دانشگاه کردستان

University of Kurdistan

یادگیری ماشین

موضوع: تحلیل اکتشافی و پیش‌پردازش داده

churn modeling dataset

استاد مدرس

دکتر صادق سلیمانی

دانشجو

ایمان قوامی

زمستان ۱۴۰۲

دیتاستی که در این تحلیل مورد بررسی قرار داده می‌شود از Kaggle به دست آمده و برای مدل‌سازی ریزش مشتریان است. این دیتاست شامل اطلاعاتی در مورد ۱۰۰۰۰ مشتری بانک است و پارامتر هدف یک متغیر دودویی است که نشان می‌دهد آیا مشتری بانک را ترک کرده است یا همچنان مشتری است. از این تعداد، ۷۹۶۳ نمونه با کلاس مثبت (maintained) و ۲۰۳۷ نمونه با کلاس منفی (exited) وجود داشت. متغیر هدف نشان‌دهنده پرچم دودویی ۱ است که نشان‌می‌دهد حساب مشتری بسته شده است و ۰ که نشان می‌دهد مشتری حفظ شده است.

جدول ۱: توضیحات مربوط به دیتاست

Feature Name	Feature Description
Row Number	Row numbers from 1 to 10000.
Customer Id	Unique Ids for bank customer identification.
Surname	Customer's last name.
Credit Score	Credit score of the customer.
Geography	The country from which the customer belongs.
Gender	Male or Female.
Age	Age of the customer.
Tenure	Number of years for which the customer has been with the bank.
Balance	Bank balance of the customer.
Num of Products	Number of bank products the customer is utilizing (savings account, mobile banking, internet banking etc.).
Has Cr Card	Binary flag for whether the customer holds a credit card with the bank or not.
Is Active Member	Binary flag for whether the customer is an active member with the bank or not.
Estimated Salary	Estimated salary of the customer in Dollars.
Exited	Binary flag 1 if the customer closed account with bank and 0 if the customer is retained.

نکته بسیار مهم در بحث پیش‌پردازش داده در همین ابتدا، نامربوط بودن آن‌ها است. داده‌ها یا ویژگی‌هایی که تأثیری روی موضوع مورد بحث ندارند، به عنوان ویژگی‌های غیرمرتبط در نظر گرفته می‌شوند. نگه‌داشتن چنین ویژگی‌هایی گاهی اوقات می‌تواند بر عملکرد طبقه‌بندی‌کننده‌ها تأثیر بگذارد. با توجه به دیتاست ریزش مشتری، ویژگی‌هایی با نام Row Number، Customer Id، Surname و Geography هیچ ارتباطی با پیش‌بینی ندارند. بنابراین، این ویژگی‌ها به صورت دستی در این مطالعه باید نادیده گرفته شوند.

مورد بعدی، بحث تبدیل داده است. تبدیل داده، فرآیند تغییر شکل دادن داده از یک فرمت به فرمت دیگر است. در این دیتاست بهتر است برای ویژگی Gender که دارای دو مقدار Female و Male است، به ترتیب مقدار ۰ و ۱ قرار دهیم.

در پردازش داده، Oversampling و Undersampling دو رویکرد برای پیکربندی توزیع کلاسیِ داده‌های معین هستند. از آنجایی که داده‌ها به شدت نامتوازن هستند (۷۹۶۳ نمونه با کلاس مثبت و ۲۰۳۷ نمونه با

کلاس منفی) و حجم نمونه‌داده در دسترس کم است، مقاله پیشنهاد داده است که از تکنیک Oversampling استفاده شود. زیرا در صورت ترجیح Undersampling، حجم داده به گونه‌ای کاهش می‌یابد که داده کافی برای ساخت مدل وجود نخواهد داشت. بنابراین، این مطالعه از روش Oversampling تصادفی با نمونه‌گیری مجدد از کلاس اقلیت (کلاس منفی) استفاده می‌کند.

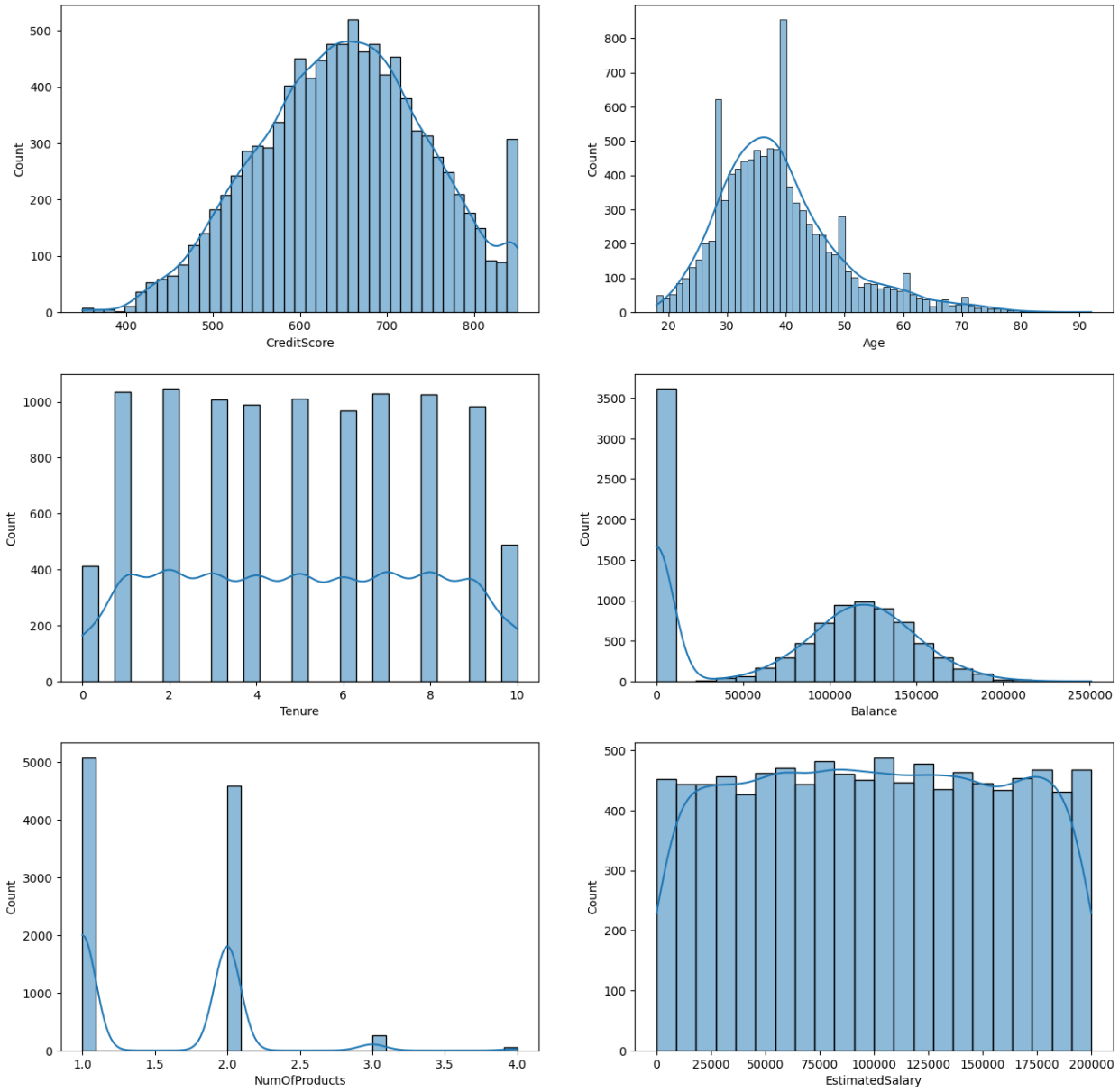
در دیتاست مربوطه بعد از بررسی، معلوم شد که هیچ داده‌ای با مقدار Null وجود ندارد و به همین دلیل لازم به مدیریت Missing value data نبود. همچنین هیچ سطر تکراری در دیتاست وجود نداشت.

```
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CreditScore      10000 non-null  int64
 1   Geography        10000 non-null  object
 2   Gender           10000 non-null  object
 3   Age              10000 non-null  int64
 4   Tenure           10000 non-null  int64
 5   Balance          10000 non-null  float64
 6   NumOfProducts    10000 non-null  int64
 7   HasCrCard        10000 non-null  int64
 8   IsActiveMember   10000 non-null  int64
 9   EstimatedSalary  10000 non-null  float64
 10  Exited           10000 non-null  int64
dtypes: float64(2), int64(7), object(2)
memory usage: 859.5+ KB
```

یک تحلیل آماری از دیتاست به شرح زیر است:

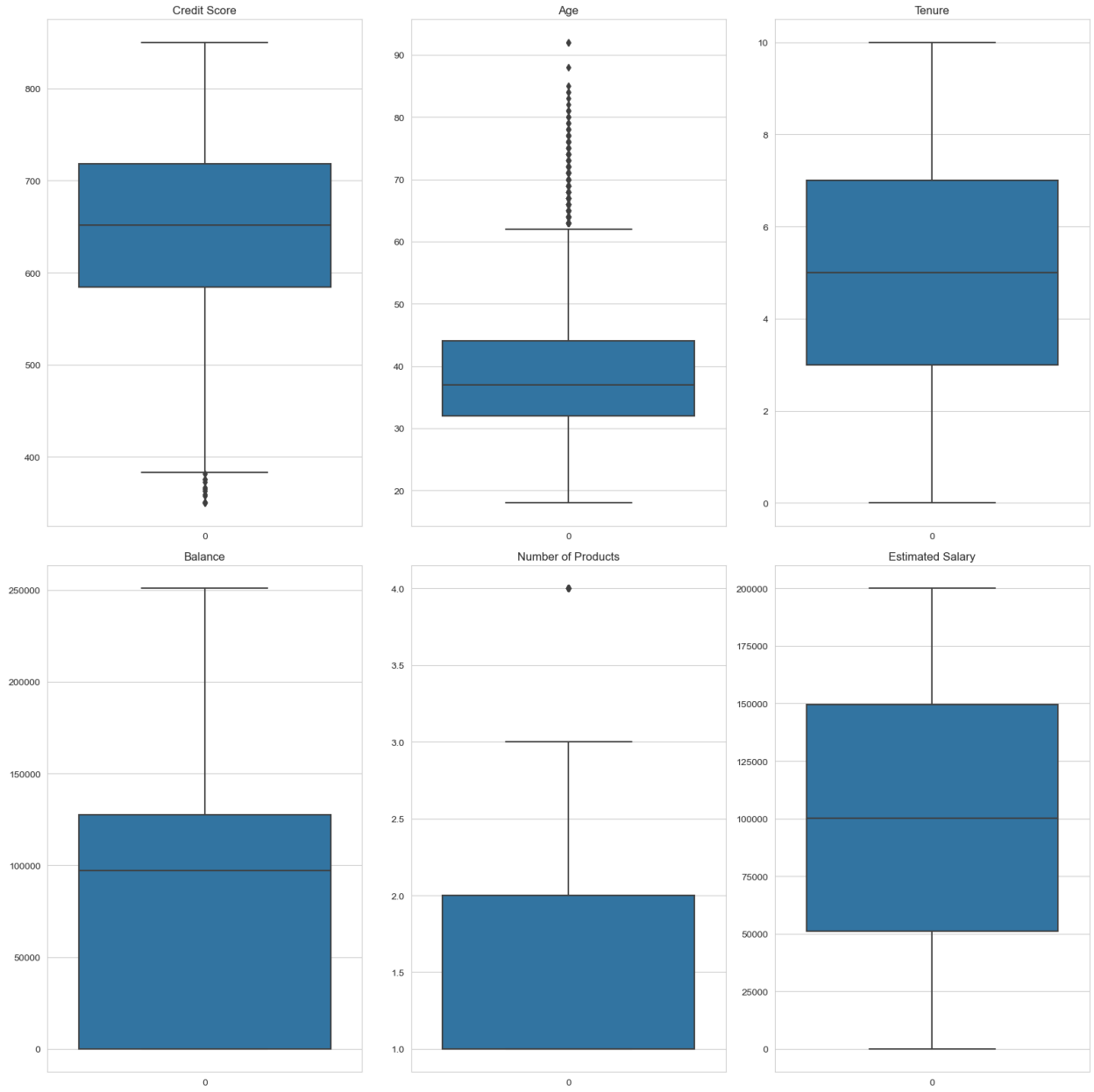
	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

برای ویژگی‌هایی که مقدار عددی و پیوسته دارند، نمودار هیستوگرام آن رسم شد و توزیع داده‌ها به شرح زیر است:



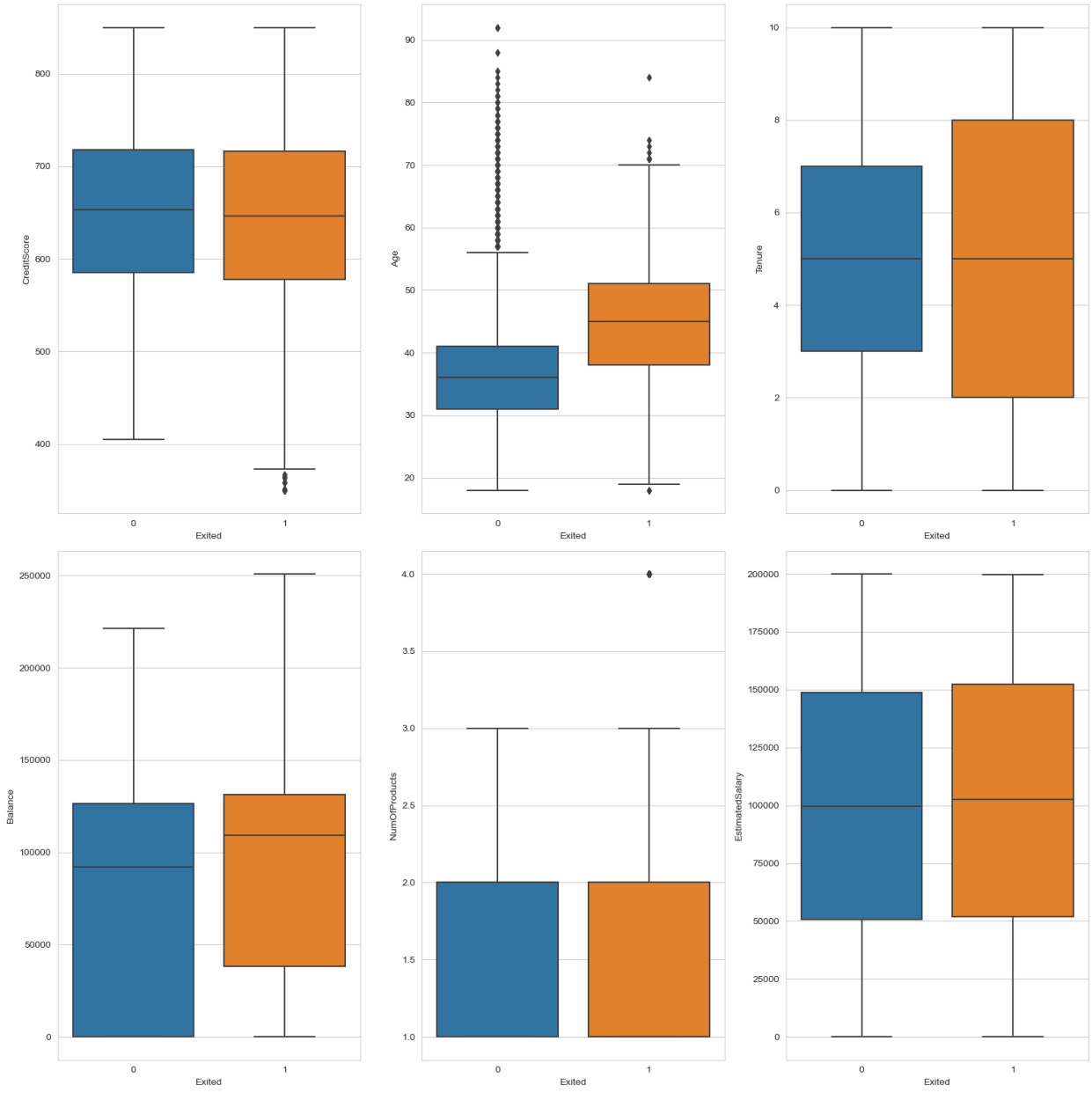
- *توزیع Age*: توزیع آن به سمت چپ خمیده است. سنین از ۲۰ تا ۴۵ سال دارای چگالی بالاتری هستند.
- *توزیع CreditScore*: توزیع آن به صورت نرمال است. امتیازهای اعتباری از ۵۰۰ تا ۷۰۰ دارای چگالی بالاتری هستند.
- *توزیع Balance و NumOfProducts*: توزیع آن‌ها به صورت نرمال است. تعداد محصولات از ۱ تا ۲ دارای چگالی بالاتری هستند.
- *توزیع EstimatedSalary و Tenure*: توزیع آن‌ها تقریبا مساوی با توزیع یکنواخت است.

نمودار جعبه‌ای ویژگی‌های عددی به صورت زیر است:

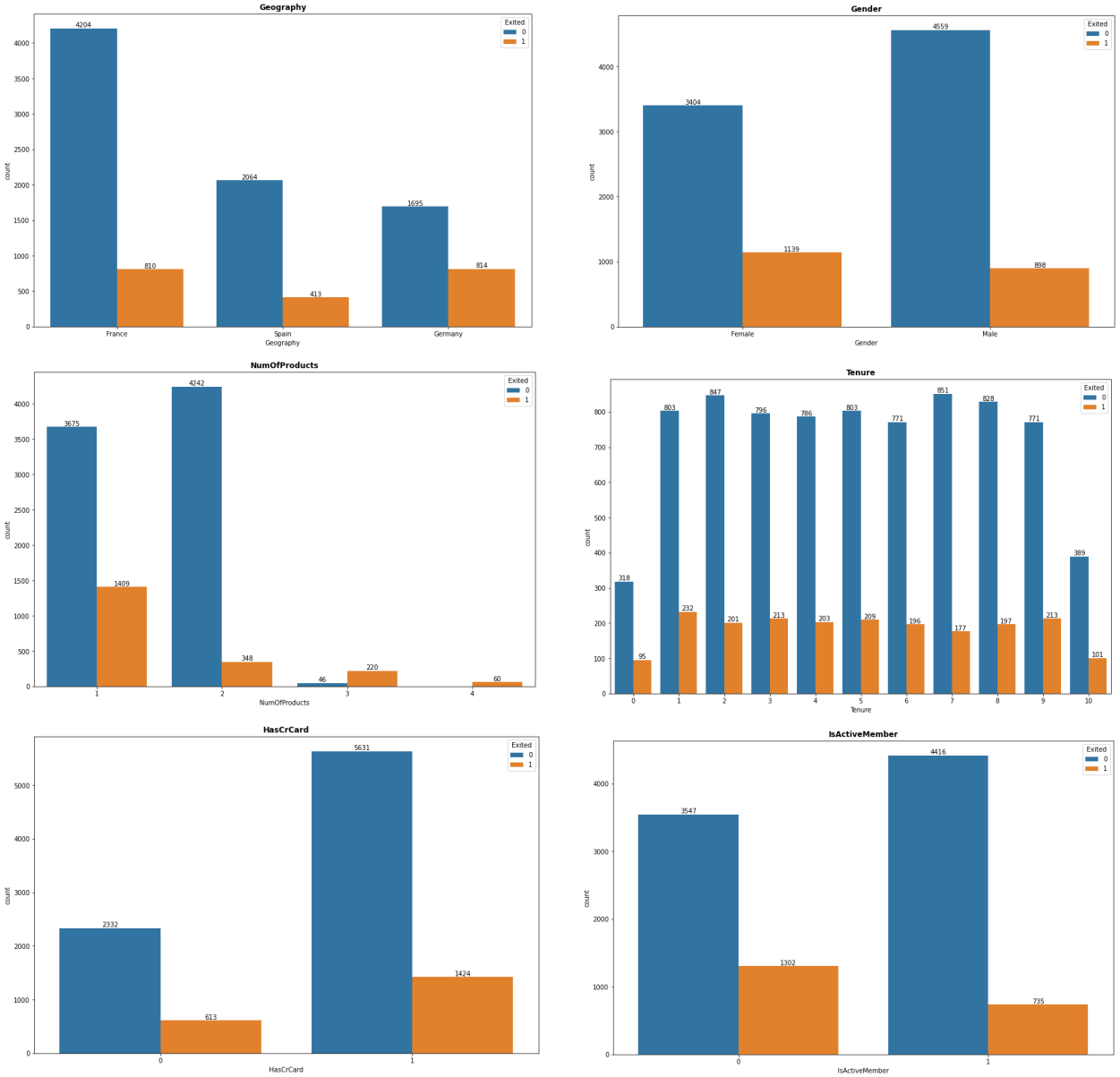


در شکل بالا کاملا واضح است که در سه ویژگی CreditScore، Age و NumberOfProduct ما مقداری Outlier داریم که باید آن‌ها را حذف کنیم. این کار در قسمت پیش‌پردازش داده انجام خواهد شد.

رسم نمودار جعبه‌ای برای هر یک ویژگی‌های فوق در کنار ویژگی هدف:

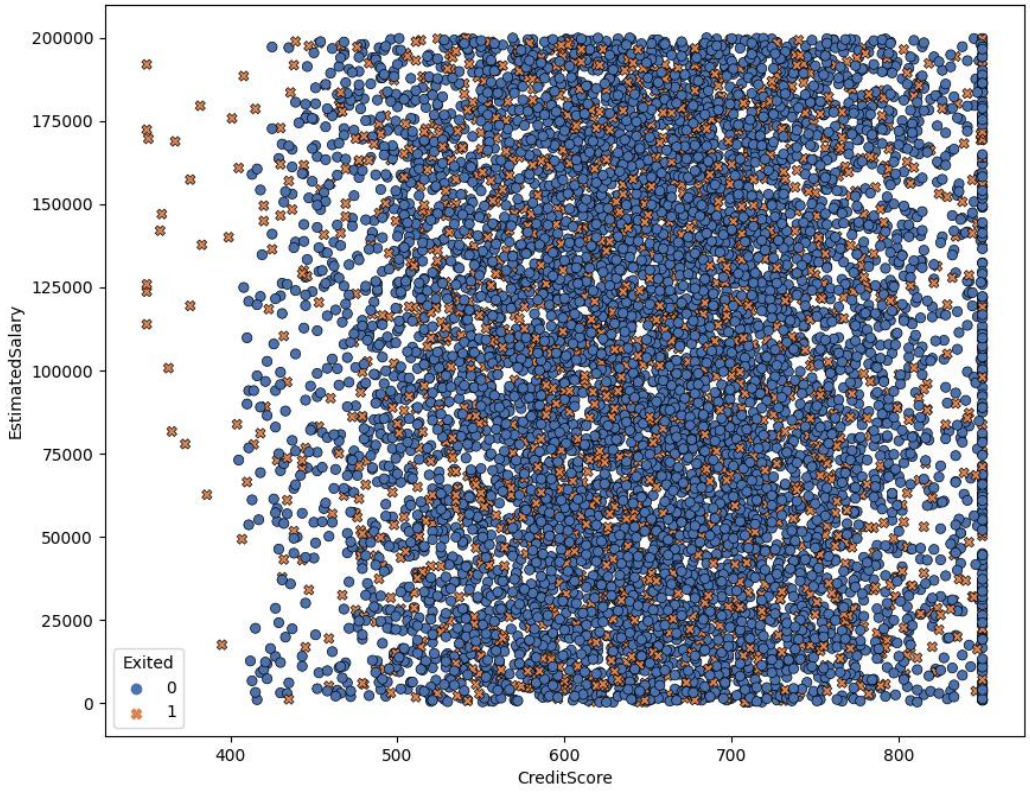


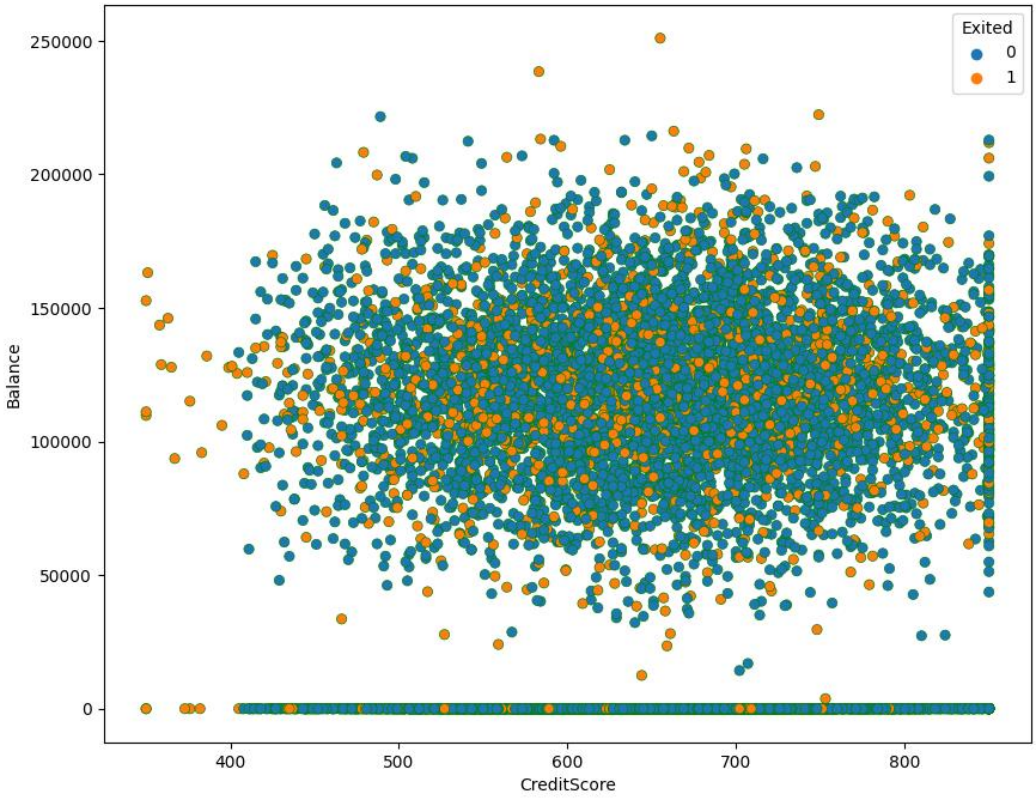
رسم نمودار شمارشی ویژگی‌های مهم:

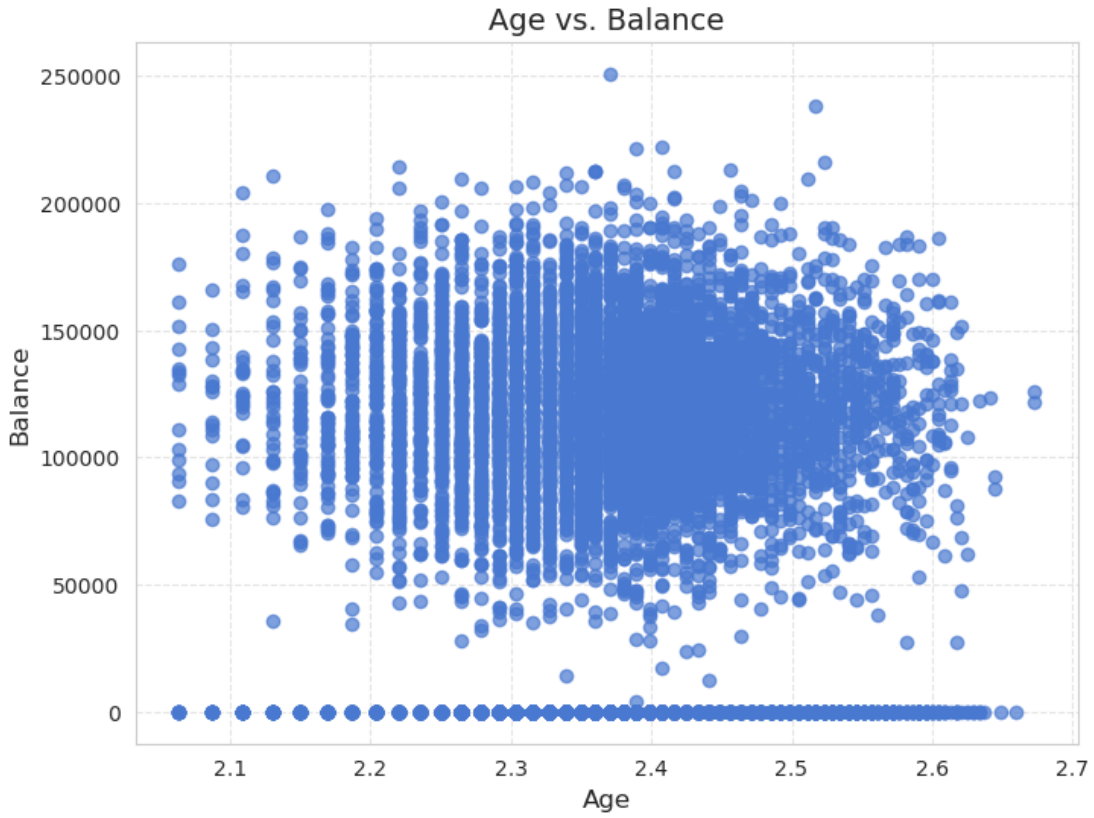


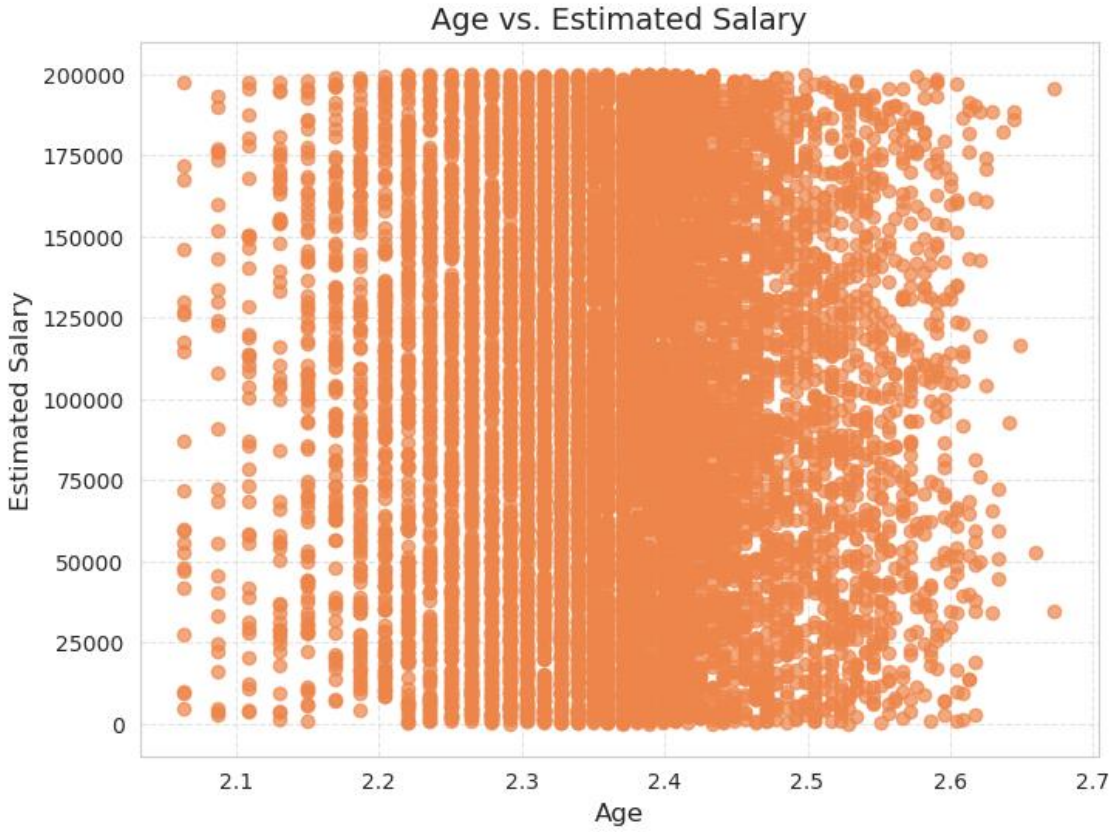
- یک سوم مشتری‌های آلمانی تصمیم به ترک بانک گرفته‌اند.
- یک سوم مشتریان خانم و یک پنجم مشتریان آقا تصمیم به ترک بانک گرفته‌اند.
- اکثریت قریب به اتفاق مشتریان بانک دارای یک یا دو محصول هستند و مشتریان با محصولات بیشتر تمایل دارند بانک را ترک کنند.
- افرادی که کارت اعتباری دارند یا ندارند، دارای نتایج یکسان بوده و ۲۵٪ تصمیم به ترک بانک گرفته‌اند.
- یک سوم از اعضای فعال در بانک و یک ششم اعضای غیرفعال تصمیم به ترک بانک گرفته‌اند.
- میانگین حقوق تخمینی مردان در اسپانیا و فرانسه بالاتر است اما در آلمان نه.

وابستگی برخی از ویژگی‌های مهم به شرح نمودارهای زیر هستند:

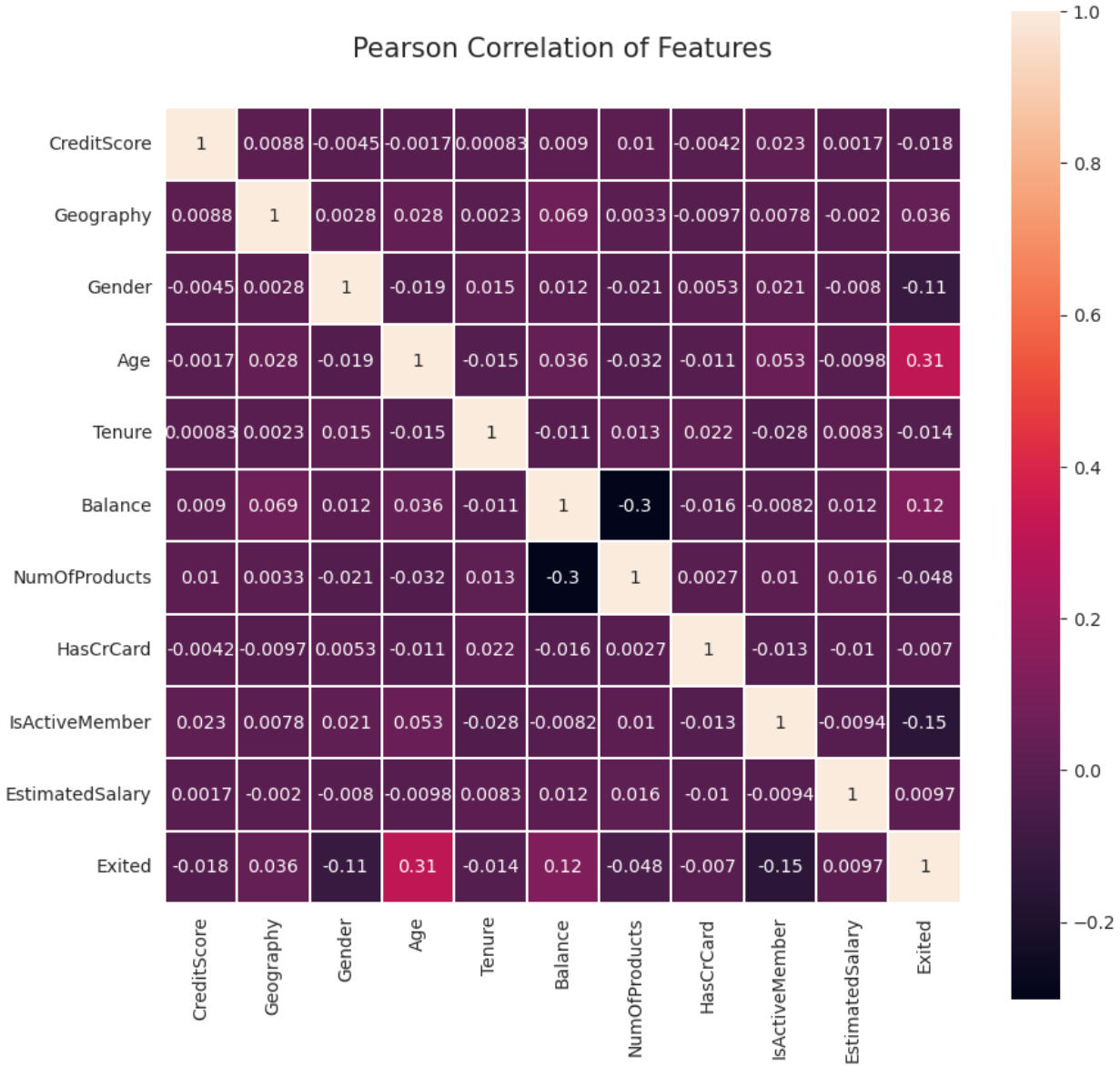





Age vs. Balance
Balance
250000
200000
150000
100000
50000
0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
Age


Age vs. Estimated Salary
Estimated Salary
200000
175000
150000
125000
100000
75000
50000
25000
0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
Age

ماتریس همبستگی تمام ویژگی‌های نیز به صورت زیر است:



در مورد بالانس بودن نمونه‌ها، در ابتدا گفته شد که از ۱۰۰۰۰ نمونه، ۷۹۶۳ نمونه با ویژگی هدف مثبت (maintained) و ۲۰۳۷ نمونه با ویژگی هدف منفی (exited) داریم. طبق بررسی مقاله، به دلیل اینکه تعداد نمونه‌ها آنقدر زیاد نیست که بخواهیم تعداد نمونه‌های مثبت را حذف کنیم، از تکنیک Oversampling استفاده می‌کنیم که طی آن تعداد هر دو نمونه‌های متعل به دو کلاس مثبت و منفی برابر شوند. به عبارتی دیگر، در نهایت ما تعداد ۱۵۹۲۶ نمونه خواهیم داشت.

In [19]:
```
X = data.drop('Exited',axis = 1)
y = data['Exited']
```

In [20]:
```
from imblearn.over_sampling import RandomOverSampler
ros = RandomOverSampler(sampling_strategy = 'minority')
X_train_balanced, y_train_balanced = ros.fit_resample(X, y)
```

In [21]:
```
plt.figure(figsize=(12,6))
plt.title("Balanced target variable", fontsize=15, ha='center')
ax = sns.countplot(x=y_train_balanced, data=data)
plt.show()
```