



## شناسایی عوامل تاثیر گذار در رویگردانی مشتریان شرکت مخابرات کردستان و ارائه مدل هایی برای پیش بینی رویگردانی با استفاده از الگوریتم های یادگیری ماشین

ویدا صادقی

فارغ التحصیل مهندسی کامپیوتر، واحد سنندج، دانشگاه آزاد اسلامی، سنندج، ایران؛ کارمند مخابرات استان کردستان

انور بهرام پور (نویسنده مسؤل)

استادیار، گروه مهندسی کامپیوتر، واحد سنندج، دانشگاه آزاد اسلامی، سنندج، ایران

Email: bahrapour@iausdj.com

سید علی حسینی

مدرس گروه مهندسی کامپیوتر، واحد سنندج، دانشگاه آزاد اسلامی، سنندج، ایران

تاریخ دریافت: ۱۴۰۲/۰۳/۰۹ \* تاریخ پذیرش ۱۴۰۲/۰۶/۰۶

### چکیده

مشتریان منبع اصلی درآمد و دارایی مهم برای هر سازمان هستند. با این دیدگاه، امروزه شرکتها تلاش بیشتری را برای حفظ مشتریان موجود آغاز کرده‌اند. از آنجا که در بسیاری از شرکتها هزینه به دست آوردن مشتری جدید بسیار بالاتر از هزینه ایجاد رضایتمندی مشتری فعلی است، رویگردانی مشتری به حوزه اصلی نگرانی این شرکتها تبدیل شده است. لذا شرکت‌های مبتنی بر مشتری از جمله شرکت‌های فعال در صنعت مخابرات به دلیل رویگردانی مشتریان با چالش بزرگی روبرو هستند. با توسعه سریع صنعت مخابرات، پیش‌بینی رویگردانی به عنوان یکی از فعالیتهای اصلی در به دست آوردن مزیت رقابتی در بازار محسوب می‌شود. پیش‌بینی رویگردانی مشتری به اپراتورها اجازه می‌دهد تا قبل از مهاجرت مشتریان فعلی به اپراتورهای دیگر، یک دوره زمانی برای اصلاح و اجرای یک سری اقدامات پیشگیرانه داشته باشند. در این پژوهش یک سیستم پشتیبانی تصمیم برای پیش‌بینی و تخمین رویگردانی مشتریان شرکت مخابرات استان کردستان (دارای ۵۲۹۰۰۰ مشترک) با روش‌های مختلف داده‌کاوی و یادگیری ماشین (شامل رگرسیون خطی چندگانه (MLR)، رگرسیون چند جمله‌ای (PR)، رگرسیون لجستیک، شبکه‌های عصبی مصنوعی، آدابوست و جنگل تصادفی) ارزیابی شده است. نتایج ارزیابی‌های انجام شده بر روی مجموعه داده‌های شرکت مخابرات استان کردستان عملکرد بالای روش‌های شبکه‌های عصبی مصنوعی با دقت ۹۹.۹٪، آدابوست با دقت ۱۰۰٪ و جنگل تصادفی با دقت ۱۰۰٪ را نشان می‌دهد.

**کلمات کلیدی:** پیش‌بینی، رویگردانی مشتریان، داده‌کاوی، یادگیری ماشین، شبکه‌های عصبی مصنوعی.

## ۱- مقدمه

گام نهادن به اقتصاد خدماتی و از میان رفتن مرزهای جغرافیایی برای کسب و کارها و به تبع آن شدت یافتن رقابت باعث شده تا مشتری به عنوان رکن اساسی و محور اصلی فعالیتهای شرکت‌ها مطرح شود. امروزه رویگردانی<sup>۱</sup> مشتری بسیاری از صنایع را تهدید می‌کند و شرکت‌ها را وادار می‌کند تا برنامه‌ها و پیشنهادات حفظ مشتری هدفمند و فعالانه را به کار گیرند. صنعت مخابرات شاهد بازارهای بسیار اشباع شده در سراسر جهان است. بنابراین، حفظ مشتریان فعلی یک نگرانی اصلی برای شرکت‌ها به منظور حفظ سود پایدار است. بر این اساس پیش‌بینی رویگردانی مشتری در صنعت مخابرات نیز یک عامل بسیار ضروری برای دستیابی به سود مناسب است و تأثیر مستقیمی بر حفظ مشتری و افزایش درآمد دارد (Wu et al., 2022). به عبارت دیگر، پیش‌بینی رویگردانی اپراتورهای مخابراتی را وادار می‌سازد تا استراتژی‌های بازاریابی موثرتری را بر اساس تجزیه و تحلیل گسترده مشتری ارائه دهند و از گردش مشتری جلوگیری کنند.

شرکت‌ها انواع مختلفی از داده‌های مشتری مانند رفتار شخصی، جمعیتی، صورت‌حساب استفاده و ... را ذخیره می‌کنند و از آنها بهره‌برداری می‌کنند. مطالعات بسیاری برای ارائه مدل‌های پیش‌بینی رویگردانی مشتریان مخابراتی بر اساس داده‌های در دسترس از مشتریان تمرکز دارند (Jafari-Marandi et al., 2020). تجزیه و تحلیل مشتری در صنعت مخابرات شامل دو جزء کلیدی، یعنی پیش‌بینی رویگردانی و تقسیم‌بندی مشتری در گروه‌های با ویژگی‌های مشابه است. هزینه به دست آوردن مشتریان جدید ۵ تا ۱۰ برابر بیشتر از هزینه حفظ مشتریان فعلی است. پیش‌بینی رویگردانی مشتری به اپراتورها اجازه می‌دهد تا قبل از مهاجرت مشتریان فعلی به اپراتورهای دیگر، یک دوره زمانی برای اصلاح و اجرای یک سری اقدامات حفظ مشتری داشته باشند. از سوی دیگر، تقسیم‌بندی مشتریان وسیله مهمی برای انجام تجزیه و تحلیل مشتریان است تا آنها را در گروه‌های مختلف با توجه به طرح‌های مختلف قرار دهد. در (Vo et al., 2021) یک چارچوب تجزیه و تحلیل مشتری یکپارچه با هدف دستیابی به تخصیص کارآمد منابع شرکت و بهبود حفظ مشتری، برای مدیریت رویگردانی در صنعت مخابرات ارائه شده است. از آنجا که ریزش مشتری برای درآمد شرکت مضر است، لذا توسعه یک مدل خوب و موثر پیش‌بینی ریزش بسیار مهم است، چالش اصلی پیش‌بینی سودآوری مشتری، عدم تقارن است، زیرا تعداد مشتریان بسیار سودآور در مقایسه با سایرین بسیار کم است و به دلیل عدم تقارن (چولگی شدید)، دقت پیش‌بینی‌ها اغلب با افزایش ارزش سود کاهش می‌یابد. لذا، از راهکارهای مبتنی بر داده‌کاوی برای غلبه بر این مشکلات استفاده می‌شود (Rogić et al., 2020).

تحقیقات متنوعی بر روی تحلیل‌های مختلف بازار برای پیش‌بینی نرخ رویگردانی بسته به روش‌های طبقه‌بندی، تحلیل آماری و تکنیک‌های هوش مصنوعی و غیره ارائه شده است. از جمله این پژوهش‌ها به می‌توان به موارد زیر اشاره کرد: توکلی و همکاران، در پژوهشی روی داده‌های شرکت بیمه، با هدف تبیین قابلیت‌های داده‌کاوی در مدیریت رویگردانی مشتری، و با بهره‌گیری از متدولوژی استاندارد داده‌کاوی CRISP-DM، به کاوش در پایگاه‌های داده یکی از شرکتهای سهامی عام بیمه‌ای در رشته بیمه آتش سوزی پرداخته است. نتایج نشان می‌دهد کانال جذب مشتری عامل اصلی پیش‌بینی کننده رویگردانی یا ماندگاری مشتری در شرکت بوده و در مراتب بعد سابقه خرید و کاربری مکان بیمه شده به عنوان عوامل پیش‌بینی کننده رویگردانی قرار می‌گیرند (Tavakoli et al., 2011).

نجمی و همکاران، با کاوش در داده‌های بانک شهر، به منظور خوشه‌بندی مشتریان، از نقشه‌های خودسازمان دهنده شبکه عصبی که یک روش یادگیری بدون نظارت است، استفاده و برای دسته‌بندی از ماشین بردار پشتیبان و درخت تصمیم استفاده شده است. روش استفاده از این ابزارها به این صورت است که ابتدا از دو مشخصه میانگین موجودی و میانگین تراکنش مشتریان در دوره سه ماهه پایانی استفاده شده و به عنوان ورودی شبکه عصبی در خوشه‌بندی مورد استفاده قرار گرفته است. پس از آن، در مرحله کلاس‌بندی، از داده‌های مربوط به تراکنش‌های نقدی و اعتباری به منظور کلاس

<sup>1</sup> Churn Prediction

بندی و پیش بینی استفاده شده است. نتایج به دست آمده حاکی از آن است که مدل پیشنهادی بیش از ۸۰٪ توانایی پیش بینی رویگردانی مشتری را داشته و ماشین بردار پشتیبان عملکرد بهتری از درخت تصمیم نشان داده است (Najmi et al., 2018).

امیری و همکاران، در پژوهشی بر روی مشتریان سرویس های اینترنت یکی از بزرگ ترین شرکت های مخابراتی ایران، الگوریتم های "جنگل تصادفی"، "ماشین بردار پشتیبان" و "K نزدیک ترین همسایگان" برای طبقه بندی مشتریان رویگردان و غیررویگردان به کار گرفتند که معیارهای ارزیابی، نشان دهنده ی برتری الگوریتم جنگل تصادفی است (Amiri et al., 2022).

اخیراً چندین طبقه بندی مانند الگوریتم یادگیری مونتاژ<sup>۲</sup> (ALA)، تکنیک نمونه برداری بیش از حداقلیت مصنوعی (SMOTE) و الگوریتم تقویت<sup>۳</sup> (BA)، برای حفظ نرخ مشتری برای بهبود فروش استفاده شده است. اگرچه برخی تحقیقات مانند الگوریتم SMOTEBA وجود دارد، روش های سنتی نیز برای توسعه نرخ مشتری بدون رویگردانی بهبود یافته اند. اما نتایج بهینه به دلیل نرخ رویگردانی نامتعادل در بازار تجارت الکترونیک حاصل نشده است. به طور کلی، این مجموعه داده ها از دو کلاس تشکیل شده اند: کلاس اقلیت و اکثریت، از نظر کلاس اکثریت تحلیل بازار، در کلاس منفی و اقلیت در منطقه مثبت مشتق می شوند (Lemmens & Gupta, 2020). هوموده و همکاران به بررسی مدل انبوه انتخابی برای پیش بینی نرخ خروج مشتریان در صنعت ارتباطات تلفن همراه می پردازد و نشان می دهد که استفاده از این مدل انبوه انتخابی با ترکیب چندین مدل ماشین یادگیری، می تواند بهبود قابل توجهی در پیش بینی نرخ خروج مشتریان نسبت به روش های سنتی داشته باشد، که این نتایج می تواند به توسعه روش های بهبود یافته در مدیریت و کاهش نرخ خروج مشتریان در صنعت ارتباطات مفید باشد (Hammoudeh et al., 2019). دوهت و همکاران از الگوریتم های نمونه برداری هندسی و آدابوست در کنار یادگیری عمیق استفاده می کند تا بهبود قابل توجهی در عملکرد سیستم ها و کاهش خطاها در حوزه تجارت الکترونیک ایجاد کند (Dhote et al., 2020).

علاوه بر این، در سال ۲۰۱۹ سیواسانکر و همکاران به بررسی روش های انتخاب ویژگی برای پیش بینی حفظ مشتریان در صنعت ارتباطات تلفنی می پردازد. این مقاله تلاش می کند تا با استفاده از تکنیک های متنوعی از جمله تحلیل ارتباط ویژگی ها، تحلیل اهمیت ویژگی ها و تحلیل خوشه بندی، به بهبود پیش بینی حفظ مشتریان در صنعت ارتباطات تلفنی بپردازد. نتایج مقاله نشان می دهد که استفاده از روش های انتخاب ویژگی مناسب می تواند بهبود قابل توجهی در دقت پیش بینی حفظ مشتریان ایجاد کند (Sivasankar & Vijaya, 2019). در مقاله با عنوان "انتخاب مدل محاسبات ابری برای شرکت های تجارت الکترونیک با استفاده از روش تصمیم گیری زبانی فازی ۲-تایی جدید"، به بحث درباره روش های حل مشکل طبقه بندی داده های نرخ رویگردانی می پردازد. سهیب و همکاران با استفاده از یک رویکرد ترکیبی که شامل یادگیری عمیق مبتنی بر آدابوست و روش نمونه برداری هندسی است و نتایج متفاوتی در مورد دقت، صحت، پیش بینی، قابلیت استفاده، ویژگی و حساسیت ارائه می دهد، مشکل طبقه بندی داده ها را حل می کنند (Sohaib et al., 2019).

هدف اصلی از انجام این مقاله پیش بینی رویگردانی مشتری و تقسیم بندی مشتریان در شرکت مخابرات استان کردستان است که یک موضوع بسیار ضروری در حفظ مشتری و درآمدهای شرکت است، عواملی که در رویگردانی مشتری نقش دارند را می توان با روش های داده کاوی و یادگیری ماشین به خوبی تجزیه و تحلیل کرد. سپس استراتژی های حفظ متفاوتی را برای گروه های مختلف مشتریان پیشنهاد کرد تا به مدیریت رویگردانی و بازاریابی دقیق دست یافت.

ساختار کلی مقاله به شرح زیر است. در بخش ۱ مقدمه و پیشینه ی تئوریک و خلاصه ای از پژوهش های انجام شده قبلی در ارتباط با موضوع بیان شده است. در بخش ۲ روش های پیشنهادی توضیح داده می شود. بخش ۳ شامل ارزیابی مدل -

<sup>2</sup> Assembly Learning Algorithm

<sup>3</sup> Boost Algorithm

های مختلف ارائه شده برای پیش‌بینی رویگردانی مشتری است. در بخش پایانی نیز نتیجه‌گیری نهایی انجام شده و پیشنهاداتی برای کارهای آتی ارائه می‌شود.

## ۲- روش‌شناسی پژوهش

با توجه به وسعت فراوانی دیتای موجود در پایگاه داده شرکت مخابرات، که در مقیاس داده‌های بزرگ<sup>۴</sup> است و لزوم پردازش و آنالیز آن برای بهبود توانایی‌های شرکت در راه رسیدن به اهداف مد نظر، نیاز به استفاده از تکنولوژی‌های نوین مبتنی بر هوش مصنوعی<sup>۵</sup> (AI) و علوم داده<sup>۶</sup> جهت پردازش دقیق و سریعتر نسبت به روشهای کلاسیک وجود دارد. با در نظر گرفتن موارد مذکور، با استفاده از متدهای داده کاوی<sup>۷</sup> در کنار دانش یادگیری ماشین<sup>۸</sup> (ML)، در پی حل مسائل و رسیدن به اهداف مورد نظر خواهیم بود.

در ابتدا داده‌های موجود بر اساس معیارهای مشخص و انواع روشهای نرمال‌سازی داده<sup>۹</sup> مورد ارزیابی و پیش پردازش قرار گرفته تا داده‌های قابل استفاده و استاندارد داشته باشیم. سپس مشتریان بر اساس مشتریان خوش‌حساب، بدحساب؛ مدت تماس و دوره‌های استفاده از اینترنت و... در یک دوره زمانی مشخص طبقه‌بندی و سطح وفاداری آنها مشخص می‌شود. طبقه‌بندی سطح وفاداری مشتریان وسیله مهمی برای انجام تجزیه و تحلیل مشتری است، که مشتریان را در چندین گروه مختلف با توجه به طرح‌های مختلف هدف قرار می‌دهد. در میان تمام مشتریانی که در شرف رویگردانی هستند، در عمل، اپراتورهای مخابراتی قرار نیست اقدامات یکسانی را برای همه مشتریان انجام دهند. از آنجایی که همه مشتریان دارای ارزش بالا نیستند، اپراتورهای مخابراتی باید منابع نگهداری بیشتری را برای مشتریان با ارزش‌تر هزینه کنند. از سوی دیگر، برای برخی از مشتریانی که برای درآمد شرکت چندان مفید نیستند، اپراتورهای مخابراتی مجبور نیستند به آنها توجه زیادی کنند.

در این پژوهش از روش‌های پیش‌بینی مختلفی شامل رگرسیون، جنگل تصادفی، شبکه‌های عصبی مصنوعی و آدبوست برای پیش‌بینی رویگردانی مشتری استفاده می‌نماییم. همچنین از روش‌های مختلف آماری در کنار متدهای مختلف بصری سازی داده برای نمایش خروجی و نتیجه نهایی بصورت جدول و نمودار استفاده می‌شود. در ادامه این بخش به معرفی مجموعه داده‌های مورد استفاده، پیش پردازش‌های لازم روی مجموعه داده شرکت مخابرات استان کردستان و مدل‌های مورد استفاده در پیش‌بینی رویگردانی مشتری می‌پردازیم.

الف) مجموعه داده‌های<sup>۱۰</sup> استفاده شده

این دیتاست مربوط به مشتریان شرکت مخابرات در استان کردستان است. این دیتا بصورت قالب CSV است و شامل اطلاعات ۵۲۹,۰۰۰ مشترک می‌باشد. هدف اصلی از استفاده از این دیتاست، پیش‌بینی رویگردانی مشتریان است که به معنای ترک خدمات شرکت مخابرات و انتقال به رقیب یا ارائه‌دهندگان دیگر است. دیتاست شامل ویژگی‌هایی نظیر کد مشترک، شماره تلفن، عنوان و ... می‌باشد که پس از حذف ویژگی‌های کم اهمیت مانند بارکد، مشخصات سجلی مشتری، آدرس، کدپستی و ... بقیه ویژگی‌های در جدول شماره (۱) قرار گرفته است. با تحلیل این دیتاست، می‌توان الگوها و رفتارهای مشتریان را شناسایی کرده و با استفاده از مدل‌سازی و پیش‌بینی، عواملی که باعث ممکن است مشتریانی که رویگردانی می‌کنند را شناسایی و استراتژی‌های مناسب برای پیش‌بینی رویگردانی مشتریان انتخاب کرد.

جدول شماره (۱) متغیرهای مجموعه داده مخابرات

<sup>4</sup> Big Data

<sup>5</sup> Artificial Intelligence

<sup>6</sup> Data Science

<sup>7</sup> Data Mining

<sup>8</sup> Machine Learning

<sup>9</sup> Data Normalization

<sup>10</sup> DataSet

نام مشخصه	توضیحات	نام مشخصه	توضیحات
Code	کد مشترک	Vsl	وصولی (پرداختی مشترک در دوره)
Telno	شماره تلفن مشترک	ADSL	هزینه زوج سیم ADSL
Title	عنوان (عادی، دولتی، مخابرات)	FTTH	هزینه فیبر نوری FTTH
Stitle	زیر عنوان (شخصی، FTTH، SIP، ...)	VDSL	هزینه زوج سیم VDSL
City	نام شهر	Hamrahaval	هرینه مکالمه همراه اول
Date	تاریخ دایری تلفن	Irancel	هرینه مکالمه با اپراتور ایرانسل
Daramad	جمع درآمد دوره جاری مشترک	Kish	هرینه مکالمه با اپراتور کیش
Hesab	صورتحساب دوره مشترک	Taliya	هرینه مکالمه با اپراتور تالیا
Bed	بدهی قبلی مشترک	Rightel	هرینه مکالمه با اپراتور رایتل
Total	مبلغ قابل پرداخت	Ohamrah	هرینه مکالمه با سایر اپراتورها
Numbed	دفعات بدهی	DLocDetail	کارکرد کل شهری
Internal	کارکرد درون استان	DNatDetail	کارکرد کل بین شهری
Outernal	کارکرد برون استان	Dovrei	هزینه های دوره ای
iInternational	هزینه مکالمه بین الملل	Taki	هزینه های تکی در دوره
Hoshmand	هوشمند کشوری سهم مخابرات	Fcp	هزینه مشترکین جدید اینترنت
Abone	آبونمان ماهانه	Pap	هزینه ماهانه شرکتهای ارایه دهنده خدمات اینترنت
Omormoshtarekin	هزینه ماهانه امور مشترکین		

(ب) پیش پردازش داده‌ها<sup>۱۱</sup>

پیش پردازش داده یکی از مراحل مهم در داده‌کاوی است که به پاکسازی و فیلتر کردن داده‌ها کمک می‌کند. بنابراین، حذف ناسازگاری‌ها و مقادیر تهی یا مقادیر از دست رفته در مجموعه داده و یکپارچه سازی و تبدیل داده‌های خام به اطلاعات و سپس کاهش داده، می‌تواند به طور کارآمد مدیریت شود.

شناسایی مناسب‌ترین داده‌ها از داده‌های خام به منظور ایجاد یک مدل پیش‌بینی رویگردانی مشتری از اهمیت ویژه‌ای برخوردار است. در این پژوهش برای شناسایی متغیر مهم از همبستگی استفاده شده است. متغیرهای کم اهمیت که در پیش بینی رویگردانی مشتریان تأثیری ندارد مانند ستون های شناسه پرداخت، شناسه قبض، بارکد هزینه‌های تکی، افزایش و کاهش دستی و مشخصات سجلی مشترکین به جهت صیانت از حقوق مشترکین، حذف می‌گردد.

مرحله پاکسازی داده‌ها<sup>۱۲</sup> جهت مدیریت مقادیر از دست رفته و تصحیح فرمت داده است که شامل تمیز کردن و فیلتر کردن داده‌ها با حذف مقادیر از دست رفته، پارامترهای غیر مرتبط و اصلاح قالب داده‌های مجموعه داده که به صورت صحیح تشخیص داده نشده، می‌باشد. در این مرحله جهت مدیریت مقادیر از دست رفته با توجه به نوع ارزش فیلدها پر کردن مقادیر از دست رفته به یکی از روش های زیر اقدام می‌کنیم:

- حذف سطر برای ویژگی‌هایی که ستون آنها مهم است مانند تلفن و مبلغ صورتحساب دوره
- نادیده گرفتن برای ستون های کم اهمیت مانند کد مشترک، مالیات، عوارض
- از محتمل ترین مقدار برای پر کردن مقدار از دست رفته استفاده می‌شود: مانند ستون آبونمان که هزینه آن برای کل یک مرکز ثابت است.
- از یک ثابت سراسری برای پر کردن مقدار از دست رفته استفاده می‌شود: برجسیبی مانند "ناشناخته"<sup>۱۳</sup> جایگزین می‌شود.

11 Data Preprocessing

12 Data Cleaning

13 Unknown

• از یک معیار تمایل مرکزی برای ویژگی (مثلاً میانگین یا میانه) مانند تعداد دوره بدهی، برای پر کردن مقدار از دست رفته استفاده می‌شود.

استانداردسازی<sup>۱۴</sup> فرآیند تبدیل داده‌ها به یک قالب رایج است که اجازه می‌دهد تا مقایسه معناداری انجام شود. بدین منظور لازم است که ستون "مبلغ کارکرد درون استانی"، "مبلغ کارکرد برون استانی" و "هزینه موبایل" را به دقیقه تبدیل نماییم. نحوه محاسبه مقادیر استاندارد ستون‌های مذکور از مجموعه داده، در روابط زیر ارائه شده است:

$$(۱) \text{ هر دقیقه مکالمه درون استانی برابر است با } ۴۳ \text{ ریال}$$

$$\text{Internal\_min} = \text{Internal}/43$$

$$(۲) \text{ هر دقیقه مکالمه برون استانی برابر است با } ۳۳۰ \text{ ریال}$$

$$\text{outernal\_min} = \text{outernal}/330$$

$$(۳) \text{ هر دقیقه مکالمه هر کدام از اپراتورهای همراه برابر است با } ۶۲۰ \text{ ریال}$$

$$\text{Mobile\_min} = (\text{HamrahAval} + \text{Irancell} + \text{Kish} + \text{Taliya} + \text{Rightel} + \text{Ohamrah})/620$$

فرآیند نرمال‌سازی<sup>۱۵</sup> داده‌ها شامل تبدیل مقادیر متغیرهای مرتبط به یک محدوده مشابه است. نرمال‌سازی‌های معمولی شامل مقیاس بندی متغیر به طوری که مقادیر متغیر از ۰ تا ۱ باشد (Alpaydin, E. (2020)). جهت نرمال‌سازی از روش باینینگ<sup>۱۶</sup> استفاده می‌کنیم، بدین صورت که یک مقدار داده مرتب شده را با مراجعه به "همسایگی" آن، یعنی مقادیر اطراف آن، هموار می‌نماییم. در واقع فرآیند تبدیل متغیرهای عددی پیوسته به دسته "bin" های گسسته برای تجزیه و تحلیل گروهی است. به عنوان مثال در این مقاله برای رسم نمودار رگرسیون رابطه بین ستون تعداد دوره بدهی و درآمد، باید ستون تعداد دوره بدهی را بین ۰ و ۱ نرمال نماییم. لذا، این مورد با استفاده از رابطه زیر انجام شده است.

(۴)

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

در مجموعه داده ها، "تعداد دوره بدهی" یک متغیر با ارزش واقعی است که از ۰ تا ۶۰ متغیر است و دارای ۶۰ مقدار منحصر به فرد است. اگر فقط به تفاوت "مبلغ بدهی" با تعداد دوره بالا، متوسط و کم (۳ نوع) اهمیت دهیم، می‌توانیم آن‌ها را در سه دسته برای ساده‌تر کردن تحلیل مرتب کنیم.

ج) متغیرهای طبقه بندی شده - متغیرهای شاخص

متغیر نشانگر (متغیر شاخص) یک متغیر عددی است که برای برچسب‌گذاری دسته‌ها استفاده می‌شود. خود اعداد معنای ذاتی ندارند اما از متغیرهای شاخص استفاده می‌کنیم تا بتوانیم از متغیرهای طبقه بندی برای تحلیل رگرسیون در مازول های بعدی استفاده کنیم.

• ستون "نوع مشترک" دارای ۷ مقدار منحصر به فرد است: "عادی"، "دولتی"، "خاص"، "همگانی"، "مخابرات"، "سیم خصوصی" و "سایر".

• ستون "رویگردانی" دارای ۲ مقدار منحصر به فرد است: "رویگردانی"، "غیر رویگردانی" جهت جهت تحلیل رگرسیون بر اساس ۰ و ۱ دسته‌بندی می‌شود.

از آنجا که رگرسیون مقادیر غیر عددی (کلمات) را نمی‌فهمد، برای استفاده از این ویژگی در تحلیل رگرسیون، «نوع مشترک» را به متغیرهای شاخص تبدیل می‌کنیم.

<sup>14</sup>Standardization

<sup>15</sup> Normalization

<sup>16</sup> Binning

- ستون "صورت حساب" را بر اساس مبلغ کارکرد به سه دسته "مشترکین با کارکرد بالا"، "مشترکین با کارکرد متوسط"، "مشترکین با کارکرد کم" را جهت ارزش گذاری مشتریان دسته بندی می کنیم.
- (د) شناسایی عوامل تاثیرگذار در رویگردانی مشتری
- متغیرهای تاثیرگذار در رویگردانی مشتری بر اساس تحلیل همبستگی در جدول شماره (۲) در دو دسته عددی پیوسته و طبقه بندی شده به ترتیب اهمیت آمده است:

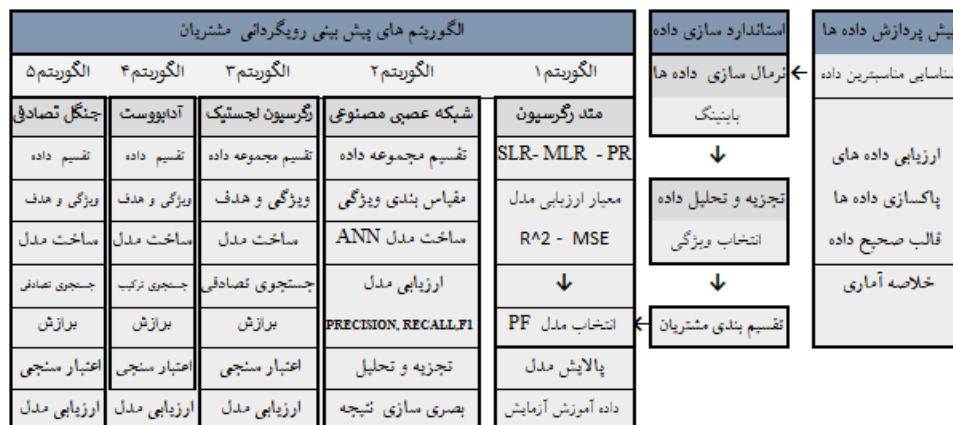
جدول شماره (۲): متغیرهای تاثیرگذار در رویگردانی مشتری

متغیرهای عددی پیوسته		متغیرهای طبقه بندی شده	
ترتیب اهمیت	ویژگی	ترتیب اهمیت	ویژگی
1	تعداد دوره بدهی	1	نوع مشترک
2	بدهی	2	نام شهر
3	درآمد	3	رویگردان
4	وصولی		
5	کارکرد موبایل		
6	سرویس اینترنت		

پیش بینی کننده های خوب رویگردانی عبارتند از: تعداد دوره بدهی، بدهی، صورتحساب مشتری، کارکرد شهری، کارکرد بین شهری، کارکرد موبایل، وصولی. تحلیل همبستگی جهت تجزیه و تحلیل تاثیر هر ویژگی بر ویژگی هدف (رویگردان) استفاده می شود تا متغیرهای تاثیرگذار در رویگردانی مشتری را شناسایی و یک مدل با استفاده از این متغیرها به عنوان متغیرهای پیش بینی توسعه دهیم.

(ر) پیش بینی رویگردانی مشتری

مشتریان را بر اساس سطح وفاداری (مشتریان در معرض رویگردانی و غیررویگردانی) و درآمد (مشتریان با درآمد بالا، متوسط و کم) گروه بندی می نماییم و مشترکین در شرف رویگردانی را با استفاده از پنج الگوریتم پیشنهادی شکل شماره (۱) پیش بینی می نماییم.



شکل شماره (۱): مراحل روش پیشنهادی

روش اول: پیش بینی رویگردانی مشترکین با استفاده از متد رگرسیون

در تجزیه و تحلیل داده ها، اغلب از توسعه مدل استفاده می کنیم تا به ما کمک کند مشاهدات آینده را از روی داده هایی که در اختیار داریم پیش بینی کنیم. یک مدل به ما کمک می کند تا رابطه دقیق بین متغیرهای مختلف و نحوه استفاده از این متغیرها برای پیش بینی نتیجه را درک کنیم. با استفاده از متغیرها یا ویژگی ها، رویگردانی مشتری را پیش بینی می کنیم.

برای استفاده از متغیرهای بیشتری در مدل خود برای پیش‌بینی رویگردانی مشتری از رگرسیون خطی چندگانه استفاده می‌کنیم. رگرسیون خطی چندگانه بسیار شبیه به رگرسیون خطی ساده است، اما از این روش برای توضیح رابطه بین یک متغیر پاسخ پیوسته (وابسته) و دو یا چند متغیر پیش‌بینی کننده (مستقل) استفاده می‌شود. اکثر مدل‌های رگرسیون دنیای واقعی شامل پیش‌بینی‌کننده‌های متعدد هستند. معادله به صورت زیر به دست می‌آید:

(۵)

$$Y_{hat} = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + \dots$$

$$Rizesh = 9.08 - 7.2 * hesab + 3.8 * mobile - 2.4 * internal - 4.4 * outernal$$

در اینجا، متغیر پاسخ (Y) به عنوان متغیری است که تحت تأثیر متغیرهای پیش‌بینی‌کننده (X) قرار می‌گیرد. عبارت a به خط رگرسیون اشاره دارد و به معنای مقدار Y در زمانی که X برابر با صفر است است. همچنین، عبارت b به شیب خط رگرسیون اشاره دارد و به معنای تغییری است که Y با افزایش یک واحدی X تجربه می‌کند.

رگرسیون چند جمله‌ای<sup>۱۷</sup> و پیپ لاین<sup>۱۸</sup> (خطوط لوله)

رگرسیون چند جمله‌ای روابط غیر خطی را با مربع کردن یا تنظیم شرایط مرتبه بالاتر متغیرهای پیش‌بینی کننده بدست می‌آوریم.

(۶)

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \dots$$

$$Z = ['hesab1', 'mobile1', 'numbed', 'bed1', 'vs11']$$

$$y = ['rizesh']$$

روش دوم: پیش‌بینی رویگردانی مشترکین با استفاده از الگوریتم شبکه‌های عصبی مصنوعی (ANN) مراحل پیش‌بینی رویگردانی مشترکین با استفاده از الگوریتم شبکه‌های عصبی مصنوعی عبارتند از خواندن مجموعه داده مخابرات و حذف ستون‌های غیر ضروری و تقسیم ویژگی‌ها و هدف برای آموزش. ویژگی‌ها شامل ۴۱ ستون می‌شوند که شامل ستون‌های تلفن، درآمد، صورت‌حساب، دوره‌های بدهی، وصولی و... است. هدف به نام "رویگردانی" است که نوع رویگردانی مشتری را با ۰ (غیر رویگردانی) و ۱ (رویگردانی) مقداردهی می‌کند. سپس ستون‌های جدیدی برای هر دسته (کم، بالا، متوسط) در ویژگی "دوره‌های بدهی" مشترکین ایجاد می‌شود. همچنین ویژگی‌های "شهرستان" و "دوره بدهی" که دسته‌بندی شده‌اند، برچسب‌گذاری می‌شوند و ستون‌های نوع رشته‌ای حذف می‌شوند. سپس مجموعه داده به داده‌های آموزش و آزمایش تقسیم می‌شود، با ۲۰٪ از داده‌ها به عنوان نمونه آزمایشی و بقیه به عنوان نمونه آموزشی. مقیاس بندی ویژگی‌ها نیز برای همه ویژگی‌های مجموعه داده آموزشی و آزمایشی اعمال می‌شود، اما به منظور جلوگیری از نشت اطلاعات، مقیاس بندی فقط بر روی مجموعه داده آموزشی انجام می‌شود. سپس یک متغیر به عنوان نمونه‌ای برای نمایش مدل شبکه‌های عصبی مصنوعی ایجاد می‌شود. در ادامه، یک پرسپترون چند لایه ساده با ۱ لایه ورودی، ۲ لایه پنهان و ۱ لایه خروجی ساخته می‌شود و مدل شبکه عصبی (ANN) کامپایل می‌شود با بهینه‌ساز adam، زبان binary\_crossentropy و معیار ارزیابی accuracy برای مجموعه داده آموزشی، مدل را به مدت ۱۰۰ دوره اجرا می‌کنیم و سپس نتایج مجموعه آزمایش را پیش‌بینی می‌کنیم و از Confusion برای مشاهده نتایج پیش‌بینی شده استفاده می‌کنیم.

<sup>17</sup> Polynomial

<sup>18</sup> Pipelines



روش سوم: پیش بینی رویگردانی مشترکین با استفاده از الگوریتم رگرسیون لجستیک (LR)<sup>۱۹</sup> پس از خواندن مجموعه داده و تقسیم ویژگی ها و هدف، یک مدل رگرسیون لجستیک ایجاد می شود. سپس مجموعه ای از فرآپارامترها برای مدل تعریف می شود و با استفاده از جستجوی تصادفی روی فرآپارامترها، بهترین مجموعه پارامترها را پیدا می کنیم. در مدل رگرسیون لجستیک، با استفاده از پارامترهای ثابتی مانند 'l2=penalty' ، 'C=6' و 'class\_weight='balanced' ، مدل را ایجاد می کنیم. در نهایت، میانگین دقت، میانگین امتیاز F1 و میانگین امتیاز AUC-ROC برای الگوریتم لجستیک محاسبه می شود.

روش چهارم: پیش بینی رویگردانی مشترکین با استفاده از الگوریتم آدا بوست<sup>۲۰</sup> مراحل پیش بینی رویگردانی مشترکین با استفاده از الگوریتم آدا بوست به شرح زیر است: ابتدا شی را با تنظیم `random_state=42` مقداردهی می کنیم. سپس یک شبکه از مقادیر فرآپارامترها را برای جستجو تعیین می کنیم که دو کلید دارد و برای هر کلید مقادیر متفاوتی برای جستجو به عنوان ترکیب بهینه وجود دارد. از روش `RandomizedSearchCV` برای تنظیم های پارامتر استفاده می کنیم. شبکه های پارامتر تعریف شده در `params_AB` و پارامترهای اعتبارسنجی متقاطع (`cv=5`) و (`random_state=42`) را استفاده می کند. همچنین `n_jobs=-1` را برای محاسبات موازی و `n_iter=30` را برای تعیین تعداد تکرارهای نمونه گیری تصادفی تنظیم می کند. سپس برازش برای آموزش طبقه بندی کننده روی داده های آموزشی استفاده می شود. در نهایت، ارزیابی با محاسبه میانگین دقت، امتیاز F1 و امتیاز AUC برای طبقه بندی کننده آدا بوست انجام می شود.

روش پنجم: پیش بینی رویگردانی مشترکین با استفاده از الگوریتم جنگل تصادفی<sup>۲۱</sup> جنگل تصادفی است که یک تکنیک یادگیری گروهی است که درختان تصمیم گیری زیادی را در بر می گیرد و یک جنگل می سازد. برای طبقه بندی نمونه ها، هر درخت در جنگل یک کلاس را پیش بینی می کند، و طبقه بندی نهایی بر روی کلاسی که بیشترین رای را دارد، پیش بینی می شود. در انجام این کار، RF از برازش بیش از حد در هنگام استفاده از درختان تصمیم گیری منفرد جلوگیری می کند (Kratsch, W., 2021).

شی طبقه بندی کننده جنگل تصادفی با `random_state=42` تنظیم می کنیم، سپس یک شی `RandomizedSearchCV "grid_RF"` جستجوی تصادفی را روی فرآپارامترهای مشخص شده انجام می شود. و تعداد فولدها برای اعتبارسنجی متقاطع بر روی ۵ تنظیم شده است. آرگومان `"return_train_score"` روی True تنظیم شده است که نمرات (امتیاز) آموزشی را همراه با امتیازهای اعتبارسنجی متقاطع برمی گرداند. در نهایت برای ارزیابی تعداد هسته های CPU برای محاسبات روی ۱- به این معنی که از تمام هسته های موجود استفاده شود. تعداد تکرارها برای جستجوی تصادفی ۲۰ تنظیم شده است. محاسبه میانگین دقت، امتیاز F1 و مساحت زیر منحنی مشخصه عملیاتی گیرنده (AUC-ROC) برای طبقه بندی جنگل تصادفی با استفاده از استراتژی اعتبارسنجی متقابل می باشد و در نهایت میانگین دقت، میانگین امتیاز F1 و میانگین امتیاز AUC-ROC برای طبقه بندی جنگل تصادفی محاسبه می شود.

### ۳- نتایج و بحث

دیتای مورد نظر با استفاده از زبان برنامه نویسی پایتون مورد تجزیه و تحلیل قرار می گیرد. همچنین از روش های مختلف آماری در کنار متدهای مختلف بصری سازی دیتا برای نمایش خروجی و نتیجه نهایی بصورت جدول و نمودار استفاده می شود.

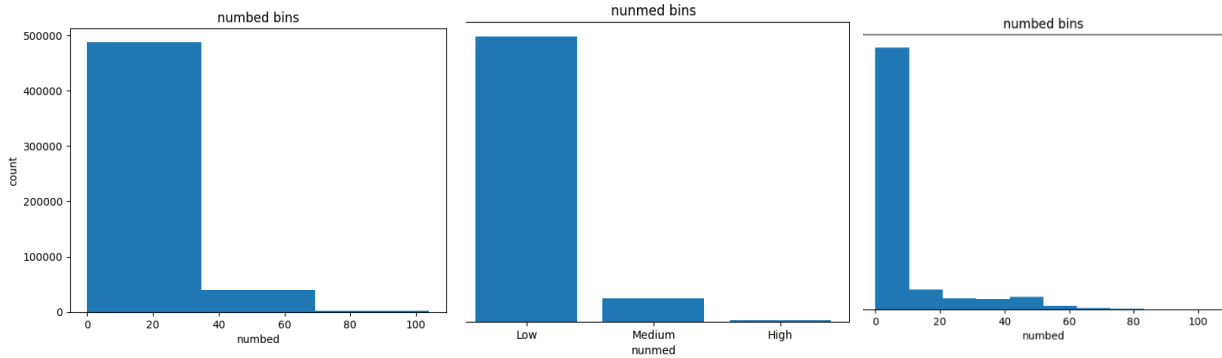
<sup>19</sup> Logistic Regression

<sup>20</sup> Ada Boost

<sup>21</sup> Random Forest

الف) نرمال سازی داده‌ها (باینینگ)

هیستوگرام یک روش مشترک در تصویرسازی توزیع دسته‌ها است که از آن برای نشان دادن تعداد و فراوانی داده‌ها در هر دسته استفاده می‌شود.



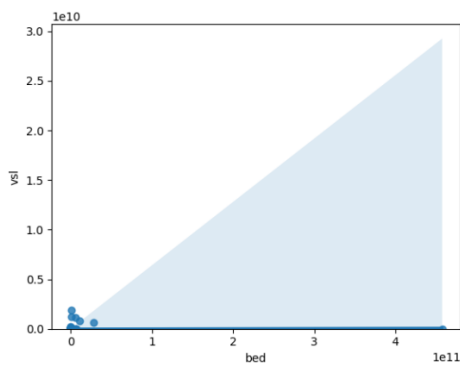
ج نرمال‌سازی (Bin)

ب- نرمال سازی (دسته‌بندی)

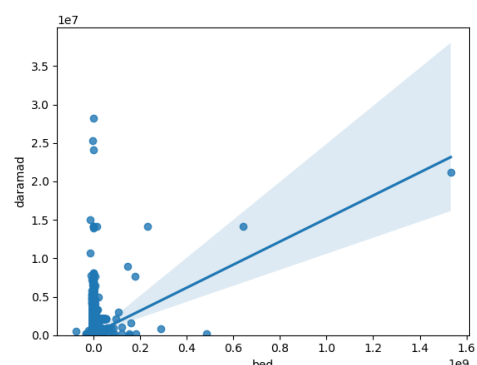
الف - قبل از نرمال سازی

نمودار شماره (۱): نرمال سازی ستون «تعداد دوره بدهی» با روش دسته‌بندی، باینینگ

ب) تجزیه و تحلیل داده‌ها (انتخاب ویژگی) و همبستگی با استفاده از نمودار رگرسیون در این مقاله، تمرکز بر تحلیل داده‌ها و انتخاب ویژگی‌های مناسب از میان متغیرها قرار دارد. به منظور انجام این فرآیند، دو روش اصلی همبستگی و رگرسیون مورد استفاده قرار می‌گیرند. با استفاده از این دو روش، تحلیل داده‌ها انجام می‌شود و ویژگی‌های مهم در نظر گرفته می‌شوند. این روش‌ها مناسب برای تحلیل داده‌های بزرگ هستند و به محققان این امکان را می‌دهند تا ویژگی‌های مهم را شناسایی کرده و در تحلیل‌های خود استفاده کنند. در ادامه مقاله، با استفاده از ترسیم نمودارهای رگرسیون، روابط بین برخی از متغیرها را مشاهده می‌کنیم. به عنوان مثال، نمودار (۲) نشان می‌دهد که بین ستون بدهی و درآمد یک همبستگی بسیار قوی با مقدار  $+0.99$  وجود دارد و متغیر "درآمد" به عنوان یک پیش‌بینی کننده برای بدهی عمل می‌کند. همچنین، نمودار (۳) نشان می‌دهد که همبستگی بین ستون بدهی و وصولی بسیار ضعیف است با مقدار  $+0.03$ ، بنابراین "وصولی" و "بدهی" رابطه معکوسی دارند. این مقاله به محققان و علاقه‌مندان در حوزه تحلیل داده‌ها و انتخاب ویژگی‌های مهم ارائه‌های مفیدی می‌کند.



نمودار شماره (۳): رگرسیون همبستگی بین ستون بدهی و وصولی

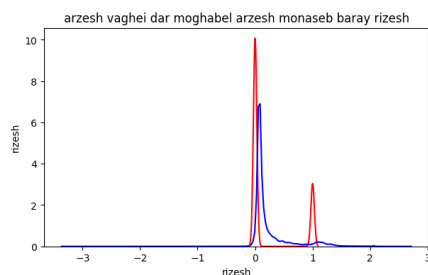


نمودار شماره (۲): رگرسیون همبستگی بین ستون بدهی و درآمد

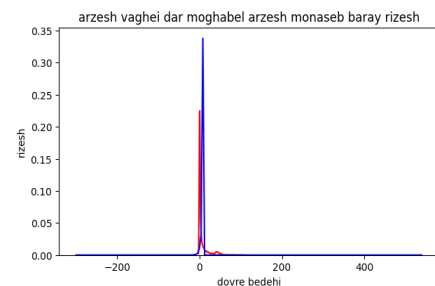
## ج) تحلیل نتایج

## نتایج پیش بینی رویگردانی با الگوریتم رگرسیون

نتایج حاصل از روش رگرسیون خطی چندگانه نشان می‌دهند که مقادیر برازش شده به طور منطقی به مقادیر واقعی نزدیک هستند، به این معنی که مدل پیش‌بینی به خوبی عمل می‌کند. همچنین، تحلیل توزیع مقادیر برازش حاصل از پیش‌بینی رویگردانی مشتریان در دو شاخص، یعنی دوره بدهی و نوع رویگردانی، انجام شده است. نمودار ۱۲ نشان می‌دهد که توزیع مقادیر برازش حاصل از پیش‌بینی رویگردانی مشتریان در دوره بدهی، با توزیع واقعی همپوشانی دارد. همچنین، نمودار ۱۳ نیز نشان می‌دهد که توزیع مقادیر برازش حاصل از پیش‌بینی رویگردانی مشتریان در نوع رویگردانی نیز با توزیع واقعی همپوشانی دارد. این همپوشانی‌ها نشان از صحت و دقت مدل پیش‌بینی در تخمین رویگردانی مشتریان می‌باشد.



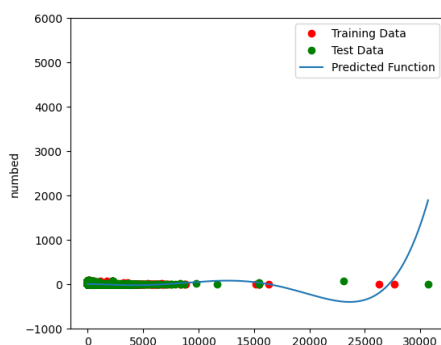
نمودار شماره (۱۳): توزیع مقادیر برازش حاصل از پیش‌بینی رویگردانی (نوع رویگردانی)



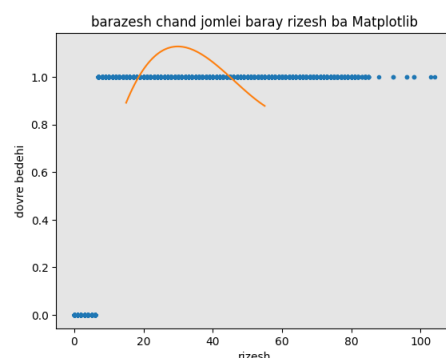
نمودار شماره (۱۲): توزیع مقادیر برازش حاصل از پیش‌بینی رویگردانی (دوره بدهی)

## نتایج رگرسیون چند جمله‌ای و پیم لاین (خطوط لوله)

رگرسیون چندجمله‌ای به ما امکان می‌دهد روابط غیرخطی را با استفاده از ترکیب متغیرهای پیش‌بینی کننده به صورت مربعی یا با تنظیم شرایط مرتبه بالاتر، مدل کنیم. این روش باعث افزایش قدرت پیش‌بینی و توصیف دقیق‌تر رابطه بین متغیرهای وابسته و مستقل می‌شود (Ostertagová, 2012). در نمودار ۱۴، ارتباط متغیر تعداد دوره بدهی و رویگردانی مشتری با استفاده از رگرسیون چندجمله‌ای نمایش داده شده است. همچنین، نمودار ۱۵ نشان دهنده استفاده از یک مدل رگرسیون چندجمله‌ای است، که در آن نقاط قرمز نمایانگر داده‌های آموزشی، نقاط سبز نمایانگر داده‌های آزمون و خط آبی نمایانگر پیش‌بینی مدل است. این روش‌ها و الگوریتم‌ها می‌توانند به شرکت‌ها در تحلیل و پیش‌بینی رویگردانی مشتریان کمک کرده و ارزش افزوده قابل توجهی را به عملکرد و استراتژی‌های بازاریابی آن‌ها اضافه نمایند.



نمودار شماره (۱۵): رگرسیون چند جمله‌ای



نمودار شماره (۱۴): چند جمله‌ای با متغیر تعداد دوره بدهی و رویگردانی مشتری

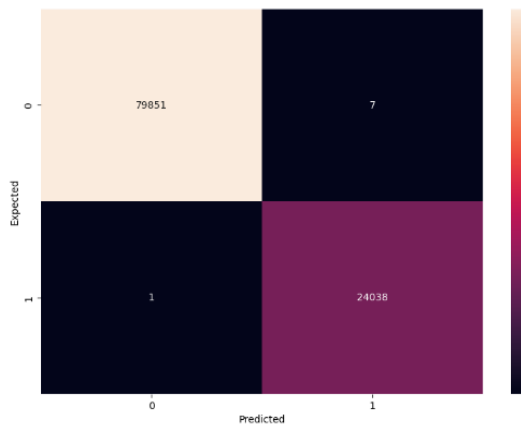
جدول شماره (۵): ارزیابی تعیین دقت مدل های رگرسیون

ارزیابی/مدل	R <sup>2</sup>	MSE
-------------	----------------	-----

مدل: رگرسیون خطی چندگانه (MLR)	٪۶۳	۰/۰۶۶
مدل: رگرسیون چند جمله ای (PR)	٪۸۵	۰/۰۲۵

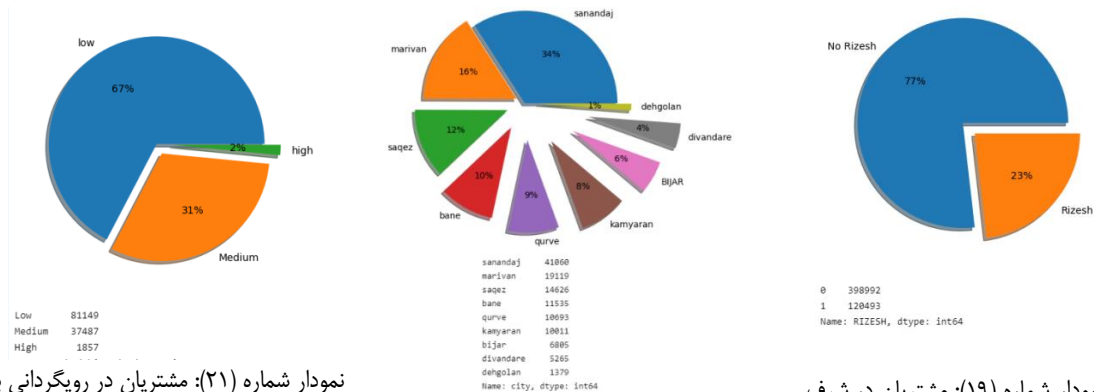
با مقایسه این دو مدل با توجه به جدول شماره (۵)، بر اساس متغیرهای  $R^2$  و MSE نتیجه می‌گیریم که مدل PR (رگرسیون چند جمله ای) از عملکرد بهتری در پیش‌بینی رویگردانی مشتری از روی مجموعه داده مورد استفاده برخوردار است. این نتیجه منطقی است زیرا در مجموع ۲۸ متغیر داریم و می‌دانیم که بیش از یکی از آن متغیرها پیش‌بینی کننده رویگردانی مشتری هستند. همچنین  $R^2$  آن بیشتر و MSE کمتر است.

نتایج پیش‌بینی رویگردانی با استفاده از شبکه‌های عصبی مصنوعی ماتریس درهم‌ریختگی یک ابزار مهم در ارزیابی عملکرد مدل‌های پیش‌بینی است. این ماتریس برای نمایش دقت پیش‌بینی‌ها و میزان درستی طبقه‌بندی‌ها استفاده می‌شود. در واقع، ماتریس درهم‌ریختگی به صورت جدولی نشان می‌دهد که مدل به چه اندازه در تشخیص درست و غلط دسته‌ها عمل کرده است. با بررسی ماتریس درهم‌ریختگی، می‌توانیم معیارهای مهمی مانند دقت (accuracy)، صحت (precision)، تفسیر (recall) و امتیاز F1 را محاسبه کنیم. این معیارها براساس ارقام موجود در ماتریس درهم‌ریختگی محاسبه می‌شوند و به ما کمک می‌کنند تا درک بهتری از کارایی و عملکرد مدل پیش‌بینی داشته باشیم. به طور کلی، ماتریس درهم‌ریختگی به ما اجازه می‌دهد تا ببینیم که مدل ما به درستی درک می‌کند که داده‌های واقعی به کدام دسته تعلق دارند و چه مقدار از پیش‌بینی‌ها صحیح و چه مقدار غلط بوده است.

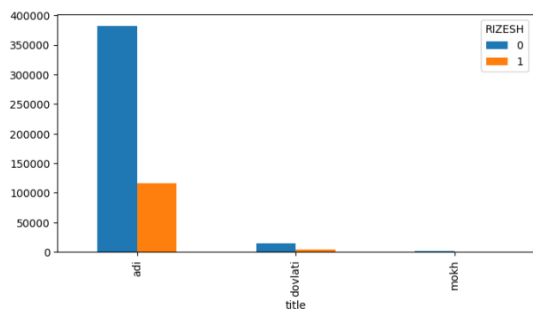


نمودار شماره (۱۸): مقایسه نتایج پیش‌بینی شده با نتایج مورد انتظار

تجزیه و تحلیل تاثیر هر ویژگی بر ویژگی هدف (رویگردانی)

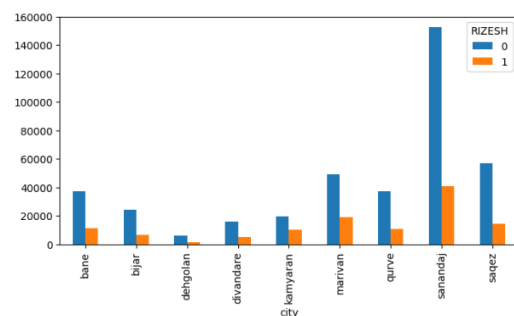


رویگردانی بر اساس شهرستان

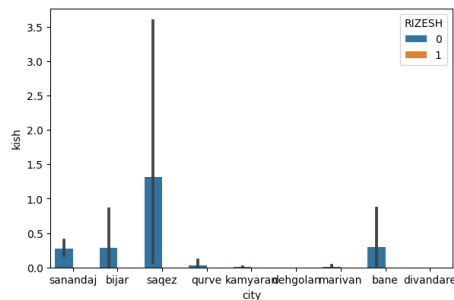
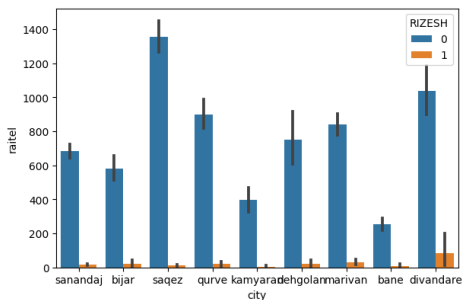
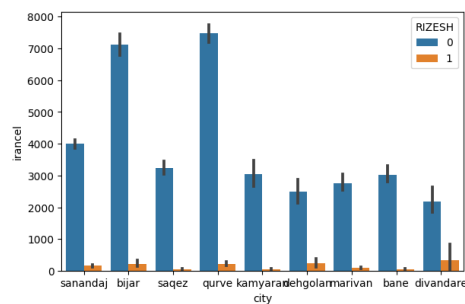
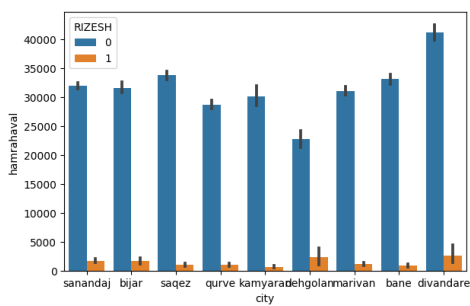


نمودار شماره (۲۳): مشتریارویگردانی و غیر رویگردانی بر اساس نوع مشترک

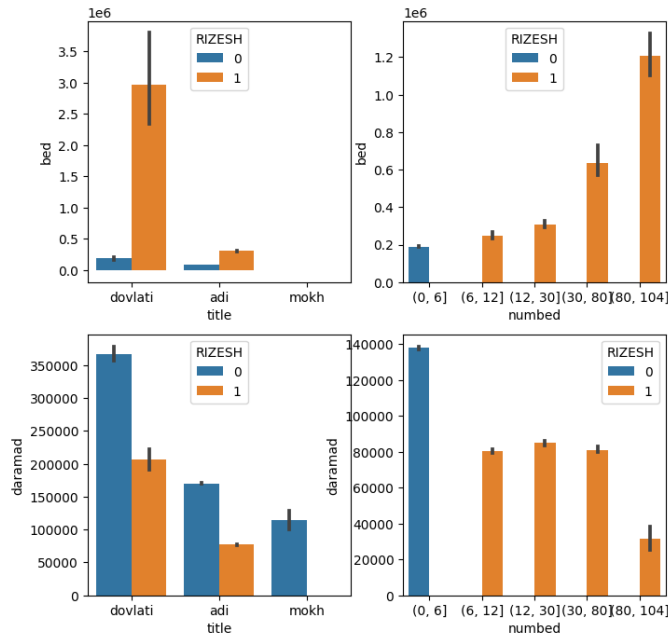
رویگردانی با غیر رویگردانی



نمودار شماره (۲۲): مشتریان رویگردانی و غیر رویگردانی بر اساس شهر



نمودار شماره (۲۴): تجزیه و تحلیل تأثیر ویژگیهای شهر و مکالمه اپراتورهای همراه در برابر ویژگی رویگردانی



نمودار شماره (۲۵): تجزیه و تحلیل تاثیر ویژگی‌های درآمد، بدهی - دوره بدهی، نوع مشترک در برابر ویژگی رویگردانی

نتایج پیش بینی رویگردانی مشترکین با الگوریتم رگرسیون لجستیک (LR)

در این تحلیل، ما مدل Logistic Regression را بررسی می‌کنیم. با بررسی میانگین دقت (Accuracy Mean)، مشاهده می‌شود که این مدل تنها با دقت ۰.۲۳ درصد توانسته است داده‌ها را به درستی دسته‌بندی کند. امتیاز F1 (Score Mean) نیز با مقدار ۰.۳۷ نشان می‌دهد که عملکرد مدل در تفسیر و دقت نیز به صورت کلی پایین است. همچنین، با محاسبه امتیاز AUC (AUC Score Mean) که برای ارزیابی توانایی تفکیک دسته‌ها استفاده می‌شود، می‌بینیم که مدل با امتیاز ۰.۵۹ عملکردی متوسط داشته است. به طور کلی، نتایج نشان می‌دهد که مدل Logistic Regression در پیش‌بینی و دسته‌بندی دارای عملکرد ضعیفی بوده است.

نتایج پیش بینی رویگردانی مشترکین با الگوریتم آدا بوست

در این مدل، نتایج عملکرد مدل AdaBoost را بررسی می‌کنیم. براساس میانگین دقت (Accuracy Mean) این مدل توانسته است با دقت ۱۰۰ درصد، داده‌ها را به درستی دسته‌بندی کند. همچنین، با محاسبه میانگین امتیاز F1 (Score Mean)، می‌توانیم ببینیم که عملکرد مدل در مورد دقت و تفسیر (Recall) برابر با ۱ است. همچنین، با محاسبه میانگین امتیاز AUC (AUC Score Mean) می‌توانیم عملکرد مدل در مورد توانایی تفکیک بین دسته‌ها را بررسی کنیم که در اینجا نیز برابر با ۱ است. این نتایج نشان می‌دهند که مدل AdaBoost با دقت و عملکرد بسیار بالا، در پیش‌بینی و دسته‌بندی عالی عمل کرده است.

نتایج پیش بینی رویگردانی مشترکین با الگوریتم جنگل تصادفی

مدل RandomForest در این مطلب مورد بررسی قرار می‌گیرد. میانگین دقت (Accuracy Mean) این مدل برابر با ۱.۰ است، که نشان می‌دهد که مدل توانسته است با دقت ۱۰۰ درصد، داده‌ها را به درستی دسته‌بندی کند. همچنین، با محاسبه میانگین امتیاز F1 (F1 Score Mean) می‌توانیم ببینیم که مدل در مورد دقت و تفسیر (Recall) نیز به امتیاز ۱.۰ رسیده است. همچنین، با محاسبه میانگین امتیاز AUC (AUC Score Mean) می‌توانیم عملکرد مدل در توانایی تفکیک بین دسته‌ها را بررسی کنیم، که در اینجا نیز به امتیاز ۱.۰ دست یافته است. این نتایج نشان می‌دهند که مدل RandomForest با دقت و عملکرد بسیار بالا، عملکرد عالی در پیش‌بینی و دسته‌بندی دارد.

جدول شماره (۶): نتایج ارزیابی کلیه الگوریتم‌ها

Model	R <sup>2</sup>	MSE	Accuracy/Precision	F1 Score	AUC/Recall	Description
1 Regression(MLR)	63%	0.066				
2 Regression(PR)	85%	0.025				
3 ANN			99.9%	99.9%	99.9%	بهترین
4 Logistic Regression			23%	37%	59%	
5 Ada Boost			100%	100%	100%	بهترین
6 Random Forest			100%	100%	100%	بهترین

با عنایت به اینکه الگوریتم شبکه‌های عصبی مصنوعی و با دقت ۹۹.۹٪، الگوریتم آدابووست با دقت ۱۰۰٪ و الگوریتم جنگل تصادفی با دقت ۱۰۰٪ رویگردانی مشتری را شناسایی نمودند و با موفقیت ارزیابی شدند در نتیجه مشخصات و ویژگی های مشتریان با ریسک رویگردانی بالا شامل مشتریان با تعداد دوره و مبلغ بدهی زیاد، بر اساس دوره بدهی (مشترکین عادی)، بر اساس مبلغ بدهی (مشترکین نوع دولتی) می باشد.

از آنجایی که مشتریان در بازار مخابراتی همیشه تمایل به اشیاع دارند، برای اپراتورها سودمندتر است که برای مشتریانی که در شرف رویگردانی هستند، استراتژی‌های حفظ پیشنهاد دهند.

در این مقاله یک سیستم پشتیبانی تصمیم مبتنی بر داده‌کاوی و با استفاده از الگوریتم یادگیری ماشین جهت تجزیه و تحلیل مشتری یکپارچه و شناسایی عوامل تاثیر گذار در رویگردانی مشتریان پیشنهاد شده است. ابتدا، پاکسازی داده‌ها، تبدیل داده‌ها و عادی سازی داده‌ها در فرآیند پیش پردازش انجام می‌شود. به دنبال آن مرحله تجزیه و تحلیل داده‌های اکتشافی و ویژگی‌های موثر تعیین شد و سپس تقسیم بندی مشتریان در صنعت مخابرات با استفاده از رگرسیون مشتریان به گروه‌های مختلفی تقسیم می‌شوند، که به بازاریابان و تصمیم‌گیرندگان اجازه می‌دهد تا استراتژی‌های حفظ را با دقت بیشتری اتخاذ کنند. تقسیم بندی بر اساس مشتریان با درآمد بالا، متوسط و کم و همچنین بر اساس مشتریان در شرف رویگردانی و غیر رویگردانی جهت تحلیل رگرسیون بر اساس مقدار ۱۰۰ تقسیم بندی نمودیم و جهت پیش بینی و تخمین رویگردانی مشتریان پنج الگوریتم پیشنهاد داده شد.

الگوریتم اول: با استفاده از رگرسیون دو مدل رگرسیون خطی چندگانه (MLR)، رگرسیون چند جمله ای (PR) را ارایه نمودیم و هر دو مدل را بر روی داده‌ها پیاده سازی کردیم و بر اساس معیارهای آماری مقدار مربع R، مقدار میانگین مربعات خطا جهت تعیین دقت ارزیابی نمودیم که با مقایسه این سه مدل، مشخص شد که مدل PR با ۸۵٪ بیشترین دقت و با ۰/۰۲۵ کمترین میانگین مربعات خطا را دارد و بهترین مدل است که بتوان رویگردانی مشتری را از روی مجموعه داده ما پیش بینی نمود. الگوریتم دوم: با استفاده از شبکه‌های عصبی مصنوعی و با دقت ۹۹.۹٪ یکی از بهترین الگوریتمها برای پیش بینی رویگردانی مشتریان می‌باشد. الگوریتم سوم: با استفاده از الگوریتم رگرسیون لجستیک دارای دقت ۲۳٪ می‌باشد. الگوریتم چهارم: با استفاده از الگوریتم رگرسیون آدابووست با دقت ۱۰۰٪ یکی از بهترین الگوریتمها برای پیش بینی رویگردانی مشتریان می‌باشد. الگوریتم پنجم: با استفاده از الگوریتم جنگل تصادفی با دقت ۱۰۰٪ یکی از بهترین الگوریتمها برای پیش بینی رویگردانی مشتریان می‌باشد.

(ب) محدودیت‌ها

برای عملکرد، "دقت" معیار خوبی است، اما اندازه گیری عملکرد فقط با "دقت" کافی نیست زیرا در مجموعه داده‌های کوچک دقت قابل پیش بینی تر است و یکسان خواهد بود. در اختیار نبودن مجموعه داده موبایل و به جهت صیانت از حقوق مشترکین و مسائل امنیتی و حجم زیاد مجموعه داده، از مجموعه داده کامل ریز مکالمات در این پژوهش از آن استفاده نشده است

در این مقاله از مجموعه داده مخابرات استان کردستان استفاده شده است که در کارهای آتی می‌توان برای روی داده کامل مخابرات کل کشور و ریز مکالمات مشتریان تلفن ثابت و همراه اجرا نمود. همچنین می‌توان از الگوریتم یادگیری عمیق CNN استفاده کرد که خود توانایی استخراج ویژگی را دارد و خود را به عنوان یک تکنیک قدرتمند برای مدل‌های برگرداندن، به‌ویژه برای مجموعه داده‌های بزرگ، تثبیت می‌کند.

#### ۴-منابع

1. Dhote, S., Vichoray, C., Pais, R., Baskar, S., & Mohamed Shakeel, P. (2020). Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in E-commerce. *Electronic Commerce Research*, 20, 259-274.
2. Hammoudeh, A., Fraihat, M., & Almomani, M. (2019). Selective ensemble model for telecom churn prediction. 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT),
3. Jafari-Marandi, R., Denton, J., Idris, A., Smith, B. K., & Keramati, A. (2020). Optimum profit-driven churn decision making: innovative artificial neural networks in telecom industry. *Neural Computing and Applications*, 32, 14929-14962.
4. Lemmens, A., & Gupta, S. (2020). *Managing churn to maximize profits*. Marketing Science, 39(5), 956-973.
5. Rogić, S., Kaščelan, L., Kaščelan, V., & Đurišić, V. (2022). Automatic customer targeting: a data mining solution to the problem of asymmetric profitability distribution. *Information Technology and Management*, 23(4), 315-333.
6. Sivasankar, E., & Vijaya, J. (2019). A study of feature selection techniques for predicting customer retention in telecommunication sector. *International Journal of Business Information Systems*, 31(1), 1-26.
7. Sohaib, O., Naderpour, M., Hussain, W., & Martinez, L. (2019). Cloud computing model selection for e-commerce enterprises using a new 2-tuple fuzzy linguistic decision-making method. *Computers & Industrial Engineering*, 132, 47-58.
8. Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems*, 212, 106586.
9. Wu, Z., Jing, L., Wu, B., & Jin, L. (2022). A PCA-AdaBoost model for E-commerce customer churn prediction. *Annals of Operations Research*, 1-18.
10. Kratsch, W., Manderscheid, J., Röglinger, M., & Seyfried, J. (2021). Machine learning in business process monitoring: a comparison of deep learning and classical approaches used for outcome prediction. *Business & Information Systems*
11. Alpaydın, E. (2020). *Introduction to machine learning*. MIT press
12. Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering*, 48, 500-506.
13. Tavakoli, Ahmad; Mortezaei, Saeed; Kahani, Mohsen; Hosseini, Zahra. (2011). Application of Data Mining Process for Customer Churn Prediction in Insurance. *Journal of Business Management Perspective*, 9(4).
14. Najmi, Parvin; Rad, Abbas; Shouar, Maryam. (2018). Customer Churn Prediction in Banks Using Data Mining Techniques. *Journal of Strategic Management in Industrial Systems. Formerly Industrial Management Journal*, 13(44), 99-111.
15. Amiri, Sahar; Hasan Zadeh, Alireza; Sahraei, Shaghayegh. (2022). A Model for Customer Churn Management in an Internet Service Provider Company. *Studies in Intelligent Business Management*, 10(39), 67-95.



## **Identifying the Influencing Factors of Customer Churn of Kurdistan Telecommunications Company and Presenting a Model for Predicting Churn Using Machine Algorithms**

**Vida Sadeghi**

Master's student in Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran - Kurdistan province telecommunication employee

**Email: Sadeghy2008@gmail.com**

**Anwar Bahrampour** (Corresponding Author)

Assistant Professor, Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran

**Email: Bahrampour@iausdj.ac.ir**

**Seyed Ali Hosseini**

Lecturer, Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran

### **Abstract**

The main sources of income and assets are important for any organization. With this in mind, companies have begun to make more efforts to maintain their financial health. Given that the cost of acquiring a new customer is often much higher than ensuring customer satisfaction, customer churn has become a primary area of concern for many companies. Client-facing companies, particularly those in the technology industry, are encountering significant challenges due to customer attrition. As the telecommunications industry rapidly evolves, predicting customer dropout has become a key activity in gaining a competitive edge in the market. Predicting customer churn provides operators with a window of opportunity to remediate issues and implement a series of preventative measures before customers migrate to other service providers. In this research, a decision support system for predicting and estimating customer churn for Kurdistan Telecommunication Company (which has 52,900 subscribers) is presented using various data-mining and machine learning methods, including simple linear regression (SLR), multiple linear regression (MLR), polynomial regression (PR), logistic regression, artificial neural networks, Adabust, and random forest. The results of the evaluations

conducted on the dataset of Kurdistan Province Telecommunication Company demonstrate the high performance of artificial neural network methods with 99.9% accuracy, Adabust with 99.9% accuracy, and random forest with 100% accuracy.

**Keywords:** prediction, customer\_churn, data\_mining, machine\_learning, artificial\_neural\_networks.