دانشکده مهندسی کامپیوتر دانشگاه اصفهان

مستندات پروژه اول درس ماشین یادگیری

محمد مولوی

۹۹۳۶۱۳۰۵۷

خرداد ۱۴۰۳

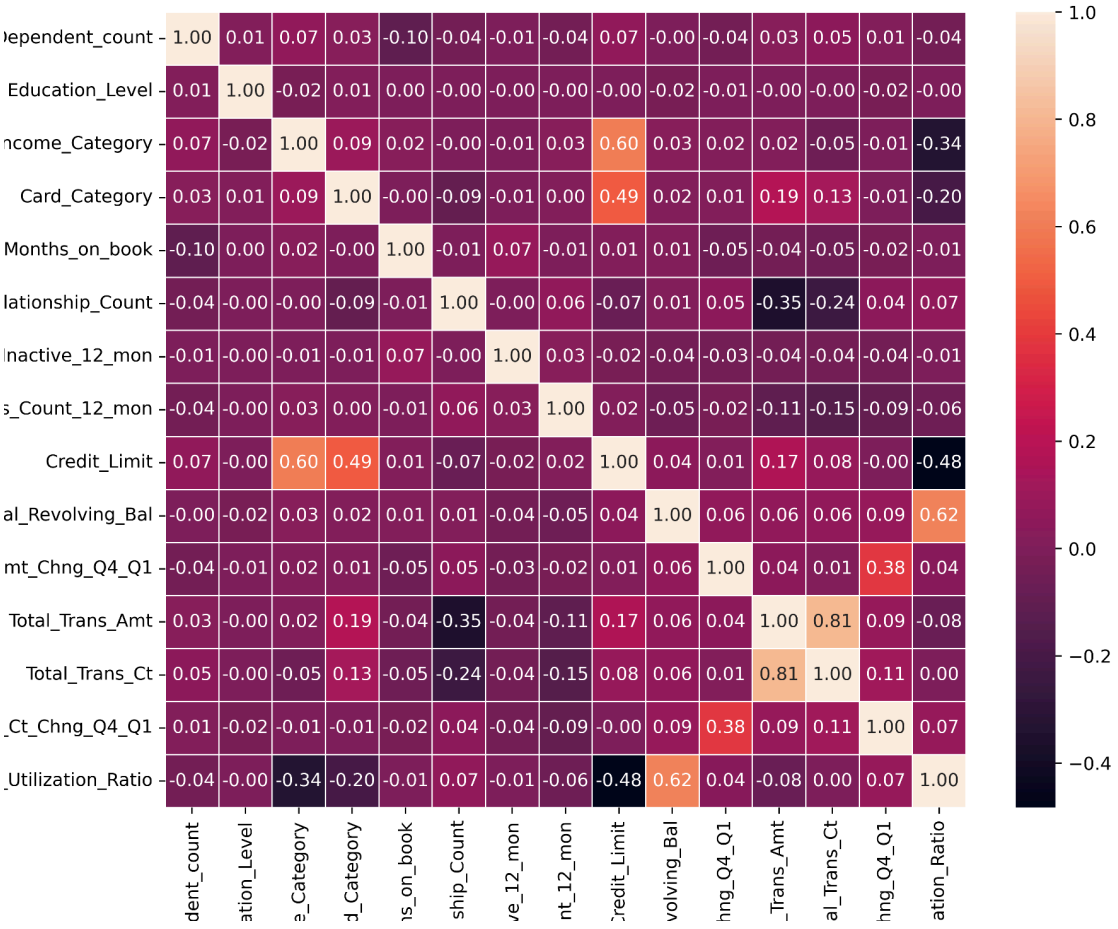
ابتدا برای خواندن فایل مقدار هایی که برابر unknown هستند را برابر nan در نظر میگیریم زیرا آنها نیز همانند مقدار های گم شده اند.

ابتدا داده های غیر مرتبط یا کم ربط را حذف میکنیم. برای اینکار ستون های

'Customer_Age','Marital_Status','Gender','Unnamed: 19','CLIENTNUM'

را حذف میکنیم.

سپس از ستون های correlation گرفته شد تا مقدار وابستگی هر ستون نسبت به credit limit که مقدار هدف است مشخص شود.



سه ستون Avg_Utilization_Ratio , income category , card category بیشترین وابستگی به مقدار هدف را دارند بنابراین میتوانند حاوی مقادیر مهمی باشند. پس برای پر کردن این ستون ها بهتر است روشی مانند clustering استفاده کنیم. در این پروژه از kmeans clustering برای پر کردن مقادیر این ستون ها استفاده شده و برای پر کردن باقی ستون ها از پرتکرار ترین مقدار همان ستون برای پر کردن استفاده

شده. (به طور حدسی اهمیت month of book را نیز زیاد در نظر گرفتم بنابراین برای پر کردن مقدار های خالی آن نیز از kmeans استفاده کردم)

استفاده از kmeans برای پر کردن مقادیر خالی به این صورت انجام شده که تعداد کل کلاستر های ۱۰ تا در نظر گرفته شده است سپس خوشه بندی بدون ستون های خالی(, card category , income category month of book) انجام شده است. سپس پر کردن این مقادیر خالی به این صورت انجام شده که مقادیر خالی هر سطر را برابر پر تکرار ترین مقدار همان ستون در خوشه در نظر گرفته شده است.

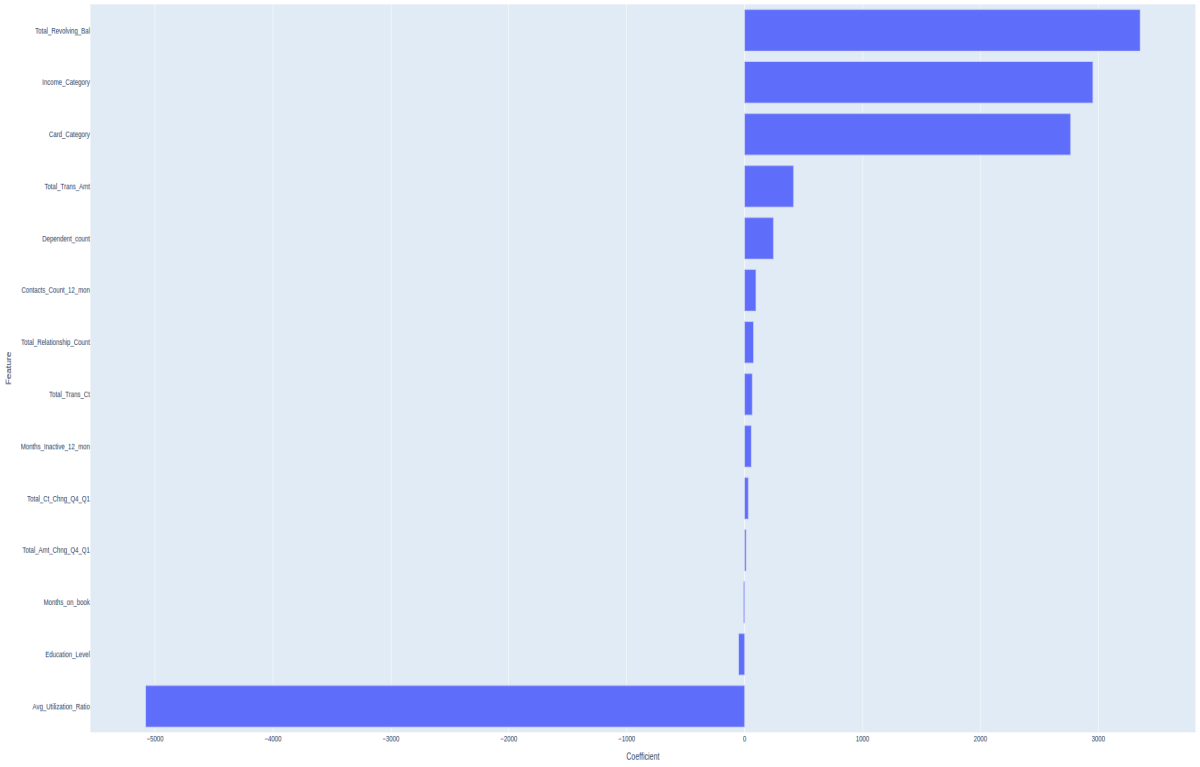
همچنین برای انجام خوشه بندی از دیتافریم با ابعاد اصلی و کاهش بعد شده استفاده کردم. و با معیار silhouette بهترین بعد برای خوشه بندی همان ابعاد اصلی انتخاب شد.

silhouette_score with out pca: 0.11471451594848583

silhouette_score with pca: 0.2625203249147169

حالا تمامی مقادیر خالی پر شده اند.

ابتدا از مدل یک رگرسیون خطی گرفتم سپس با r2 score دقت 0.607482123903178 دریافت شد. سپس Coefficient هر ستون را گرفتم. نمودار آن به صورت زیر است.



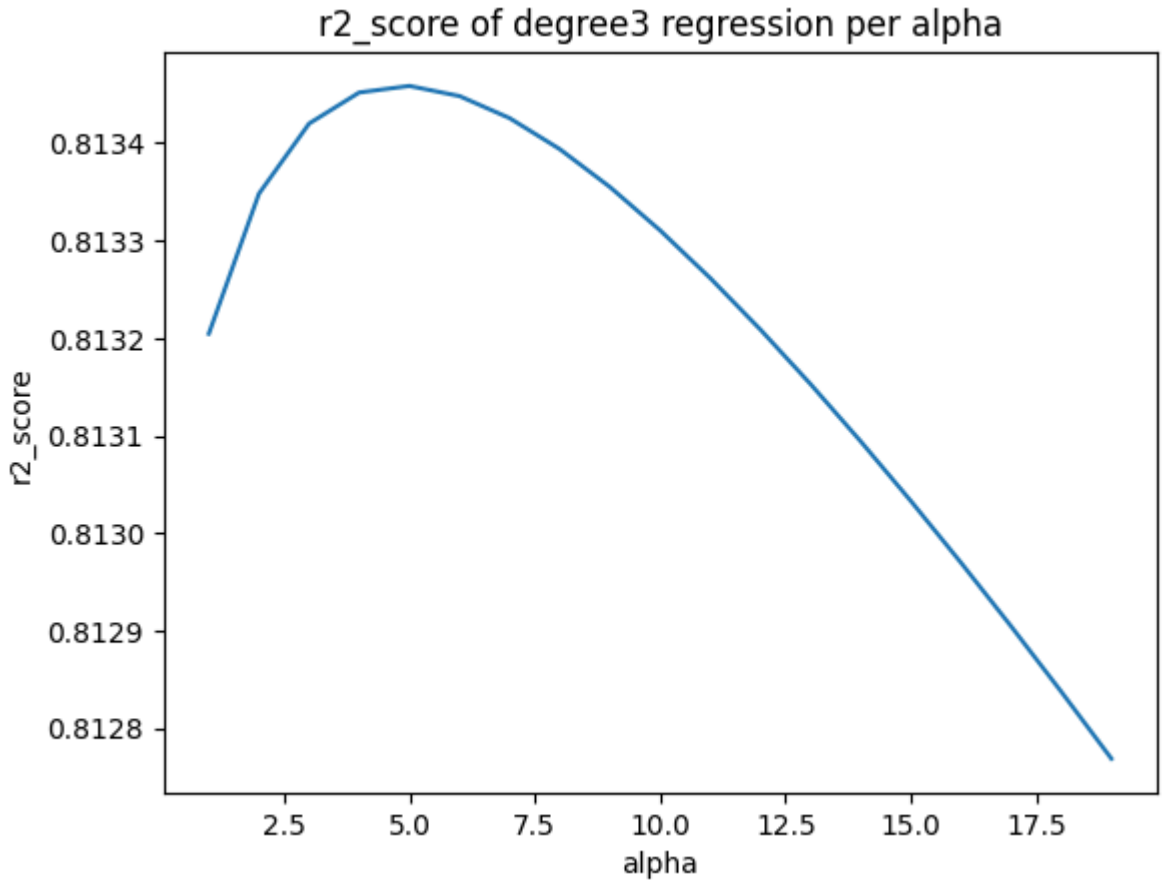
سه ستون اول به ترتیب:

Dependent_count , Education_Level , Income_Category

پس باید مدل را به صورت ridge بگذاریم و یک ضریب alpha به آن اختصاص دهیم. همچنین میتوانیم از رگرسیون چند جمله ای استفاده کنیم. در رگرسیون سه جمله ای بدون استفاده از پارامتر alpha دقت برابر 0.8129019821089438 شد. اعداد زیر مرتبط با r2_score از هر درجه میباشد.

```
r2 score of linear :  0.607482123903178
r2 score of degree 2 :  0.7681213298577707
r2 score of degree 3 :  0.8129019821089438
r2 score of degree 4 :  -4.815402253556546e+17
```

بنابراین از ridge با alphaهای مختلف استفاده میکنیم.



میبینیم بهترین عمکرد در alpha = 5 میباشد سپس مدل overfit میشود.

max score:

degree: 3 alpha: 5

score: 0.8134584105623583

سپس از مدل random forest regressor استفاده کردم.

n_estimators: 39

score: 0.8641181814824872