International Journal of Industrial Engineering & Production Management (2016)

December 2016, Volume 27, Number 4
pp. 635-653

http://IJIEPM.iust.ac.ir/

Applying Data Mining Techniques for Customer Churn Prediction in Insurance Industry

Samira Rezaei Navaei *& Hamidreza Koosha

Samira Rezaei Navaei*, Faculty of Industrial Engineering, Sadjad University of Technology*
Hamidreza Koosha*, Department of Industrial Engineering, Faculty of Engineering, Ferdowsi University of Mashhad Iran*

Keywords

Churn prediction,
Classification,
Support vector machine,
Feature Selection technique,
Data Mining

ABSTRACT

Due to the recent severe competition in insurance industry, customer churn prediction is a very important issue. This research applies some of data mining classification techniques for customer churn prediction in an insurance company. Researchers also apply support vector machine - an optimization based classification method- to predict customer churn in insurance industry for the first time. In this reseach, we applied genetic algorithm to find effective attributes. After modeling, the parameters of the model are optimized using grid search and k-fold cross-validation. Then the predictive performance of support vector machine is compared with decision tree, neural networks, logistic regression, random forest, naive bayes classifier, k-nearest neighbor. The results of the study show that the support vector machine performs better than the others and purchase history, how familiar the customer with the company and desire to purchase were determined as the most important customer churn predictors in our model. Finally, we suggested churn preventive actions according to the most important churn predictors.

نشـــریه بین المللـــی مهنـــدسی صنـــایع و مدیـــریت تولیـــد
شمـــاره ۴، جلـــد ۲۷، زمستان ۱۳۹۵
صفحـــه ۶۵۳ -۶۳۶
ISSN: 2008-4870
http://IJIEPM.iust.ac.ir/

به‌کارگیری و ارزیابی تکنیک‌های داده‌کاوی جهت پیش‌بینی رویگردانی مشتری در صنعت بیمه

سمیرا رضائی نوائی*، حمیدرضا کوشا

کلمات کلیدی

پیش‌بینی رویگردانی، دسته‌بندی، ماشین بردار پشتیبان، تکنیک انتخاب مشخصه، داده‌کاوی

چکیده:

با توجه به رقابتی شدن صنعت بیمه در سال‌های اخیر، توجه به پیش‌بینی رویگردانی مشتری در این صنعت اهمیت ویژه‌ای یافته است. در این مقاله، تعدادی از تکنیک‌های شناخته شده دسته‌بندی داده‌کاوی برای پیش‌بینی رویگردانی مشتری در صنعت بیمه به کار گرفته شده است. برای نخستین بار پیش‌بینی رویگردانی مشتری در یک سازمان بیمه‌ای با استفاده از یکی از رویکردهای مبتنی بر تحقیق در عملیات دسته‌بندی یعنی تکنیک ماشین بردار پشتیبان (SVM) انجام می‌شود. در این مقاله نخست از الگوریتم ژنتیک برای انتخاب مشخصه‌های تأثیرگذار استفاده شده است. پس از مدل‌سازی مسأله، پارامترهای مدل ماشین بردار پشتیبان با استفاده از دو روش جستجوی شبکه و اعتبارسنجی متقابل K لایه، بهینه می‌شوند. در ادامه عملکرد پیش‌بینی روش SVM با روش‌های درخت تصمیم، شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته‌بندی‌کننده بیزی و K نزدیک‌ترین همسایگی مقایسه شده است. یافته‌های این تحقیق نشان می‌دهد که روش ماشین بردار پشتیبان از عملکرد بالاتری نسبت به سایر روش‌ها برخوردار است. در مدل پیشنهادی مبتنی بر این روش، مشخصه‌های سابقه خرید، نحوه آشنایی با سازمان و تمایل به خرید، به عنوان مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی مشتری شناسایی شدند. در انتها با توجه به مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی، به ارائه راهکارهای پیشگیرانه پرداختیم.

۱. مقدمه

امروزه بیشتر سازمان‌ها، به‌دلیل وجود فضای اشباع‌شده و رقابتی شدید بازار، بر مدیریت ارتباط با مشتری[i] تمرکز کرده‌اند. رفتار آینده مشتریان، برای مدیریت ارتباط با مشتری بسیار مهم محسوب می‌شود. بنابراین، کشف تصمیم‌های آینده مشتریان توسط سازمان، به‌منظور انجام اقدامات به‌هنگام و پیشگیرانه، از اهمیت بالایی برخوردار است. از آن جایی‌که هزینه نگه‌داری **تاریخ** مشتریان موجود بسیار کمتر از هزینه جذب مشتریان جدید است، سازمان‌ها بیش از پیش به جلوگیری از رویگردانی مشتریان روی آورده‌اند [۱] [۲] [۳]. محققان به این نتیجه رسیده‌اند که اندکی تغییر در نرخ نگه‌داری، می‌تواند تأثیر بسیار زیادی بر بهبود تجارت بگذارد. پیش‌بینی رویگردانی، ابزاری مناسب برای توصیف فرایند نگه‌داری مشتریان یک سازمان به شمار رفته و هدف از آن، شناسایی گروهی از مشتریان که مستعد رویگردانی هستند، است. با شناسایی این گروه و انجام اقدامات پیشگیرانه می‌توان نقش مهمی در جلوگیری از رویگردانی مشتریان ایفا کرد. در چنین شرایطی، پیش‌بینی رویگردانی توجه زیادی را در مطالعات مدیریت و بازاریابی پیدا کرده است. به‌منظور مدیریت کارآمد پیش‌بینی رویگردانی در درون یک سازمان، ارائه مدل پیش‌بینی اثربخش و با صحت بالای رویگردانی، بسیار اهمیت دارد. مدل‌های پیش‌بینی‌کننده متعددی برای این کار وجود دارند. در این میان، تکنیک‌های داده‌کاوی می‌توانند به طور مؤثری مشتریان متمایل به رویگردانی را شناسایی کنند. این تکنیک‌ها می‌توانند الگوها و رابطه‌های درونی داده‌ها را کشف کرده و به

تاریخ وصول: ۹۳/۰۸/۰۳
تاریخ تصویب: ۹۴/۰۲/۰۹

سمیرا رضائی نوائی، دانشکده مهندسی صنایع، دانشگاه صنعتی سجاد، مشهد، samira.rezae@gmail.com
نویسنده مسئول: حمیدرضا کوشا، گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه فردوسی مشهد، koosha@um.ac.ir

دسته‌بندی و پیش‌بینی رفتار مدل براساس داده‌های در دسترس بپردازد. به عبارت دیگر، داده‌کاوی یک مبحث میان‌رشته‌ای بوده که با به‌کارگیری الگوریتم‌های مختلف، به کشف الگوهای پنهانی مجموعه داده‌های گسترده می‌پردازد[۴].

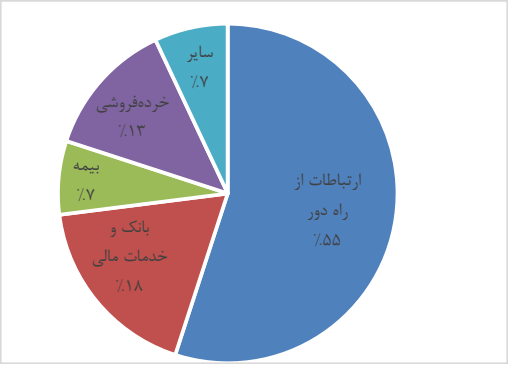
قابلیت‌های داده‌کاوی را می‌توان در چند دسته، تحت عنوان رویکردهای داده‌کاوی، دسته‌بندی کرد. در یک دسته‌بندی وسیع که توسط انگای و همکاران انجام شده است، رویکردهای داده‌کاوی در هفت گروه دسته‌بندی، خوشه‌بندی، پیش‌بینی، رگرسیون، کشف توالی، همبستگی و مصورسازی تعریف شده است[۵.] مسأله رویگردانی مشتریان، زیرمجموعه‌ای از مسائل مربوط به رویکرد دسته‌بندی بوده و از این رویکرد داده‌کاوی، در مقالات مربوط به پیش‌بینی رویگردانی استفاده می‌شود [۶]. در واقع دسته‌بندی عبارت است از ساخت مدل‌هایی برای پیش‌بینی رفتار آتی پدیده مورد مطالعه، از طریق نگاشت رکوردهای پایگاه داده به تعدادی دسته‌های از پیش تعریف‌شده، براساس معیارهای معین [۵].

با توجه به مطالعات پیشین، از تکنیک‌های دسته‌بندی متفاوتی مانند درخت تصمیم، شبکه‌های عصبی، ماشین بردار پشتیبان [i i]، رگرسیون لجستیک و غیره برای تحقق هدف پیش‌بینی رویگردانی استفاده می‌شود. همچنین باید این نکته را نیز در نظر داشت که بیشتر تحقیقات انجام‌شده در این زمینه، به دسته‌بندی مشتریان و مقایسه صحت تکنیک‌های مختلف در پیش‌بینی رویگردانی در صنایع مختلف پرداخته و تحقیقات کمتری در زمینه شناسایی دلایل رویگردانی مشتریان انجام شده است [۷]. با توجه بررسی‌های انجام شده، می‌توان مطالعات پیشین را از نظر صنعت مورد مطالعه، تکنیک مورد استفاده و نوع بررسی رویگردانی مقایسه کرد. خلاصه بررسی‌های انجام شده روی ۴۹ مقاله مربوط به پیش‌بینی رویگردانی مشتری، در شکل‌های ۱، ۲ و ۳ آمده است.

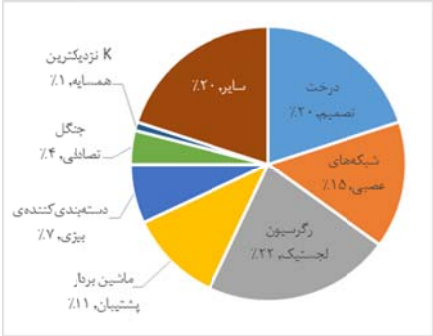
از جمله تحقیقات مربوط به دسته‌بندی مشتریان و مقایسه‌ی صحت تکنیک‌های مختلف دسته‌بندی در پیش‌بینی رویگردانی مشتری می‌توان به مادن و همکاران [i i i] اشاره کرد که از مدل پرابیت دوجمله‌ای [i v] برای بیان احتمال رویگردانی مشتریان در شرکت‌های سرویس‌دهنده اینترنت استفاده کردند [۸]. وی و چیو [v] با استفاده از تکنیک درخت تصمیم، مشتریان صنعت مخابرات را به دو گروه فعال و غیر فعال دسته‌بندی و قدرت پیش‌بینی مدل ارائه‌شده را نیز رضایت‌بخش ارزیابی کرده‌اند [۹]. بارز و ون دن پل [vi] نیز به مقایسه صحت سه تکنیک رگرسیون لجستیک، زنجیره مارکوف و جنگل تصادفی در پیش‌بینی رویگردانی مشترکین تلویزیون‌های پولی پرداخت [۱۰]. گلیدی و همکارانش [v i i] به مقایسه صحت پیش‌بینی رویگردانی مشتریان یک سازمان خدمات مالی با استفاده از روش‌های رگرسیون لجستیک، شبکه‌های عصبی، درخت تصمیم و دو الگوریتم حساس به هزینه آداکاست [v i i i] و درخت تصمیم حساس به هزینه [i x] پرداختند [۱۱]. شیم و همکاران [x] از تکنیک‌های درخت تصمیم، شبکه‌های عصبی و رگرسیون لجستیک برای دسته‌بندی و پیش‌بینی رویگردانی مشتریان خریدهای اینترنتی استفاده کردند [۱۲]. چن و همکاران رویکرد ماشین بردار پشتیبان با کرنل چندگانه سلسله‌مراتبی [xi] را ارائه داده و آن را با روش‌های ماشین بردار پشتیبان با کرنل چندگانه [x i i] و ماشین بردار پشتیبان مقایسه کردند [۱۳]. میگوئز و همکاران [x i i i] نیز به مقایسه‌ی عملکرد دو روش رگرسیون لجستیک و و رگرسیون تطبیقی چندمتغیره [x i v] در پیش‌بینی رویگردانی مشتریان صنعت خرده‌فروشی پرداختند [۶]. کیم و همکاران [xv]، با استفاده از اطلاعات تماس مشتریان با یکدیگر و آموزش مدل با استفاده از دو روش رگرسیون لجستیک و شبکه‌های عصبی، قدرت پیش‌بینی مدل را بهبود دادند [۱۴].

تعدادی از تحقیقاتی که به شناسایی دلایل رویگردانی مشتری پرداختند، عبارتند از: ون دن پل ولاریویر [xvi] که با استفاده از روش تحلیل بقا [x v i i] دریافتند که مشخصات جمعیت‌شناختی و تغییرات محیطی، از مهم‌ترین نگرانی‌های تحلیل رویگردانی سازمان خدمات مالی مورد مطالعه است [۱۵]. کیم و یون [x v i i i] با استفاده از تکنیک رگرسیون لجستیک در پنج شرکت اپراتور عمده خدمات تلفن همراه دریافتند که عوامل وابسته به سطح رضایت او از ویژگی‌های خدمات اپراتور مانند: کیفیت تماس، سطح تعرفه‌ها، اعتبار نشان تجاری و طول زمان اشتراک از مشخصه‌های رویگردانی مشتریان هستند [۱۶]. آن و همکاران [x i x] با استفاده از روش رگرسیون لجستیک در صنعت مخابرات دریافتند که نارضایتی مشتریان، هزینه‌های تغییر و میزان مصرف خدمات بر تصمیم مشتریان مبنی بر ماندگاری یا رویگردانی مؤثر است [۱۷]. وانگ و همکاران [xx] در پژوهشی با استفاده از تکنیک درخت تصمیم روی داده‌های مشتریان یک شرکت ارتباطات بی‌سیم، دریافتند که دو متغیر فاصله زمانی از آخرین اتصال مشتری به شبکه اینترنت و تعداد دفعات اتصال، دارای قابلیت پیش‌بینی رفتار رویگردانی هستند [۱۸]. تیسای و چن [xxi] با استفاده از دو روش درخت تصمیم و شبکه‌های عصبی در یک شرکت خدمات چندرسانه‌ای دریافتند که درخت تصمیم بهتر از شبکه‌های عصبی عمل کرده و دو متغیر مدت زمان اتصال مشتری به شبکه و میزان تخفیف نرخ خدمات، مهم‌ترین عوامل پیش‌بینی‌کننده رویگردانی مشتری از شرکت هستند [۱۹]. نی و همکاران نیز با استفاده از درخت تصمیم و رگرسیون لجستیک دریافتند که متغیرهای مربوط به اطلاعات کارت و میزان

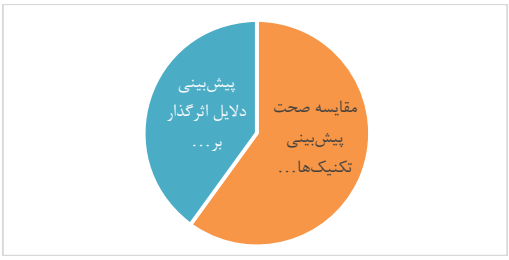
تعاملات، در مدل پیش‌بینی رویگردانی مشتریان کارت‌های اعتباری بسیار مؤثر هستند [۱].



شکل ۱. مقایسه مطالعات پیشین از دیدگاه صنعت



شکل ۲. مقایسه مطالعات پیشین از دیدگاه تکنیک



شکل ۳. مقایسه مطالعات پیشین از دیدگاه نوع بررسی رویگردانی

با بررسی شکل‌های (۱)، (۲) و (۳) درمی‌یابیم که بیشتر مطالعات مربوط به پیش‌بینی رویگردانی مشتری در صنایعی چون ارتباطات از راه دور و بانک پیاده سازی شده است و به‌ندرت در صنایعی چون بیمه و خرده‌فروشی دیده می‌شود. علی‌رغم عملکرد قابل قبول تکنیک‌های جنگل تصادفی، ماشین بردار پشتیبان و دسته‌بندی‌کننده‌ی بیزی با توجه مطالعات پیشینه، توجه کمتری نسبت به آن‌ها در مقایسه با سایر تکنیک‌ها شده است. همچنین اکثر مطالعات به مقایسه صحت پیش‌بینی تکنیک‌های مختلف پرداخته و در زمینه مطالعات مربوط به پیش‌بینی دلایل رویگردانی مشتری شکاف دیده می‌شود. شکاف بعدی نیز به توجه کمتر مطالعات به فرایند پیش‌پردازش داده‌ها و استفاده ناچیز از تکنیک‌های ترکیبی مربوط می‌شود. با توجه به شکاف‌های تحقیقاتی موجود، در این مطالعه توجه به روش ماشین بردار پشتیبان، به‌دلیل ماهیت کمینه‌سازی ریسک دسته‌بندی و عملکرد بالای آن [۲۰]، معطوف شده است. این روش یک تکنیک دسته‌بندی بر مبنای تحقیق در عملیات است که در سال ۱۹۹۵ توسط وپنیک xxi i ارائه شد. در حالت دسته‌بندی به دو دسته‌ی مجزا، ماشین بردار پشتیبان به‌دنبال یافتن ابرصفحه بهینه‌ای است که داده‌های دو دسته را با در نظر گرفتن بیشترین حاشیه‌ی بین آن‌ها دسته‌بندی کند. برای رسیدن به مدل ابرصفحه بهینه، باید یک مسأله بهینه‌سازی درجه دوم حل شود. در عمل امکان عدم دسته‌بندی خطی داده‌ها وجود دارد. بدین منظور داده‌ها با استفاده از یک نگاشت غیر خطی، از فضای ورودی xxi i i به فضای دیگر و با ابعاد بالاتر منتقل می‌شوند. این انتقال با استفاده از تابع کرنل xxi v انجام شده و در فضای جدید، داده‌ها به صورت خطی دسته‌بندی می‌شود [۲۰] و [۲۱]. در این پژوهش کاربرد تکنیک ماشین بردار پشتیبان را در صنعت بیمه بررسی خواهیم کرد و برای بهبود عملکرد این تکنیک نسبت به مطالعات پیشین، از الگوریتم ژنتیک برای انتخاب مشخصه‌های تأثیرگذار استفاده خواهیم کرد. همچنین، مقایسه عملکرد پیش‌بینی روش ماشین بردار پشتیبان با روش‌های درخت تصمیم، شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته‌بندی‌کننده بیزی و K نزدیک‌ترین همسایگی و بهینه‌سازی پارامترهای مربوط به کلیه روش‌های نام‌برده، به‌منظور عملکرد بهتر آن‌ها انجام می‌شود.

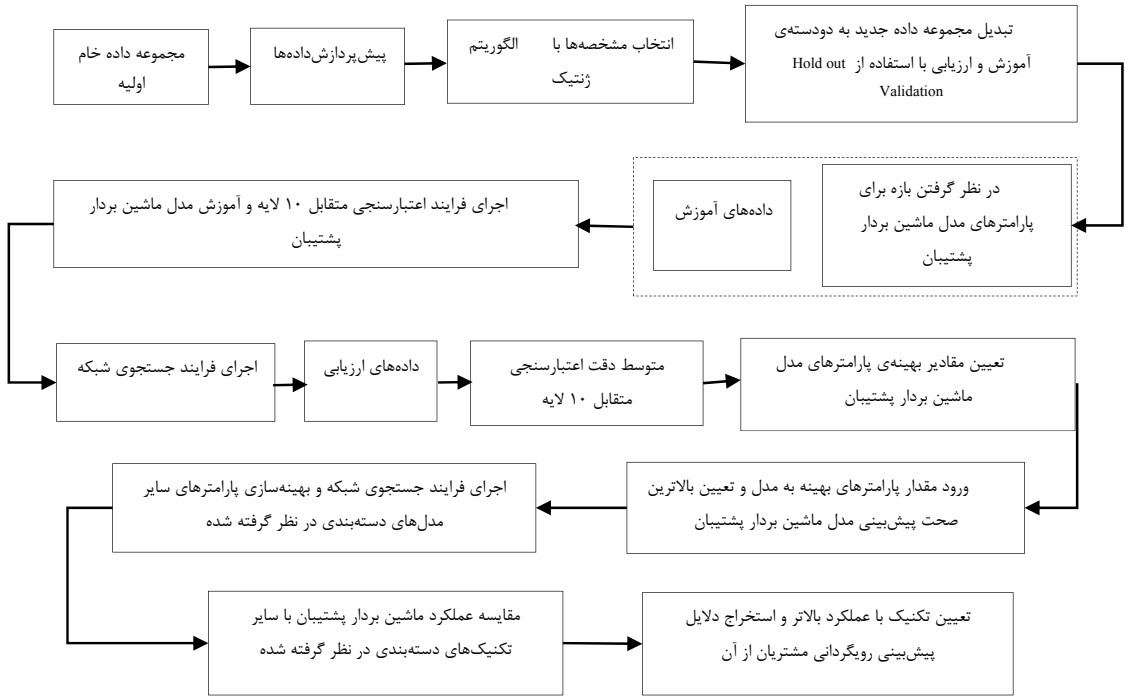
در این مقاله، ابتدا در قسمت بعد به فرایند پالایش داده‌ها پرداخته می‌شود و سپس، تجزیه و تحلیل داده‌ها و مدل‌سازی پیش‌بینی رویگردانی با روش ماشین بردار پشتیبان، تشریح می‌گردد. در این بخش توضیح مختصری درباره‌ی سایر تکنیک‌های دسته‌بندی برای مقایسه با ماشین بردار پشتیبان نیز بیان می‌شود. بعد از آن، پارامترهای مدل ماشین بردار پشتیبان بهینه می‌شوند. در ادامه فرایند ارزیابی و اعتبارسنجی مدل ارائه می‌شود. سپس، نتایج محاسباتی حاصل از مقاله بیان می‌شود. راهکارهای پیشگیرانه به منظور جلوگیری از رویگردانی مشتریان در ادامه ارائه می‌شود. در نهایت، نتیجه‌گیری و جمع‌بندی نیز بیان می‌گردد. شکل (۴) چارچوب اجرایی تحقیق را به صورت خلاصه نشان می‌دهد.

۱. روش تحقیق

روش انجام این تحقیق، بر مبنای یک فرایند استاندارد داده‌کاوی است. در این مقاله، مدل‌سازی پیش‌بینی رویگردانی مشتریان به عنوان هدف اساسی مطالعه حاضر بوده و این فرایند در دو گام اصلی پالایش داده‌ها و مدل‌سازی مسأله، اجرا می‌شود.

۱-۲. پالایش داده‌ها

در گام پالایش داده‌ها ابتدا باید به درک کلی از هدف مسأله و داده‌های مورد نیاز رسید. هدف اصلی مقاله، رسیدن به مدلی برای پیش‌بینی رویگردانی مشتریان بیمه بدنه خودرو است. متغیر خروجی (متغیر هدف)، وضعیت رویگردانی مشتریان تعریف شده و هدف، پیش‌بینی دسته‌ی مشتریان در دو دسته رویگردان و غیر رویگردان است. متغیرهای ورودی که پیش‌بینی‌کننده متغیر هدف هستند نیز به دو گروه کلی ویژگی‌های جمعیت‌شناختی و ویژگی‌های رفتاری مشتری تقسیم شدند. سپس برای رسیدن به درکی از داده‌های مورد نیاز، با توجه به درک پژوهش‌گر و نظرات مسئولان، متغیرهایی از پایگاه داده سازمان که مرتبط با هدف پژوهش بود، انتخاب شد. در این گام، یک بانک اطلاعاتی به تعداد ۳۵۱۸ رکورد از مشتریان بیمه بدنه خودرو که هر کدام بیان‌کننده عقد قرارداد بیمه بدنه خودرو در بازه زمانی ۱۳۹۰/۱/۱۵ تا ۱۳۹۳/۳/۱۴ بوده است، از سازمان مورد مطالعه گرفته شد. در این بانک، ۴۰ متغیر متمایز که حاوی اطلاعات جمعیت‌شناختی و رفتاری مشتریان بود، وجود دارد و برای درک بهتر نمونه کوچکی از متغیرهای ثبت‌شده در قرارداد بیمه بدنه خودرو به همراه دامنه تغییرات آن در جدول ۱ آمده است. بیان این نکته قابل ذکر است که اطلاعات مربوط به توضیحات برخی از متغیرها به‌دلیل محرمانه بودن آن در جدول (۱) نیامده است.



شکل ۴. چارچوب اجرایی تحقیق

جدول ۱. نمونه‌ای از متغیرهای ثبت‌شده در قرارداد بیمه بدنه خودرو و دامنه آن

ردیف	متغیر	توضیحات یا دامنه مقادیر	ردیف	متغیر	توضیحات یا دامنه مقادیر
۱	شماره بیمه‌نامه		۷	شماره پلاک	۰ الی ۷
۲	شماره بیمه‌نامه سال قبل		۸	تعداد سال عدم خسارت	
۳	تاریخ شروع بیمه‌نامه سال قبل		۹	شرکت بیمه سال قبل	

ردیف	متغیر	توضیحات یا دامنه مقادیر	ردیف	متغیر	توضیحات یا دامنه مقادیر
۴	تاریخ انقضا بیمه‌نامه سال قبل	۹۰/۱/۱۵ الی ۹۳/۳/۱۲	۱۰	مدل خودرو	۱۳۵۲ الی ۱۳۹۳
۵	تاریخ صدور	۹۰/۱/۱۵ الی ۹۳/۳/۱۴	۱۱	رنگ خودرو	
۶	تاریخ پایان	۹۱/۱/۱۵ الی ۹۴/۳/۱۴	۱۲	مدت قرارداد	۲۷ الی ۳۶۶ روز

سپس، متغیرهایی که اطلاعات آن‌ها برای رسیدن به هدف مفید بود، با توجه به درک پژوهش‌گر، از مجموعه داده خام استخراج شده است. در این پایگاه داده، خصیصه‌هایی مانند پلاک خودرو وجود داشت که در فرایند مدل‌سازی کمکی به محقق نخواهد کرد؛ زیرا نمی‌توان آن‌ها را به عنوان یک متغیر ورودی محتمل که در رویگردانی مشتریان اثر دارد، به حساب آورد. بنابراین این ستون‌ها در این مرحله حذف شدند. مشتریانی وجود داشتند که اطلاعات مربوط به تمامی ستون‌ها تکمیل نشده بود و بنابراین حذف شدند. مشتریانی بودند که

در سال آخر بازه در نظر گرفته شده به سازمان مورد مطالعه روی آورده بودند. بنابراین مشخص نبود که آیا این گروه از مشتریان در سال‌های آتی مشتری سازمان هستند یا خیر. بنابراین این مشتریان نیز از مجموعه داده حذف شدند. از دیگر فعالیت‌های انجام شده در مرحله پالایش داده، تعیین وضعیت رویگردانی مشتریان کنونی است. مشتریانی که در سال‌های متوالی بازه مورد بررسی بیمه‌نامه خود را تمدید کرده‌اند، در دسته مشتریان غیر رویگردان و سایر مشتریان در دسته مشتریان رویگردان قرار می‌گیرند. با توجه به حجم زیاد داده‌ها به کمک کد خودرو، که متمایز کننده مشتریان از هم بود، وضعیت تمدید قرارداد یا عدم تمدید را در سه سال متوالی بازه‌ی مورد بررسی جستجو کردیم و مشتریان در دسته‌های رویگردان و غیر رویگردان دسته‌بندی شدند. به عبارت دیگر همان‌طور که در قسمت مقدمه بیان شد، ماهیت رویکرد دسته‌بندی، تعریف دسته‌های از پیش تعیین شده و مشخص کردن جایگاه هر عضو در دسته خودش می‌باشد. در این رویکرد ما در ابتدا وضعیت رویگردانی مشتریان را مشخص می‌کنیم. سپس با استفاده از متغیرهای ورودی در نظر گرفته شده در صدد آنیم تا بتوان فهمید که با چه صحت پیش‌بینی می‌توان دسته مشتریان کنونی را پیش‌بینی کرد. به این معنا که مشتریان با چه صحتی در دسته خودشان پیش‌بینی خواهند شد. اگر پیش‌بینی ما صحت بالایی داشته باشد نشان‌دهنده این موضوع است که متغیرهای تعریف شده توانایی پیش‌بینی وضعیت رویگردانی را خواهند داشت. با توجه به ماهیت و نیاز روش‌های درخت تصمیم ID3، درخت تصمیم CHAID، جنگل تصادفی، دسته‌بندی‌کننده بیزی و K نزدیک‌ترین همسایگی، باید داده‌هایی از نوع اسمی به عنوان ورودی وارد این مدل‌ها شوند. بنابراین، در این مرحله تمام رکوردها به صورت اسمی تعریف شدند. سپس با توجه به ماهیت عددی داده‌ها در روش‌های ماشین بردار پشتیبان، شبکه‌های عصبی و رگرسیون لجستیک، تمامی مشخصه‌های اسمی تعریف‌شده در حالت قبل، کدگذاری شده و با استفاده از رابطه (۱) در بازه ۰ تا ۱ نرمال شدند [۲۲]. بنابراین در این مرحله دو مجموعه داده اسمی و عددی نرمال‌شده با توجه به نیاز تکنیک‌های مختلف ایجاد شدند.

$$x' = \frac{x - \min_A}{\max_A - \min_A} \quad (۱)$$

در این رابطه، x مقداری است که قرار است نرمال شود. minA ، کمترین مقدار در متغیر A، maxA بیشترین مقدار در متغیر A و 'x مقدار نهایی نرمال شده است.

اقدام دیگری که در این گام صورت گرفته است، انتخاب مشخصه‌های مهم از بین کلیه مشخصه‌ها بوده است. انتخاب مشخصه به فرایندی گفته می‌شود که به تعیین مرتبط‌ترین مشخصه‌ها با کم‌ترین خطای دسته‌بندی، از میان مجموعه مشخصه‌ها می‌پردازد. انتخاب مشخصه‌های مهم و حذف مشخصه‌های کم‌اهمیت و غیر ضروری، علاوه بر کاهش بعد داده، باعث افزایش صحت پیش‌بینی خواهد شد. در هر روش انتخاب مشخصه، دو گام اصلی وجود دارد. اولین گام، فرایند جستجو برای معرفی زیرمجموعه‌ای از مشخصه‌ها بوده و گام دوم یک روش ارزیابی، به‌منظور آزمایش یکپارچگی مشخصه‌ها است. یکی از الگوریتم‌هایی که محققان زیادی را به خود جلب کرده است، الگوریتم ژنتیک می‌باشد. تحقیقات نشان داده است که الگوریتم ژنتیک می‌تواند صحت بالاتری نسبت به روش‌های دیگر ارائه دهد [۲۳].

فرایند انتخاب مشخصه با استفاده از الگوریتم ژنتیک در نرم‌افزار رپیدماینرXXV، نسخه ۶، بر روی مشخصه‌های موجود پیاده‌سازی و مشخصه‌های کم‌اهمیت حذف شده است. فرایندی که الگوریتم ژنتیک برای کاهش ابعاد داده دنبال می‌کند بدین صورت است که ابتدا در مرحله آموزش به الگوریتم، یک مجموعه مشخصه تصادفی از مجموعه داده به‌عنوان جمعیت اولیه، توسط الگوریتم ژنتیک انتخاب می‌شود. سپس مدل‌سازی مسأله با استفاده از تکنیک دسته‌بندی مربوط اجرا شده، مقادیر تابع برازندگی که معادل صحت پیش‌بینی رویگردانی است، تعیین شده و جواب‌های اولیه به‌صورت نزولی مرتب می‌شوند. در این میان،

بهترین مجموعه مشخصه که دارای بیشترین صحت پیش‌بینی است، تعیین می‌شود. سپس با ایجاد نسل بعدی الگوریتم ژنتیک با استفاده از عملگرهای تقاطع و جهش، مجموعه مشخصه جدید تعریف می‌شود. این فرایند به‌گونه‌ای است که همواره برای حفظ بهترین جواب، بدترین جواب نسل کنونی با بهترین جواب نسل قبل جایگزین می‌شود. ایجاد نسل بعدی را تا رسیدن الگوریتم به شرط توقف یا همان حداکثر تعداد نسل ادامه داده و مجموعه مشخصه با بهترین صحت پیش‌بینی به‌عنوان مجموعه مشخصه بهینه تعیین می‌شود. در گام آخر، مدل با مجموعه مشخصه بهینه عملیات دسته‌بندی را انجام داده و خطای دسته‌بندی محاسبه می‌شود [۲۳]. در جدول (۲)، فهرست متغیرهای بهینه نهایی وارد شده به مرحله مدل‌سازی آمده است.

جدول ۲. متغیرهای بهینه مدل‌سازی

ردیف	عنوان متغیر	نام متغیر
۱	نوع خودرو	متغیر ورودی
۲	مدل خودرو	متغیر ورودی
۳	تخفیف عدم خسارت	متغیر ورودی
۴	مدت قرارداد	متغیر ورودی
۵	سابقه خرید	متغیر ورودی
۶	تمایل به خرید	متغیر ورودی
۷	نحوه آشنایی با سازمان	متغیر ورودی
۸	وضعیت رویگردانی مشتری	متغیر هدف

۲-۲. مدل‌سازی

از آن جا که هدف نهایی این مقاله، نگاشت مشتریان به دو گروه رویگردان و غیر رویگردان و پیش‌بینی مشخصه‌های تأثیرگذار بر رویگردانی است، رویکرد داده‌کاوی متناسب با این هدف، "دسته‌بندی" خواهد بود. روش ماشین بردار پشتیبان، که از روش‌های مربوط به رویکرد دسته‌بندی است، به عنوان روش اصلی مدل‌سازی، و روش‌های دسته‌بندی درخت تصمیم ID3، درخت تصمیم CHAID، شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته‌بندی‌کننده‌ی بیزی و K نزدیک‌ترین همسایگی نیز، برای مقایسه با نتایج ماشین بردار پشتیبان انتخاب شدند.

۲-۲-۱. ماشین بردار پشتیبان

روش ماشین بردار پشتیبان یکی از ابزارهای قدرتمند و شناخته‌شده در دسته‌بندی داده‌ها است. فرض کنید $X=\{(x_i,y_i)\}_{i=1}^{n}$ بردارهای ویژگی یا الگوهای یادگیری باشند که در آن هر $x_i \in R$ یک بردار ویژگی d بُعدی در فضای ورودی بوده که دارای برچسب y_i می‌باشد و $y_i \in \{-1,+1\}$. یعنی n نمونه یادگیری وجود دارد که هر کدام دارای d ویژگی بوده و هر نمونه متعلق به دسته ۱+ یا ۱- است. هدف از حل این مسأله، یافتن ابرصفحه بهینه جداکننده دو دسته (H*)، با حداکثر حاشیه است. یعنی تا آنجا که ممکن است، ابرصفحه جداکننده دور از نقاط داده‌های هر دو دسته باشد [۲۵] [۲۴].

مسأله ماشین بردار پشتیبان در دو حالت سطح تصمیم‌گیری خطی و غیر خطی حل می‌شود. در حالت سطح تصمیم خطی، بسته به نحوه توزیع داده‌های یادگیری، مسأله‌ی ماشین بردار پشتیبان به دو شکل حاشیه سخت xxvi و حاشیه نرم xxvi i تقسیم‌بندی می‌شود. در حالت حاشیه سخت، داده‌های یادگیری به صورت خطی جداپذیر بوده و در حالت حاشیه‌ی نرم، به صورت خطی جداناپذیر هستند. حال اگر نحوه قرارگیری داده‌ها به‌گونه‌ای باشد که به هیچ وجه امکان دسته‌بندی خطی وجود نداشته باشد، از حالت غیر خطی استفاده می‌شود [۲۴]. در این مطالعه، از مدل ماشین بردار پشتیبان در حالت حاشیه نرم خطی و غیر خطی استفاده می‌شود. بنابراین، به بررسی این دو حالت خواهیم پرداخت.

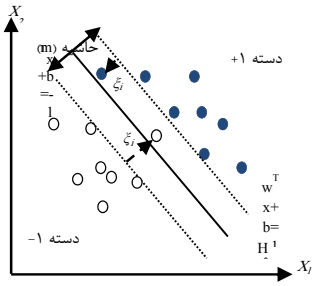
در مدل ماشین بردار پشتیبان با حاشیه نرم، داده‌ها جداپذیر خطی نیستند. یعنی تعداد کمی از نمونه‌ها در داخل دسته دیگر قرار دارند. ولی می‌توان با تحمل مقداری خطا آن‌ها را به صورت خطی جدا کرد. در واقع در این حالت به جداکننده اجازه داده می‌شود که دارای خطا باشد. برای این منظور، همان‌طور که در شکل (۵) دیده می‌شود به هر کدام از نمونه‌ها یک جریمه ξi نسبت داده می‌شود. هر چقدر مجموع مقادیر متغیرهای ξi بیشتر شود، مدل از حالت بهینه دورتر و خطا بیشتر می‌شود. بدیهی است که اگر $\xi_i = 0$ باشد، یعنی داده xi درست دسته‌بندی شده و در دسته خودش قرار دارد. در واقع با در نظر گرفتن خطا برای دو ابرصفحه مرزی، حاشیه می‌تواند برای برخی نمونه کمی جابه‌جا شود [۲۶].

معادله هر ابرصفحه جداکننده به صورت (۲) نمایش داده می‌شود. در این رابطه، w بردار وزن و b مقدار ثابت است. در صورت وجود ابرصفحه بهینه، مقادیر w و b بهینه در این رابطه جایگذاری می‌شود. شکل (۵)، دسته‌بندی را در دو بعد و با داشتن دو مشخصه X1 و X2 نشان می‌دهد [۲۴].

$$f(X,w,b)=w^TX+b=w_1x_1+w_2x_2+....+w_nx_n+b=0 \qquad (۲)$$

پس در نتیجه مطابق شکل (۵)، برای تمام نمونه‌های دسته ۱+ و ۱- روابط زیر صدق می‌کند [۲۷] [۲۵] [۲۴].

$$\left.\begin{array}{l} w^T x_i + b \geq 1-\xi_i, \quad y_i = +1 \\ w^T x_i + b \leq -1+\xi_i, \quad y_i = -1 \end{array}\right\} \Rightarrow y_i(w^T X + b) \geq 1-\xi_i, \quad w \in R^n, b \in R, \xi_i \geq 0 \qquad (۳)$$



شکل ۵. ماشین بردار پشتیبان در حالت حاشیه نرم [۲۵]

ثابت می‌شود که فاصله بین این دو نامعادله و یا به‌عبارتی، مقدار حاشیه m برابر $2/\|w\|$ است. پس برای به‌دست آوردن ابرصفحه

(۵)

در این رابطه، که λ_i و β_i ضرایب لاگرانژ هستند. با توجه به شرایط کان تاکر، مشتق L نسبت به w، ξ و b برابر صفر است. با مساوی قرار دادنِ مشتق برابر با صفر و جایگذاری مقادیر در (۵)، مسأله بهینه‌سازی به صورت (۶) خواهد شد [۲۷] [۲۴].

$$\begin{array}{l} Maximize \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j \\ S.t. \\ \sum_{i=1}^{n} \lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C \qquad i = 1,\ldots,n \end{array} \qquad (۶)$$

بنابراین، تابع تصمیم برای دسته‌بندی یک الگوی ورودی x_t به فرم (۷) است [۲۷] [۲۴].

$$D(x) = sign(w^T x_t + b) = sign\left(\sum_{i=1}^{n} \lambda_i y_i x_i^T x_t + b\right) \qquad (۷)$$

اگر نمونه‌ها جداپذیر خطی نباشند آنگاه با استفاده از (۷)، نمی‌توان این نمونه‌ها را به صورت صحیح به دو دسته تقسیم کرد. برای این منظور نمونه‌ها از فضای ورودی به فضای ویژگی XXVi i i که دارای ابعاد بیشتر است، با استفاده از توابع کرنل رده شده و در فضای جدید می‌توان نمونه‌ها را با یک ابرصفحه جداکننده خطی جدا کرد. با توجه به این‌که عملیات خطی در فضای ویژگی معادل عملیات غیر خطی در فضای ورودی است، پس عمل تبدیل باعث می‌شود که نوع عملیات راحت‌تر شود. از طرف دیگر عمل تبدیل باعث می‌شود تا عمل دسته‌بندی در فضای ویژگی به‌راحتی انجام

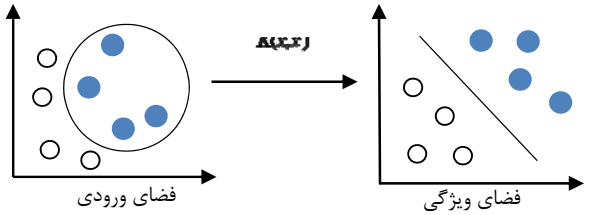
جداساز بهینه باید مسأله بهینه‌سازی زیر که به آن ماشین بردار پشتیبان با حاشیه نرم می‌گویند، حل شود [۲۶].

$$\begin{array}{l} Minimize \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ S.t. \\ y_i(w^T X + b) \geq 1 - \xi_i \\ \xi \geq 0 \qquad i = 1,\ldots,n \end{array} \qquad (۴)$$

رابطه (۴) یک مسأله برنامه‌ریزی درجه دوم از نوع محدب است، که برای حل آن، تابع لاگرانژ زیر تشکیل می‌شود و ضرایب لاگرانژ λ_i به‌دست می‌آید. تابع لاگرانژ در این حالت به صورت (۵) است [۲۷].

$$L(w,b,\xi,\lambda,\beta) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} C_i\xi_i + \sum_{i=1}^{n} \lambda_i(1-\xi_i - y_i((w^T X + b)) - \sum_{i=1}^{n} \beta_i\xi_i$$

شود درحالی‌که این عمل در فضای ورودی به‌سختی انجام می‌شود [۲۳]. شکل (۶) نمونه‌ای از تبدیل فضا را نشان می‌دهد.



شکل ۶. تبدیل فضای ورودی به فضای ویژگی [۲۰]

باید به این نکته توجه داشت که فضای ویژگی در عمل ابعاد بزرگتری نسبت به فضای ورودی دارد. با استفاده از تابع کرنل $K(x_i, x_j) = \phi(x_i)\phi(x_j)$، مسأله دوگان (۶)، در فضای ویژگی به صورت (۸) خواهد بود [۲۶].

$$\begin{array}{l} Maximize \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(x_i^T x_j) \\ S.t. \\ \sum_{i=1}^{n} \lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C \qquad i = 1,\ldots,n \end{array} \qquad (۸)$$

بنابراین، تابع تصمیم برای دسته‌بندی یک الگوی ورودی x_t به صورت (۹) است [۲۷] [۲۰].

مدل‌سازی ماشین بردار پشتیبان در دو حالت خطی و غیر خطی در نرم‌افزار متلبXXİ X، نسخه ۲۰۱۰ انجام شده است. برای مدل‌سازی ماشین بردار پشتیبان، سه نوع تابع کرنل خطیXXX،

چند جمله‌ای XXXi و گوسی XXXi i را که در این مقاله از آن‌ها استفاده شده‌اند، در جدول (۳) بررسی خواهیم کرد.

جدول ۳. انواع توابع کرنل [۲۵]

تابع کرنل	فرمول ریاضی
خطی	$K(x,x')=x^Tx'$
چند جمله‌ای	$K(x,x')=(x^Tx'+1)^d$
گوسی	$k(x,x')=exp(-\frac{1}{2\sigma^2}\|x-x'\|^2)$

۲-۲-۲. درخت تصمیم

درخت تصمیم از معروف‌ترین تکنیک‌های دسته‌بندی است. هر درخت تصمیم از تعدادی گره و یال تشکیل شده است. در ساخته شدن درخت و تشکیل هر گره، الگوریتم درخت تصمیم به دنبال انتخاب بهترین مشخصه برای شکستن درخت به دو یا چند زیر درخت است. درخت تصمیم، مشاهدات واردشده به مدل را با مرتب کردن آن‌ها در درخت از گره ریشه، گره‌ای که در بالای درخت قرار دارد، به سمت گره‌های برگ، گره‌های انتهایی درخت که فقط از یک سو به سایر گره‌ها متصل هستند، دسته‌بندی می‌کند. هر گره داخلی (غیر برگ) از درخت، متناظر با یک مشخصه از مشاهدات بوده و هر یالی که از آن خارج می‌شود متناظر با یک مقدار برای آن مشخصه است. در نهایت هر گره برگ، در یک دسته متغیر هدف، دسته‌بندی می‌شود [۱۹]. یکی از الگوریتم‌های درخت تصمیم که به‌منظور مقایسه با تکنیک ماشین بردار پشتیبان در نظر گرفته شده است، الگوریتم ID3XXXi i i می‌باشد. این الگوریتم، از الگوریتم‌های ساده درخت‌های تصمیم‌گیری بوده که توسط کوییلنV XXXi ارائه شده است. ایده‌ی اساسی این روش، فرایند جستجوی حریصانه بالا به پائین در مجموعه داده، به‌منظور ارزیابی هر مشخصه در هر گره بوده و در آن انتخاب‌های قبلی هرگز مورد بازبینی قرار نمی‌گیرند [۲۸]. از دیگر الگوریتم‌های معروف درخت تصمیم، الگوریتم CHAID XXXV می‌باشد. تنها تفاوت این الگوریتم با درخت تصمیم ساده این است که، این روش برای ساخت درخت، سطح اهمیتی از آزمون مربع‌کای XXXVi را مشخص می‌کند تا رشد درخت را متوقف کند [۲۹]

۳-۲-۲. شبکه‌های عصبی

مفهوم شبکه‌های عصبی از مغز انسان گرفته شده است و برای مسائل متعددی مانند دسته‌بندی و پیش‌بینی استفاده شده است. شبکه‌های عصبی استفاده وسیعی در شناسایی الگوها دارند؛ زیرا قابلیت آن را دارند که به‌طور عمومی به ورودی‌های غیر منتظره نیز پاسخ دهند. در طول ساخت، نرون‌ها می‌آموزند که چگونه الگوهای ویژه گوناگون را تشخیص دهند. اگر الگویی پذیرفته شود در حالی که در طول اجرا ورودی با خروجی مرتبط نباشد، نرون از مجموعه‌ای از الگوهایی که قبلا آموخته، آن خروجی را که بیش‌ترین شباهت را به الگو داشته و کم‌ترین تفاوت را با ورودی دارد، انتخاب می‌کند. این فرایند از سه لایه تشکیل شده است: لایه ورودی، لایه میانی، لایه خارجی. لایه ورودی، اطلاعات را از یک منبع خارجی دریافت می‌کند و لایه خارجی نیز اطلاعات را به سیگنال‌هایی برای استفاده منبع خارجی تبدیل می‌کند. لایه میانی نیز پلی بین لایه‌های ورودی و خروجی است و در حقیقت، فرایند پردازش داده را انجام می‌دهد [۳۱] [۳۰].

۴-۲-۲. رگرسیون لجستیک

رگرسیون لجستیک یکی از تکنیک‌های کاربردی برای تحلیل داده‌های دسته‌بندی شده است. زمانی‌که متغیر هدف، متغیری کیفی با دو سطح باشد، دیگر مدل‌های رگرسیون معمولی قابل استفاده نیستند. در این‌گونه موارد، از رگرسیون لجستیک استفاده می‌شود. هدف رگرسیون لجستیک، تعیین احتمال شرطی مربوط به مشاهده‌های مشخص یک دسته با توجه به متغیرهای مستقل است. به عبارت ساده‌تر، با گرفتن متغیر ورودی، مقدار متغیر وابسته به آن را پیش‌بینی می‌کند [۳۲].

۵-۲-۲. جنگل تصادفی

درخت‌های تصمیم به‌دلیل سهولت کاربرد و تفسیر آن‌ها در زمینه‌ی دسته‌بندی‌های دوسطحی بسیار معروف شده‌اند. کاربرد این درخت‌ها، امکان استفاده از متغیرهای پیش‌بینی‌کننده با مقیاس‌های متفاوت را فراهم می‌کند. یکی از مشکلات کاربرد درخت تصمیم، عدم ثبات و پایداری آن‌ها و ایجاد راه‌حل‌های بهینه محلی است. تکنیک جنگل تصادفی از دیگر تکنیک‌های دارای رویکرد دسته‌بندی می‌باشد که برای رفع مشکلات موجود در تکنیک درخت تصمیم ارائه شده است. در این تکنیک، مجموعه‌ای از درخت‌های تصمیم ایجاد شده و هر درخت به مشهورترین دسته رأی می‌دهد. با ادغام رأی درخت‌های مختلف، برای هر نمونه یک دسته پیش‌بینی می‌شود. در این روش که برای افزایش صحت درخت تصمیم طراحی شده است، تعداد بیشتری درخت تولید می‌شود تا برای پیش‌بینی دسته با هم اقدام به رأی‌گیری کنند. این روش یک دسته‌بندی‌کننده جمعی XXXVi i است که از تعدادی درخت تصمیم تشکیل شده است و نتیجه نهایی، میانگین نتیجه تک‌تک درخت‌ها است [۳۳].

۶-۲-۲. دسته‌بندی‌کننده‌ی بیزی

دسته‌بندی‌کننده بیزی، مدل مقدماتی از مدل احتمال بیزی است. این مدل بر پایه احتمال وقوع یا عدم وقوع پدیده‌ها است. عملکرد آن، بر فرضیات استقلال قوی استوار بوده و احتمال رخداد یک صفت روی احتمال سایر صفت‌ها بی‌تأثیر است. به این معنی که در این مدل، احتمال رخداد نتیجه نهایی، بر اساس احتمالات رخداد متغیرهای مستقل به‌شرط رخداد همان نتیجه به‌دست آمده و احتمال رخداد هر یک از متغیرهای مستقل به‌شرط رخداد یک

نتیجه نهایی خاص، مستقل از احتمال رخداد سایر متغیرهای مستقل به‌شرط رخداد همان نتیجه است. به این ترتیب، یک مسأله چندمتغیره p بعدی به تخمین p مسأله یک‌متغیره کاهش پیدا می‌کند. این امر، باعث کاهش پیچیدگی‌های محاسبات می‌شود [۳۴].

۷-۲-۲. K نزدیک‌ترین همسایگی

هدف از تکنیک K نزدیک‌ترین همسایگی، دسته‌بندی یک عضو جدید براساس ویژگی نمونه‌های آموزش‌دهنده است. در این تکنیک نمونه‌ی جدید براساس اکثریت K دسته که نزدیک‌ترین همسایگی‌ها را با آن نمونه داشته باشند، تقسیم‌بندی می‌شود. به‌طور کلی می‌توان بیان کرد که روش K نزدیک‌ترین همسایگی، یک روش تشخیص الگوهای غیرپارامتری می‌باشد که تعداد K تا از نزدیک‌ترین الگوهای مشابه را پیدا کرده و براساس آن‌ها، ارزش نمونه مورد مطالعه را پیش‌بینی می‌کند. این الگوریتم براساس حداقل فاصله نمونه مورد بررسی تا نمونه‌های موجود دیگر برای تعیین K نزدیک‌ترین همسایگی‌ها کار می‌کند و نمونه را متعلق به دسته‌ای می‌داند که بیشترین آرا را در بین K نزدیک‌ترین همسایه داشته باشد [۳۵].

۳. انتخاب پارامترهای بهینه مدل

عملکرد ماشین بردار پشتیبان به چندین پارامتر مستقل وابسته می‌باشد. مانند پارامتر C در مدل خطی، که به منظور کنترل رابطه‌ی بین حداکثر حاشیه و حداقل خطا است، پارامترهای C و σ در کرنل گوسی و پارامترهای C و d در کرنل چندجمله‌ای. روش ماشین بردار پشتیبان وابستگی شدیدی به مقادیر این پارامترها داشته و انتخاب مقدار هر پارامتر، عملکرد حیاتی در این روش دارد. انتخاب نامناسب هر یک از پارامترها، می‌تواند منجر به رسیدن جواب‌های نامطلوبی برای مدل شود. به همین دلیل، برای رسیدن به عملکرد خوب در روش ماشین بردار پشتیبان، بهینه‌سازی پارامترها بسیار مهم است. در عمل، اگر اندازه مجموعه داده‌ها بزرگ باشد، فرایند پیدا کردن پارامتر خوب بسیار وقت‌گیر خواهد بود. معمولا مدل‌های پیش‌بینی رویگردانی باید حجم انبوهی از مجموعه داده‌ها را بررسی کنند. بنابراین، پیدا کردن یک روش کارا برای تنظیم پارامترها، به‌منظور رسیدن به نتایج بهتر، ضروری است. یکی از روش‌های ممکن، جستجوی شبکه است. این روش، یک روش عددی است و عملکرد هر مقدار را در فضای پارامترهای از پیش‌تعیین‌شده، برای مدل ارزیابی کرده و در نهایت نقطه‌ای با بهترین عملکرد انتخاب خواهد شد. به عنوان مثال، با توجه به وجود دو پارامتر در کرنل گوسی، روش جستجوی شبکه برای بهینه‌سازی پارامترها، تمامی ترکیب‌های ممکن بین مقدارهای این دو پارامتر را در نظر گرفته، عملکرد روش را برای هر ترکیب محاسبه کرده و در گام آخر، آن ترکیبی از پارامترها را که منجر به بالاترین عملکرد پیش‌بینی روش می‌شود، به‌عنوان پارامترهای بهینه انتخاب خواهد کرد [۲۰]. بدین منظور، برای انتخاب پارامترهای بهینه مدل ماشین بردار پشتیبان در سه حالت کرنل خطی، چند جمله‌ای و گوسی، از ترکیب دو روش جستجوی شبکه و اعتبارسنجی متقابل k لایه XXXvi i i، با استفاده از نرم‌افزار متلب، نسخه ۲۰۱۰، استفاده خواهد شد. در این مطالعه، k برابر با ۱۰ در نظر گرفته شده است.

برای استفاده از روش جستجوی شبکه در حالت خطی، باید مقدار پارامتر ثابت C، که تنها پارامتر مربوط به حالت خطی است را بهینه نمود. بنابراین، ابتدا بازه (۲۱۰،...،۲۱،۲۲،۲۳) برای مقادیر پارامتر C انتخاب می‌شود. برای هر مقدار C، فرایند اعتبارسنجی متقابل ۱۰ لایه اجرا شده و با استفاده از معیار دقت، عملکرد دسته‌بندی هر حالت به دست خواهد آمد. سپس بین ۱۰ مقدار صحت به دست آمده، میانگین گرفته خواهد شد. این رویکرد برای تمامی مقادیر در نظر گرفته شده برای پارامتر C تکرار خواهد شد. در آخر برای تعیین مقدار C بهینه، میان تمامی دقت‌های به دست آمده ناشی از C های متفاوت، بیشینه گرفته خواهد شد و مقدار C مربوط به بیشترین دقت، به عنوان C بهینه وارد مدل خواهد شد. الگوریتم ماشین بردار پشتیبان با کرنل خطی و مقدار بهینه پارامتر C اجرا شده و عملکرد پیش‌بینی نهایی مدل به دست خواهد آمد. در این حالت، مقدار ۴ به عنوان مقدار بهینه پارامتر C در کرنل خطی به دست آمده است.

سپس برای مقایسه عملکرد دو حالت خطی و غیر خطی، کدنویسی با استفاده از دو کرنل غیر خطی چند جمله‌ای و گوسی نیز انجام شده است. . همان روال حالت خطی، برای دو حالت غیر خطی کرنل چند جمله‌ای و گوسی نیز انجام شده است؛ با این تفاوت که از آن جایی که نوع کرنل‌ها تفاوت دارد، پارامترهای مربوط به هر کدام نیز فرق می‌کند. به همین دلیل، ابتدا به بررسی نتایج حاصل از کرنل غیر خطی چند جمله‌ای پرداخته و سپس نتایج حاصل از کرنل غیر خطی گوسی را مورد بررسی قرار می‌دهیم. در حالت کرنل چند جمله‌ای علاوه بر پارامتر C، پارامتر d نیز، که نشان‌دهنده درجه تابع چند جمله‌ای است، وجود دارد. بنابراین، در حالت کرنل چند جمله‌ای باید به بهینه‌سازی مقدار این دو پارامتر پرداخت. در تابع کرنل چند جمله‌ای، بازه (۲۶،...،۲۱،۲۲) برای مقدار پارامتر C و بازه‌ی (۷،...،۲،۳) برای مقدار پارامتر d، در نظر گرفته شده است. بدین صورت که مدل تمامی ترکیب‌های ممکن را برای دو پارامتر C و d در نظر گرفته و برای هر ترکیب، فرایند اعتبارسنجی متقابل ۱۰ لایه را اجرا کرده و آن ترکیبی را که منجر به بهترین عملکرد پیش‌بینی خواهد شد را انتخاب می‌کند. در این حالت، مقدارهای ۲ و ۲ به ترتیب برای پارامترهای C و d در کرنل چند جمله‌ای به دست آمده است.

کرنل گوسی نیز پارامترهای مخصوص به خود را دارد که باید برای استفاده از روش جستجوی شبکه، مدل را با توجه به پارامترهای آن تغییر داد. کرنل گوسی دارای پارامتر ثابت C و پارامتر σ است. بنابراین در این حالت، به بهینه‌سازی پارامترهای C و σ موجود در مدل ماشین بردار پشتیبان با کرنل غیر خطی گوسی می‌پردازیم. به‌منظور استفاده از روش جستجوی شبکه، بازه (۲۶،...،۲۲،۲۱) به پارامتر C و بازه (۲۴،...،۲-۲،۳-۲) به مقدار پارامتر σ اختصاص داده شده است. روش جستجوی شبکه تمامی ترکیب‌های ممکن بین دو پارامتر C و σ را در نظر گرفته و اعتبارسنجی متقابل ۱۰ لایه را روی هر ترکیب ممکن اجرا می‌کند. در نهایت مقادیر بهینه پارامترهای C و σ که نیل به بالاترین صحت پیش‌بینی مدل می‌شود، به دست خواهد آمد. این مقادیر در تابع کرنل گوسی به ترتیب برای پارامترهای C و σ، ۲ و ۱۶ به دست آمده است.

۴. ارزیابی و اعتبارسنجی مدل

بعد از ساخت مدل، برای پیش‌بینی رفتار آینده مشتریان از آن استفاده می‌شود. بنابراین، ارزیابی و اعتبارسنجی مدل، فرایند بسیار مهمی محسوب می‌شود [۱۹]. اعتبارسنجی متقابل، روشی آماری برای ارزیابی و مقایسه الگوریتم‌های یادگیری است که داده‌ها را به دو بخش متمایز تقسیم می‌کند: یک بخش برای یادگیری یا آموزش مدل و دیگری برای ارزیابی مدل استفاده می‌شود. Hold-Out Validation، از روش‌های اعتبارسنجی مدل است که از هم‌پوشانی بین داده‌های آموزش و ارزیابی جلوگیری می‌کند. در این روش به طور معمول داده‌های موجود را با نسبت ۷۰ به ۳۰، به دو بخش غیرهم‌پوشان تقسیم می‌کنند: ٪۷۰ داده‌ها برای آموزش و ٪۳۰ دیگر برای ارزیابی مدل استفاده خواهند شد. داده‌های ارزیابی کنار گذاشته شده و در طول آموزش بررسی نخواهند شد. این روش بدین صورت است که همه داده‌های موجود را استفاده نمی‌کند و نتایج تا حد زیادی به نمونه‌های آموزش و ارزیابی انتخاب‌شده وابسته است. همچنین ممکن است که داده‌های موجود در مجموعه ارزیابی برای آموزش ارزشمند باشند و اگر آن‌ها را در عملکرد پیش‌بینی قرار دهیم به انحراف نتایج منجر شود. این مشکل می‌تواند تا حدی با چندین بار تکرار Hold-Out Validation و متوسط‌گیری نتایج برطرف شود. در غیر این صورت، این تکرار به روش نظام‌مندی اجرا می‌شود. برخی از داده‌ها ممکن است چندین بار در مجموعه ارزیابی قرار گیرند درحالی‌که داده‌های دیگر، اصلا در این بخش وارد نشوند. یا برعکس برخی داده‌ها ممکن است همیشه در مجموعه ارزیابی قرار گیرند و هیچ وقت فرصت برای پیوستن به مرحله آموزش نداشته باشند. به‌منظور مقابله با این چالش‌ها و استفاده از داده‌های موجود، به طور معمول از روش K-fold Cross-validation استفاده می‌شود. اعتبارسنجی متقابل k لایه، یکی از روش‌های محبوب ارزیابی مدل است. در این روش، داده‌ها به صورت تصادفی به k زیر مجموعه مجزا تقسیم شده و k بار آموزش و ارزیابی انجام می‌شوند؛ به این صورت که هر بار یکی از زیر مجموعه‌ها برای ارزیابی مدل نگه‌داشته شده و k-۱ زیر مجموعه دیگر، برای آموزش مدل استفاده می‌شود. این فرایند k بار تکرار شده؛ به طوری که هر یک از زیرمجموعه‌ها دقیقاً یک بار برای ارزیابی مدل به کار برده می‌شوند. در نهایت، نتیجه این k تکرار برای دست‌یابی به یک برآورد نهایی، میانگین‌گیری می‌شود. بدین صورت، همه‌ی داده‌ها در هر دو گروه آموزش و ارزیابی قرار گرفته و از این منظر، روش ارزیابی دقیق‌تری محسوب شدهق است. به طور کلی، فرایند اعتبارسنجی متقابل ۱۰ لایه برای برآورد صحت پیشنهاد می‌شود [۲۲].

بنابراین، در این مقاله ابتدا با استفاده از روش Hold-Out Validation، به طور تصادفی داده‌های موجود با نسبت ۷۰ به ۳۰، به دو بخش غیرهم‌پوشان تقسیم شدند. به‌طوری‌که ٪۷۰ داده‌ها که حدودا ۱۰۶۰ مشتری است، برای آموزش و ٪۳۰ دیگر که معادل ۴۶۰ مشتری است، برای ارزیابی و تأیید مدل مدل در نظر گرفته شده است. سپس، با استفاده از روش اعتبارسنجی متقابل ۱۰ لایه، داده‌های آموزش به صورت تصادفی به ۱۰ زیرمجموعه مجزا تقسیم شدند. به این صورت که یکی از زیرمجموعه‌ها برای ارزیابی مدل و ۹ زیر مجموعه دیگر، برای آموزش مدل در نظر گرفته شده است. سپس، مدل با استفاده از ۹ زیر مجموعه در نظر گرفته شده آموزش دیده و از یک زیر مجموعه باقی‌مانده دیگر برای پیش‌بینی رفتار مدل و ارزیابی آن استفاده می‌شود. این فرایند با استفاده از معیار دقت، که نسبت پیش‌بینی‌های درست به تعداد کل پیش‌بینی‌ها است، عملکرد پیش‌بینی مدل را تعیین می‌کند. این روند ۱۰ مرتبه تکرار می‌شود؛ به طوری که هر یک از زیر مجموعه‌ها دقیقاً یک بار برای ارزیابی مدل انتخاب و با استفاده از معیار دقت، عملکرد پیش‌بینی مدل را ارزیابی می‌کند. بعد از آن، میانگین نتیجه این ۱۰ تکرار محاسبه می‌شود. مقدار میانگین‌گیری‌شده، عملکرد پیش‌بینی نهایی مدل را بر اساس روش اعتبارسنجی متقابل ۱۰ لایه نشان می‌دهد. این فرایند برای هر سه نوع تابع کرنل خطی، چند جمله‌ای و گوسی در نرم‌افزار متلب اجرا و نتایج اعتبارسنجی و ارزیابی روش ماشین بردار پشتیبان با استفاده از فرایند اعتبارسنجی متقابل ۱۰ لایه و معیار صحت در جدول (۳) ارائه شده است.

در انتها برای بررسی صحت نتایج روش ارائه‌شده، عملکرد پیش‌بینی حاصل از آن با عملکرد پیش‌بینی روش‌های درخت تصمیم ID3، درخت تصمیم CHAID، شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته‌بندی‌کننده‌ی بیزی و K نزدیک‌ترین همسایگی مقایسه شده است. برای پیاده‌سازی این تکنیک‌ها، از نرم‌افزار رپیدماینر، که از نرم‌افزارهای قدرتمند حوزه داده‌کاوی است، استفاده شده است. نتایج حاصل از مقایسه عملکرد

پیش‌بینی روش‌های نام‌برده در جدول (۳)، ارائه شده است. در پایان این بخش مدل‌سازی مسأله به روش ماشین بردار پشتیبان، بهینه‌سازی پارامترها، مقایسه ماشین بردار پشتیبان با سایر تکنیک‌های دسته‌بندی و اعتبارسنجی و ارزیابی مدل صورت گرفته است.

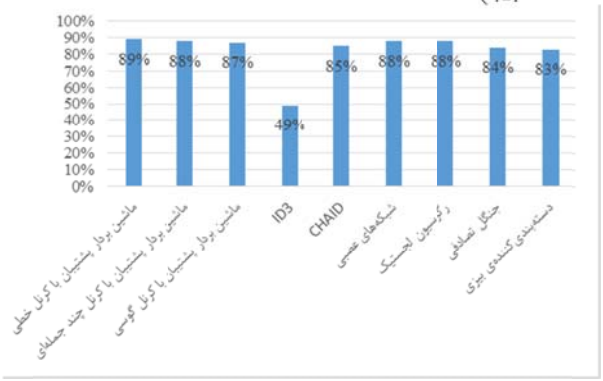
۵. نتایج محاسباتی و تحلیل یافته‌ها

مدل‌سازی، بهینه‌سازی پارامترها و اعتبارسنجی مسأله ماشین بردار پشتیبان در دو حالت خطی و غیر خطی، با در نظر گرفتن سه تابع کرنل خطی، چند جمله‌ای و گوسی، در نرم‌افزار متلب ۲۰۱۰، بر روی مجموعه داده موجود انجام شده است. جدول (۴) عملکرد پیش‌بینی روش ماشین بردار پشتیبان را بر اساس اعتبارسنجی متقابل ۱۰ لایه و معیار صحت نشان می‌دهد. همچنین برای بررسی کارایی مدل ماشین بردار پشتیبان، این روش با روش‌های ID3، CHAID، شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته‌بندی‌کننده‌ی بیزی و K نزدیک‌ترین همسایگی مقایسه شده است. از آن جایی که روش نمونه‌گیری به کار برده شده برای تکنیک‌های دسته‌بندی پیش رو نیز مانند روش نمونه‌گیری روش ماشین بردار پشتیبان، تصادفی است، امکان ایجاد نمونه‌گیری متفاوت با نمونه‌های آموزش و ارزیابی روش ماشین بردار پشتیبان وجود دارد. به همین دلیل، برای کاهش خطای نمونه‌گیری تصادفی و افزایش صحت مقایسه نتایج تکنیک‌های مختلف با هم، از همان نمونه‌های آموزش و ارزیابی که به‌منظور پیش‌بینی رویگردانی با روش ماشین بردار پشتیبان در نرم‌افزار متلب ایجاد شده است، استفاده خواهد شد. برای رسیدن به عملکرد بهتر این روش‌ها، پارامترهای مربوط به کلیه روش‌های نام‌برده نیز، با استفاده از روش جستجوی شبکه در نرم‌افزار رپیدماینر بهینه‌سازی شده است. بهینه‌سازی پارامترها در نرم‌افزار، از روش جستجوی شبکه پیروی می‌کند. بدین صورت که ابتدا با در نظر گرفتن بازه برای هر پارامتر، روش جستجوی شبکه تمام ترکیب‌های ممکن قرارگیری پارامترها با هم را در نظر می‌گیرد. سپس آن ترکیبی از پارامترها را که منجر به بالاترین صحت پیش‌بینی رویگردانی مشتری خواهد شد، به عنوان پارامترهای بهینه هر تکنیک معرفی می‌کند. جدول (۴) مقایسه نتایج حاصل از مدل‌سازی مسأله به روش ماشین بردار پشتیبان و سایر تکنیک‌های دسته‌بندی را ارائه می‌دهد. همچنین مقایسه صحت پیش‌بینی تکنیک‌های مورد استفاده به طور مصور در شکل (۷) آمده است.

$$D(x) = sign\left(w^T \varphi(x_t) + b\right) = sign\left(\sum_{i=1}^{n} \lambda_i y_i \varphi(x_i)^T \varphi(x_t) + b\right) = sign\left(\sum_{i=1}^{n} \lambda_i y_i K\left(x_i^T, x_t\right) + b\right) \quad (۹)$$

جدول ۴. مقایسه صحت تکنیک‌های به کار رفته در مدل‌سازی

تکنیک دسته‌بندی	صحت پیش‌بینی
ماشین بردار پشتیبان با کرنل خطی	۸۹٪
ماشین بردار پشتیبان با کرنل چند جمله‌ای	۸۸٪
ماشین بردار پشتیبان با کرنل گوسی	۸۷٪
ID3	۴۹٪
CHAID	۸۵٪
شبکه‌های عصبی	۸۸٪
رگرسیون لجستیک	۸۸٪
جنگل تصادفی	۸۴٪
دسته‌بندی‌کننده‌ی بیزی	۸۳٪



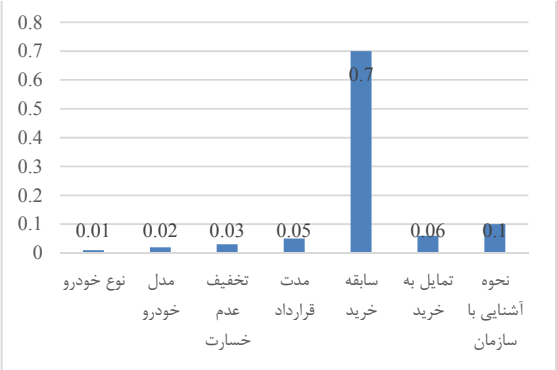
شکل ۷. مقایسه تکنیک‌های به کار رفته در مدل‌سازی

همان‌طور که از جدول (۴) و شکل (۷) مشخص است، تمامی تکنیک‌ها به جز تکنیک درخت تصمیم ID3 از صحت نسبتا بالایی در پیش‌بینی رویگردانی مشتریان برخوردار هستند. این بدین معنا است که تکنیک‌های دسته‌بندی مورد استفاده قابلیت خوبی در پیش‌بینی رویگردانی دارند و صحت‌های پیش‌بینی بالا به دست آمده در اکثریت تکنیک‌ها گواهی بر تایید توانایی یکدیگر در پیش‌بینی رویگردانی هستند. در این بین، تکنیک ماشین بردار پشتیبان با کرنل خطی، بالاترین صحت پیش‌بینی رویگردانی مشتریان را در بین هفت تکنیک دسته‌بندی دیگر دارد. بعد از آن

ماشین بردار پشتیبان با کرنل چند جمله‌ای، شبکه‌های عصبی و رگرسیون لجستیک نیز دارای صحت خوب و نزدیک به ماشین بردار پشتیبان با کرنل خطی هستند. با توجه به نتایج به دست آمده بیان این نکته صحت بالاتر تکنیک ماشین بردار پشتیبان با کرنل خطی، قدرت بالای این تکنیک را در پیش‌بینی رویگردانی مشتریان در صنعت بیمه نشان می‌دهد و دو فرایند انتخاب مشخصه و انتخاب پارامترهای بهینه نقش مهمی را ایفا کرده‌اند. همچنین استفاده از دو روش Hold Out Validation و 10-fold cross validation نیل به صحت پیش‌بینی دقیق‌تری شده است. زیرا نتایج روش Hold Out Validation، وابستگی زیادی به انتخاب نمونه‌های آموزش و ارزیابی داشته و از آن جایی که روش نمونه‌گیری نمونه‌های آموزش و ارزیابی به صورت تصادفی است، بازه‌ای وسیع‌تری از دقت‌های پیش‌بینی مختلف به دست خواهد آمد. برای رفع این مشکل، در این پژوهش استفاده از روش 10-fold cross validation به‌همراه روش Hold Out Validation منجر به جواب‌های دقیق‌تر شده است. این روش به‌دلیل تکرار ۱۰ مرتبه‌ای فرایند و میانگین‌گیری از دقت‌های پیش‌بینی به دست آمده، به صحت واقعی‌تر و مستقل از نمونه‌های آموزش و ارزیابی انتخاب‌شده، رسیده است. بنابراین، به‌دلیل عملکرد بالاتر روش ماشین بردار پشتیبان با کرنل خطی در پیش‌بینی رویگردانی، روش قابل استنادتری نسبت به سایر روش‌ها بوده و از این جهت، روش پیش‌بینی مبنا قرار می‌گیرد. بنابراین، برای تعیین عوامل اثرگذار در پیش‌بینی رویگردانی مشتریان، به روش ماشین بردار پشتیبان با کرنل خطی مراجعه کرده و دلایل رویگردانی مشتریان را از این روش استخراج خواهیم کرد. علاوه براین، با توجه به این‌که ماشین بردار پشتیبان با کرنل چند جمله‌ای، شبکه‌های عصبی و رگرسیون لجستیک نیز صحت بالایی در پیش‌بینی رویگردانی داشته‌اند، به تعیین مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی مشتری با استفاده از این تکنیک‌ها نیز پرداخته و نتایج آن را با نتایج روش ماشین بردار پشتیبان با کرنل خطی مقایسه خواهیم کرد. جدول (۵) میزان تأثیر هر یک از مشخصه‌ها را در متغیر هدف پیش‌بینی رویگردانی، با استفاده از مدل‌سازی به روش ماشین بردار پشتیبان با کرنل خطی در نرم‌افزار متلب و شکل (۸) نیز اهمیت مشخصه‌ها را به‌صورت مصور نشان می‌دهد.

جدول ۴. اهمیت مشخصه‌ها در پیش‌بینی رویگردانی مشتریان

عنوان مشخصه	درجه اهمیت
نوع خودرو	۰.۰۱
مدل خودرو	۰.۰۲
تخفیف عدم خسارت	۰.۰۳
مدت قرارداد	۰.۰۵
سابقه خرید	۰.۷
تمایل به خرید	۰.۰۶
نحوه آشنایی با سازمان	۰.۱



شکل ۸. اهمیت مشخصه‌ها در پیش‌بینی رویگردانی مشتریان

همان‌گونه که از جدول (۵) و شکل (۸) بر می‌آید، مشخصه‌های سابقه خرید و نحوه آشنایی با سازمان، به‌ترتیب بیشترین درجه اهمیت را بر متغیر هدف داشته و در اولویت بعدی مشخصه تمایل به خرید، به عنوان مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی مشتریان معرفی خواهند شد. سایر مشخصه‌ها دارای درجه اهمیت ناچیزی هستند و به همین دلیل در پیش‌بینی رویگردانی مشتری در نظر گرفته نخواهند شد. از سوی دیگر، نتایج ماشین بردار پشتیبان با کرنل چند جمله‌ای، شبکه‌های عصبی و رگرسیون لجستیک به‌ترتیب مشخصه‌های سابقه خرید، نحوه آشنایی با سازمان و تمایل به خرید را به عنوان مشخصه‌های اصلی پیش‌بینی‌کننده‌ی رویگردانی مشتری برگزیده است. این بدین معنا است که نتایج این سه روش با نتایج روش ماشین بردار پشتیبان هم‌خوانی داشته و می‌توان با تکیه بر این سه مشخصه به پیش‌بینی وضعیت آینده رفتار رویگردانی مشتریان سازمان پرداخت. بنابراین، با توجه به نتایج به دست آمده، مشخصه سابقه خرید مهم‌ترین مشخصه پیش‌بینی‌کننده رویگردانی با بیشترین ضریب است و با توجه به آن که درجه اهمیت مشخصه‌ی سابقه خرید اختلاف مقداری زیادی با دو مشخصه‌ی نحوه آشنایی با سازمان و تمایل به خرید دارد، مشخصه‌ی غالب بر دو مشخصه‌ی دیگر در پیش‌بینی رویگردانی مشتری محسوب می‌شود. نتایج پژوهش نشان می‌دهد که بیمه‌گذارانی که سابقه خرید از سازمان را نداشته‌اند و یا به عبارتی دیگر، جزء دسته بیمه‌گذاران خرید اول یا جذب‌شده از سازمان‌های دیگر بودند، بیشترین رویگردانی را داشته‌اند. مشخصه سابقه خرید، مهم‌ترین عامل پیش‌بینی‌کننده رویگردانی در سازمان مورد مطالعه پیش‌بینی شده است. بنابراین در گام اول، سازمان باید بیمه‌گذاران خرید اول و جذب‌شده از سازمان‌های دیگر را، به عنوان گروه هدف قرار دهد و با استفاده از راهکارهای پیشگیرانه، از رویگردانی این دو گروه جلوگیری کند. مشخصه پیش‌بینی‌کننده بعدی، نحوه آشنایی با سازمان است. با استناد بر نتایج، می‌توان بیان کرد که میزان رویگردانی در بیمه‌گذارانی که به‌دلیل داشتن

قرارداد با سازمان، بیمه‌نامه‌ی خود را در موعد قرارداد تمدید می‌کنند، بسیار کمتر از بیمه‌گذارانی است که داوطلبانه به سازمان مراجعه می‌کنند. بنابراین گروه هدف بعدی، بیمه‌گذاران غیر قراردادی سازمان هستند. از آن جایی که میزان رویگردانی در این گروه نیز زیاد است، سازمان باید اقدامات پیشگیرانه را در این گروه نیز اعمال کند. مشخصه بعدی که با میزان درجه اهمیت کمتری به عنوان پیش‌بینی‌کننده رویگردانی انتخاب شده است، مشخصه تمایل به خرید می‌باشد. نتایج نشان می‌دهد که بیمه‌گذارانی که تمایل کمتری برای تمدید بیمه‌نامه از خود نشان می‌دهند، بیمه‌گذارانی که فاصله زمانی میان اتمام قرارداد قبلی آن‌ها و تمدید قرارداد جدیدشان بسیار طولانی است، رویگردان‌تر هستند. بنابراین باید دسته بیمه‌گذاران با تمایل به خرید کم را به عنوان گروه هدف بعدی انتخاب کرد.

باید به این نکته نیز توجه کرد که گروه هدف قرار دادن دسته‌های نام‌برده، به‌معنی کم‌توجهی و ارائه خدمات نامناسب به دسته‌های بیمه‌گذاران سابقه خرید از سازمان، قراردادی و تمایل به خرید متوسط به بالا نیست. از آن جایی که این دسته از مشتریان بسیار وفادار به سازمان هستند، می‌توان با برخورد مناسب با آن‌ها و یا یک هدیه کوچک همراه با تمدید بیمه‌نامه‌ها، آن‌ها را برای سالیان سال برای سازمان حفظ کرد. با این توضیح، به سراغ بانک اطلاعاتی سازمان می‌رویم و با توجه به مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی، به پیش‌بینی رفتار نمونه‌ای از مشتریان سازمان می‌پردازیم. از آن جایی که سه مشخصه سابقه خرید، نحوه آشنایی با سازمان و تمایل به خرید، به عنوان مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی تعیین شده و تأثیر سایر مشخصه‌ها در پیش‌بینی رویگردانی بسیار ناچیز بوده است، برای پیش‌بینی رفتار آینده مشتری توجه خود را تنها به این سه مشخصه اصلی معطوف خواهیم کرد. باید به این نکته نیز توجه داشت که درجه اهمیت مشخصه سابقه خرید دارای مقدار عددی بالایی بوده و به همین دلیل مشخصه غالب بر دو مشخصه دیگر محسوب می‌شود.

حال برای پیش‌بینی رفتار آینده مشتریان سازمان، به صورت تصادفی سه مشتری را از میان مشتریان سازمان انتخاب کرده و با توجه به نتایج به دست آمده از مدل ماشین بردار پشتیبان، به پیش‌بینی وضعیت رفتار رویگردانی مشتریان می‌پردازیم. باید به این نکته توجه داشت که پیش‌بینی انجام شده توسط مدل ماشین بردار پشتیبان مبتنی بر آموزشی است که مدل در مرحله آموزش دیده است. در مرحله آموزش، مدل با استفاده از مجموعه داده آموزشی، به الگو نهفته در آن پی برده و سپس با استفاده از این الگو به پیش‌بینی وضعیت مجموعه داده ارزیابی می‌پردازد. مدل ماشین بردار پشتیبان با کرنل خطی از رابطه $D(x)=sign\,(w^T x_t+b)=sign\left(\sum_{i=1}^{n}\lambda_i y_i x_i^T x_t+b\right)$ برای تعیین درجه اهمیت مشخصه‌ها و پیش‌بینی وضعیت رویگردانی مشتری استفاده می‌کند. بر اساس نتایج به دست آمده، مدل ماشین بردار پشتیبان می‌تواند با صحت ٪۸۹ این پیش‌بینی را انجام دهد. در جدول (۶) وضعیت رفتار رویگردانی سه مشتری با توجه به مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی بر مبنای مدل ماشین بردار پشتیبان پیش‌بینی شده است.

جدول ۵. پیش‌بینی وضعیت رفتار رویگردانی مشتری با استفاده ازمدل ماشین بردار پشتیبان

شماره مشتری	سابقه خرید	نحوه آشنایی با سازمان	تمایل به خرید	پیش‌بینی وضعیت رویگردانی مشتری
مشتری اول	دارد	قراردادی	مثبت	غیر رویگردان
مشتری دوم	ندارد	قراردادی	معمولی	رویگردان
مشتری سوم	ندارد	غیر قراردادی	مثبت	رویگردان

با توجه به جدول (۶) مدل ماشین بردار پشتیبان برای پیش‌بینی وضعیت رویگردانی مشتری اول در ابتدا به مشخصه سابقه خرید وی توجه می‌کند. مشتری اول دارای سابقه خرید از سازمان بوده و بنابراین با احتمال بالایی در دسته مشتریان غیر رویگردان قرار خواهد گرفت. سپس به مشخصه نحوه آشنایی با سازمان توجه کرده و درمی‌یابیم که جزء مشتریان دارای قرارداد با سازمان است. بعد از آن مشخصه تمایل به خرید را مورد بررسی قرار داده و به تمایل مثبت این مشتری به خرید پی خواهد برد. بنابراین، با توجه به نتایج به دست آمده، مشتری اول دارای تمامی خصوصیات مشتریان دسته‌ی غیر رویگردان بوده و پیش‌بینی می‌شود که این مشتری در دسته مشتریان غیر رویگردان قرار بگیرد. مشتری دوم دارای سابقه خرید از سازمان نیست. این مشتری جزء مشتریان قراردادی سازمان و تمایل به خرید وی نیز معمولی ارزیابی شده است. با توجه به این شرایط، مدل ماشین بردار پشتیبان مشتری دوم را به‌دلیل دارا بودن تنها یک خصوصیت قراردادی با درجه اهمیت پایین، در دسته مشتریان رویگردان پیش‌بینی می‌کند. مشتری سوم نیز دارای عدم سابقه خرید از سازمان و غیر قراردادی است. وی دارای تمایل به خرید مثبت است. با توجه به این‌که مشتری سوم تنها تمایل به خرید مثبت داشته و این مشخصه دارای درجه اهمیت کمتری در بین دو مشخصه دیگر است، مغلوب دو مشخصه دیگر خواهد شد. در نتیجه مدل ماشین بردار پشتیبان این مشتری را در دسته مشتریان رویگردان قرار خواهد داد.

با توجه به مشخصه‌های اصلی تعیین‌شده به عنوان پیش‌بینی‌کننده‌های رویگردانی مشتری در بخش قبل، سازمان

خواهد توانست مشتریان متمایل به رویگردانی را شناسایی کند. بعد از تعیین مشتریان متمایل به رویگردانی، می‌توان با اجرای راهکارهای پیشگیرانه از رویگردانی این دسته از مشتریان جلوگیری کرد. بنابراین، با توجه به عوامل تأثیرگذار بر رویگردانی مشتریان سازمان مورد مطالعه، راهکارهای پیشگیرانه به‌منظور جلوگیری از رویگردانی مشتریان و افزایش سودآوری سازمان ارائه خواهد شد.

۱) دو مشخصه اول پیش‌بینی‌کننده رویگردانی در سازمان مورد مطالعه، مشخصه سابقه خرید و نحوه آشنایی با سازمان است. مشتریان دارای عدم سابقه خرید و مشتریانی که به صورت داوطلبانه به سازمان مراجعه می‌کنند، نسبت به مشتریان دارای سابقه خرید از سازمان و مشتریان قراردادی، تمایل به رویگردانی بیشتری دارند. طراحی بسته‌های تخفیفی، اقدام پیشگیرانه برای این دو دسته از مشتریان است. در این شرایط می‌توان با مشتریان عدم سابقه خرید و غیر قراردادی سازمان، تعهدنامه‌ای مبتنی بر میزان تخفیف بیشتر به شرط تمدید بیمه‌نامه برای دو سال متوالی آینده منعقد کرد. بدین معنی که اگر مشتری هنگام تمدید قرارداد و یا همان ابتدای مراجعه به سازمان، تعهد خرید بیمه برای دو سال متوالی آینده را بدهد، به او تخفیفی بیشتر در طی این سال‌ها تعلق خواهد گرفت. این بسته‌های تخفیفی می‌تواند برای سه سال و پنج سال متوالی نیز تعمیم داده شود که باید به این نکته نیز توجه داشت که بسته به تعداد سال‌های تعهد خرید، تخفیف بیشتری هم به مشتری تعلق خواهد گرفت.

۲) از دیگر راهکارهای پیشگیرانه برای گروه مشتریانی که سابقه خرید از سازمان ندارند، اطلاع رسانی به این دسته از مشتریان و آگاه کردن آن‌ها از وضعیت موجود در سازمان‌های بیمه و بیان شرایط، مزیت‌ها و تسهیلات سازمان خود است. بدین ترتیب می‌توان از رویگردانی مشتریانی که به‌دلیل عدم آگاهی از شرایط موجود و با هدف بررسی شرایط سایر سازمان‌ها به سازمان دیگر روی می‌آورند، جلوگیری کرد.

۳) مشخصه‌ی اثرگذار بعدی تمایل به خرید است. مشتریانی که فاصله‌ی زمانی بین اتمام قرارداد قبلی و تمدید قرارداد جدید آن‌ها طولانی است، تمایل بیشتری به رویگردانی از خود نشان می‌دهند. راهکار پیشگیرانه برای جلوگیری از رویگردانی این دسته از مشتریان، در نظر گرفتن طرح تخفیفی برای زمان تمدید قرارداد بیمه‌نامه است. این طرح بدین صورت است که بیمه‌گذارانی که در مدت زمان کمتر از یک هفته بعد از زمان اتمام قرارداد خود، برای تمدید قرارداد سال بعد به سازمان مراجعه کنند، مشمول طرح تخفیفی ویژه‌ای خواهند شد. این طرح برای بیمه‌گذارانی که در زمانی بیشتر از یک هفته بعد از پایان قرارداد خود به سازمان مراجعه کنند، برقرار نخواهد بود. بدین صورت، انگیزه‌ی بیشتری برای تمدید زودتر قرارداد برای مشتریان با تمایل به خرید پایین به وجود آمده و از رویگردانی آن‌ها جلوگیری خواهد شد.

۴) طراحی سیستم رسیدگی به شکایات مشتری، راهکار پیشگیرانه بعدی است که می‌تواند از رویگردانی تمام دسته‌های مشتریان متمایل به رویگردانی سازمان جلوگیری کند. با طرحی این سیستم به صورت اینترنتی و یا به صورت یک بخش رسیدگی به شکایات در درون سازمان، این امکان برای مشتریان فراهم می‌شود که انتقادات خود را چه از نظر برخورد کارکنان و چه از نظر نحوه خدمت‌دهی و کیفیت آن، به گوش مسئولان سازمان برسانند. این گونه مدیران و مسئولان می‌توانند از نظرات و پیشنهادات مشتریان آگاه شوند و سعی در رفع کمبودها و کاستی‌ها برآیند. علاوه بر این، فرصتی به سازمان داده می‌شود تا از رویگردانی ناگهانی مشتریان خود جلوگیری کنند.

۵. نتیجه‌گیری و جمع‌بندی

در این مقاله، توانایی تکنیک ماشین بردار پشتیبان در پیش‌بینی رویگردانی مشتریان بیمه‌ی بدنه‌ی خودرو نشان داده شد. در این تحقیق، داده‌های یک سازمان خدمات بیمه‌ای به عنوان سازمان مورد مطالعه انتخاب شد. پس از تعیین جمعیت بیمه‌گذاران نهایی از میان مجموعه کلی رکوردهای مربوط به عقد قراردادها در بازه زمانی سه سال و سه ماه، انتخاب مشخصه‌های بهینه از میان مشخصه‌های موجود، با استفاده از الگوریتم ژنتیک صورت گرفت. انتخاب مشخصه با حذف مشخصه‌های کم‌اهمیت، علاوه بر کاهش بعد مجموعه داده، بر افزایش صحت مدل می‌افزاید. نتایج الگوریتم ژنتیک نشان‌دهنده‌ی اهمیت بالاتر مشخصه‌های نوع خودرو، مدل خودرو، تخفیف عدم خسارت، مدت قرارداد، سابقه خرید، تمایل به خرید و نحوه آشنایی با سازمان، برای پیش‌بینی رویگردانی مشتریان می‌باشد. استفاده از الگوریتم ژنتیک برای انتخاب مشخصه، از نوآوری‌های مطالعه‌ی حاضر بوده و تاکنون در مطالعات پیش‌بینی رویگردانی دیده نشده است.

سپس به مدل‌سازی مسأله با استفاده از تکنیک ماشین بردار پشتیبان در دو حالت خطی و غیر خطی و با در نظر گرفتن سه تابع کرنل خطی، چند جمله‌ای و گوسی پرداختیم. این روش به‌دلیل ماهیت کمینه‌سازی ریسک دسته‌بندی و عملکرد بالا، به عنوان روش اصلی تحقیق معرفی شده است. ترکیب دو روش

ماشین بردار پشتیبان و الگوریتم ژنتیک نیز برای اولین بار در این تحقیق به چشم می‌خورد. علاوه بر این، دو روش جستجوی شبکه و ۱۰-fold cross-validation برای دست‌یابی به مقدار بهینه‌ی پارامتر C تابع کرنل خطی، پارامترهای C و d تابع کرنل غیر خطی چند جمله‌ای و پارامترهای C و σ تابع کرنل غیر خطی گوسی ماشین بردار پشتیبان پیاده‌سازی شده است. از آن جایی‌که عملکرد روش ماشین بردار پشتیبان به شدت به مقادیر پارامترهای توابع کرنل وابسته می‌باشد، فرایند بهینه‌سازی پارامترها تأثیر فراوانی بر عملکرد آن داشته است. به عبارت دیگر می‌توان بیان کرد که تکنیک انتخاب پارامتر برای مدل‌سازی مسأله یک امر حیاتی است. چرا که اندکی تغییر در عملکرد پیش‌بینی رویگردانی، می‌تواند تأثیر زیادی بر بازگشت سرمایه‌ی فعالیت‌های نگه‌داری بازاریابی داشته باشد [۱۵]. کاربرد روش ماشین بردار پشتیبان در صنعت بیمه و ترکیب آن با الگوریتم ژنتیک نیز از نوآوری‌های این مطالعه به شمار می‌آید.

بعد از آن با استفاده از روش ۱۰-fold cross-validation و معیار دقت، ارزیابی و اعتبار سنجی مدل ماشین بردار پشتیبان انجام شد. مدل ماشین بردار پشتیبان با کرنل خطی توانسته است با صحت حدود ٪۸۹ به پیش‌بینی رویگردانی مشتریان بپردازد. همچنین برای بررسی صحت اعتبار نتایج حاصل از این روش، عملکرد آن با هفت روش دسته‌بندی معروف و پرکاربرد درخت تصمیم ID3، درخت تصمیم CHAID، شبکه‌های عصبی، رگرسیون لجستیک، جنگل تصادفی، دسته‌بندی کننده‌ی بیزی و K نزدیک‌ترین همسایگی نیز مقایسه شده است. برای بهبود عملکرد تکنیک‌های دسته‌بندی نام‌برده، پارامترهای مربوط به کلیه‌ی روش‌ها نیز با روش جستجوی شبکه بهینه شده‌اند. مقایسه‌ی عملکرد ماشین بردار پشتیبان با هفت تکنیک نام‌برده و بهینه‌سازی پارامترهای آنان نیز، نیز از دیگر نوآوری‌های این مطالعه است. نتایج پژوهش نشان می‌دهد که تکنیک ماشین بردار پشتیبان از صحت بالاتری نسبت به روش‌های دیگر برای پیش‌بینی رویگردانی مشتریان سازمان مورد مطالعه برخوردار است. این موضوع نشان از کارایی بالای تکنیک ماشین بردار پشتیبان در صنعت بیمه است. از سوی دیگر، با توجه به نتایج حاصل از مدل‌سازی مسأله به روش ماشین بردار پشتیبان، مشخصه‌های سابقه خرید، نحوه آشنایی با سازمان و تمایل به خرید، به عنوان مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی مشتری به دست آمدند. بعد از تعیین مشخصه‌های اصلی پیش‌بینی‌کننده رویگردانی، چهار راهکار پیشگیرانه به‌منظور جلوگیری از رویگردانی مشتریان سازمان ارائه شدند. ارائه‌ی راهکارهای پیشگیرانه در این تحقیق نیز از دیگر نوآوری‌ها محسوب می‌شود.

به عنوان پیشنهاد برای تحقیقات آتی می‌توان به‌منظور در نظر گرفتن تأثیر متغیرهای مختلف دیگر در پیش‌بینی رویگردانی، از اطلاعات ثبت‌شده در واحدهایی علاوه بر واحد صدور مانند واحد مالی، خسارت، رسیدگی به شکایات (در صورت وجود) نیز استفاده نمود. همچنین می‌توان با استفاده از یک رویکرد ترکیبی داده‌کاوی و پرسش‌نامه‌ای، نقش شبکه‌ی فروش را در انتخاب و حفظ ارتباط با یک سازمان جویا شد؛ زیرا در بازار بیمه ایران نقش واسطه‌ها، نماینده‌ها و کارگزاران غیر قابل انکار است و گروه زیادی از مشتریان به‌جای وفاداری به شرکت بیمه، به واسطه‌ها وفادارند. مورد بعدی برداشت برش‌های زمانی متناوب از پایگاه‌های داده شرکت و پیش‌بینی رویگردانی مشتریان با در دست داشتن این بانک اطلاعاتی و اجرای فرایند پیش‌بینی رویگردانی پویا است. نتایج این مدل از این لحاظ که روند تغییرات در ترجیحات مشتریان را منعکس می‌کند، می‌تواند نقش شایانی در برنامه‌ریزی‌های سازمان ایفا کند. همچنین ترکیب تکنیک‌های دسته‌بندی مانند تکنیک ماشین بردار پشتیبان با تکنیک‌های مربوط به کارکرد خوشه‌بندی نیز می‌تواند دید متفاوتی را برای پژوهش‌گران به‌منظور پیش‌بینی رویگردانی مشتریان ایجاد کرده و در ارائه‌ی سیاست‌های پیشگیرانه مؤثر باشد.

پی نوشت

1. Transfer points location problem
2. The multiple location of transfer point
3. Facility and transfer point location problem
4. Customer Relationship Management (CRM)
5. Support Vector Machine (SVM)
6. Madden et al.
7. Binomial Probit Model
8. Wei and Chiu
9. Burez and Van Den Poel
10. Glady et al.
11. AdaCost
12. Cost-sensitive decision tree
13. Shim et al.
14. Hierarchical Multiple Kernel Support Vector Machine (H-MK-SVM)
15. Multiple Kernel Support Vector Machine (MK-SVM)
16. Migueis et al.
17. Multivariate Adaptive Regression Splines (MARS)
18. Kim et al.
19. Van Den Poel and Lariviere
20. Survival Analysis
21. Kim and Yoon
22. Ahn et al.
23. Wang et al.
24. Tsai and Chen
25. Vapnik

26. Input Space
27. Kernel Function
28. RapidMiner
29. Hard Margin
30. Soft Margin
31. Feature Space
32. Matlab
33. Linear Kernel
34. Polynomial Kernel
35. Gaussian Radial Basis Function Kernel (RBF)
36. Interactive Dichotomizer version ۳
37. Quinlan
38. Chi Square Automatic Interaction Detector
39. Chi Square Test
40. Ensemble Classifier
41. k-fold cross-validation

مراجع

[1] Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y., "Credit card churn forecasting by logistic regression and decision tree", Expert Systems with Applications, 2011,Vol. 38, pp. 15273-15285.

[۲] سپهری، محمد مهدی.، کارگری، مهرداد.، «بهبود الگوریتم خوشه‌بندی مشتریان برای توزیع قطعات یدکی با رویکرد داده‌کاوی (k-means)»، نشریه بین‌المللی مهندسی صنایع و مدیریت تولید دانشگاه علم و صنعت ایران، (۱۳۹۰)، جلد ۲۳، شماره ۲، صفحه ۱۷۹-۱۷۲.

[۳] سید حسینی، سید محمد.، غلامیان، محمدرضا.، ملکی، آناهیتا.، «طراحی یک متدولوژی بر مبنای RFM جهت سنجش وفاداری مشتری بر مبنای تکنیک‌های داده‌کاوی»، نشریه بین‌المللی مهندسی صنایع و مدیریت تولید دانشگاه علم و صنعت ایران، (۱۳۹۰)، جلد ۲۲، شماره ۲، صفحه ۲۴۹-۲۴۰.

[4] Tsai, Ch. F., & Lu, Y. H., "Customer churn prediction by hybrid neural networks", Expert Systems with Applications, (2009), Vol. 36, pp. 12457-12553.

[5] Ngai, E. W. T., Xiu, L., & Chau, D. C. K., "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications, (2009), Vol. 36, pp. 2592-2602.

[6] Migueis, V. L., Camanho, A., & Cunha, J. F., "Customer attrition in retailing: An application of Multivariate Adaptive Regression Splines", Expert Systems with Applications, (2013), Vol. 40, pp. 6225-6232.

[7] Gur Ali, O., & Ariturk, U., "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking", Expert Systems with Applications, (2014), Vol. 41, pp. 7889-7903.

[8] Madden, G., Savage, S. J., & Coble-Neal, G., "Subscriber churn in the Australian ISP market", Information Economics and Policy, (1999), Vol. 11, pp. 195–207.

[9] Wei, Ch. P., & Chiu, I. T., "Turning telecommunications call details to churn prediction: a data mining approach", Expert Systems with Applications, (2002), Vol. 23, pp. 103-112.

[10] Burez, J., & Van den Poel, D., "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services", Expert Systems with Applications, (2007), Vol. 32, pp. 277-288.

[11] Glady, N., Baesens, B., & Croux, Ch., "Modeling churn using customer lifetime value", European Journal of Operational Research, (2009), Vol. 197, pp. 402–411.

[12] Shim, B., Choi, K., & Suh, Y., "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns", Expert Systems with Applications, (2012), Vol. 39, pp. 7736–7742.

[13] Chen, Zh. Y., Fan, Zh. P., & Sun, M., "A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data", European Journal of Operational Research, (2012), Vol. 223, pp. 461-472.

[14] Kim, K., Jun, Ch. H., Lee, J., "Improved churn prediction in telecommunication

Morgan Kaufmann, (2006).

[23] Hadden, J., Tiwari, A., Roy, R., & Ruta, D., "Computer assisted customer churn management: State-of-the-art and future trends", Computers & Operations Research, (2005), Vol. 34, pp. 2902-2917.

[24] Vapnik, V. N., The Nature of Statistical Learning Theory, Springer-Verlag, New York, (1995).

[25] Burges, Ch. J. C., "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, (1998), Vol. 2, pp. 121–167.

[26] Cortes, C., Vapnik, V., "Support-Vector Networks, Machine Learning", (1995), Vol. 20, pp. 273-297.

[27] Vapnik, V. N., Statistical Learning Theory, Wiley-Interscience Publication, New York, 1998).

[28] Quinlan, J. R., C4.5: Programs for machine learning, San Francisco, CA: Morgan Kaufman, (1993).

[29] Michael, J.A., & Gordon, S.L., Data Mining Technique: For Marketing, Sales and Customer Support, Wiley, New York, (1997).

[30] Au,W. H., Chan, K. C. C., & Yao, X., "A Novel Evolutionary Data Mining Algorithm With Application to Churn Prediction", IEEE Transactions on Evolutionary Computation, (2003), Vol. 7, pp. 532-545.

[31] Bigus, J. P., Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support, McGraw-Hill, New York, (1996).

[32] Berry, M. J. A., & Linoff, G. S., Data Mining Techniques for Marketing, Sales and Customer Relationship Management, 2nd edition, Wiley, USA, (2004).

[33] Cutler, A., Cutler, D. R., & Stevens, J. R., Ensemble Machine Learning: Random Forests, Springer US, (2012).

industry by analyzing a large network", Expert Systems with Applications, (2014), Vol. 41, pp. 6575-6584,.

[15] Van den Poel, D., & Lariviere, B., "Customer attrition analysis for financial services using proportional hazard models", European Journal of Operational Research, (2004), Vol. 157, pp. 196-217.

[16] Kim, H. S., & Yoon, Ch. H., "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market", Telecommunications Policy, (2004), Vol. 28, pp. 751-765.

[17] Ahn, J. H., Han, S. P., & Lee, Y. S., "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry", Telecommunications Policy, (2006), Vol. 30, pp. 552-568.

[18] Wang, Y. F., Chiang, D. A., Hsu, M. H., Lin, Ch. J., & Lin, I. L., "A recommender system to avoid customer churn: A case study", Expert Systems with Applications, (2009), Vol. 36, pp. 8071-8075.

[19] Tsai, Ch. F., & Chen, M. Y., "Variable selection by association rules for customer churn prediction of multimedia on demand", Expert Systems with Applications, (2010), Vol. 37, pp. 2006–2015.

[20] Coussement, K., & Van den Poel, D., "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques", Expert Systems with Applications, (2008), Vol. 34, pp. 313-327.

[۲۱] کاظمی، ابوالفضل.، ابوطالب، سیامک.، «ارائه یک مدل بهینه‌سازی ریاضی چندهدفه برای طبقه‌بندی در ریاضی»، نشریه بین‌المللی مهندسی صنایع و مدیریت تولید دانشگاه علم و صنعت ایران، (۱۳۹۱)، جلد ۲۳، شماره ۴، صفحه ۴۸۶-۵۰۱.

[22] Han, J., & Kamber, M., Data Mining Concepts and Techniques, Second Edition,

[34] Russel, S. J., & Norvig, P., Artificial intelligence: a modern approach, Pearson Education International, USA, 2nd edition, (2004).

[35] Thanuja, V., Venkateswarlu, B., & Anjaneyulu, G. S. G. N., "Applications of Data Mining in Customer Relationship Management", J. Comp. & Math. Sci, (2011), Vol. 2, pp. 423-433.