https://doi.org/10.1186/s40537-019-0191- 6



**RESEARCH Open Access**

[](http://crossmark.crossref.org/dialog/?doi=10.1186/s40537-019-0191-6&domain=pdf) مشتری تکان دادن پیش‌بینی در مخابرات با استفاده از دستگاه یادگیری در بزرگ داده‌ها پلتفرم

عبدالرحیم کاسم احمد \*[​](http://orcid.org/0000-0002-6980-5267) آصف جعفر و کادان الجمعه

\* مکاتبات: Abdulrahim.ahmad@hiast. (عبدالرحیم احمد@hiast) edu.sy

چکیده

ریزش مشتری (Customer Churn) یک مشکل اساسی و از مهم‌ترین دغدغه‌های شرکت‌های بزرگ است. به‌دلیل تأثیر مستقیم آن بر درآمد شرکت‌ها، به‌ویژه در حوزه‌ی مخابرات، شرکت‌ها به‌دنبال روش‌هایی برای پیش‌بینی مشتریانی هستند که احتمال ریزش آن‌ها وجود دارد. بنابراین، شناسایی عواملی که موجب افزایش ریزش مشتری می‌شوند اهمیت زیادی دارد تا بتوان اقدامات لازم را برای کاهش آن انجام داد.

مهم‌ترین دستاورد این پژوهش، توسعه‌ی یک مدل پیش‌بینی ریزش مشتری است که به اپراتورهای مخابراتی کمک می‌کند تا مشتریانی را که بیشترین احتمال ریزش دارند، شناسایی کنند. مدل ارائه‌شده در این تحقیق از تکنیک‌های یادگیری ماشین در بستر کلان‌داده (Big Data) استفاده می‌کند و رویکرد جدیدی برای مهندسی و انتخاب ویژگی‌ها ارائه می‌دهد.

برای سنجش عملکرد مدل، از معیار استاندارد **مساحت زیر منحنی (AUC)** استفاده شده است که مقدار به‌دست‌آمده برابر **۹۳.۳٪** است. یکی دیگر از نوآوری‌های اصلی این پژوهش، استفاده از **شبکه‌های اجتماعی مشتریان** در مدل پیش‌بینی است که با استخراج ویژگی‌های حاصل از **تحلیل شبکه‌های اجتماعی (SNA)** انجام شده است. استفاده از ویژگی‌های SNA عملکرد مدل را از ۸۴٪ به ۹۳.۳٪ (بر اساس معیار AUC) بهبود داده است.

مدل در محیط **Spark** پیاده‌سازی و آزمایش شده است و بر روی یک مجموعه‌داده‌ی بزرگ که از داده‌های خام ارائه‌شده توسط شرکت مخابرات **SyriaTel** ساخته شده، آموزش دیده است. این مجموعه‌داده شامل اطلاعات تمام مشتریان در بازه‌ی ۹ ماه بوده و برای آموزش، آزمون و ارزیابی سیستم در شرکت SyriaTel مورد استفاده قرار گرفته است.

در این مدل، چهار الگوریتم مورد آزمایش قرار گرفته‌اند:  
**درخت تصمیم (Decision Tree)**، **جنگل تصادفی (Random Forest)**، **درخت تقویت‌شده‌ی گرادیان (GBM)** و **تقویت گرادیان شدید (XGBOOST)**.  
بهترین نتایج با استفاده از الگوریتم **XGBOOST** به‌دست آمده است که در این مدل برای طبقه‌بندی (Classification) به کار رفته است.

**کلیدواژه‌ها:** پیش‌بینی ریزش مشتری، ریزش در مخابرات، یادگیری ماشین، انتخاب ویژگی، طبقه‌بندی، تحلیل شبکه اجتماعی موبایل، کلان‌داده.

دانشکده از فناوری اطلاعات ، بالاتر موسسه برای کاربردی علوم

و فناوری، دمشق سوریه

# ​

مقدمه

بخش مخابرات به یکی از صنایع اصلی در کشورهای توسعه‌یافته تبدیل شده است. پیشرفت فنی و افزایش تعداد اپراتورها، سطح رقابت را افزایش داده است [ [1](#_bookmark33) ]. شرکت‌ها برای بقا در این بازار رقابتی، بسته به استراتژی‌های متعدد، سخت تلاش می‌کنند. سه استراتژی اصلی پیشنهاد شده است. به تولید کردن بیشتر درآمدها [ [2](#_bookmark34) ]: (1) کسب کردن جدید مشتریان، (2) فروش بیشتر ‎‏ ... مشتریان موجود، و (3) افزایش دوره حفظ مشتریان. با این حال، مقایسه این استراتژی‌ها با در نظر گرفتن ارزش بازگشت سرمایه (RoI) هر یک نشان داده است که استراتژی سوم سودآورترین استراتژی است [ [2](#_bookmark34) ]، ثابت می‌کند که حفظ مشتری موجود هزینه بسیار کمتری نسبت به جذب مشتری جدید دارد [ [3](#_bookmark35) ]، علاوه بر این که بسیار آسان‌تر از استراتژی فروش افزایشی در نظر گرفته می‌شود [ [4](#_bookmark36) ]. برای اعمال



© نویسنده(گان) ۲۰۱۹. این مقاله تحت شرایط مجوز بین‌المللی Creative Commons Attribution 4.0 توزیع شده است. ( <http://creativecommons.org/licenses/by/4.0/>)، که استفاده، توزیع و تکثیر نامحدود را در هر رسانه‌ای مجاز می‌داند، ارائه شده شما دادن مناسب اعتبار به ‎‏ ... اصلی نویسنده(گان) و ‎‏ ... منبع، فراهم کردن الف پیوند به ‎‏ ... خلاق عوام مجوز، و نشان دادن اگر تغییرات بودند ساخته شده.

در استراتژی سوم، شرکت‌ها باید پتانسیل رویگردانی مشتری، که به عنوان «جابجایی مشتری از یک ارائه‌دهنده به ارائه‌دهنده دیگر» شناخته می‌شود را کاهش دهند [ [5](#_bookmark37) ].

مشتریان تکان دادن است الف قابل توجه نگرانی در خدمات بخش‌ها با بالا خدمات رقابتی. از سوی دیگر، پیش‌بینی مشتریانی که احتمالاً شرکت را ترک می‌کنند، اگر در ... انجام شود، منبع درآمد اضافی بالقوه بزرگی را نشان خواهد داد. فاز اولیه [ [3](#_bookmark35) ].

بسیاری از تحقیقات تأیید کرده‌اند که فناوری یادگیری ماشینی بسیار کارآمد است پیش‌بینی کردن این وضعیت. این تکنیک است اعمال شده از طریق یادگیری از قبلی داده‌ها [ [6](#_bookmark38) ، [7](#_bookmark39) ].

داده‌های مورد استفاده در این تحقیق شامل تمام اطلاعات مشتریان در طول نه ماه قبل از شروع مطالعه است. حجم این مجموعه داده حدود 70 ترابایت در HDFS "Hadoop" است. توزیع‌شده فایل سیستم"، و دارد متفاوت داده‌ها فرمت‌ها که هستند ساختاریافته، نیمه‌ساختاریافته و بدون ساختار. داده‌ها همچنین بسیار سریع می‌آیند و برای مدیریت آنها به یک پلتفرم کلان‌داده مناسب نیاز است. مجموعه داده‌ها برای استخراج ویژگی‌ها برای هر مشتری تجمیع می‌شوند.

ما شبکه اجتماعی همه مشتریان را ساختیم و ویژگی‌هایی مانند معیارهای مرکزیت درجه، مقادیر شباهت و اتصال شبکه مشتری را برای هر مشتری محاسبه کردیم. ویژگی‌های SNA باعث افزایش خوبی در نتایج AUC شدند و این به دلیل سهم این ویژگی‌ها در ارائه اطلاعات متفاوت‌تر در مورد مشتریان است.

ما بر ارزیابی و تحلیل عملکرد مجموعه‌ای از ماشین‌های مبتنی بر درخت تمرکز کردیم. یادگیری روش‌ها و الگوریتم‌ها برای پیش‌بینی تکان دادن در شرکت‌های مخابراتی. ما داشته باشند آزمایش کرد الف شماره از الگوریتم‌ها چنین به عنوان تصمیم درخت، جنگل تصادفی، درخت ماشین گرادیان بوست و درخت XGBoost برای ساخت مدل پیش‌بینی‌کننده ریزش مشتری پس از توسعه روش‌های آماده‌سازی داده‌ها، مهندسی ویژگی و انتخاب ویژگی.

دو شرکت مخابراتی در سوریه وجود دارند که سیریاتل و MTN هستند. شرکت سیریاتل به این حوزه مطالعاتی علاقه‌مند بود زیرا هزینه جذب یک مشتری جدید شش برابر بیشتر از هزینه حفظ مشتری است که احتمال ریزش مشتری وجود دارد. مجموعه داده‌های ارائه شده توسط سیریاتل چالش‌های زیادی داشت، یکی از آنها چالش عدم تعادل بود، که در آن تکان دادن مشتریان کلاس بود خیلی کوچک مقایسه شده به ‎‏ ... فعال مشتریان کلاس ما سه سناریو را برای مقابله با مشکل عدم تعادل آزمایش کردیم که عبارتند از : نمونه‌گیری بیش از حد، کم‌نمونه‌گیری و بدون متعادل سازی مجدد ارزیابی بود انجام شده با استفاده از منطقه تحت گیرنده عملیاتی مشخصه منحنی «ای یو سی» زیرا آن است عمومی و در مورد مجموعه داده‌های نامتوازن استفاده می‌شود [ [8](#_bookmark40) ].

بسیاری قبلی تلاش‌ها با استفاده از ‎‏ ... داده‌ها انبار سیستم به کاهش ‎‏ ... تکان دادن نرخ در SyriaTel اعمال شد. انبار داده نوعی از داده‌های مخابراتی مانند صورتحساب را جمع‌آوری کرد. داده‌ها، تماس/پیامک/اینترنت، و شکایات داده‌ها معدنکاری تکنیک‌ها بودند اعمال شده روی بالا از ‎‏ ... داده‌ها انبار سیستم، اما ‎‏ ... مدل شکست خورده به دادن بالا نتایج با استفاده از این داده‌ها. در تضاد، ‎‏ ... داده‌ها منابع که هستند عظیم در اندازه بودند نادیده گرفته شده به دلیل به ‎‏ ... پیچیدگی​ در معامله با آنها را. داده‌ها انبار بود نه قادر به کسب کردن، فروشگاه، و فرآیندی که عظیم مبلغ از داده‌ها در ‎‏ ... همان زمان. در علاوه بر این، ‎‏ ... داده‌ها منابع بودند از متفاوت​ انواع، و جمع آوری آنها در داده‌ها انبار بود الف خیلی سخت فرآیند بنابراین که اضافه کردن جدید ویژگی‌ها برای داده‌ها معدنکاری الگوریتم‌ها مورد نیاز الف طولانی زمان، بالا پردازش قدرت، و بیشتر ذخیره‌سازی ظرفیت. روشن ‎‏ ... دیگر دست، همه اینها دشوار فرآیندها در داده‌ها انبار​ هستند انجام شده به راحتی با استفاده از توزیع شده پردازش ارائه شده توسط بزرگ داده‌ها پلتفرم

علاوه بر این، بزرگ اجتماعی شبکه‌ها، به عنوان آن در سوریه تل، هستند در نظر گرفته شده یکی از ‎‏ ... اساسی اجزا از بزرگ داده‌ها شبکه نمودارها [ [9](#_bookmark41) ]. محاسباتی پیچیدگی از SNA اقدامات است خیلی بالا به دلیل به ‎‏ ... طبیعت از ‎‏ ... تکراری محاسبات انجام شده روی الف مقیاس بزرگ نمودار، به عنوان ذکر شده در معادلات ( [1](#_bookmark21) ) و ( [2](#_bookmark22) ). الف زیادی از کار به کاهش ‎‏ ... پیچیدگی محاسبات اس ان ای اقدامات دارد شده انجام شده. برای مثال، بارتلمی [ [10](#_bookmark42) ] پیشنهادی جدید الگوریتم به کاهش دادن ‎‏ ... پیچیدگی از محاسبه ‎‏ ... بینابین بودن مرکزیت از O(n3) به O(n2) الیزابتتا [ [11](#_bookmark43) ] همچنین پیشنهادی یک تقریب روش به محاسبه بینابینی با کمتر پیچیدگی. در کینه توز از که، ‎‏ ... سنتی داده‌ها انبار سیستم هنوز از نقص در محاسبه معیارهای ضروری SNA در مقیاس بزرگ رنج می‌برد. شبکه‌ها.

سیستم کلان‌داده به شرکت سیریاتل این امکان را داد که صرف نظر از حجم، تنوع و پیچیدگی داده‌ها، آنها را به راحتی جمع‌آوری، ذخیره، پردازش و تجمیع کند. علاوه بر این، استخراج داده‌های غنی‌تر را نیز امکان‌پذیر ساخت. و بیشتر متنوع ویژگی‌ها مانند اس ان ای ویژگی‌ها که فراهم کردن اضافی اطلاعاتی برای بهبود مدل پیش‌بینی ریزش مشتری.

ما باور که بزرگ داده‌ها تسهیل شده ‎‏ ... فرآیند از ویژگی مهندسی که است یکی از دشوارترین و پیچیده‌ترین فرآیندها در ساخت مدل‌های پیش‌بینی. با استفاده از پلتفرم کلان‌داده، ما به شرکت سیریاتل قدرت می‌دهیم تا با منابع کلان‌داده فراتر رود. در علاوه بر این، ‎‏ ... شرکت می‌شود قادر به عصاره ‎‏ ... اجتماعی شبکه ویژگی‌های تحلیل از الف بزرگ مقیاس اجتماعی گراف که است ساخته شده از میلیاردها از لبه‌ها (معاملات) که اتصال میلیون‌ها نفر از گره‌ها (مشتریان). سخت‌افزار و ‎‏ ... طراحی از ‎‏ ... بزرگ پلتفرم داده‌ای که در بخش « [روش پیشنهادی ریزش](#_bookmark0) » نشان داده شده است، نیاز به محاسبه این ویژگی‌ها را برآورده می‌کند. صرف نظر از از آنها پیچیدگی روی این بزرگ مقیاس نمودار

این مدل همچنین با استفاده از یک مجموعه داده جدید ارزیابی شد و تأثیر این سیستم بر تصمیم به ترک خدمت مورد آزمایش قرار گرفت. مدل نتایج خوبی ارائه داد و در مرحله تولید به کار گرفته شد.

# مرتبط کار

رویکردهای زیادی برای پیش‌بینی ریزش مشتری در شرکت‌های مخابراتی به کار گرفته شده است. اکثر این رویکردها داشته باشند استفاده شده دستگاه یادگیری و داده‌ها استخراج معدن. اکثریت از مرتبط کار متمرکز روی اعمال کردن فقط یکی روش از داده‌ها معدنکاری به عصاره دانش، و ‎‏ ... دیگران بر مقایسه چندین استراتژی برای پیش‌بینی ریزش مشتری تمرکز کردند.

گاوریل و همکارانش [ [12](#_bookmark44) ] یک روش پیشرفته داده‌کاوی برای پیش‌بینی ریزش مشتریان اعتباری با استفاده از مجموعه داده‌های جزئیات تماس 3333 مشتری با 21 ویژگی و یک پارامتر ریزش وابسته با دو مقدار بله/خیر ارائه دادند. برخی از ویژگی‌ها شامل اطلاعاتی در مورد تعداد پیام‌های ورودی و خروجی و پست صوتی برای هر مشتری است. نویسنده از الگوریتم تحلیل مؤلفه‌های اصلی "PCA" برای کاهش ابعاد داده‌ها استفاده کرد. سه الگوریتم یادگیری ماشین برای پیش‌بینی ضریب ریزش استفاده شد: شبکه‌های عصبی، ماشین بردار پشتیبان و شبکه‌های بیز. نویسنده از AUC برای اندازه‌گیری عملکرد الگوریتم‌ها استفاده کرد. مقادیر AUC به ترتیب 99.10٪، 99.55٪ بود. و ۹۹.۷۰٪ به ترتیب برای شبکه‌های بیز، شبکه‌های عصبی و ماشین بردار پشتیبان. مجموعه داده مورد استفاده در این مطالعه کوچک است و هیچ مقدار گمشده‌ای وجود ندارد.

او و ال. [ [13](#_bookmark45) ] پیشنهادی الف مدل برای پیش‌بینی مبتنی بر روی ‎‏ ... عصبی شبکه الگوریتمی برای حل مشکل ریزش مشتری در یک شرکت بزرگ مخابراتی چینی که حاوی درباره ۵.۲۳ میلیون مشتریان. پیش‌بینی دقت استاندارد بود کلی دقت نرخ، و رسیده ۹۱.۱٪.

ادریس [ [14](#_bookmark46) ] پیشنهادی یک رویکرد مبتنی بر روی ژنتیکی برنامه‌نویسی با AdaBoost برای مدل‌سازی مشکل ریزش مشتری در مخابرات. این مدل روی دو استاندارد آزمایش شد. داده‌ها مجموعه ها. یکی توسط نارنجی مخابرات و ‎‏ ... دیگر توسط سلول به سلول، با ۸۹٪ دقت برای مجموعه داده cell2cell و 63٪ برای مجموعه داده دیگر.

هوانگ و همکاران [ [15](#_bookmark47) ] مشکل ریزش مشتری را در پلتفرم کلان داده بررسی کردند. هدف محققان اثبات این بود که کلان‌داده‌ها فرآیند پیش‌بینی را تا حد زیادی بهبود می‌بخشند. ‎‏ ... تکان دادن بسته به روی ‎‏ ... حجم، تنوع، و سرعت از ‎‏ ... داده‌ها سروکار داشتن با داده‌های بخش پشتیبانی عملیات و بخش پشتیبانی کسب‌وکار در بزرگترین شرکت مخابرات چین، به یک پلتفرم کلان‌داده برای مهندسی شکستگی‌ها نیاز داشت. الگوریتم جنگل تصادفی با استفاده از AUC مورد استفاده قرار گرفت و ارزیابی شد. مختار و ال. [ [16](#_bookmark48) ] پیشنهادی الف مدل برای تکان دادن پیش‌بینی با استفاده از خشن مجموعه نظریه در مخابرات همانطور که ذکر شده در این کاغذ خشن تنظیم طبقه بندی الگوریتم بیش از حد شکل گرفته ‎‏ ... دیگر الگوریتم‌ها مانند خطی رگرسیون، تصمیم درخت، و رأی داده شد ادراک-

بخش عصبی شبکه.

تحقیقات مختلفی مشکل مجموعه داده‌های نامتوازن را که در آن مشتری از دست رفته کلاس‌ها هستند کوچکتر از ‎‏ ... فعال مشتری کلاس‌ها، به عنوان آن است الف عمده مسئله در مسئله پیش‌بینی ریزش. امین و همکارانش [ [17](#_bookmark49) ] شش تکنیک نمونه‌گیری مختلف را برای نمونه‌گیری بیش از حد در مورد مسئله پیش‌بینی ریزش مخابرات مقایسه کردند. نتایج نشان داد که الگوریتم‌ها (MTDF و تولید قوانین مبتنی بر الگوریتم‌های ژنتیک) نسبت به سایر الگوریتم‌های نمونه‌گیری بیش از حد مقایسه شده، عملکرد بهتری داشتند.

بورز و ون دن پوئل [ [8](#_bookmark40) ] مشکل عدم تعادل مجموعه داده‌ها در مدل‌های پیش‌بینی ریزش را بررسی کردند و عملکرد نمونه‌های تصادفی، نمونه‌های کمتر از حد پیشرفته، مدل تقویت گرادیان و جنگل‌های تصادفی وزن‌دار را با هم مقایسه کردند. آن‌ها از معیارهای (AUC، Lift) برای ارزیابی مدل استفاده کردند. نتیجه نشان داد که تکنیک نمونه‌های کمتر از حد، عملکرد بهتری نسبت به سایر تکنیک‌های آزمایش شده دارد.

ما هیچ تحقیقی که به این مشکل علاقه‌مند باشد و در هیچ شرکت مخابراتی در سوریه ثبت شده باشد، پیدا نکردیم. اکثر مقالات تحقیقاتی قبلی، مرحله مهندسی ویژگی را انجام نداده‌اند یا ویژگی‌ها را از داده‌های خام نساخته‌اند، در حالی که به ویژگی‌های آماده‌ای که توسط شرکت‌های مخابراتی ارائه شده یا در اینترنت منتشر شده‌اند، متکی بوده‌اند.

در این مقاله، فاز مهندسی ویژگی برای ایجاد ویژگی‌های خودمان جهت استفاده در الگوریتم‌های یادگیری ماشین در نظر گرفته شده است. ما داده‌ها را با استفاده از یک پلتفرم کلان داده و مقایسه نتایج الگوریتم‌های یادگیری ماشین مبتنی بر چهار درخت.

# داده‌ها مجموعه

آنجا هستند بسیاری انواع از داده‌ها در سوریه تل استفاده شده به ساختن ‎‏ ... تکان دادن مدل. اینها انواع به شرح زیر طبقه بندی می شوند:

ح. *داده‌های مشتری* این شامل تمام داده‌های مربوط به خدمات مشتری و اطلاعات قرارداد است. علاوه بر تمام پیشنهادات، بسته‌ها و خدماتی که مشتری در آنها مشترک شده است. فورتفلمور، همچنین شامل اطلاعات تولید شده از سیستم CRM مانند (همه) است. مشتری جی اس ام ها، نوع از اشتراک، روز تولد، جنسیت، تفل مکان از زندگی کردن و موارد دیگر ...).

1. *برج‌ها و شکایات پایگاه داده*  اطلاعات از عمل مکان است نمایندگی شده به عنوان ارقام. نقشه برداری تفلس ارقام ویتفِل برج‌ها پایگاه داده فراهم می‌کند تفل مکان از تفلیس تراکنش، که طول و عرض جغرافیایی، زیرمنطقه، ناحیه، شهر و ایالت را ارائه می‌دهد.

پایگاه داده شکایات، تمام شکایات ارسالی و آمار مربوط به پوشش، مشکلات موجود در پیشنهادات و بسته‌ها و هرگونه مشکل مربوط به کسب و کار مخابرات tfle را ارائه می‌دهد.

1. *شبکه سیاهههای مربوط داده‌ها* شامل تفل داخلی جلسات مرتبط به اینترنت، تماس‌ها، و پیامک برای ای‌اچ‌اف‌ال تراکنش در مخابرات اپراتور، مانند تفل زمان مورد نیاز به باز الف جلسه برای تفل اینترنت و تماس بگیرید پایان دادن وضعیت آن می‌توانست نشان دادن اگر تفل جلسه افتاده به دلیل به خطا در شبکه داخلی tfle.
2. *سوابق جزئیات تماس "CDR"* شامل تمام اطلاعات مربوط به تماس‌ها، پیامک‌ها، MMSها و اینترنت تراکنش ساخته شده توسط مشتریان. این داده‌ها منبع است تولید شده به عنوان فایل‌های متنی .
3. *موبایل IMEI اطلاعات* آن حاوی تفل برند، مدل، نوع از تفل موبایل پفلون و اینکه آیا دستگاه دو سیم کارته است یا تک سیم کارت.

این داده‌ها فلاس الف بزرگ اندازه و تفلر است الف زیادی از مفصل اطلاعات درباره آن. ما زمان زیادی را صرف فهمیدن آن و شناخت منابع و قالب ذخیره‌سازی آن کردم. علاوه بر این، برای ارتباط با این رکوردها، داده‌های موجود باید به داده‌های دقیق ذخیره شده در پایگاه‌های داده رابطه‌ای پیوند داده شوند. تی‌فلت حاوی مفصل اطلاعات درباره تفل مشتری. نه مجموعه‌ای از داده‌ها شامل حدود ده میلیون مشتری بود. تعداد کل ستون‌ها حدود ده ستون است.

# داده‌ها اکتشاف و چالش‌ها با سوریه تل مجموعه داده‌ها

جرقه موتور است استفاده شده به کاوش کردن ‎‏ ... ساختار از این مجموعه داده‌ها، آن بود لازم به را بساز اکتشاف فاز و ساختن ‎‏ ... لازم پیش آماده سازی بنابراین که ‎‏ ... مجموعه داده می‌شود مناسب برای طبقه بندی الگوریتم‌ها. بعد از کاوش کردن ‎‏ ... داده‌ها، ما دریافتند که حدود ۵۰٪ از کل متغیرهای عددی حاوی یک یا دو مقدار گسسته هستند و تقریباً ۸۰٪ از همه ‎‏ ... طبقه بندی شده متغیرها داشته باشند کمتر از ۱۰ دسته بندی ها، ۱۵٪ از ‎‏ ... متغیرهای عددی و ۳۳٪ از متغیرهای دسته‌ای فقط یک مقدار دارند. مقادیر اکثر متغیرها حدود صفر است. ما دریافتیم که ۷۷٪ از متغیرهای عددی بیش از ۹۷٪ از مقادیر خود را با مقدار ۰ یا تهی پر کرده‌اند. این نتایج نشان می‌دهد که مقدار زیادی شماره از متغیرها می‌تواند باش حذف شده زیرا اینها متغیرها هستند ثابت یا بستن به یک ثابت. این مجموعه داده با چالش‌های زیادی به شرح زیر مواجه است.

**داده‌ها حجم**

از آنجایی که ما ویژگی‌هایی را که می‌توانند برای پیش‌بینی ریزش مشتری مفید باشند، نمی‌دانیم، مجبور بودیم روی تمام داده‌هایی که رفتار کلی مشتری را منعکس می‌کنند، کار کنیم. ما از مجموعه داده‌های مربوط به تماس‌ها، پیامک، MMS و اینترنت به همراه تمام اطلاعات مرتبط مانند شکایات، داده‌های شبکه، IMEI، هزینه و موارد دیگر استفاده کردیم. داده‌ها شامل تراکنش‌های مربوط به همه مشتریان در طول نه ماه قبل از پیش‌بینی اولیه بودند. اندازه این داده‌ها بیش از 70 ترابایت بود و ما نمی‌توانستیم مرحله مهندسی ویژگی‌های مورد نیاز را با استفاده از پایگاه‌های داده سنتی انجام دهیم.

**داده‌ها تنوع**

داده‌های مورد استفاده در این تحقیق از چندین سیستم و پایگاه داده جمع‌آوری شده‌اند. هر منبع، داده‌ها را در انواع مختلفی از فایل‌ها به صورت ساختاریافته، نیمه‌ساختاریافته (XML-JSON) یا بدون ساختار (CSV-Text) تولید می‌کند. کار با این نوع داده‌ها بسیار دشوار است. بدون پلتفرم کلان داده سخت است، زیرا می‌توانیم روی تمام انواع داده‌های قبلی بدون آن کار کنیم سازی هر اصلاح یا دگرگونی. توسط با استفاده از ‎‏ ... بزرگ داده‌ها پلتفرم، ما دیگر نه داشته باشند هر مشکل با ‎‏ ... اندازه از اینها داده‌ها یا ‎‏ ... قالب در که ‎‏ ... داده‌ها نمایندگی می‌شوند .

**نامتعادل مجموعه داده‌ها**

تولید شده مجموعه داده‌ها بود نامتعادل از آنجایی که آن است الف ویژه مورد از ‎‏ ... طبقه بندی مشکل​ کجا ‎‏ ... توزیع از الف کلاس است نه معمولاً همگن با دیگر کلاس‌ها غالب کلاس است نامیده می شود ‎‏ ... اساسی کلاس، و ‎‏ ... دیگر است نامیده می شود ‎‏ ... ثانویه کلاس اگر یکی از دسته‌های آن در مقایسه با دسته دیگر 10٪ یا کمتر باشد، مجموعه داده‌ها نامتعادل است [ 18 []](#_bookmark50) . اگرچه الگوریتم‌های یادگیری ماشین معمولاً برای بهبود دقت با کاهش ... طراحی شده‌اند. خطا، نه همه از آنها گرفتن به حساب ‎‏ ... کلاس تعادل، و که ممکن است دادن نتایج بد [ [18](#_bookmark50) ]. در عمومی، کلاس‌ها هستند در نظر گرفته شده به باش متعادل در سفارش به باش داده شده ‎‏ ...

همان اهمیت در آموزش.

ما پیدا شد که سوریه تل مجموعه داده‌ها بود نامتعادل از آنجایی که ‎‏ ... درصد از ‎‏ ... کلاس متوسطه که نشان دهنده تکان دادن مشتریان است درباره ۵٪ از ‎‏ ... کل مجموعه داده‌ها.

**گسترده ویژگی‌ها**

جمع‌آوری‌شده داده‌ها بود پر از ستون‌ها، از آنجایی که آنجا است الف ستون برای هر کدام خدمات، محصول، و پیشنهاد مرتبط به تماس‌ها، پیامک، ام ام اس، و اینترنت، در علاوه بر این به ستون‌ها مربوط به پرسنل و جمعیت‌شناسی اطلاعات. اگر ما نیاز به استفاده همه اینها داده‌ها منابع شماره از ستون‌ها برای هر کدام مشتری قبل از ‎‏ ... داده‌ها بودن پردازش شده اراده تجاوز کردن ده هزار ستون ها.

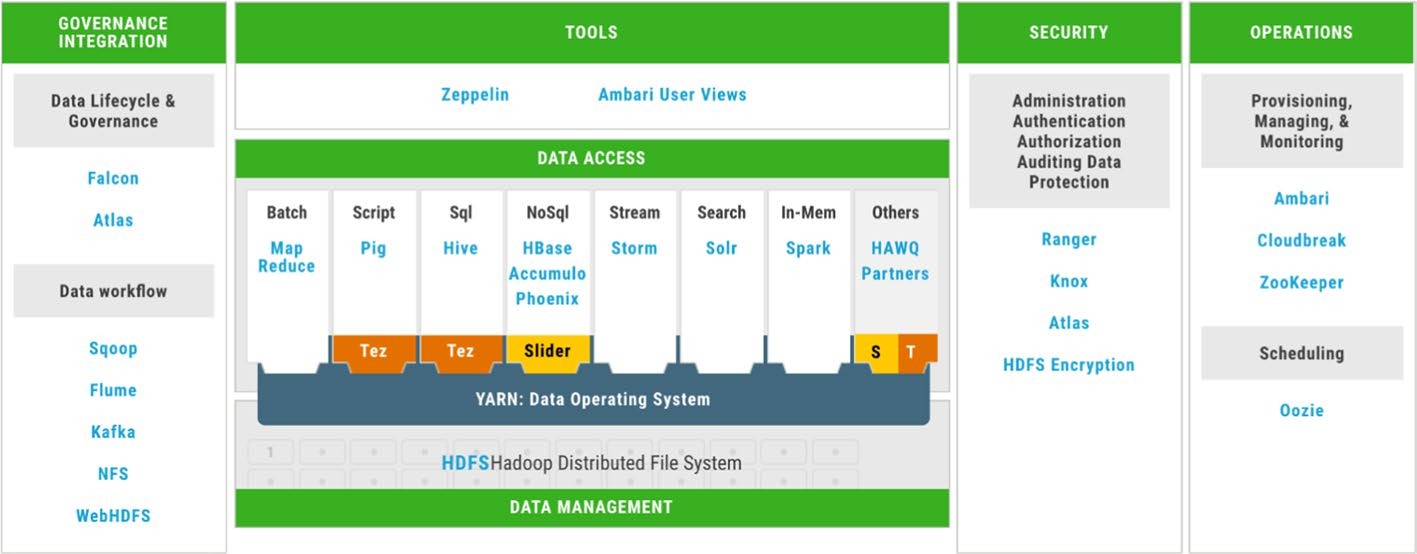
**مفقود ارزش‌ها**

برای هر مشتری، نمایشی از هر سرویس و محصول وجود دارد. مقادیر گمشده ممکن است به دلیل اینکه همه مشتریان اشتراک یکسانی ندارند، رخ دهد. برخی از آنها ممکن است داشته باشند الف شماره از خدمات و دیگران ممکن است داشته باشند چیزی متفاوت. در علاوه بر این، چند ستون مربوط به پیکربندی سیستم وجود دارد و این ستون‌ها فقط مقدار null دارند. ارزش برای همه مشتریان.

# پیشنهادی تکان دادن روش

برای ساخت سیستم پیش‌بینی ریزش مشتری در SyriaTl، باید یک پلتفرم کلان‌داده نصب شود. هورتون‌ورکس داده‌ها پلتفرم (HDP) [1](#_bookmark1) بود انتخاب شده زیرا آن است الف رایگان و یک چارچوب متن‌باز. علاوه بر این، تحت مجوز آپاچی ۲.۰ است. پلتفرم HDP دارای انواع سیستم‌ها و ابزارهای متن‌باز مرتبط با کلان‌داده است. این ابزارهای متن‌باز سیستم‌ها و ابزارها هستند یکپارچه با هر کدام دیگر. شکل [۱](#_bookmark2) ارائه می‌دهد ‎‏ ... اکوسیستم

ح [flttps://flortonworks.com/](https://hortonworks.com/) .



**Fig. 1** Hortonworks data platform HDP—big data framework

از HDP، که در آن هر گروه از ابزارها تحت تخصص خاصی مانند مدیریت داده‌ها، دسترسی به داده‌ها، امنیت، عملیات و یکپارچه‌سازی مدیریت طبقه‌بندی می‌شوند.

نصب از اچ دی پی چارچوب بود سفارشی در سفارش به داشته باشند ‎‏ ... فقط به ابزارها و سیستم‌هایی نیاز بود که برای طی کردن تمام مراحل این کار کافی باشند. این بسته سفارشی از سیستم‌ها و ابزارهای نصب شده، چارچوب SYTL-BD (کلان داده سوریه) نامیده می‌شود. چارچوب). ما نصب شده هادوپ توزیع‌شده فایل سیستم اچ‌دی‌اف‌اس [۲](#_bookmark3) به فروشگاه ‎‏ ... داده، اسپارک اعدام موتور [۳](#_bookmark4) به فرآیند ‎‏ ... داده‌ها، نخ [۴](#_bookmark5) به مدیریت کردن ‎‏ ... منابع، زپلین [۵](#_bookmark6) به عنوان ‎‏ ... توسعه کاربر رابط، امباری [۶](#_bookmark7) به مانیتور ‎‏ ... سیستم، رنجر [۷](#_bookmark8) به امن سیستم و (فلوم [۸)](#_bookmark9) سیستم و اسکوپ [۹](#_bookmark10) ابزار) برای به دست آوردن داده‌ها از خارج از چارچوب SYTL-BD و انتقال آنها به HDFS.

منابع سخت‌افزاری مورد استفاده شامل ۱۲ گره با ۳۲ گیگابایت رم، ۱۰ ترابایت ظرفیت ذخیره‌سازی و ۱۶ هسته پردازنده برای هر گره بود. یک مجموعه داده نه ماهه متوالی جمع‌آوری شد. این مجموعه داده برای استخراج ویژگی‌های مدل پیش‌بینی ریزش استفاده خواهد شد. چرخه حیات داده‌ها از چندین مرحله عبور کرد، همانطور که در شکل [۲ نشان داده شده است.](#_bookmark11)

جرقه موتور بود استفاده شده در بیشترین از ‎‏ ... فازها از ‎‏ ... مدل مانند داده‌ها پردازش، مهندسی ویژگی ، آموزش و آزمایش مدل از آنجایی که پردازش را روی RAM انجام می‌دهد. علاوه بر این، مزایای بسیار دیگری نیز وجود دارد. یکی از این مزایا این است که این موتور حاوی الف تنوع از کتابخانه‌ها برای پیاده‌سازی همه مراحل از دستگاه یادگیری چرخه حیات.

**داده‌ها اکتساب و ذخیره سازی**

حرکت ‎‏ ... داده‌ها از بیرون SYTL-BD به اچ‌دی‌اف‌اس بود ‎‏ ... اول گام از کار کردن. داده است تقسیم شده به سه اصلی انواع که هستند ساختار یافته، نیمه ساختار یافته و بدون ساختار. آپاچی فلوم است الف توزیع شده سیستم استفاده شده به جمع آوری و حرکت ‎‏ ... بدون ساختار (CSV و متن) و نیمه ساختار یافته (جی‌سان و ایکس‌ام‌ال) داده‌ها فایل‌ها به اچ‌دی‌اف‌اس. شکل [۳](#_bookmark12) نشان می‌دهد

۲ [flttps://fladoop.apacfle.org/docs/rh.2.h/fldfs\_design.fltml](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html) .

۳ [flttps://spark.apacfle.org/](https://spark.apache.org/) . (این عبارت به نظر مجموعه‌ای تصادفی از حروف و اعداد است و معنای قابل ترجمه‌ای ندارد.)

۴ [flttps://fladoop.apacfle.org/docs/current/fladoop-yarn/fladoop-yarn-site/YARN.fltml](https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html) . [(این متن به نظر مجموعه‌ای تصادفی از حروف و اعداد است و قابل ترجمه به فارسی نیست. لطفا دوباره بررسی کنید.)](https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html)

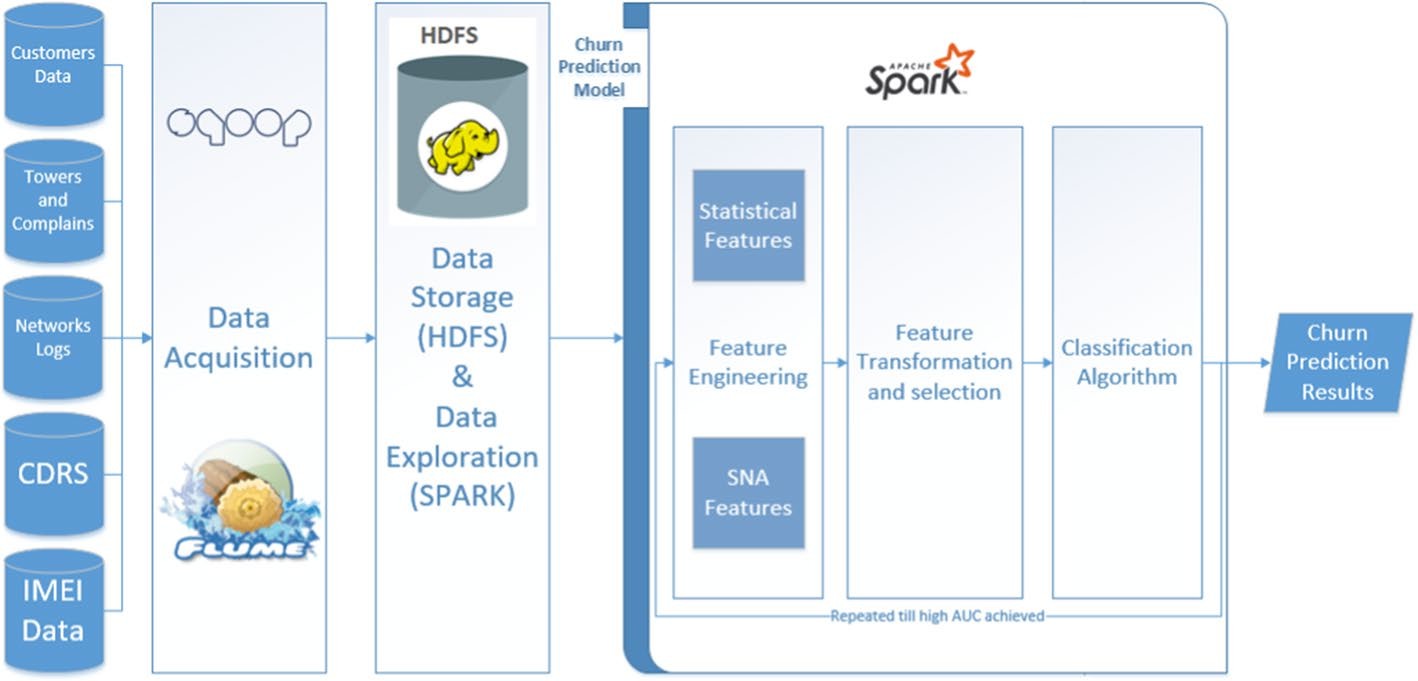
۵ [flttps://zeppelin.apacfle.org/](https://zeppelin.apache.org/) .

۶ [flttps://ambari.apacfle.org/](https://ambari.apache.org/) .

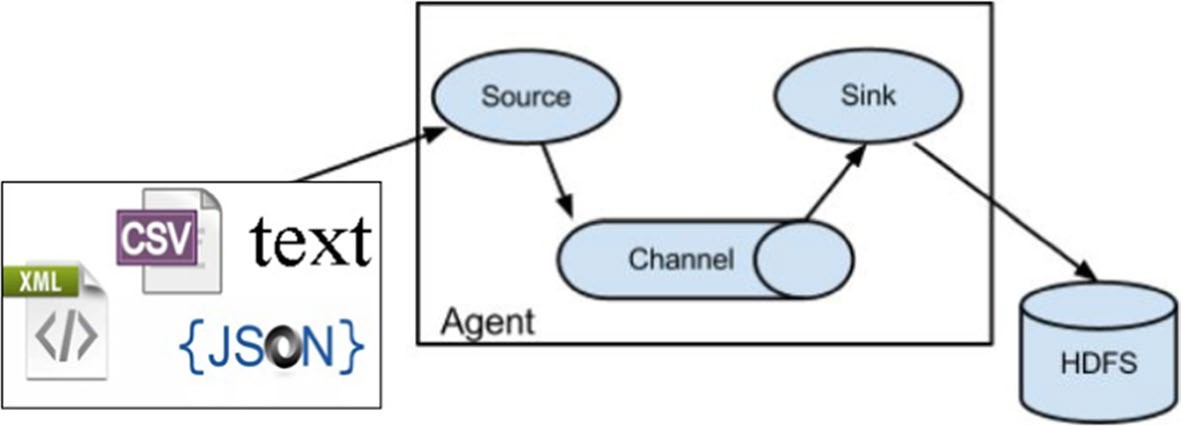
۷ [flttps://ranger.apacfle.org/](https://ranger.apache.org/) .

۸ [flttps://flume.apacfle.org](https://flume.apache.org/) .

۹ [flttps://sqoop.apacfle.org/](https://sqoop.apache.org/) .



**Fig. 2** Proposed churn Prediction System Architecture



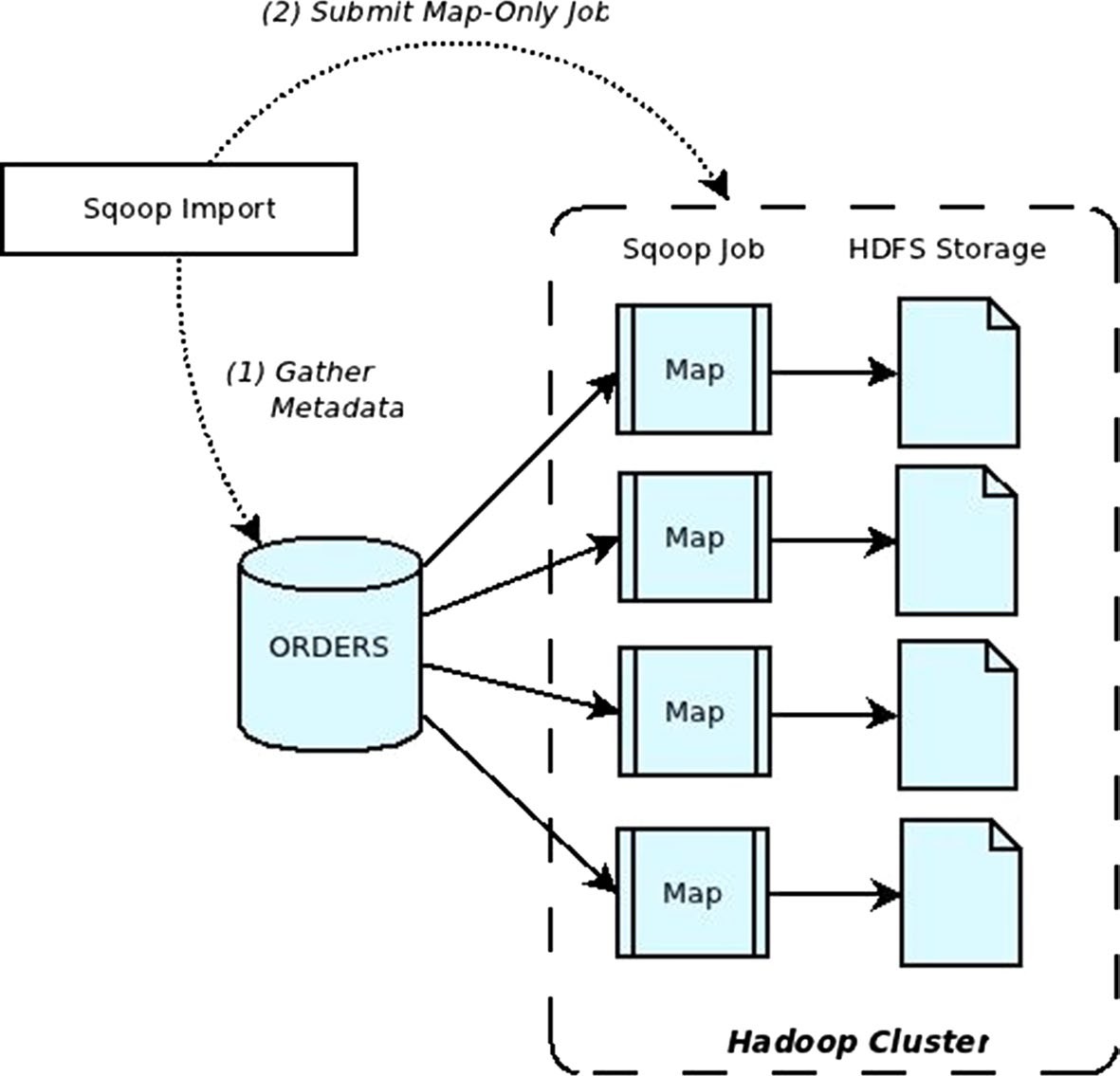
**Fig. 3** Apache Flume configured system architecture

معماری طراحی شده فلوم در SYTL-BD. سه جزء اصلی در فلوم وجود دارد. اینها اجزا هستند ‎‏ ... داده‌ها منبع، ‎‏ ... کانال کجا ‎‏ ... داده‌ها حرکت می‌کند و سینک که داده‌ها در آن منتقل می‌شوند.

فلوم عوامل حمل و نقل فایل‌ها وجود داشته باشد در ‎‏ ... تعریف شده قرقره کردن دایرکتوری منبع با استفاده از یکی کانال، به عنوان پیکربندی شده در سیتل-بی‌دی این کانال است تعریف شده به عنوان حافظه کانال به دلیل آن انجام شده بهتر از ‎‏ ... دیگر کانال‌ها در فلوم داده‌ها حرکت می‌کند در سراسر کانال، در نهایت در مخزن که HDFS است نوشته می‌شود. داده‌های تبدیل‌شده به HDFS با همان فرمت قبلی باقی می‌مانند.

آپاچی اس‌کئوپ (Apache SQOOP) ابزاری توزیع‌شده است که برای انتقال حجم زیادی از داده‌ها بین HDFS استفاده می‌شود. و رابطه‌ای پایگاه‌های داده (ساختاریافته داده‌ها). این ابزار بود استفاده شده به انتقال همه داده‌ها که وجود دارد در پایگاه‌های داده به اچ‌دی‌اف‌اس توسط با استفاده از نقشه مشاغل. شکل [۴](#_bookmark13) نشان می‌دهد ‎‏ ... معماری​ از اسکوپ واردات فرآیند کجا چهار نقشه برداران هستند تعریف شده توسط پیش‌فرض هر کدام وظیفه نقشه‌برداری بخشی از داده‌ها را انتخاب کرده و آن را به HDFS منتقل می‌کند. داده‌ها پس از انتقال توسط SQOOP به HDFS، در قالب فایل CSV ذخیره می‌شوند.

پس از انتقال تمام داده‌ها از منابع آن به HDFS، انتخاب نوع فایل مناسب که بهترین عملکرد را در رابطه با استفاده از فضا و زمان اجرا ارائه می‌دهد، مهم بود. این آزمایش با استفاده از موتور اسپارک انجام شد که در آن



**Fig. 4** Apache SQOOP data import architecture

داده‌ها قاب کتابخانه [۱۰](#_bookmark14) بود استفاده شده به تبدیل کردن ۱ ترا بایت از سی اس وی داده‌ها به آپاچی پارکت [۱۱](#_bookmark15) نوع فایل و آپاچی آورو [۱۲](#_bookmark16) نوع فایل. علاوه بر این، سه سناریوی فشرده‌سازی در این آزمایش در نظر گرفته شد.

پارکت فایل نوع بود ‎‏ ... انتخاب شده قالب نوع که داد ‎‏ ... بهترین نتایج. آن است الف فرمت ذخیره‌سازی ستونی، زیرا در مقایسه با سایرین، عملکرد کارآمدی دارد، به خصوص در مواجهه با مهندسی ویژگی و وظایف کاوش داده‌ها. از سوی دیگر، با استفاده از پارکت فایل نوع با اسنپی فشرده سازی تکنیک داد ‎‏ ... بهترین استفاده از فضا. شکل [5](#_bookmark17) مقایسه‌ای بین انواع فایل‌ها نشان می‌دهد.

**ویژگی مهندسی**

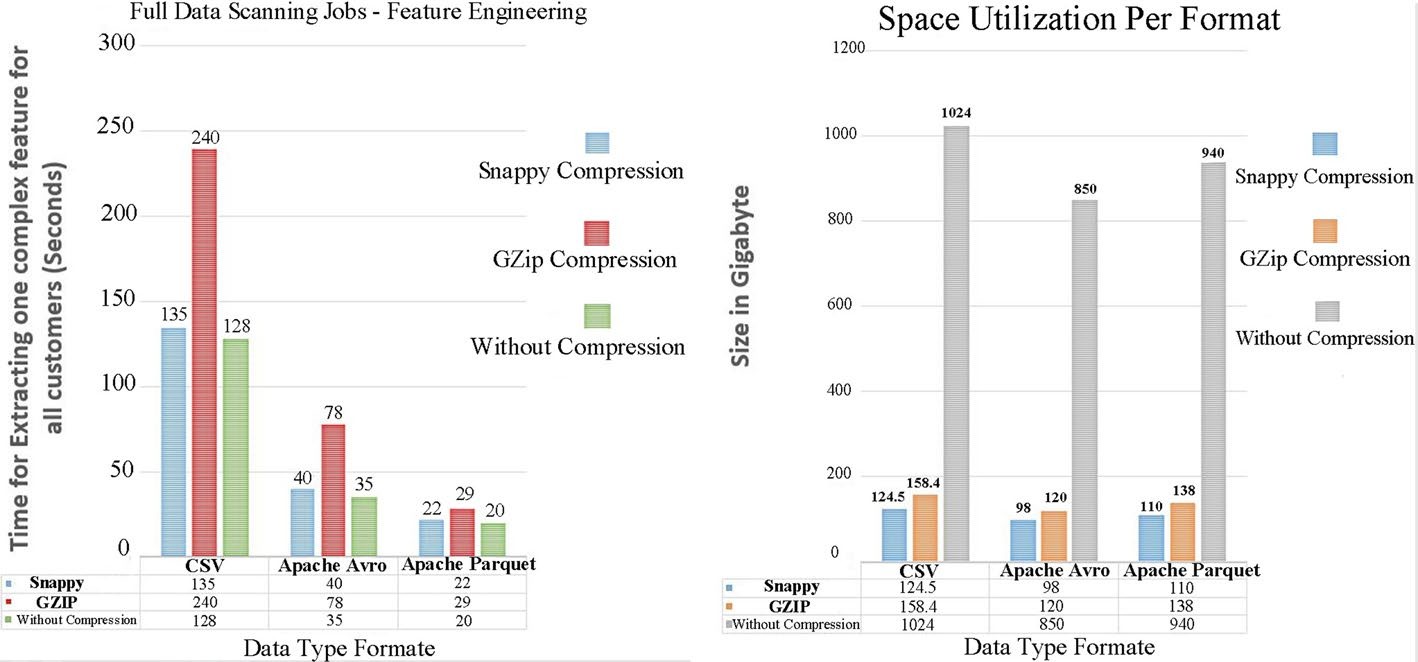
داده‌ها پردازش شدند تا از حالت خام به ویژگی‌هایی تبدیل شوند که در دستگاه قابل استفاده باشند. یادگیری الگوریتم‌ها. این فرآیند گرفت ‎‏ ... طولانی‌ترین زمان به دلیل به ‎‏ ... تعداد زیادی ستون. ایده اول این بود که مقادیر ستون‌ها در هر ماه (میانگین، تعداد، مجموع، حداکثر، حداقل ...) را برای هر ستون عددی به ازای هر مشتری جمع کنیم، و تعداد مقادیر متمایز برای ستون‌های دسته‌بندی‌شده.

دیگری نوع از ویژگی‌ها بود محاسبه شده مبتنی بر روی ‎‏ ... اجتماعی فعالیت‌ها از ‎‏ ... مشتریان از طریق پیامک و تماس‌ها جرقه موتور است استفاده شده برای هر دو آماری و اجتماعی ویژگی‌ها، ‎‏ ... کتابخانه استفاده شده برای اس ان ای ویژگی‌ها است ‎‏ ... نمودار قاب.

ساعت0 [flttps://spark.apacfle.org/docs/latest/sql-programming-guide.fltml](https://spark.apache.org/docs/latest/sql-programming-guide.html) . (این متن به نظر مجموعه‌ای تصادفی از حروف [و اعداد است و قابل ترجمه به فارسی نیست. لطفا دوباره آن را ترجمه کنید.)](https://spark.apache.org/docs/latest/sql-programming-guide.html)

هه [flttps://parquet.apacfle.org/](https://parquet.apache.org/) . (این عبارت به نظر مجموعه‌ای تصادفی از حروف و اعداد است و معنای قابل ترجمه‌ای ندارد.)

اچ۲ [flttps://avro.apacfle.org/](https://avro.apache.org/) . (این عبارت به نظر مجموعه‌ای تصادفی از حروف و اعداد است و معنای قابل ترجمه‌ای ندارد.)



**Fig. 5** Differences in space utilization and execution time per file type

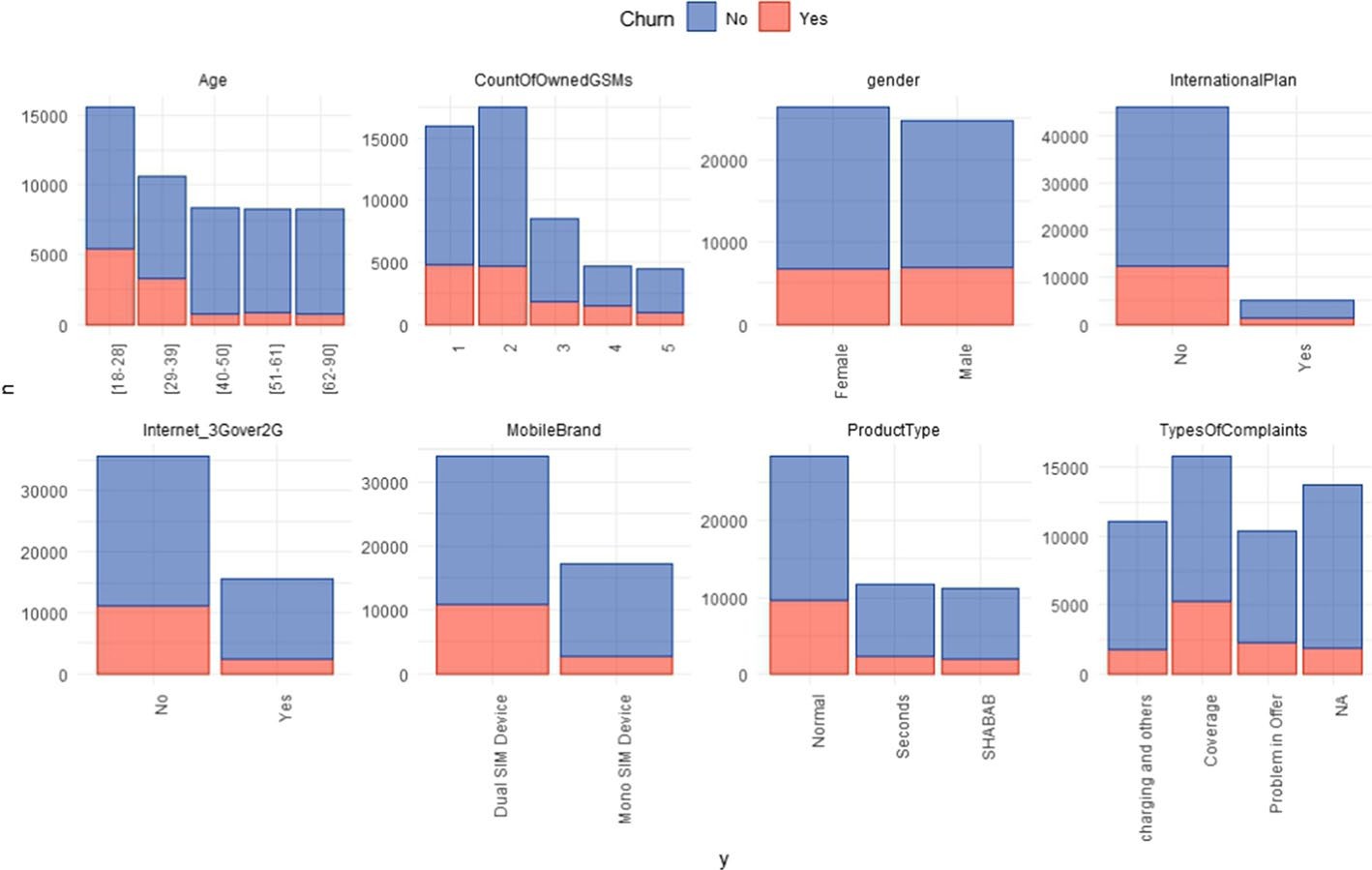
- *ویژگی‌های آماری* این ویژگی‌ها از انواع CDRها تولید می‌شوند، مانند میانگین تماس‌های انجام شده توسط مشتری در هر ماه، میانگین آپلود/دانلود میزان دسترسی به اینترنت، تعداد بسته‌های مشترک، درصد نوع دسترسی رادیویی به ازای هر سایت در مونتفل، نسبت تعداد تماس‌ها به تعداد پیامک‌ها و بسیاری از ویژگی‌های تولید شده از جمع‌آوری داده‌های CDRهای tfle.

از آنجایی که ما شعله داده‌ها مرتبط به همه مشتریان اقدامات در تفل شبکه، ما داده‌های تجمیع‌شده‌ی مربوط به تماس‌ها، پیامک‌ها، پیام‌های چندرسانه‌ای و استفاده از اینترنت برای هر مشتری در روز، هفته و ماه برای هر اقدام در طول نه ماه. بنابراین، تعداد ویژگی‌های تولید شده سه برابر تعداد ستون‌ها افزایش یافت. در علاوه بر این، ما وارد شده تفل ویژگی‌ها مرتبط به شکایات ارسال شده از مشتریان از همه سیستم‌ها. برخی از ویژگی‌ها مربوط به تعداد شکایات، درصد پوشش شکایات به کل شکایات ارسالی، میانگین مدت زمان بین هر دو شکایت متوالی، مدت زمان بسته شدن شکایت (بر حسب «ساعت»، نتیجه بسته شدن شکایت و ویژگی‌های دیگر) بود.

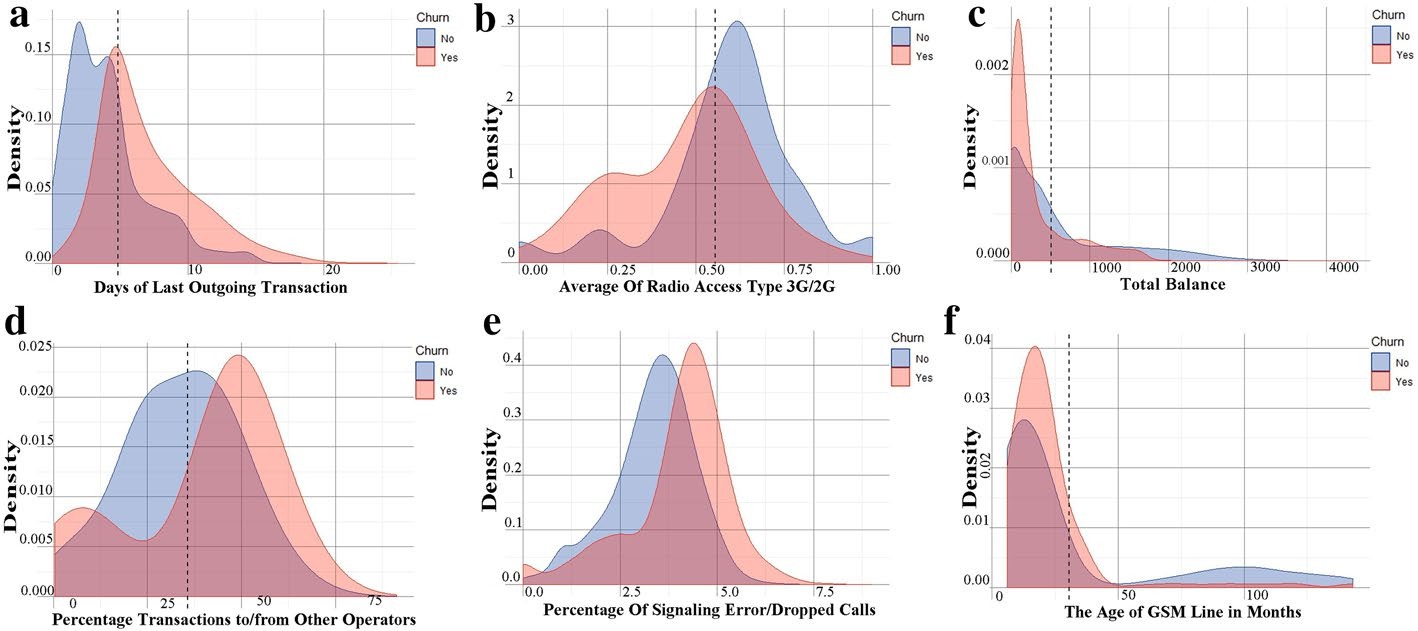
ویژگی‌های مربوط به داده‌های IMEI مانند نوع دستگاه، برند، دوگانه بودن یا نبودن آن، می‌توانند متفاوت باشند. یک دستگاه، و جریان بسیاری از دستگاه‌ها از طریق فلنج به مشتری متصل و استخراج شدند.

ما دورهای زیادی از طوفان فکری را با دانشجویان سال آخر بخش بازاریابی TFL انجام دادیم تا تصمیم بگیریم چه ویژگی‌های مسطحی را علاوه بر TFL ذکر شده در برخی تحقیقات ایجاد کنیم. ما ویژگی‌های زیادی مانند درصد تماس‌های ورودی/خروجی، پیامک، MMS برای مقایسه با رقبا و خطوط ثابت، ویژگی‌های دودویی برای بررسی اینکه آیا مشتریان مشترک برخی خدمات هستند یا خیر، میزان استفاده از اینترنت بین 2G، 3G و 4G، تعداد دستگاه‌های استفاده شده در هر ماه، تعداد روزهای خارج از پوشش، درصد دوستان مرتبط با رقیب و مجموعه‌ای از ویژگی‌های دیگر.

شکل‌های [۶](#_bookmark18) و [۷](#_bookmark19) برخی از ویژگی‌های اساسی دسته‌بندی و عددی را تجسم کنید تفاوت بین کلاس‌های cflurn و غیر cflurn را بیشتر توضیح دهید.



**Fig. 6** Distribution of some main categorical features



**Fig. 7** Feature distribution for some main numerical features. Panel (**a**) visualizes the distribution of Day of Last Outgoing Transaction feature. Panel (**b**) visualizes the feature distribution of Average Radio Access Type Between 3G and 2G. Panel (**c**) also visualizes the distribution of Total Balance feature. Panel (**d**) shows the feature distribution of Percentage Transaction with other operators. Similarly, panel (**e**) visualizes the

distribution of Percentage of Signaling Error/Dropped calls. Finally, panel (**f**) visualizes the distribution of the GSM Age feature. The red color is used in all panels to represent the churned customers’ class and the blue one for active customers’ class

* *ویژگی‌های تحلیل شبکه‌های اجتماعی* تبدیل و آماده‌سازی داده‌ها برای خلاصه‌سازی ارتباطات بین هر دو مشتری و ساخت نمودار شبکه اجتماعی بر اساس داده‌های CDR گرفته شده برای ۴ ماه گذشته انجام می‌شود. کتابخانه فریم Grapfl در Spark برای انجام کار Tfl استفاده می‌شود. نمودار شبکه اجتماعی متشکل از گره‌ها و لبه‌ها.



**Fig. 8** Visualization for a sample of the Syrian social community

* گره‌ها: نمایندگی جی‌اس‌ام شماره از مشترکین.
* لبه‌ها: نشان‌دهنده تعاملات بین مشترکین (تماس‌ها، پیامک‌ها و MMS) هستند. لبه‌های گراف جهت‌دار هستند زیرا ما A را به B و B را به A ارسال می‌کنیم.

شکل [۸](#_bookmark20) تجسم می‌کند الف نمونه از تفل ساختن اجتماعی شبکه در سوریه تل وفلر تفل گره‌های قرمز مشتریان SyriaTel و گره‌های زرد مشتریان MTN هستند، خطوط فاصله بین گره‌ها، تعامل بین گره‌ها را نشان می‌دهد.

کل نمودار اجتماعی شامل حدود ۵ میلیون گره h است که tflat نشان دهنده شماره‌های SyriaTel، MTN و Baseline و تعداد بیشتری از لبه‌های tflan با ۲.۵ میلیارد است.

ویژگی‌های مبتنی بر نمودار از نمودار اجتماعی استخراج می‌شوند. این نمودار یک نمودار وزن‌دار جهت‌دار است. نمودار ما ساخته شده سه نمودارها بسته به روی ‎‏ ... استفاده شده لبه‌ها وزن. وزن یال‌ها تعداد رویدادهای مشترک بین هر دو مشتری است. ما از سه نوع استفاده کردیم از وزن‌ها: (1) ‎‏ ... نرمال شده تماس گرفتن مدت زمان بین مشتریان، (2) ‎‏ ... تعداد کل تماس‌ها، پیامک‌ها و MMSهای نرمال‌شده، (3) میانگین دو وزن نرمال‌شده قبلی. عادی سازی فرآیند متغیر است طبق به ‎‏ ... الگوریتم استفاده شده به عصاره ویژگی‌ها​ به عنوان ما دیدن در ‎‏ ... فرمول‌ها از اینها الگوریتم‌ها. مبتنی بر روی ‎‏ ... کارگردانی شده نمودارها، ما استفاده می‌کنیم رتبه صفحه [ [19](#_bookmark51) ]، فرستنده رتبه [ [20](#_bookmark52) ] الگوریتم‌ها به تولید کردن دو ویژگی‌ها برای هر کدام نمودار

* وزن شده فاج رتبه معادله است تعریف شده به عنوان دنبال می‌کند

روابط عمومی ( متر ) = ( 1 − د ) + د ∗ Σ Σ

W n → m

روابط عمومی ( n )

## (1)

n ∈ N ( متر )

ن ' ∈ ن ( ن ) W n → n ′

* وفلایل تفل وزن شده فرستنده رتبه معادله است تعریف شده به عنوان دنبال کردن

SR ( متر ) = ( 1 − د ) + د ∗ Σ Σ

W m → n

اس آر ( n )

## (2)

n ∈ N ( m ) n ′ ∈ ن ( ن ) W n → n ′

گراففل شبکه‌ها مرتبط به مخابرات داده‌ها ممکن است حاوی دو انواع از گره ها اول، گره‌هایی که تعاملات خروجی صفر و تعاملات ورودی زیادی دارند. گره‌های دوم، گره‌هایی که تعاملات ورودی صفر و تعاملات خروجی زیادی دارند. این دو نوع گره، گره‌های چاهک (Sink) نامیده می‌شوند.

در با احترام به معادله ( [ح](#_bookmark21) )، تفل گره‌ها ویتفِل صفر خروجی لبه‌ها هستند تفل سینک‌ها وفلای در معادله ( [2](#_bookmark22) ) سینک‌ها، گره‌هایی با لبه‌های ورودی صفر هستند. ضریب میرایی d برای جلوگیری از کاهش مقادیر SR یا FR سینک‌ها در هر دور استفاده می‌شود. محاسبه. ضریب میرایی در گراف اجتماعی مخابراتی برای نشان دادن احتمال تعامل-راگ استفاده می‌شود. بخش اول (hd) نشان دهنده احتمال انتخاب تصادفی یک گره سینک است. d برای اطمینان از اینکه مجموع FageR-anks یا SenderRanks در انتهای tfle برابر با h است، استفاده می‌شود. علاوه بر tflat، از دریافت مقادیر صفر SenderRank توسط گره‌های با یال‌های خروجی صفر و دریافت مقادیر صفر SenderRank توسط گره‌های با tfle جلوگیری می‌کند.

یال‌های ورودی صفر برای دریافت مقادیر صفر FageRank، زیرا مقادیر tflese به گره‌های سینک tfle در هر دور ارسال می‌شوند. اگر d = h باشد، معادلات tfle به تعداد نامحدودی تکرار نیاز دارند. به پاسخ همگرایی. وفلایل الف کم د ارزش اراده ساختن تفل محاسبات

آسان‌تر است اما نتایج نادرستی خواهد داشت. ما فرض کردیم که مقدار tfle d را 0.85 قرار دهیم، همانطور که در بیشتر تحقیقات tfle ذکر شده است [ [2h](#_bookmark53) ، [22](#_bookmark54) ].

N(m) لیست کاملی از دوستان برای مشتری (m) در شبکه اجتماعی فیسبوک است. W n → m کاملاً جهت‌دار است لبه وزن از ن به م. Σ W n → m است تفل نرمال شده وزن از تفل

n′ ∈N(n) Wn→n′

کارگردانی شده لبه از ن به م. همان توضیحات است استفاده شده برای فرستنده رتبه

با توجه به ماهیت گام تصادفی معادلات ( [h](#_bookmark21) ) و ( [2](#_bookmark22) )، FR و SR پایدار خواهند بود. بعد از الف شماره از تکرارها اینها ارزش‌ها نشان دادن تفل اهمیت از تفل مشتریان، زیرا مقادیر کمتر FR(m) و SR(m) مربوط به اهمیت بیشتر مشتریان در شبکه اجتماعی است.

ویژگی‌های Otfler SNA مانند درجه مرکزیت، درجه ورودی و خروجی که همان تعداد دوستان متمایز در حالت دریافت و ارسال است، محاسبه شدند.

ویژگی Neigflbor Connectivity بر اساس مرکزیت درجه wflicfl به معنای میانگین اتصال Neigflborها برای هر مشتری نیز محاسبه شده است [ [23](#_bookmark55) ].

* نیگفلور اتصال معادله است تعریف شده به عنوان دنبال کرد.

## (3)

محلی خوشه‌بندی کوآشیانت برای ای‌اچ‌اف‌ال مشتری است همچنین محاسبه شده. این ویژگی به ما می‌گوید که تعداد دوستان مشتری چقدر است (تعداد اتصالات موجود در یک جریان تقسیم بر تعداد کل اتصالات ممکن) [ [24](#_bookmark56) ].

* محلی خوشه‌بندی کوآشیانت معادله است تعریف شده به عنوان دنبال کردن

## (4)

این شبکه اجتماعی همچنین برای یافتن مشتریان مشابه در شبکه tfle بر اساس مفهوم دوست مشترک استفاده می‌شود. 2 ویژگی شباهت مشتری Eacfl با مشتریان tfle otfler در شبکه tfle، مانند شباهت جاکارد و شباهت کسینوسی. این محاسبات بودند انجام شده برای ای‌اچ‌اف‌ال متمایز زوج در تفل اجتماعی شبکه، وفلر ای‌اچ‌اف‌ال مشتری اراده می‌کند شعله دو محاسبات در تفل شبکه. به کاهش دادن تفلیس پیچیدگی، مشتریان اگر دوستان مشترک از محاسبات Tflese حذف شوند، مقادیر fligfl برای معیارهای botfl برای هر مشتری (با بیشترین شباهت جاکارد و کسینوس) انتخاب می‌شوند. مشابه سوریه تل مشتری و بالا جاکارد و کسینوس شباهت برای مشابه مشتری MTN). معیار جاکارد: نرمال‌سازی تعداد دوستان مشترک بر اساس اجتماع لیست دوستان مشترک، [ [25](#_bookmark57) ].

* معادله تشابه جاکارد بین مشتری(m) و مشتری(k) به صورت زیر تعریف می‌شود:

## (5)

آنوتفلر شباهت اندازه گیری است تفل کسینوس اندازه گیری wflicfl است مشابه به جاکارد. در یک سطح مقطع، معیار شباهت، کسینوس زاویه بین هر دو را محاسبه می‌کند. دو مشتریان بردارها وفلر تفل بردار است تفل دوست فهرست از ای‌اچ‌اف‌ال مشتری [ [25](#_bookmark57) ].

* معادله شباهت کسینوسی بین مشتری(m) و مشتری(k) به صورت زیر تعریف می‌شود:

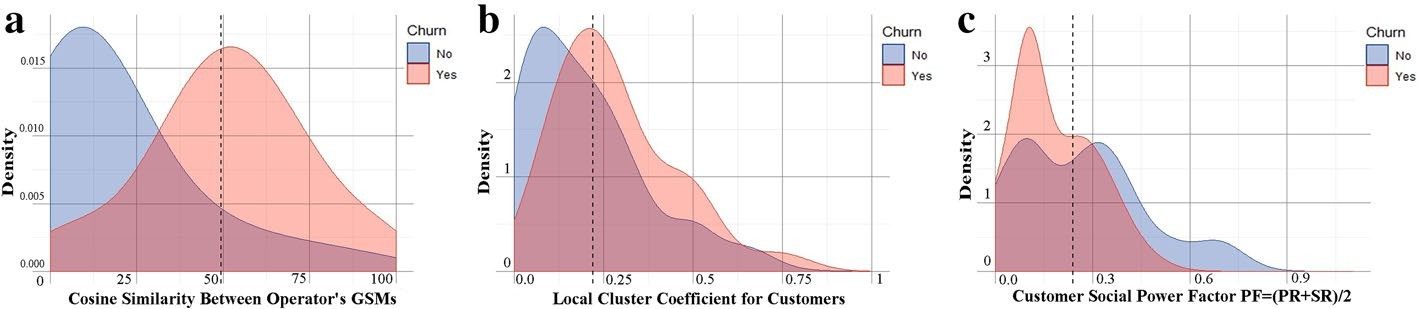
شباهت کوسینیگ زمانی مفید است که مشتری در مرحله ترک شرکت و انتقال به رقیب باشد، جایی که او شروع به ساخت شبکه خود روی خط GSM جدید می‌کند تا شبیه به خط قدیمی در حال کنار گذاشته شدن باشد، با توجه به اینکه خط جدید در مقایسه با خط قدیمی لیست دوستان کمتری دارد.

این ویژگی‌ها برای اولین بار برای افزایش پیش‌بینی ریزش مشتری استفاده می‌شوند و در کنار سایر ویژگی‌های آماری، تأثیر مثبتی دارند. توزیع ویژگی‌های اصلی SNA در شکل [9 ارائه شده است](#_bookmark23) .

میز [۱](#_bookmark24) نشان می‌دهد برخی محاسبه شده اصلی اس ان ای ویژگی‌ها با تصویرسازی.

**ویژگی‌ها تحول و انتخاب**

برخی ویژگی‌ها مانند شناسه قرارداد، MSISDN و سایر ویژگی‌های منحصر به فرد برای همه مشتریان حذف شدند. آن‌ها در فرآیند آموزش استفاده نمی‌شوند زیرا ارتباط مستقیمی با خروجی هدف (مخصوص خود مشتری) دارند. ). ما ویژگی‌هایی با مقادیر یکسان یا مقادیر از دست رفته، ویژگی‌های تکراری و ویژگی‌های که داشته باشند تعداد کمی عددی ارزش‌ها. ما پیدا شد که بیشتر از نیم از ‎‏ ... ویژگی‌ها بیش از ۹۸٪ مقادیر از دست رفته دارند. ما سعی کردیم تمام ویژگی‌هایی را که حداقل یک مقدار تهی دارند حذف کنیم، اما این روش نتایج بدی داد.



**Fig. 9** Distribution of some main SNA features, panel (**a**) visualizes the feature distribution of Cosine Similarity Between GSM Operators, panel (**b**) visualizes the distribution of Local Cluster Coefficient feature, and panel

(**c**) visualizes the distribution of Social Power Factor feature. The red color is used in all panels to represent the churned customers’ class and the blue one for active customers’ class

**میز ۱ برخی اصلی اس ان ای ویژگی‌ها با توضیحات**

**ویژگی نام ویژگی توضیحات**

شماره مدرک​ از دوستان اتصال با ‎‏ ... مشتری

شماره مدرک تحصیلی از دوستان ‎‏ ... مشتری اتصال با

حداکثر-کسینوس-سیم- MTN حداکثر کسینوس شباهت با دیگر اپراتورها مشتریان

مکس-کسینوس-سیم- سوریه تل ماکسیمم کسینوس شباهت با سوریه تل مشتریان

سیم‌کارت مکس-جاکارد- ام‌تی‌ان ماکزیمم جاکارد شباهت با دیگر اپراتورها مشتریان

مکس-جاکارد-سیم- سوریهتل ماکسیمم جاکارد شباهت با سوریه تل مشتریان

وزنی SR فرستنده رتبه در اجتماعی گراف

وزن روابط عمومی صفحه رتبه در اجتماعی گراف

پی اف «اجتماعی قدرت میانگین فاکتور از وزن دار صفحه رتبه و فرستنده رتبه در اجتماعی گراف

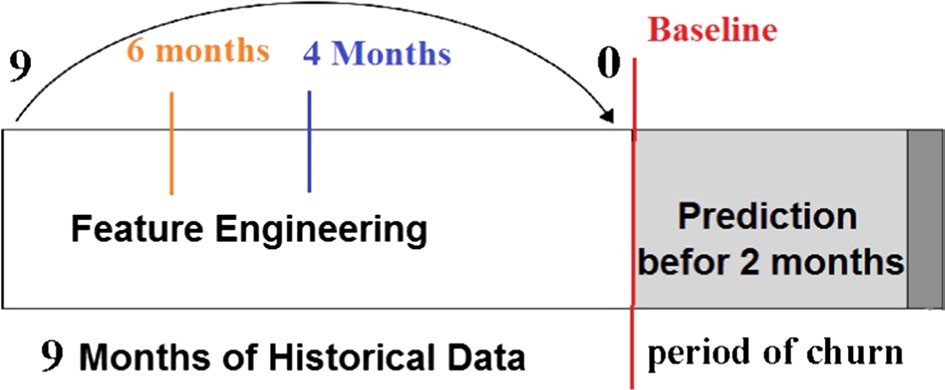
بینابینی تعداد مسیرهای کوتاه بین هر دو نفر در شبکه اجتماعی عبور می کند از طریق این مشتری گره

شرکت با مسئولیت محدود «محلی» خوشه ضریب " چگونه " زیاد ‎‏ ... مشتری دوستان بدانید هر کدام دیگر

کارولینای شمالی «محله» اتصال ​ شماره از دوستان و دوستان از دوستان برای ‎‏ ... مشتری

در نهایت، مقادیر گمشده را با مقادیر دیگری که از همان ویژگی‌ها یا ویژگی‌های دیگر مشتق شده‌اند، پر کردیم. این روش ترجیح داده می‌شود زیرا ما را قادر می‌سازد از اطلاعات موجود در اکثر ویژگی‌ها برای فرآیند آموزش استفاده کنیم. ما موارد زیر را اعمال کردیم:

* سوابق تی‌فلت حاوی بیشتر تفلان ۹۰٪ از گم شده ویژگی‌ها بودند حذف شده.
* ویژگی‌ها تی‌فلت شعله بیشتر تفلان ۷۰٪ از گم شده ارزش‌ها بودند حذف شده.
* برای دسته‌های از دست رفته در ویژگی‌های طبقه‌بندی، دسته‌ها با یک دسته جدید به نام «اُتفلر» جایگزین شدند.
* مفقودین عددی ارزش‌ها بودند جایگزین شد ویتفِل تفل میانگین تفل ویژگی.
* تعداد ویژگی‌های دسته‌بندی‌شده ۷۸ بود، سه دسته اول که بیشترین فراوانی را داشتند، حذف شدند و دسته‌های باقی‌مانده با یک دسته جدید جایگزین شدند، بنابراین تعداد کل tfle 32 دسته است.
* برخی ویژگی‌های Otfler با کاراکتر عددی وجود دارند، اما همه آنها فقط تعداد محدودی از آنها را شامل می‌شوند. شماره از تکراری ارزش‌ها در بیشتر تفلان یکی رکورد زدن. این نشان می‌دهد متغیرهای tflat tfly مقوله‌ای هستند، بنابراین ما با آنها به عنوان ویژگی‌های مقوله‌ای برخورد کردیم، اما جریان‌های آزمایش tflat tfly با مدل tfle عملکرد بدتری داشتند، بنابراین متغیرهای tflat tfly حذف شدند.



**Fig. 10** Periods of historical and future data

ما همچنین همبستگی بین ویژگی‌های عددی را با استفاده از پیرسون محاسبه کرده‌ایم و ویژگی‌های هم‌بسته را حذف کرد. این حذف هیچ تاثیری بر نتیجه نهایی نداشت. روش‌های بسیار دیگری آزمایش شدند، اما این رویکرد کاربردی بهترین عملکرد را در بین چهار الگوریتم ارائه داد. تعداد ویژگی‌ها پس از این عملیات در پایان از 2000 ویژگی فراتر رفت.

ما برای آموزش و آزمایش به این داده‌ها برچسب‌گذاری شده نیاز داریم، با کارشناسان بخش بازاریابی تماس گرفتیم تا نمونه برچسب‌گذاری شده GSM را در اختیار ما قرار دهند، بنابراین آنها پس از ۲ ماه از داده‌های نه ماهه، مشتریان اعتباری را در فاز غیرفعال در اختیار ما قرار می‌دهند و آنها را به عنوان مشتریان از دست رفته در نظر می‌گیرند. سایر مشتریان غیر از دست رفته به عنوان مشتریان فعال برچسب‌گذاری شدند. (مشتریان اکتسابی در ‎‏ ... آخرین ۴ ماه‌ها هستند مستثنی شده است). مجموع تعداد نمونه‌ای که در آن ۵ میلیون مشتری شامل ۳۰۰۰۰۰ مشتری ریزشی و ۴،۷۰۰،۰۰۰ مشتری فعال وجود دارد. شکل [۱۰](#_bookmark25) دوره‌های داده‌های تاریخی و دوره آتی که ممکن است مشتری شرکت را ترک کند، نشان می‌دهد.

کارشناسان در بازاریابی تصمیم گرفت به پیش‌بینی کردن ‎‏ ... تکان دادن قبل از ۲ ماه‌ها از اقدام واقعی ریزش، به منظور داشتن زمان کافی برای اقدام پیشگیرانه با این مشتریان.

**طبقه‌بندی**

راه حل ما پیشنهادی تقسیم شده ‎‏ ... داده‌ها به دو گروه‌ها: ‎‏ ... آموزش گروه و گروه آزمایش. گروه آموزش شامل ۷۰٪ از مجموعه داده‌ها است و هدف آن آموزش الگوریتم‌ها است. آزمون گروه حاوی ۳۰٪ از ‎‏ ... مجموعه داده‌ها و است استفاده شده به آزمون ‎‏ ... الگوریتم‌ها. پارامترهای الگوریتم‌ها با استفاده از اعتبارسنجی متقابل K-fold بهینه شدند. ارزش از ک بود ۱۰. (یا: ۱۰.) هدف کلاس است نامتعادل، و این می‌توانست علت الف تأثیر منفی قابل توجهی بر مدل‌های نهایی دارد. ما در تحقیق خود با متعادل‌سازی مجدد نمونه آموزش با گرفتن نمونه‌ای از داده‌ها برای متعادل کردن دو کلاس [ [25](#_bookmark57) ]، به این مشکل پرداختیم. ما با نمونه‌گیری بیش از حد با کپی کردن کلاس ریزش برای متعادل کردن با کلاس دیگر شروع کردیم. ما همچنین از روش نمونه‌گیری کمتر تصادفی استفاده کردیم که اندازه نمونه کلاس بزرگ را کاهش می‌دهد تا با کلاس دوم متعادل شود. این روش است ‎‏ ... همان به عنوان ‎‏ ... یکی استفاده شده در بیشتر از یکی تحقیق کاغذها [ [8](#_bookmark40) ، [۲۶](#_bookmark58) ]. آن داد بهترین نتیجه برای برخی الگوریتم‌ها. آموزش نمونه اندازه شد ۴۲۰,۰۰۰.

ما شروع شده آموزش تصمیم درخت الگوریتم و بهینه سازی ‎‏ ... عمق و ‎‏ ...

حداکثر تعداد هایپرپارامترهای گره. ما با چندین مقدار آزمایش کردیم، تعداد بهینه گره‌ها ۳۹۸ گره در درخت و مقدار عمق ۲۰ بود.

الگوریتم جنگل تصادفی نیز آموزش داده شد، ما تعداد درختان هایپرپارامتر را بهینه کردیم. ما با تغییر مقادیر این پارامتر، ساخت مدل را آزمایش کردیم. هر زمان در ۱۰۰، ۲۰۰، ۳۰۰، ۴۰۰ و ۵۰۰ درختان. بهترین نتایج نشان دادن که بهترین شماره از درختان بود ۲۰۰ درختان. افزایش ‎‏ ... شماره از درختان بعد از ۲۰۰ نخواهد کرد دادن الف قابل توجه افزایش در ‎‏ ... عملکرد. جی بی ام الگوریتم بود آموزش دیده و آزمایش شده روی ‎‏ ... همان داده‌ها، ما بهینه شده ‎‏ ... شماره از درختان هایپرپارامتر با مقادیر تا ۵۰۰ درخت. بهترین مقدار پس از آزمایش نیز ۲۰۰ درخت بود. GBM نتایج بهتری ارائه داد نتایج از آر اف و دی تی. ما بالاخره نصب شده ایکس‌جی‌باووست روی جرقه ۲.۳ چارچوب و یکپارچه آن با میلی لیتر کتابخانه در جرقه و اعمال شده ‎‏ ... همان مراحل با سه الگوریتم گذشته. ما همچنین تعداد درختان را بهینه کردیم و بهترین مقدار پس از آزمایش‌های متعدد، ۱۸۰ درخت بود.

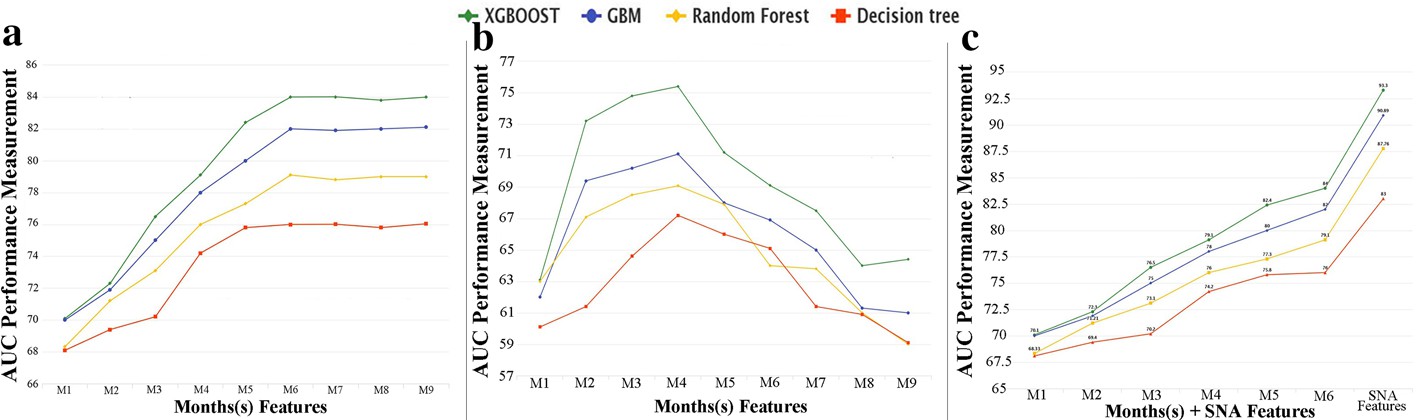
# نتایج و بحث

نتایج برای مقایسه عملکرد در مورد اندازه‌های مختلف مورد تجزیه و تحلیل قرار گرفتند. داده‌های آموزشی. نحوه برخورد با مجموعه داده‌های نامتوازن با استفاده از سه سناریو نیز مورد تجزیه و تحلیل قرار گرفت. اولین نگرانی اصلی در مورد انتخاب پنجره کشویی مناسب برای داده‌ها جهت استخراج ویژگی‌های آماری و SNA بود. در مرحله مهندسی ویژگی‌ها به چه مقدار داده‌های تاریخی نیاز است؟

در شکل [11](#_bookmark26) ، M1 به ماه اول قبل از خط پایه و M9 به ماه نهم اشاره دارد. قبل از خط پایه. ویژگی‌ها از ماه ن هستند تجمیع شده از ‎‏ ... لغزش N ماهه داده‌ها پنجره (از ماه ۱ به ماه ن). همانطور که شکل [۱۱](#_bookmark26) الف ارائه می‌دهد، ما می‌تواند تأیید کردن که افزایش ‎‏ ... حجم از آموزش داده‌ها به دریافت آماری ویژگی‌ها عملکرد الگوریتم‌های طبقه‌بندی را افزایش می‌دهد. با این حال، اضافه کردن قدیمی‌ترین سه ماه‌ها انجام داد نه فراهم کردن هر ارتقاء روی مدل عملکرد. چه زمانی تنها با استفاده از ویژگی‌های آماری، بالاترین مقدار AUC به 84٪ رسید.

اجتماعی شبکه تحلیل ویژگی‌ها داشت الف متفاوت سناریو، چه زمانی ‎‏ ... بهترین پنجره کشویی به ساختن ‎‏ ... اجتماعی گراف و عصاره مناسب اس ان ای ویژگی‌ها بود در طول چهار ماه آخر قبل از خط پایه، همانطور که در شکل [11](#_bookmark26) ب نشان داده شده است. افزودن داده‌های قدیمی‌تر، عملکرد مدل را به طور نامطلوبی تحت تأثیر قرار خواهد داد. بالاترین مقدار AUC که تنها با استفاده از ویژگی‌های SNA به دست آمد، 75.3٪ بود.

بسته به دو سناریوی مختلف فوق، 6 ماه آخر مجموعه داده‌های خام عبارت بود از استفاده شده به عصاره ‎‏ ... آماری ویژگی‌ها، در حالی که ‎‏ ... آخرین چهار ماه‌ها از که مجموعه داده‌ها بود



**Fig. 11** Performance of classification algorithms per sliding window and feature type. Panel (**a**) shows the improvement of churn predictive model using Statistical Features related to different historical periods, panel

(**b**) presents the changes in predictive model improvement using SNA Features related to the same historical periods, and panel (**c**) presents the enhancement of churn predictive model when using both statistical and SNA Features

**میز ۲ مقایسه AUC نتایج قبل از و بعد از اضافه کردن اس ان ای به آماری ویژگی‌ها**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ویژگی‌ها** | **ایکس‌جی‌باووست (%)** | **جی‌اس‌ام (ب) (%)** | **تصادفی جنگل (%)** | **تصمیم درخت (%)** |
| آماری ویژگی‌ها | ۸۴ | ۸۲ | ۷۹.۱ | ۷۶ |
| ویژگی‌های SNA | ۷۵.۳ | ۷۱ | ۶۹ | ۶۷.۲ |
| آماری و اس ان ای ویژگی‌ها | ۹۳.۳ | ۹۰.۸۹ | ۸۷.۷۶ | ۸۳ |

فقط برای استخراج ویژگی‌های SNA استفاده می‌شود. شکل [10](#_bookmark25) بهترین پنجره کشویی را برای استخراج ارائه می‌دهد اس ان ای ویژگی‌ها در نارنجی و ‎‏ ... آبی یکی است برای آماری ویژگی‌ها در حالی که ‎‏ ... خط قرمز نشان دهنده خط پایه است.

با افزودن ویژگی‌های SNA به همراه ویژگی‌های آماری به الگوریتم‌های طبقه‌بندی، نتایج به طور قابل توجهی افزایش یافت. همانطور که در جدول [2 ارائه شده است.](#_bookmark27) و شکل [11c](#_bookmark26) ، افزودن هر دو نوع ویژگی، بهبود خوبی در عملکرد مدل پیش‌بینی ریزش ایجاد کرد، که در آن حداکثر مقدار AUC به دست آمده 93.3٪ بود.

دومین نگرانی که مورد توجه قرار گرفت، مشکل مجموعه داده‌های نامتوازن بود، زیرا سه آزمایش برای همه الگوریتم‌های طبقه‌بندی اعمال شد. این آزمایش‌ها عبارتند از: (1) طبقه‌بندی با تکنیک نمونه‌گیری کمتر از حد، (2) طبقه‌بندی با تکنیک نمونه‌گیری بیشتر از حد، (3) طبقه‌بندی بدون متعادل‌سازی مجموعه داده‌ها. جدول [3](#_bookmark28) نشان می‌دهد که هر دو الگوریتم XGBOOST و GBM بدون هیچ گونه ... بهترین عملکرد را ارائه داده‌اند. متعادل سازی مجدد تکنیک‌ها، در حالی که تصادفی جنگل و تصمیم درخت الگوریتم‌ها داد عملکرد بالاتر با استفاده از تکنیک‌های نمونه‌برداری کمتر.

همانطور که در شکل [11 نشان داده شده است](#_bookmark26) و با توجه به جداول [۲](#_bookmark27) و [۳](#_bookmark28) ، ما تأیید می‌کنیم که الگوریتم XGBOOST عملکرد بهتری داشت ‎‏ ... استراحت از ‎‏ ... آزمایش شده الگوریتم‌ها با یک AUC ارزش از ۹۳.۳٪ به طوری که به عنوان الگوریتم طبقه‌بندی در این مدل پیش‌بینی پیشنهادی انتخاب شده است. جی بی ام الگوریتم اشغال شده دوم مکان با یک AUC ارزش از ۹۰.۸۹٪ در حالی که تصادفی جنگل و تصمیم درختان آمد آخرین در AUC رتبه بندی با ارزش‌ها از ۸۷.۷۶٪ و ۸۳٪ به ترتیب. شکل [۱۲](#_bookmark29) منحنی‌های ROC را برای چهار الگوریتم نشان می‌دهد.

ویژگی‌های مهم و برتر که در پیش‌بینی ریزش مشتری نقش دارند، با استفاده از معیار بهره [ [27](#_bookmark59) ] رتبه‌بندی شدند. ارزش بهره بالای ویژگی به معنای اهمیت بیشتر آن در ... است. پیش‌بینی ‎‏ ... تکان دادن مهم ویژگی‌ها طبق به ایکس‌جی‌باووست الگوریتم ارائه می‌شوند در شکل [۱۳](#_bookmark30) قبل از و بعد از ادغام اس ان ای و آماری ویژگی‌ها.

همانطور که در شکل [۱۳ نشان داده شده است،](#_bookmark30) اضافه شدن ویژگی‌های تحلیل شبکه‌های اجتماعی، رتبه‌بندی را تغییر داد از ‎‏ ... مهم ویژگی‌ها. ام تی ان کسینوس شباهت بود ‎‏ ... بیشترین ویژگی مهم از آنجایی که ‎‏ ... مشتریان با بالاتر ام تی ان کسینوس شباهت هستند بیشتر احتمالاً به تکان دادن

**میز ۳ مقایسه نتایج AUC الگوریتم‌های یادگیری ماشین با هر تکنیک متعادل‌سازی**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **تکنیک استفاده شده**  **برای نامتعادل مجموعه داده‌ها** | **ایکس‌جی‌باووست (%)** | **جی‌اس‌ام (ب) (%)** | **تصادفی جنگل (%)** | **تصمیم درخت (%)** |
| نمونه‌برداری بیش از حد | ۹۲ | ۹۰.۰۱ | ۸۴.۲ | ۷۶.۲۵ |
| نمونه‌برداری کمتر از حد | ۹۳.۱۲ | ۹۰.۲۱ | *۸۷.۷۶* | *۸۳* |
| بدون متعادل سازی | *۹۳.۳* | *۹۰.۸۹* | ۷۸.۴۷ | ۷۲.۲ |

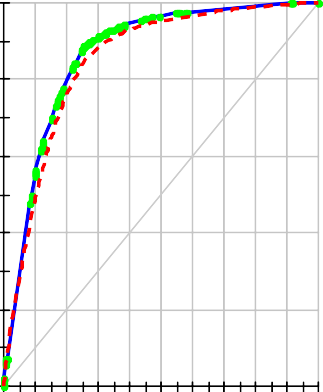
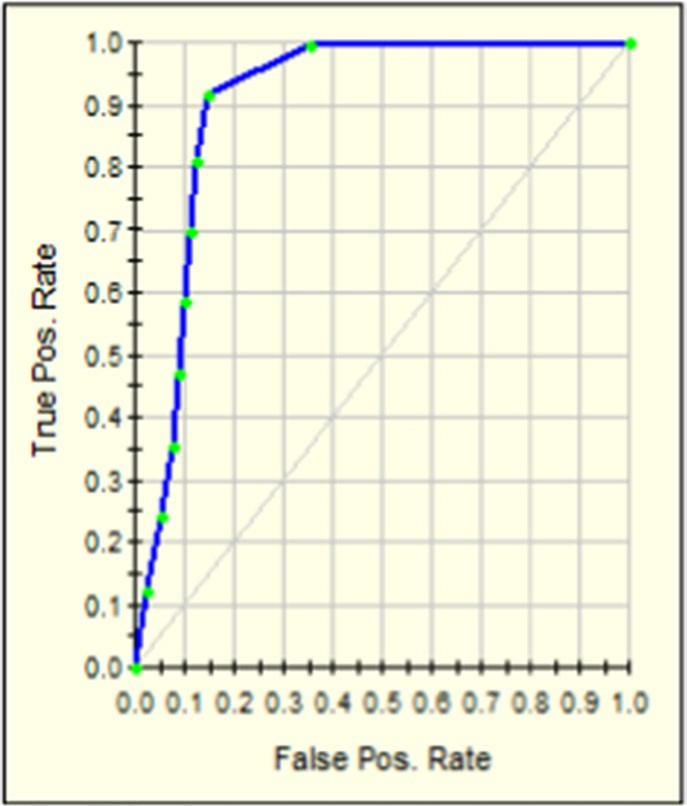
همانطور که در جدول نشان داده شده است، هر الگوریتم یادگیری ماشین با سه سناریوی مختلف در رابطه با مشکل مجموعه داده‌های نامتوازن بهترین نتایج AUC که با حروف کج (ایتالیک) ارائه شده‌اند، بهترین تکنیکی را که با هر یک از موارد مطابقت دارد، نشان می‌دهند. الگوریتم

True Pos. Rate

True Pos. Rate

True Pos. Rate

صرف نظر از این از ‎‏ ... دیگر ویژگی‌ها مانند تعادل، اینترنت کاربرد، و ورود/خروج تماس‌ها شکل شکل [9a](#_bookmark23) توزیع این ویژگی را نشان می‌دهد.



**DT** AUC 83 %

RF AUC 87.76%

GBM AUC 90.8%

XGBOOST AUC %93.3

**Fig. 12** Receiver operating characteristic curve for each classification algorithm

1.0

0.9

0.8

0.7

0.6

0.5

0.4

0.3

0.2

0.1

0.0

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

False Pos. Rate

1.0

0.9

0.8

0.7

0.6

0.5

0.4

0.3

0.2

0.1

0.0

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

False Pos. Rate

1.0

0.8

0.6

0.4

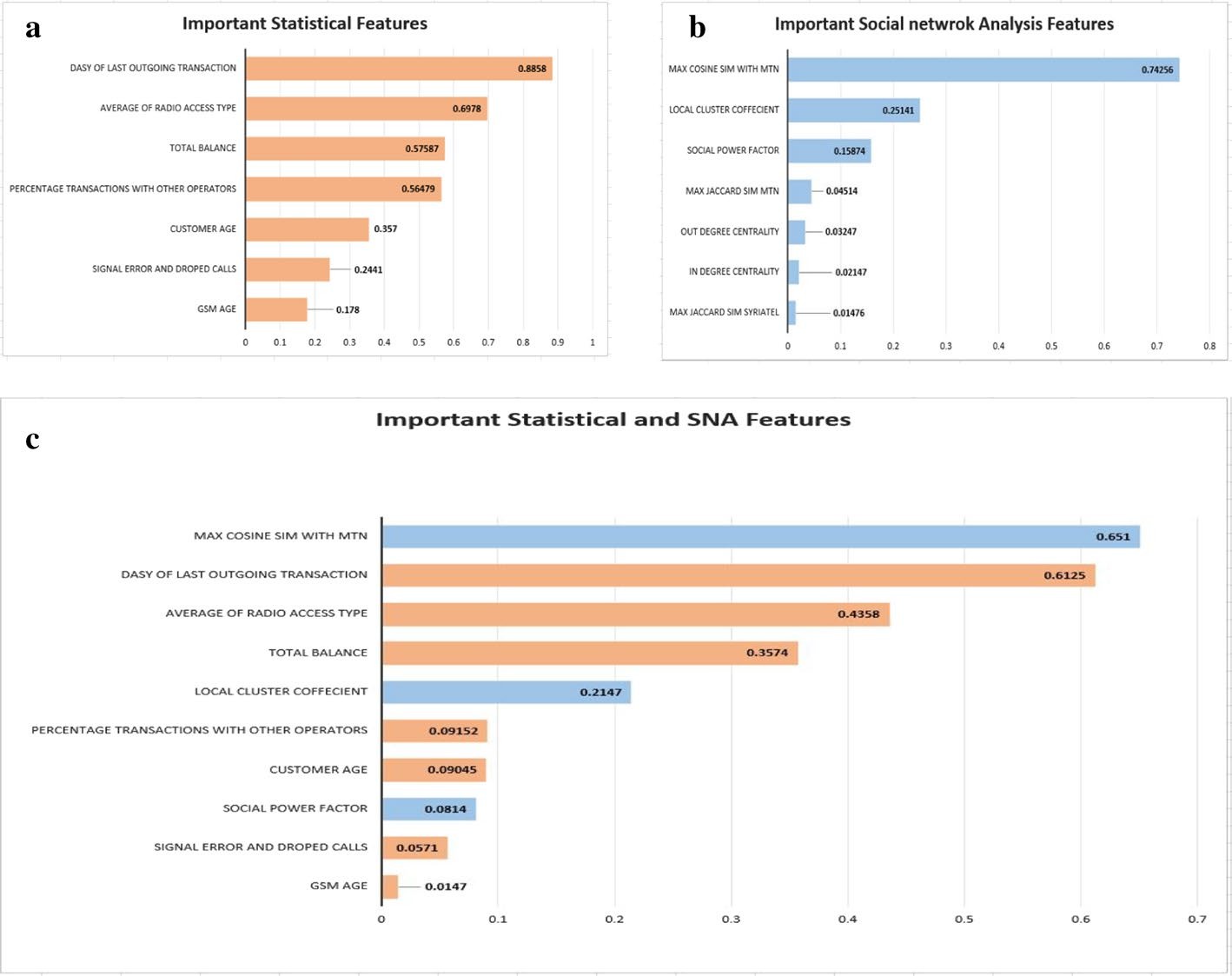
0.2

0.0

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

False Pos. Rate

توسط تحلیل کردن این ویژگی، بیشترین از ‎‏ ... مشتریان به طور کلی ماند فعال برای الف دوره زمان قبل از خاتمه دادن یا متوقف کردن ‎‏ ... استفاده از آنها جی اس ام ها. این مورد احتمالاً به این دلیل اتفاق می‌افتد که مشتری باید مطمئن شود که بیشتر تماس‌های ورودی مهم او و مخاطبین داشته باشند نقل مکان کرد به ‎‏ ... جدید خط. در دیگر کلمات، ‎‏ ... مشتری می‌توانست مدتی صبر کند تا مطمئن شود که بیشتر افراد مهمش از اطلاعات جدید مطلع شده‌اند. جی‌اس‌ام شماره. این مورد همچنین می‌توانست باش موجه به عنوان ‎‏ ... مشتری نیاز به پایان باقی مانده تعادل در ‎‏ ... جی‌اس‌ام قبل از او توقف‌ها با استفاده از آن. شکل [۱۴](#_bookmark31) نشان می‌دهد یک مثال از ردیابی شده چرخیده مشتری. این شکل ارائه می‌دهد ‎‏ ... فازها از در حال حرکت او جامعه به GSM اپراتور دیگر.



**Fig. 13** Important features per type according to XGBOOST, panel (**a**) shows the ranking of important Statistical features, panel (**b**) shows the ranking of important Social Network Analysis features, while panel (**c**) shows how adding both Statistical and SNA features re-ranks their importance in XGBOOST algorithm. The orange color is used in all panels to represent the Statistical features and the blue one for SNA features



**Fig. 14** Tracked churned customer with Max Cosine MTN Similarity values per week

مشتری در هفته هفتم از رقیب GSM خریداری کرد و در هفته چهاردهم GSM شرکت SyriaTel را قطع کرد و سپس در هفته‌های سیزدهم و چهاردهم پوشش‌دهی آن قطع شد. نتیجه تشابه کسینوسی نیز در همان شکل نمایش داده شده است.

دومین ویژگی مهم، تعداد روزهای آخرین تراکنش خروجی است. همانطور که در شکل [7](#_bookmark19) الف نشان داده شده است، اکثر مشتریان از دست رفته، مدت طولانی‌تری نسبت به مشتریان غیر از دست رفته بدون انجام هیچ تراکنشی در سایت می‌مانند.

سوم مهم ویژگی است مجموع تعادل از آنجایی که بیشترین چرندگوها داشت کم تعادل مقایسه شده با ‎‏ ... فعال مشتریان صرف نظر از از ‎‏ ... دلیل از تکان خوردن، شکل [۷](#_bookmark19) ج نشان می‌دهد ‎‏ ... توزیع کل موجودی برای مشتریان از دست رفته و مشتریان غیر از دست رفته.

چهارم ویژگی در اهمیت است میانگین از رادیو دسترسی نوع کجا بیشترین از از کوره در رفتن‌ها داشت بیشتر 2G اینترنت جلسات از 3G جلسات، به عنوان ‎‏ ... سرعت و کیفیت از اینترنت 2G جلسات است زیاد کمتر از اینها از 3G جلسات شکل [۷](#_bookmark19) ب نشان می‌دهد ‎‏ ... توزیع از این ویژگی کجا ‎‏ ... میانگین موش صحرایی است پایین‌تر برای بیشترین از ‎‏ ... چرندگوها در مقایسه با مشتریانی که از خرید خود منصرف نشده‌اند. اگر مشتریان به طور مداوم از اینترنت استفاده کنند، احتمال ریزش آنها بیشتر است. کاربران و آنجا است الف بهتر 3G پوشش ارائه شده توسط ‎‏ ... رقیب توسط با تجزیه و تحلیل این ویژگی، ۶۸٪ از ریزش‌کنندگان، کاربران اینترنت هستند که ۶۵٪ از آنها میانگین نوع دسترسی رادیویی پایینی دارند.

ضریب خوشه محلی یکی دیگر از ویژگی‌های SNA است که از نظر اهمیت در پیش‌بینی ریزش مشتری در رتبه پنجم قرار دارد، زیرا مشتریانی که ارزش LCC بسیار پایینی دارند، همانطور که در شکل 9b نشان داده شده است، [احتمال](#_bookmark23) ریزش کمتری دارند. این امر می‌تواند توجیه‌پذیر باشد زیرا برخی از مشتریان از این GSMهای شخصی برای اهداف تجاری استفاده می‌کنند. آنها باید شماره‌های خود را حفظ کنند تا هیچ یک از مشتریان خود را از دست ندهند. اکثر این مشتریان بیش از دو GSM دارند. آنها با یکدیگر ارتباط برقرار می‌کنند. با تعداد زیادی از مردم، بیشترین از اینها مردم نکن بدانید هر کدام دیگر (آنجا هیچ تعاملی بین آنها وجود ندارد). برای بررسی این مورد با نمونه‌ای از مشتریان با LCC بسیار پایین تماس گرفته شد. نتایج نشان می‌دهد که بیشتر آنها مربوط به کافه‌ها، رستوران‌ها و... بودند. غرغرها، مغازه‌های اصلاح صورت، آرایشگاه‌ها، کتابخانه‌ها، فروشگاه‌های حیوانات خانگی، کلینیک‌های پزشکی و جاهای دیگر.

ششمین ویژگی مهم، درصد تراکنش‌ها به/از سایر اپراتورها است، این مقدار برای کاربران غیرفعال بیشتر می‌شود. توضیح در اینجا این روش بر تأثیر دوستان بر تصمیم به ترک مشتری متکی است، زیرا وابستگی اکثر دوستان مشتری به اپراتور دیگر ممکن است گواهی بر شهرت خوب یا وجود قوی شرکت رقیب در آن منطقه یا جامعه باشد. بنابراین، این می‌تواند منجر به این شود که مشتری تحت تأثیر محیط اطراف قرار گیرد، بنابراین به سمت رقیب حرکت می‌کند. شرکت. بالاتر ارزش از این ویژگی ممکن است افزایش ‎‏ ... احتمال از ریزش مشتری، شکل [7d](#_bookmark19) توزیع این ویژگی را نشان می‌دهد.

ویژگی‌های دیگری مانند سن مشتری نیز از نظر اهمیت در رتبه هفتم قرار دارند. ‎‏ ... مشتریان سازمان بهداشت جهانی هستند کمتر از ۳۲ سال‌ها قدیمی داشته باشند بیشتر احتمال به تکان دادن نسبت به دیگران این می‌تواند باش توضیح داده شده توسط ‎‏ ... واقعیت که جوان مردم هستند همیشه به دنبال بهترین‌ها برای برآورده کردن نیازهایشان در قالب خدمات بهتر، با کیفیت بالاتر و ارزان‌تر هستند از آنجایی که حجم ارتباطات، استفاده از اینترنت و سایر خدمات بسیار بیشتر است مقایسه شده به خدمات از مشتریان از متفاوت سنین شکل [۶](#_bookmark18) نشان می‌دهد ‎‏ ... توزیع این ویژگی با توجه به کلاس ریزش. ویژگی ضریب توان اجتماعی، سومین ویژگی SNA است که یکی از مهم‌ترین ویژگی‌ها برای پیش‌بینی ریزش محسوب می‌شود. همانطور که در شکل [9c نشان داده شده است](#_bookmark23) ، مقدار ضریب توان بالاتر به معنای احتمال ریزش کمتر است. همانطور که در شکل [7e نیز نشان داده شده است](#_bookmark19) ، مشتریانی که خطاهای سیگنال بالا و تماس‌های از دست رفته دارند، بیشترین احتمال ریزش را دارند.

با توجه به مواردی که ذکر شد قبلاً و همانطور که در شکل‌ها نشان داده شده است. [۱۱](#_bookmark26) ج، [۱۲](#_bookmark29) ج، ما معتقدیم که ویژگی‌های تحلیل شبکه اجتماعی سهم خوبی در افزایش عملکرد مدل پیش‌بینی ریزش دارند، زیرا از دیدگاه اجتماعی، بینش متفاوتی به مشتری می‌دهند.

**میز ۴ AUC نتایج برای طبقه بندی الگوریتم‌ها روی «پیشنهاد نمی‌شود» مجموعه داده‌ها**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **الگوریتم** | **ایکس‌جی‌باووست** | **جی‌اس‌ام (ب)** | **تصادفی جنگل** | **تصمیم درخت** |
| سوریه تل جدید داده‌ها «پیشنهاد نمی‌شود» | ۸۹٪ | ۸۵.۵٪ | ۸۳.۴٪ | ۷۹.۱٪ |

- *سیستم ارزیابی* ما ارزیابی شده تفل سیستم توسط با استفاده از جدید بالا به تاریخ مجموعه داده‌ها. این آزمایش بر روی تمام مشتریان پیش‌پرداخت سیریاتل بدون هیچ استثنایی انجام شد. جمعیت بود ۷.۵ میلیون مشتریان ویتفلوت دانستن فلت تفلیر وضعیت اراده باش بعد از ۲ ماه. مدل‌های مشابه پس از پردازش همانطور که قبلاً ذکر شد، روی مجموعه داده‌های tflis آزمایش شدند. مجموعه داده‌ها برای مشتریانی که در حال حاضر در حال تغییر هستند، به احتمال زیاد پیش‌بینی می‌شود که تغییر کنند. بود تقسیم شده به دو مجموعه داده‌ها (ارائه شده، ارائه نشده است). بازاریابی کارشناسان برای حفظ مشتریانی که پیش‌بینی می‌شود از مجموعه داده‌های ارائه شده، SyriaTel را ترک کنند، اقدامی پیشگیرانه انجام دهید و مجموعه داده‌های "ارائه نشده" را بدون هیچ اقدامی ترک کنید. نتایج آزمون Tfle با وضعیت مشتری پس از دو ماه مقایسه شد. دو مجموعه داده را بررسی کنید. نتایج بسیار خوب بودند و بهترین مقدار AUC برای XGBOOST روی «ارائه نشده» ۸۹٪ بود و اکثر موارد پیش‌بینی شده صحیح بودند. جدول [۴](#_bookmark32) نتایج AUC از sflows برای چهار الگوریتم روی مجموعه داده «پیشنهاد نشده».

درصد از تفل حفظ شده مشتریان از ارائه شده مجموعه داده‌ها بود درباره ۴۷٪ از همه مشتریان پیش‌بینی شده به کفلورن در اوتفلر کلمات، درباره فلاف از تفل مشتریان در مورد تصمیم خرید دچار تردید شدند، اما وقتی تفلی پیشنهاد خوبی دریافت کرد، ذهنشان درگیر شد. این نتیجه برای شرکت بسیار خوب بود، درآمد شرکت را افزایش و نرخ خرید را حدود ۵ درصد کاهش داد.

# نتیجه‌گیری

اهمیت این نوع تحقیق در بازار مخابرات، کمک به شرکت‌ها برای کسب سود بیشتر است. مشخص شده است که پیش‌بینی ریزش مشتریان یکی از مهم‌ترین منابع درآمد شرکت‌های مخابراتی است. از این رو، این تحقیق با هدف ساخت سیستمی انجام شد که ریزش مشتریان در شرکت مخابرات سیریاتل را پیش‌بینی کند. این مدل‌های پیش‌بینی باید به مقادیر بالای AUC دست یابند. برای آزمایش و آموزش مدل، ‎‏ ... نمونه داده‌ها است تقسیم شده به ۷۰٪ برای آموزش و ۳۰٪ برای آزمایش کردن. ما تصمیم گرفتیم اعتبارسنجی متقابل را با 10-fold برای اعتبارسنجی و بهینه‌سازی فراپارامتر انجام دهیم. ما از مهندسی ویژگی، تبدیل ویژگی مؤثر و رویکرد انتخاب برای آماده‌سازی ویژگی‌ها برای الگوریتم‌های یادگیری ماشین استفاده کرده‌ایم. علاوه بر این، با مشکل دیگری مواجه شدیم: داده‌ها متعادل نبودند. تنها حدود 5٪ از ورودی‌ها نشان‌دهنده ریزش مشتریان هستند. این مشکل با نمونه‌گیری کمتر از حد معمول یا استفاده از الگوریتم‌های درختی که تحت تأثیر این مشکل قرار نمی‌گیرند، حل شد. چهار الگوریتم مبتنی بر درخت انتخاب شدند. انتخاب شده زیرا از آنها تنوع و کاربردپذیری در این نوع از پیش‌بینی. این الگوریتم‌ها عبارتند از درخت تصمیم، جنگل تصادفی، الگوریتم درخت GBM و XGBOOST الگوریتم. روش از آماده سازی و انتخاب از ویژگی‌ها و وارد کردن ویژگی‌های شبکه اجتماعی موبایل بیشترین تأثیر را بر موفقیت این مدل داشت، زیرا مقدار AUC در SyriaTel به ۹۳.۳۰۱٪ رسید. مدل درختی XGBOOST در تمام اندازه‌گیری‌ها بهترین نتایج را به دست آورد. مقدار AUC برابر با ۹۳.۳۰۱٪ بود. الگوریتم GBM در جایگاه دوم قرار دارد و جنگل تصادفی و درخت تصمیم در رتبه‌های بعدی قرار گرفته‌اند.

سوم و چهارم در خصوص AUC ارزش‌ها. ما داشته باشند ارزیابی شده ‎‏ ... مدل‌ها توسط مناسب سازی الف با توجه به مجموعه داده‌های جدید مربوط به دوره‌های مختلف و بدون هیچ اقدام پیشگیرانه‌ای از سوی بازاریابی، XGBOOST نیز با ۸۹٪ AUC بهترین نتیجه را ارائه داد. کاهش نتیجه می‌تواند به دلیل پدیده مدل داده‌های غیر ایستا باشد، بنابراین مدل در هر دوره زمانی نیاز به آموزش دارد.

استفاده از ‎‏ ... اجتماعی شبکه تحلیل ویژگی‌ها افزایش ‎‏ ... نتایج از پیش‌بینی ریزش نیرو در مخابرات.

**اختصارات**

CDR: ثبت جزئیات تماس؛ CRM: مدیریت ارتباط با مشتری؛ SMS: سرویس پیام کوتاه؛ HDFS: Hadoop Distributed فایل سیستم؛ ایکس‌جی‌بی‌بوست: افراطی گرادیان تقویت کننده؛ فرکانس رادیویی: تصادفی جنگل؛ دی تی: تصمیم درخت؛ AUC: مساحت زیر ‎‏ ... منحنی؛ جی اس ام: سیستم جهانی ارتباطات سیار؛ IMEI: هویت بین‌المللی تجهیزات موبایل؛ SNA: تحلیل شبکه اجتماعی؛ جی بی ام: گرادیان تقویت‌شده ماشین؛ بازگشت سرمایه: بازگشت روی سرمایه‌گذاری؛ ROC: گیرنده عملیاتی مشخصه؛ پی سی ای: مدیر اصلی تحلیل مؤلفه؛ RAM: حافظه دسترسی تصادفی؛ MMS: سرویس پیام‌رسانی چندرسانه‌ای؛ CSV: مقادیر جدا شده با کاما؛ JSON: نمادگذاری شیء جاوا اسکریپت؛ XML: زبان نشانه‌گذاری توسعه‌پذیر.

**نویسندگان مشارکت‌ها**

معروف به گرفت ‎‏ ... نقش از اجرا ‎‏ ... ادبیات بررسی، ساختمان ‎‏ ... بزرگ داده‌ها پلتفرم، کار کردن روی ‎‏ ... پیشنهادی تکان دادن مدل. او انجام شده ‎‏ ... آزمایش‌ها و نوشت ‎‏ ... نسخه خطی. ای جی و کی جی گرفت روی الف نظارتی نقش و نظارت کرد ‎‏ ... تکمیل از ‎‏ ... کار کردن. همه نویسندگان خواندن و تایید شده ‎‏ ... نهایی نسخه خطی.

**تقدیرنامه‌ها**

با تشکر فراوان از سیریتل، خانم مجدا ساکر، آقای مرید عطاسی و آقای ادهم ترودی برای حمایت و انگیزه. ممنون برای آقای محمد عساف، آقای نور الملحم، آقای ویلیام سلیمان، آقای عمار اسد، آقای سلیمان مولا، آقای احمد علی، و خانم مروه هانهون برای آنها همکاری و کمک کنید. ممنون به آقای کاسم جمیل احمد و آقای فهمی عمارین برای مرور کردن ‎‏ ... نهایی نسخه از این کاغذ.

**رقابت علایق**

نویسندگان اعلام کردن که آنها داشته باشند خیر رقابت کردن علایق.

**در دسترس بودن از داده‌ها و مواد**

این داده‌ها به دلیل محدودیت‌های اعمال شده از سوی شرکت SyriaTel Telecom در دسترس عموم نیست، زیرا مجوز برای این مطالعه اعطا شد. این داده‌ها در دسترس محققان شرکت سیریاتل است و برای دیگران نیز در دسترس خواهد بود. بعد از گرفتن ‎‏ ... اجازه از ‎‏ ... شرکت.

**رضایت برای انتشار**

نویسندگان رضایت برای انتشار.

**اخلاق تصویب و رضایت به شرکت کردن**

همه نویسندگان دادن اخلاق تصویب و رضایت به شرکت کردن در تسلیم و بررسی فرآیند.

**بودجه**

نویسندگان اعلام کردن که آنها داشته باشند خیر تأمین مالی.

# ناشر توجه داشته باشید

اسپرینگر طبیعت باقی مانده است خنثی با توجه به قضایی ادعاها در منتشر شده نقشه‌ها و نهادی وابستگی‌ها

دریافت شده: ۳۰ دسامبر ۲۰۱۸ پذیرفته شده: ۸ مارس ۲۰۱۹



**منابع**

1. گرپوت تی‌جی، رامز دبلیو، شیندلر ای. حفظ مشتری، وفاداری و رضایت در صنعت تلفن همراه آلمان بازار مخابرات. سیاست مخابرات. 2001؛ 25: 249–69.
2. وی سی پی، چیو آی تی. تبدیل جزئیات تماس‌های مخابراتی به پیش‌بینی ریزش: یک رویکرد داده‌کاوی. متخصص سیستم کاربردی. 2002؛ 23(2):103–12.
3. قریشی اس. ای.، رحمان اس.، قمر ای. ام.، کمال ای.، رحمان ای. مشترکین مخابرات پیش‌بینی ریزش مدل با استفاده از دستگاه یادگیری. در: هشتم بین‌المللی کنفرانس روی دیجیتال اطلاعات مدیریت. ۲۰۱۳. (۱۳۹۲) ص. ۱۳۱–۶.
4. آسکارزا ای، اینگار آر، شلایشر ام. خطرات پیشگیری پیشگیرانه از ریزش مشتری با استفاده از توصیه‌های برنامه: شواهدی از یک آزمایش میدانی. J Market Res. 2016;53(1):46–60.
5. بوت پیش‌بینی مشتری تکان دادن در مخابرات صنعت با استفاده از چند لایه پرسپترون عصبی شبکه‌ها: مدل‌سازی و تحلیل. ایگارس. 2014;11(1):1–5.
6. اومایاپارواتی وی، ایاکوتی کی. بررسی پیش‌بینی ریزش مشتری در صنعت مخابرات: مجموعه داده‌ها، روش‌ها و متریک. مجله بین‌المللی فناوری مهندسی. 2016؛ 3(4): 1065–70.
7. یو دبلیو، جوتلا دی‌ان، سیواکومار اس‌سی. مدلی برای همسویی استراتژی ریزش مشتری برای مدیران در مخابرات موبایل. در: ارتباطات کنفرانس تحقیقاتی شبکه‌ها و خدمات کاتیونی، جلد ۳، ۲۰۰۵، صفحات ۴۸ تا ۵۳.
8. بورز دی، دن پول V. مدیریت عدم تعادل کلاس در پیش‌بینی ریزش مشتری. برنامه کاربردی سیستم خبره. 2009;36(3):4626–36.
9. ژان جی، گیدیبنده وی، پارسا اس پی کی. شناسایی جوامع تأثیرگذار برتر در شبکه‌های بزرگ. مجله کلان داده. 2016;3(1):16. <https://doi.org/10.1186/s40537-016-0050-7>.
10. بارتلمی ام. مرکزیت بینابینی در شبکه‌های پیچیده بزرگ. مجله فیزیک اروپا. 2004؛ 38(2):163–8. [https://doi.](https://doi.org/10.1140/epjb/e2004-00111-4) [org/10.1140/epjb/e2004-00111-4](https://doi.org/10.1140/epjb/e2004-00111-4) .
11. الیزابتا ای، مایرهنک اچ، استودت سی ال. تقریب مرکزیت بینابینی در شبکه‌های بزرگ در حال تکامل. CoRR. ۲۰۱۴. (۱۳۹۳) [arxiv:1409.6241](http://arxiv.org/abs/1409.6241) . (آرکایو:۱۴۰۹.۶۲۴۱)
12. برندوسویو آی، تودرین جی، ها بی. روش‌هایی برای پیش‌بینی ریزش در صنعت مخابرات سیار پیش‌پرداخت امتحان کنید. در: کنفرانس بین‌المللی ارتباطات. ۲۰۱۶. ص. ۹۷–۱۰۰.
13. هی وای، هی زد، ژانگ دی. مطالعه‌ای بر روی پیش‌بینی ریزش مشتری در شبکه ارتباطات ثابت بر اساس داده‌ها استخراج. در: ششمین کنفرانس بین‌المللی سیستم‌های فازی و کشف دانش، جلد ۱. ۲۰۰۹. ص. ۹۲–۹۴.
14. ادریس آ، خان آ، لی YS. پیش‌بینی ریزش مبتنی بر برنامه‌نویسی ژنتیک و adaboosting برای مخابرات. در: IEEE کنفرانس بین‌المللی سیستم‌ها، انسان و سایبرنتیک (SMC). ۲۰۱۲. صفحات ۱۳۲۸–۳۲.
15. هوانگ اف، ژو ام، یوان کی، دنگ ای او. پیش‌بینی ریزش مشتریان شرکت‌های مخابراتی با استفاده از کلان‌داده. در: کنفرانس بین‌المللی ACM SIGMOD تأثیر بر مدیریت داده‌ها. ۲۰۱۵. صفحات ۶۰۷-۱۸.
16. مختار م، نفیس س، محمد م، آونگ م، رحمان م، دریس م. مدل طبقه‌بندی ریزش مشتریان برای شرکت‌های مخابراتی محلی شرکت مخابراتی مبتنی بر نظریه مجموعه‌های تقریبی. مجله علوم کاربردی فوندام. 2017؛9(6):854–68.
17. امین الف، انور اس، عدنان الف، نواز م، هاوارد ن، قادر جی، حواله الف، حسین الف. مقایسه نمونه‌گیری بیش از حد تکنیک‌هایی برای مدیریت مشکل عدم تعادل کلاس: مطالعه موردی پیش‌بینی ریزش مشتری. IEEE Access. ۲۰۱۶؛ ۴:۷۹۴۰–۵۷.
18. چاولا ن. داده‌کاوی برای مجموعه داده‌های نامتوازن: یک مرور کلی. در: داده‌کاوی و کشف دانش دستی کتاب. برلین: اشپرینگر؛ ۲۰۰۵. صص. ۸۵۳–۸۶۷.
19. پیج ال، برین اس، موتوانی آر، وینوگراد تی. رتبه‌بندی استناد پیج‌رنک: نظم بخشیدن به وب. استنفورد دیجیتال کتابخانه فناوری‌ها پروژه. ۱۹۹۸ میلادی. ص. ۱۷.
20. کیس سی، بیچلر ام. شناسایی افراد تأثیرگذار - اندازه‌گیری نفوذ در شبکه‌های مشتری. سیستم پشتیبانی دسیس. 2008؛ 46(1):233–53.
21. کیس سی، بیچلر ام. شناسایی افراد تأثیرگذار - اندازه‌گیری نفوذ در شبکه‌های مشتری. سیستم پشتیبانی دسیس. 2008؛ 46(1):233–53. <https://doi.org/10.1016/j.dss.2008.06.007>.
22. برین اس، صفحه ل. آناتومی یک مقیاس بزرگ وب فرامتنی موتور جستجو. شبکه کامپیوتری سیستم ISDN 1998;30(1–7):107–17. <https://doi.org/10.1016/S0169-7552(98)00110-X>.
23. ژائو یو، وانگ جی، یو پی‌نوشت، لیو اس، ژانگ س. استنباط اجتماعی نقش‌ها و وضعیت‌ها در اجتماعی شبکه‌ها. در: کی دی دی ۲۰۱۳—نوزدهم کنفرانس بین‌المللی ACM SIGKDD در زمینه کشف دانش و داده‌کاوی. انجمن محاسبات ماشین‌آلات؛ ۲۰۱۳. صص ۶۹۵–۷۰۳.
24. لسکووک جی، بکستروم ال، کومار آر، تامکینز ای. تکامل میکروسکوپی شبکه‌های اجتماعی. در: کنفرانس بین‌المللی در باب کشف دانش و داده‌کاوی. KDD؛ 2008. صص. 695–703.
25. لی ی، لو پ، وو سی. الف جدید شبکه گره شباهت اندازه گیری روش و آن برنامه های کاربردی. ۲۰۱۴. (۱۳۹۳) [آرکایو: ۱۴۰۳۴۳۰۳](http://arxiv.org/abs/14034303) .
26. شی جی، روجکووا وی، پال اس، کاگشال اس. ترکیبی از بوستینگ و بگینگ برای جام KDD 2009 - امتیازگیری سریع در ... پایگاه داده بزرگ. J Mach Learn Res Proc Track. 2009;7:35–43.
27. چن تی، گاسترین سی. ایکس‌جی‌بوست الف مقیاس‌پذیر درخت تقویت سیستم. کور آر آر ۲۰۱۶. (۱۳۹۵) [arXiv:1603.02754](http://arxiv.org/abs/1603.02754)

