

پروژه مبانی داده کاوی

بررسی رفتار کاربران دریک بانک جهت افتتاح حساب امیررضا حسینی - پاشا احمدی بهار ۱۴۰۲ دانشگاه صنعتی اصفهان دانشکده برق و کامپیوتر

فهرستمطالب

۱ - مقدمه

- مشكل چيست؟
- سوال داده کاوی که برای حل مشکل مطرح شده است چیست؟
 - داده های جمع آوری شده
 - نحوه ارزیابی

۳- نتایج

- نتیجه اجرای روش نهایی و لیست حقایق کشف شده در این مرحله
 - لیست نتایج نهایی

۲- آزمایشها

- اکتشافات دادهای و لیست حقایق کشف شده در این مرحله
- اکتشافات مدل سازی و لیست حقایق کشف شده در این مرحله

۲- پیشنهادات

• پیشنهادات برای نو آوری و بهبود



۵- مراجع

• ليست مراجع

مقدمه

- كمپينهاى تبليغاتى ساليانه
- سپردههای کوتاه مدت و بلند مدت
 - مشتری خوب؟ مشتری بد؟
 - جذب مشتری و سوددهی
 - پیشنهادات متنوع



بررسی رفتار کاربران دریک بانک جهت افتتاح حساب

- معرفی حوزه مسئله
- سوال داده کاوی برای حل مشکل

بیان مختصر ایده تیم

- حذف مقادیر Missing و شناسایی داده های پرت و حذف آنها
 - نرمال سازی و استاندارد سازی ویژگی ها
 - سبد بندی و reclassify
- در نظر گرفتن ویژگی های با بیشترین همبستگی با ویژگی هدف برای تحلیل روابط
 - استخراج ویژگی های جدید با همبستگی بالا
- حذف ویژگی هایی با همبستگی کم با متغیر هدف و جایگزینی آن با ویژگی های جدید استخراج شده



ماركتينگ موسسه مالى پرتقال

دادههای مربوط به مشتریان	سن	شغل	وضعيت تاهل	موجودى		
دادههای تماس کمپین قبلی	تعداد تماس	روز هفته	مدت زمان تماس	ماه		
ویژگی هدف	آیا مشتری حساب کوتاه مدت باز کرده یا نه؟					

مشخصات دیتاست و ویژگی های آن

دادههای مربوط به مشتریان شامل:

- سن: (عددی)
- شغل: نوع شغل (دستهای: 'ادمین', 'کارگر', 'کارآفرین', 'خانهدار', 'مدیر', ' بازنشسته', 'خود اشتغال', 'خدمات', 'دانشجو', 'تکنسین', 'بیکار', 'نامشخص')
 - وضعیت تأهل: وضعیت تأهل (دستهای: 'طلاق گرفته'، 'متأهل'، 'مجرد'، 'نامشخص')،
 - تحصیلات: (دستهای: سیکل، دیپلم، عالی، و نامشخص)
 - قرضالحسنه: أيا شخص حساب قرضالحسنه دارد؟ (دستهای: 'خير'، 'بله'، 'نامشخص')
 - وام مسكن: أيا شخص وام مسكن دارد؟ (دستهاى: 'خير'، 'بله'، 'نامشخص')
 - وام شخصی: آیا شخص وام شخصی دارد؟ (دستهای: 'خیر'، 'بله'، 'نامشخص')
- موجودی: کل موجودی حساب شخص. (یک متغیر عددی که میتواند برحسب بدهکاری یا طلبکاری منفی یا مثبت باشد.)

مشخصات دیتاست و ویژگی های آن

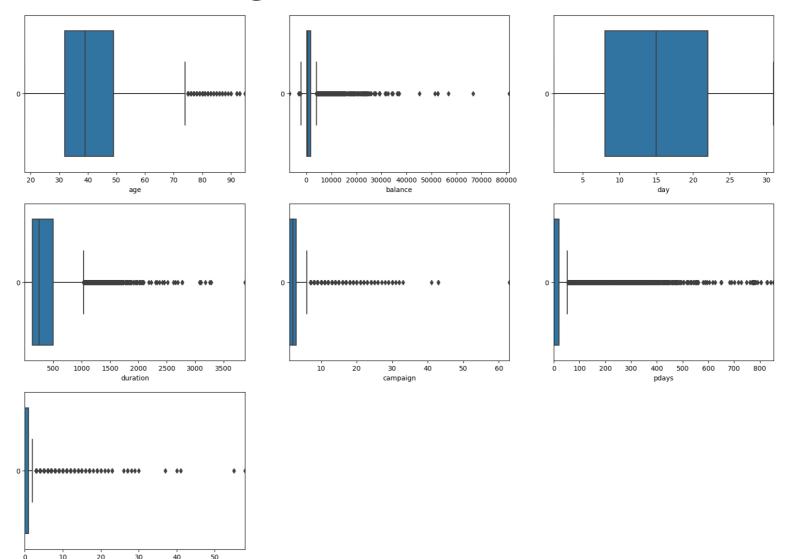
دادههای مربوط به آخرین تماس با هر مشتری در کمپین فعلی:

- نوع بستر ارتباط تماس (دستهای: 'گوشی موبایل'، 'تلفن ثابت')
- آخرین ماه تماس در سال (دستهای: 'ژانویه'، 'فوریه'، 'مارس'، ..., 'نوامبر'، 'دسامبر')
- آخرین روز هفته تماس (دستهای: 'دوشنبه'، 'سهشنبه'،،'چهارشنبه'، 'پنجشنبه'، 'جمعه')
 - مدتزمان آخرین تماس گرفته شده با مشتری بر حسب ثانیه (عددی).
- تعداد تماسهای انجام شده در طول این کمپین و برای این مشتری (عددی، شامل آخرین تماس)
- تعداد روزهایی که پس از تماس آخرین با مشتری از کمپین قبلی گذشته است. (عددی؛ عدد ۹۹۹ به معنای عدم تماس قبلی با مشتری است)
 - تعداد تماسهای انجام شده قبل از این کمپین و برای این مشتری (عددی)
 - نتیجه کمپین بازاریابی قبلی (دستهای: 'شکست'، 'ناموجود'، 'موفقیت')

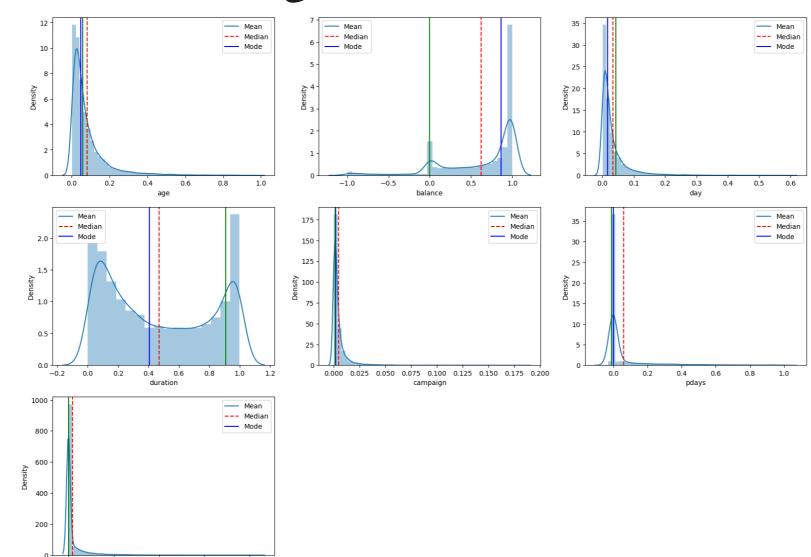
برچسب خروجی (Y) شامل:

آیا مشتری سپرده کوتاهمدت بازکرده است؟ (باینری شامل بله و خیر)





previous



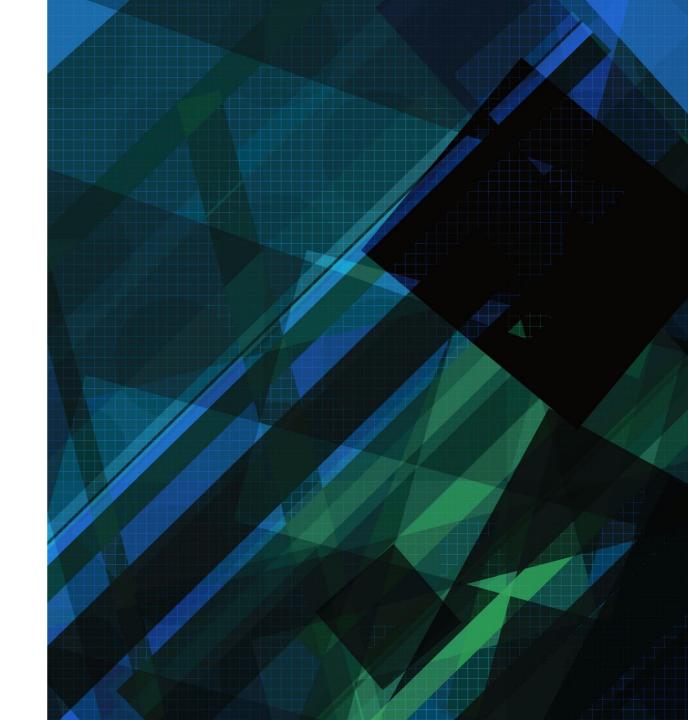
0.02

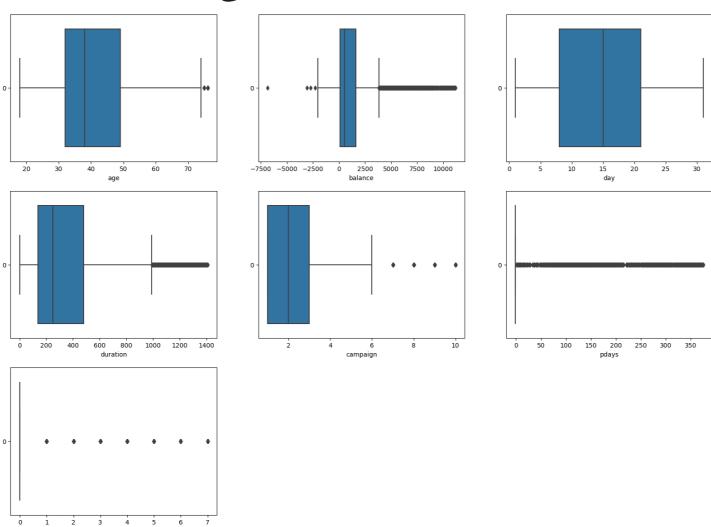
previous

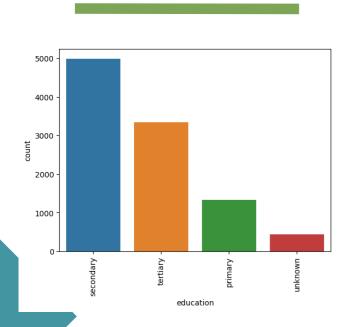
0.04

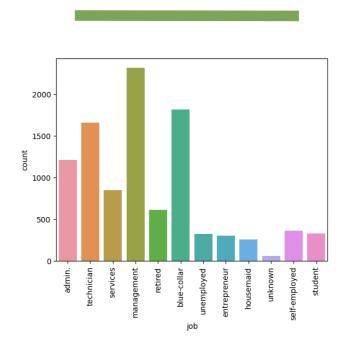
حذف دادهها پرت

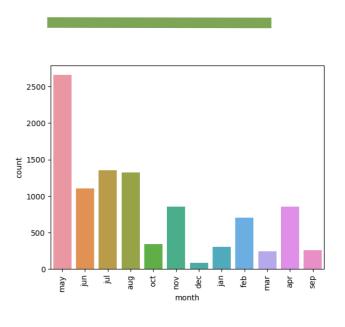
Z-score IQR

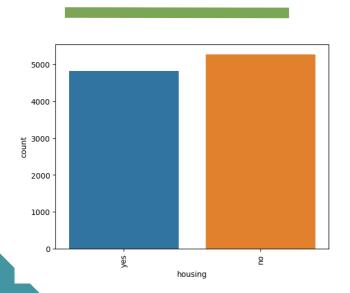


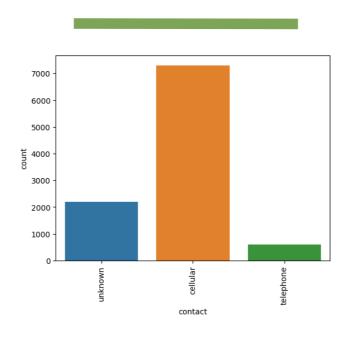


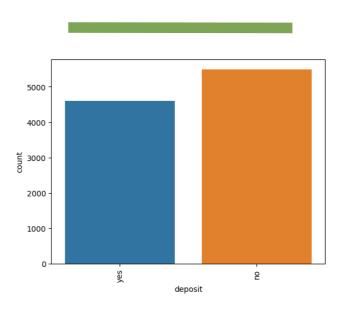




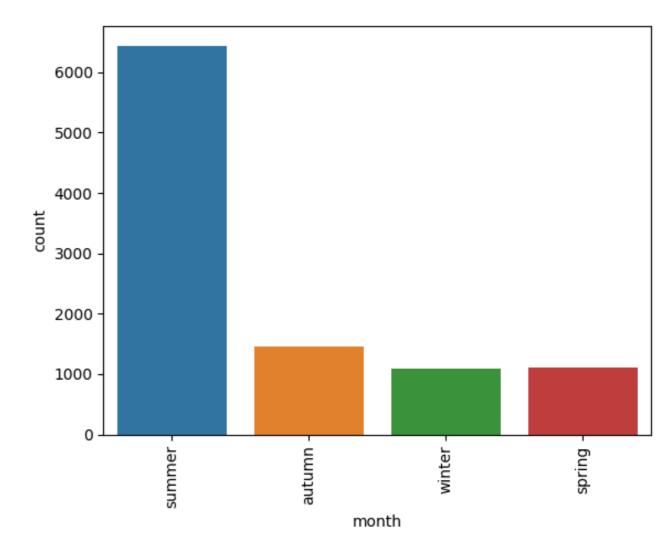








كوچك كردن تعداد بازهها به فصل



اكتشافات دادهاى - سبدبندى

Cut

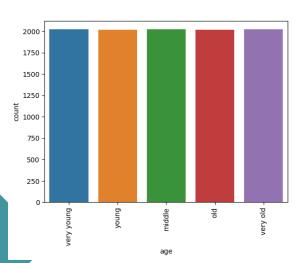
Qcut

Natural Break

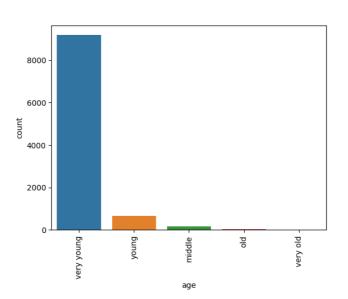


سبد بندی

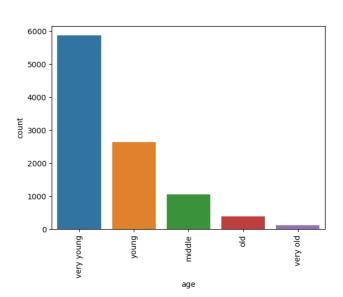
Qcut

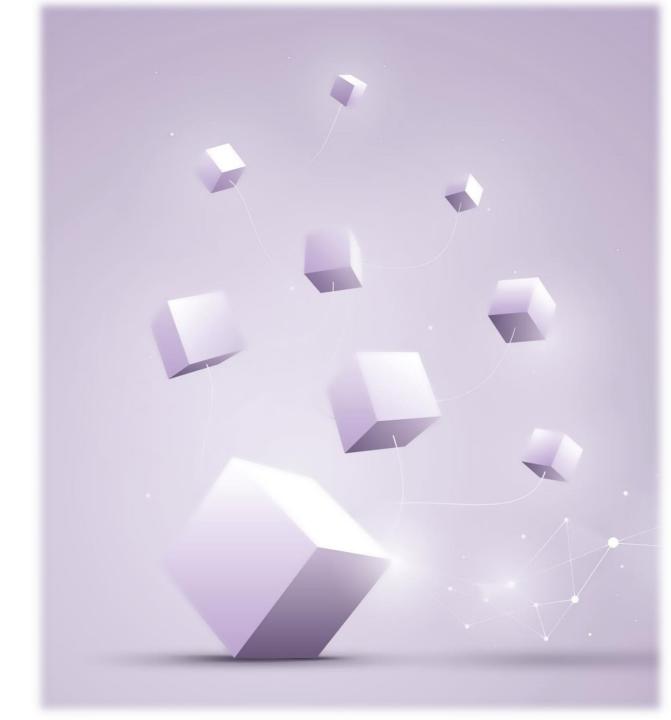


CUT



break





همیستگی و پژ گیها با یکدیگر

-0.03 -0.11 -0.056 0.069 -0.39 -0.0079 0.045 0.086 -0.018 0.37 0.005 0.048 0.11 -0.28 iob - -0.03 0.081 0.15 -0.0085 0.044 -0.14 -0.067 -0.086-0.0087-0.075 -0.02 -0.0087-0.012 0.00940.0038 0.062 marital - -0.11 0.081 0.12 -0.012 0.051 -0.039 -0.064 -0.059 -0.0170.000450.0051 -0.027 0.038 0.045 -0.039 0.072 education -- 0.056 0.15 0.12 -0.015 0.067 -0.12 -0.074 -0.14 -0.02 -0.064 -0.048 -0.029 0.019 0.037 -0.043 0.11 default - 0.069 -0.0085-0.012 -0.015 **-0.2** 0.0088 0.078 0.034 0.062 0.0022 0.086 0.05 **-**0.027 **-**0.03 0.042 **-**0.034 -0.39 0.044 0.051 0.067 -0.2 -0.12 -0.13 -0.038 -0.34 -0.0085 -0.68 -0.26 -0.13 -0.091 -0.043 0.056 housing -0.0079 -0.14 -0.039 -0.12 0.0088 -0.12 0.074 0.23 -0.0026 0.23 0.038-0.0021 0.072-0.0075 -0.05 -0.21 loan - 0.045 -0.067 -0.064 -0.074 0.078 -0.13 0.074 0.0037 0.036 0.025 0.057 0.047 0.0028-0.0039 0.02 -0.11 contact - 0.086 -0.086 -0.059 -0.14 0.034 -0.038 0.23 0.0037 0.8 -0.0087-0.017 -0.02 0.062 -0.34 -0.0026 0.036 0.045 -0.033 0.31 -0.027 0.0085 0.12 -0.25 -0.035 -0.054 0.024 0.01 -0.042 -0.043 month --0.018 -0.0750.00045-0.064 0.0022-0.0085 0.23 0.025 0.3 -0.033 1 duration - 0.37 -0.02 0.0051-0.048 0.086 -0.68 0.038 0.057 0.01 0.31 -0.035 0.22 0.0057 0.048 0.12 0.12 0.71 -0.0087-0.027 -0.029 0.05 -0.26 -0.0021 0.047 0.053 -0.054 0.22 -0.039-0.0025 0.11 -0.23 pdays - 0.005 -0.012 0.038 0.019 -0.027 -0.13 0.072 0.0028 -0.2 -0.027 0.024 0.0057-0.039 -0.69 0.044 previous - 0.048 0.0094 0.045 0.037 -0.03 -0.091-0.00750.0039 -0.17 0.0085 0.01 0.048-0.0025 0.69 -0.54 0.095 poutcome - 0.11 -0.0038-0.039 -0.043 0.042 -0.043 -0.05 0.02 0.25 0.12 -0.042 0.12 0.11 -0.69 -0.54 deposit - -0.28 0.062 0.072 0.11 -0.034 0.056 -0.21 -0.11 -0.26 -0.25 -0.043 0.12 -0.23 0.044 0.095 -0.11

- 0.8

- 0.6

- 0.4

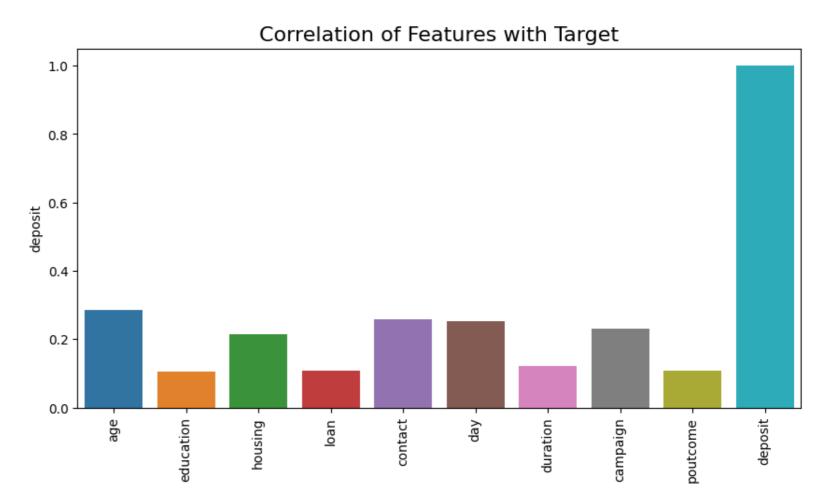
- 0.2

- 0.0

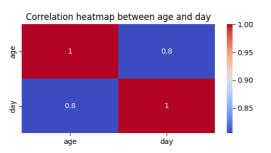
- -0.2

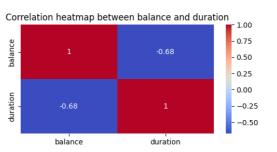
21

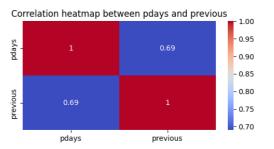
همبستگی ویژگیها با متغیر هدف

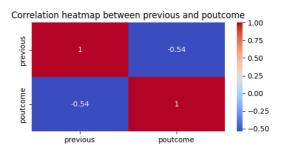


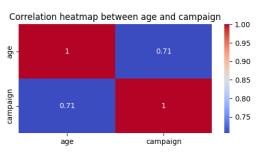
متغیر هایی با بیشترین همبستگی

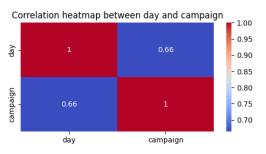


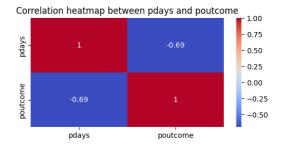


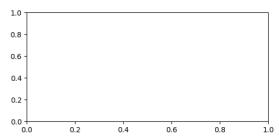




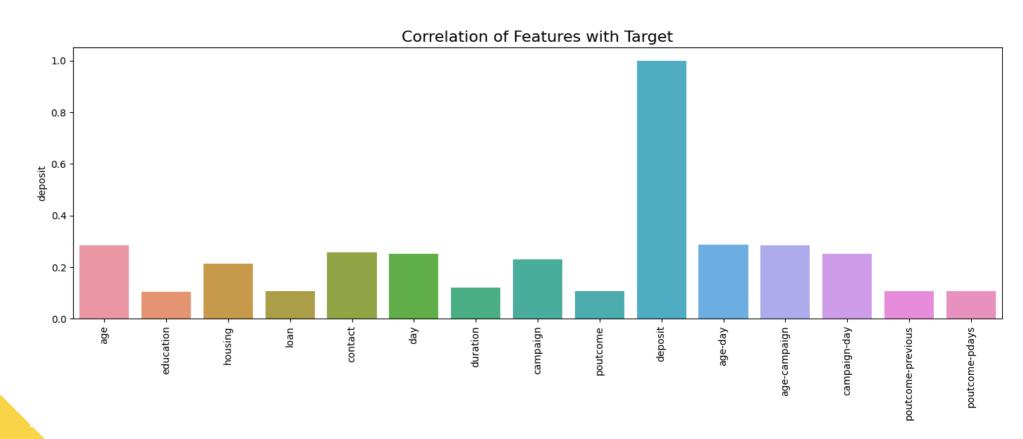


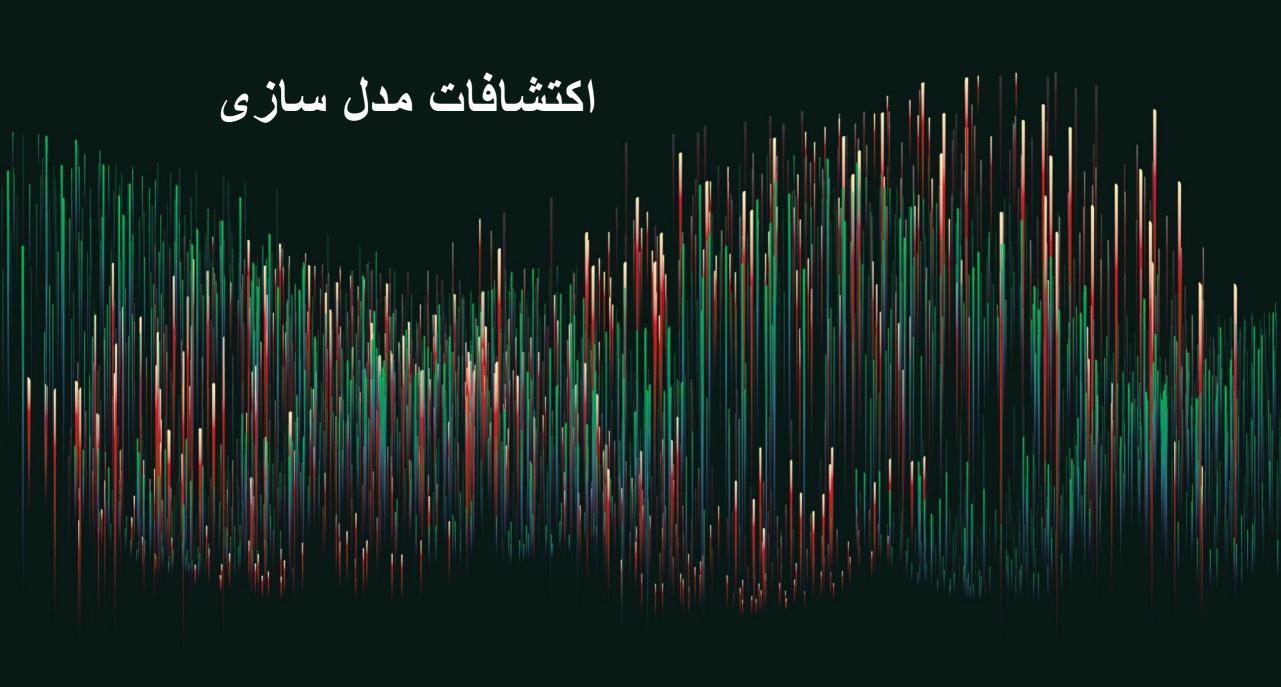






همبستگی ویژگیها با متغیر هدف





اكتشافات مدلسازى -فاز پيشمدل

تقسیم دادهها به ۳۰ درصد تست و ۷۰ درصد آموزش

cross validation

بالانس دادهها با standard scaler

اكتشافات مدلسازى -انتخاب مدل

خوشه بندى

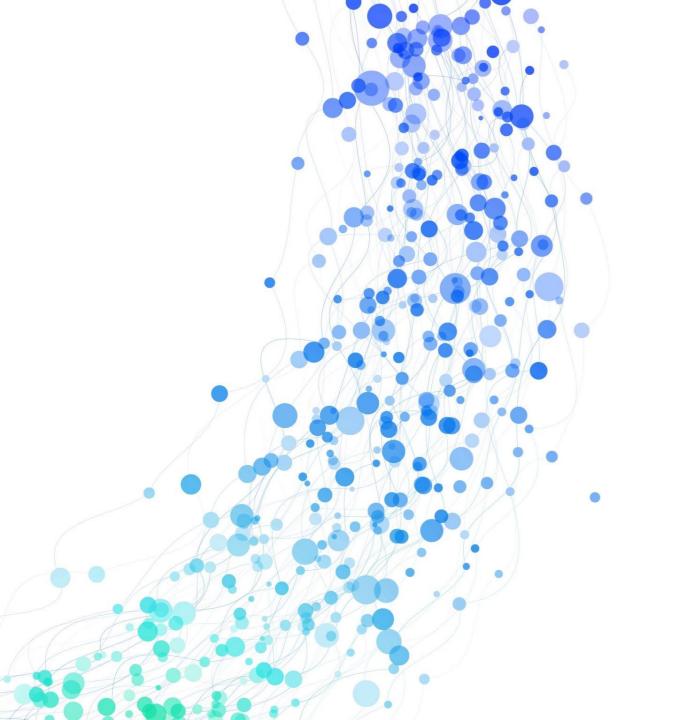
دسته بندی

دسته بندی -اکتشافات مدلسازی

XGBoost Neural Network Random Forest SVM Stochastic Gradient Boosting Machine Classifier

دسته بندی -اکتشافات مدلسازی

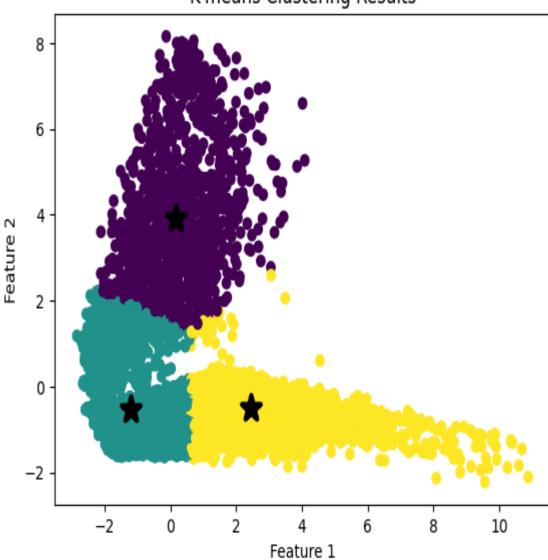
· test accuracy: 0.822277227722							
	precision	recall	f1-score	support			
0	0.86	0.81	0.83	1109			
1	0.78	0.84	0.81	911			
accuracy			0.82	2020			
macro avg	0.82	0.82	0.82	2020			
weighted avg	0.82	0.82	0.82	2020			



اكتشافات مدلسازى-خوشه بندى مشتريان

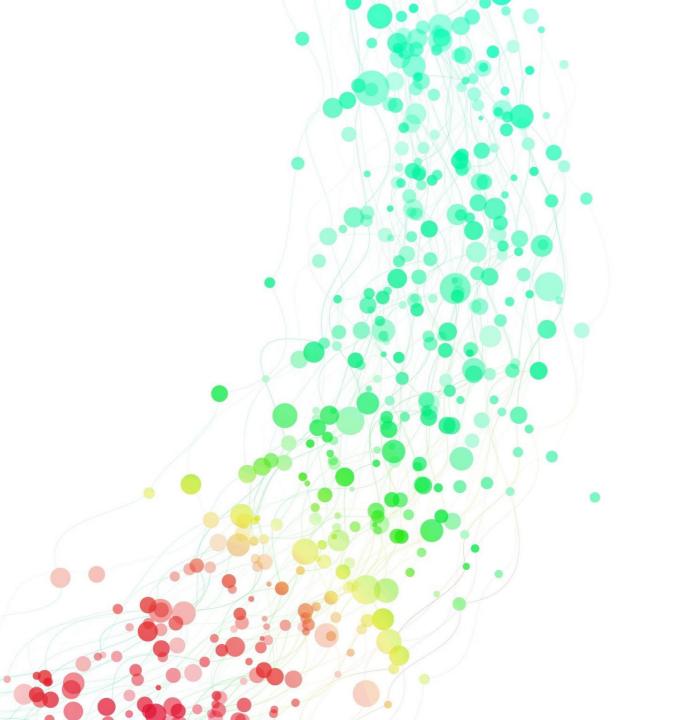
با استفاده از ++K-means کاهش ویژگیها به دوتا روش PCA

K-means Clustering Results



اكتشافات مدلسازى-خوشه بندى مشتريان

با استفاده از ++K-means کاهش ویژگیها به دوتا روش PCA



نتايج

پیشنهادات بیزنسی

ماه برگزاری کمپین

بهترین ماه: ماه می

بیشترین رد تماس: ماه می

بهترین ماهها برای برگزاری کمپین به ترتیب: مارچ، سپتامبر،اکتبر

ایجاد پرسشنامه

افزایش طول مکالمه با استفاده از پرسشنامه برای کاربران

مهمترین ویژگی استخراج شده در ریشه درخت



- [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
- S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
- S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]
- archive.ics.uci.edu/ml/datasets/Bank+Marketing
- https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset/code?datasetId=4471

