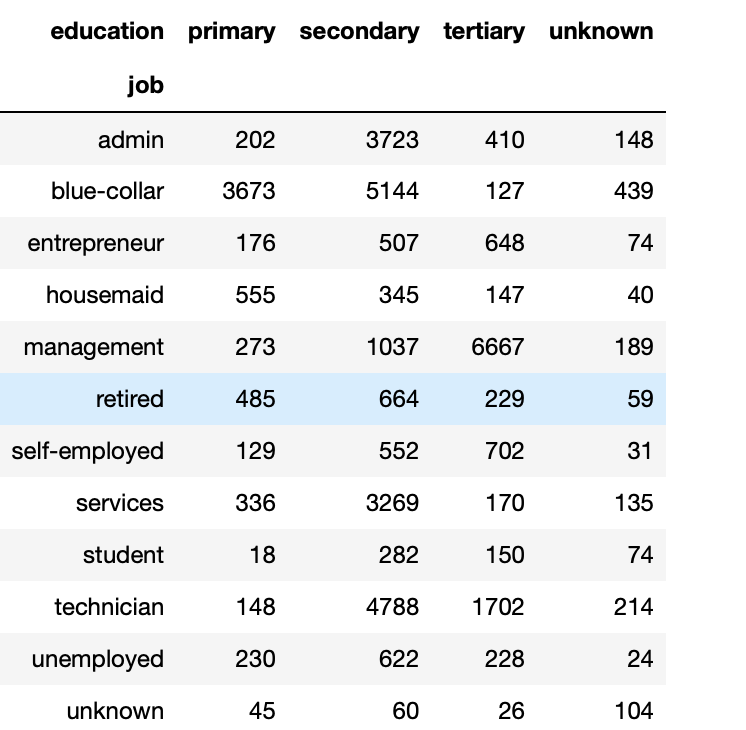
**Case Study**

**Data Preprocessing**

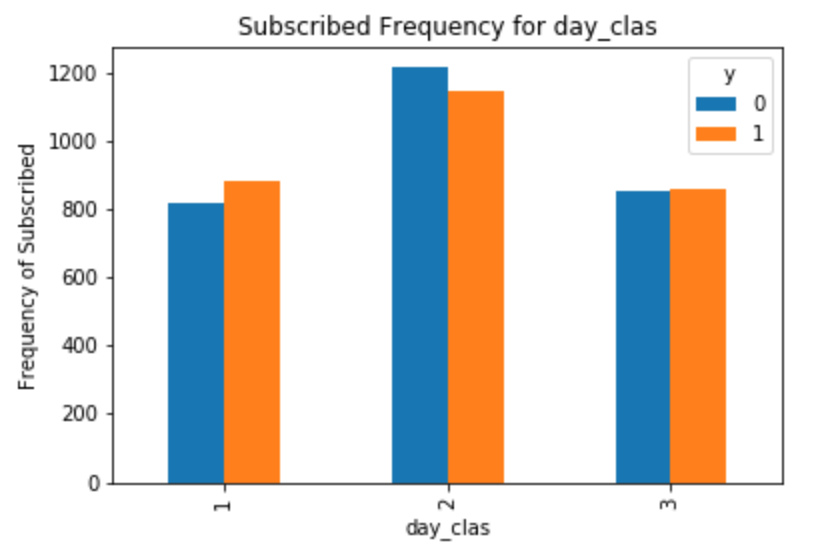
Veriyi incelediğimizde education alanında unknown olan değerler görünmektedir. Bu unkown değerleri doldurabilmek için kişinin mesleği değerlendirildi. Veri setinde İlgili kişinin mesleğine sahip olan diğer kullanıcıların eğitim. Durumu incelendi. Meslek grubuna göre eğitim durumu aşağıdaki tabloda gösterilmiştir. Bu tablodan yararlanılarak ilgili meslek grubunda en çok görünen eğitim durumuna göre unkown satırları dolduruldu.



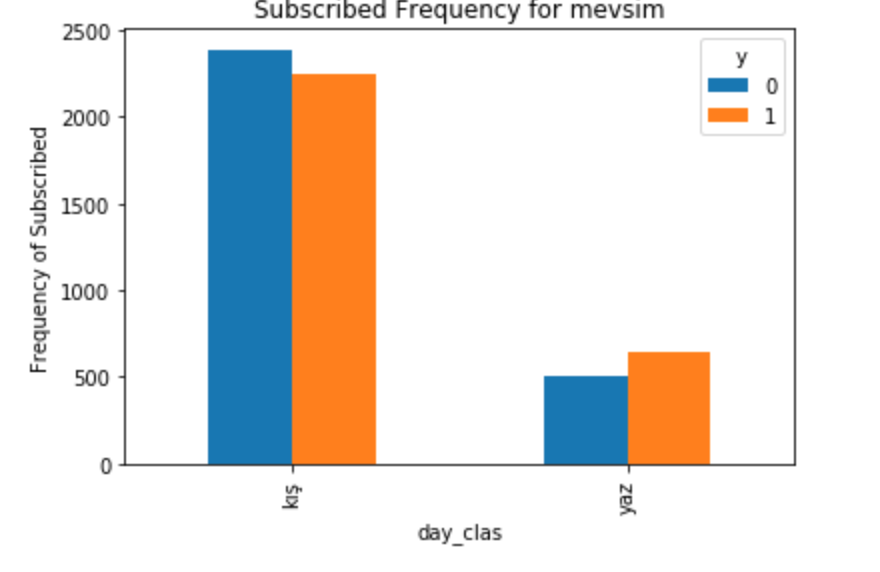
Bu işlemden sonra eğitim durumu unknown olan 104 kullanıcı kaldı. Artan eğitim durumuna göre sayısal değer ataması yapacağımız için kalan bu 104 kullanıcı. Veri setinden elendi. Sonuç olarak eğitim durumu primary olanlara 1, secondary olanlara 2 tertiary olanlara ise 3 değeri atandı.

Teklifi kabul eden müşteriler veri setimizin yaklaşık %7’lik kısmını oluşturmaktadır. Bu haliyle makine öğrenmesi modellerine koyarsak bias olacaktır. Modelimiz müşterinin teklifi kabul etmeme bilgisini iyi öğrenirken, kabul etme davranışını yeterince öğrenemeyecektir. Bu sebeple daha sağlıklı bir model kurabilmek için kabul edilen teklif sayısı kadar kabul edilmeyen telifler rassal olarak seçildi. Böylelikle son durumda veri setimizde kabul edilmiş etiket ve kabul edilmemiş etiket sayısı eşit duruma getirilmiştir.

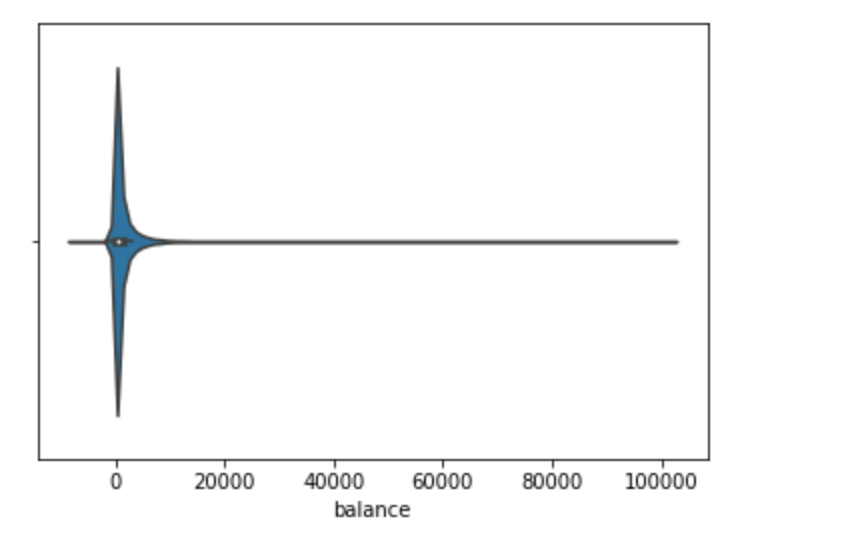
Day verisi kullanılarak ayın başı ortası ve sonu olacak şekilde ay 3’e bölündü. Ayın ilk 10 günü 3 ile etiketlenirken ortası 2 sonu da 1 ile etiketlenmiştir. Ayrılan bu üç sınıfa göre etiket frekansı aşağıdaki grafikte çizdirilmiştir. Buna göre ayın sonuna doğru kabul edilen tekliflerin oranı artmış görülmektedir.



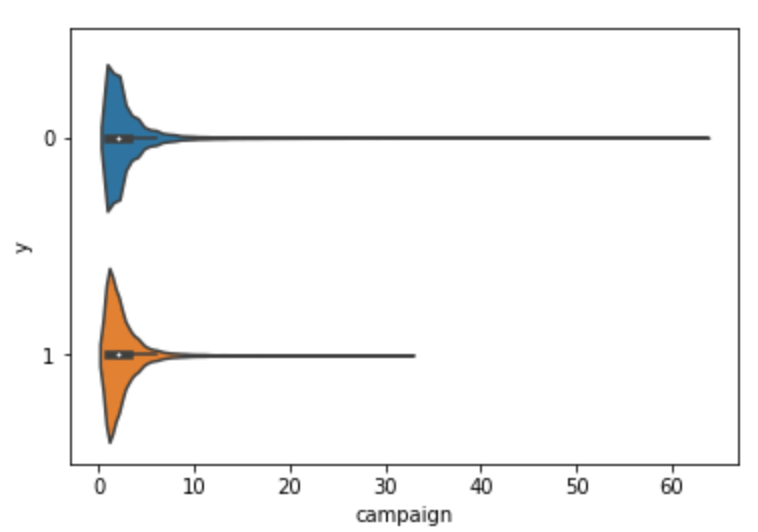
Month bilgisi kullanılarak veri seti yaz ve kış olacak şekilde iki gruba etiketlendi. Yaz ve kış gruplarına göre teklif kabul etme frekansı aşağıdaki gibidir. elimizdeki kalan veriye göre yaz etiketinde oran olarak daha fazla kabul eden müşteri verisi vardır diyebiliriz.



Veri setimizdeki Balance değerinin dağılımını anlayabilmek için violin plot grafiğini çizersek aşağıdaki gibi sonuç elde ediyoruz. Buna göre veri setimizde outlier diye nitelendirilecek datalar mevcut. Fakat veri setimizi küçültmemek için outlier elemesi yapmayacağız(outlier elemesi yapıldığında model performansları düşmektedir)



Kampanya süresince müşteri ile iletişime geçme sayısını ifade eden campaign bilgisinin teklifi kabul edip etmeme etiketine göre dağılımını çizdiğimizde aşağıdaki grafiği elde ediyoruz. Buna göre teklifi kabul etmeyen grupta campaign verisinin üst sınırı daha yüksek değerlere ulaşmaktadır.



Tahmin modelleri uygulayabilmek için Veri setimizdeki kategorik değişkenler dummy variable’a çevrildi.

Veri serimizde müşteri ile ilerişim sayısını gösteren campaign datası bulunmaktadır. Bu metriği Frequency olarak düşünebiliriz. Bunun yanında müşterinin balance değeri de bulunmaktadır. Bu metriği de Monitoring metriği olarak düşünebiliriz. Sonuç olarak elimizde F ve M metrikleri bulunmaktadır. RFM modeli uygulayabilmek için tek eksiğimiz müşteriyi son arama tarihidir. Tracsection datası olsaydı R metriğini de elde edebilirdik. Bu şekilde RFM modelini uygulayarak müşterileri gruplayabilirdik. Bu gruplama sonucu modelin performansı iyileşebilir.

**Model Uygulamaları**

1. **Logistic regresyon modeli**

Veri setimize logistic regresyon modelini 5 Cross validation(cv) ile uyguladığımızda elde ettiğimiz ortalama doğruluk(accuracy) performansı %77.8 olarak çıkmaktadır. Ortalama Precission değeri %81.6 olarak ölçülmüştür. Yani gerçekte 1 olarak tahminlediklerimizin %81.6’i gerçekte 1’dir. Recall parametresi de %71.5 olarak ölçülmüştür. Yani gerçekte 1 olan verilerin %71.5’ini 1 olarak tahminlemişiz. Recall ve Precission değerlerinin harmonik ortalamasını veren F1-score değeri ise %77.5 olarak ölçülmüştür.

Modelin çıktısına göre attribute’ların log-odd oran ve odd oran değerleri aşağıdaki tabloda verilmiştir.



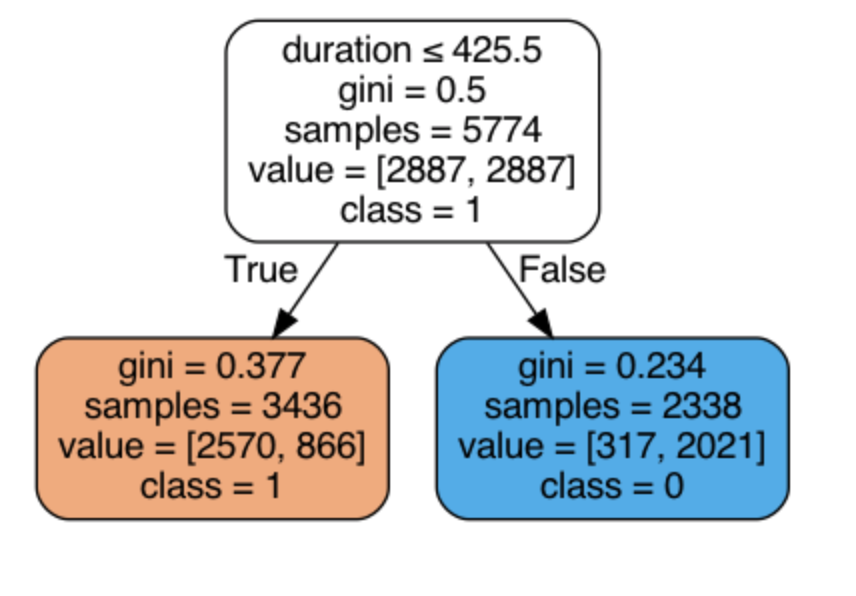
Attribute’ların odd oran değerleri “exp” isimli kolanda yer almaktadır. Odd-oran değerleri 1’den büyük olan attribute’lar müşetirinin kabul etme davranışını olumlu yönde etkilerken 1’den küçük olan attribute’lar olumsuz yönde etkilemektedir.

Logistik regresyon modeline göre etiketin pozitif olmasını etkiyen en güçlü attribute mevsimin yaz olmasıdır. Yaz mevsiminin odd-oran değeri 1.43’tür. yani mevsimin yaz olması müşterinin teklifi kabul etme ortalamasını %43 oranında arttırmaktadır. Bunun yanında meslek grubunun retired yani emekli olması müşterinin teklifi kabul etme olasılığını %28 oranında arttırmaktadır. Müşteriye ulaşım şeklinin de telefon kanalından olması da kabul ortalamasını arttırmaktadır.

Contact şeklinin unknown olması kabul şansını %60 oranında düşürmektedir(odd-oran değeri 0.4) bunun yanında müşterinin house\_yes attribute’una sahip olması yani ev sahibi olması teklifi kabul etme ortalamasını %58 oranında azaltmaktadır.

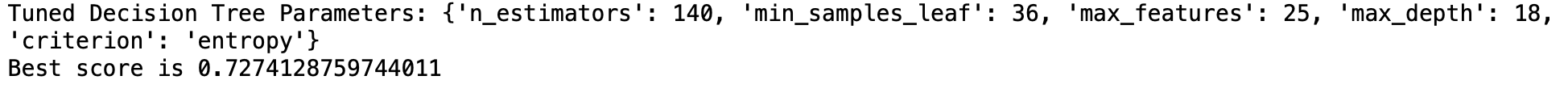
**Karar Ağacı modeli**

Veri setine karar ağacı uygularken 5 CV uygulandı. En optimal hyper parametreleri saptayabilmek için GridSearh metodu uygulandı. Buna göre en iyi accuracy performansını veren hyperparametre değerleri saptanıldı. Model sonucunda en iyi ortalama accuracy değeri %79 olarak ölçüldü. En iyi performans gösteren Karar ağacının görseli aşağıdaki gibidir. Buna göre tek karar kriteri müşteri ile konuşma süresini ifade eden duration attribute değeridir. Müşteri ile konuşma süresi 425.5 değerinden büyük olan denemeleri pozitif etiketlemiştir. Karar ağacını siplit ederken gini index kullanılmıştır.



**Random Forest Modeli**

Ensemble modellerinden olan Random Forest modelini uygulandı. En iyi hyperparametre’lerin belirlenmesi için yine GridSearch metdou kullanıldı. Gridsearch metoduna göre en iyi accuracy performansını veren hyperparametre değerleri aşağıdaki gibidir.



En iyi accuracy değeri %72 olarak belirlenmiştir. Yani karar ağacından daha düşük bir performans elde edimiştir.

**Neural Network Modeli**

Neural network modelini kurabilmek için Keras kütüphanesinden yararlanıldı. Cross-validasyon yapabilmek için Sckleran kütüphanesinin K-Fold’ undan yararlanıldı. Modele göre 4 hidden layer katmanı oluşturuldu. Bu layer’ların aktivasyon fonksiyonları sırası ile sigmoid, tanh, sigmoid ve sigmoid dir. Bu layerlardaki node sayısı sırası ile 28, 20, 15 ve 10 dur. Epoh sayısı da 80 olarak belirlenmiştir. Aynı anda optimize edilen parametre sayısını ifade eden batch\_size parametresi de 35 olarak belirlenmiştir. Loss fonksiyonu olarak binary\_crossentropy kullanılırken optimizasyon modeli olarak adam yöntemi uygulanmıştır.

Model veri setimizde 5 CV ile koşulduktan sonra elde edilen ortalama Accuracy değeri %81 olarak ölçülmüştür.