PROJE 3

SINIFLANDIRMA MODELİ

Veri Bilimi

Batuhan Yıldız – Betül Uyar Can

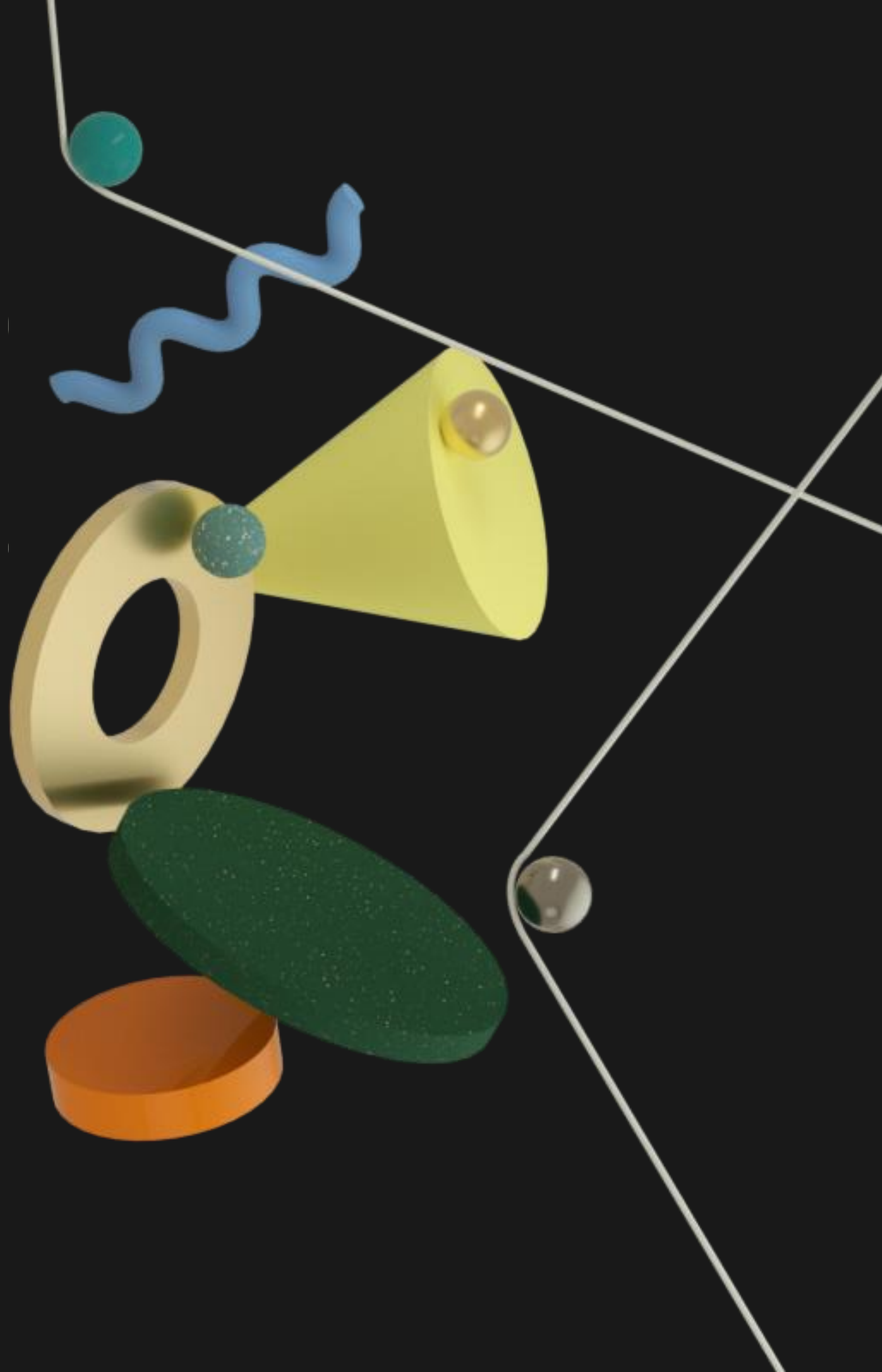
İÇERİK

İnsan hedef, gerçeklik ise manipüle edilmeye çalışılan bir olgudur.

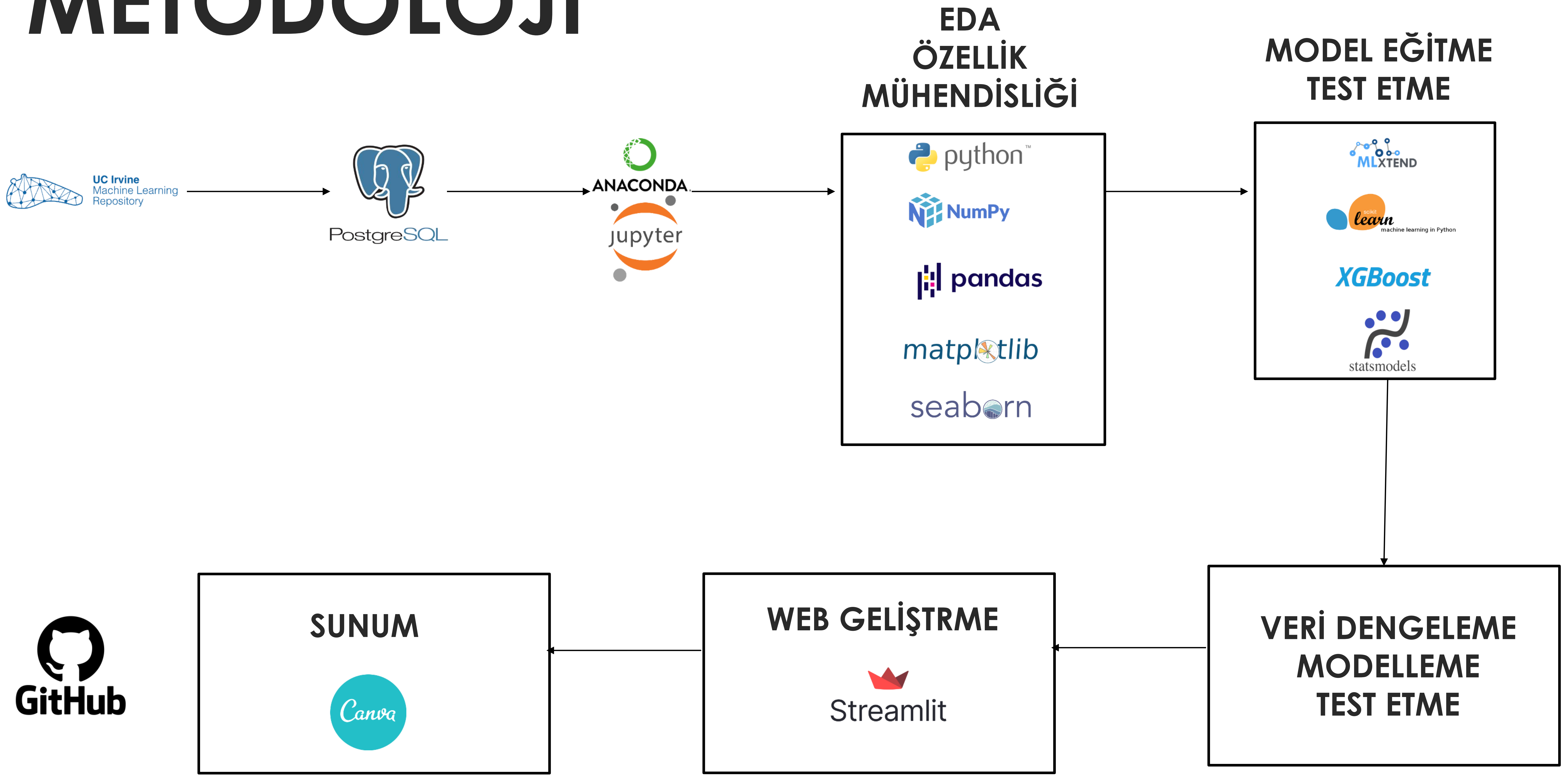
Akan Abdula - Öngörülemeyenler

AMACIMIZ

UCI Machine Learning Repository sitesinden çektiğimiz Bank Marketing veri setini içeriyor. Portekizli bir bankanın doğrudan pazarlama kampanyalarına yönelik telefon görüşmelerinden elde ettiği verilerle müşterilerin vadeli mevduata abone olup olmayacağını tahmin etmeye yönelik bir sınıflandırma projesidir.



METODOLOJİ


UC Irvine
Machine Learning
Repository
PostgreSQL
ANACONDA
jupyter
EDA
ÖZELLİK
MÜHENDİSLİĞİ
python
NumPy
pandas
matplotlib
seaborn
MODEL EĞİTME
TEST ETME
MLXTEND
scikit learn
machine learning in Python
XGBoost
statsmodels
VERİ DENGELEME
MODELLEME
TEST ETME
WEB GELİŞTRME
Streamlit
SUNUM
Canva
GitHub

EDA

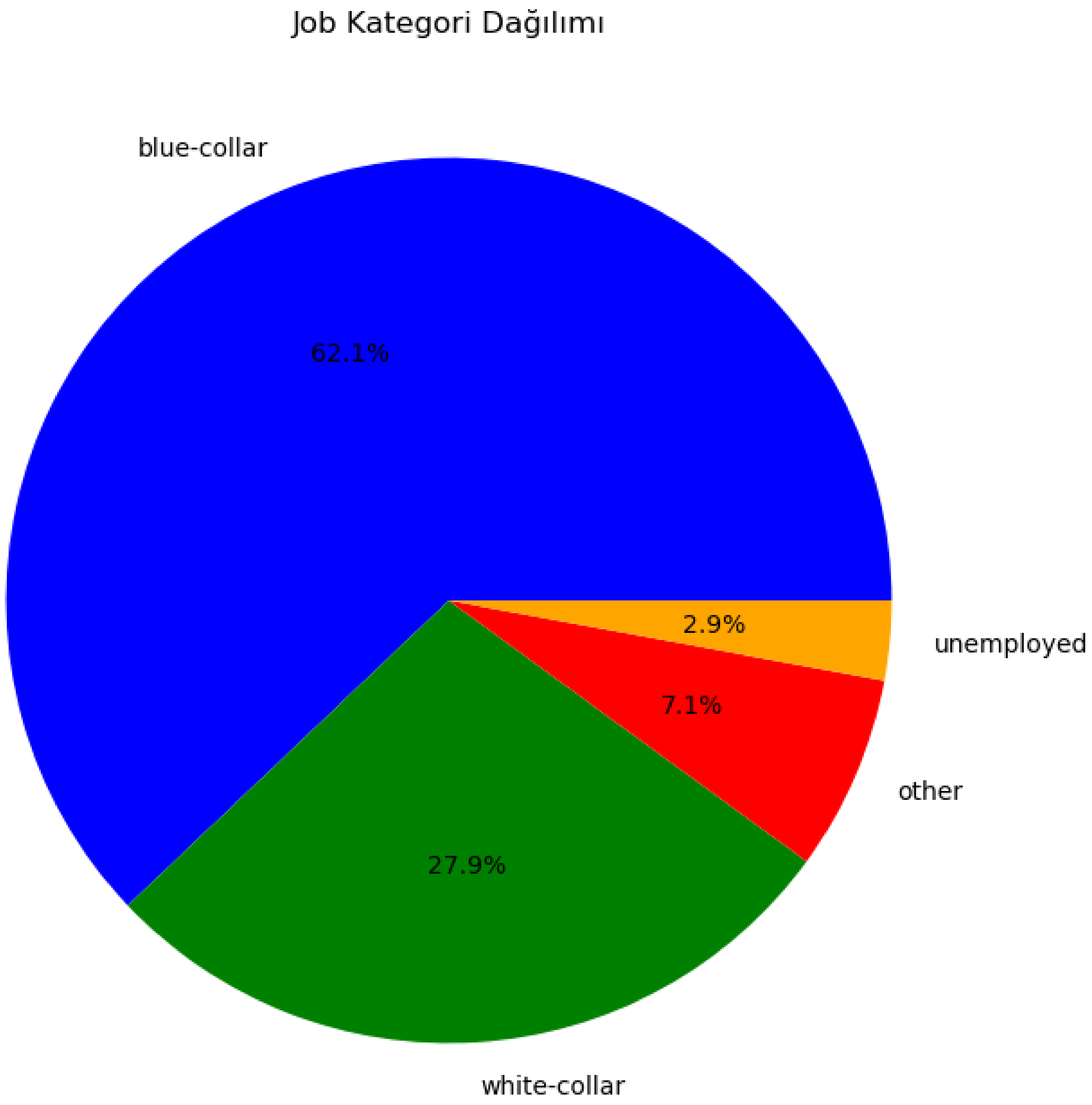
(Keşifsel Veri Analizi)

Veri Çerçevesi Özeti

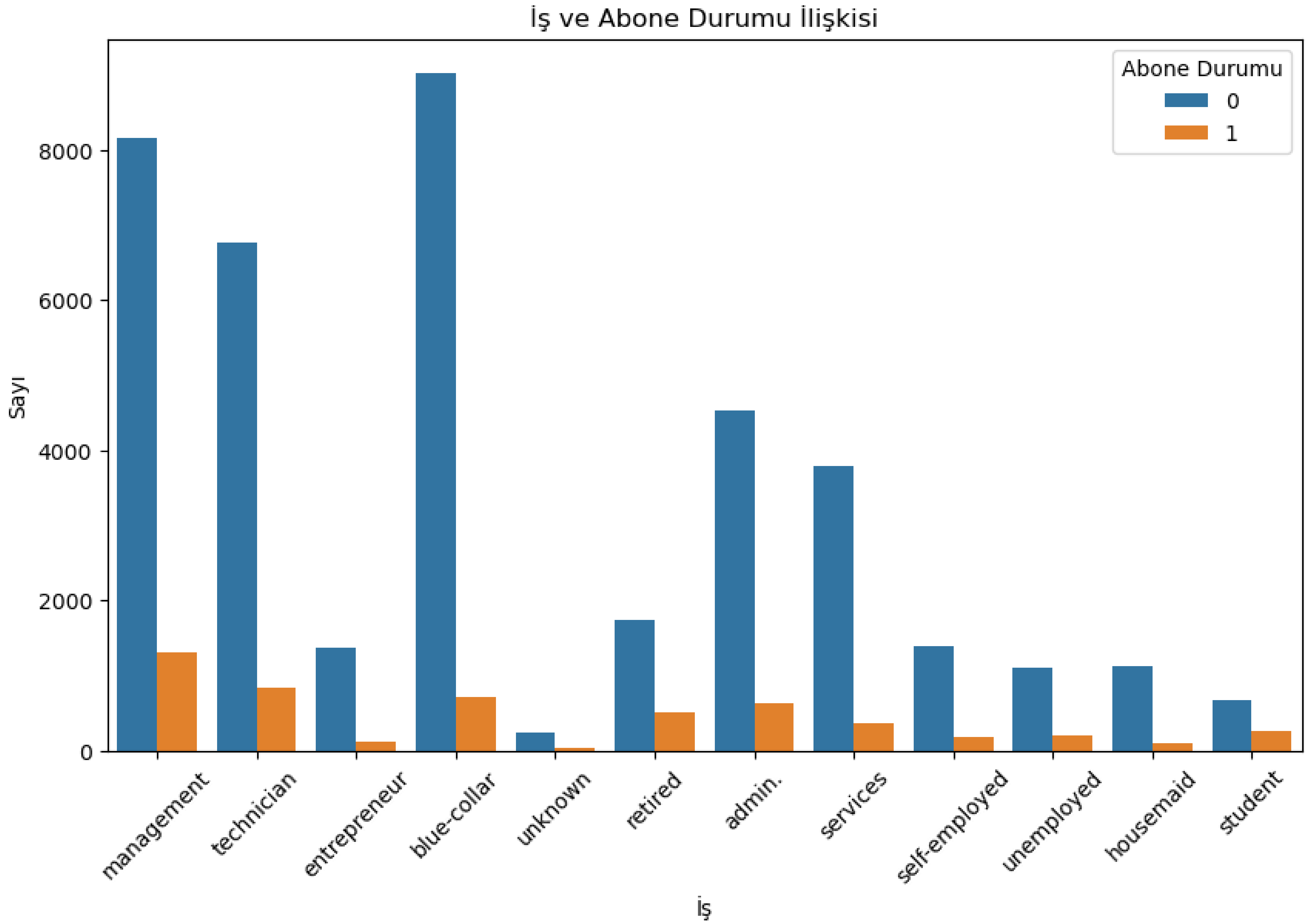
In [44]:
```python
#elimizde 17 sütun var.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   age        45211 non-null  int64 
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64 
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64 
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64 
 12  campaign   45211 non-null  int64 
 13  pdays      45211 non-null  int64 
 14  previous   45211 non-null  int64 
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

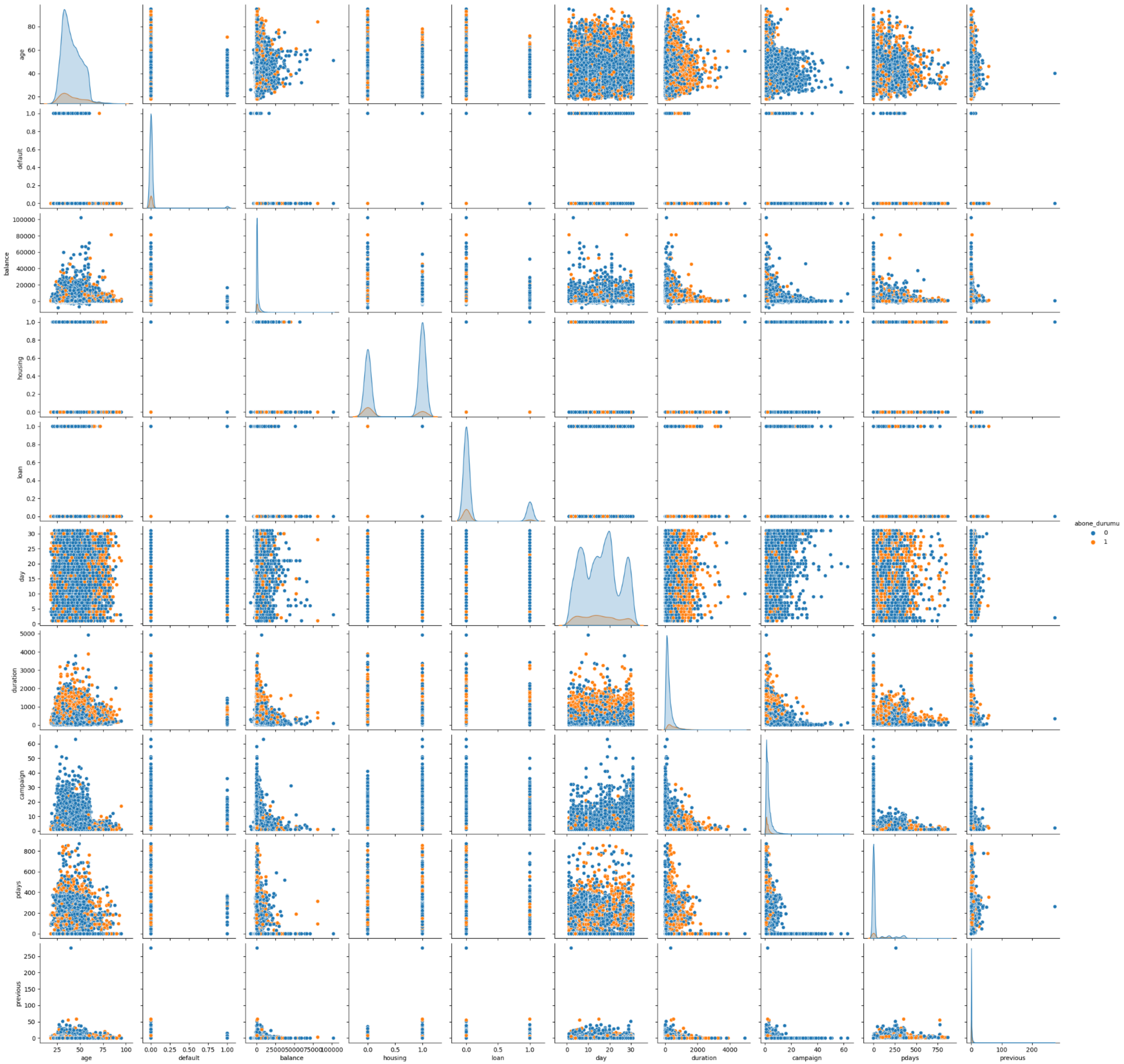
Veri Türü	Boş Olmayan Sayısı	Benzersiz Değer Sayısı	Örnek Değerler
object	4	4	['blue-collar', 'white-collar', 'other', 'unemployed']
int64	4	4	[27894, 12524, 3202, 1303]



Job sütunundaki değerler kategorize edildi.

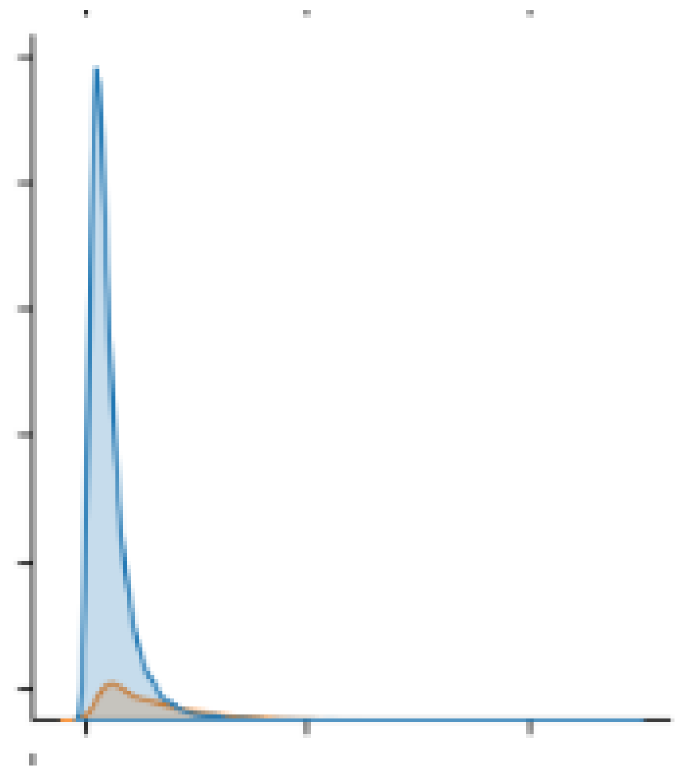


İş ve Abone Durumu İlişkisi
Abone Durumu
0
1
Sayı
0
2000
4000
6000
8000
management
technician
entrepreneur
blue-collar
unknown
retired
admin.
services
self-employed
unemployed
housemaid
student
İş



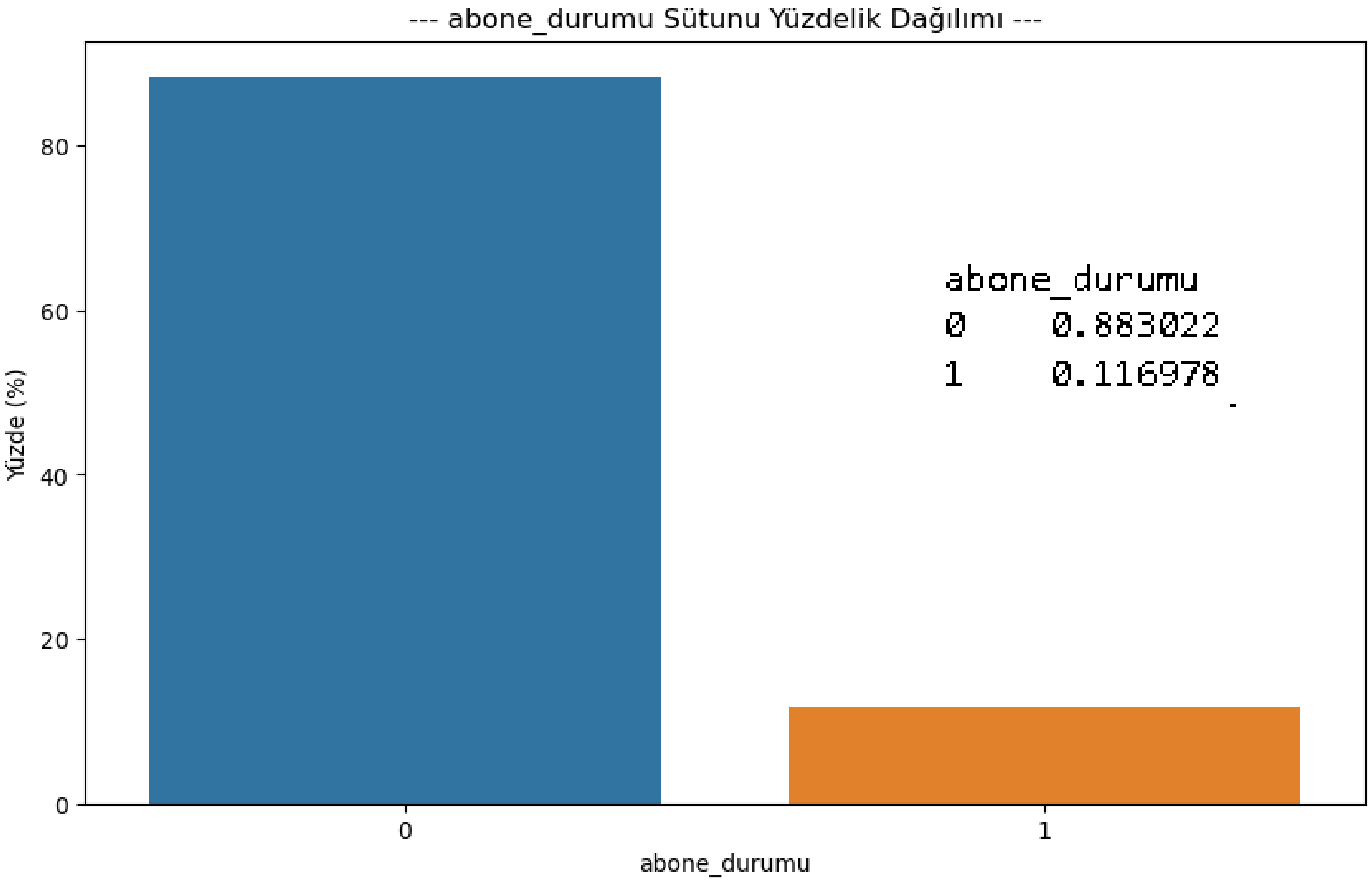
Değişkenler arasındaki korelasyonu ve ilişkiyi anlamak için pairplot çizdirdik.

Duration ön plana çıktı.



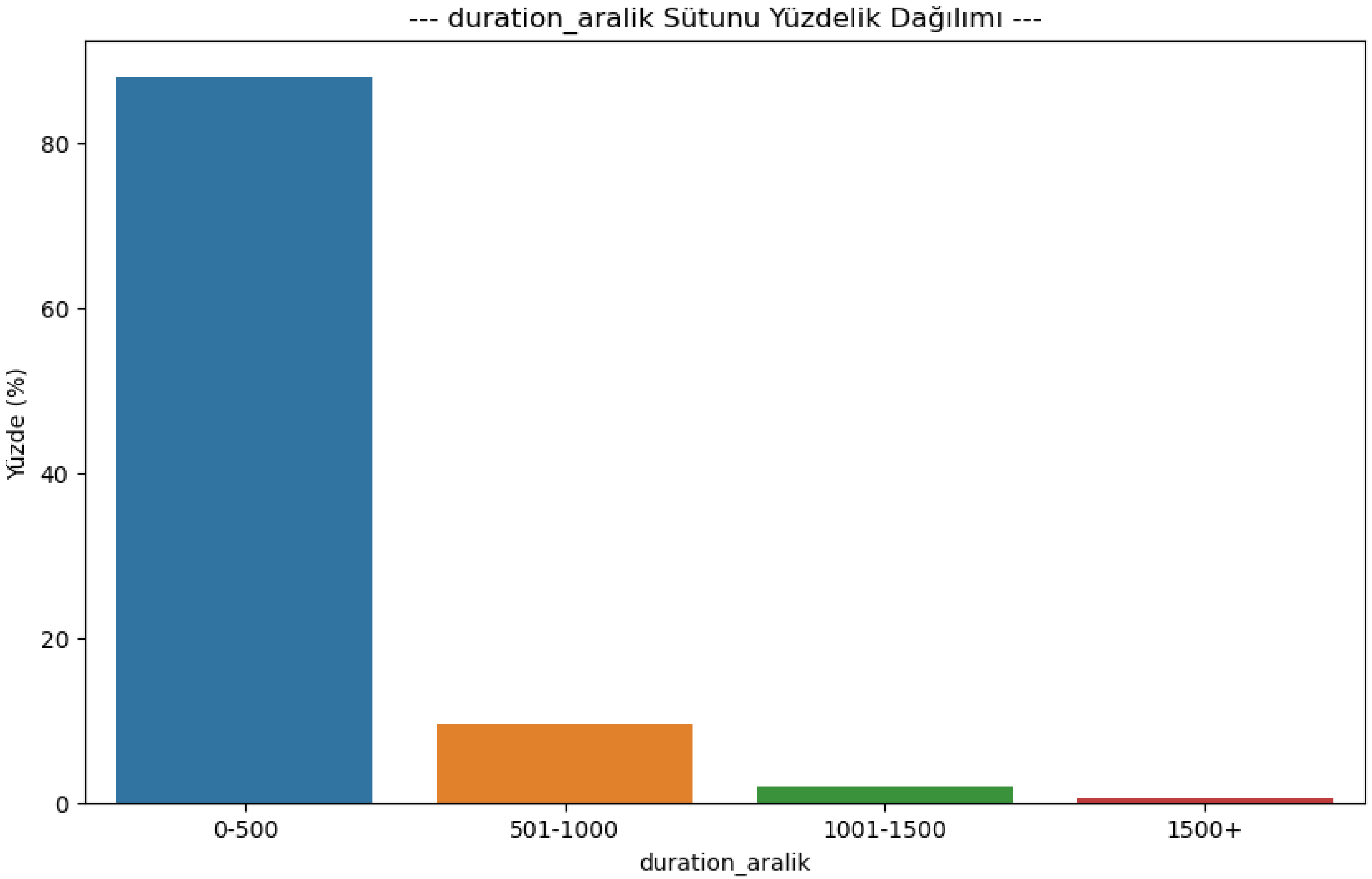
Sayısal olmayan değerler için One-Hot Encoding uyguladık.

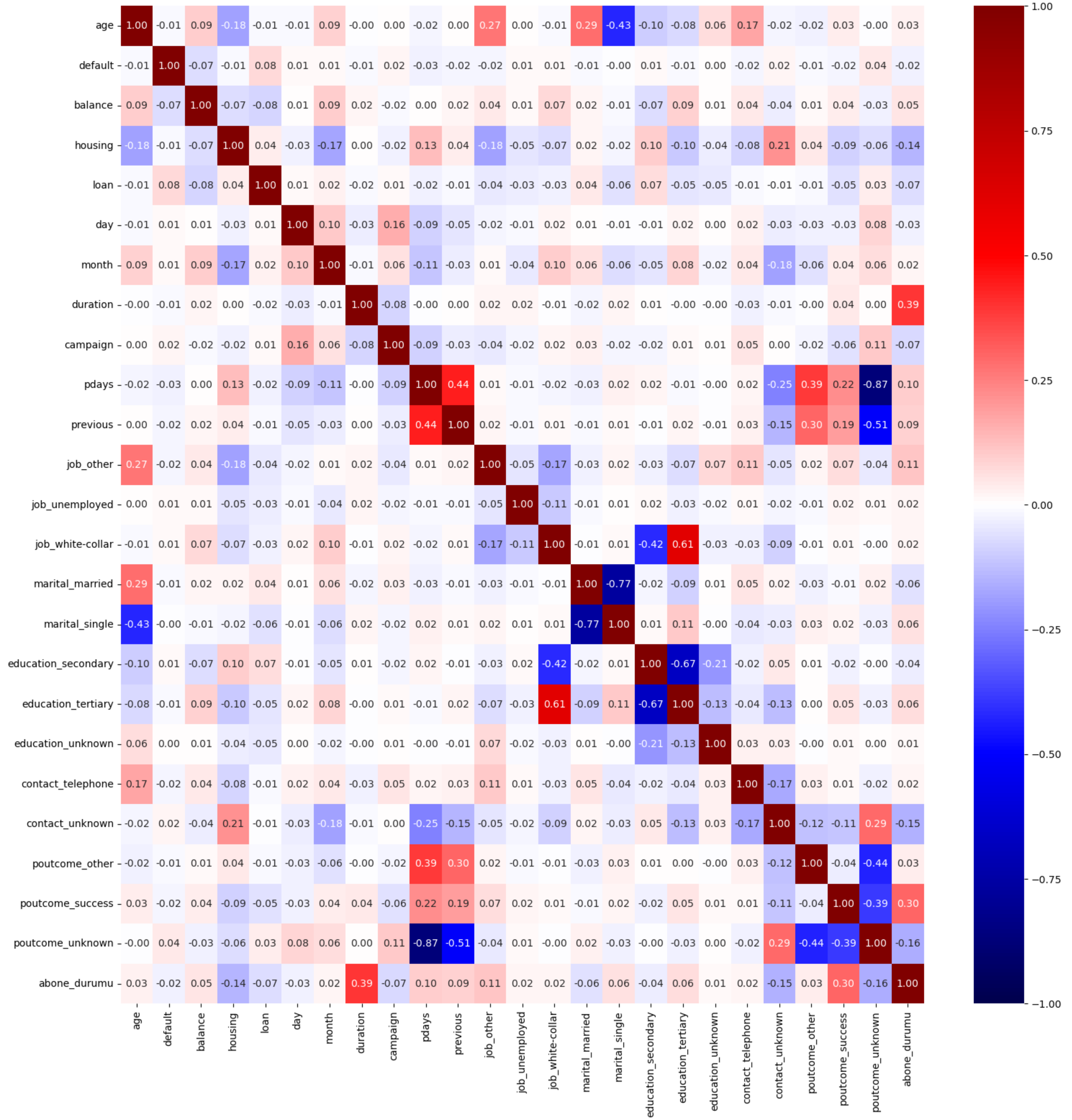
```
In [65]: df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 44923 entries, 0 to 45210
Data columns (total 25 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   abone_durumu         44923 non-null  int64
 1   age                  44923 non-null  int64
 2   default              44923 non-null  int64
 3   balance              44923 non-null  int64
 4   housing              44923 non-null  int64
 5   loan                 44923 non-null  int64
 6   day                  44923 non-null  int64
 7   month                44923 non-null  int32
 8   duration             44923 non-null  int64
 9   campaign             44923 non-null  int64
 10  pdays                44923 non-null  int64
 11  previous             44923 non-null  int64
 12  job_other            44923 non-null  int32
 13  job_unemployed       44923 non-null  int32
 14  job_white-collar     44923 non-null  int32
 15  marital_married      44923 non-null  int32
 16  marital_single       44923 non-null  int32
 17  education_secondary  44923 non-null  int32
 18  education_tertiary   44923 non-null  int32
 19  education_unknown    44923 non-null  int32
 20  contact_telephone    44923 non-null  int32
 21  contact_unknown      44923 non-null  int32
 22  poutcome_other       44923 non-null  int32
 23  poutcome_success     44923 non-null  int32
 24  poutcome_unknown     44923 non-null  int32
dtypes: int32(14), int64(11)
memory usage: 6.5 MB
```



Değerlerin yüzdelik dağılımı incelendi.

Dengesiz veri setine sahip olduğumuz gözlemlendi.

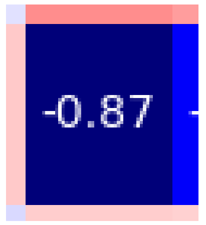




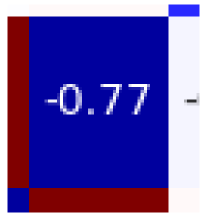
Değişkenler arasındaki korelasyonu ve ilişkiyi anlamak için heatmap çizdirdik.

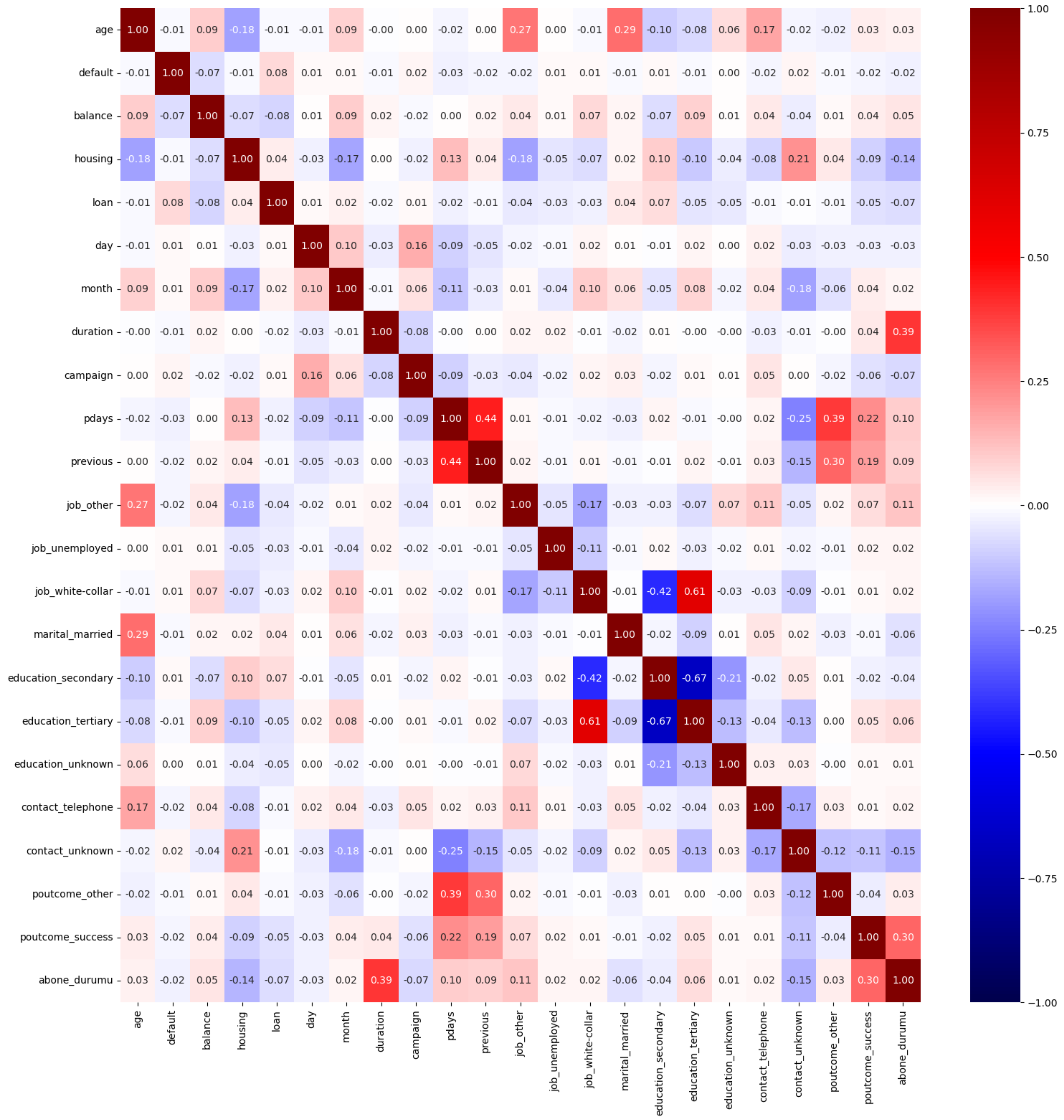
Multicollinearity!

Poutcome_unknown – pdays



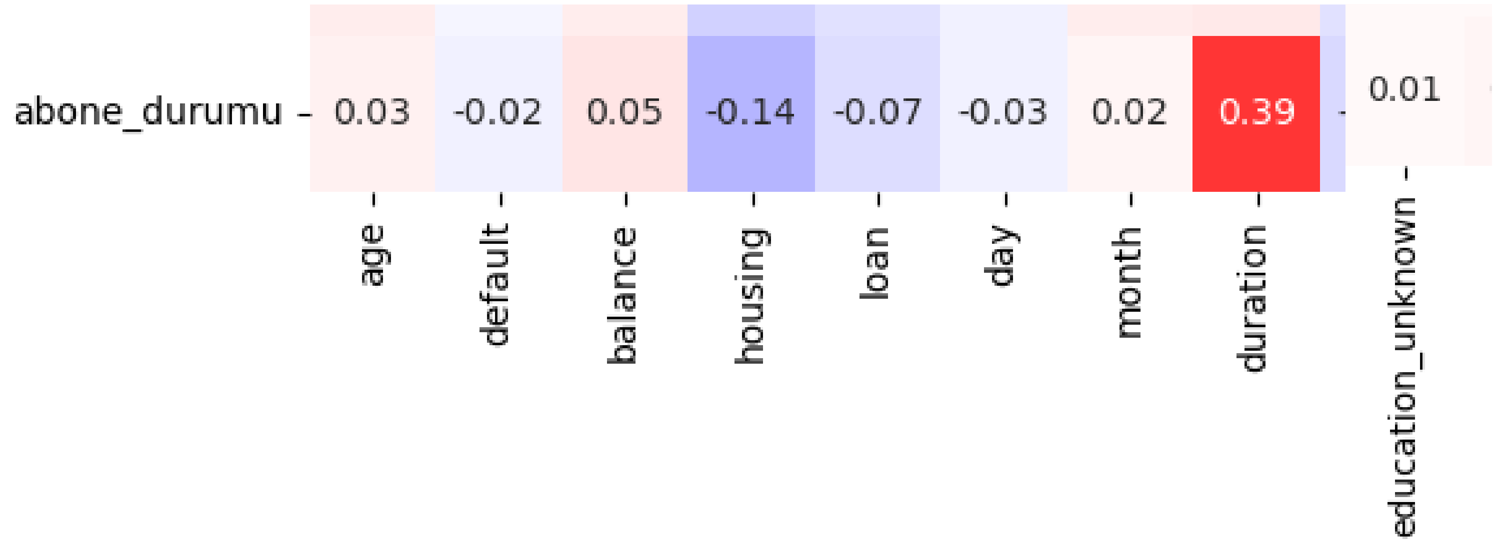
Marital_single / marital_married



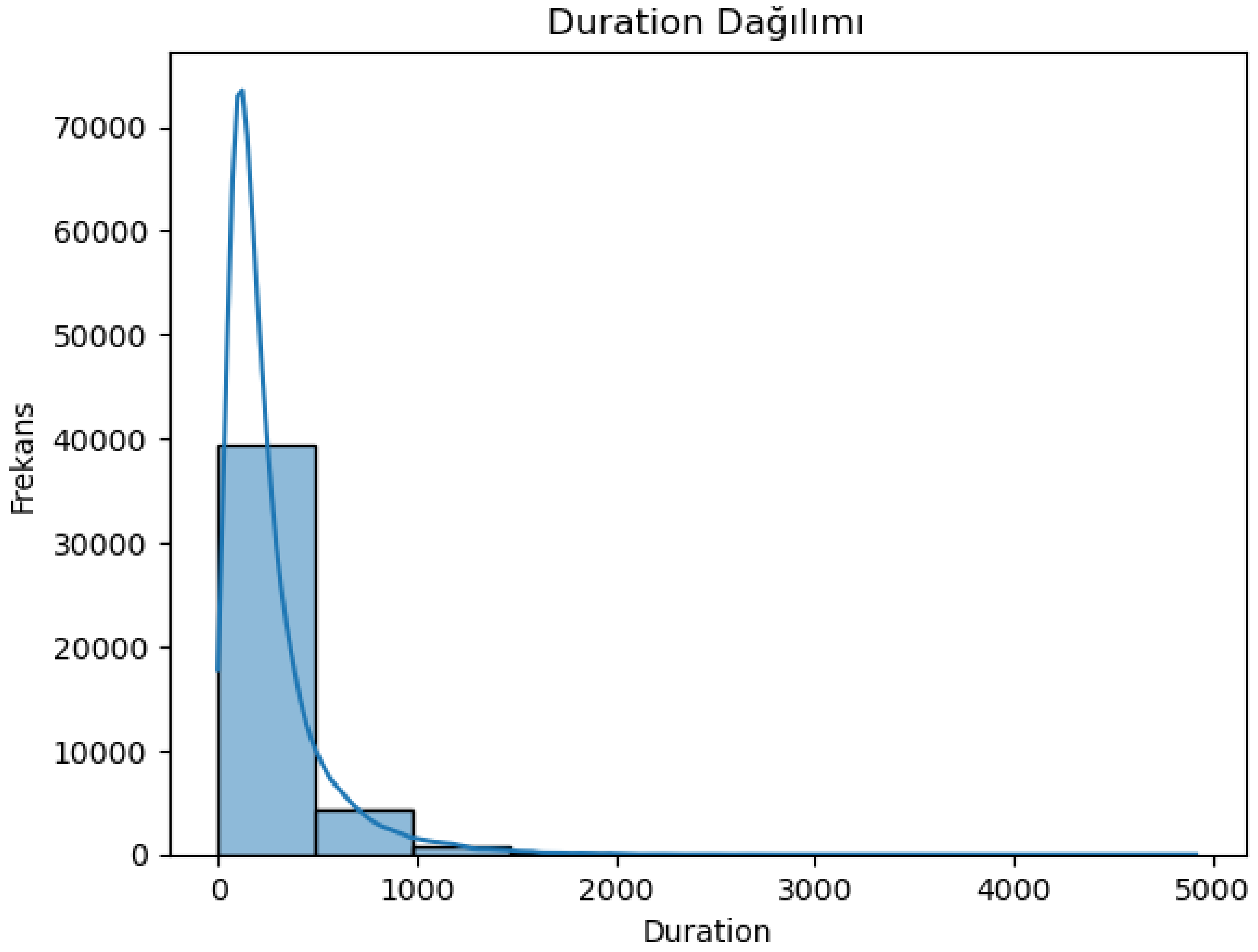


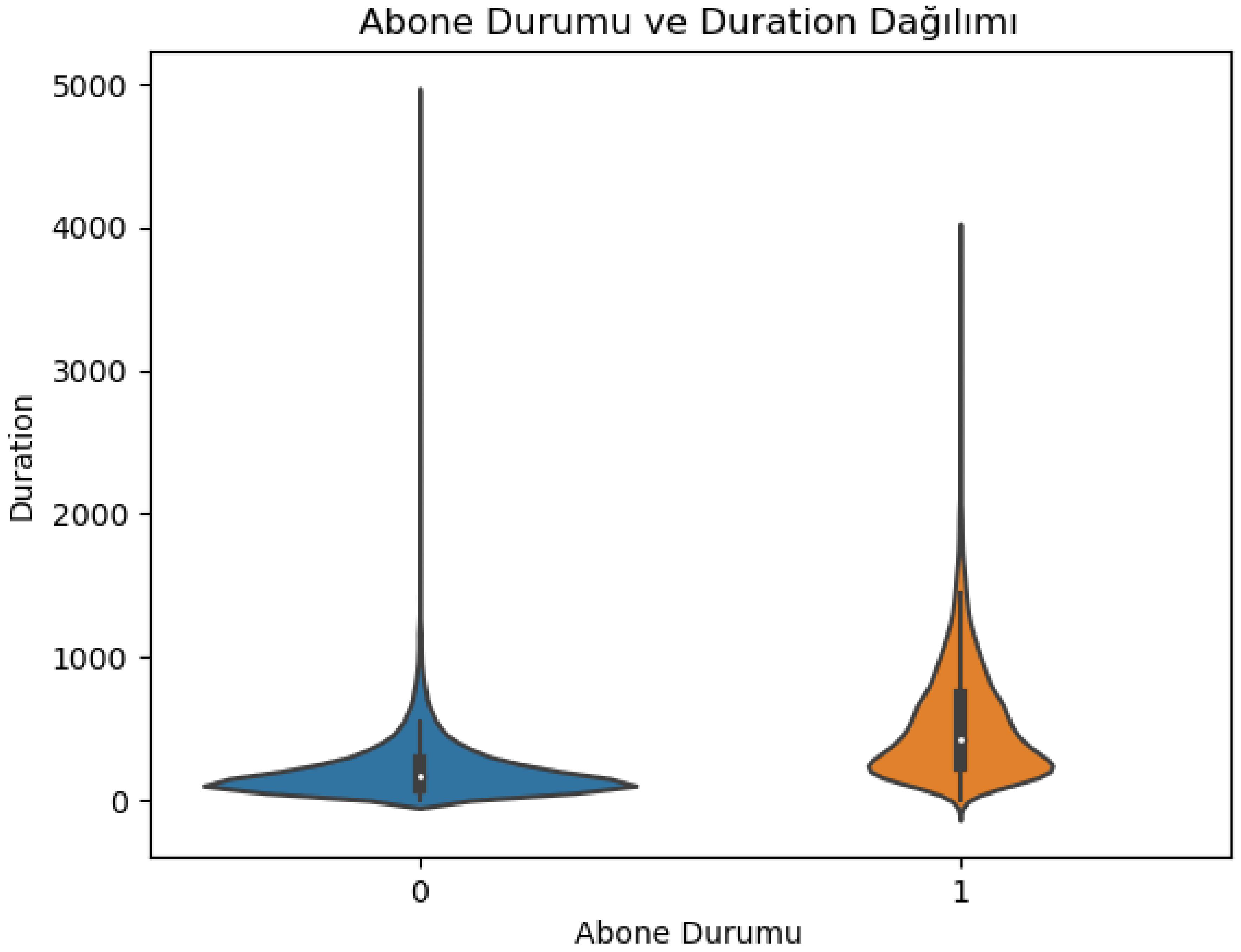
Heatmap

İnceleme Sonuçları;



Duration özelliğinin dağılımı incelendi.





MODEL EĞİTME
TEST ETME

Out[18]:

Model	Cross Validated Score (Mean)	Cross Validated Score (Std)
XGBoost	90.56	0.16
Random Forest	90.47	0.14
Logistic Regression	89.99	0.26
SVC	89.82	0.19
KNN	89.35	0.18
Decision Tree	87.53	0.22
Perceptron	84.09	1.04

Accuracy – standart sapma

Random Forest

```
In [20]: #En iyi sonucu Random Forest döndürüyor gibi standart sapması en düşük.
         rf = RandomForestClassifier()

         rf.fit( x_train_scaled, y_train)
         rf.score(x_test_scaled, y_test)

Out[20]: 0.9052865887590429
```

```
In [23]: evaluate_model(rf, x_test_scaled, y_test, pred_label=0)
```

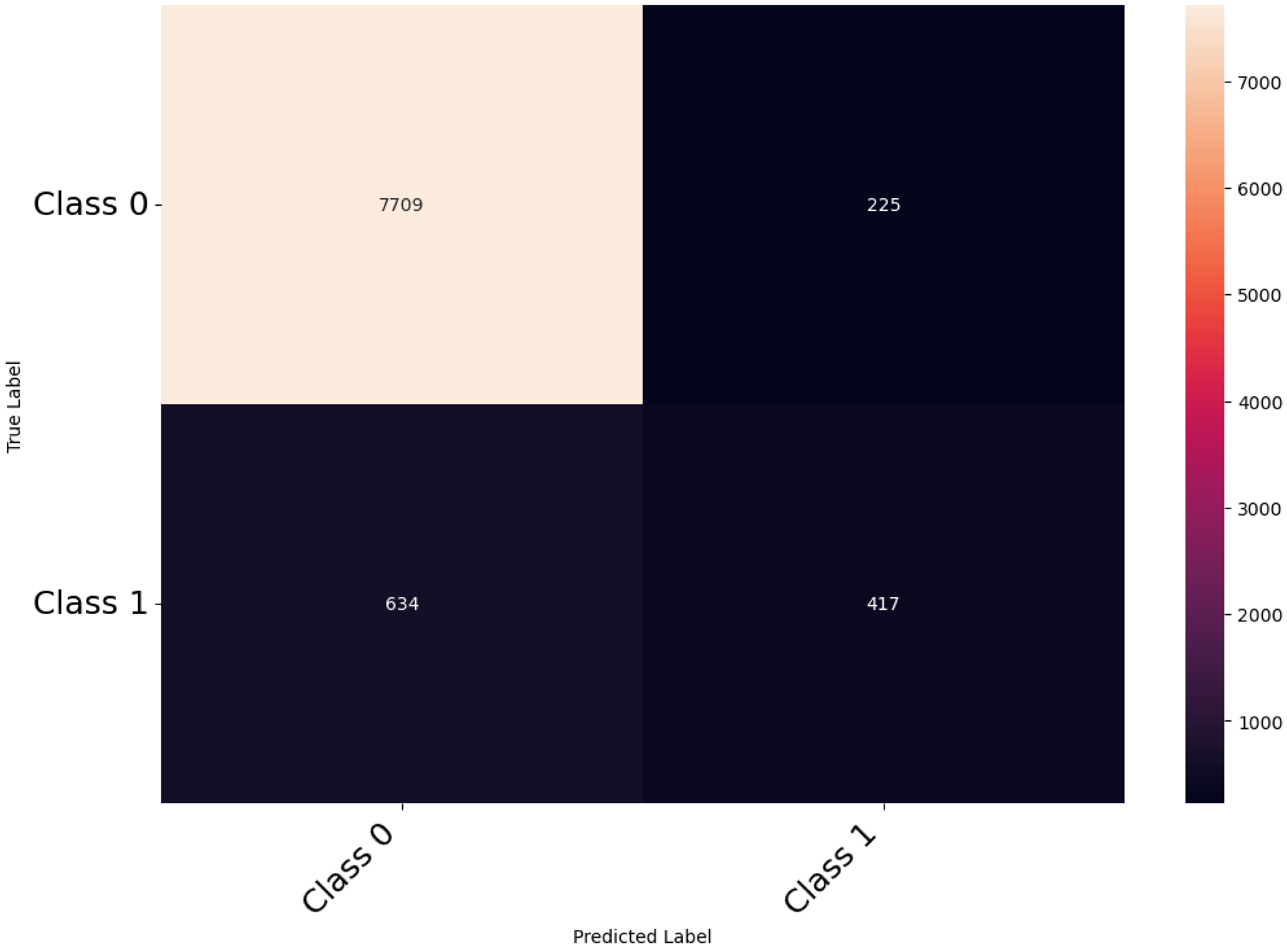
Out[23]:

	Accuracy	Recall	Precision	F1 Score	AUC Score
Class 0	0.905	0.972	0.924	0.948	0.686

```
In [24]: #Henüz dengelenmemiş veri setinde amacımız 1 (abone olma) durumunu tahmin etmek ve 1 sınıfında RECALL %40'larda düşük.
         #Bir de AUC Score değerinin 1'e yakın olması önemli şu an düşük derecede. Accuracy bu veri setinde gerçek başarı değil.
         evaluate_model(rf, x_test_scaled, y_test, pred_label=1)
```

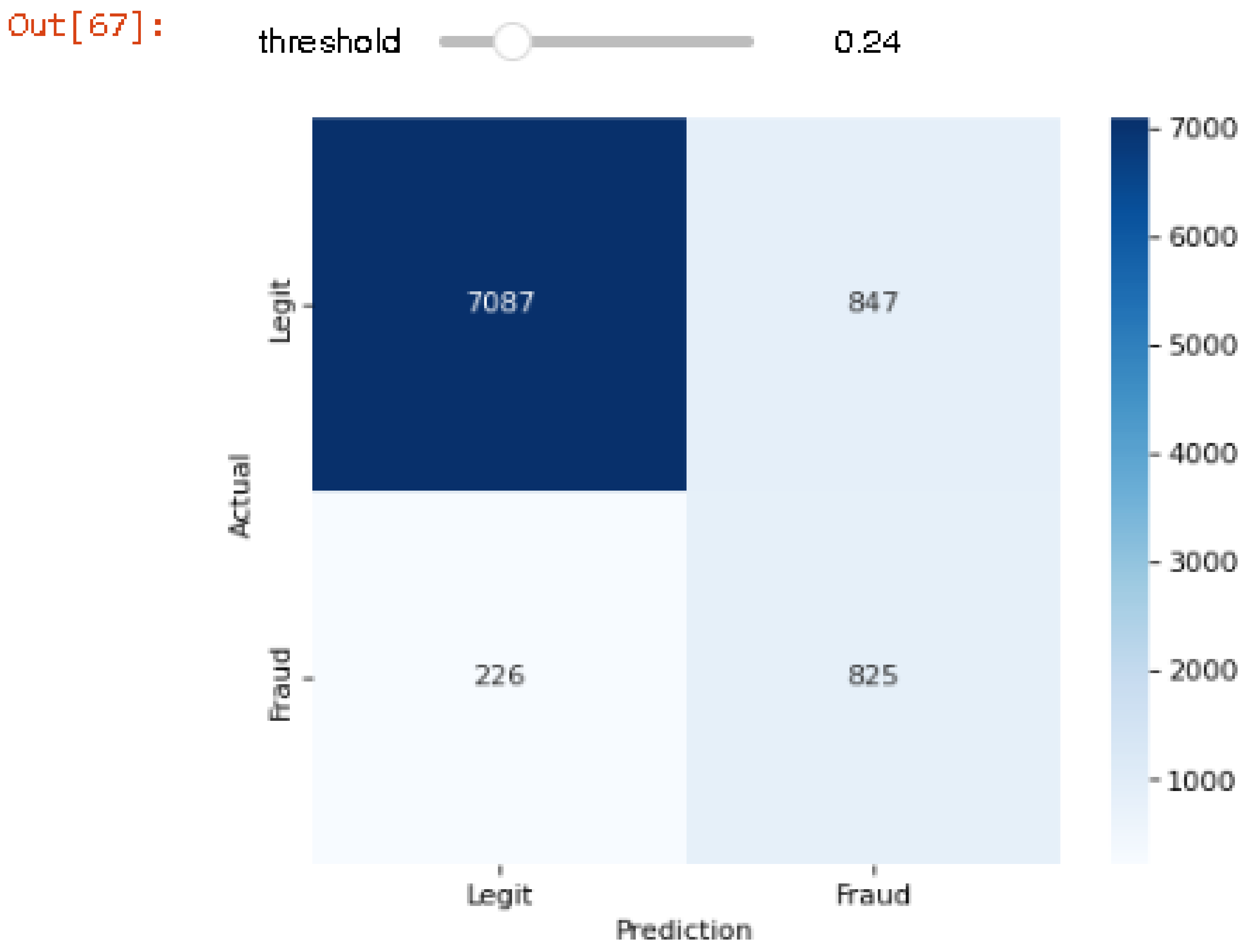
Out[24]:

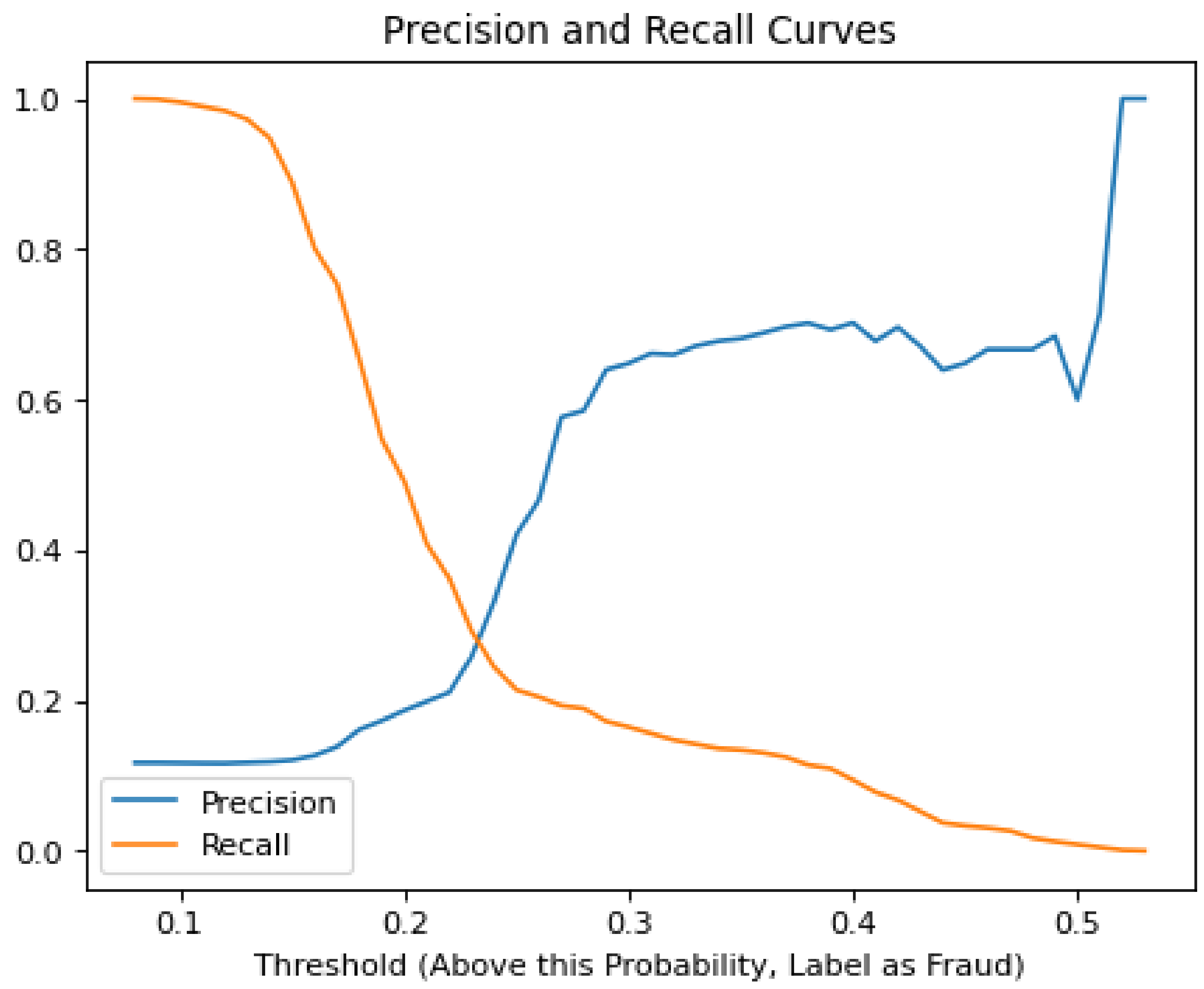
	Accuracy	Recall	Precision	F1 Score	AUC Score
Class 1	0.905	0.401	0.656	0.497	0.686



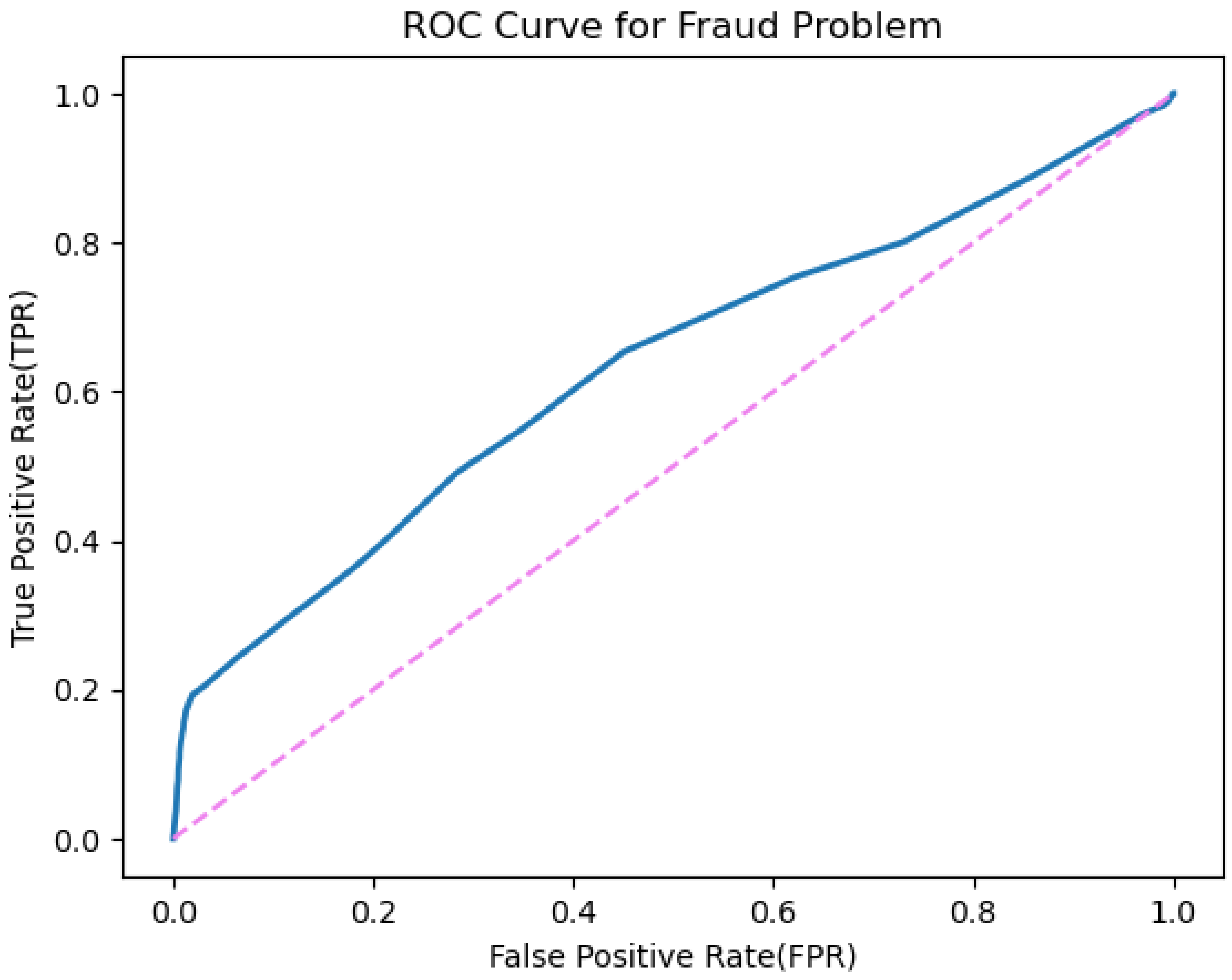
7709
225
634
417
Class 0
Class 1
True Label
Class 0
Class 1
Predicted Label
7000
6000
5000
4000
3000
2000
1000

Threshold değeri için en iyi aralık incelemesi yapıldı.





Precision, recall ve ROC AUC Score metrikleri görselde incelendi.



Out[56]:

Feature	Importance
duration	0.292
balance	0.115
age	0.107
day	0.099
month	0.083
poutcome_success	0.055
pdays	0.049
campaign	0.042
previous	0.026
housing	0.023
marital_married	0.015
contact_unknown	0.014
job_white-collar	0.013
education_secondary	0.012
education_tertiary	0.011
loan	0.010
job_other	0.010
contact_telephone	0.008
job_unemployed	0.005
education_unknown	0.005
poutcome_other	0.004
default	0.002

duration!

Feature importance

Duration (saniye cinsinden son görüşme süresi) **ön plana çıktı.**

Random Forest

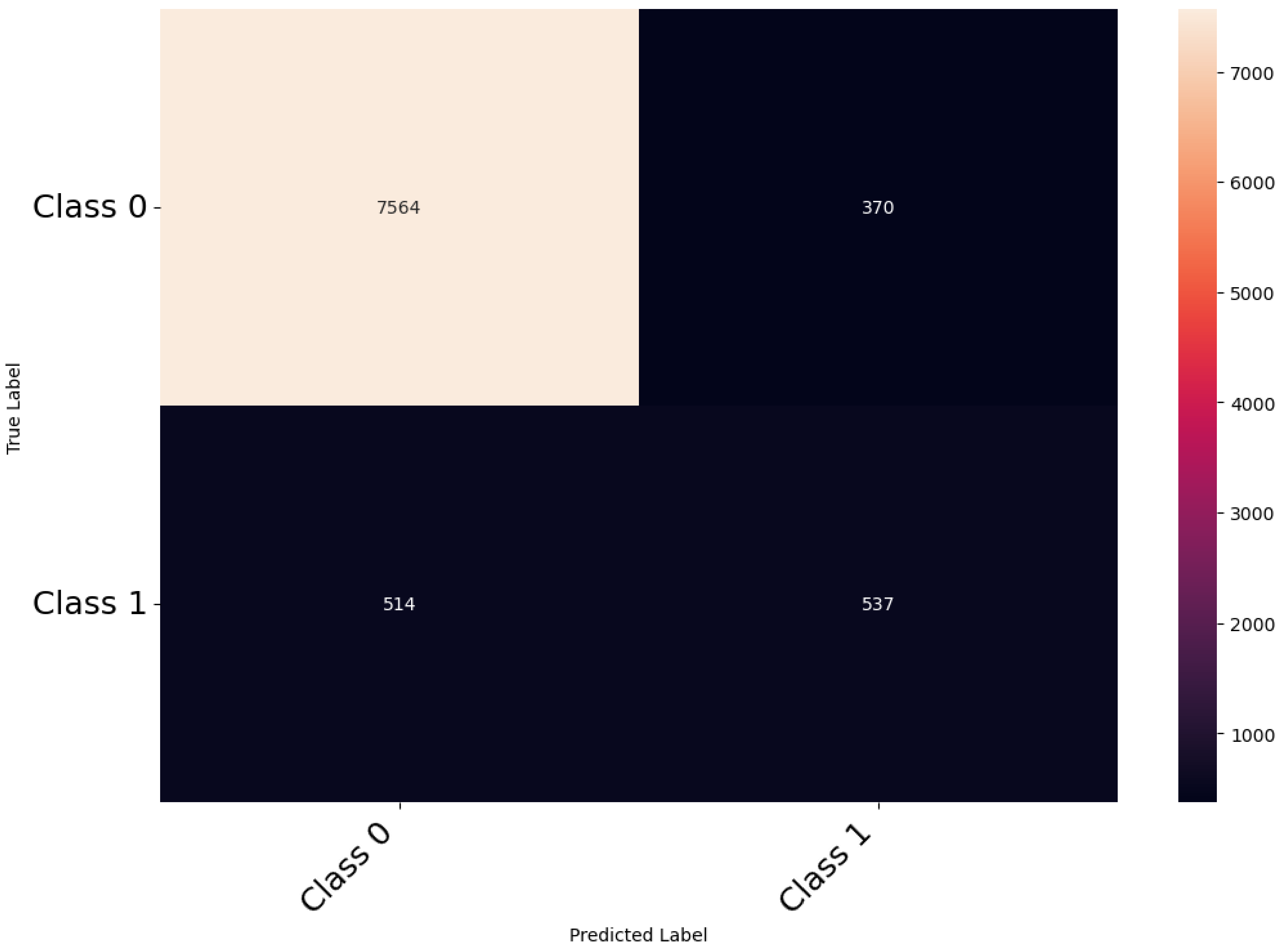
```
: print("Training Accuracy:", round(lm1.score(X_train[['duration']].values, y_train), 3))
  print("Testing Accuracy:", round(lm1.score(X_test[['duration']].values, y_test), 3))

Training Accuracy: 0.9
Testing Accuracy: 0.882
```

DENGESİZ VERİ KÜMESİ EĞİTME - TESTETME

Random Oversampling

Gerçekte Abone Olma Durumu

417’den 537’ye artmış

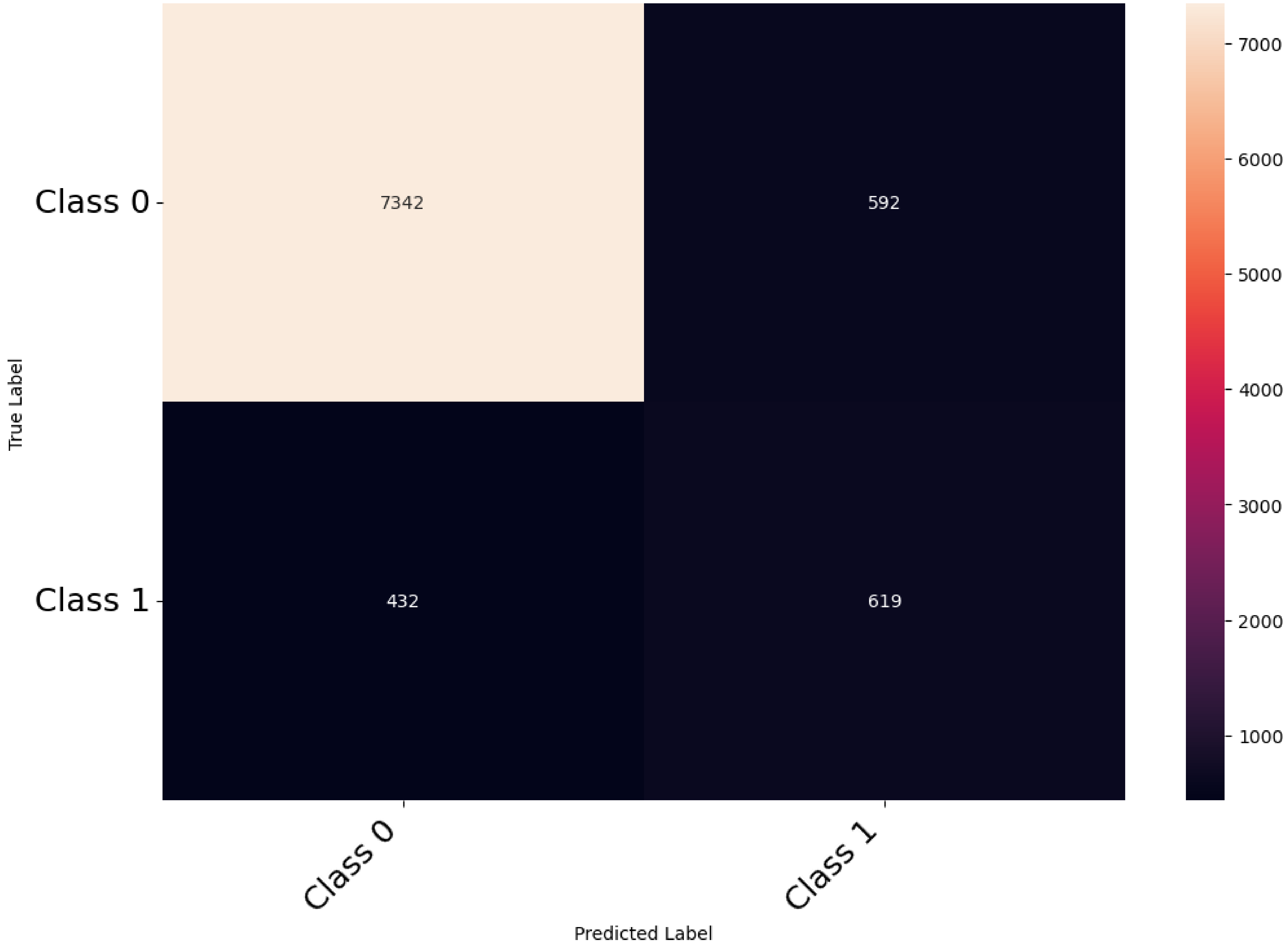


SMOTE Oversampling

Gerçekte Abone Olma Durumu

417’den 619’a artmış

Random, BORDERLINE ve ADASYN’e oranla daha iyi



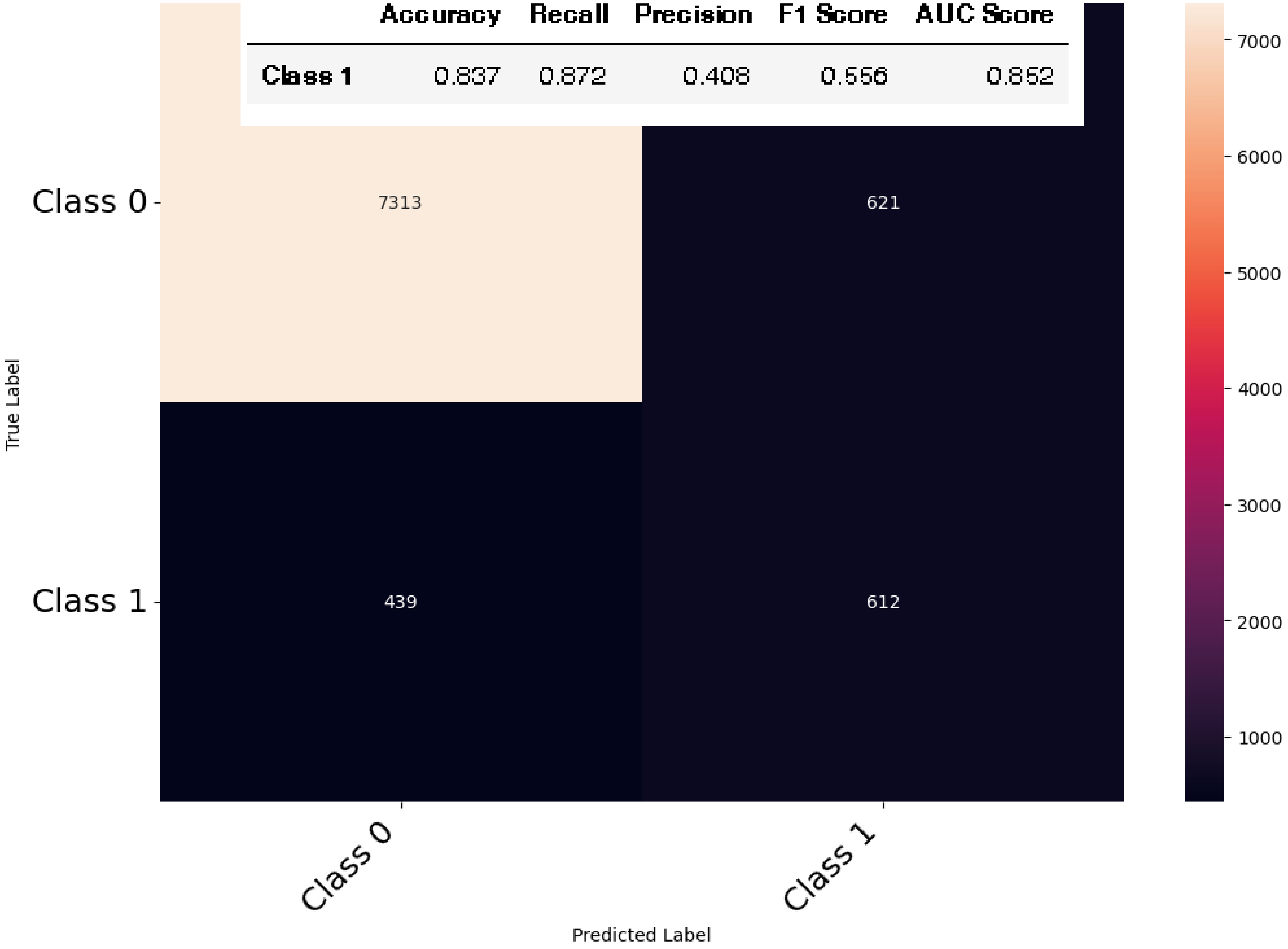
Adasyn Oversampling

Gerçekte Abone Olma Durumu

417’den 612’ye artmış

Random ve BORDERLINE’a oranla daha iyi

	Accuracy	Recall	Precision	F1 Score	AUC Score
Class 1	0.837	0.872	0.408	0.556	0.852

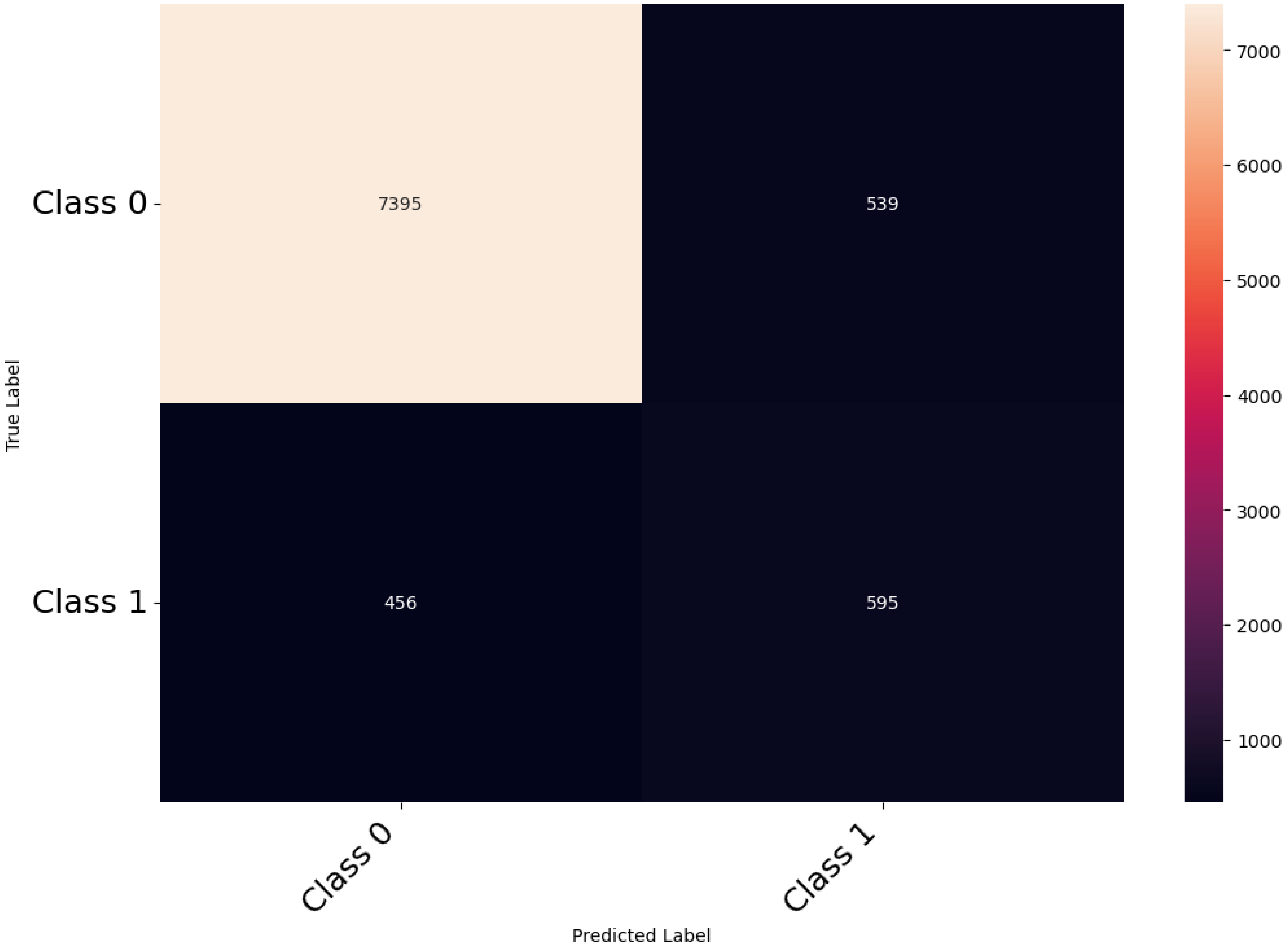


BORDERLINE Oversampling

Gerçekte Abone Olma Durumu

417’den 595’e artmış

Random’a oranla daha iyi

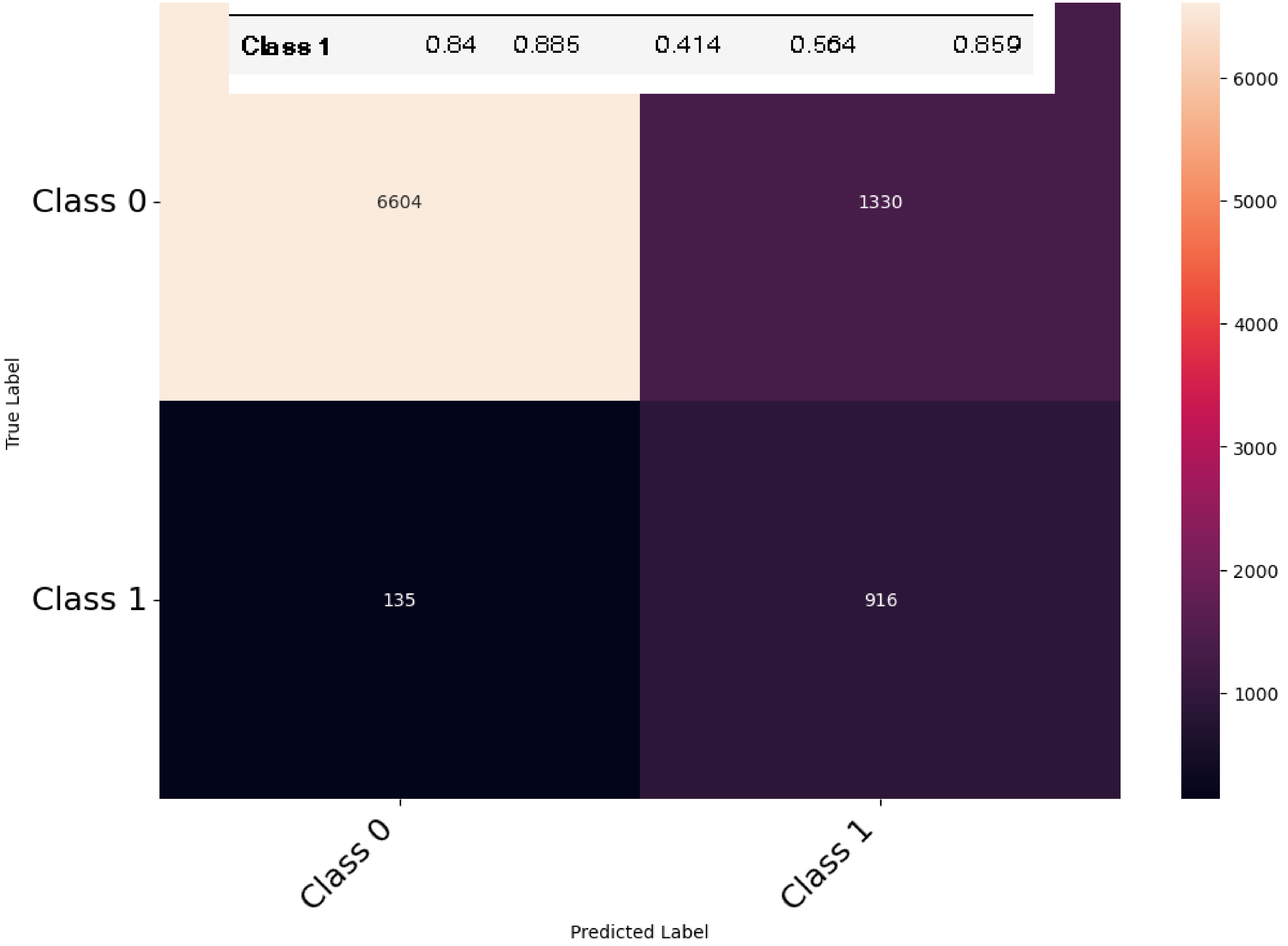


Random Undersampling

Gerçekte Abone Olma Durumu

417’den 916’ya artmış çok iyi ama sentetik veriler üretti.

	Accuracy	Recall	Precision	F1 Score	AUC Score
Class 1	0.84	0.885	0.414	0.564	0.859



Undersampling ve Oversampling model test denemeleri sonucunda Oversampling – Adasyn –Random Forest algoritması ile devam ettik.

Yalnızca duration’ı kullandık.

Ek olarak denenen özellikler

XBalance

XHousing

MODEL TEST KARŞILAŞTIRMA

Out[59]:

	abone_durumu	duration
0	0	261
1	0	151
2	0	76
3	0	92
4	0	139
...	...	...
44918	1	977
44919	1	456
44920	1	1127
44921	0	508
44922	0	361

MODEL TEST KARŞILAŞTIRMA

Dağılım Dengesizken

In [40]:
```
lm1.predict([[76],[456],[508],[977],[1127]])
```

Out[40]:
```
array([0, 0, 0, 1, 1], dtype=int64)
```

In [41]:
```
lm1.predict_proba([[76],[456],[508],[977],[1127]])
```

Out[41]:
```
array([[0.99421254, 0.00578746],
       [0.78600088, 0.21399912],
       [0.83277099, 0.16722901],
       [0.41088889, 0.58911111],
       [0.05004762, 0.94995238]])
```

Dağılım Dengeliyken

In [66]:
```
lm1.predict([[76],[456],[508],[977],[1127]])
```

Out[66]:
```
array([0, 1, 1, 1, 1], dtype=int64)
```

In [67]:
```
lm1.predict_proba([[76],[456],[508],[977],[1127]])
```

Out[67]:
```
array([[0.91520791, 0.08479209],
       [0.36147196, 0.63852804],
       [0.3705631 , 0.6294369 ],
       [0.1137455 , 0.8862545 ],
       [0.        , 1.        ]])
```

STREAMLIT

VERİ SETİ

- **age**: Müşterinin yaşı.
- **default**: Kredi temerrüt durumu (0 = Hayır, 1 = Evet).
- **balance**: Ortalama yıllık bakiye.
- **housing**: Konut kredisi varlığı (0 = Hayır, 1 = Evet).
- **loan**: Kişisel kredi varlığı (0 = Hayır, 1 = Evet).
- **day**: İletişim günü (ayın günü).
- **month**: İletişim ayı.
- **duration**: Son iletişim süresi (saniye cinsinden).
- **campaign**: Bu kampanya sırasında yapılan iletişim sayısı.
- **pdays**: Önceki kampanyadan bu yana geçen gün sayısı (numeric; -1 = müşteri daha önce temas edilmedi).
- **previous**: Bu kampanya öncesinde yapılan iletişim sayısı.

4. Modelin Çalışma Prensibi

Modelimiz, yukarıda belirtilen özellikleri kullanarak müşteri davranışını tahmin eder. Bu model, `RandomForestClassifier` kullanılarak eğitilmiştir ve müşteri verileri ile tahmin yapar:

- **Tahmin**: Müşterinin vadeli mevduat ürünü abone olup olmayacağını (subscribed veya failed) tahmin eder.
- **Olasılıklar**: Müşterinin abone olma veya olmama olasılıklarını yüzdelik olarak verir.

Choose the features below to see the result!

Age

25,00

Default

1

Balance

20,00

Housing

1

Loan

1

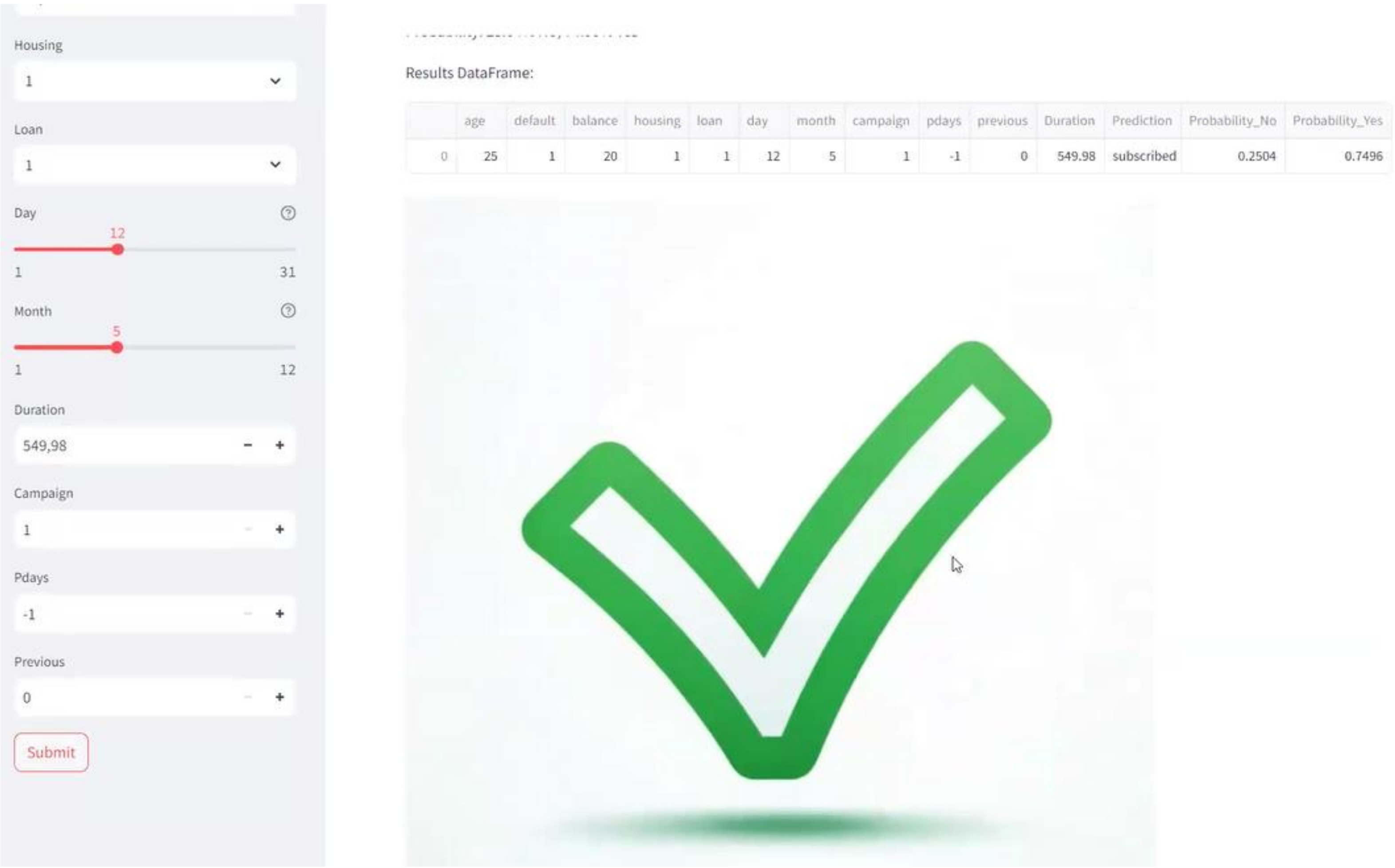
Day

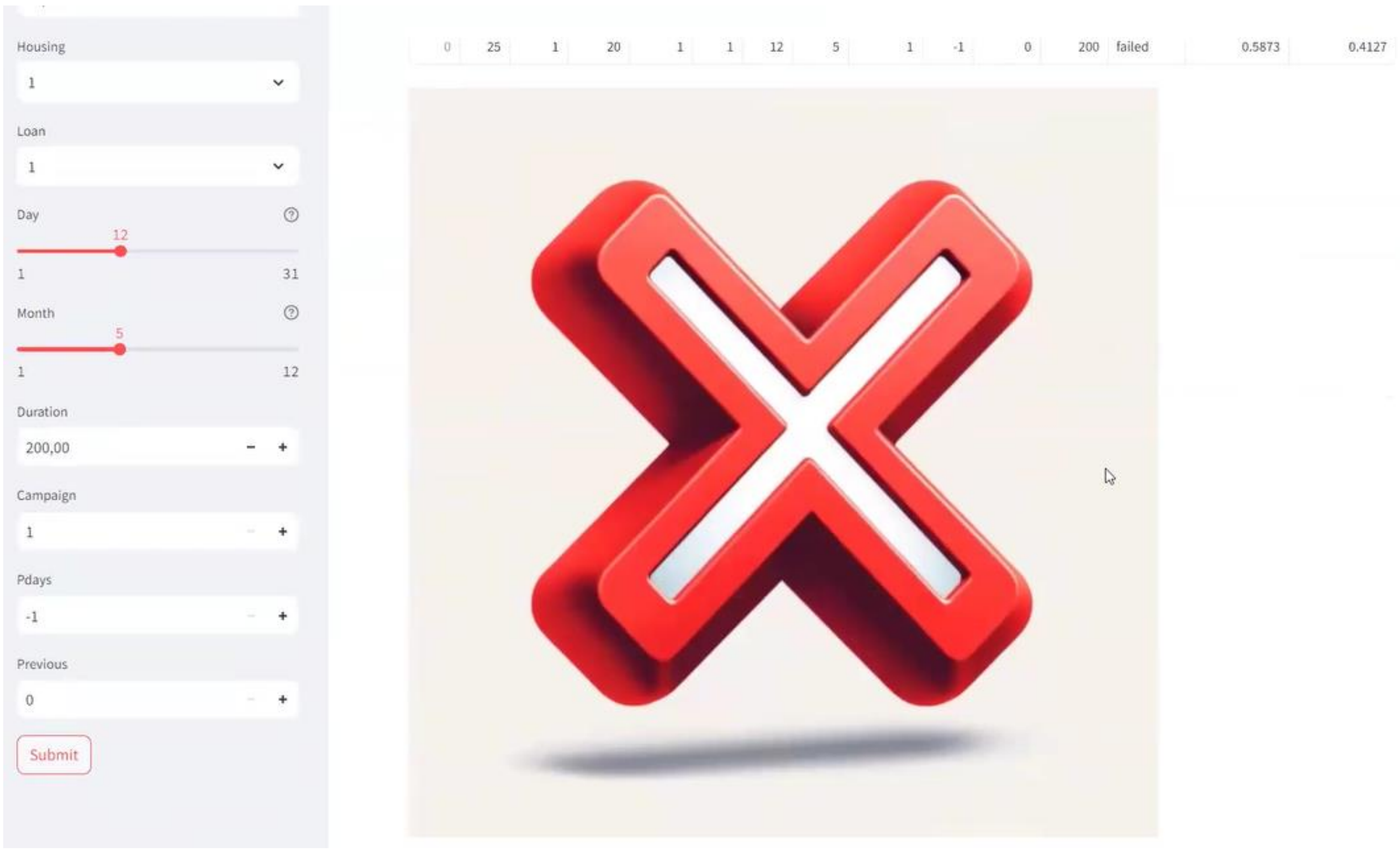
12

Goal: Predict if the client will subscribe a term deposit.

Banking Marketing Project

Web sitesinde değerler girilerek denendi.





Teşekkür Ederiz

Batuhan Yıldız

Betül Uyar Can