# توضیحات روش کار مقاله: بررسی تأثیر ارزش طول عمر مشتری در وفاداری مشتریان

## شرح مدل

ما در این مقاله قصد داریم میزان ارتباط میان دو شاخص مهم در رتبه بندی مشتریان را مورد بررسی قرار دهیم. این دو شاخص به ترتیب ارزش طول عمر مشتری و وفاداری مشتریان هستند. در اکثر پژوهشات علمی تنها به یکی از این دو شاخص توجه شده است و هر کدام به طور جداگانه تحلیل شدند. اما شاید بتوان با در نظر ارتباط و تأثیر پذیری آن‌ها از یکدیگر رتبه بندی بهتری از مشتریان ایجاد نمود. به این منظور لازم است ابتدا وجود یا عدم وجود رابطه مثبت یا منفی بین آن‌ها بررسی شود. در این مقاله منظور از رابطه بین دو شاخص وفاداری و ارزش طول عمر مشتری این است که آیا رتبه بندی مشتریان طبق شاخص وفاداری عیناً همان رتبه بندی را طبق شاخص ارزش طول عمر مشتری ایجاد می‌کند؟ به عبارت دیگر آیا اگر یک مشتری طبق شاخص وفاداری، به عنوان مشتری وفادار لحاظ شود، از نظر ارزش طول عمر مشتری نیز سودمندی بالایی برای شرکت ایجاد خواهد کرد؟ و یا برعکس اگر مشتری ارزش طول عمر بالایی داشته باشد لزوماً مشتری وفاداری برای شرکت خواهد بود؟ بنابراین فرضیه ما وجود رابطه مثبت یا منفی میان این دو شاخص است. در این مقاله قصد آزمودن این فرضیه را داریم.

لازمه حصول نتایجی با دقت بالا و نزدیک به واقعیت به کارگیری داده‌های جمع‌اوری شده  در یک شرکت مشخص با حجم قابل قبولی از داده‌ها در بازه زمانی طولانی مدت (به طور مثال بیشتر از ده سال) است. اما به دلیل عدم امکان دسترسی به چنین داده‌هایی، از یک پایگاه داده عمومی منتشر شده در وب‌سایت کگل استفاده شده است. این داده‌ها مربوط به یک بانک ناشناس است و برای هر مشتری ویژگي‌هايي همچون نام، شماره، نام خانوادگي، امتياز اعتباري، سود، مكان جغرافيايي و … ثبت شده است.

ابتدا با استفاده از اين مجموعه داده و نيز الگوريتم‌هاي يادگيري ماشين مناسب، به پيش‌بيني وفاداري و ارزش طول عمري مشتريان به طور جداگانه مي‌پردازيم. براي پيش‌بيني وفاداري مشتريان مي‌توان يكي از ستونهاي اين جدول داده كه به نام "Exited" را به عنوان برچسب داده‌ها در نظر گرفت كه قرار است ماندن يا خارج شدن مشتري در سيستم را به مدل يادگيري ماشين آموزش دهد. بنابراين اين ستون به پيش‌بيني وفاداري مشتريان كمك خواهد كرد. از طرف ديگر مي‌توان يك ستون جديدي تحت عنوان "target\_Lable\_remainstatue" ايجاد نمود كه بر اساس ستون "Tenure" ماندگاري را عدم ماندگاري هر مشتري را مشخص مي‌سازد. ستون "Tenure" به معناي تعداد سال‌هايي است كه يك مشتري در سيستم باقي مانده و فعاليتي از او ثبت شده است. بنابراين براي ايجاد ستون هدف جديد مي‌توان گفت هر مشتري كه بيشتر از پنج سال در اين بانك فعاليت داشته است ماندگار(وفادار) و در غيراينصورت غيرماندگار(غيروفادار) است. پس يكبار با استفاده از ستوني كه از قبل در اين جدول موجود بود به آموزش مدلهاي ماشين لرنينگ اقدام مي‌كنيم و يكبار هم بر اساس ستوني كه جديداً ايجاد كرديم اين فرايند را انجام مي‌دهيم. از ميان نتايج اين دو فرايند هر كدام كه دقت مدل بالاتري داشته باشند به عنوان مدل نهايي پيش‌بيني وفاداري مشتري انتخاب مي‌شود.

در رابطه فوق نماد نشان دهنده پیش‌بینی ماندگاری مشتری i طبق مدل m ام است. نیز وضعیت وفاداری یا عدم وفاداری مشتری i است.

پس از اتمام پيش‌بيني وفاداري مشتريان به سراغ پيش‌بيني ارزش طول عمر مشتريان مي‌رويم. در اين مرحله لازم است دو پارامتر اصلي ارزش پولي حاصل از مشتري و مدت زمان باقي ماندن مشتري در سيستم پيش‌بيني شوند. سپس از حاصلضرب اين دو پارامتر شاخص ارزش طول عمر مشتريان محاسبه خواهد شد. بنابراین ارزش طول عمر مشتری از رابطه زیر محاسبه خواهد شد:

در رابطه فوق نماد نشان دهنده پیش‌بینی مدت زمان باقی ماندن مشتری i در سیستم است. نماد نشان دهنده پیش‌بینی مقدار سود حاصل از مشتری i است. نیز ارزش طول عمر مشتری i است.

در جدول داده‌اي كه در اختيار داريم ستون "Tenure" نقش مدت زمان ماندگاري مشتري () را دارد و ستون "Balance" همان سود حاصل از مشتري () است. بنابراين يكبار ستون "Tenure" به عنوان هدف(ليبل) قرار مي‌گيرد ومقدار T به ازای هر مشتری با استفاده از مدلهاي ماشين لرنينگ پيش‌بيني مي‌شود و يكبار هم اين فرايند براي پيش‌بيني مقادیر به ازای هر مشتری و از طریق لیبل "Balance" انجام ميگیرد. در نهایت برای هر مشتری مقادیر پیش‌بینی شده هر دو پارامتر در یکدیگر ضرب خواهند شد. این مقدار معادل با ارزش طول عمر هر مشتری خواهد بود.

از آنجایی که در دنیای واقعی هیچ داده‌ای قطعی نیست و همواره داده‌ها در حال تغییر هستند بنابراین نمی‌توان تنها مقادیر حاصل از فرایندهای فوق را قطعی در نظر گرفت. رویکردهای مختلفی برای کار با داده‌های غیرقطعی وجود دارد تا بتوان به نحوی تأثیرات ناشی از احتمالی بودن پارامترهای ورودی را در روند حل مسأله و مدلهای ماشین لرنینگ وارد ساخت. بدین منظور در این تحقیق سعی شده است که از رویکرد سناریو محور بهره‌گیری شود. در این رویکرد چندین سناریو که شامل مقادیر مختلفی از پارامترها هستند در نظر گرفته می‌شود. برای هر سناریو تمامی فرایندهایی که پیش‌تر ذکر شد انجام می‌گردد. در نهایت اگر فرض شود به تعداد n سناریو تعریف شده باشد، لذا به تعداد n مدل پیش‌بینی برای شاخص‌های وفاداری (یکبار بر حسب لیبل "Exited" و یکبار بر حسب لیبل "target\_Lable\_remainstatue")و ارزش طول عمر مشتری ( شامل پیش‌بینی بر اساس لیبلهای "Balance" و "Tenure") تولید خواهد شد. ئر نهایت میان مقادیر پیش‌بینی شده از این مدلها میانگین گیری می‌شود و این مقدار به عنوان مقدار پیش‌بینی محتمل‌تر در محاسبه شاخص‌های مورد نظر استفاده می‌شوند.  بنابراین مقدار مورد انتظار از هر پارامتر از رابطه زیر محاسبه خواهد شد.

در رابطه فوق X می‌تواند یکی از پارامترهای مورد بررسی شامل ، و یا باشد. همچنین n تعداد سناریوهای تعریف شده است.

در نهایت ارزش طول عمر پیش‌بینی شده مورد انتظار و وفاداری موردانتظار مشتریان به ترتیب از روابط زیر محاسبه خواهند شد:

## پیاده‌سازی روش

در این بخش به پیاده‌سازی روش تشریح شده بر روی پایگاه داده از اطلاعات یک بانک ناشناس می‌پردازیم. این اطلاعات برگرفته از داده‌های منبع باز وب‌سایت کگل است. همانند اغلب مطالعات علمی که در حوزه داده کاوی انجام می‌شوند، لازم است یک چارچوب عمومی برای اجرای مرحله به مرحله عملیات در نظر گرفته شود. مراحل اين چارچوب در ادامه تشريح مي‌شود.

**کسب شناخت کافی از مسأله:** از آنجایی که عوامل گوناگونی بر شاخص‌های وفاداری مشتریان و ارزش طول عمر مشتری تأثیر گذارند بنابراین لازم است پایگاه داده مورد استفاده حاوی تمام اطلاعات مفید و مؤثر در هر دو شاخص باشد. بنابراین پیش از اقدام به جمع‌آوري داده‌ها لازم است شناخت كافي نسبت به مسأله كسب نمود و تا حد ممكن تمام عوامل مؤثر بر اين شاخص‌های مورد بررسی را تعيين نمود. به طور مثال در اين پژوهش سعي گرديد پايگاه داده‌ای انتخاب شود که شامل ویژگی‌های مهم و تأثیر گذاری همچون تاریخ تراکنشات مشتری، سن، تحصیلات، درآمد، سود حاصل از مشتری، مکان جغرافیایی، جنسیت، خدمات یا محصولات مورد استفاده باشد.

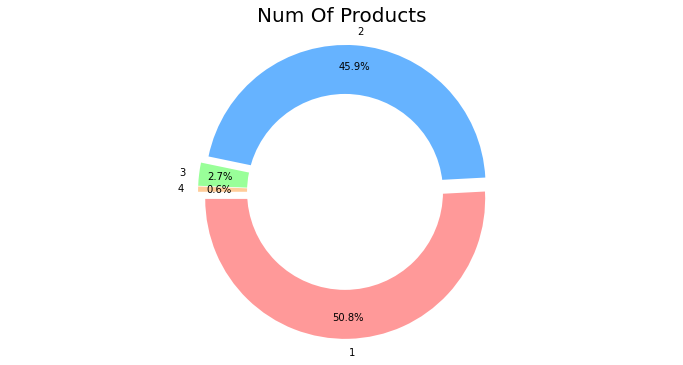
**جمع‌آوری داده‌های مناسب:** این مرحله مهمترین گام در اجرای سودمند روش پیشنهادی است. کیفیت داده ها و همچنین حجم آن‌ها نیز بسیار مهم و قابل توجه است. در صورتی که بتوان پایگاه داده‌ای با حجم اطلاعات به قدر کافی بزرگ جمع‌آوری نمود به طوری که کمترین میزان داده‌های ناقص و از دست رفته را شامل شود می‌توان با اتکا به چنین داده‌های مفیدی تخمین دقیق‌تری از شاخص‌ها ایجاد نمود. از طرف دیگر از آن‌جایی که در شاخص ارزش طول عمر مشتری پارامتر زمان نقش برجسته‌ای را ایفا می‌كند در نتيجه لازم است كه داده‌هاي گردآوري شده مربوط به بازه زماني طولاني و چند سال اخير شركت باشد. معمولاً چنين داده‌هايي يا برحسب تعداد سال گردآوري مي‌شوند يا به طور جزئي و بر حسب فصول سال، ماه‌ها یا روزهای سال تنظیم می‌شوند. در هر کدام از این حالتها اگر بتوان بازه زمانی طولانی و نزدیک به زمان فعلی را بررسی نمود در نهایت خروجی معتبر تری حاصل خواهد شد.

**‌پیش‌پردازش داده‌ها:** زمان‌بر ترین و یکی دیگراز مهمترین گام‌های مطالعات داده کاوی همین مرحله است. همواره پایگاه‌های اطلاعاتی و یا داده‌های گردآوری شده حاوی داده‌های از دست رفته، ناقص و انواع دیگر از عوامل کاهش دهنده کیفیت داده‌ها هستند. علاوه بر این معمولاً داده‌ها ترکیبی از فرمتهای مختلف هستند که مانع از اعمال مدلهای یادگیری ماشین بر روی آن‌ها می‌شود. به طور مثال در یک پایگاه داده ممکن است داده‌های عددی و متن همزمان وجود داشته باشند به هطوری که داده‌های متنی نیز به عنوان یکی از ویژگی‌های مهم در پیش‌بینی باشند که نمی‌توان از آنها چشم پوشی کرد. بنابراین در این مرحله اصلاحاتی بر روی داده‌ها انجام گرفته می‌شود تا بتوان آن را به فرم قابل قبولی در مدلهای ماشین لرنینگ تبدیل نمود. اصولاً تبدیل داده‌های متنی به عددی، حذف ستونهای غیرقابل استفاده و نیز هم مقیاس سازی داده‌ها رایج‌ترین عملیات پیش‌پردازش داده‌ها هستند. به طور مثال در این پژوهش ستونهای جدول اولیه عبارتند از CustomerId، CreditScore، Geography، Gender، Age، Tenure، Balance، NumOfProducts، HasCrCard، IsActiveMember، Exited، EstimatedSalary، Surname و RowNumber . از میان این ستونها فیچرهایی که در مدلهای پیش‌بینی مؤثر هستند انتخاب خواهند شد. این فیچرها عبارتند از CustomerId، CreditScore، Geography، Gender، Age، Tenure، Balance، NumOfProducts، HasCrCard، IsActiveMember، Exited، EstimatedSalary.

**تجزیه و تحلیل داده‌ها:** در این مرحله با استفاده از ابزارهای نموداری و بصری یک بینش و درکی از نحوه تأثیرگذاری داده‌ها بر یکدیگر ایجاد خواهد شد. در این مرحله ویژگی‌ای که بیشترین تأثیر را در فراوانی داده‌ها بر اساس سایر ویژگی‌ها دارد به عنوان ویژگی اصلی انتخاب می‌شود. به عبارت دیگر ابتدا فراوانی و سهم داده‌ها طبق ویژگی‌ها (فیچرها) بررسی می‌شود سپس ارتباط میان دو به دوی فیچرها بر روی فراوانی داده‌ها بررسی می‌شود. به طور مثال فیچر Tenure مورد تحلیل قرار داده می‌شود. در این تحلیل مشخص خواهد شد که مشتریان از نظر Tenure به چند دسته تقسیم می‌شوند. حداقل و حداکثر و میانگین این فیچر چه مقداری است. در هر دسته چه تعدادی از مشتریان قرار می‌گیرند. چنین تحلیلهایی برای سایر فیچرها نیز انجام خواهد گرفت.

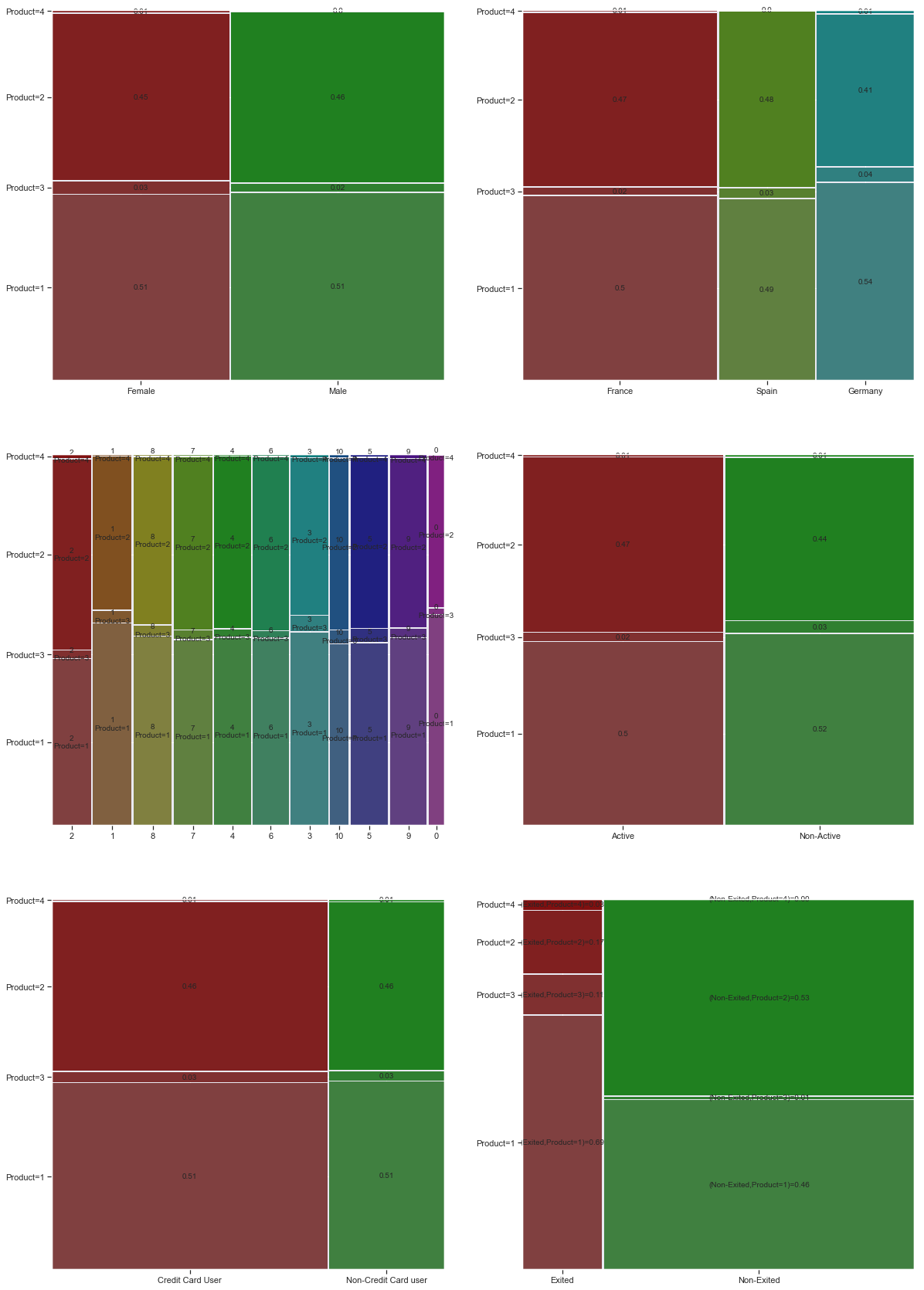
در این پژوهش تأثیر تعداد خدمت مورد استفاده توسط مشتری در سایر فیچرها مورد نظر قرار داده شده است. بدین منظور ابتدا بررسی شده است که مشتریان از نظر تعداد خدمت مود استفاده به چند دسته تقسیم می‌شوند و هر دسته چه بخشی از مشتریان را شامل می‌شود. همانطور که در شکل ‏1‑1 مشاهده می‌شود مشتریان از نظر تعداد خدمت استفاده شده به چهار دسته تقسیم می‌شوند. یعنی مشتریان ممکن است در یکی از حالات یک تا چهار خدمت باشند. اما از این میان مشتریان تک خدمته و سپس مشتریان دو محصوله بیشترین بخش از جامعه مشتریان را تشکیل می‌دهند. تعداد مشتریانی که سه یا چهار محصول را استفاده می‌کنند بسیار ناچیز است.

شکل ‏1‑1: دسته بندی مشتریان بر اساس فیچر تعداد خدمت مورد استفاده

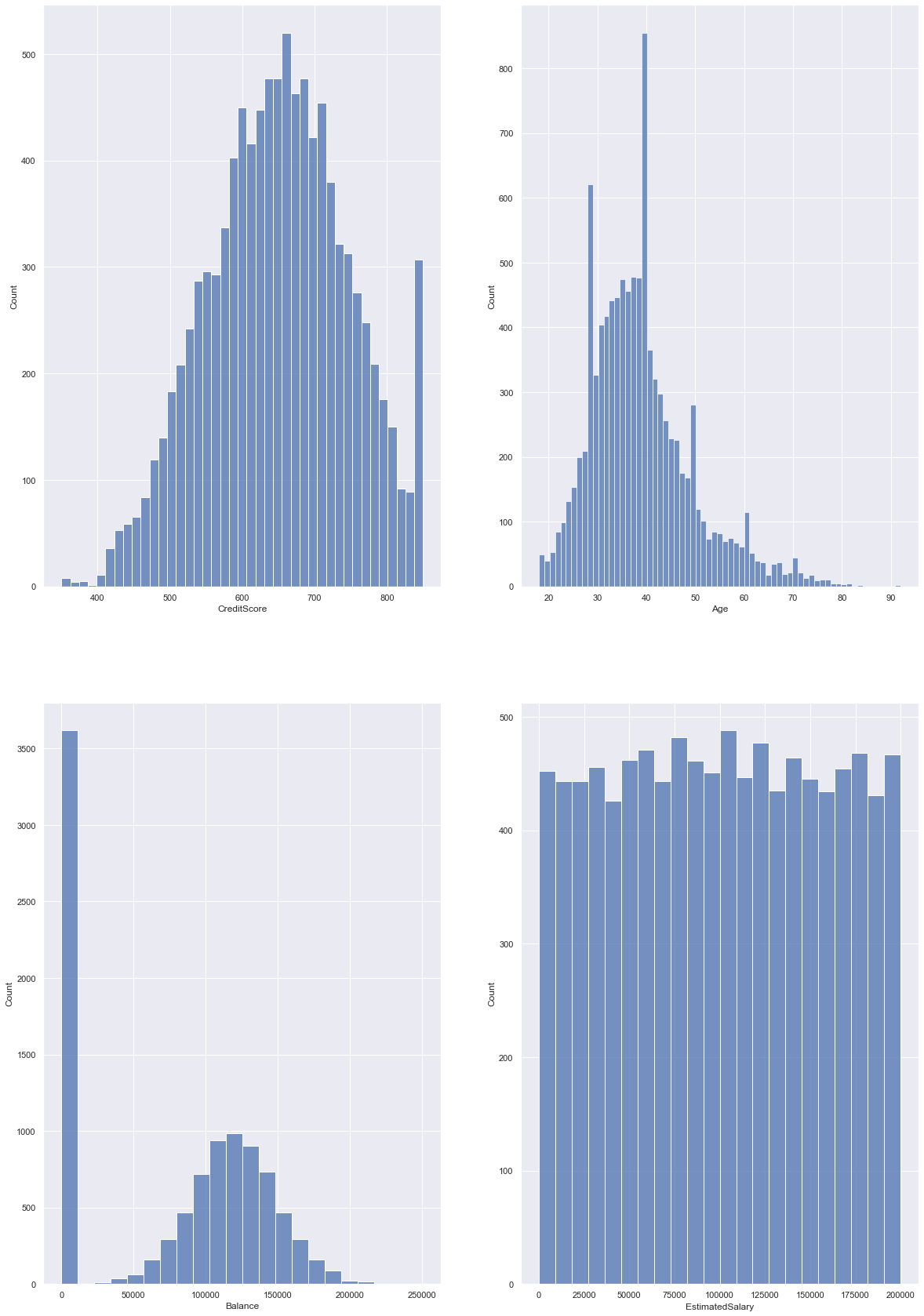


سپس به بررسی زوجی سهم تعداد خدمات مورد استفاده با سایر فیچرها می‌پردازیم. پایگاه داده مورد بررسی شامل دو دسته فیچر از نوع categorical و non\_categorical است. به منظور تحلیل آن‌ها از دو روش مختلف نموداری استفاده شده است. در شکل ‏1‑2 و شکل ‏1‑3 نمودارهای مربوط به تحلیل زوجی فیچرها با فیچر تعداد خدمات مورد استفاده ارائه شده است.

شکل ‏1‑2: مقایسه زوجی فیچرهای categorical و تعداد خدمات مورد استفاده از نظر تعداد مشتریان در هر دسته



شکل ‏1‑3: مقایسه زوجی فیچرهای non\_categorical و تعداد خدمات مورد استفاده از نظر تعداد مشتریان در هر دسته



ایجاد سناریوها: در این مرحله با استفاده از تحلیل جفتی بین فیچر تعداد محصولات مورد استفاده و سایر فیچرها در هر دو نوع categorical و non categorical بر این اساس که چه تعداد مشتری یا چه درصدی از مشتریان از هر جفت دسته مشخص استفاده کردند می‌توان سناریو های مختلف را ایجاد نمود. به طور مثال یکبار می‌توان مشتریان را بر اساس جفت فیچر تعداد خدمات مورد استفاده و مکان جغرافیایی در سیستم تقسیم بندی نمود. در این مورد تعداد مشتریانی که از یک خدمت استفاده کردند و نیز فرانسوی بودند 50 درصد از کل مشتریانی هستند که از یک خدمت استفاده کردند. بدین ترتیب با محاسبه درصد تعداد مشتریان از دیدگاه جفت فیچرها به طوری که همواره فیچر تعداد خدمات مورد استفاده با سایر فیچرها مقایسه گردد می‌توان نسبتهای ثابتی برای هر جفت فیچر محاسبه نمود. با در دست داشتن این نسبتهای ثابت به منظور ایجاد سناریوی جدید کافی است درصدهای اولیه تعداد مشتریان بر اساس فیچر تعداد خدمات مورد استفاده تغییر داده شود و بر اساس رابطه تناسبی با سایر فیچرها درصد مشتریان برای سایر فیچرها بر مبنای فیچر خدمات مورد استفاده محاسبه شود.

**ایجاد مدلهای ماشین لرنینگ:** در صورتی که مراحل قبل به خوبی انجام گیرند این مرحله با سرعت بیشتری انجام خواهد گرفت. در این مرحله برای هر سناریو به طور جداگانه تنها کافی است داده‌ها به دو دسته آموزش و آزمون تقسیم شوند و مدلهایی متناسب با نوع داده‌ها و نیز لیبل آن‌ها ایجاد شود. به طور مثال در نگاه اول برای پیش‌بینی وفاداری مشتریان و نیز برای پیش‌بینی مدت زمان ماندگاری مشتریان مدل لجستیک رگرسیون مناسب است همچنین برای پیش‌بینی سود حاصل از مشتریان نیز مدل رگرسیون خطی مناسب است. علاوه بر اینها می‌توان از سایر مدلها با اعمال برخی تغییرات مختصر استفاده نمود.

**ارزیابی عملکرد مدلها و انتخاب بهترین مدل:** پس از تشکیل مدلهای پیش‌بینی می‌توان اعتبار و دقت آن‌ها را با استفاده از شاخص‌هاي متدوالي همچون دقت، صحت و ... محاسبه نمود. در اين پژوهش از روشهاي متنوعي همچون K-Fold validation، cross-validation و confusion matrix استفاده گرديده است.

**تشکیل پایگاه‌داده‌های پیش‌بینی شده:** از كنار هم قرار دادن مقادير پيش‌بيني شده براي پارامترهاي مورد نظر مي‌توان يك پايگاه داده جديد تشكيل داد كه قرار است از اين داده‌ها براي محاسبه شاخص‌هاي وفاداري مشتريان و ارزش طول عمر مشتريان استفاده نمود.

**محاسبه شاخص‌های وفاداری و ارزش طول عمر مشتری:** با استفاده از جدول مقادير پيش‌بيني شده و از طريق روابط ذكر شده براي هر شاخص مقدار آن به ازاي هر مشتري محاسبه مي‌شود و به عنوان يك ستون جديد به جدول پيش‌بيني شده ها افزوده مي‌شود.

**مقایسه نتایج و بررسی آزمون فرض:** در اين مرحله مشتريان به طور جداگانه يكبار بر اساس شاخص وفاداري و يكبار بر اساس شاخص ارزش طول عمر مشتري رتبه‌بندي مي‌شوند. سپس به ازاي هر مشتري بررسي مي‌شود كه آيا مشتري كه در دسته وفاداران بوده است رتبه ارزش طول عمر بالايي نيز داشته است يا خير. بدين ترتيب مي‌توان فرضيه مورد نظر خود يعني وجود رابطه ميان دو شاخص ارزش طول عمر مشتري و وفاداري مشتريان را سنجيد.

## بررسي نتايج عددي

جدول ‏1‑1: سهم تعداد محصولات مورد استفاده در هر سناريو

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **4** | **3** | **2** | **1** | **NO.product**  **NO.senario** |
| 0.6 | 2.7 | 45.9 | 50.8 | 1 |
| 10 | 10 | 50 | 30 | 2 |
|  |  |  |  | 3 |

جدول ‏1‑2: پنج سطر اول داده‌هاي سناريوهاي مختلف

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **IsActiveMember** | **salary ($)** | **HasCrCard** | **Exited** | **Geography** | **Gender** | **Tenure** | **Balance ($)** | **Age** | **NomOfProducts** | **senario** |
| 1 | 101348.88 | 1 | 1 | 1 | 1 | 2 | 0 | 42 | 1 | 1 |
| 1 | 112542.58 | 0 | 0 | 3 | 1 | 1 | 83807.86 | 41 | 1 |
| 0 | 113931.57 | 1 | 1 | 1 | 1 | 8 | 159660.80 | 42 | 3 |
| 0 | 93826.63 | 0 | 0 | 1 | 1 | 1 | 0 | 39 | 2 |
| 1 | 79084.10 | 1 | 0 | 3 | 1 | 2 | 125510.82 | 43 | 1 |
|  |  |  |  |  |  |  |  |  |  | 2 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | 3 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |