و شبکهٔ بیزین (منبع: یافته‌های پژوهش)

Figure 6: ROC Diagram Comparing Two C5.0 Decision Tree Algorithms and Bayesian Network



شکل ۷: نمودار ROC درخت تصمیم C5.0 (منبع: یافته‌های پژوهش)

Figure 7: ROC Diagram of C5.0 Decision Tree

همچنین، مطابق جدول ۴ مقدار خروجی و یا مقدار پیش‌بینی‌شده که با مدل ارائه‌شده به‌دست‌آمده، نشان می‌دهد درخت تصمیم C5.0 با میزان دقت ۹۷/۵۹ درصد نسبت به شبکهٔ بیزین با میزان دقت ۹۳/۹۸ درصد دقت بیشتری را دارد. به‌طورکلی، باتوجه به دو معیار دقت طبقه‌بندی و منحنی ROC می‌توان نتیجه گرفت که مدل درخت تصمیم C5.0 دقت بیشتری دارد.

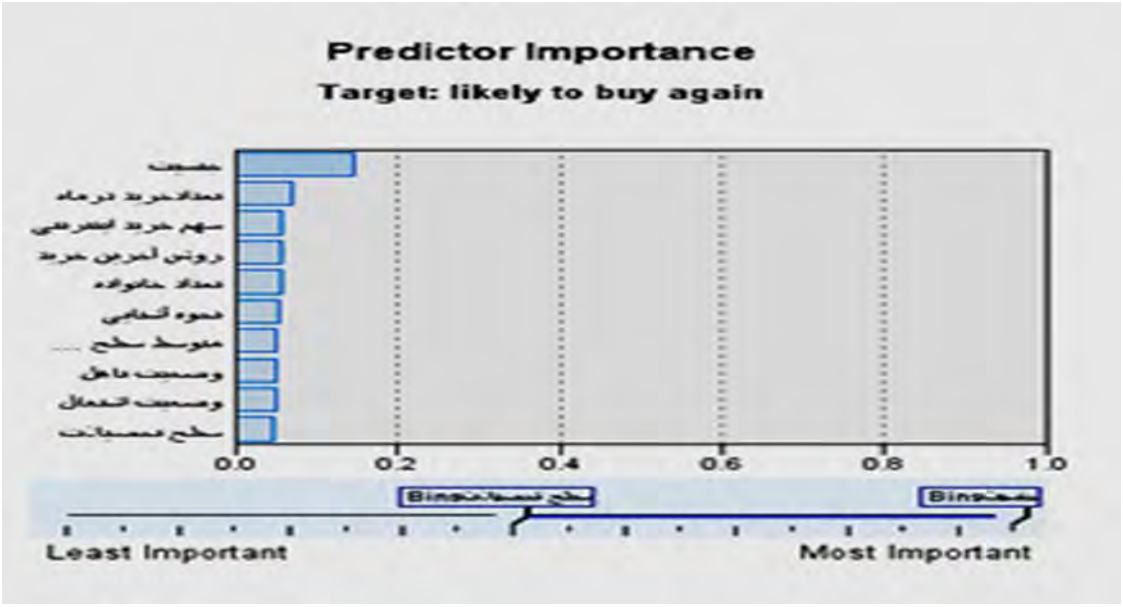
جدول ٤: مقایسهٔ دقت دو مدل ارائه‌شده

Table 4: Comparison of the Accuracy of the Two Presented Models

الگوریتم	مجموعه داده یادگیری		مجموعه داده آزمون	
	درصد پیش‌بینی درست	درصد پیش‌بینی نادرست	درصد پیش‌بینی درست	درصد پیش‌بینی نادرست
درخت تصمیم C5.0	۱۰۰	۰	۹۷/۵۹	۲/۴۱
شبکهٔ بیزین	۹۶/۷۷	۳/۲۳	۹۳/۹۸	۶/۰۲

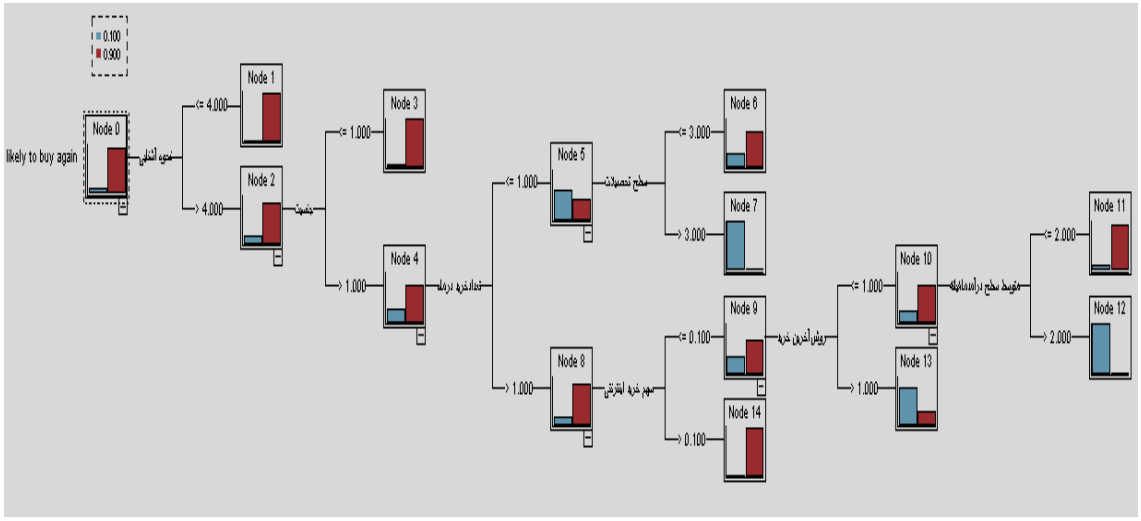
منبع: یافته‌های پژوهش

درنهایت، پس از اولویت‌بندی معیارها با الگوریتم شبکهٔ بیزین (شکل ۸)، نتایج حاصل به‌عنوان ورودی درخت تصمیم C5.0 به کار گرفته می‌شود. نتایج حاصل از پیاده‌سازی درخت تصمیم در شکل ۹ نمایش داده شده است. براساس قوانین استخراج‌شده و متغیرهای تأثیرگذار بر پیش‌بینی، آن دسته از مشتریان که متوسط سطح درآمد پایین‌تری دارند و به تعبیری میزان خرید کمی از سازمان دارند، احتمال بیشتری دارد که از سازمان دریافت‌کنندهٔ خدمت روی‌گردان شوند. ازسویی برخلاف آنچه تصور می‌شد، متغیر نحوهٔ آشنایی جزء کم‌تأثیرترین متغیرها بر پیش‌بینی روی‌گردانی مشتریان است.



شکل ۸: ویژگی‌های انتخاب‌شده با الگوریتم شبکهٔ بیزین (منبع: یافته‌های پژوهش)

Figure 8: Selected Features with Bayesian Network Algorithm



شکل ۹: نتایج طبقه‌بندی با الگوریتم درخت تصمیم C5.0 (منبع: یافته‌های پژوهش)

Figure 9: Classification Results with C5.0 Decision Tree Algorithm

٤-٤. پیشنهادهای مدیریتی

نتایج پژوهش حاکی از آن است که مشخصه‌های دموگرافیک و سوابق خرید مشتریان در رفتار روی‌گردانی مشتریان اثرگذار است. مبتنی بر نتایج حاصل از طبقه‌بندی در پژوهش حاضر و قوانین ارائه‌شده با درخت تصمیم (شکل *)، ۸ عامل کلیدیِ شناسایی‌شده در روی‌گردانی مشتریان با تأثیر فراوانی برای شهرستان مشهد بررسی و درادامه، تحلیل شده است. پیشنهادهای کاربردی به تأیید خبرگان پژوهش رسیده است.

٤-١-١. متوسط سطح درآمد ماهیانه

نتایج طبقه‌بندی انجام‌شده نشان می‌دهد که افراد با درآمد ماهیانه پنج میلیون تومان و بیشتر جزء آن دسته از

مشتریانی هستند که امکان خرید مجدد آنها وجود دارد. از سوی مقابل، افراد با درآمد ماهیانه کمتر از پنج میلیون تومان جزء مشتریان روی‌گردان هستند؛ از این رو پیشنهادهای کاربردی ذیل ارائه می‌شود.

- پیشنهاد می‌شود به تأمین کالاهای مشابه با گروه قیمتی متفاوت دقت شود تا مشتریان با درآمد پایین نیز توانایی خرید را داشته باشند و درنهایت، از روی‌گردانی مشتریان فروشگاه جلوگیری شود.
- پیشنهادهای قیمتی ویژه برای خرید مانند انواع روش‌های تخفیف مبتنی بر اندازهٔ خرید، زمان و حجم خرید برای مشتریان با سطح درآمد پایین‌تر ارائه شود تا این دسته از مشتریان به خرید از فروشگاه تمایل بیشتری پیدا کنند و بدین نحو از روی‌گردانی آنها اجتناب شود.

۴-۱-۲. تعداد خرید در ماه

نتایج پژوهش حاکی از آن است که افرادی که تعداد خرید آنها بیشتر از ۲ بار در ماه بوده است، جزء آن دسته از مشتریان وفادار قرار می‌گیرند؛ زیرا احتمال خرید مجدد آنها وجود دارد و آن دسته از مشتریانی که احتمال خرید آنها کمتر از ۲ بار بوده است، امکان خرید مجدد آنها وجود ندارد و جزء مشتریان روی‌گردان خواهند بود؛ از این رو پیشنهادهای کاربردی ذیل ارائه می‌شود.

- برگزاری جشنواره‌های فروش ویژه نظیر جشنوارهٔ پایان فصل و ارائهٔ تخفیف‌های جذاب و اطلاع‌رسانی ازطریق باشگاه مشتریان با هدف ترقیب مشتریان به تکرار خرید برگزار شود؛
- طرح‌های تحفیف سفارشی‌سازی‌شده برای هر گروه از مشتریان باتوجه به اطلاعات کسب‌شده از سوابق خرید گذشته با هدف نزدیکی بیشتر به مشتریان و ایجاد وفاداری ارائه شود؛
- خدمات باارزش به هر دسته از مشتریان با عنوان باشگاه مشتریان و دسته‌بندی آنها درقالب‌های مختلف مانند طلایی، نقره‌ای و برنزی ارائه شود.

۴-۱-۳. سهم خرید اینترنتی در طی سال

نتایج کسب‌شده نشان می‌دهد که افراد با سهم خرید اینترنتی بیشتر از ۲۰ درصد جزء مشتریان وفادار و افراد با سهم خرید پایین‌تر جزء مشتریان روی‌گردان هستند. از آنجایی که امروزه افزایش فروش و جذب مشتریان جدید وابستگی زیادی به استراتژی‌های بازاریابی دیجیتال سازمان دارد، مجموعه پیشنهادهای کاربردی ذیل ارائه می‌شود.

- فعالیت و تولید محتوا در شبکه‌های اجتماعی با تعداد مخاطب زیاد انجام شود؛
- سایت اینترنتی و نرم‌افزارهای موبایلی برای خرید از فروشگاه و ارائهٔ اطلاعات از فعالیت‌های فروشگاه، طرح‌های تخفیف و غیره توسعه یابند؛
- طرح‌های فروش مانند تخفیف ویژهٔ خرید اینترنتی و یا تسهیلات ویژه برای تحویل خرید اینرنتی ارائه شود تا انگیزهٔ مشتریان به انتخاب خرید اینترنتی بیشتر شود.

۴-۱-۴. جنسیت

نتایج پژوهش نشان می‌دهد که میزان خرید بانوان بیشتر از آقایان بوده است و بانوان وفاداری بیشتری نسبت به آقایان داشته‌اند؛ از این رو پیشنهادهایی درزمینهٔ توجه ویژه به جامعهٔ مشتریان بانوان در تدوین طرح‌های فروش مانند ارائهٔ تخفیفات و تسهیلات برای تکرار خرید و افزایش دفعات و ارزش کلی خرید می‌شود. همچنین، پیشنهاد می‌شود که در ارسال محتوای تبلیغاتی و اطلاع‌رسانی به ویژگی جنسیت نیز توجه شود.

٥-١-٤. وضعیت تأهل

باتوجه به نتایج به‌دست‌آمده، افراد متأهل درصد بیشتری از مشتریان فروشگاه را دارند و نسبت به افراد مجرد وفاداری بیشتری نیز دارند؛ از این رو پیشنهاد می‌شود که در ارسال محتوای تبلیغاتی و اطلاع رسانی به ویژگی جنسیت نیز توجه شود.

٦-١-٤. نوع بازار

نتایج حاصل حاکی از آن است که میزان کلی خرید افراد از فروشگاه زنجیره‌ای بسیار بیشتر از سوپرمارکت سنتی است. دلیل این پدیده را می‌توان در ارائهٔ طیف متنوعی از محصولات هم‌خانواده به‌صورت متمرکز در یک محل فیزیکی جست‌وجو کرد. پاسخگویی به طیف گسترده از نیازهای مشتریان، ارائهٔ خدمات همزمان مانند محصولات مصرفی و غیرمصرفی در یک مکان، درنظرگرفتن خدمات تفریحی و سرگرمی برای مشتریان همچون رستوان، محل بازی و نگهداری کودکان و غیره همگی در ارتقا جذابیت خرید از یک فروشگاه زنجیره‌ای مؤثر است. از دیگر دلایل افزایش خرید افراد در این فروشگاه‌ها این است که افراد می‌توانند مدت زمان بیشتری و ارزش خرید بیشتری را در فروشگاه داشته باشند.

٧-١-٤. سن

نتایج نشان می‌دهد افرادی که بیش از ۲۰ سال دارند، امکان خریدشان بیشتر است و افرادی که کمتر از بیست سال دارند، امکان خرید مجددشان اندک است؛ از این رو پیشنهاد می‌شود برای جذب مشتریان در بازهٔ سنی کمتر، ضمن شناسایی نیازهای مصرفی خریداران (نظیر گروه‌ها و سبک‌های کالایی تقاضاشده) به این جامعه از مشتریان به‌عنوان مشتریان بالقوهٔ آینده نیز توجه شود.

٨-١-٤. نحوهٔ آشنایی با فروشگاه

به‌طور کلی، چهار روش آشنایی افراد با فروشگاه در این پژوهش بررسی شده است (جدول ۵).

جدول ٥: روش‌های آشنایی مشتریان با فروشگاه

Table 5: Ways to Familiarize Customers with the Store

شمارهٔ روش	نام روش
۱	دوستان و آشنایان
۲	کارکنان فروشگاه
۳	صفحات مجازی فروشگاه
۴	تبلیغات تلویزیونی

منبع: یافته‌های پژوهش

نتایج حاکی از آن است که آشنایی افراد با فروشگاه ازطریق معرفی دوستان و آشنایان، کارکنان فروشگاه، صفحات مجازی و تبلیغات تلویزیونی به‌ترتیب بیشترین اهمیت را در آشنایی افراد با فروشگاه و امکان خرید آنها دارد؛ از این رو، پیشنهاد استفاده از ابزار تبلیغات دهان‌به‌دهان ارائه می‌شود. همچنین، پیشنهاد می‌شود سرمایه‌گذاری در تبلیغات باتوجه به درجهٔ اهمیت هریک از شیوه‌های آشنایی باشد. در این میان، اثر تبلغیات دهان‌به‌دهان مهم است. این در حالی است که این اثر سنتی در فضای مجازی نیز نقش بازی کرده است و برنامه‌ریزان باید به اثر دهان‌به‌دهان اکترونیکی توجه کنند.

۵. نتیجه‌گیری

تحلیل و پیش‌بینی رفتار مشتریان اهمیت فراوانی دارد؛ زیرا هزینهٔ از دست دادن مشتری برای یک سازمان بسیار زیاد است. در همین راستا، در پژوهش حاضر کوشش شد تا برای پیش‌بینی تحلیل روی‌گردانی مشتریان، روش ترکیبی شبکهٔ بیزین و درخت تصمیم C5.0 توسعه داده شود. برای این منظور، داده‌های مشتریان یک فروشگاه زنجیره‌ای در شهرستان مشهد به‌عنوان مطالعهٔ موردی بررسی شد. متغیر احتمال خرید مجدد مشتریان به‌عنوان متغیر وابسته درنظرگرفته و سپس بااهمیت‌ترین متغیرهای مستقل برای پیاده‌سازی درخت تصمیم C5.0 و شبکهٔ بیزین با گره انتخاب ویژگی واقع در نرم‌افزار IBM SPSS MODELER 18.0 شناسایی شد. نتایج پژوهش حاضر نشان می‌دهد که به‌کارگیری الگوریتم انتخاب ویژگی می‌تواند دقت مدل طبقه بندی را بهبود دهد. همچنین، انتخاب بهترین مدل طبقه‌بندی و تمرکز بر روی متغیرها با بالاترین اهمیت بر کیفیت پیش‌بینی روی‌گردانی مشتریان اثرگذار است. نتایج پژوهش حاکی از آن است که ۸ عامل سن، وضعیت تأهل، متوسط سطح درآمد ماهیانه، تعداد خرید در ماه، نحوهٔ آشنایی با فروشگاه، نوع بازار، سهم خرید اینترنتی و فروش ویژه از مهم‌ترین عوامل مؤثر بر روی‌گردانی است. باتوجه به مقایسهٔ دو الگوریتم شبکهٔ بیزین و درخت تصمیم‌گیری C5.0 که مبتنی بر نتایج نمودار ROC[1] است، نتایج حاصل بر این مسئله تأکید دارد که درخت تصمیم C5.0 با بیشترین دقت، عملکرد بهتری در شناسایی مشتریان روی‌گردان دارد. درنهایت، در پژوهش حاضر مجموعه‌ای از پیشنهادهای کاربردی مبتنی بر نتایج پژوهش برای تدوین طرح‌های بازاریابی، فروش و مواجه با انواع مشتریان ارائه شد.

برای انجام‌دادن پژوهش‌های آتی پیشنهاد می‌شود که طیف گسترده‌تری از متغیرهای اثرگذار مرتبط با مشتریان (هزینه و زمان دسترسی به فروشگاه)، فروشگاه ارائه‌دهندهٔ خدمت (تنوع ارائهٔ محصولات مصرفی و غیرمصرفی، خدمات رفاهی و سرگرمی) و رقبا (کیفیت خدمات ارائه‌شدهٔ رقبا و طرح‌های فروش) در توسعهٔ مدل پیش‌بینی درنظر گرفته شود. علاوه بر آن، استفاده از دیگر روش‌های پیش‌بینی و ارزیابی عملکرد آن نیز به محققان پیشنهاد شود.

منابع

اسماعیلی، مهدی (۱۳۹۲). *مفاهیم و تکنیک‌های داده‌کاوی*. کاشان: دانشگاه آزاد اسلامی واحد کاشان.

امامی، لطیف، پوراشرف، یاسان اله، و طولابی، زینب (۱۳۹۴). ارائهٔ مدلی برای رویگ‌ردانی مشتریان از بانک ملی با استفاده از معادلات ساختاریافته (مطالعهٔ موردی: شعب بانک ملی استان ایلام). *مدیریت بازاریابی*، ۱۰ (۲۶)، ۲۵-۴۶.

بحرینی‌نژاد، اردشیر، و سروش، علیرضا (۱۳۸۸). *هوشمندی کسب وکار و داده کاوی: یک استراتژی برای به‌کارگیری داده‌ها و برگشت سرمایه*. تهران: انتشارات ناقوس.

حبیبی‌پور، اعظم، طالبی، علی، کریمیان، علی‌اکبر، دهقانی، فرهاد، و مختاری، محمدحسین (۱۳۹۶). تعیین روش بهینهٔ پیش‌پردازش داده‌ها به‌منظور افزایش دقت شبیه‌سازی‌های شوری خاک سطحی (مطالعهٔ موردی: منطقهٔ مروست). *آب و خاک*، ۳۱ (۳)، ۹۱۵-۹۲۸. Doi: 10.22067/jsw.v31i3.55462

صنیعی آباده، محمد، محمودی، سینا، و طاهر پرور، محدثه (۱۳۹۱). *داده‌کاوی کاربردی*. تهران: انتشارات نیاز دانش.

علیـزاده، سـمیه، و ملـک محمـدی، سـمیرا (۱۳۹۳). *داده‌کاوی و کشف دانش گام‌به‌گام با نـرم‌افـزار Clementine*. تهـران: انتشـارات دانشـگاه صـنعتی خواجه‌نصیرالدین طوسی.

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 1-24. Doi: 10.1186/s40537-019-0191-6

Alizadeh, S., & Malek Mohammadi, S. (2014). *Data mining and knowledge discovery step by step with Clementine software*. Tehran: Khajeh Nasiruddin Tousi University of Technology [In Persian].

Amin, A., Adnan, A., & Anwar, S. (2023). An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes. *Applied Soft Computing*, *137*, 110103. Doi: 10.1016/j.asoc.2023.110103

Baghla, S., & Gupta, G. (2022). Performance evaluation of various classification techniques for customer churn prediction in e-commerce. *Microprocessors and Microsystems*, *94*, 104680. Doi: 10.1016/j.micpro.2022.104680

Bahraini Nejad, A., & Soroush, A. (1388). *Business intelligence and data mining: A strategy for data utilization and return on investment*. Tehran: Naghoos Publication [In Persian].

Brown, S. A., & Coopers, P. W. (1999). *Customer relationship management: A strategic imperative in the world of e-business*. John Wiley & Sons.

Chung, B. D., Park, J. H., Koh, Y. J., & Lee, S. (2016). User satisfaction and retention of mobile telecommunications services in Korea. *International Journal of Human-Computer Interaction*, *32*(7), 532-543. Doi: 10.1080/10447318.2016.1179083

Emami, L., Pourashraf, Y., & Toulabi, Z. (2016). A model for customers switching from national bank by applying structured equations (Case study: National bank branches in Ilam province). *Journal of Marketing Management*, *10*(26), 25-46 [In Persian].

Fenton, N. E., & Neil, M. D. (2007). *Managing risk in the modern world: Applications of Bayesian networks*. London Mathematical Society.

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing*, *64*(3), 65-87. Doi: 10.1509/jmkg.64.3.65.18028

Gattermann-Itschert, T., & Thonemann, U. W. (2022). Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests. *Journal of Industrial Marketing Management*, *107*, 134-147. Doi: 10.1016/j.indmarman.2022.09.023

Geiler, L., Affeldt, S., & Nadif, M. (2022). An effective strategy for churn prediction and customer profiling. *Journal of Data & Knowledge Engineering*, *142*, 102100. Doi: 10.1016/j.datak.2022.102100

Habibipoor, A., Talebi, A., Karimian, A. A., Dehghani, F., & Mokhtari, M. H. (2017). Investigation of the optimal method of data processing to increase the accuracy of simulation of surface soil salinity (Case study: Marvast). *Water and Soil*, *31*(3), 915-928. Doi: 10.22067/jsw.v31i3.55462 [In Persian].

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed). University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

Ismaili, M. (2012). *Data mining concepts and techniques*. Kashan: Islamic Azad University Press [In Persian].

Keramati, A., Ghaneei, H., Mirmohammadi, S.

customer reviews. *Journal of Air Transport Management*, *83*, 101760. Doi: 10.1016/j.jairtraman.2019.101760

Mohanty, R., & Naga Ratna Sree, C. (2018). Churn and non-churn of customers in banking sector using extreme learning machine. In *Proceedings of the Second International Conference on Computational Intelligence and Informatics: ICCII 2017* (pp. 51-58). Springer Singapore. Doi: 10.1007/978-981-10-8228-3_6

Mousavi, S. M., Sangari, M. S., & Keramati, A. (2018). An integrative framework for customer switching behavior. *The Service Industries Journal*, *38*(15-16), 1067-1094. Doi: 10.1080/02642069.2018.1428955

Nadkarni, S., & Shenoy, P. P. (2004). A causal mapping approach to constructing Bayesian networks. *Decision Support Systems*, *38*(2), 259-281. Doi: 10.1016/S0167-9236(03)00095-2

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Second Edition. New York, USA: Springer Science & Business Media.

Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence*, *32*(2), 245-257. Doi: 10.1016/0004-3702(87)90012-9

Rosa, N. B. D. C. (2019). *Gauging and foreseeing customer churn in the banking industry: A neural network approach*. Master Thesis. The New University of Lisbon

Sanii Abadeh, M., Mahmoudi, S., & Taher Parvar, M. (2011). *Applied data mining*. Tehran: Niaz Danesh Publication [In Persian].

Sayed, H., Abdel-Fattah, M. A., & Kholief, S. (2018). Predicting potential banking customer churn using Apache Spark ML and MLlib packages: A comparative study. *International Journal of Advanced* M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*, 10. Doi: 10.1186/s40854-016-0029-6

Khan, Y., Shafiq, S., Naeem, A., Ahmed, S., Safwan, N., & Hussain, S. (2019). Customers churn prediction using artificial neural networks (ANN) in telecom industry. *International Journal of Advanced Computer Science and Applications*, *10*(9), 132-142. Doi: 10.14569/IJACSA.2019.0100918

Kim, S., & Lee, H. (2022). Customer churn prediction in influencer commerce: An application of decision trees. *Procedia Computer Science*, *199*, 1332-1339. Doi: 10.1016/j.procs.2022.01.169

Kjaerulff, U. B., & Madsen, A. L. (2008). Bayesian networks and influence diagrams. *Springer Science & Business Media*, *200*, 114. Doi: 10.1007/978-0-387-74101-7

Kotler, P., & Keller, K. L. (2006). *Marketing management* (12th ed). New Jersey.

Kumar, V. A., & Elavarasan, N. (2014). A survey on dimensionality reduction technique. *International Journal of Emerging Trends and Technology in Computer Science*, *3*(6), 36-41.

Lin, C. S., Tzeng, G. H., & Chin, Y. C. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Journal of Expert Systems with Applications*, *38*(1), 8-15. Doi: 10.1016/j.eswa.2010.05.039

Ljubičić, K., Merćep, A., & Kostanjčar, Z. (2023). Churn prediction methods based on mutual customer interdependence. *Journal of Computational Science*, *67*, 101940. Doi: 10.1016/j.jocs.2022.101940

Lucini, F. R., Tonetto, L. M., Fogliatto, F. S., & Anzanello, M. J. (2020). Text mining approach to explore dimensions of airline customer satisfaction using online

Computer Science and Applications, *9*(11), 674-677.

Shobana, J., Gangadhar, C., Arora, R. K., Renjith, P. N., Bamini, J., & devidas Chincholkar, Y. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, *27*, 100728. Doi: 10.1016/j.measen.2023.100728

Shrestha, S. M., & Shakya, A. (2022). A customer churn prediction model using XGBoost for the telecommunication industry in Nepal. *Procedia Computer Science*, *215*, 652-661. Doi: 10.1016/j.procs.2022.12.067

Tomar, D., & Agarwal, S. (2014). A survey on pre-processing and post-processing techniques in data mining. *International Journal of Database Theory and Application*, *7*(4), 99-128.

Tsai, C. F., & Chiou, Y. J. (2009). Earnings management prediction: A pilot study of combining neural networks and decision trees. *Expert Systems with Applications*, *36*(3), 7183-7191. Doi: 10.1016/j.eswa.2008.09.025

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547-12553. Doi: 10.1016/j.eswa.2009.05.032

Wu, X., Li, P., Zhao, M., Liu, Y., Crespo, R. G., & Herrera-Viedma, E. (2022). Customer churn prediction for web browsers. *Expert Systems with Applications*, *209*, 118177. Doi: 10.1016/j.eswa.2022.118177

Xue, H., & Lu, W. C. (2011). Research of customer churn prediction model in a supermarket. In *2011 International Conference on E-Business and E-Government (ICEE)* (pp. 1-5). IEEE. Doi: 10.1109/ICEBEG.2011.5886886

Ye, L., Qiuru, C., Haixu, X., Yijun, L., & Guangping, Z. (2013). Customer segmentation for telecom with the k-means clustering method. *Information Technology Journal*, *12*(3), 409.