



## گزارش پروژه مبانی داده کاوی

اعضای گروه:

امیررضا حسینی

پاشا احمدی

بهار ۱۴۰۲

دانشگاه صنعتی اصفهان – دانشکده مهندسی برق و کامپیوتر

۱	معرفی حوزه مسأله .....
۳	فاز آماده سازی داده ها .....
۱۳	فاز تحلیل اکتشافی داده ها .....
۱۷	فاز پیش مدل .....
۱۸	فاز مدل سازی .....
۲۰	نتایج .....
۲۱	جمع بندی .....

## مقدمه (توضیح مسئله)

معرفی حوزه مسأله:



مجموعه دادگان ما حاوی داده‌های مربوط به یک کمپین تبلیغاتی در بانک‌های کشور پرتغال می‌باشد. این اطلاعات شامل 11162 رکورد (11162 مشتری) و 17 ستون است. ستون‌ها یا ویژگی‌های ما عبارتند از سن، شغل، وضعیت تأهل، تحصیلات، قرض الحسنه، وام مسکن، وام شخصی، موجودی، نوع بستر ارتباطی برای تماس، آخرین ماه تماس در سال، آخرین روز هفته تماس، مدت زمان آخرین تماس گرفته شده با مشتری، تعداد تماس‌های انجام شده در طول این کمپین برای هر مشتری، تعداد روز‌های سپری شده از آخرین تماس با مشتری از کمپین قبلی، تعداد تماس‌های انجام شده قبل از این کمپین برای هر مشتری، نتیجه کمپین قبلی و در نهایت ویژگی هدف آیا مشتری سپرده کوتاه مدت باز کرده است؟ یا خیر.

حوزه مسئله شما داده کاوی در بانک جهت افتتاح حساب است. به طور دقیق‌تر، شما قصد دارید با استفاده از داده‌های موجود در مجموعه دادگان خود که شامل اطلاعات مشتریان بانکی می‌باشد، الگوهای رفتاری کاربران را درک کنید و پیش‌بینی کنید که کدام مشتریان ممکن است در آینده سپرده کوتاه مدت باز کنند و کدام مشتریان نمی‌توانند آن را انجام دهند.



هدف ما در این پروژه بررسی رفتار کاربران در کمپین تبلیغاتی به منظور افتتاح حساب است. به عبارتی دیگر ما قصد داریم با استفاده از داده‌های موجود از اطلاعات مشتریان بانکی، الگوهای رفتاری آنها را درک کنیم تا بر اساس آن خدمات خود را بهبود بخشیده و بهترین تجربه را برای مشتریان خود رقم بزنیم. علاوه بر این با این کار می‌توانیم پیشنهادات مناسب‌تری به مشتریان خود ارائه دهیم و به نحو بهتری محصولات خود را تبلیغ کنیم.



# A data-driven approach to predict the success of bank telemarketing

Sérgio Moro<sup>a</sup>  , Paulo Cortez<sup>b</sup>, Paulo Rita<sup>a</sup>

Show more 

+ Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.dss.2014.03.001> 

[Get rights and content](#) 

Corpus ID: 13326443

## Using data mining for bank direct marketing: an application of the CRISP-DM methodology

Sérgio Moro, Raul M. S. Laureano, P. Cortez • Published 1 October 2011 • Business

The increasingly vast number of marketing campaigns over time has reduced its effect on the general public. Furthermore, economical pressures and competition has led marketing managers to invest on directed campaigns with a strict and rigorous selection of contacts. Such direct campaigns can be enhanced through the use of Business Intelligence (BI) and Data Mining (DM) techniques. This paper describes an implementation of a DM project based on the CRISP-DM methodology. Real-world data were... [Expand](#)

## بررسی مختصر ایده‌های تیم:

با بررسی دیتاست، متوجه شدیم، که خوشبختانه داده **missing** نداریم اما در برخی از سطر ها داده های پرت وجود دارد که باید حذف شوند.

در قدم بعدی، متوجه شدیم که برخی از ویژگی‌ها مقادیر بسیار کوچکی دارند و برخی از ویژگی‌ها با اعداد بسیار بزرگ سر و کار دارند. این موضوع می‌تواند در زمان آموزش و استفاده از مدل مشکل‌ساز شود. برای رفع این مشکل، می‌توانیم از روش‌های نرمال‌سازی و استانداردسازی استفاده کنیم.

در دیتاست مورد بررسی، هم ویژگی‌های عددی و هم ویژگی‌های دسته‌ای داشتیم. برای ویژگی‌های دسته‌ای از روش **Label Encoder** به منظور تبدیل دسته‌ها به مقادیر عددی و برای ویژگی‌های عددی از روش‌های سبب‌بندی به منظور گروه‌بندی مقادیر آنها استفاده کرده ایم.

کار دیگری که در این پروژه انجام دادیم بررسی، **correlation** دو به دو هر یک از ویژگی‌ها و استخراج ویژگی‌های جدید از آنهاست که با هم بیشترین **correlation** داشتند و در نهایت با مقایسه **correlation** بین ویژگی‌ها و ویژگی هدف با **correlation** بین ویژگی جدید استخراج شده و متغیر هدف، می‌توانیم تصمیم بگیریم که آیا ویژگی استخراج شده را جایگزین ویژگی‌های تشکیل دهنده آن کنیم یا خیر.

## روش (داده‌ها- نحوه پردازش و انجام کار و...):

### ۱) گزارش تشخیص داده‌های پرت با استفاده از روش‌های **z-score** و **IQR**

در این بخش، به بررسی دو روش برای تشخیص داده‌های پرت می‌پردازیم. روش اول استفاده از تکنیک **IQR** و روش دوم استفاده از **z-score** است.

در روش **IQR**، داده‌ها به چهار چارک تقسیم می‌شوند. سپس با کم کردن انتهای چارک اول از انتهای چارک سوم، حاصل **IQR** محاسبه می‌شود. سپس داده‌هایی که از  $1.5 \times IQR$  برابر کمتر و از  $1.5 \times 3Q$  بیشتر است به عنوان داده‌های پرت شناسایی میشوند.

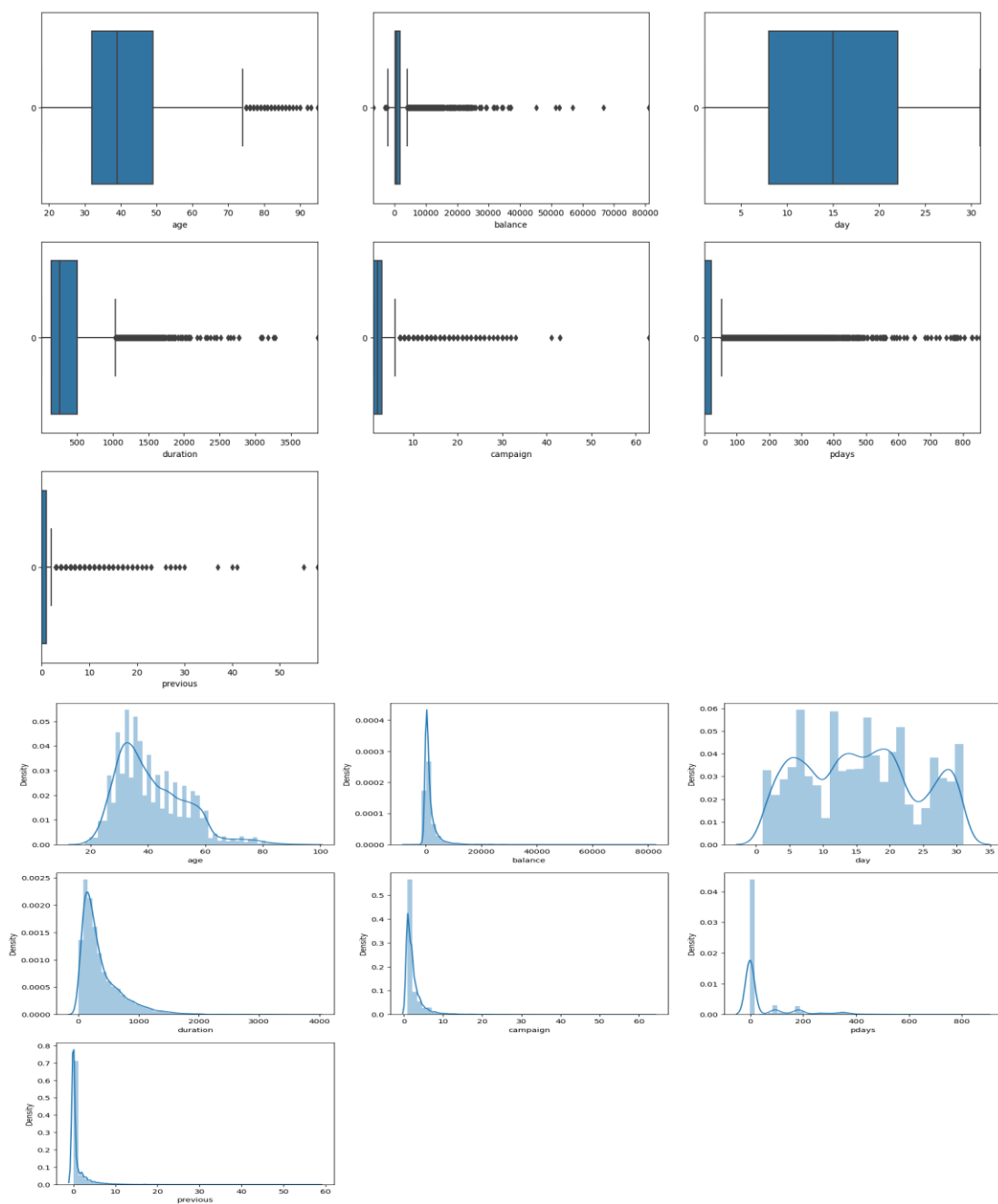
در روش **z-score**، ابتدا برای هر ویژگی از داده‌ها، امتیاز **z-score** محاسبه می‌شود. سپس با بررسی قدر مطلق این امتیاز برای هر داده در مقایسه با عدد ثابت ۳، داده‌های پرت شناسایی می‌شوند.

برای بررسی تأثیر این دو روش در پاکسازی داده‌های پرت، از نمودار **distplot** و **boxplot** در قبل و بعد پاکسازی داده‌ها استفاده می‌شود.

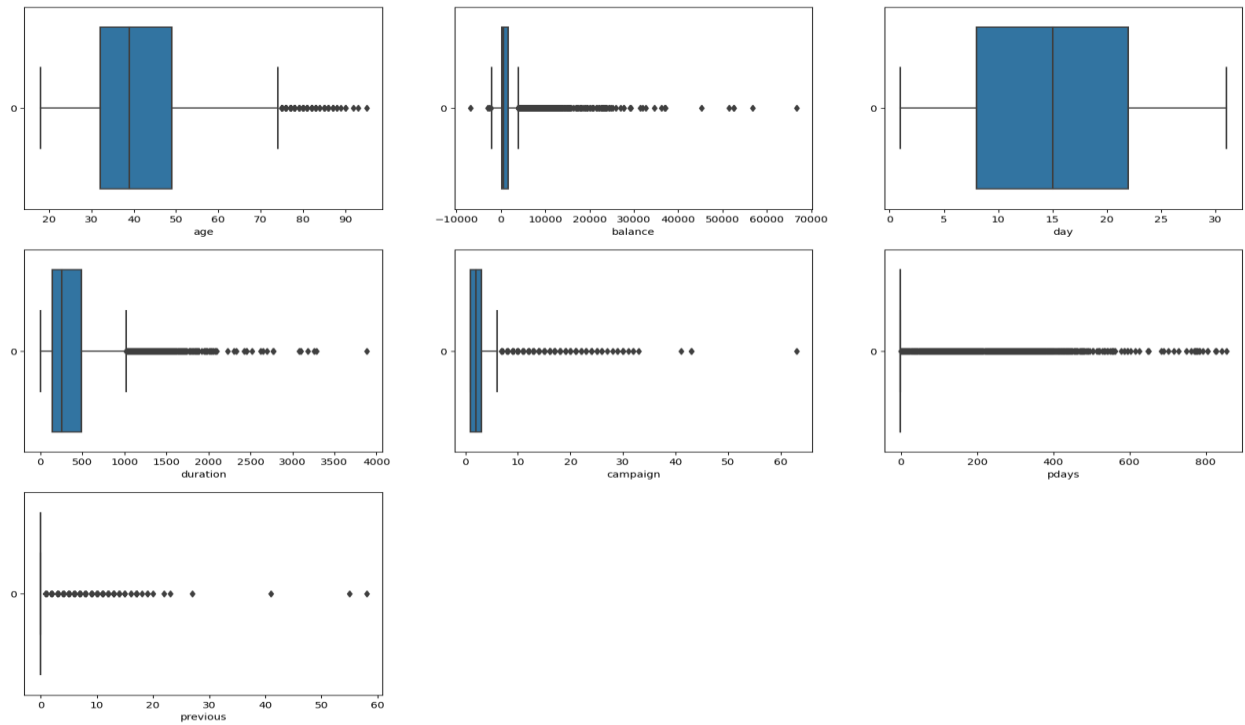
این روش‌ها باعث بهبود کیفیت داده‌ها شده و تأثیر مثبتی بر روی دقت و صحت نتایج تحلیل داده‌ها دارند.

نتایج به دست آمده نشان می‌دهد که روش **z-score** در مقایسه با روش **IQR**، روش مناسب‌تری برای تشخیص داده‌های پرت خواهد بود بنابراین ما در ادامه کار، از داده‌های بدست آمده از این روش استفاده خواهیم کرد.

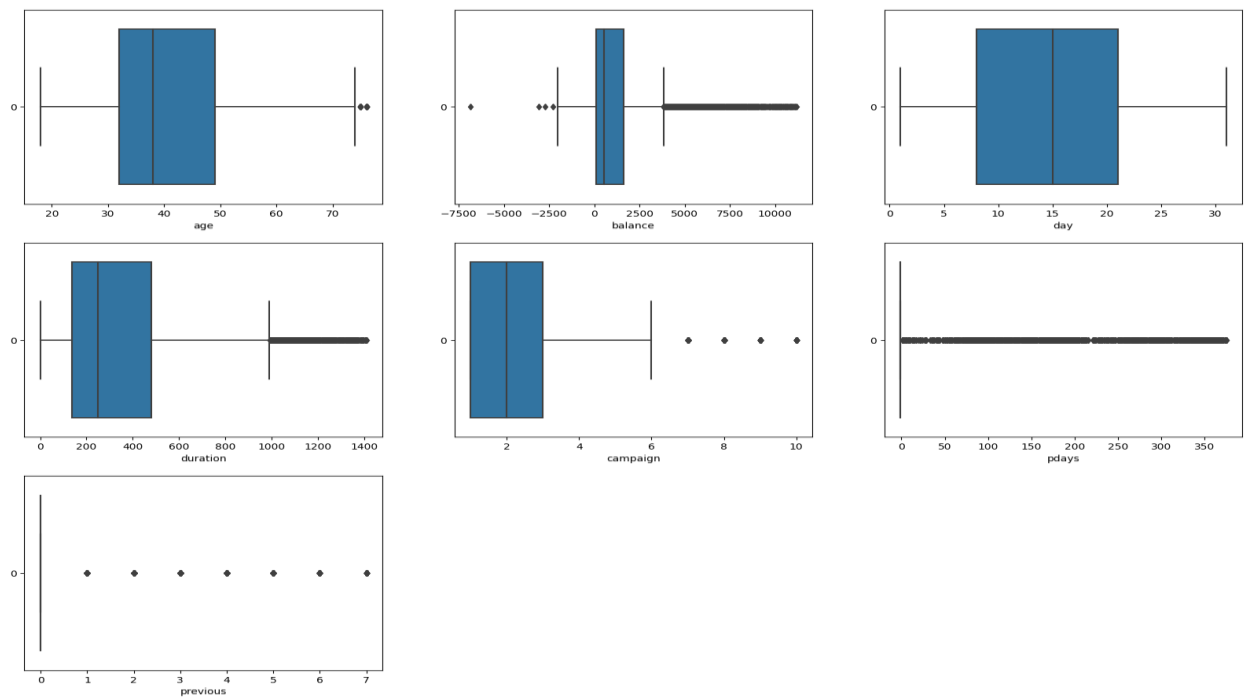
به عنوان نمونه این روش‌ها را برای چند ویژگی بررسی خواهیم کرد. توزیع اولیه داده‌ها عبارت است از:



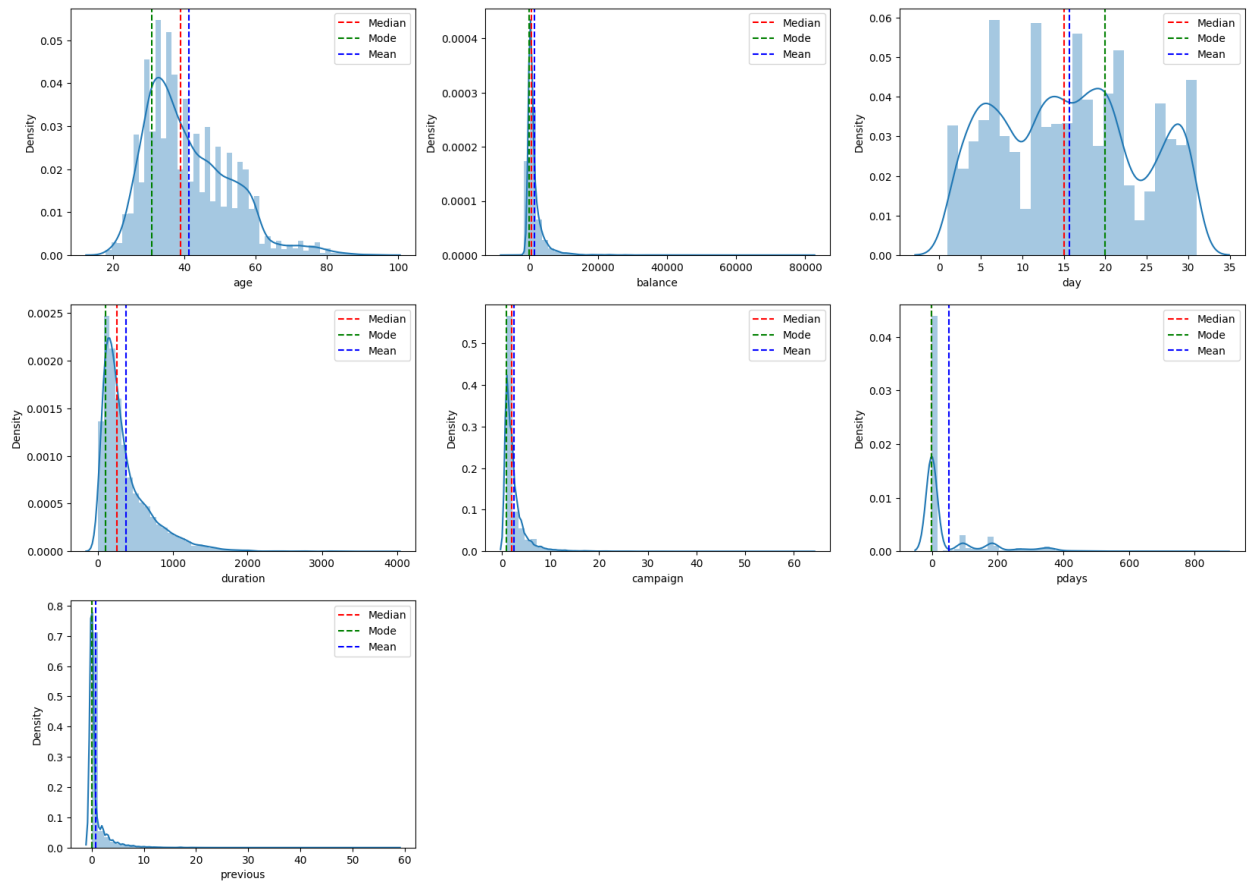
حذف داده‌های پرت به روش IQR:



حذف داده‌های پرت به روش Z-Score:



همانطور که مشاهده میشود، Z-Score توانایی بهتری برای حذف داده‌های پرت دارد و در نهایت نمودار توزیع پراکندگی داده‌ها برای ویژگی‌های عددی به صورت زیر میباشد:

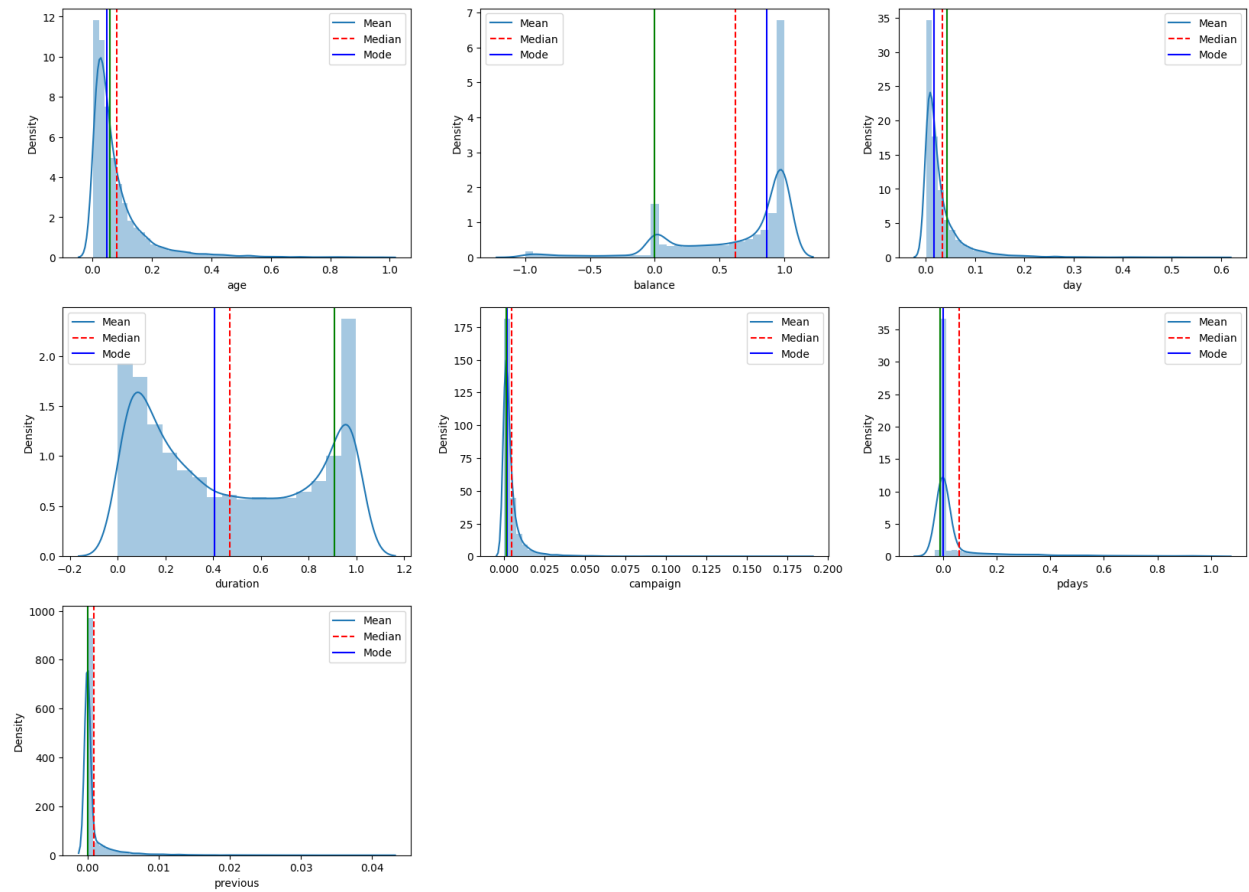


### تبدیل و استاندارد سازی داده ها:

با توجه به اینکه تبدیل و استاندارد سازی داده ها، یک مرحله مهم در پیش پردازش داده ها در پروژه های داده کاوی محسوب می شود استفاده از روش های تبدیل و استاندارد سازی داده ها می تواند به بهبود کیفیت داده ها و بهبود عملکرد الگوریتم های داده کاوی کمک کند.

برای این دیتاست، ما برای استاندارد سازی داده ها از تابع `normalize` از پکیج `preprocessing` در کتابخانه `Sckit-Learn` استفاده کرده ایم. این تابع، برای استاندارد سازی داده های عددی به کار می رود. به صورت پیش فرض، این تابع داده های عددی را به یک دامنه `[0,1]` تبدیل می کند. بدین صورت که مقدار هر ویژگی را بر تعداد ویژگی ها تقسیم می کند. این کار، به کاهش اثرات تغییرات مقیاس در ویژگی ها کمک می کند و باعث می شود همه ویژگی ها در یک مقیاس یکسان باشند.

در تصویر زیر می توانید نمودار `distplot` مربوط به داده های عددی برای نمایش توزیع داده ها بعد از استاندارد سازی مشاهده کنید:

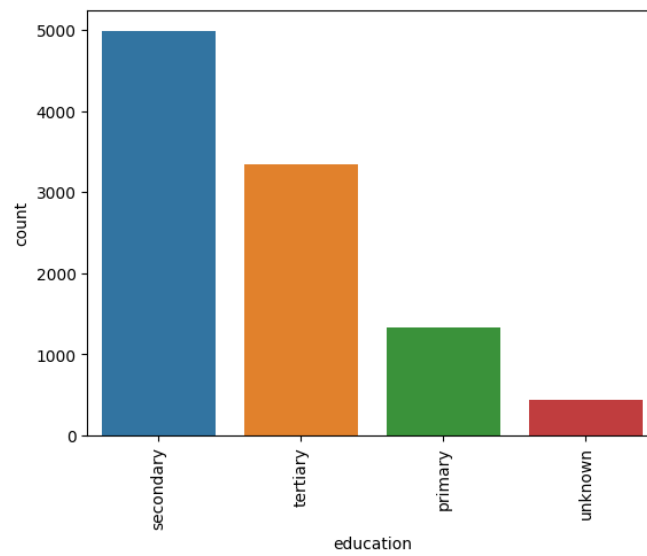
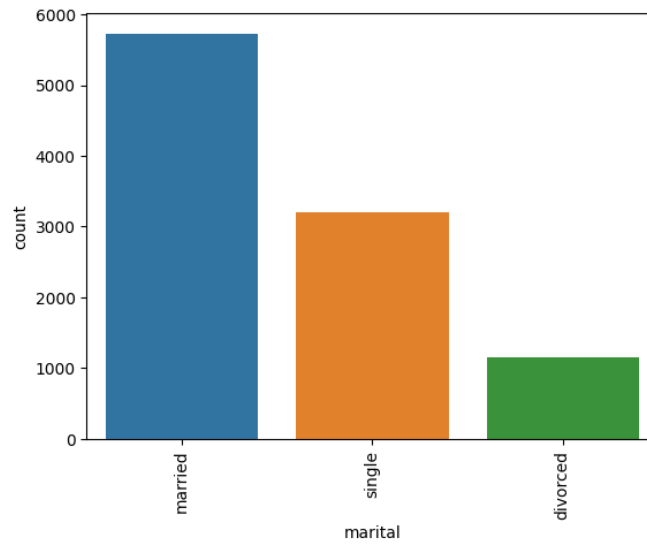
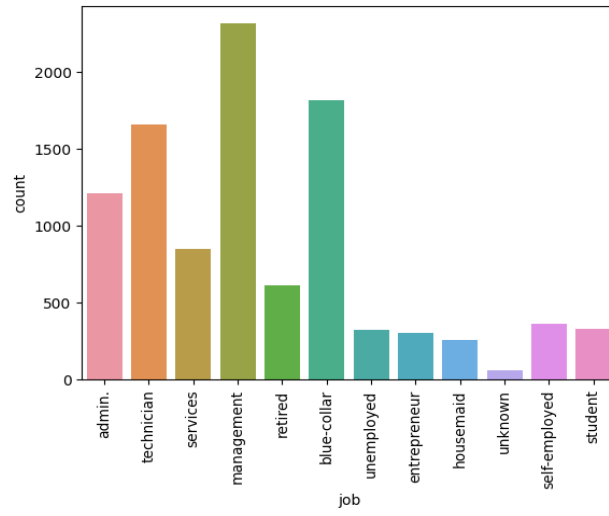


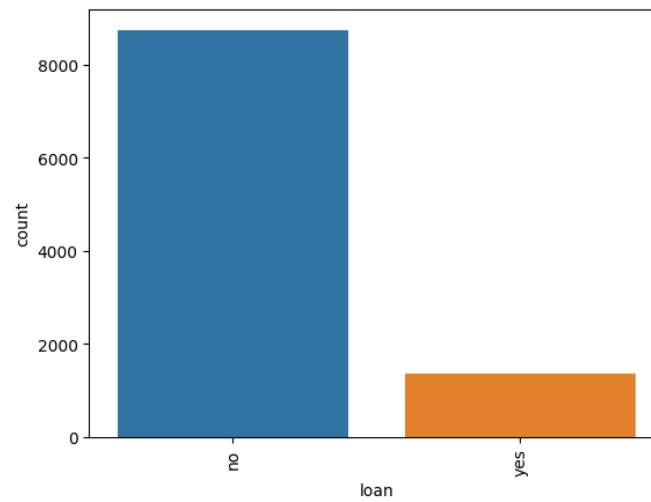
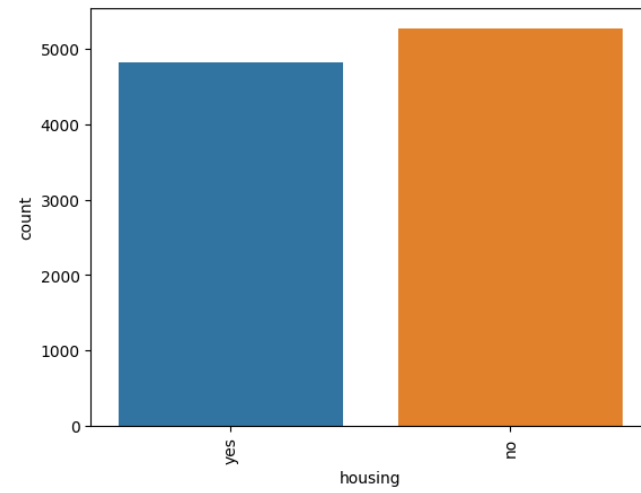
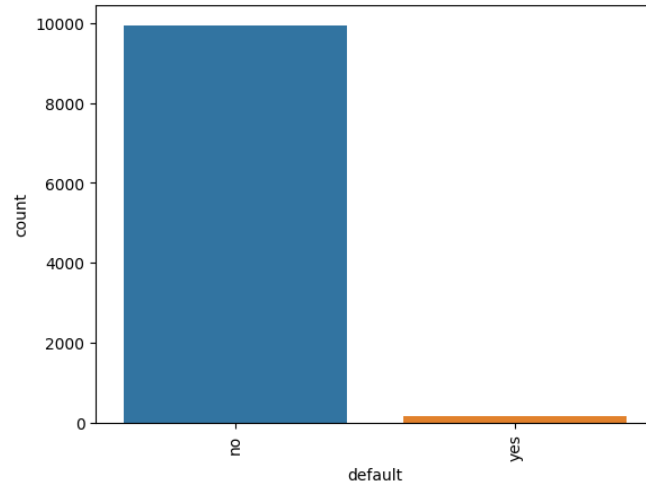
### دسته بندی مجدد (Reclassify) متغیر های دسته ای:

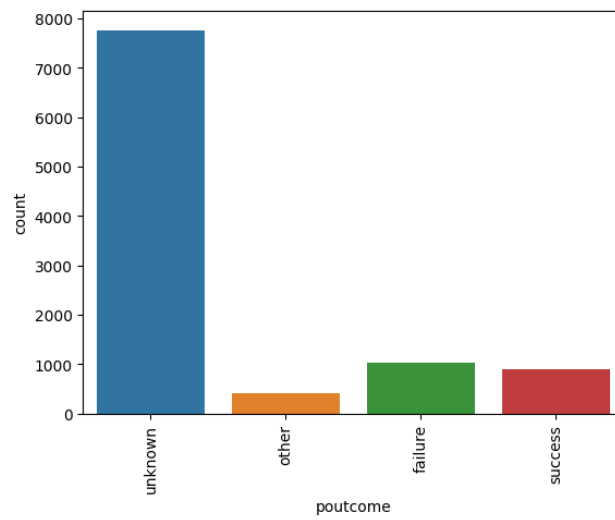
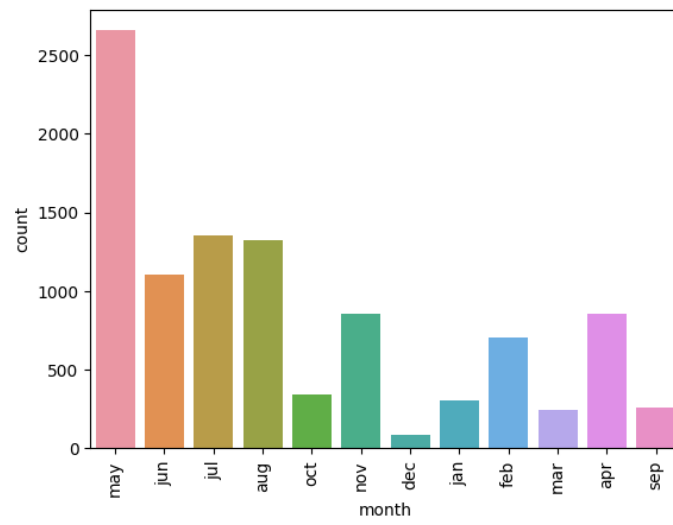
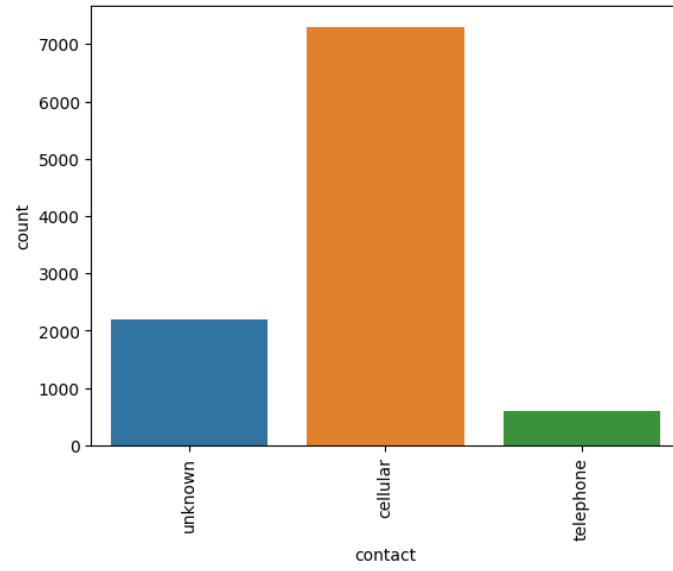
در فرآیند دسته بندی مجدد، ویژگی های دسته ای که در ابتدا با توجه به اطلاعات موجود تعریف شده اند، ممکن است نیاز به بازنگری و تغییر کلاس داشته باشند. درواقع گاهی با تحلیل دقیق تر داده ها، ممکن است متوجه شویم که این دسته بندی مناسب نیست و نیاز به تغییر دارد. در دسته بندی مجدد، معمولا، هدف تعیین دسته های مناسب برای ویژگی های دسته ای است تا در آنالیز داده ها و دسته بندی بهتر اطلاعات به ما کمک کند.

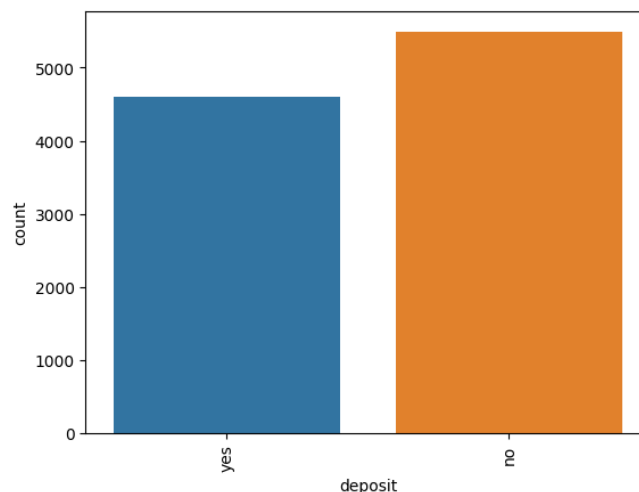
نمایش متغیر های دسته ای به صورت تعداد در هر دسته با استفاده از تابع `cat_summary`:



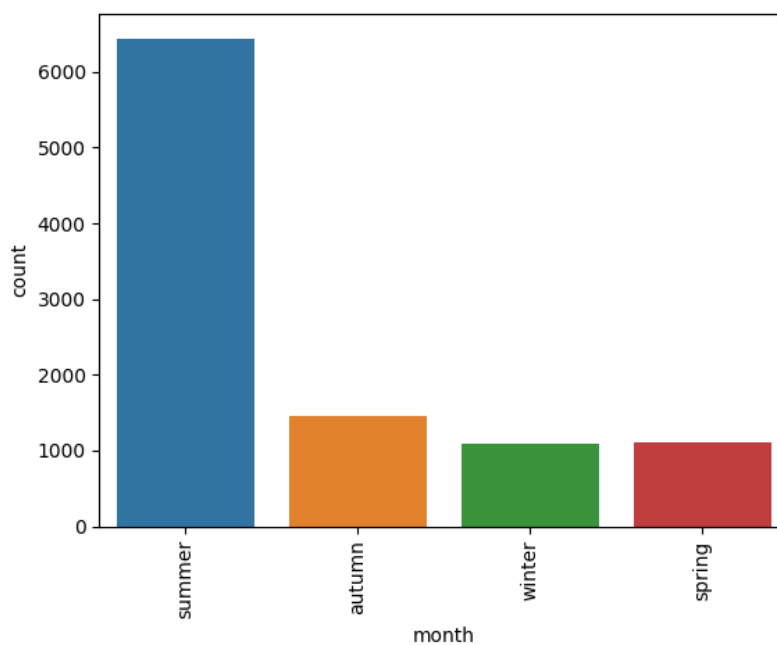








در این دیتاست با توجه به اهمیت هر یک از دسته های ویژگی های دسته ای، اقدامی برای تغییر دسته بندی انجام ندادیم اما به عنوان یک نمونه دسته های مختلف مربوط به ویژگی **month** را بر اساس فصل، دسته بندی مجدد (**reclassify**) کرده ایم:

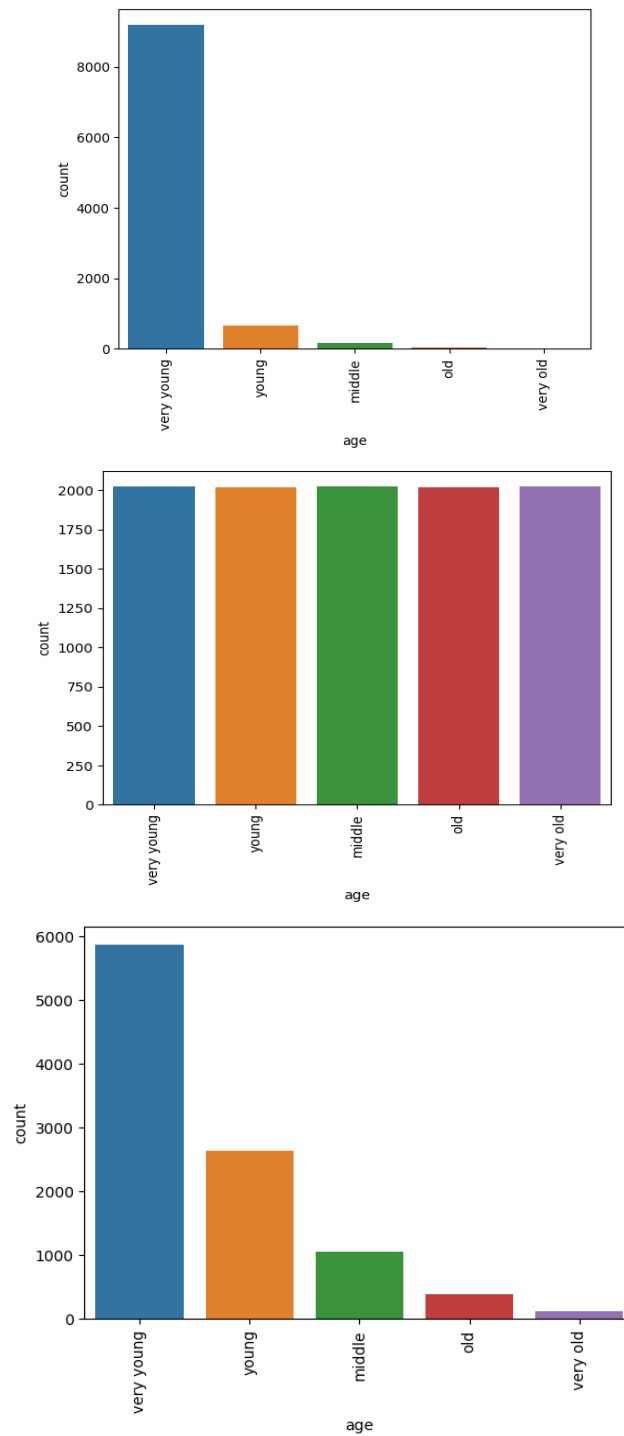


### سبد بندی متغیر های عددی:

سبندبندی یا **binning** یکی از روش های پرکاربرد در داده کاوی است که به منظور گروه بندی مقادیر یک ویژگی یا متغیر استفاده می شود. این روش به منظور کاهش تعداد مقادیر یک ویژگی و کاهش اثر نویزهای کوچک در داده ها استفاده می شود. استفاده از روش سبندبندی می تواند به عنوان یک قدم اولیه در پیش پردازش داده ها مفید باشد، به طوری که برای مقادیر پراکنده یک ویژگی، می توان مقادیر را در بازه هایی تقسیم کرد تا اطلاعاتی درباره یک متغیر را به شکل خلاصه تری ارائه داد.

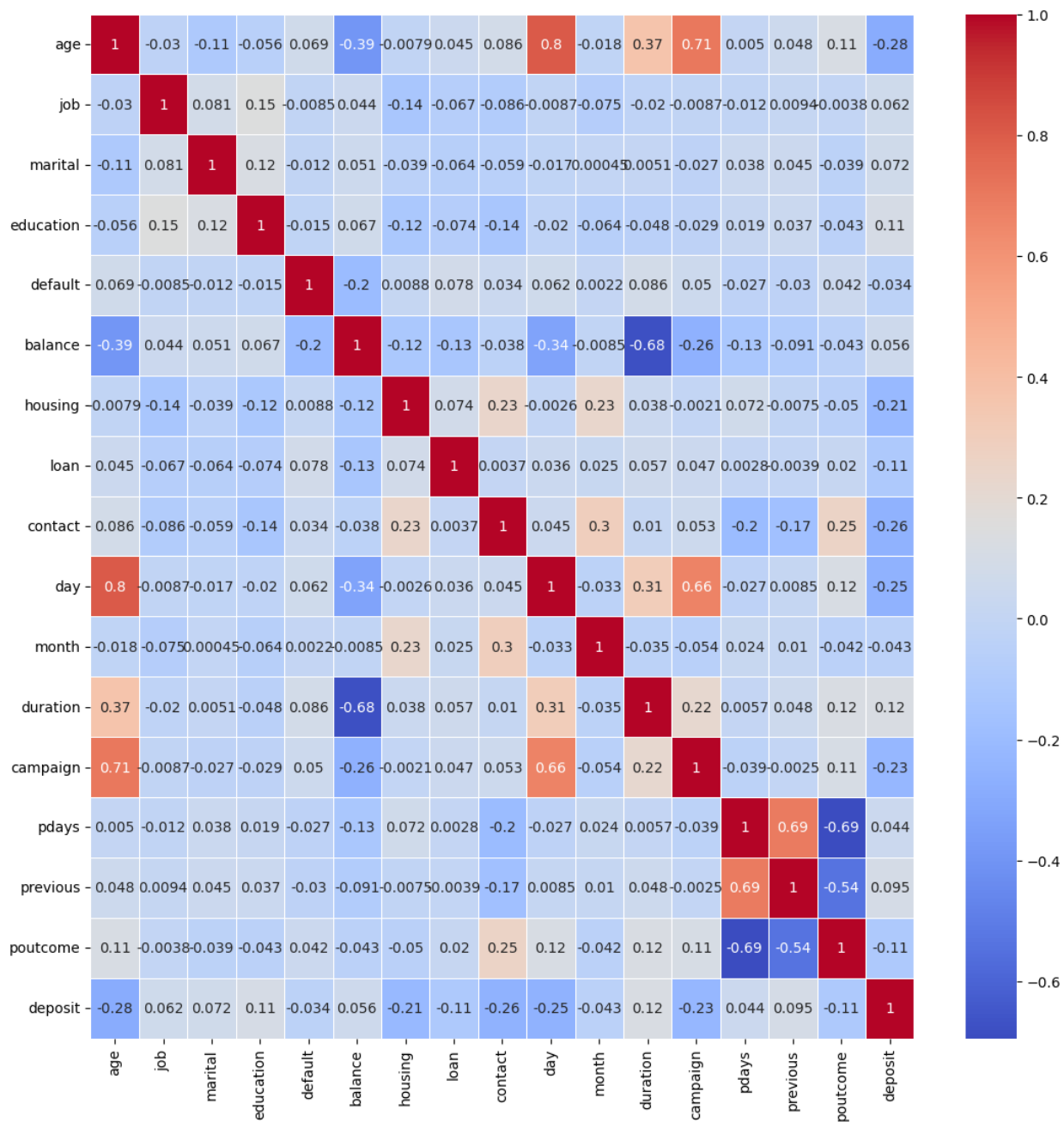
بنابراین، روش سبندبندی به منظور خلاصه‌تر کردن داده‌ها، افزایش سرعت عملیات روی داده‌ها، افزایش دقت مدل‌سازی و از بین بردن نویز در داده‌ها استفاده می‌شود.

در تصاویر زیر می‌توانید سبندبندی ویژگی Age را به سه روش: cut سپسس qcut و درنهایت natural break به ترتیب مشاهده کنید:

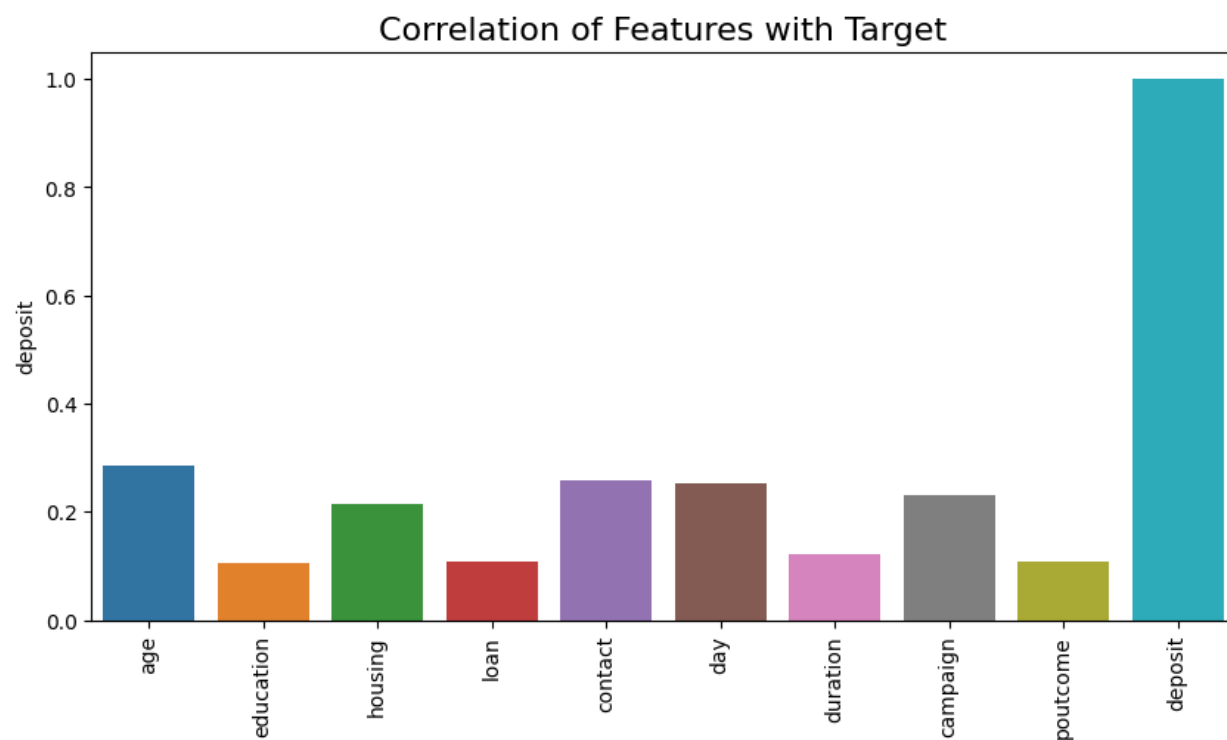


فاز تحلیل اکتشافی داده ها (EDA) :

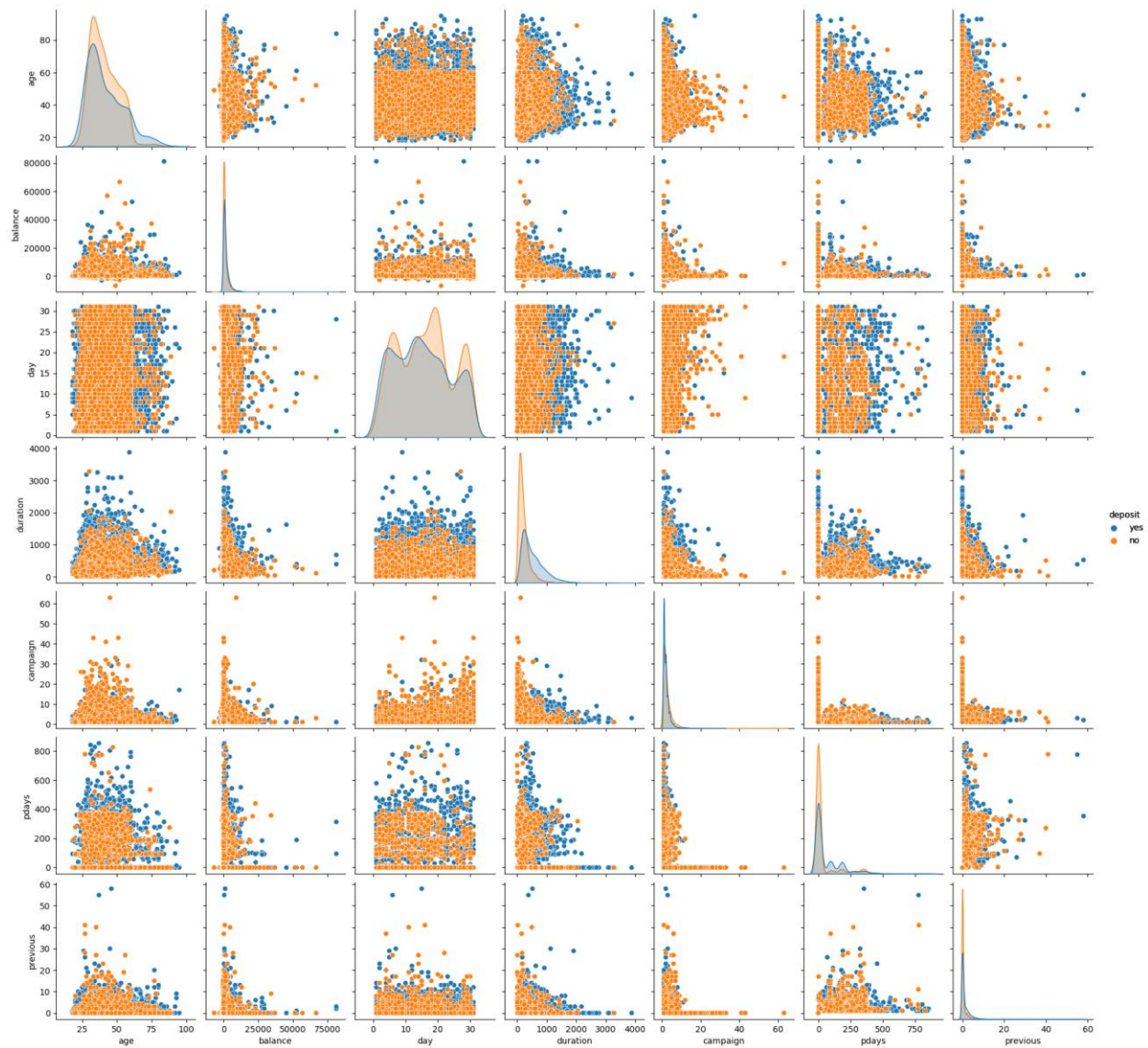
در تصویر زیر میتوانید نمودار heatmap مربوط همبستگی (Correlation) دو به دو هر یک از ویژگی را مشاهده کنید:



همچنین در تصویر زیر میتوانید میزان همبستگی هر یک از ویژگی ها با ویژگی هدف که بیشتر از 0.1 است را مشاهده کنید:

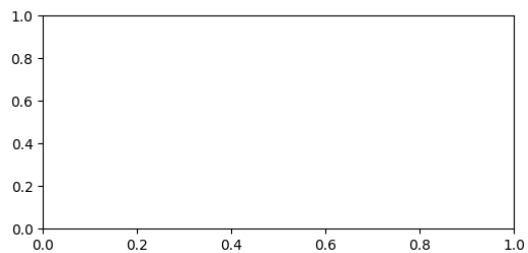
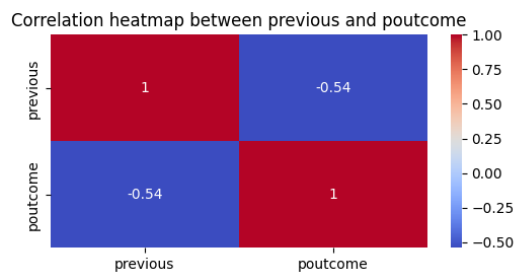
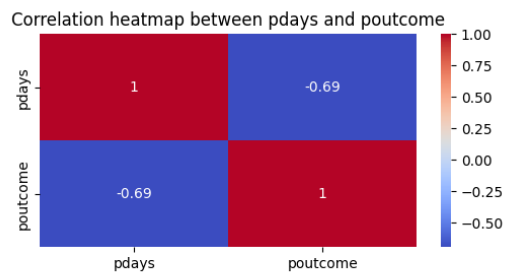
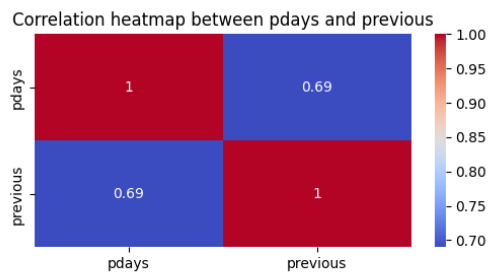
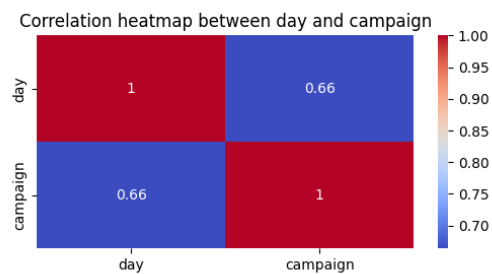
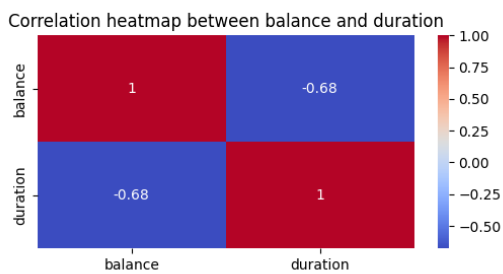
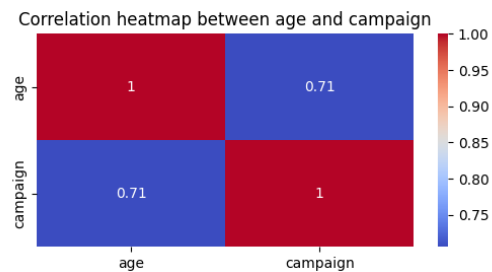
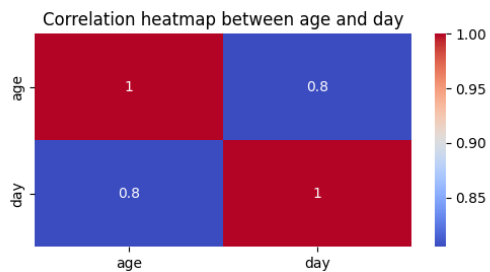


همچنین در تصویر زیر میتوانید با استفاده از متد **pairplot**، روابط چند متغیره بین متغیر ها را مشاهده کنید:

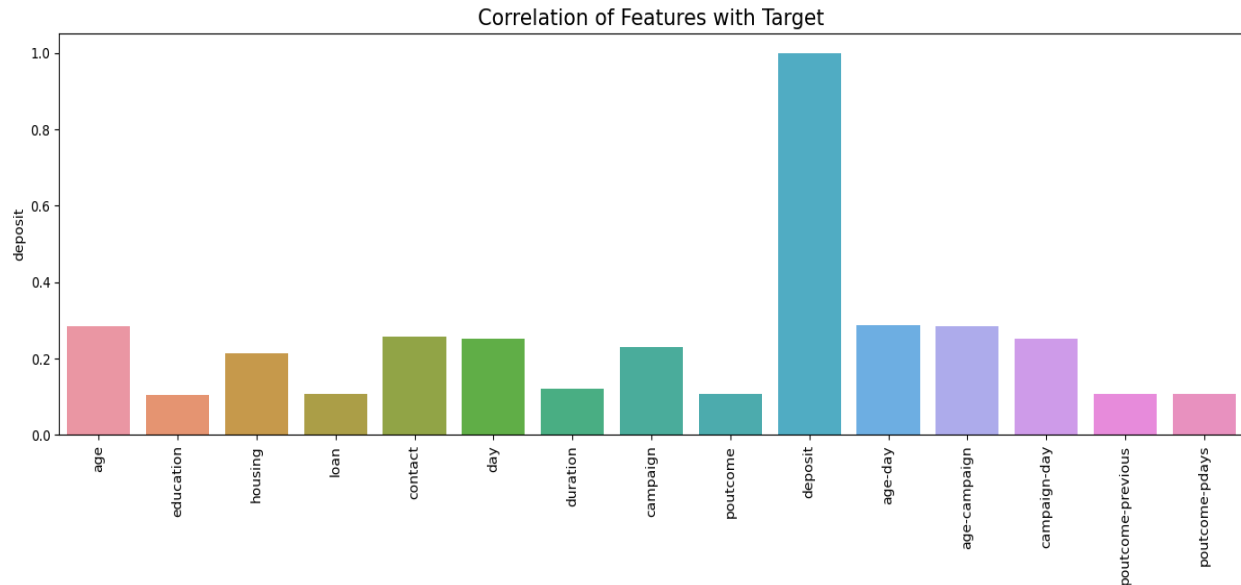


در مرحله بعد برای استخراج ویژگی های جدید بر اساس ترکیب متغیر های موجود لازم است میزان **correlation** بین هر دو ویژگی عددی را دو به دو به صورت جداگانه بررسی کنیم. در تصویر زیر میتوانید نمودار **heatmap** مربوط به **correlation** دو به دو ویژگی های عددی که همبستگی بیشتر از 0.4 داشته اند را مشاهده کنید:





با انجام این کار احتمال می‌دهیم که از بین ویژگی‌های بالا که دارای همبستگی بیشتر از 0.4 باهم دارند بتوانیم ویژگی‌های جدیدی استخراج کنیم، به همین منظور لازم است میزان **correlation** بین ویژگی‌های جدید استخراج شده با ویژگی هدف را بررسی کنیم و در نهایت آن ویژگی‌های استخراج شده از که دارای همبستگی بیشتری با ویژگی هدف در مقایسه با ویژگی‌های تشکیل دهنده آن و متغیر هدف دارند را جایگزین کنیم:



با توجه به تصویر بالا میتوان مشاهده کرد با استخراج برخی ویژگی ها مانند **age-day** میزان **correlation** آن با ویژگی هدف یعنی **deposit** در مقایسه با **correlation** بین هر یک از ویژگی های **age** و **day** نسبت به ویژگی هدف بیشتر خواهد بود.

## آزمایشات

فاز پیش مدل:

انجام **corss validation**:

برای شروع مدل سازی، مرحله اول، تقسیم دیتافریم به دو بخش است: بخش اول که شامل تمامی ویژگی ها به جز ویژگی هدف است و بخش دوم که شامل ویژگی هدف می باشد.

برای انجام **cross validation** در این پروژه، از دو روش استفاده شده است. در روش اول، داده ها به نسبت 70 به 30 به صورت تصادفی به دو دسته **train** و **test** تقسیم شده و مدل روی داده های **train** آموزش داده می شود. سپس عملیات ارزیابی و تست مدل روی داده های **test** صورت می گیرد. در روش دوم، از روش **fold cross-validation-20** استفاده شده است که در آن داده ها به 20 بخش تقسیم شده و هر بار یک بخش به عنوان داده ی آزمایشی (**test**) در نظر گرفته می شود و داده های باقی مانده به عنوان داده ی آموزشی (**train**) استفاده می شوند. سپس مدل روی داده های **train** آموزش داده شده و روی داده های **test** ارزیابی می شود. این عملیات به صورت 20 بار تکرار و برای هر بخش به عنوان داده ی آزمایشی، نتیجه ی میانگین به دست می آید.

استانداردسازی داده ها:

در صورتی که رنج عددی فیلهای مختلف در مجموعه داده با یکدیگر تفاوت دارد، ابتدا از روش **StandardScaler** برای نرمال سازی متغیرهای **X\_train** و **X\_test** استفاده می شود.

بالانس داده‌ها:

پس از بررسی ویژگی هدف، متوجه می‌شویم که تعداد ۴۶۰۵ نمونه از کلاس **yes** و ۵۴۹۳ نمونه از کلاس **no** در مجموعه داده وجود دارد. این نشان می‌دهد که مجموعه داده بالانس نیست و به بالانس کردن نیاز دارد. برای این منظور، از روش **over sampling** و به خصوص روش **SMOTE** استفاده می‌شود. پس از اعمال این روش، تعداد نمونه‌های هر دو کلاس به ۴۳۸۴ نمونه تعدیل شده و بالانس مجموعه داده حفظ می‌شود.

تعیین **baseline**:

در این بخش، با استفاده از الگوریتم **Dummy Classifier** یک مدل **baseline** ساخته شده است. در این مدل، استراتژی انتخاب کلاس به صورت یکنواخت (**uniform**) تعیین شده است. دقت مدل **baseline** در داده‌های تست برابر با ۵۰٪ و با استفاده از روش **fold cross-validation-20** دقت حدود ۷۶٪ به دست آمده است.

فاز مدل سازی:

در این قسمت الگوریتم‌های زیر برای مدل سازی استفاده شده‌اند. هریک از این الگوریتم‌ها با استفاده از پارامترهای پیش فرض بر روی داده‌ی تست آموزش داده شده‌اند:

- Random Forest
- SVM
- Stochastic Gradient Descent
- XGBoost
- Light Gradient Boosting Machine
- Neural Network
- Naive Bayes

اطمینان عملکرد نسبت به **baseline**:

تمامی موارد ذکر شده فوق، عملکرد بهتری نسبت به مدل **Dummy Classifier** از خود در مواجهه با داده‌های تست نشان می‌دهند. به عبارت دقیق‌تر خواهیم داشت:

• Random Forest:

دقت به دست آمده برای این مدل روی داده‌ی تست برابر با ۸۳٪ و میانگین دقت به دست آمده با روش **20fold cross-validation** برابر با ۸۴٪ است که در هر دو حالت بهتر از **baseline** است.

• SVM:

دقت به دست آمده برای این مدل روی داده‌ی تست برابر با ۷۹٪ و میانگین دقت به دست آمده با روش **20fold cross-validation** برابر با ۸۰٪ است که در هر دو حالت بهتر از **baseline** است.

- **Stochastic Gradient Descent:**

دقت به دست آمده برای این مدل روی داده ی تست برابر با 76٪ و میانگین دقت به دست آمده با روش 20fold cross-validation برابر با 76٪ است که در هر دو حالت بهتر از baseline است.

تنظیم بهینه ی مدل و hyper-parameterها:

در این قسمت برای هر یک از مدل های قسمت قبل hyper-parameterها به کمک GridSearchCV تنظیم میشوند. سپس مدل ها مجددا با استفاده از پارامترهای به دست آمده آموزش داده میشوند و دقت محاسبه میشود

- **Random Forest:**

دقت به دست آمده برای این مدل روی داده ی تست برابر با 81٪ و بهترین معیارهای پیدا شده عبارت است از:

```
{ 'criterion': 'gini',  
  'max_depth': 8,  
  'max_features': 'auto',  
  'n_estimators': 100 }
```

- **SVM:**

دقت به دست آمده برای این مدل روی داده ی تست برابر با 81٪ میباشد.

- **Stochastic Gradient Descent:**

دقت به دست آمده برای این مدل روی داده ی تست برابر با 78٪ و بهترین معیارهای پیدا شده عبارت است از:

```
{ 'alpha': 0.001, 'learning_rate': 'optimal', 'loss': 'hinge',  
  'penalty': 'l1' }
```

استفاده از حالت های پیشرفته:

- **XGBoost:**

دقت به دست آمده برای این مدل روی داده ی تست برابر با 83٪ میباشد که بهتر از baseline است.

- **Light Gradient Boosting Machine:**

دقت به دست آمده برای این مدل روی داده ی تست برابر با 84٪ میباشد که بهتر از baseline است.

- **Neural Network:**

دقت به دست آمده برای این مدل روی داده ی تست برابر با 81٪ میباشد که بهتر از baseline است.

بهترین پارامترهای یافت شده توسط جست و جوی حریصانه برای این مدل عبارت است از:

```
{ 'activation': 'relu',  
  'alpha': 0.05,  
  'hidden_layer_sizes': (100, ),
```

```
'learning_rate': 'constant',  
'solver': 'adam'}
```

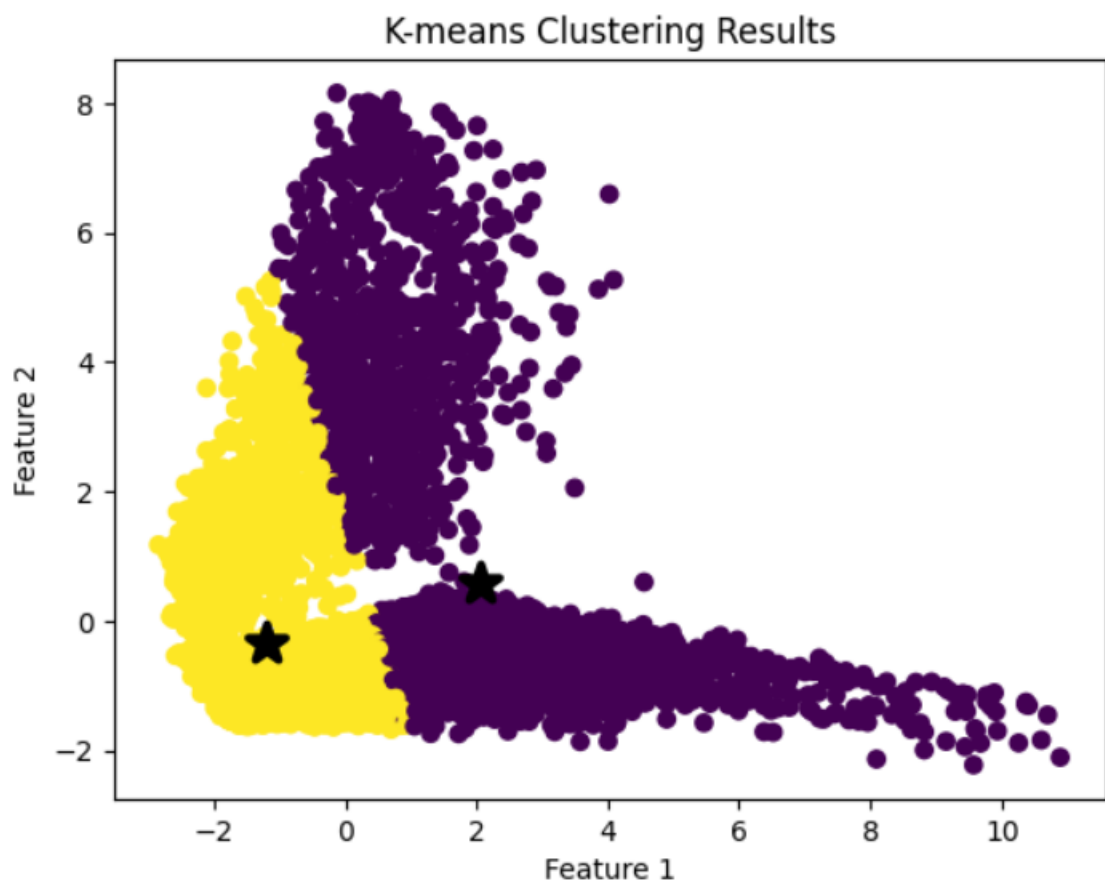
#### • Naive Bayes:

دقت به دست آمده برای این مدل روی داده ی تست برابر با 64٪ میباشد که بهتر از baseline است.

#### نتایج

دسته بندی مشتریان:

با استفاده از کاهش ویژگی ها و استفاده از آن در الگوریتم خوشه بندی، میتوانیم مشتریان را به دو دسته تقسیم کنیم. این کار سبب میشود تا ویژگی های نزدیک بهم مشتریان مشخص شود و با استفاده از آن بتوانیم برنامه های تبلیغاتی متفاوتی برای هر گروه ایجاد نماییم. در اینجا به عنوان مثال اگر تعداد دسته ها را ۲ فرض کنیم، مشتریان به شکل زیر دسته بندی میشوند.



میتوان با ترکیب این اطلاعات با اطلاعات قبلی بدست آمده، plan های نوین تبلیغاتی ایجاد نمود که در بالا بردن سوددهی بانک به موسسه مالی کمک کند.

## جمع بندی و نتیجه گیری

### پیشنهادهای بیزنسی:

برای بهبود فرآیند مارکتینگ در بانکداری، باتوجه به مواردی که از داده‌ها استخراج کردیم، میتوان موارد زیر را به موسسه بانکی پیشنهاد کرد.

تماس در فصول مناسب: ما مشاهده کردیم که ماه با بیشترین فعالیت بازاریابی، ماه می بود. با این حال، در این ماه مشتریان پتانسیل به پیشنهادات سپرده‌های تضمینی رد می‌کردند. برای کمپین بازاریابی بعدی، بهتر است بانک در ماه‌های مارس، سپتامبر، اکتبر و دسامبر تمرکز کند. (دسامبر باید مورد بررسی قرار گیرد زیرا کمترین فعالیت بازاریابی در آن ماه بود، ممکن است دلیلی وجود داشته باشد که دسامبر کمترین فعالیت را داشته است).

تعداد تماس: باید سیاستی اجرا شود که بیان کند حداکثر ۳ تماس برای هر مشتری اعمال شود تا برای جذب مشتریان دارای پتانسیل، زمان و تلاش صرفه‌جویی شود. باید به یاد داشته باشیم که هر چه بیشتر به یک مشتری دارای پتانسیل تماس بگیریم، احتمال اینکه او موافقت کند تا سپرده تضمینی باز کند، کمتر می‌شود.

ایجاد پرسشنامه: همانطور که مشاهده کردیم، یکی از ویژگی‌هایی که ارتباط مستقیم با متغیر هدف دارد، ویژگی **duration** می‌باشد. لذا یکی از راه‌کارهایی که میتوان با آن طول مکالمه را افزایش داد، ایجاد یک پرسشنامه جذاب است.

بر این باوریم که با انجام کارهایی از این قبیل، در آینده بهبود قابل توجه‌ای در تمایل مشتریان به باز کردن سپرده در بانک پس از هربار تبلیغات را مشاهده کنیم.

### تحلیل نقاط قوت و ضعف کار انجام شده و پیشنهاد بهبود آینده:

در این پروژه سعی بر آن بود که با پیش پردازش مناسب و تحلیل درست داده‌ها بهترین نتیجه حاصل شود.

یکی از نقاط ضعف کار تسلط کم ما بر روی مفهوم ویژگی‌ها در دنیای واقعی بود. با درک بهتر از مفهوم

ویژگی‌ها و آشنایی بهتر با شیوه‌های مفید تبلیغات در بانکداری نوین احتمال رسیدن به نتایج بهتر و قابل اعتماد تر وجود دارد.