**BURSA TEKNİK ÜNİVERSİTESİ**

**Selçuk Şan**

**18360859049**

**BANK MARKETING VERİ SETİ**

**SINIFLANDIRMA (CLASSIFICATION) PROJESİ**

**VERİ MADENCİLİĞİ DERSİ PROJESİ**

**06, 2022**

**İÇİNDEKİLER**

**Sayfa**

[KISALTMALAR iii](#_Toc105879244)

[GİRİŞ 4](#_Toc105879245)

[1. PROJE SÜRECİ VE KULLANILAN METODOLOJİ 4](#_Toc105879246)

[1.1 Keşifçi Veri Analizi 4](#_Toc105879247)

[1.2 Veri Ön İşleme 7](#_Toc105879248)

[1.2.1 Eksik Değer Analizi 7](#_Toc105879249)

[1.2.2 Aykırı Değer Analizi 7](#_Toc105879250)

[1.2.3 Under Sampling 8](#_Toc105879251)

[1.2.4 Discretization 9](#_Toc105879252)

[1.2.5 Label Encoding 9](#_Toc105879253)

[1.2.6 One Hot Encoding 9](#_Toc105879254)

[2. MODELLEME 10](#_Toc105879255)

[2.1 Model Seçimi 10](#_Toc105879256)

[2.1.1 Random Forest (Rassal Orman) 10](#_Toc105879257)

[2.2 Model Validation (Model Doğrulama) 11](#_Toc105879258)

[2.2.1 Holdout Yöntemi 11](#_Toc105879259)

[2.2.2 K-Katlı Çapraz Doğrulama (K Fold Cross Validation) 11](#_Toc105879260)

[2.2.3 Bootstrap Yöntemi 11](#_Toc105879261)

[2.2.4 Leave One Out Yöntemi 11](#_Toc105879262)

[2.3 Model Başarı Değerlendirme Metrikleri 12](#_Toc105879263)

[2.4 Hiperparametre Optimizasyonu 13](#_Toc105879264)

[2.4.1 GridSearchCV 13](#_Toc105879265)

[2.4.2 RandomizedSearchCV 13](#_Toc105879266)

[3. MODEL SONUÇLARI 13](#_Toc105879267)

[3.1 Base Model Performansı 13](#_Toc105879268)

[3.2 Hiperparametre Optimizasyonu Sonrası Model Performansı 13](#_Toc105879269)

[3.3 Model Sonuçlarının Başka Çalışmalarla Karşılaştırılması 14](#_Toc105879270)

[3.3.1 KNN, Lineer Regresyon ve Logistik Regresyon Modelleri ile Karşılaştırma 14](#_Toc105879271)

[3.3.2 Naive Bayes Modeli İle Karşılaştırma 14](#_Toc105879272)

[3.4 Model Değişken Önem Düzeyleri (Feature Importances) 14](#_Toc105879273)

[4. SONUÇ 15](#_Toc105879274)

[KAYNAKLAR 16](#_Toc105879275)

KISALTMALAR

**TP : True Pozitive**

**FP : False Pozitive**

**TN : True Negative**

**FN : False Negative**

GİRİŞ

Projede UCI Machine Learning Repository sitesinden alınan Bank Marketing veri seti analiz edildi. Veriler, bir Portekiz bankacılık kurumunun doğrudan pazarlama kampanyalarıyla ilgilidir.

Pazarlama kampanyaları telefon görüşmelerine dayanmaktadır. Çoğu zaman, müşterinin ürüne (banka vadeli mevduat) abone olup olmayacağına ('yes') veya abone olmayacağına ('no') erişmek için aynı müşteriyle birden fazla iletişim gerekmektedir.

Mevcutta dört veri kümesi bulunmaktadır. Bu veri kümesinin eski sürümü için tüm örnekleri içeren bank-full.csv veri kümesi seçildi.

# PROJE SÜRECİ VE KULLANILAN METODOLOJİ

Bu çalışmada veri madenciliği yöntemleri kullanıldı. Veri madenciliği, büyük miktarda veriden anlamlı bilgiler çıkarmaktır. Veriler, bilgiyi bulmak için işlendiklerinde anlam kazanır.

Günümüzde büyük miktarda veriyi işleyebilen veri madenciliği tekniklerinin kullanılması daha önemli hale gelmektedir. Anlamlı bilgiyi bulmak veri madenciliği ile mümkündür. Veri madenciliği, verilerden bilinmeyen anlamlı bilgiler çıkarmak için betimleyici ve tahmin edici teknikleri içerir.

Bank Marketing verilerini analiz etmek için Python programlama dili ve frameworkleri kullanıldı.

Python açık kaynak kodlu, interpreted, yüksek seviyeli ve nesne yönelimli bir programlama dilidir. Veri Madenciliği ve Makine öğrenmesi uygulamaları için kullanılan en iyi dillerden birisidir.

## Keşifçi Veri Analizi

Veri seti, 10 kategorik ve 7 sayısal değişken olmak üzere 17 değişken 45211 gözlem içermektedir. Sadece iki değeri olanlar ikili değerler olarak sınıflandırılmaktadır.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 1 : Bank marketing veri setinin değişkenleri ve açıklamaları

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 2 : Değişkenlerin tip bilgileri.

metin, bilgisayar, dosya içeren bir resim

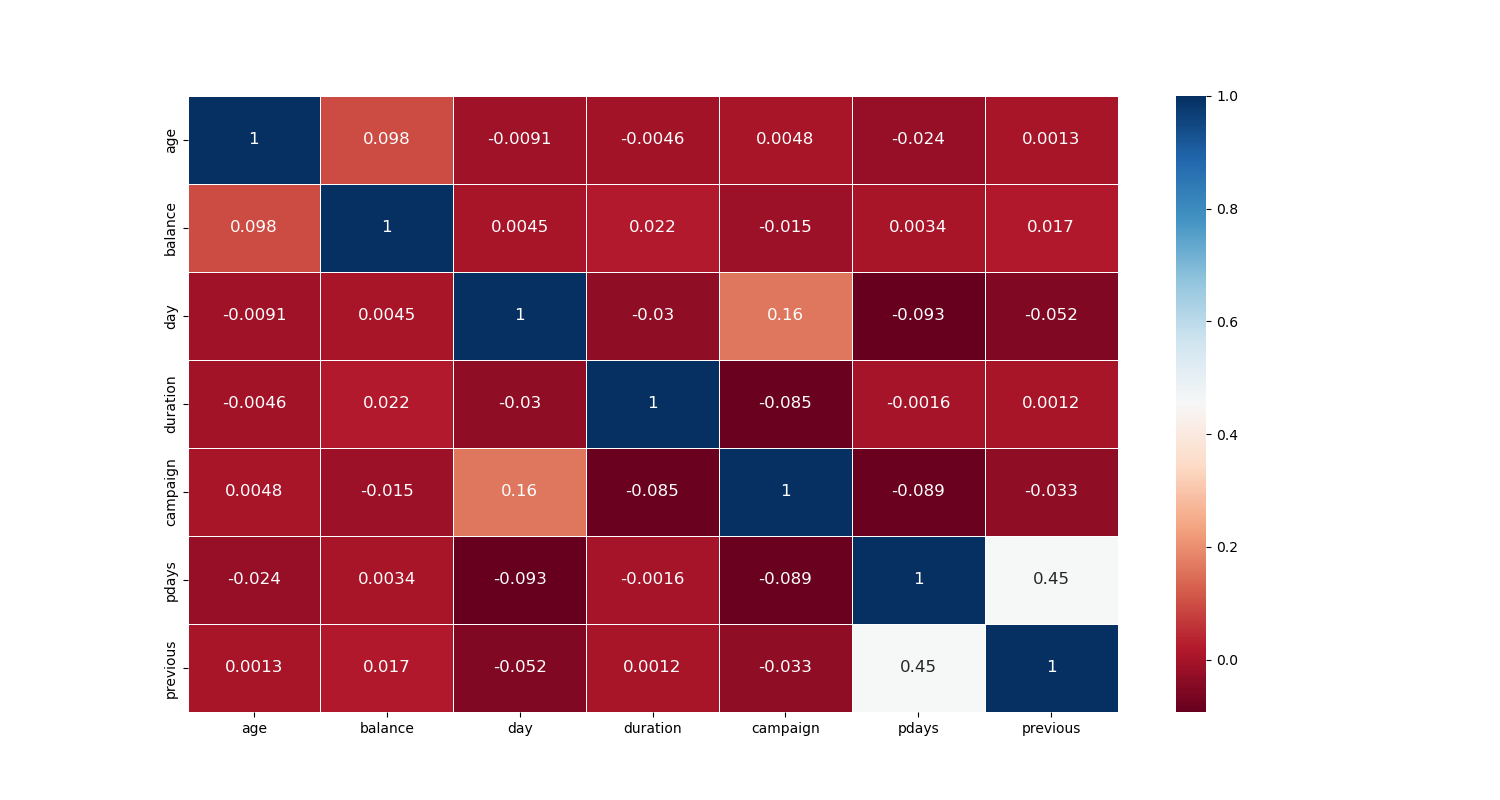
Açıklama otomatik olarak oluşturuldu

Şekil 3 : Veri setinin ilk 10 gözlem değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 4 : Sayısal değişkenlerin betimsel istatistikleri



Şekil 5 : Sayısal değişkenler arasındaki korelasyon değerlerinin ısı haritası ile gösterilmesi.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 6 : Eksik değerlerin incelenmesi

## Veri Ön İşleme

Bir veri seti doğru formatta olsa bile, veri madenciliği algoritmalarını uygulayabilmek ve analiz sonucunun kalitesini artırmak için ön işlemeye ihtiyaç duyulur. Birçok veri ön işleme tekniği vardır.

Bu çalışmada kullanılan teknikler eksik değer analizi, aykırı değerleri baskılama, under sampling, encoding ve discretization işlemleridir.

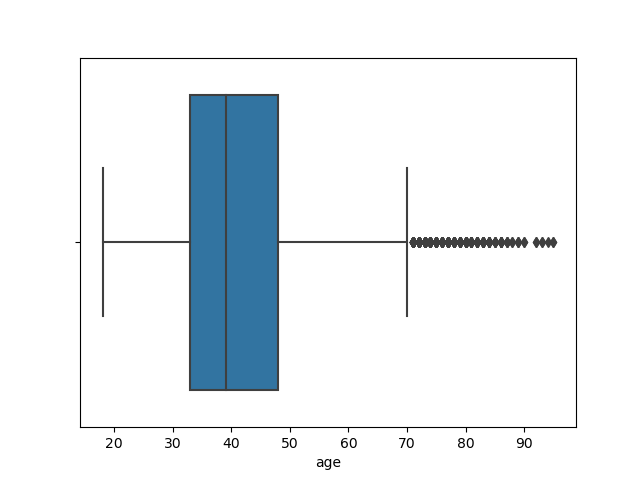
Eksik değer olmadığı için herhangi bir silme veya doldurma işlemi gerçekleştirilmemiştir. Yalnızca aykırı değerler baskılama yöntemi ile çözülmüştür.

### Eksik Değer Analizi

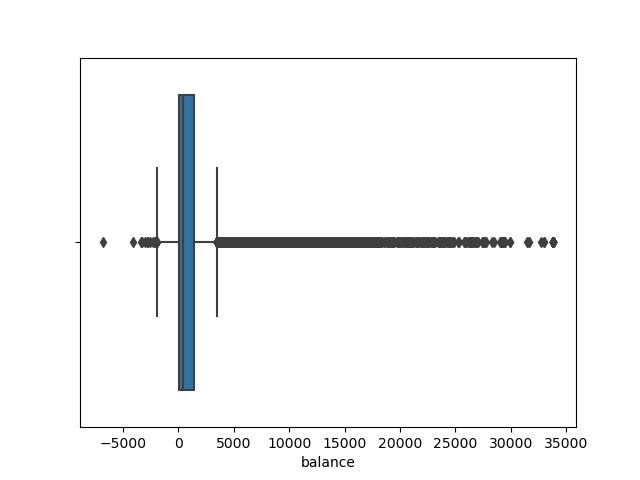
Veri setinde eksik değer bulunmamaktadır.

### Aykırı Değer Analizi

Aykırı değer (outlier), istatistik analizi sürecinde verilerde bulunan aşırı uç değerlerden oluşan gözlemleri temsil etmektedir. Verilerin genel yapısına göre aşırı ölçüde yüksek veya düşük gözlemler aykırı değer olarak tanımlanabilir.



Şekil 7 : Age değişkeninin aykırı değerlerinin boxplot ile gösterilmesi.

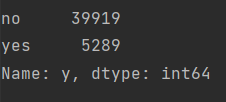


Şekil 8 : Balance değişkeninin aykırı değerlerinin boxplot ile gösterilmesi.

### Under Sampling

Eksik örnekleme (Under Sampling), sınıf dağılımları eşit olana kadar çoğunluk olan sınıfının örneklerini ortadan kaldırarak veri kümesini yeniden dengelemeyi amaçlar. Bu yöntemin en büyük dezavantajı, yetersiz gözlem sayısı olan projelerde katkı sağlayabilecek gözlemleri de veri setinden kaldırmasıdır.

Ayrıca, az sayıda gözlemin olduğu durumunda örnek uzayın rastgeleliği zarar görebilir.



Şekil 9 : Under Sampling işlemi öncesi hedef değişkenin sınıf dağılımı.

metin, cihaz, metre içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 11 : Under Sampling işlemi sonrası hedef değişkenin sınıf dağılımı.

### Discretization

Veri ayrıklaştırma (Discretization), çok sayıda veri değerini daha küçük değerlere dönüştürerek verilerin değerlendirilmesini ve yönetimini kolaylaştıran bir yöntemi ifade eder. Başka bir deyişle, veri ayrıklaştırma, sürekli verilerin nitelik değerlerini minimum veri kaybıyla sonlu bir aralık kümesine dönüştürme yöntemidir.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 10 : Discretization işlemi sonrası oluşan yeni değişkenin değerleri.

### Label Encoding

Label Encoding, kategorik değişkenleri işlemek için popüler bir kodlama tekniğidir. Bu teknikte, her bir veriye alfabetik sıralamaya göre benzersiz bir tam sayı atanır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 12 : Label Encoding işlemi sonrası hedef değişkeninin değerleri.

### One Hot Encoding

One-Hot Encoding, esasen kategorik değişkenlerin ikili vektörler olarak temsilidir. Bu kategorik değerler ilk olarak tamsayı değerlere eşlenir. Her bir tamsayı değeri daha sonra tamamı 0 olan bir ikili vektör olarak temsil edilir.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 13 : One Hot Encoding işlemi sonrası değişkenlerin isimleri ve değerleri.

# MODELLEME

## Model Seçimi

Temelde bir yöntem ön plana çıkmaktadır.

* Oluşabilecek değişken kombinasyonları ile oluşturulan modeller arasında en iyi modeli seçimek.

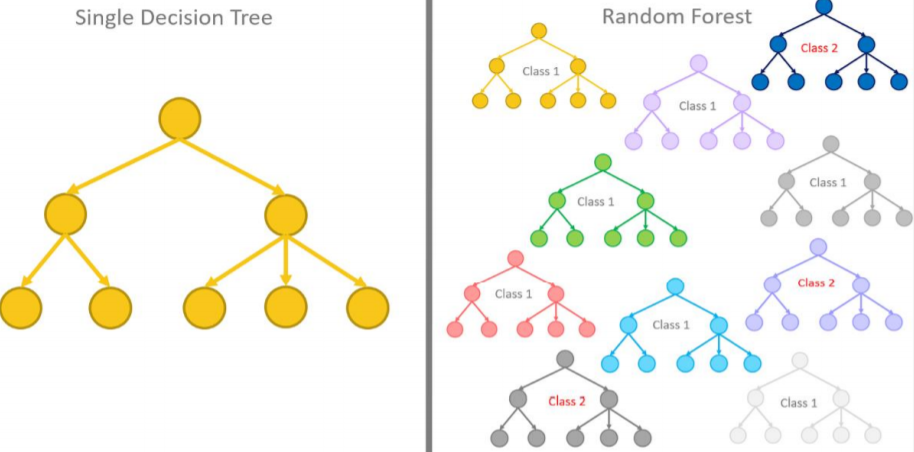
### Random Forest (Rassal Orman)

Random forest algoritması, denetimli sınıflandırma algoritmalarından biridir. Algoritma, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler.

Random forest algoritması birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir. Aşağıdaki görsel bu olayı net bir şekilde ifade etmektedir.

Ağaç sayısı arttıkça kesin bir sonuç elde etme oranı artmaktadır.

Karar ağaçları algoritması ile arasındaki temel fark, random forest algoritmasında kök düğümü bulma ve düğümleri bölme işleminin rastgele olmasıdır.



Şekil 14 : Random Forest ile Karar Ağacı algoritmasının farkı.

Random forest, elinde yeterli miktarda ağaç varsa aşırı öğrenme sorununu azaltır. Az oranda bir veri hazırlığına ihtiyaç duyar.

## Model Validation (Model Doğrulama)

Model doğrulama, modellerin beklendiği gibi performans gösterdiğini doğrulamayı amaçlayan süreç ve faaliyetler kümesidir.

### Holdout Yöntemi

Hold-out, verisetini “eğitim” ve “test” kümesi olarak ikiye ayırma yöntemidir. Eğitim seti, modelin eğitildiği verilerken test seti, modelin eğitilmeyen veriler üzerinde ne kadar iyi performans gösterdiğini görmek için kullanılan verierldir.

### K-Katlı Çapraz Doğrulama (K Fold Cross Validation)

Cross-validation, makine öğrenmesi modelinin görmediği veriler üzerindeki performansını mümkün olduğunca objektif ve doğru bir şekilde değerlendirmek için kullanılan istatistiksel bir yeniden örnekleme(resampling) yöntemidir.

### Bootstrap Yöntemi

Yerine koymalı bir şekilde veri seti içerisinden veri türetmek için kullanılır ve oluşan yeni verilerin her birisi üzerinden model kurulur. Bu modeller test edilir ve buna göre sonuçlar değerlendirilir.

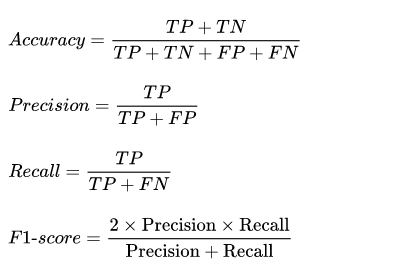
### Leave One Out Yöntemi

Jackknife yöntemi her aşamada bir obje ya da bireye ait veriyi dışarıda bırakarak aynı testin tekrar edilmesidir.

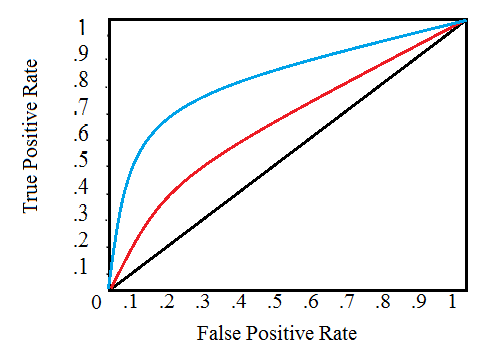
Bu nedenle bu yöntem birini dışarıda bırak (leave one out) yöntemi olarak da isimlendirilmektedir.

## Model Başarı Değerlendirme Metrikleri

Accuracy, Precision, Recall, F1 skor, ROC Curve ve AUC değerlerine dayalı olarak sınıflandırıcının performansını değerlendirilecektir.



Şekil 14 : Performans metriklerinin formülleri.



Şekil 15 : ROC Eğrisi.

* Accuracy bir modelin başarısını ölçmek için çok kullanılan ancak tek başına yeterli olmadığı görülen bir metriktir.
* Kesinlik (Precision) ise Positive olarak tahminlediğimiz değerlerin gerçekten kaç adedinin Positive olduğunu göstermektedir.
* Duyarlılık (Recall) ise Positive olarak tahmin etmemiz gereken işlemlerin ne kadarını Positive olarak tahmin ettiğimizi gösteren bir metriktir.
* F1 Score değeri bize Kesinlik (Precision) ve Duyarlılık (Recall) değerlerinin harmonik ortalamasını göstermektedir.
* Bir ROC eğrisi (receiver operating characteristic curve), tüm sınıflandırma eşiklerinde bir sınıflandırma modelinin performansını gösteren bir grafiktir.
* ROC bir olasılık eğrisidir ve altında kalan alan olan AUC ayrılabilirliğin derecesini veya ölçüsünü temsil eder.

## Hiperparametre Optimizasyonu

Hiperparametre optimizasyonu, bir makine öğrenmesi algoritması için belirlenen başarı metriğine göre en uygun hiperparametre kombinasyonunu bulma işlemidir.

* Hiperparametre optimizasyonu ile model karmaşıklığı dengelenerek overfitting ve underfitting dengesi sağlanabilir.
* Yine modelin esnekliğinden kaynaklanan aşırı öğrenme problemi hiperparametreler ile getirilen kısıtlamalarla çözülebilir.

### GridSearchCV

Modelde denenmesi istenen hiperparametreler ve değerleri için bütün kombinasyonlar ile ayrı ayrı model kurulur ve belirtilen metriğe göre en başarılı hiperparametre seti belirlenir.

### RandomizedSearchCV

Rastgele olarak bir hiperparametre seti seçilir ve cross-validation ile model kurularak test edilir.

Belirlenen hesaplama süresi limitine ya da iterasyon sayısına ulaşıncaya kadar bu adımlar devam eder.

# MODEL SONUÇLARI

## Base Model Performansı

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metrik | Accuracy | Recall | Precision | F1 Score | AUC Score |
| Score | 0.7181 | 0.6564 | 0.749 | 0.6994 | 0.775 |

## Hiperparametre Optimizasyonu Sonrası Model Performansı

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metrik | Accuracy | Recall | Precision | F1 Score | AUC Score |
| Score | 0.7405 | 0.6451 | 0.7975 | 0.7132 | 0.7997 |

## Model Sonuçlarının Başka Çalışmalarla Karşılaştırılması

### KNN, Lineer Regresyon ve Logistik Regresyon Modelleri ile Karşılaştırma

“Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data” isimli makalede aşağıdaki tablodaki değerler gözlenmiştir.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

### Naive Bayes Modeli İle Karşılaştırma

“Data Analysis of a Portuguese Marketing Campaign using Bank Marketing data Set” isimli makalede aşağıdaki tablodaki değerler gözlenmiştir.

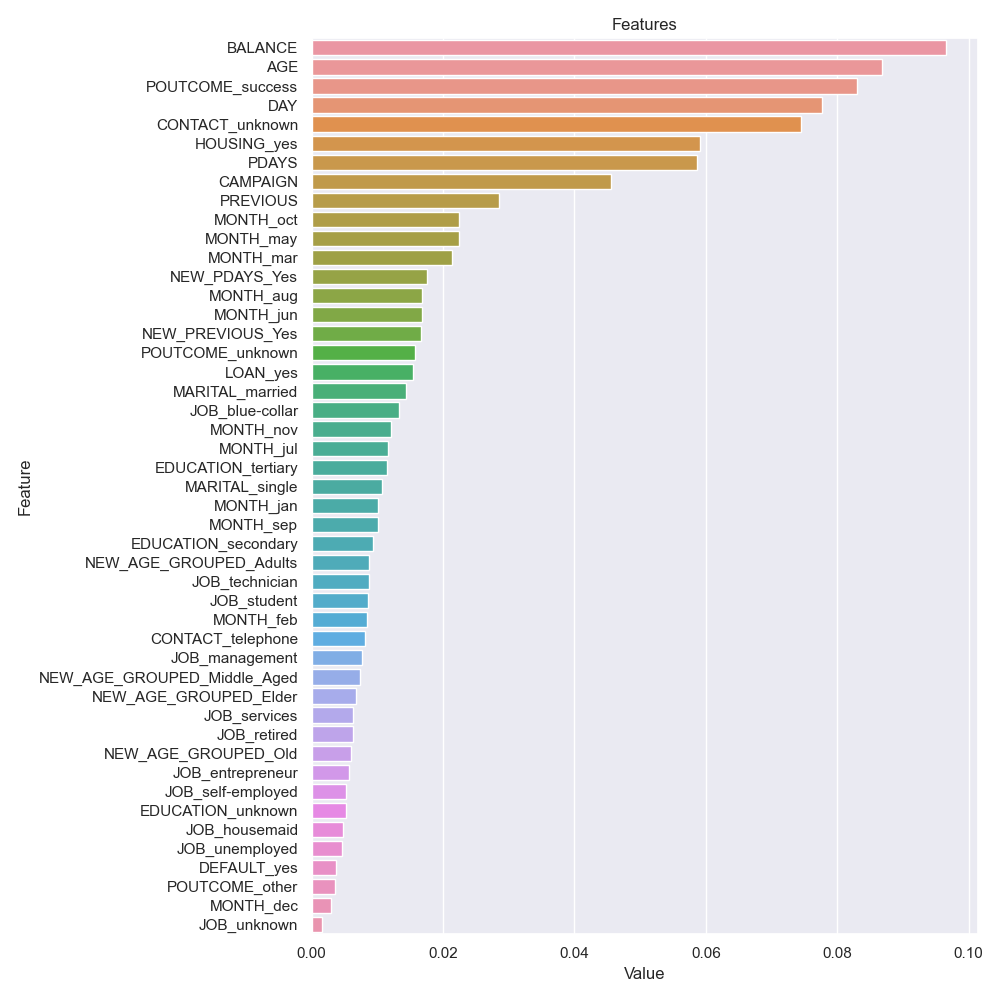
**P**: Precision **R**: Recall **F**: F1 Score

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

## Model Değişken Önem Düzeyleri (Feature Importances)

Model için değişkenlerin önem düzeyleri bar grafiği çoktan aza doğru sıralanmıştır.



Şekil 16 : Değişkenlerin model için önem düzeyleri.

# SONUÇ

Ekonomik koşullar ticari kuruluşları ve bankacılık sektörlerini etkilemektedir. Buna göre pazarlama yöneticilerinin pazarlama kampanyalarını artırma ihtiyacı vardır. En zorlu süreçler, büyük miktarda veri kaydı nedeniyle ortaya çıkar.

Veri madenciliği, bu büyük miktardaki veriyi analiz etmede kilit bir role sahiptir. Yöneticiler, veri madenciliği araçlarını kullanarak işlerini ve kampanya stratejilerini yeniden şekillendirebilir.

Bu projede, Bank Marketing veri seti üzerinde veri analizi, ön işleme ve modelleme adımları gerçekleştirerek veri setini en iyi temsil edecek bir model oluşturduk.

KAYNAKLAR

* [**https://medium.com/analytics-vidhya/classification-in-machine-learning-ed30753d9461**](https://medium.com/analytics-vidhya/classification-in-machine-learning-ed30753d9461)
* [**https://medium.datadriveninvestor.com/classification-in-machine-learning-db33514c77ad**](https://medium.datadriveninvestor.com/classification-in-machine-learning-db33514c77ad)
* [**https://towardsdatascience.com/8-metrics-to-measure-classification-performance-984d9d7fd7aa**](https://towardsdatascience.com/8-metrics-to-measure-classification-performance-984d9d7fd7aa)
* [**https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007**](https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007)
* [**https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-models-e2f0d8009d69**](https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-models-e2f0d8009d69)
* [**https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b**](https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b)
* [**https://towardsdatascience.com/understanding-random-forest-58381e0602d2**](https://towardsdatascience.com/understanding-random-forest-58381e0602d2)
* [**https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-random-forest-algoritmas%C4%B1-a79b044bbb31#**](https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-random-forest-algoritmas%C4%B1-a79b044bbb31)
* **https://medium.com/data-science-tr/makine-%C3%B6%C4%9Frenmesi-dersleri-5-bagging-ve-random-forest-2f803cf21e07**