دانشکده مهندسی صنایع

پروژه درس مبانی داده‌کاوی

استاد پروژه: جناب دکتر خدمتی

اعضای گروه:

امیرحسین قناعتیان ۹۷۱۰۴۵۸۳

سجاد عابد ۹۷۱۰۴۵۱۵

تابستان ۱۴۰۱

فهرست

مقدمه

«داده‌کاوی»[1] این مساله را با فراهم کردن روش‌ها و نرم‌افزارهایی برای خودکارسازی تحلیل‌ها و اکتشاف مجموعه داده‌های بزرگ و پیچیده، حل می‌کند. پژوهش‌ها در زمینه داده‌کاوی در گستره‌ی وسیعی از موضوعات شامل آمار، علوم کامپیوتر[2]، «یادگیری ماشین» (Machine Learning)، «مدیریت پایگاه داده» (Database Management) و «بصری‌سازی داده‌ها» (Data Visualization) دنبال می‌شود. روش‌های داده‌کاوی و یادگیری، در زمینه‌هایی غیر از آمار نیز توسعه داده شده‌اند، که از جمله آن‌ها می‌توان به یادگیری ماشین و «پردازش سیگنال» (signal processing) اشاره کرد.

به مجموعه‌ای از روش‌های قابل اعمال بر پایگاه داده‌های بزرگ و پیچیده به منظور کشف الگوهای پنهان و جالب توجه نهفته در میان داده‌ها، داده‌کاوی گفته می‌شود. روش‌های داده‌کاوی تقریبا همیشه به لحاظ محاسباتی پر هزینه هستند. علم میان‌رشته‌ای داده‌کاوی، پیرامون ابزارها، متدولوژی‌ها و تئوری‌هایی است که برای آشکارسازی الگوهای موجود در داده‌ها مورد استفاده قرار می‌گیرند و گامی اساسی در راستای کشف دانش محسوب می‌شود. دلایل گوناگونی پیرامون چرایی مبدل شدن داده‌کاوی به چنین حوزه مهمی از مطالعات وجود دارد.

در سال ۱۹۶۰، کارشناسان آمار از اصطلاحات «صید داده»[3] و «لایروبی داده»[4] برای ارجاع به فعالیت‌های «تحلیل داده»[5] استفاده می‌کردند. اصطلاح «داده‌کاوی» در حدود سال ۱۹۹۰ در جامعه پایگاه‌داده مورد استفاده قرار گرفت و به محبوبیت قابل توجهی دست پیدا کرد. عنوان مناسب‌تر برای فرآیند داده‌کاوی، «کشف دانش از داده»[6] است.

در حال حاضر، یادگیری آماری، «تحلیل داده» و «علم داده»[7] از دیگر عباراتی هستند که با معنای مشابه داده‌کاوی مورد استفاده قرار می‌گیرند، حال آنکه گاه تفاوت‌های ظریفی میان این موارد وجود دارد.

این یک واقعیت انکارناپذیر است که داده‌ها در همه جا ما را احاطه کرده‌اند. ما واقعا خوش شانس هستیم که در این عصر زندگی می‌کنیم. عصری که در آن، شاهد رشد اینترنت هستیم و مزایای برخاسته از اشتراک اطلاعات قابل دسترس را به چشم می‌بینیم. اگر بخواهیم منطقی باشیم، حضور آنلاین ما در اینترنت، تمام کلیک‌هایی که توسط ما انجام می‌شود، وب سایت‌هایی که بازدید می‌کنیم، مدت زمانی که در هر وب سایت سپری می‌کنیم و ...، همگی داده‌هایی هستند که ما تولید می‌کنیم.

---

[1] Data Mining
[2] Computer Science
[3] Data Fishing
[4] Data Dredging
[5] Data Analytics
[6] Knowledge Discovery from Data
[7] Data Science

با استفاده از ابزارها و امکانات پردازشی مناسب، می توان داده‌های تولید شده را پاکسازی، و به یک بینش مرتبط تبدیل کرد. این بینش، منجر به تصمیم گیری شرکت‌های بزرگ می‌شود و منافع آن‌ها را تامین می‌کند. افرادی که در این حوزه‌ها فعالیت می‌کنند، به راحتی می‌توانند اصطلاحاتی مانند داده کاوی و تحلیل داده را متوجه شوند. با این حال، برای کسانی که در این زمینه‌ها نیستند، به دست آوردن یک درک اساسی از این اصطلاحات احتمالاً گیج کننده خواهد بود.

داده کاوی و تحلیل داده، گام‌های اساسی در هر پروژه داده محور هستند. داده کاوی و تحلیل داده باید برای اطمینان از موفقیت پروژه، به صورت تمام و کمال انجام شوند. گسترش حجم داده‌ها به صورت نمایی، منجر به انقلاب اطلاعات و دانش شده است. این روزها، جمع آوری اطلاعات قابل توجه و کسب دانش عمیق از داده‌های موجود، یک جنبه اصلی توسعه تحقیق و استراتژی به حساب می‌آید.

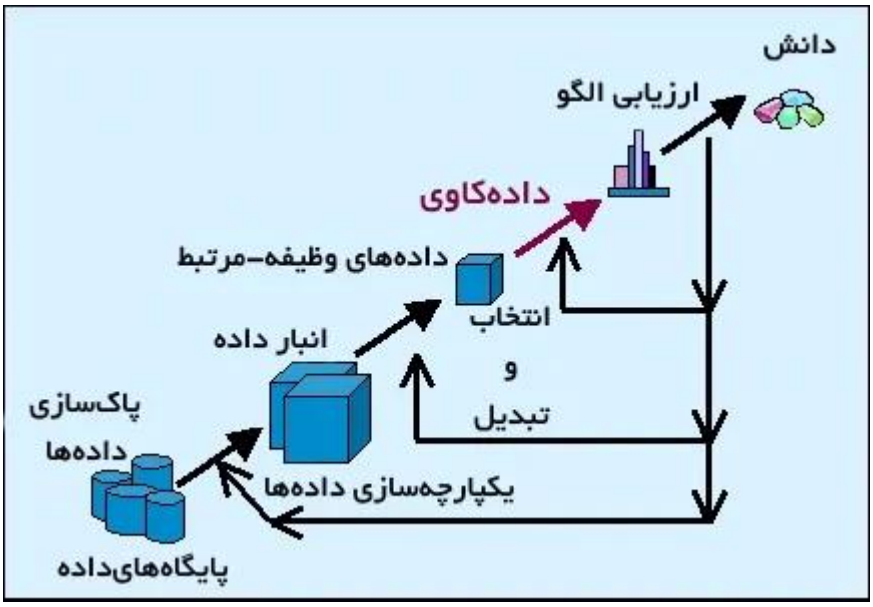
تمامی این اطلاعات، در یک انبار داده نگهداری می‌شود و سپس برای اهداف هوش تجاری مورد استفاده قرار می‌گیرد. تعاریف و دیدگاه‌های بی شماری در این زمینه وجود دارد، اما همگی موافق‌اند که تحلیل داده و داده کاوی، دو زیر مجموعه هوش تجاری هستند. با توجه به نزدیک بودن این دو حوزه، یافتن تفاوت بین داده کاوی و تحلیل داده می‌تواند بسیار چالش برانگیز باشد.

داده کاوی، فرایند استخراج داده‌های قابل استفاده از یک مجموعه بزرگتر، متشکل از داده‌های خام است. داده کاوی یک زیر مجموعه از تحلیل داده است؛ یک روش کارآمد و پیوسته برای شناخت و کشف الگوها و داده‌های پنهان در یک مجموعه داده بزرگ. علاوه بر این، داده کاوی برای ساخت مدل‌های یادگیری ماشین که بعدا در هوش مصنوعی استفاده می‌شود، مورد استفاده قرار می‌گیرد. داده کاوی، از الگوریتم‌های پیچیده ریاضی برای تقسیم بندی داده‌ها و ارزیابی احتمال وقایع آینده استفاده می‌کند. داده کاوی با عنوان کشف دانش در داده‌ها نیز شناخته می‌شود.

داده کاوی به طور کلی بخشی از تحلیل داده است که هدف و قصد آن، تعیین یا کشف صرف الگو از یک مجموعه داده است. از طرف دیگر، تحلیل داده به عنوان یک بسته کامل برای ایجاد معنی از پایگاه داده انجام می‌شود که ممکن است شامل داده کاوی باشد یا نباشد. هر دو زمینه، نیاز به مجموعه مهارت‌ها، توانایی‌ها و تخصص مشخص و متمایزی دارند. در سال‌های آینده، هر دو حوزه، چه از نظر داده و منابع و چه از نظر مشاغل، به شدت مورد توجه و تقاضا قرار خواهند گرفت.

مرور ادبیات

داده‌کاوی که با عنوان «کشف دانش از داده»[8] نیز شناخته شده است، فرایند استخراج اطلاعات و دانش از داده‌های موجود در پایگاه‌داده یا انبارداده است.[3]



فرآیند داده‌کاوی شامل چندین گام است. این فرآیند از داده‌های خام آغاز می‌شود و تا شکل‌دهی دانش جدید ادامه دارد. فرآیند بازگشتی داده‌کاوی شامل گام‌های زیر است:

- «پاک‌سازی داده» (Data Cleaning)
- «یکپارچه‌سازی داده» (Data Integration)
- «انتخاب داده» (Data Selection)
- «تبدیل داده» (Data Transformation)
- «کاوش داده» (Data Mining)
- «ارزیابی الگو» (Pattern Evaluation)
- «ارائه دانش» (Knowledge Representation)

[8] Knowledge Discovery from Data | KDD

در ادامه به توضیح جداگانه هر کدام از موارد فوق می‌پردازیم.(منبع شماره 4)

پاک‌سازی داده

در این فاز «نویز» (نوفه) از مجموعه حذف و تدابیری برای «داده‌های ناموجود»[9] اندیشیده می‌شود. برای مطالعه بیشتر در این رابطه، مطلب «پاکسازی داده»[10] در پایتون با استفاده از NumPy و Pandas — راهنمای جامع» پیشنهاد می‌شود.

یکپارچه‌سازی داده‌ها

در اغلب مسائل داده‌کاوی، داده‌ها از منابع داده گوناگون باید به یکباره مورد تحلیل قرار بگیرند. مثال خوبی از این مورد پایگاه داده‌های شعب مختلف یک فروشگاه زنجیره‌ای در شهرها و کشورهای گوناگون جهان است. برای تحلیل این داده‌ها باید آن‌ها را به صورت یکپارچه در یک «انبار داده»[11] گردآوری کرد، این کار در فاز یکپارچه‌سازی انجام می‌شود.

انتخاب داده

در فاز انتخاب داده، باید داده‌های مرتبط با تحلیل انتخاب، و از مجموعه داده برای انجام تحلیل‌ها بازیابی شوند. در مطلب «انتخاب ویژگی[12] در داده‌های ابعاد بالا — خودآموز ساده» به این مساله همراه با جزئیات پرداخته شده است. همچنین، مطالعه «الگوریتم کاهش ابعاد t-SNE با مثال‌های پایتون — آموزش کاربردی» نیز به علاقمندان پیشنهاد می‌شود.

تبدیل داده

تبدیل داده یک روش تثبیت داده نیز هست. در این فاز، داده‌های انتخاب شده به فرم دیگری تبدیل می‌شوند. این کار به ساده‌تر شدن، بهبود صحت و دقت فرآیند کاوش کمک می‌کند.

داده‌کاوی

در این فاز از روش‌های هوشمندانه برای استخراج الگوهای مهم و جالب توجه از میان داده‌ها استفاده می‌شود.

ارزیابی الگو

در این فرآیند، الگوهای حاصل شده در گام قبل، از جنبه‌های گوناگونی شامل دقت، صحت و قابلیت تعمیم و دیگر موارد مورد ارزیابی قرار می‌گیرند.

---

[9] Missing Values
[10] Data Cleaning
[11] Data Warehouse
[12] Feature Selection

ارائه دانش

ارائه دانش فاز نهایی فرآیند داده‌کاوی است. در این فاز، دانش کشف شده به شیوه قابل درک به کاربر ارائه می‌شود. در این گام حیاتی و بسیار مهم، روش‌های بصری‌سازی مورد استفاده قرار می‌گیرند. این کار به کاربران در درک و تفسیر نتایج داده‌کاوی کمک می‌کند.

متدولوژی K-means

الگوریتم K-Means یک الگوریتم یادگیری بدون نظارت است که برای حل مشکلات خوشه بندی در یادگیری ماشین یا علم داده استفاده می شود. در این مبحث به بررسی الگوریتم خوشه بندی K-Means می پردازیم. خوشه‌بندی K-Means روشی در کمی‌سازی بردارهاست که در اصل از پردازش سیگنال گرفته شده و برای آنالیز خوشه بندی در داده کاوی محبوب است. هدف الگوریتم K-Means خوشه بندی k مشاهده به n خوشه است که در آن هر یک از مشاهدات متعلق به خوشه‌ای با نزدیکترین میانگین به آن است، این میانگین به عنوان نمونه استفاده می‌شود.

الگوریتم یادگیری بدون نظارت[۱۳]

الگوریتم یادگیری بدون نظارت: در یادگیری بدون نظارت یا یادگیری نظارت نشده مسئله به این صورت است که در این حالت داده‌هایی که داریم که پاسخ صحیح آن‌ها مشخص نیست و این داده‌ها برچسبی یکسان دارند و یا اصلاً برچسبی ندارند. پس مجموعه‌ای از داده در اختیار الگوریتم قرار می گیرد که ساختار مشخصی ندارند؛ سپس الگوریتم یادگیری بدون نظارت (انواع مختلفی دارد مانند الگوریتم k-means، الگوریتم خوشه بندی سلسه مراتبی و...) تشخیص می‌دهدکه چه داده‌هایی باید در یک خوشه قرار بگیرد.

الگوریتم خوشه‌بندی[۱۴]

الگوریتم خوشه‌بندی یکی از متداول‌ترین روش تجزیه و تحلیل داده های اکتشافی است که برای دریافت شهودی در مورد ساختار داده‌ها استفاده می‌شود. این روش می‌تواند به عنوان وظیفه‌ی شناسایی زیرگروه‌ها در داده‌ها تعریف شود به این صورت که نقاط داده در یک زیر گروه (خوشه) بسیار شبیه به هم هستند در حالی که نقاط داده در خوشه‌های مختلف بسیار متفاوت هستند.

به عبارت دیگر، ما سعی می‌کنیم زیرگروه‌های همگن را در داده‌ها پیدا کنیم، به این ترتیب که نقاط داده در هر خوشه با توجه به اندازه گیری شباهت مانند فاصله مبتنی بر اقلیدس یا فاصله مبتنی بر همبستگی، تا حد امکان شبیه هستند. تصمیمی که برای اندازه‌گیری شباهت از چه نوع فاصله‌ای استفاده شود خاص هر داده و برنامه است.

الگوریتم K-Means یک الگوریتم بر پایه‌ی تکراری است که سعی می‌کند مجموعه داده‌ها را به زیرگروه‌های متمایز بدون همپوشانی تعریف کند که به این زیر گروه‌ها خوشه گفته‌می شود؛ که در این گروه‌ها هر نقطه داده فقط به یک گروه تعلق دارد. در

---

[13] Unsupervised Learning
[14] Clustering

این الگوریتم سعی می‌شود نقاط داده درون خوشه‌ای را تا حد ممکن شبیه به هم ساخت و در عین حال خوشه‌ها را تا حد امکان متفاوت (دور) از هم تعریف کرد.

این الگوریتم داده‌ها را به یک خوشه اختصاص می‌دهد به طوری که مجموع فاصله مربع شده بین نقاط داده و مرکز گروه (میانگین محاسبه تمام نقاط داده‌ای که به آن خوشه تعلق دارند) در حداقل باشد. هرچه تنوع کمتری در خوشه‌ها داشته باشیم، نقاط داده در یک خوشه همگن (مشابه) هستند. رویکردی که K-Means که برای حل مسئله دنبال می‌کند، Expectation-Maximization نامیده می‌شود.

در اینجا K تعداد خوشه‌های از پیش تعریف شده‌ای را که باید در این فرآیند ایجاد شوند تعریف می‌کند، کما اینکه K=2، دو خوشه وجود دارد و برای K=3، سه خوشه و غیره وجود دارد. K-Means یک الگوریتم تکراری است که مجموعه داده بدون برچسب را به k خوشه‌ی مختلف تقسیم می‌کند به گونه‌ای که هر مجموعه داده فقط متعلق به یک گروه است که دارای ویژگی‌های مشابه است.

به ما امکان می‌دهد تا داده‌ها را در گروه‌های مختلف قرار دهیم و راهی مناسب برای کشف دسته‌های گروه‌ها در مجموعه داده‌های بدون برچسب و بدون نیاز به هیچ گونه آموزش وجود دارد. این یک الگوریتم مبتنی بر مرکز است که در آن هر خوشه با یک مرکز ارتباط دارد. هدف اصلی این الگوریتم به حداقل رساندن مجموع فواصل بین نقطه داده و خوشه های مربوطه است.

الگوریتم K-Means ، مجموعه داده بدون برچسب را به عنوان ورودی می‌گیرد، مجموعه داده را به تعداد k خوشه تقسیم می‌کند و روند را تکرار می‌کند تا زمانی که بهترین خوشه‌ها را پیدا نکند الگوریتم ادامه پیدا می‌کند. مقدار k باید در این الگوریتم از پیش تعیین شده باشد. عملکرد الگوریتم خوشه بندی k-mean به صورت زیر است:

با استفاده از یک فرایند تکرار، بهترین مقدار را برای نقاط مرکز تعیین می‌کند. هر نقطه داده را به نزدیکترین مرکز خود اختصاص می‌دهد. آن نقاط داده‌ای که نزدیک مرکز k هستند، خوشه‌ای را ایجاد می‌کنند. از این رو هر خوشه دارای نقاط داده با برخی نقاط مشترک است و از خوشه‌های دیگر دور است.

نحوه کار الگوریتم K-Means در مراحل زیر توضیح داده شده است:

مرحله ۱: برای تصمیم گیری در مورد تعداد خوشه‌ها ، تعداد K را انتخاب می‌شود.

مرحله ۲: K تا از نقاط را به صورت تصادفی یا با محاسبه انتخاب می‌شود. (این می‌تواند غیر از مجموعه داده ورودی باشد). بر اساس کد زیر از فاصله‌ی اقلیدوسی برای انتخاب مراکز استفاده شده است.

مرحله ۳: هر نقطه داده را به نزدیکترین مرکز خود اختصاص می‌دهد، که خوشه‌های K از پیش تعریف شده را تشکیل می‌دهد.

مرحله ۴: میانگین را محاسبه کرده و یک مرکز جدید برای هر خوشه قرار می‌دهد.

مرحله ۵: مراحل سوم را تکرار می‌شود، به این معنی که هر پایگاه داده را به جدیدترین و نزدیکترین مرکز هر خوشه اختصاص می‌دهد.

مرحله ۶: اگر تغییر مجددی اتفاق افتاد، سپس مرحله 4 مجدد اجرا می‌شود و الگوریتم به پایان می‌رسد.

مرحله ۷: مدل آماده است.

متدولوژی PCA

تحلیل مولفه اساسی به بیان ساده، روشی برای استخراج متغیرهای مهم (به شکل مولفه) از مجموعه‌ی بزرگ متغیرهای موجود در یک مجموعه داده است. تحلیل مولفه اساسی در واقع یک مجموعه با بُعد پایین از ویژگی‌ها را از یک مجموعه دارای بُعد بالا استخراج می‌کند تا به ثبت اطلاعات بیشتر با تعداد کمتری از متغیرها کمک کند. بدین شکل، بصری‌سازی داده‌ها نیز معنادارتر می‌شود. تحلیل مولفه اساسی هنگامی که با داده‌های دارای سه یا تعداد بیشتری بُعد سروکار داشته باشید، کاربردپذیرتر است. تحلیل مولفه اساسی همیشه روی ماتریس کوواریانس یا همبستگی اعمال می‌شود. این یعنی داده‌ها باید عددی و استاندارد شده باشند.

یک مولفه اساسی یک ترکیب خطی نرمال شده از پیش‌بین‌های اصلی موجود در مجموعه داده است. در شکل ۱، 1PC و 2PC مولفه‌های اساسی هستند. فرض می‌شود یک مجموعه از پیش‌بین‌ها به صورت $X^1, X^2 \ldots, Xp$ وجود دارد. مولفه‌های اساسی این مجموعه از پیش‌بینی‌ها را می‌توان بدین شکل نوشت:

$$Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + \ldots. + \Phi p1Xp$$

که در آن:

Z^1 اولین مولفه اساسی است.

$\Phi^1 p$ بردار بار شامل بردارهای بار $(\Phi^2,\Phi^1 \ldots)$ اولین مولفه اساسی است. بردارهای بار به مجموع مربعات مساوی با یک محدود شده‌اند. دلیل این امر آن است که داشتن مقادیر بار بزرگ ممکن است منجر به ایجاد واریانس بسیار بزرگ شود. این مقدار همچنین جهت مولفه اساسی (Z^1) را در جهتی که داده‌ها بیشترین تنوع را دارند، تعریف می‌کند. نتیجه این امر یک خط در فضای p بُعدی است که نزدیک‌ترین مقدار به n نمونه را دارد. میزان نزدیکی به وسیله محاسبه میانگین مربعات فاصله‌های اقلیدسی اندازه‌گیری می‌شود.

$X^1..Xp$ پیش‌بینی‌های نرمال شده هستند. میانگین پیش‌بین‌های نرمال شده برابر با صفر و انحراف معیار آن‌ها برابر با یک است.

بنابراین:

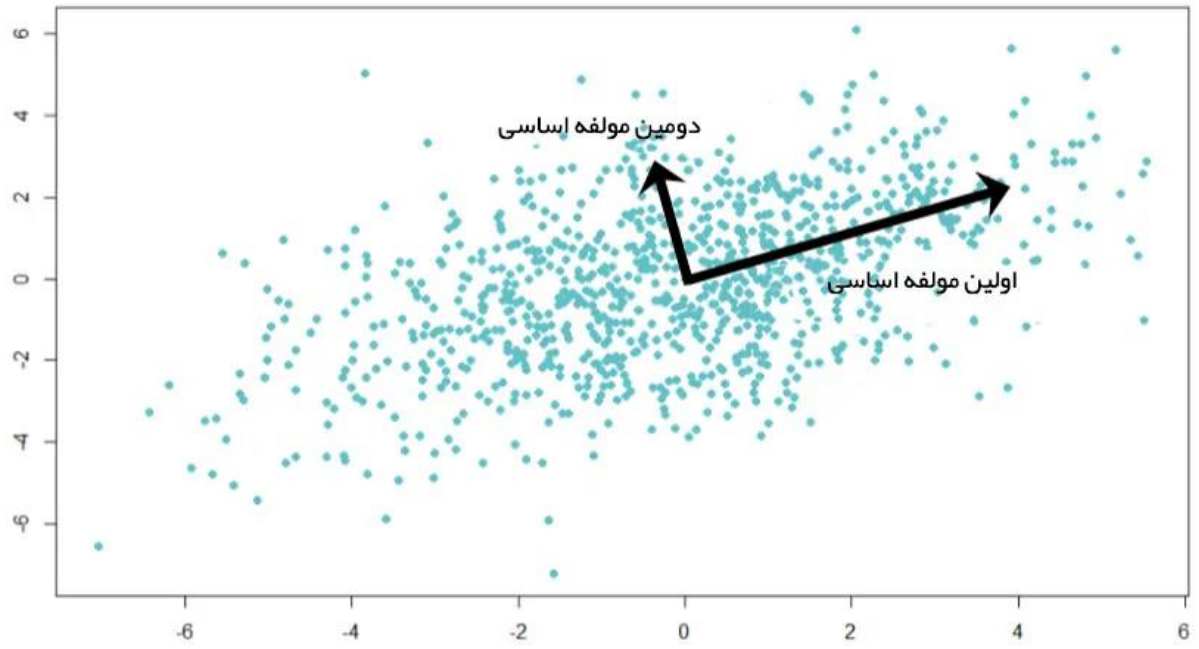
اولین مولفه اساسی، یک ترکیب خطی از پیش‌بین‌های اصلی است که بیشترین واریانس موجود در مجموعه داده‌ها را در بر می‌گیرد. این مولفه، جهت بیشترین تغییرات در داده‌ها را تعیین می‌کند. هرچه دامنه تغییرات موجود در اولین مولفه بالاتر باشد، اطلاعات موجود در این مولفه بیشتر است. هیچ مولفه دیگری نمی‌تواند بیش از مولفه اساسی اول دامنه تغییرات داشته باشد. نتیجه محاسبه اولین مولفه اساسی، خطی است که نزدیک‌ترین خط به داده‌ها محسوب می‌شود. در واقع این خط مجموع مربع فواصل را بین یک نقطه داده و خط، به کمینه مقدار می‌رساند.

مولفه اساسی دوم را نیز به روش مشابهی می‌توان به دست آورد:

دومین مولفه اساسی (Z^2) نیز یک ترکیب خطی از پیش‌بین‌های اصلی است که واریانس باقی‌مانده در مجموعه داده را در خود حفظ می‌کند و با مقدار Z^1 ناهمبسته است. به عبارت دیگر، همبستگی بین مولفه اساسی اول و دوم صفر است. مولفه اساسی دوم را می‌توان به شکل زیر نمایش داد:

$$Z^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + \ldots + \Phi p2Xp$$

اگر دو مولفه ناهمبسته باشند، جهت‌های آن‌ها باید متعامد (مانند شکل ۲) باشد. شکل ۲ براساس داده‌های شبیه‌سازی شده با دو ویژگی ترسیم شده است. جهت مولفه‌ها، چنان‌که انتظار می‌رود به صورت متعامد است و این یعنی مقدار همبستگی آن‌ها برابر با صفر است.



تحلیل مولفه اساسی روی نسخه نرمال شده پیش‌بین‌های اصلی قابل انجام است. این امر به آن دلیل است که پیش‌بینی‌های اصلی ممکن است مقیاس‌های گوناگونی داشته باشند. به عنوان مثال می‌توان به یک مجموعه داده که شامل متغیرهایی با یکاهای گالون، کیلومتر، سال نوری و دیگر انواع واحدها است، اشاره کرد. واضح است که مقدار واریانس این متغیرها اعداد بزرگی خواهد بود. انجام PCA روی متغیرهای نرمال نشده منجر به بارهای فوق‌العاده بزرگی برای متغیرهای دارای واریانس بالا می‌شود و این امر به نوبه خود می‌تواند منجر به وابستگی مولفه اساسی به متغیرهای دارای واریانس بالا شود که بسیار نامطلوب است.

چنانکه در شکل ۳ می‌توان دید، PCA دو بار روی مجموعه داده اجرا گشته (با متغیرهای نرمال شده و نرمال نشده). مجموعه داده به‌کار برده شده در این مثال دارای ۴۰ ویژگی است. چنانکه مشهود است، اولین مولفه اساسی تحت سیطره متغیر MRP قرار گرفته است. دومین مولفه اساسی نیز تحت تسلط متغیر Item_Weight قرار گرفته است. این اتفاقات به دلیل بالا بودن واریانس متغیر است. هنگامی‌که متغیرها نرمال می‌شوند، بصری‌سازی آن‌ها در فضای دو بُعدی به شکل بهتری انجام‌پذیر است.

متدولوژی t-SNE

«t-توکاری همسایگی تصادفی توزیع شده»[15]، روش نظارت نشده غیر خطی است که برای اکتشاف و بصری‌سازی داده‌ها مورد استفاده قرار می‌گیرد. به بیان ساده‌تر، t-SNE به کاربر درکی از اینکه داده‌ها چگونه در فضای ابعاد بالا سازمان‌دهی شده‌اند ارائه می‌کند. این روش توسط «لورنز ون در ماتنز»[16] و «جفری هینتون»[17] در سال ۲۰۰۸ ساخته شد.

کاربرانی که با تحلیل مولفه اساسی[18] آشنایی دارند ممکن است چنین پرسشی را طرح کنند که بین این الگوریتم و t-SNE چه تفاوتی هست. اولین موردی که می‌توان به آن اشاره کرد این است که PCA در سال ۱۹۹۳ ساخته شد، در حالیکه t-SNE در سال ۲۰۰۸ ظهور پیدا کرد. از سال ۱۹۹۳ به بعد چیزهای زیادی در دنیای علم داده تغییر کرده که مهم‌ترین آن‌ها توان محاسباتی (و ابزارهای محاسباتی) و اندازه داده‌ها محسوب می‌شود.

دومین مساله این است که PCA یک روش کاهش ابعاد خطی است که در تلاش برای بیشینه کردن واریانس و حفظ فاصله‌های زیاد دوتایی‌ها از یکدیگر است. به عبارت دیگر، چیزهایی که با یکدیگر متفاوت هستند به صورت دور از هم به پایان می‌رسند. این امر می‌تواند منجر به بصری‌سازی ضعیف به ویژه هنگام کار با ساختارهای غیرخطی خمینه می‌شود. ساختار خمینه می‌تواند به صورت یک شکل جغرافیایی مثلا استوانه، کره، منحنی و دیگر موارد باشد.

نظر به اینکه PCA به حفظ فاصله‌های دوتایی‌های بزرگ برای بیشینه کردن واریانس می‌پردازد، t-SNE با حفظ فاصله‌های کوچک دوتایی‌ها یا شباهت محلی از PCA متمایز می‌شود. لورنز تفاوت PCA و t-SNE را با استفاده از مجموعه داده

---

[15] Distributed Stochastic Neighbor Embedding
[16] Laurens van der Maatens
[17] Geoffrey Hinton
[18] PCA | Principal Components Analysis

Swiss Roll به خوبی در شکل زیر نشان داده. می‌توان مشاهده کرد که به دلیل غیرخطی بودن این مجموعه داده (خمینه) و حفظ فواصل بزرگ، PCA ساختار داده‌ها را به اشتباه حفظ می‌کند.

اکنون که مشخص شد چرا t-SNE ممکن است بر PCA غلبه کند، چگونگی عملکرد این الگوریتم تشریح می‌شود. الگوریتم t-SNE یک سنجه مشابهت را بین جفت نمونه‌ها در داده‌های ابعاد بالا و در فضای ابعاد کم محاسبه و سپس برای بهینه‌سازی این دو سنجه مشابهت با استفاده از یک تابع هزینه تلاش می‌کند. آنچه بیان شد را می‌توان به سه گام اولیه که در زیر بیان شده‌اند شکست.

۱. در گام اول، مشابهت بین نقاط در فضای ابعاد بالا اندازه‌گیری می‌شود. برای درک بهتر موضوع، می‌توان دسته‌ای از نقاط داده پراکنده شده در یک فضای دوبُعدی را مانند آنچه در شکل زیر آمده، در نظر گرفت. برای هر نقطه داده (xi) توزیع گاوسی حول محور آن نقطه توسط کاربر متمرکز می‌شود. سپس، چگالی همه نقاط (xj) تحت آن توزیع گاوسی محاسبه می‌شود. پس از آن، «بازبهنجارسازی»[19] برای همه نقاط داده انجام می‌شود. این امر یک مجموعه از احتمالات (Pij) برای کلیه نقاط داده را به دست می‌دهد. این احتمالات متناسب با مشابهت‌ها هستند. این در واقع بدان معناست که اگر نقطه داده 1x و 2x مقادیر یکسانی تحت دایره گاوسی داشته باشند، نسبت‌ها و مشابهت‌های آن‌ها مساوی و بنابراین مشابهت‌های محلی در ساختار فضای ابعاد بالا فراهم است. توزیع یا دایره گاوسی با استفاده از آنچه «سرگشتگی»[20] نامیده می‌شود، قابل دستکاری کردن محسوب می‌شود و واریانس توزیع (اندازه دایره) و اساسا تعداد نزدیک‌ترین همسایه‌ها را تحت تاثیر قرار می‌دهد. دامنه نرمال برای سرگشتگی بین ۵ و ۵۰ است.

۲. گام ۲ مشابه گام ۱ است، اما به جای توزیع گاوسی، توزیع «تی-استیودنت»[21] با یک درجه آزادی مورد استفاده قرار می‌گیرد که با عنوان «توزیع کوشی»[22] نیز شناخته می‌شود (به عبارت دیگر، هنگامی که درجه آزادی در توزیع تی-استیودنت برابر با یک باشد به توزیع کوشی تبدیل می‌شود). این کار دومین مجموعه از احتمالات (Qij) را در فضای ابعاد پایین به دست می‌دهد. چنانچه از تصویر زیر مشهود است، توزیع تی-استیودنت دارای دم سنگین‌تری نسبت به توزیع نرمال است. دم سنگین امکان مدل‌سازی بهتر فواصل طولانی‌تر از هم را فراهم می‌کند.

۳. گام آخر این است که این مجموعه از احتمالات از فضای ابعاد پایین (Qij) آن‌هایی که در فضای ابعاد بالای (Pij) قرار دارند را به بهترین شکل ممکن منعکس کنند. خواسته آن است که ساختار هر دو نقشه مشابه باشد. تفاوت بین توزیع‌های احتمال فضای دوبُعدی با استفاده از «معیار واگرایی کولبک-لیبلر»[23] قابل محاسبه است. در این مطلب به مبحث KL به طور جزئی پرداخته نخواهد شد و تنها به بیان این نکته که یک رویکرد نامتقارن است که به طور موثر مقادیر Pij و Qij را مقایسه می‌کند اکتفا خواهد شد. در نهایت، از «گرادیان نزولی» برای کاهش تابع هزینه KL استفاده می‌شود.

---

[19] renormalize
[20] Perplexity
[21] Student's t-distribution
[22] Cauchy distribution
[23] Kullback-Liebler divergence | KL

متدولوژی Feature Importance

انتخاب ویژگی را می‌توان به عنوان فرآیند شناسایی ویژگی‌های مرتبط و حذف ویژگی‌های غیر مرتبط و تکراری با هدف مشاهده زیرمجموعه‌ای از ویژگی‌ها که مساله را به خوبی و با حداقل کاهش درجه کارایی تشریح می‌کنند تعریف کرد. این کار مزایای گوناگونی دارد که برخی از آن‌ها در ادامه بیان شده‌اند.

بهبود کارایی الگوریتم‌های یادگیری ماشین

درک داده، کسب دانش درباره فرآیند و کمک به بصری‌سازی آن

کاهش داده کلی، محدود کردن نیازمندی‌ها ذخیره‌سازی و احتمالا کمک به کاهش هزینه‌ها

کاهش مجموعه ویژگی‌ها، ذخیره‌سازی منابع در دور بعدی گردآوری داده یا در طول بهره‌برداری

سادگی و قابلیت استفاده از مدل‌های ساده‌تر و کسب سرعت

به همه دلایل گفته شده، در سناریوهای «تحلیل کلان داده»، انتخاب ویژگی نقشی اساسی ایفا می‌کند.

ویژگی مرتبط

برای تشخیص یک «ویژگی مرتبط»[۲۴] با مساله، از این تعریف استفاده می‌شود: «یک ویژگی مرتبط است اگر شامل اطلاعاتی پیرامون هدف باشد». به بیان رسمی‌تر، «جان»[۲۵] و کوهاوی[۲۶] ویژگی‌ها را به سه دسته جدا از هم تقسیم کرده‌اند که «به شدت مرتبط» (strongly relevant)، «به طور ضعیف مرتبط» (weakly relevant) و «ویژگی غیرمرتبط» (irrelevant features) نامیده می‌شوند.

در رویکرد این پژوهشگران، ارتباط ویژگی X به صورت یک دسته‌بندی بیزی ایده‌آل تعریف می‌شود. ویژگی X، هنگامی که حذف آن منجر به آسیب دیدن صحت پیش‌بینی دسته‌بندی بیزیی ایده‌آل شود، به شدت مرتبط محسوب می‌شود. این ویژگی به طور ضعیف مرتبط نامیده می‌شود اگر به شدت مرتبط نباشد و یک زیرمجموعه از ویژگی‌های S وجود داشته باشد، به طوری که کارایی دسته‌بندی ایده‌آل بیزی روی S بدتر از کارایی $SU\{X\}$ باشد. یک ویژگی نامرتبط تعریف می‌شود اگر به شدت و به طور ضعیف مرتبط نباشد.

[24] Feature Relevance
[25] John
[26] Kohavi

افزونگی ویژگی

یک ویژگی معمولا در صورت وجود همبستگی بین ویژگی‌ها دارای افزونگی[۲۷] محسوب می‌شود. این مفهوم که دو ویژگی نسبت به هم دارای افزونگی هستند اگر مقادیر آن‌ها کاملا همبسته باشد توسط پژوهشگران زیادی پذیرفته شده، اما در عین حال امکان دارد تشخیص افزونگی ویژگی‌ها هنگامی که یک ویژگی با یک مجموعه از ویژگی‌ها مرتبط است کار ساده‌ای نباشد.

مطابق با تعریف ارائه شده توسط جان و کوهاوی، یک ویژگی در صورتی دارای افزونگی است و در نتیجه باید حذف شود که به طور ضعیف مرتبط و دارای پوشش مارکوف[۲۸] درون مجموعه ویژگی‌های کنونی باشد. از آنجا که ویژگی‌های غیرمرتبط باید به هر سو حذف شوند، پاک‌سازی آن‌ها بر اساس این تعریف انجام می‌شود.

[27] Feature Redundancy
[28] Markov blanket

تشریح داده‌ها

دیتاست[29] استفاده شده در پروژه، شامل ۱۸ ویژگی[30] است که در زیر به تشریح آن‌ها پرداخته شده است.

۱. شناسه مشتری
۲. موجودی حساب
۳. فرکانس به روزرسانی حساب
   اگر ۱ باشد یعنی به طور مرتب به روزرسانی می‌شود در غیر این صورت ۰
۴. مبلغ خرید
   کل مبلغ خریداری شده توسط حساب
۵. حداکثر مبلغ خرید
۶. مبلغ خرید اقساط انجام شده
۷. پول نقدی داده شده توسط مشتری
۸. فرکانس انجام خرید
   اگر ۱ باشد یعنی به طور مرتب خرید انجام می‌شود در غیر این صورت صفر
۹. فرکانس پرداخت کل مبلغ خرید به طور یکجا
   اگر ۱ باشد یعنی به طور مرتب خرید به طور کامل انجام می‌شود در غیر این صورت صفر
۱۰. فرکانس انجام خرید اقساط
   اگر ۱ باشد یعنی به طور مرتب خرید به طور اقساط انجام می‌شود در غیر این صورت صفر
۱۱. فرکانس استفاده از پول نقد
۱۲. تعداد تراکنش‌های انجام شده توسط پول نقد
۱۳. تعداد کل تراکنش‌های انجام شده
۱۴. محدودیت مبلغ تراکنش
۱۵. مبلغ کل خرید توسط مشتری
۱۶. حداقل مبلغ خرید
۱۷. درصد پرداخت کل مبلغ خرید به طور یک‌جا
۱۸. فهرست خدمات کارت اعتباری برای کاربر

[29] Credit Card Dataset for Clustering
[30] Attributes

پیش پردازش داده‌ها

ابتدا داده‌ها را وارد می‌کنیم، نوع داده‌های هر ستون را بررسی می‌کنیم. مطابق شکل شماره ۱ ، ستون CUST_ID را به علت Object بودن حذف می‌کنیم. البته دلیل اصلی این کار این است که شناسه‌ی کاربران به ما کمکی جهت شناسایی رفتار آنان نمی‌کند و نشأت گرفته از رفتار آنان نیست و همچنین اگر به صورت عددی دربیایند و در الگوریتم باشند، ممکن است دسته بندی را با خطا مواجه کنند.

```
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   CUST_ID                           8950 non-null   object
 1   BALANCE                           8950 non-null   float64
 2   BALANCE_FREQUENCY                 8950 non-null   float64
 3   PURCHASES                         8950 non-null   float64
 4   ONEOFF_PURCHASES                  8950 non-null   float64
 5   INSTALLMENTS_PURCHASES            8950 non-null   float64
 6   CASH_ADVANCE                      8950 non-null   float64
 7   PURCHASES_FREQUENCY               8950 non-null   float64
 8   ONEOFF_PURCHASES_FREQUENCY        8950 non-null   float64
 9   PURCHASES_INSTALLMENTS_FREQUENCY  8950 non-null   float64
 10  CASH_ADVANCE_FREQUENCY            8950 non-null   float64
 11  CASH_ADVANCE_TRX                  8950 non-null   int64
 12  PURCHASES_TRX                     8950 non-null   int64
 13  CREDIT_LIMIT                      8949 non-null   float64
 14  PAYMENTS                          8950 non-null   float64
 15  MINIMUM_PAYMENTS                  8637 non-null   float64
 16  PRC_FULL_PAYMENT                  8950 non-null   float64
 17  TENURE                            8950 non-null   int64
dtypes: float64(14), int64(3), object(1)
memory usage: 1.2+ MB
```

شکل شماره ۱

در مرحله بعد داده‌های تکراری[31] را بررسی می‌کنیم، هیچ داده تکراری در دیتاست موجود نیست.

سپس تعداد Nullها را بررسی می‌کنیم. شکل ۲ نتیجه را نشان می‌دهد. با توجه به اینکه درصد داده‌های خالی زیاد نیست، به راحتی برای پر کردن آن‌ها از میانگین هر ستون استفاده می‌کنیم. همچنین به همان دلیل ذکرشده، نیازی به حذف تاپل‌ها نیست.

[31] Duplications

```
BALANCE                             0
BALANCE_FREQUENCY                   0
PURCHASES                           0
ONEOFF_PURCHASES                    0
INSTALLMENTS_PURCHASES              0
CASH_ADVANCE                        0
PURCHASES_FREQUENCY                 0
ONEOFF_PURCHASES_FREQUENCY          0
PURCHASES_INSTALLMENTS_FREQUENCY    0
CASH_ADVANCE_FREQUENCY              0
CASH_ADVANCE_TRX                    0
PURCHASES_TRX                       0
CREDIT_LIMIT                        1
PAYMENTS                            0
MINIMUM_PAYMENTS                  313
PRC_FULL_PAYMENT                    0
TENURE                              0
dtype: int64
```

شکل شماره ۲

سپس از دستور describe برای توصیف داده‌ها استفاده می‌کنیم. شکل شماره ۳، توصیفی آماره از دیتاست است. با توجه به نتایج حاصل شده از این شکل، لازم است که داده‌های پرت را حذف کنیم تا انحراف استاندارد داده‌ها را کاهش دهیم.

برای حذف داده‌های پرت می‌توانیم از دو روش IQR [32] و Z-Score [33] استفاده کنیم. استفاده از روش اول باعث حذف ۷۵ درصد داده‌ها می‌شود در صورتی که روش دوم ۱۷ درصد داده‌ها را حذف می‌کند که به همین علت روش دوم را انتخاب می‌کنیم.

در این مرحله یک دیتا فریم جدید تعریف کرده و در ادامه‌ی کار از آن استفاده می‌کنیم.

به منظور یافتن الگوها و استفاده‌ی بهتر از داده‌ها، نیاز است که داده‌ها را نرمال کنیم.

برای حذف تکرار[34] در ستون‌ها[35] مقدار ضریب همبستگی را برای تمام جفت ویژگی‌ها رسم می‌کنیم. شکل شماره ۴ نتیجه کار است.

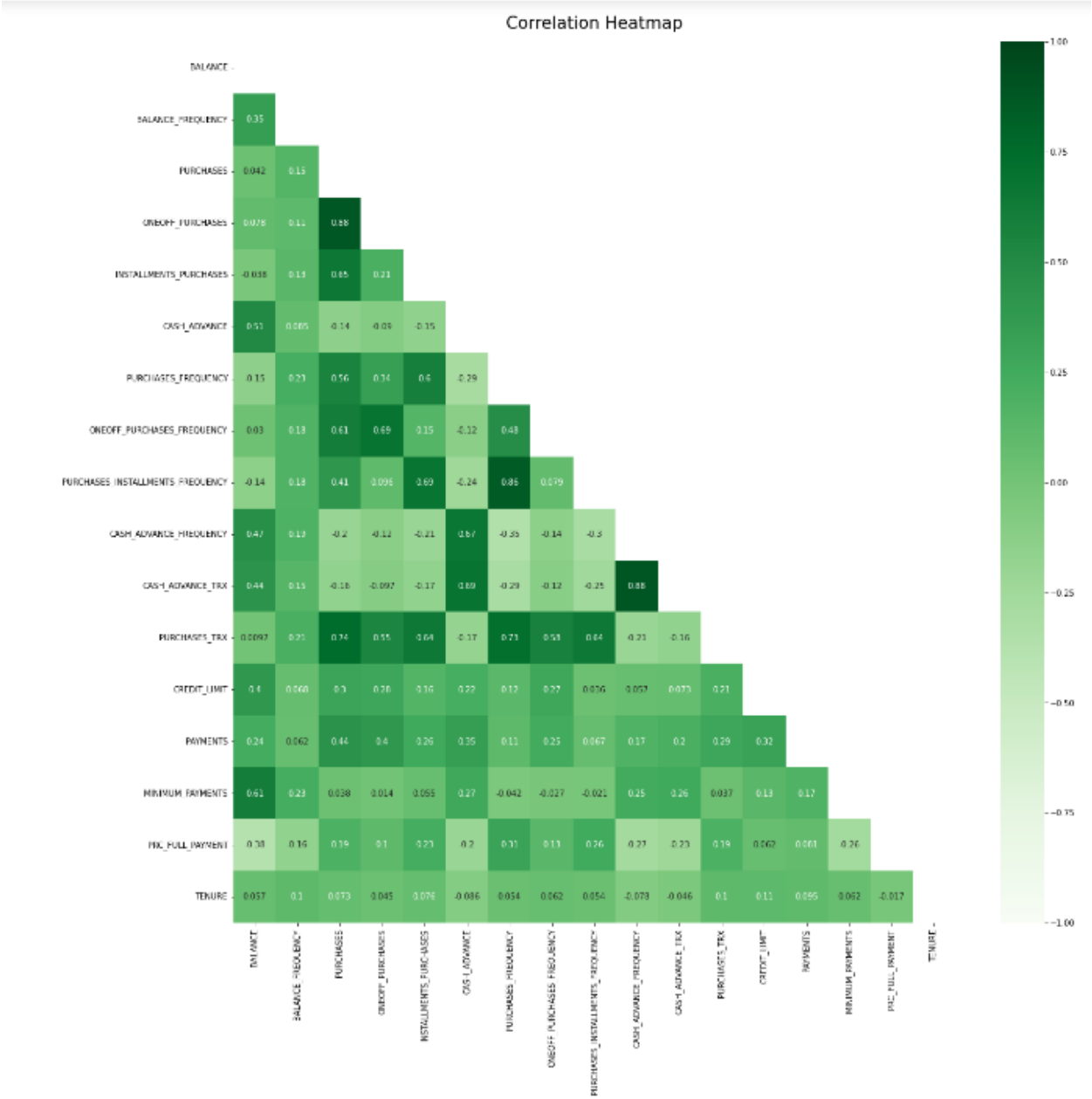
[32] IQR Methods
[33] Z-Score Method
[34] Redundancy
[35] Attributes

	count	mean	std	min	25%	50%	75%	max
BALANCE	8950.0	1564.474828	2081.531879	0.000000	128.281915	873.385231	2054.140036	19043.13856
BALANCE_FREQUENCY	8950.0	0.877271	0.236904	0.000000	0.888889	1.000000	1.000000	1.00000
PURCHASES	8950.0	1003.204834	2136.634782	0.000000	39.635000	361.280000	1110.130000	49039.57000
ONEOFF_PURCHASES	8950.0	592.437371	1659.887917	0.000000	0.000000	38.000000	577.405000	40761.25000
INSTALLMENTS_PURCHASES	8950.0	411.067645	904.338115	0.000000	0.000000	89.000000	468.637500	22500.00000
CASH_ADVANCE	8950.0	978.871112	2097.163877	0.000000	0.000000	0.000000	1113.821139	47137.21176
PURCHASES_FREQUENCY	8950.0	0.490351	0.401371	0.000000	0.083333	0.500000	0.916667	1.00000
ONEOFF_PURCHASES_FREQUENCY	8950.0	0.202458	0.298336	0.000000	0.000000	0.083333	0.300000	1.00000
PURCHASES_INSTALLMENTS_FREQUENCY	8950.0	0.364437	0.397448	0.000000	0.000000	0.166667	0.750000	1.00000
CASH_ADVANCE_FREQUENCY	8950.0	0.135144	0.200121	0.000000	0.000000	0.000000	0.222222	1.50000
CASH_ADVANCE_TRX	8950.0	3.248827	6.824647	0.000000	0.000000	0.000000	4.000000	123.00000
PURCHASES_TRX	8950.0	14.709832	24.857649	0.000000	1.000000	7.000000	17.000000	358.00000
CREDIT_LIMIT	8950.0	4494.449450	3638.612411	50.000000	1600.000000	3000.000000	6500.000000	30000.00000
PAYMENTS	8950.0	1733.143852	2895.063757	0.000000	383.276166	856.901546	1901.134317	50721.48336
MINIMUM_PAYMENTS	8950.0	864.206542	2330.588021	0.019163	170.857654	335.628312	864.206542	76406.20752
PRC_FULL_PAYMENT	8950.0	0.153715	0.292499	0.000000	0.000000	0.000000	0.142857	1.00000
TENURE	8950.0	11.517318	1.338331	6.000000	12.000000	12.000000	12.000000	12.00000

شکل شماره ۳

با توجه به این موضوع که ضریب همبستگی هیچ‌کدام از جفت داده‌ها بزرگ‌تر از ۰.۹ نیست، نیازی به حذف هیچکدام از ستون‌ها نداریم و می‌توانیم از تمام ستون‌ها استفاده کنیم.

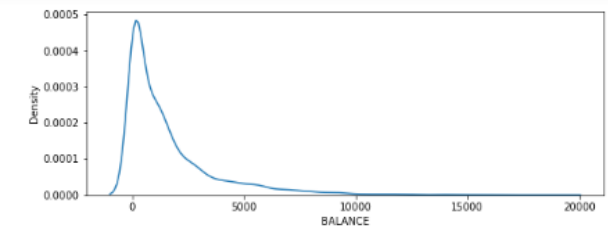
داده‌های این مرحله در فایل "Preproccessed.csv" موجود است.

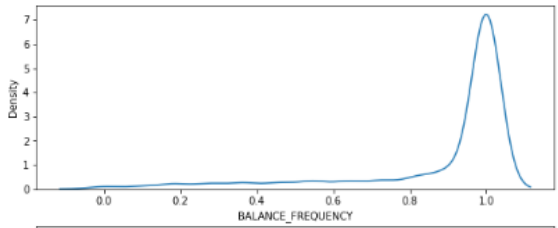

Correlation Heatmap
BALANCE
BALANCE_FREQUENCY
PURCHASES
ONEOFF_PURCHASES
INSTALLMENTS_PURCHASES
CASH_ADVANCE
PURCHASES_FREQUENCY
ONEOFF_PURCHASES_FREQUENCY
PURCHASES_INSTALLMENTS_FREQUENCY
CASH_ADVANCE_FREQUENCY
CASH_ADVANCE_TRX
PURCHASES_TRX
CREDIT_LIMIT
PAYMENTS
MINIMUM_PAYMENTS
PRC_FULL_PAYMENT
TENURE
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

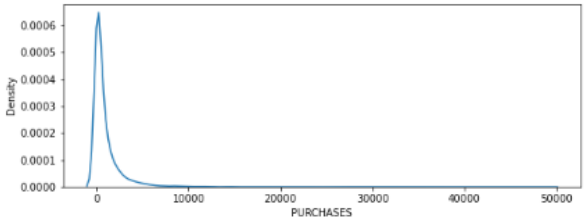
شکل شماره ۴

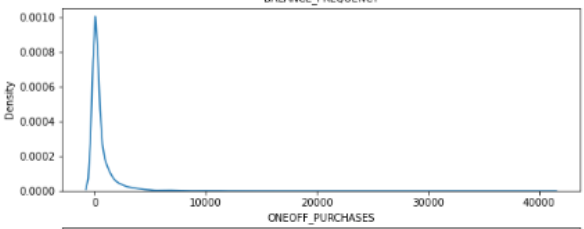
مصور کردن داده‌ها

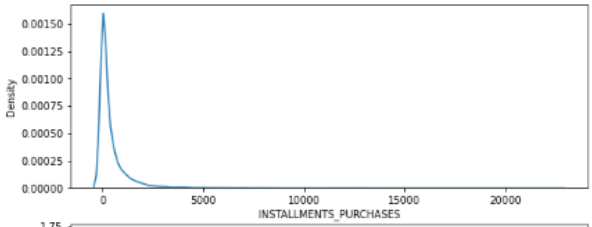
در این قسمت، برای هر ۱۷ ستون، نمودار فراوانی رسم می‌کنیم.

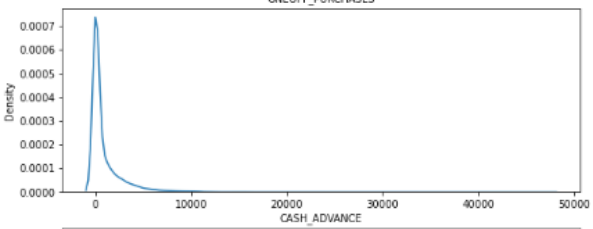


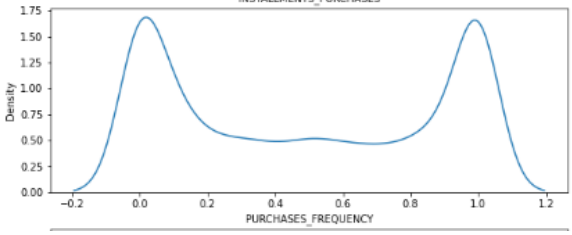


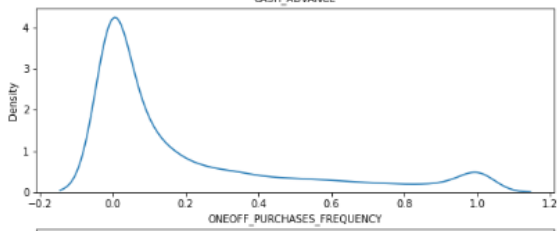


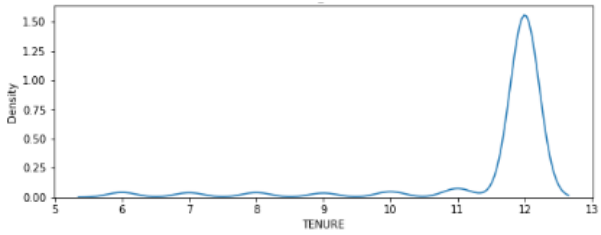


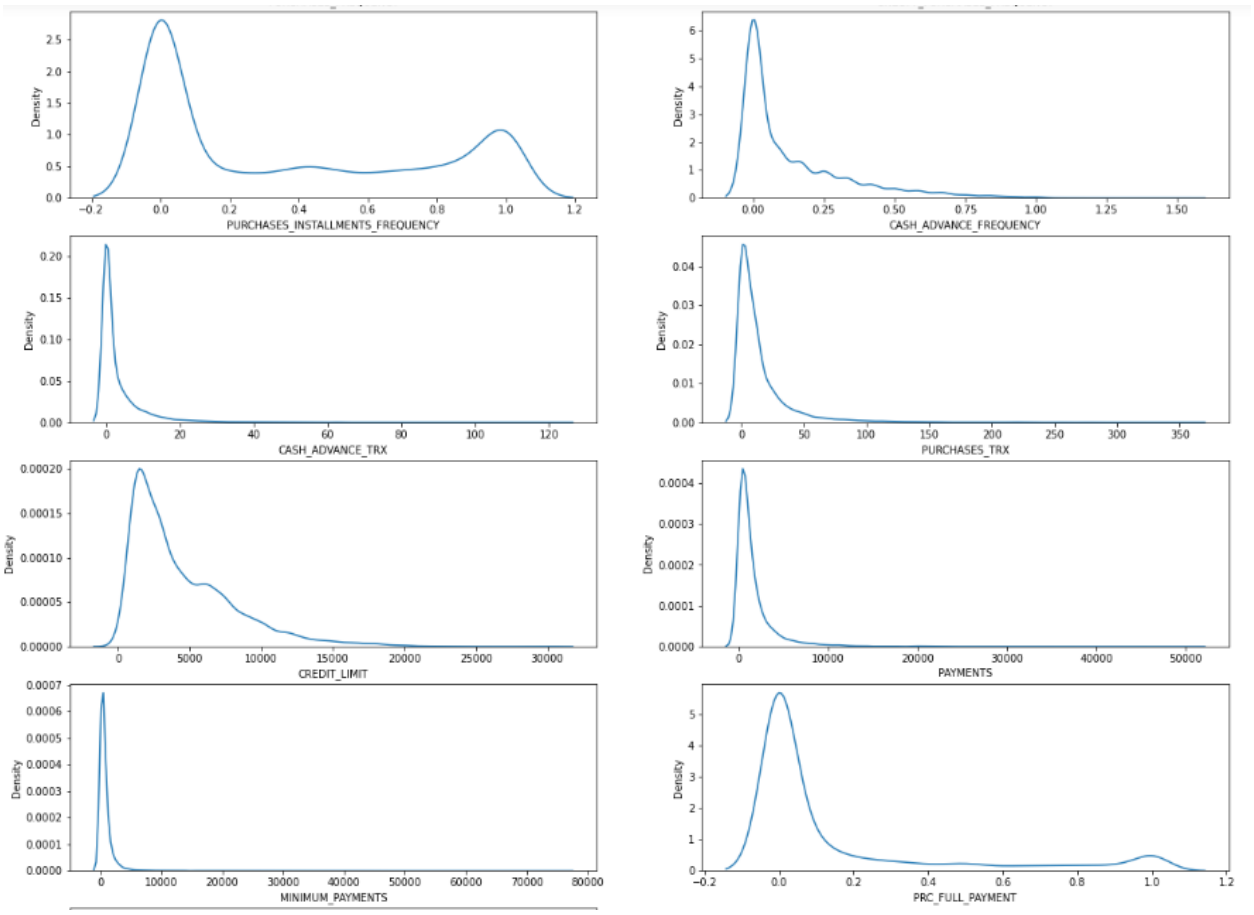








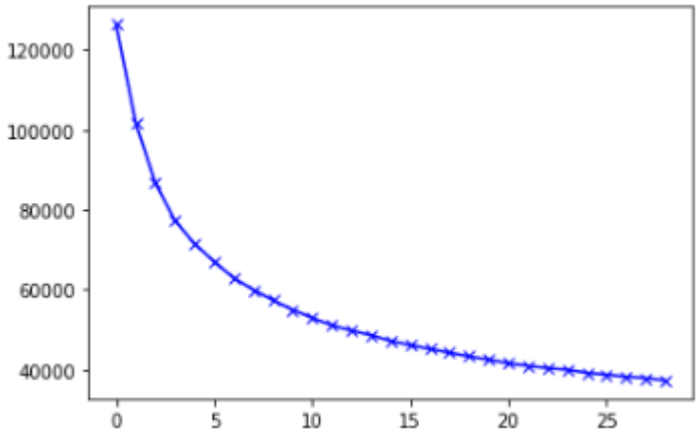



Density
PURCHASES_INSTALLMENTS_FREQUENCY
CASH_ADVANCE_FREQUENCY
CASH_ADVANCE_TRX
PURCHASES_TRX
CREDIT_LIMIT
PAYMENTS
MINIMUM_PAYMENTS
PRC_FULL_PAYMENT

داده‌کاوی و شناسایی الگوهای پنهان

روش KMeans

در این روش ابتدا لازم است که کقدار K را بیابیم. در شکل شماره ۵ این موضوع روشن است که مقدار ۶ مناسب است.

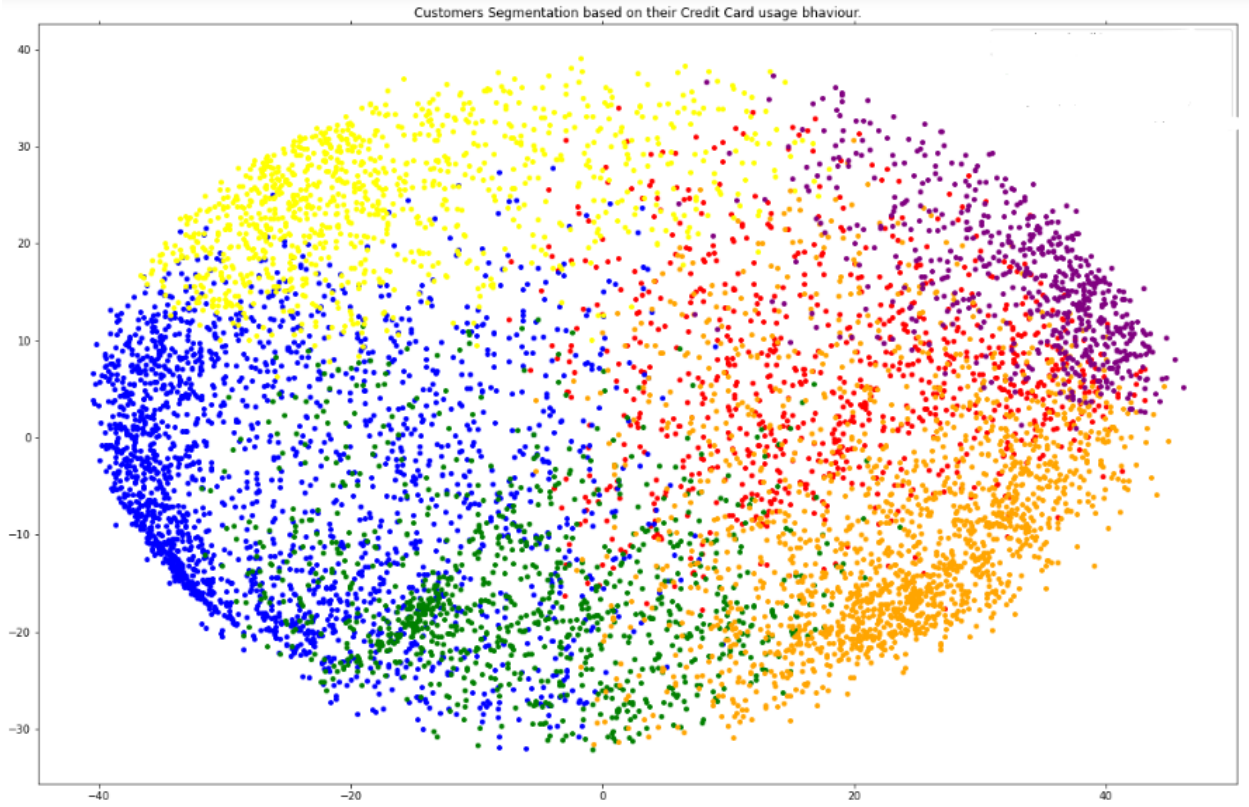


شکل شماره ۵

حال این روش را با استفاده از تابع موجود در پایتون اجرا می‌کنیم. داده‌های موجود به ۶ دسته تقسیم بندی می‌شوند که در فایل "Final1.csv" این خوشه بندی قابل مشاهده است.

۱. افرادی که همه نوع خریدی انجام می‌دهند
۲. افرادی که پرداخت دو مرحله‌ای انجام می‌دهند
۳. افرادی که به صورت اقساط خرید انجام می‌دهند
۴. افرادی که خرید را با پول نقد انجام می‌دهند
۵. افرادی که خریدهای گران انجام می‌دهند
۶. افرادی که مبلغ پایینی خرید انجام می‌دهند

شکل شماره ۶ این خوشه بندی را نشان می‌دهد


Customers Segmentation based on their Credit Card usage bhaviour.
40
30
20
10
0
−10
−20
−30
−40
−20
0
20
40

شکل شماره ۶

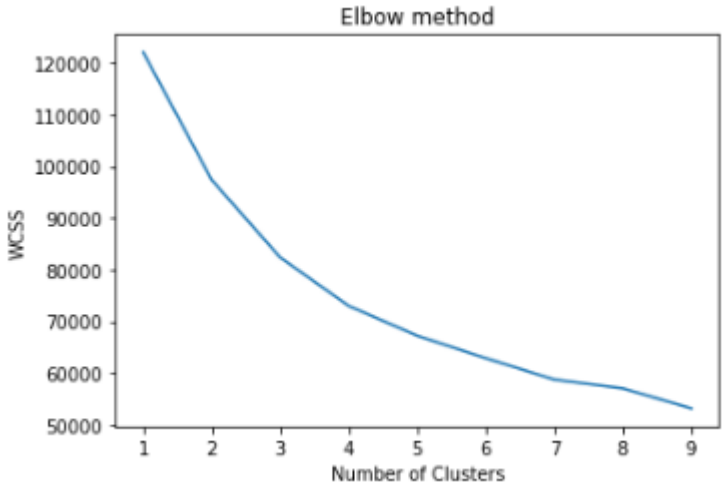
روش PCA

در این روش، با استفاده از کاهش تعداد متغیرهای تصمیم، دو متغیری که بیشترین واریانس با یکدیگر دارند را انتخاب می‌کنیم. شکل شماره ۷ نشان دهنده کاهش بعد[36] است. در مرحله بعدی با استفاده از روش Elbow method تعداد خوشه‌ها را می‌یابیم. در شکل شماره ۸ نقاط ۲ تا ۴ در حال شکست هستند. برای بررسی بیشتر تابع silhouette را برای تعداد خوشه‌های ۲ و ۳ و ۴ محاسبه می‌کنیم و نمودار آن را در شکل شماره ۹ مشاهده می‌کنیم. بیشترین امتیاز برای تعداد خوشه ۳ می‌باشد.

	x	y	label
0	-21.633944	-27.196819	1
1	-17.866336	24.351481	3
2	14.175958	10.286162	0
3	-11.380663	3.569994	1
4	-26.246978	-19.816754	1
...	...	...	...
7429	-4.415311	-2.434460	1
7430	5.853298	-9.683619	4
7431	-17.100110	-2.113640	1
7432	5.476105	-8.172672	4
7433	-15.916147	0.332234	1

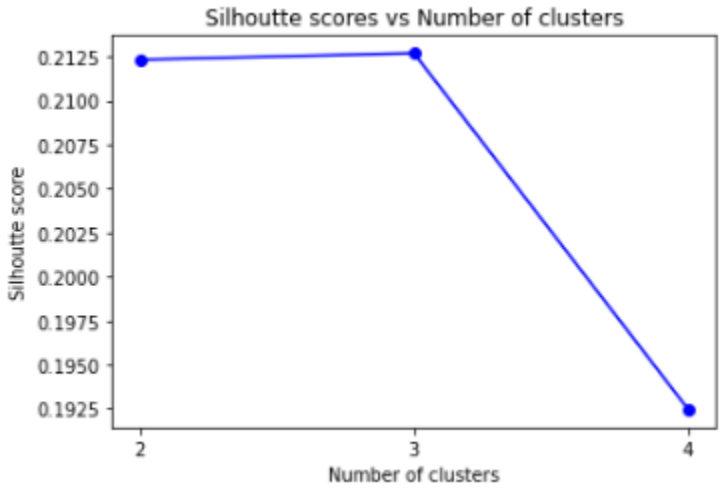
7434 rows × 3 columns

شکل شماره ۷



[36] Dimension reduction

شکل شماره ۸



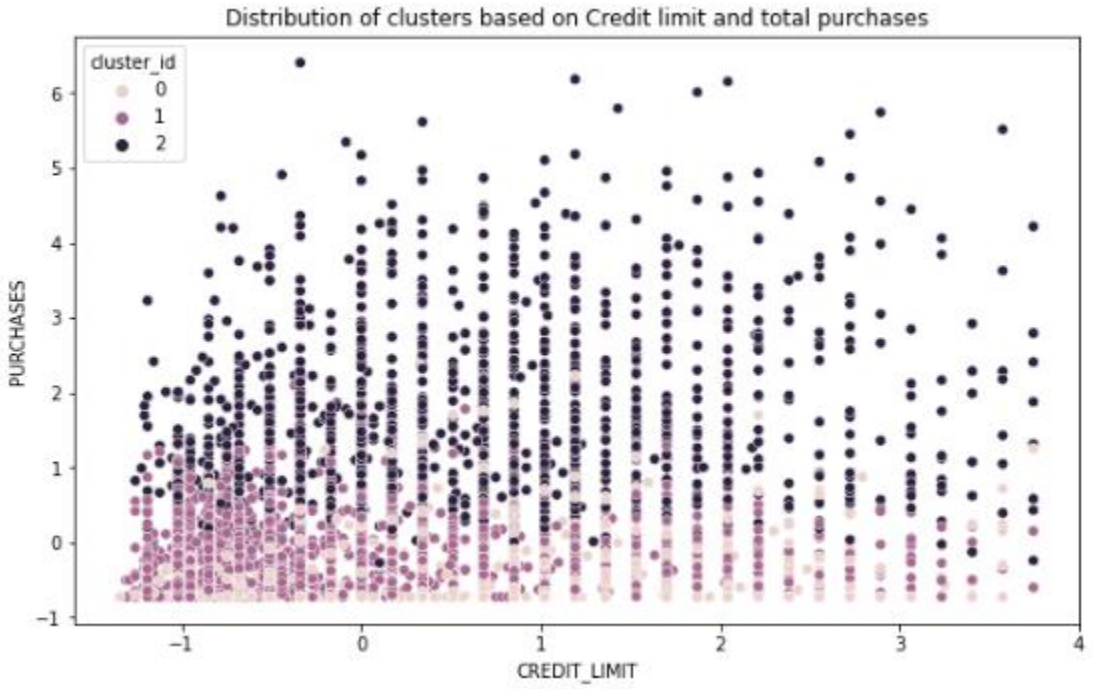
شکل شماره ۹

خروجی این روش ۳ خوشه است که در فایل "Final2.csv" قابل مشاهده است.

در نمودار شماره ۱۰ توزیع خوشه‌ها را بر اساس ستون خرید و ستون خرید به صورت یک مرحله‌ای مشاهده می‌کنیم.

شکل شماره ۱۰

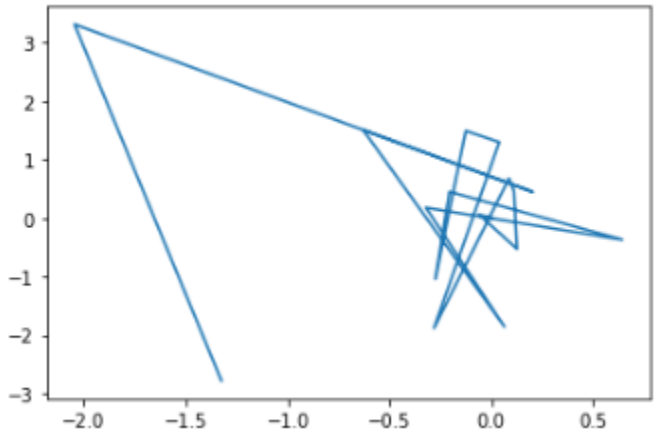
شکل شماره ۱۱ ستون‌های خرید و محدودیت خرید مورد بررسی قرار گرفته‌اند.



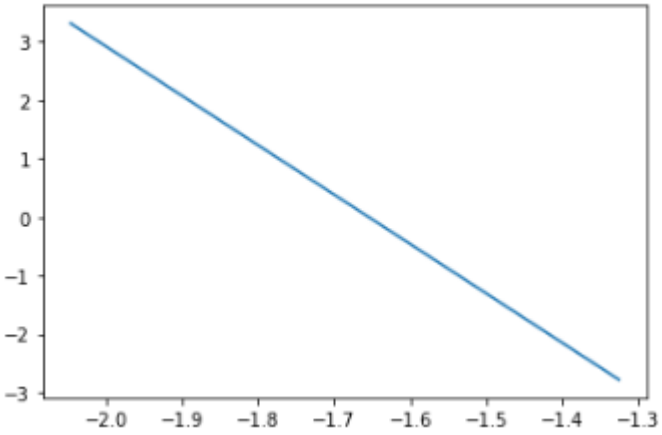
شکل شماره ۱۱

روش TSNE

در این روش ابتدا با استفاده از pca نموداری از پراکندگی کل ستون‌ها را رسم می‌کنیم که در شکل شماره ۱۲ قابل مشاهده است. در مرحله بعد با استفاده از کاهش ابعاد نمودار شماره ۱۲ به شکل شماره ۱۳ تبدیل می‌شود.

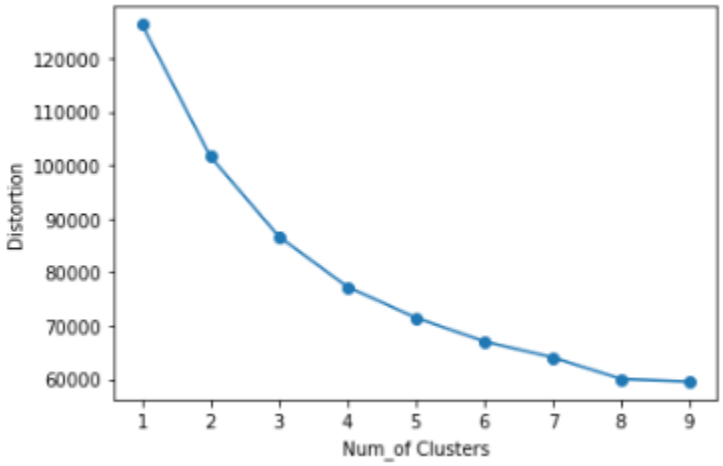


شکل شماره ۱۲



شکل شماره ۱۳

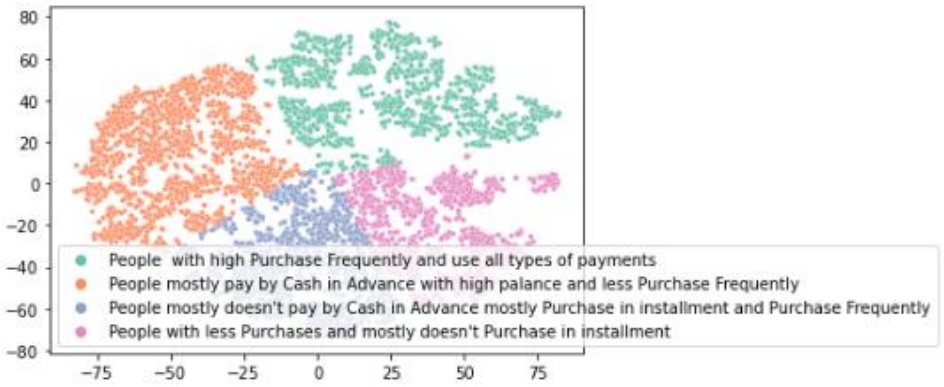
در این مرحله به سراغ یافتن تعداد خوشه‌ها می‌رویم و با مشاهده شکل شماره ۱۴ در N=4 مواجه با شکست شدیم پس ۴ خوشه خواهیم داشت.



شکل شماره ۱۴

حال می‌توانیم خوشه بندی را انجام دهیم. نتیجه‌ی این خوشه بندی در شکل شماره ۱۵ آمده است. تفسیر این نمودار در زیر آمده است:

۱. افرادی که به صورت پر تکرار خرید می‌کنند و از پول نقد استفاده نمی‌کنند، خرید به صورت اقساط.
۲. افرادی که به صورت پر تکرار خرید می‌کنند و از انواع پرداخت استفاده می‌کنند.
۳. افرادی که به تمام حالات ممکن پرداخت می‌کنند، پر تکرار خرید می‌کنند.
۴. افرادی که به صورت اقساط خرید نمی‌کنند و خرید زیادی ندارند.

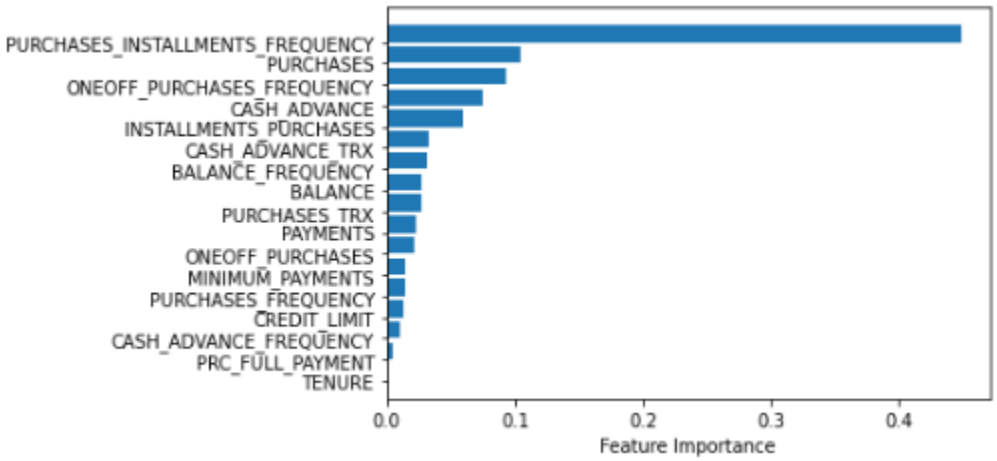


شکل شماره ۱۵

ارزیابی الگوهای شناسایی شده

خوشه‌بندی شماره ۱

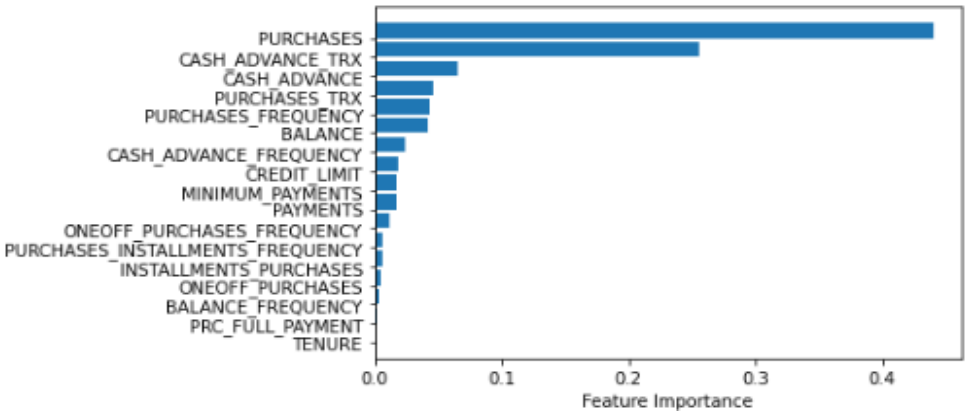
طبق مسائل مطرح شده در بخش مرور ادبیات با رسم نمودار Feature Importance متوجه می‌شویم در این خوشه بندی ستون PURCHASES_INSTALLMENTS_FREQUENCY بیشترین تاثیر را داشته است که با توجه به نتیجه‌گیری انجام شده صحت این موضوع اثبات می‌شود



شکل شماره ۱۶

خوشه‌بندی شماره ۲

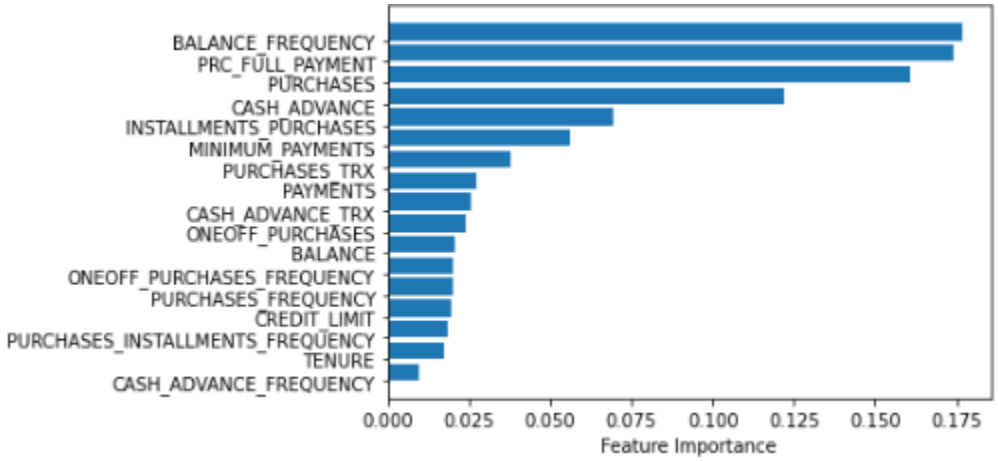
طبق مسائل مطرح شده در بخش مرور ادبیات با رسم نمودار Feature Importance متوجه می‌شویم در این خوشه بندی ستون CASH_ADVANCE_TRX و PURCHASES بیشترین تاثیر را داشته است که با توجه به نتیجه‌گیری انجام شده صحت این موضوع اثبات می‌شود



شکل شماره ۱۷

خوشه‌بندی شماره ۳

طبق مسائل مطرح شده در بخش مرور ادبیات با رسم نمودار Feature Importance متوجه می‌شویم در این خوشه بندی ستون BALANCE_FREQUENCYو PRC_FULL_PAYMENTبیشترین تاثیر را داشته است که با توجه به نتیجه‌گیری انجام شده صحت این موضوع اثبات می‌شود



شکل شماره ۱۸

نتیجه‌گیری

در این پروژه با استفاده از ۳ روش مختلف، دیتاست مربوط به تراکنش‌های مشتری‌های بانک را بررسی کردیم. این نکته بسیار حائز اهمیت است که روی داده‌ها پیش‌پردازش‌هایی از جمله جایگزینی داده‌های خالی و حذف داده‌های پرت و... . همچنین استاندارد کردن داده‌ها برای انجام خوشه‌بندی بسیار ضروری است.

الگوریتم K-Means یک الگوریتم بسیار بهینه است و مناسب برای داده‌هایی است که تعداد بعد زیادی دارند اما این نکته بسیار مهم است که انتخاب K نقش زیادی را در این الگوریتم ایفا می‌کنند. از طرفی اگر داده‌ها نویز زیادی داشته باشند این روش مناسب نیست.

در روش PCA، از بین تمام ستون‌ها لازم است دو ستون با واریانس و پراکندگی بالا را انتخاب کرده و با این کاهش ابعاد می‌توانیم خوشه‌بندی را راحت‌تر کنیم. این روش برای داده‌ها با حجم کم اما نویز زیاد مناسب است. چرا که پارامترهای مورد نیاز برای خوشه‌بندی روش مناسبی دارد. موارد ذکر شده به معنی از دست رفتن داده‌ها[۳۷] نیست.

لازم به ذکر است که PCA یک روش برای خوشه‌بندی نیست اما می‌تواند خوشه‌بندی را بهتر کند و جواب را بهبود ببخشد. در واقع می‌توان به PCA به عنوان پیش‌نیاز روش K-Means نگاه کرد تا بتوان تعداد خوشه‌ها را به طور صحیح‌تری انتخاب کرد.

حال به این پروژه از دید دیگری نگاه می‌کنیم. تقسیم بندی مشتری[۳۸] نقش بزرگی را در کسب‌وکار ایفا می‌کند. در صورتی که این خوشه‌بندی با وجود unsupervised بودن خوب عمل کند و بتواند مشتری‌ها را به گروه‌هایی تقسیم کند که بتواند بر اساس آن‌ها برنامه ریزی کند، بانک می‌تواند سیاست‌گذاری پر سودتری انجام دهد و استراتژی‌های مناسب‌تری را اتخاذ کند.

خوشه‌بندی سبب می‌شود الگوهایی که در داده‌های خام پنهان هستند را بیابیم. الگوهایی که به راحتی قابل دریافت نیستند و برای رسیدن به آن‌ها محاسبات زیادی لازم است انجام شود.

[37] loss of information
[38] Customer segment

منابع و مراجع

1. *https://hamruyesh.com/what-is-data-analysis-data-mining-vs-data-analysis-gudie/*
2. *https://blog.faradars.org/%D8%AF%D8%A7%D8%AF%D9%87-%DA%A9%D8%A7%D9%88%DB%8C-data-mining-%D8%A7%D8%B2-%D8%B5%D9%81%D8%B1-%D8%AA%D8%A7-%D8%B5%D8%AF/*
3. *Knowledge Discovery and Data Mining: Towards a Unifying Framework.*
4. *Introduction to knowledge discovery and data mining*[39]
5. *https://blog.faradars.org/feature-selection-and-feature-extraction/*
6. *https://virgool.io/@TabaMojj/%D8%A8%D8%B1%D8%B1%D8%B3%DB%8C-dimensionality-reduction-%D8%AF%D8%B1-%D9%85%D8%A7%D8%B4%DB%8C%D9%86-%D9%84%D8%B1%D9%86%DB%8C%D9%86%DA%AF-nt43owl9vumt*
7. *https://raahbord.com/k-means/*
8. *https://blog.faradars.org/practical-guide-principal-component-analysis-python-r/*
9. https://ieeexplore.ieee.org/abstract/document/1578784/
10. https://www.researchgate.net/profile/Azad-Abdulhafedh/publication/349094412_Incorporating_K-means_Hierarchical_Clustering_and_PCA_in_Customer_Segmentation/links/601f494292851c4ed554724d/Incorporating-K-means-Hierarchical-Clustering-and-PCA-in-Customer-Segmentation.pdf
11. *https://upg-bulletin-se.ro/old_site/archive/2010-3/7.%20Schiopu.pdf*

[39] PDF