به نام خدا

عنوان پروژه: دسته بندی مشتریان بر اساس رفتار آنها جهت ارائه خدمات

گردآورندگان:مهندس مونا رستگار، مهندس اشکان بویری

شرح پروژه:

- برای اینکه کار کردن با ستون ها راخت تر به وسیله پایتون انجام بشه نام آنها با معادل انگلیسی شان جایگزین شده
- نوع بعضی از ستون ها از فلوت٦٤[1]به اینت ٦٤ [2]تغییر داده شده.
- همبستگی بین ویژگی ها بدست آمده.
- مقادیر نامعلوم[3]در میان داده ها وجود داشته که از داده ها حذف شدند.
- با رسم نمودار های متوجه شدیم که ویژگی ها زیر نه تنها تاثیر چندانی روی نتیجه نهایی مدل نخواهند داشت بلکه مدل رو دچار اشتباه خواهند کرد و هزینه محاسباتی اضافی را ایجاد می کنند با توجه به موارد ذکر شده این ستون ها از داده ها حذف شدند:
  میانگین فاصله بین روزهای انجام تراکنش و تعداد روز عدم فعالیت از آخرین تراکنش تا تاریخ روز و تعداد روز از افتتاح سپرده تا تاریخ آخرین تراکنش و ردیف و تعداد حساب و مانده در تاریخ ۳۱-٦-۹۸

- از ویژگی های خام دیتاست، ویژگی های جدید ، موثرتری استخراج شدند. با استفاده از تفاضل بین مقادیر ستون تعداد تراکنش بستانکار از ستون تعداد تراکنش بدهکار ستونی

[1] Float64
[2] Int64
[3] Nan

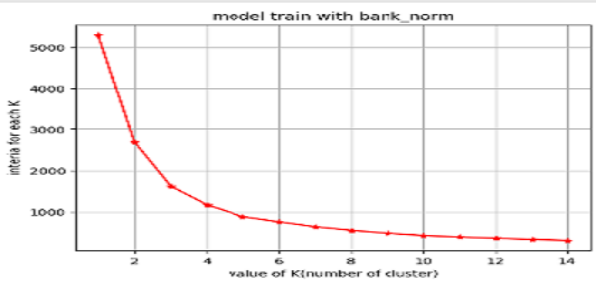
تحت نام دی بی ان سی دی[4]ساختم که بر خالف انتظارم کرولیشن باالای با هیچ کدوم از ویژگی ها نداره اما این ستون را حذف نکردیم و ستون دیگه ای از تفاضل مقادیر میانگین تراکنش های بستانکار از ستون میانگین تراکنش بدهکار تحت نام دی اف ای سی دی[5] ساختم این ستون کرولیشن مناسب با ستون میانگین موجودی(۹۵,۰)دارد،ویژگی های دیگری را میتوان از میان داده ها موجوداستخراج کرد.

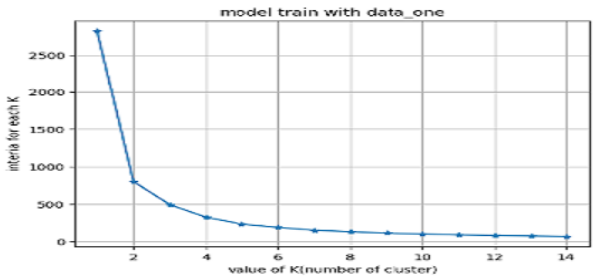
- از آنجاکه ما در این پروژه از الگوریتم های خوشه بندی استفاده خواهیم کرد و این الگوریتم ها به مقیاس داده ها حساسیت بالای دارند باید داده هایمان در یک مقیاس باشند تا این الگوریتم ها عملکرد بهتری داشته باشند ما از هر دو روش استانداردسازی[6]و نرمالیز کردن[7] استفاده می کنیم یا استفاده از استاندارسازی مجموعه داده بانک_اس سی[8]بدست آمده و با استفاده از نرمالیز مجموعه داده بانک_نرم بدست آمده که ما برای آموزش مدل ها از داده های بانک_نرم[9]استفاده میکنیم.
- برای ارزیابی خوشه بندی های صورت گرفته با الگوریتم خوشه بند کی مینز و انتخاب تعداد خوشه مناسب از معیار اینرسی[10]و روش آرنج استفاده کردیم همچنین برای سنجش میزان اعتبار خوشه بندی با تعداد خوشه های مختلف در محدوده ۲تا۱۵ از معیار سایلیت اسکور[11]استفاده کرده ایم
- در گام بعدی برای اینکه دید بهتری از خوشه بندی های انجام شده بگیری از نمودار اسکتر پلات استفاده کردیم.
- سپس ارتباط بین ویژگی ها و خوشه ها را با معیار کرولیشن بدست آوردیم.

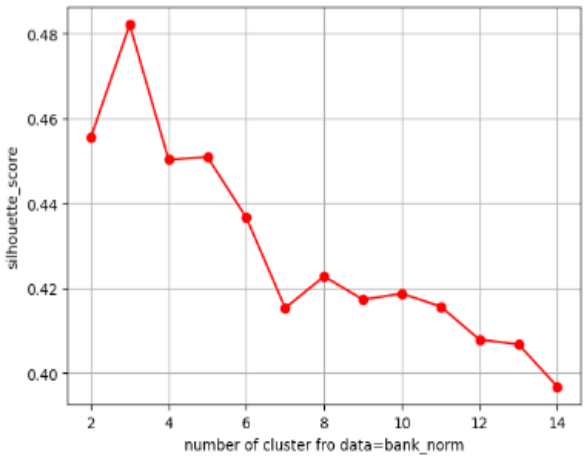
---

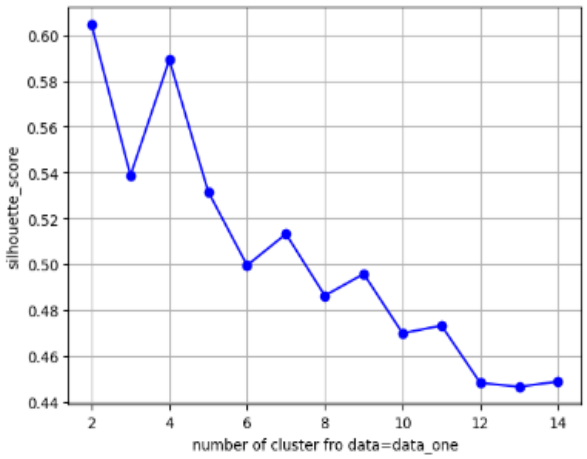
[4] DBNCD
[5] DFACD
[6] Standard Scaler
[7] Normalize
[8] Bank_sc
[9] Bank_norm
[10] Interia
[11] Silhouette Score

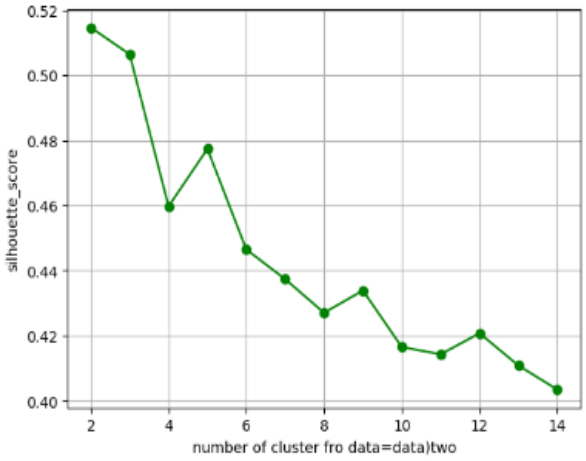
- با تحلیل های صورت گرفته به این نتیجه رسیدیم که داده های موجود در خوشه دو نشان دهنده افراد خوش حساب بوده و داده های خوشه یک و صفر به ترتیب هر کدام نماین گر میان حساب و بد حساب می باشند.



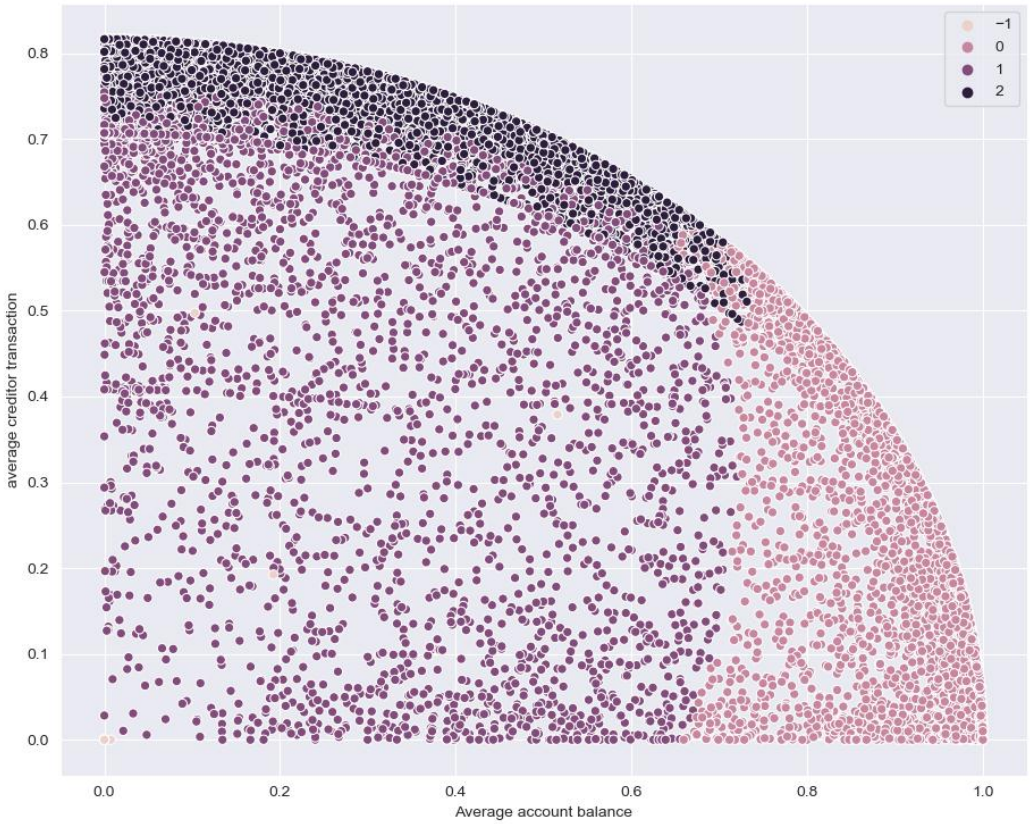








- ۱۱ ردیف از داده ها با استفاده از الگوریتم دی بی اسکن نویز تشخیص داده شدند و از مجموعه داده ها حذف شدند.
- همچنین الگوریتم دی بی اسکن ترتیب خوشه ها را عوض کرده بدین صورت که داده ها در خوشه صفر، یک و دو به ترتیب بیان گر خوش حساب ها،میان حساب ها و بد حساب ها می باشند.



- در بخش دسته بندی از داده های بانک_نرم_دی بی استفاده کرده ایم، با استفاده از تکنیک های داده کاوی زیر مجموعه ای از داده ها ایجاد کرده والگوریتم های دسته بندی مختلفی مانند جنگل تصادفی،ماشین بردار پشتیبان و.... را که پارامتر های این الگوریتم ها با استفاده از تابع گرید سرچ سی وی بهینه شده است با استفاده از داده های انتخاب شده آموزش دادیم. مدل های توسعه داده شده همگی دقت و عملکرد بالایی داشته.