LOGİSTİC REGRESYON VE LİGHTGBM MAKİNE ÖĞRENME MODELLERİNE DAYALI KREDİ RİSKİ ANALİZİ

Zeynep PAMUK

**Özet**

Kredi riski analizi, potansiyel borçluların temerrüde düşme olasılığını belirlemelerini sağladığı için finansal kurumlar için çok önemli bir görevdir. Bu raporda, Loan Applicant Data for Credit Risk Analysis veri seti üzerinde logistic regresyon ve diğer makine öğrenimi modellerini kullanarak kredi riskini analiz ediyoruz. Amacımız, kredi temerrütlerini tahmin etmede en iyi performans gösteren modeli belirlemek ve kredi riski analizinde en önemli değişkenleri tespit etmektir. Çalışmamız, LightGBM’in 0,93 accuracy, 0,97 precision, 0,82 recall, 0,94 F1-score değeri ile en iyi performans gösteren model olduğunu göstermektedir. Logistic Regresyon, diğer modellere kıyasla düşük performans göstermiştir ve özellikle LightGBM’in gerisinde kalmıştır. LightGBM’in ROC eğrisi Logistic Regresyona kıyasla daha yukarıda ve eğri daha keskin bir şekilde 1’e yaklaşmaktadır. Bulgularımız, kredi temerrütlerini tahmin etmede en önemli değişkenlerin loan\_percent\_income(gelirin krediye oranı), person\_income(kişinin geliri) ve loan\_grade(kredi harf notu) olduğunu göstermektedir. Ayrıca, borçlunun yaşının kredi almasındaki amacın tahmin etmede önemli bir faktör olmadığını buluyoruz. Bu durum, kredi riskini analiz ederken diğer değişkenlerin de dikkate alınmasının önemini vurgulamaktadır. Çalışmamız, finansal kuruluşların kredi riski analiz modellerini geliştirmeleri için pratik çıkarımlar sağlamaktadır. LightGBM gibi makine öğrenimi tekniklerini kullanarak kredi riskini daha iyi belirleyebilir ve yönetebilirler, böylece temerrütlerden kaynaklanan kayıplarını azaltabilirler.

**Anahtar Kelimeler:** Kredi risk analizi, LightGBM, logistic regresyon, makine öğrenmesi.

1. **Giriş**

Kredi riski analizi, finansal kuruluşlar için kritik bir görevdir, çünkü kredi ödemelerini temerrüde düşürme olasılığı daha yüksek olan potansiyel borçluları belirlemelerine olanak tanır. Başvuru sahiplerini iyi veya kötü olarak sınıflandırmak için bankacılık sistemi temerrüt olasılığını değerlendirir. İyi kategoride yer alan başvuru sahiplerinin kredilerini bankaya geri ödeme olasılığı yüksek olarak kabul edilirken, kötü kategoride yer alanların geri ödeme olasılığı düşük olarak değerlendirilmekte ve temerrüde düşmüş olarak kabul edilmektedir. Güvenilir kredi riski veri setinin faydaları, kredi puanlama maliyetini azaltması, çok daha kısa sürede iyi karar vermesi ve kredi tahsilatı ile ilgili daha az riskten kaçınmasıdır (Pandey ve diğerleri, 2017).

Geleneksel olarak, finans firmaları kredi riskini modellemek için klasik doğrusal, logit ve probit regresyonları kullanmışlardır (Altman, 1968). Son yıllarda, kredi riski analizi için makine öğrenimi modellerinin kullanımı, büyük miktarda veriyi ve değişkenler arasındaki karmaşık ilişkileri ele alma yetenekleri nedeniyle önemli ölçüde dikkat çekmiştir. Galindo ve Tamayo 2000 yılında ipotekli kredi verilerindeki temerrütleri tespit etmek için CART karar ağacı, KNN ve probit modellerini kullanmıştır. Khandani ve diğerleri (2010) tüketici kredilerinde karar ağaçları ve DVM kullanan bir makine öğrenimi yöntemi önermektedir. Yazarlar bu tekniği gerçek kredi verileri üzerinde test etmiş ve %25'e varan maliyet tasarrufu sağladığını bulmuşlardır. Butaru ve diğerleri 2016 yılında lojistik regresyon, karar ağaçları ve rastgele orman algoritmalarını kullanarak altı farklı bankadaki tüketici temerrütlerini incelemiştir.

Bu raporda, Loan Applicant Data for Credit Risk Analysis veri setini kullanarak kredi riskini analiz etmek için logistic regresyon ve diğer makine öğrenimi modellerinin uygulanmasına odaklanıyoruz. Figür 1’de görüldüğü üzere çalışmada iki farklı yol izlendi. Gradient Boosted ve Logistic Regresyon algoritmaları için iki farklı veri bilimi iş akışı gerçekleştirildi.

**Figür 1.**

**metin, ekran görüntüsü, yazı tipi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Veri seti, bireylerin yaşı, yıllık geliri, ev sahipliği durumu, işte çalışma süresi, kredi amacı, kredi notu, kredi miktarı, kredi faiz oranı, kredi-gelir oranı, temerrüt geçmişi ve kredi geçmişi süresi gibi demografik ve finansal özellikler ile kredi durumları hakkında bilgi içermektedir. Farklı makine öğrenimi modellerinin kredi riskini tahmin etmedeki performansını karşılaştırmayı ve kredi temerrüdüne katkıda bulunan faktörleri değerlendirmeyi amaçlıyoruz. Özellikle, tahmin modelleri oluşturmak ve bunların doğruluğunu ve yorumlanabilirliğini değerlendirmek için lojistik regresyon, rastgele ormanlar ve gradyan destekli modeller kullanacağız.

Analizimizin sonuçları, kredi riskini yönetme ve temerrüdü önlemek için hedefe yönelik müdahaleler tasarlama konusunda finansal kuruluşlar için değerli bilgiler sağlayabilir. Ayrıca, çalışmamız kredi riski analizi için makine öğreniminin kullanımına ilişkin literatüre katkıda bulunmakta ve gerçek dünya veri setine uygulanmasına ilişkin pratik bir örnek sunmaktadır.

1. **Teorik Arka Plan**
   1. **Modeller**
      1. **Logistic Regresyon**

Lojistik regresyon, bağımlı değişken iki olası değere sahip ikili bir değişken olduğunda uygulanan ve tipik olarak bir sonuç olayın varlığını veya yokluğunu gösteren bir regresyon analizi türüdür. Bağımsız değişkenler sürekli, kategorik veya her ikisinin bir kombinasyonu olabilir. Ayrıca, bağımlı değişken veya hata terimleri için normal dağılımlar hakkında varsayımlarda bulunmaz.

Modelin şekli şöyledir:

Tahmin edici değişken değerlerini olasılıklara dönüştürmek için lojistik bir fonksiyon kullanılır.

Bu, basitliği nedeniyle kredi skorlama uygulamalarında yaygın olarak kullanılmaktadır (Satchidananda ve diğerleri, 2006). İkili lojistik regresyon Altman ve diğerleri tarafından kullanılmıştır.

Sabato (2007) ve Yazdanfar & Nilsson (2008) KOBİ'lerin temerrüde düşme olasılığını tahmin etmek için kullanmıştır. Yapay zeka yöntemlerinin geleneksel istatistiksel tekniklere kıyasla kredi riski analizinin sonuçlarını iyileştirdiğine dair kanıtlara rağmen (Abellan & Castellano, 2017), lojistik regresyon basit uygulaması ve dengeli hata dağılımı nedeniyle hala popülerdir.

* + 1. **Gradient boosted Makineler**

Karar ağaçları, verimliliği, doğruluğu ve yorumlanabilirliği nedeniyle yaygın olarak kullanılan bir makine öğrenimi algoritması olan gradyan destekli makinelerin merkezi unsurudur. Gradyan artırımlı karar ağacı (GBDT) olarak da bilinen bu algoritma, çok sınıflı sınıflandırma, tıklama tahmini ve sıralamayı öğrenme gibi çeşitli makine öğrenimi görevleri için en iyi performans gösteren teknik olarak kabul edilir.

XGBoost, CatBoost veya LightGBM gibi yaygın olarak kullanılan bazı ünlü gradyan artırma modelleri vardır. XGBoost, Extreme Gradient Boosting'in kısaltmasıdır ve tahmin görevleri için sıklıkla kullanılan bir makine öğrenimi modelidir. Bu model, ABD'deki Carnegie Mellon Üniversitesi'nde doçent Chen (2016) tarafından geliştirilmiştir. CatBoost, gradient boosting kütüphanesindeki en yeni modellerden biri olan Category Boosting'in kısaltmasıdır. Algoritma, bir grup olası özellik-bölünme çiftini bölünme olarak yaprağa atar ve en küçük cezaya neden olan bölünmeyi seçer. Dengeli ağaçlar daha hızlı hesaplama ve aşırı uyum kontrolü gibi avantajlar sunar. LightGBM, Light Gradient Boosted Machine olarak da adlandırılan bir makine öğrenimi modelidir. XGBoost gibi, asimetrik bir karar ağacı kullanır. Ancak, XGBoost'tan farklı olarak, ağacı seviye bazında değil yaprak bazında büyütür. Bu, LightGBM'nin daha küçük ve daha hızlı modeller ürettiği anlamına gelir. Bu model Microsoft tarafından 2017 yılında geliştirilmiştir, ilk olarak aynı yıl G.Ke ve arkadaşlarının makalesinde tanıtılmıştır.

1. **Veri**
   1. **Veri Açıklaması**

Bu raporda kullanılan veri kümesi, kredi başvuru sahipleri ve özellikleri hakkında temel bilgiler sağlar. Amaç, bu basitleştirilmiş özelliklere dayanarak kredi temerrüdü olasılığını belirlemek için öngörücü modeller geliştirmektir. Veri setine [buradan](https://www.kaggle.com/datasets/nanditapore/credit-risk-analysis/) ulaşılabilir.

Veri kümesi 32581 satır ve 12 sütundan oluşmaktadır. Tablo 1'de veri kümesindeki her bir sütunun ayrıntılı açıklaması yer almaktadır.

**Tablo 1. Sütunlar Açıklama**

|  |  |  |
| --- | --- | --- |
| No | Sütunlar | Açıklama |
| 1 | person\_age | Kişinin yaşı (yıl olarak) |
| 2 | person\_income | Kişinin yıllık geliri |
| 3 | person\_home\_ownership | Ev sahipliği türü (Kiracı, ev sahibi, ipotek, diğer) |
| 4 | person\_emp\_length | Kişinin işte çalışma süresi (yıl olarak) |
| 5 | loan\_intent | Kredinin amacı (kişisel, eğitim, tıbbi, girişim, ev iyileştirme, borç yapılandırma) |
| 6 | loan\_grade | Kredi notu (A, B, C, D, E, F, G) |
| 7 | loan\_amnt | Kredi miktarı |
| 8 | loan\_int\_rate | Kredi faiz oranı |
| 9 | loan\_percent\_income | Kredinin gelire oranı |
| 10 | cb\_person\_default\_on\_file | Kişinin daha önce bir temerrüt geçmişine sahip olup olmadığı (Evet; Hayır) |
| 11 | cb\_person\_cred\_hist\_length | Kişinin kredi geçmişi süresi |
| 12 | loan\_status | Kredinin şu an temerrütte olup olmadığını gösterir (1: Temerrüt, 0: Temerrütte değil) |

Veri setimizin hedef değişkeni, değerleri 0 ve 1 olan ikili bir değişken olan “loan\_status” dur. Kredi temerrüt riski, bireylerin ödünç verilen parayı zamanında geri ödeyememe olasılığıdır. Veri çerçevesinde 4 kategorik öznitelik ve 8 sayısal öznitelik bulunmaktadır.

* 1. **Veri Ön İşleme**

Veri Ön işlemeye başlamadan önce gerekli kütüphane ve modüllerin yüklenmesi ve içe

aktarılması gerekmektedir.

* + 1. **Ver Setinin İncelenmesi**

Veri ön işleme adımlarına öncelikle verimizi yüklemek ve incelemekle başlıyoruz.

Kategorik ve sayısal öznitelikler ayrı ayrı değerlendirilmiştir. Kategorik değişkenlerin detaylarına Tablo 2'de yer verilmiştir.

**Tablo 2.**

|  |  |
| --- | --- |
| person\_home\_ownership | Ev sahipliği türü |
| loan\_intent | Kredinin amacı |
| loan\_grade | Kredi notu (A, B, C, D, E, F, G) |
| cb\_person\_default\_on\_file | Kişinin daha önce bir temerrüt geçmişine sahip olup olmadığı (Y; N) |

**df = pd.read\_csv('loan\_data.csv') *🡪*** Veri setini yükleme

**pd.set\_option('display.max\_columns', None) *🡪*** Tüm satırlar gözlemlenir

**df.head(20) *🡪*** Veri setinin ilk 20 satırı getirilir

**df.shape[0],df.shape[1] *🡪*** Verinin kaç satır ve sütun olduğuna bakılır

**df.info() *🡪*** Veri çerçevesinin genel yapısını özetler

**df.describe() *🡪*** Sayısal sütunların istatistiksel özetini verir

* + 1. **Eksik Değerlerin Ele Alınması**

Kategorik özniteliklerde eksik değer bulunmadığından bu değişkenler için herhangi bir işlem uygulanmamıştır. Eksik değerler olması durumunda, mod (en sık tekrar eden değer) gibi yöntemlerle doldurma işlemi gerçekleştirilebilirdi.

**print(df.isnull().sum()) *🡪***  Her satırdaki eksik veri kontrol eder

metin, ekran görüntüsü, yazı tipi, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**mean\_emp\_length = df['person\_emp\_length'].mean() *🡪***  Ortalamayı döndürür

**df['person\_emp\_length'].fillna(mean\_emp\_length, inplace=True) *🡪*** Eksik verileridolu satırların ortalamasıyla doldurulmuştur

**loan\_int\_rate** sütunundaki değerlerin temerrüt üzerinde önemli rol oynadığından bu boşlukların olduğu satırlar silindi.

**df = df.dropna() *🡪*** Eksik kısımları siler

* + 1. **Tekrar Eden Verilerin Ele Alınması**

Eksik değerler temizlendikten sonra, veri setindeki tekrar eden satırlar analiz edilmiştir. Bu aşama tamamlandığında, veri seti eksik ve tekrarlayan verilerden arındırılarak daha tutarlı bir hale getirilmiştir.

**df.duplicated().sum() 🡪** Tekrarlı verileri kontrol eder



**df.drop\_duplicates(inplace=True) 🡪** Tekrarlı verileri siler.

Eksik değerlerin doldurulması, silinmesi ve tekrarlı verilerin de silinmesi işlemlerinden sonra elimizdeki **29,321** satıra düşmüştür.

* + 1. **Aykırı Değerlerin Ele Alınması**

Aykırı değer analizi, IQR yöntemi kullanılarak gerçekleştirilmiştir.

**def detect\_outliers\_iqr(df, column) 🡪** fonksiyonunu kullanarak sayısal sütunlardaki aykırı değerleri tespit edildi. Ardından plotbox grafiklerini çizdirilerek tespit edilen aykırı değerler görselleştirilmiştir. İşte birkaç örnek:

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**df['person\_age'].value\_counts().sort\_index() 🡪** Artan şekilde yaşlar sıralanmıştır ve sayısı gözlemlenmiştir.

**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**person\_age** sütununda, birçok bankanın kredi üst yaş sınırı dikkate alınarak, 70 yaş üstü veriler temizlenmiştir.

**df = df.loc[df["person\_age"] <= 70, :] 🡪** loc koşulu ile istenilen koşuulu sağlamayanlar veri setinden çıkarılmıştır.

**person\_emp\_length s**ütununda, 60 tan fazla çalışan kişiler silinmiş bu şekilde olan yalnızca 123 yıl çalışan 2 kişi vardır.

**df.loc[df['person\_emp\_length'] <= 60, :] 🡪** loc koşulu ile istenilen koşuulu sağlamayanlar veri setinden çıkarılmıştır.

Bu işlemler sonucunda, veri setindeki tutarsızlıklar giderilerek daha güvenilir bir analiz altyapısı oluşturulmuştur ve satır sayımız 29,309 ‘a düşmüştür.

* + 1. **Variables Encoding**

Logistic Regression modeline one hot encoding uygulandığında **multicollinearity** sorunu ortaya çıkar.

One-hot encoding, kategorik değişkenleri her bir sınıf için ayrı sütunlara böler ve bu sütunlardan biri diğer sütunların doğrusal bir kombinasyonu haline gelebilir (dummy variable trap). Bu durum, bağımsız değişkenler arasında yüksek bir ilişkiye yol açarak modelin katsayılarını doğru şekilde tahmin etmesini zorlaştırır ve istatistiksel güvenilirliği düşürür.

Bu sorunun üstesinden gelmek için, **label encoding** veya **binary encoding** gibi alternatif yöntemler kullanılabilir. Label encoding, kategorilere sayısal değerler atayarak daha az sütun oluşturur, binary encoding ise her kategoriyi ikili (binary) formatta kodlayarak hem bilgi kaybını azaltır hem de multicollinearity riskini düşürür. Bu yöntemlerle, Logistic Regression gibi modeller için daha uygun bir çözüm sunulmuş olur.

Bu sebepten bu aşamada veriables encoding kısmında modellere özel farklı tekniklerle iki farklı çalışma yapılmıştır.

* + - 1. **Logistic Regression Modeli için Encoding**

Logistic Regression’da one hot ile oluşabilecek multicollinearity probleminden dolayı binary ve label encoding yöntemleri tercih ettik.

Kategorik sütunlarda belirli bir sıralamaya ve düzene sahip olmayan aralarında bir sıralanamaz(nominal) bir ilişki olan sütunlar için **binary encoding** yapılır.

Binary encoding, her kategori için ikili (binary) format oluşturur ve kategorilerin arasında ilişki ya da sıralama olmadığı için veriyi daha kompakt ve etkili şekilde temsil eder. Aynı zamanda one-hot encoding'e kıyasla sütun sayısını azaltarak multicollinearity riskini azaltır

Kategorik verilerde belirli bir sıralamaya ve düzene sahip aralarında bir sıralanabilir(ordinal) ilişki olan sütunlar için **label encoding** yapılır.

Label Encoding kategorilere sıralamalarına göre sayısal değerler atar (örneğin: lise = 1, üniversite = 2, yüksek lisans = 3). Bu, sıralama bilgisini koruduğu için ordinal kategorilerde doğru bir temsili sağlar.

**Sıralı olanlar**:

*loan\_grade* 🡪 kredi notu (A, B,C,D,E,F,G) en iyi norun A dan başladığını varsayaeak onun katsayısını daha ağır olacak şekilde verdik bu durumda A = 7 …. G = 1 olur.

*cb\_person\_default\_on\_file 🡪*   Kişinin temerrüt geçmişi varsa (Yes, No)Y=0, N=1 olur.

*person\_home\_ownership* 🡪 Ev sahipliği türü (Own, Mortgage, Rent, Other) Rent=1, Mortgage=2, Own=3, Other=0 olarak ağırlık verdim.

**Sırasız olanlar:**

*loan\_intent* 🡪  Kişinin kredi için niyeti (Education,personal, medical, Venture, Debtconsolidation, Homeimprovement)

def fonksiyonları kullanılarak label encoding işlemlerini yaptırıldı.

**df["…"] = df["…"].apply(SC\_LabelEncoder1) à** Sütun isimlerini içe aktardık ilgili fonksiyona yolladık.

ekran görüntüsü, metin, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**encoder = ce.BinaryEncoder(cols=['loan\_intent']) 🡪**Binary encoding işlemini gerçekleştiren bir encoder sınıfıdır. loan\_intent binary encoding ile 3 sütun haline gelir.

⌈log2​(6)⌉=⌈2.58⌉=3

**df\_encoded = encoder.fit\_transform(df) 🡪** df isimli DataFrame üzerinde binary encoding işlemini gerçekleştirir

ekran görüntüsü, metin, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

* 1. **Keşifsel Veri Analizi (EDA)**

EDA kısmında tüm çalışmalara ait verileri burada birlikte paylaşacağız.

**Şekil 1**’ deki tabloda Korelasyon tablosu bulunmaktadır.

Modelin türüne göre (örneğin, lineer regresyon, karar ağaçları, vs.) bazı algoritmalar korelasyona daha duyarlıdır. Lineer modellerde yüksek korelasyon ciddi sorunlar yaratabilirken, ağaç tabanlı modeller (örneğin, karar ağaçları, rastgele ormanlar) yüksek korelasyonlu değişkenlerle daha iyi başa çıkabilir.

Korelasyon matrisi, veri setindeki sayısal değişkenler arasındaki ilişkileri ölçmek için kullanılan bir tablodur. Bu matris, her bir değişken çifti için **Pearson korelasyon katsayısını** gösterir. Korelasyon katsayısı, -1 ile 1 arasında bir değer alır:

* **1'e yakın değerler**: Güçlü pozitif ilişki (bir değişken artarken diğeri de artar).
* **-1'e yakın değerler**: Güçlü negatif ilişki (bir değişken artarken diğeri azalır).
* **0'a yakın değerler**: İki değişken arasında anlamlı bir ilişki yoktur.

Tabloda yalnızca cb\_person\_hist\_length kısmı ve person\_age kısmı kırmızı ile işaretlenmiştir. Bu iki sütun arasında güçlü bir benzerlik olduğunu gösterir dolayısıyla iki sütundan birinin kaldırılması gerekir. Aralarındaki ilişki çok güçlü çıkan iki sütundan birini kaldırmamızın temel nedeni, **multicollinearity** (çoklu doğrusal ilişki) probleminden kaçınmaktır bu problem modelin doğruluğu ve kararlılığını azaltır.

**cb\_person\_cred\_hist\_length:**  kredi geçmişi uzunluğu genellikle daha iyi kredi puanına işaret eder ve kredi risk analizinde kritik bir faktör olarak kabul edilir.

**person\_age**, dolaylı bir etkiye sahiptir. Yaş tek başına kredi riski açısından yeterli bilgi sunmayabilir; ancak kredi geçmişi uzunluğu bireyin ödeme geçmişi hakkında doğrudan bilgi sağlar.

Bu sebeplerden dolayı person\_age sütunu silinir.

**df = df.drop(columns=['person\_age']) 🡪** person\_age satırını siler

Şekil 1.

**metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Figure 2**’de “loan\_status” hedef değişkenini incelediğimizde, temerrüde düşenlerin oranı %21,8 iken, temerrüde düşmeyenlerin oranı %78,2'dir.

**Figur 3’**te Veri setinde en yoğun yaş grubu 20-30 aralığında bulunuyor. Bu da müşterilerin çoğunluğunun çalışma yaşında olduğu ve kredi aldıklarında muhtemelen en az bir işte çalıştıkları anlamına gelmektedir.

Grafik sağa doğru uzun bir kuyruk göstermektedir. Bu, veri setinde daha az sayıda yaşlı kişinin bulunduğunu ifade eder. (Silinmeden önce age değerlendirmesi)

**Figür 3. Yaş Dağılımı**

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Figure 4’te** kredi harf notu incelendiğinde kayda değer bir gözlem yapılabilir. Kredi puanı yüksek olan bireylerin kredilerini ödememe eğilimi daha düşükken, kredi puanı düşük olanların ödememe olasılığı daha yüksektir. Bu durum, kredi puanının kredi temerrüdünü tahmin etmede önemli bir değişken olabileceğini göstermektedir.

**Figür 4. Temerrüt Durumunun** dikdörtgen, diyagram, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**Kredi Notuna Göre İncelenmesi**

**Figür 5** Temerrüt Durumunun Kredinin Gelire Oranına Göre İncelenmesi.Kredinin gelire oranı incelendiğinde geliri daha çok olan ve veya kredi tutarı az olan müşterilerin yani kredinin gelire oranı düşük olduğu durumlarda müşterilerin kredilerini ödeme eğiliminin daha yüksek olduğunu görüyoruz. Etiketin sıfır olduğu durumdaki aykırılıklar göze çarpabilir.

**Figure 5**

diyagram, çizgi, dikdörtgen, ekran görüntüsü içeren bir resim

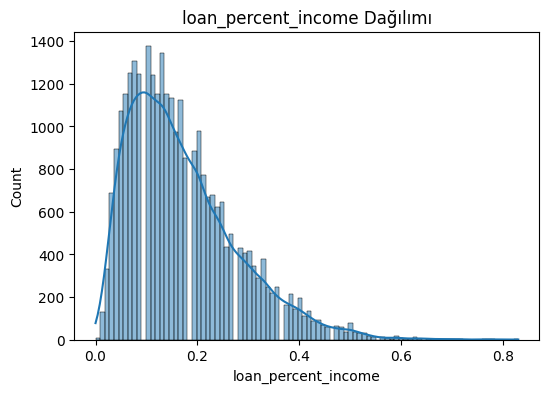
Açıklama otomatik olarak oluşturuldu

**Figür 6** ‘da gördüğümüz üzere Grafik, loan\_percent\_income değişkeninin dağılımını göstermektedir. Bu değişken, bireylerin gelirlerinin yüzde kaçının kredi ödemelerine ayrıldığını ifade etmektedir.

Değişkenin dağılımı asimetrik bir yapıya sahiptir ve pozitif çarpıklık (right-skewed) göstermektedir. Bu durum, çoğu bireyin gelirlerinin küçük bir yüzdesini kredi ödemelerine ayırdığını, ancak bazı bireylerde bu oranın oldukça yüksek olabileceğini göstermektedir. En yüksek yoğunluk, loan\_percent\_income değişkeninin 0.1 ile 0.2 arasındaki değerlerinde gözlemlenmiştir.

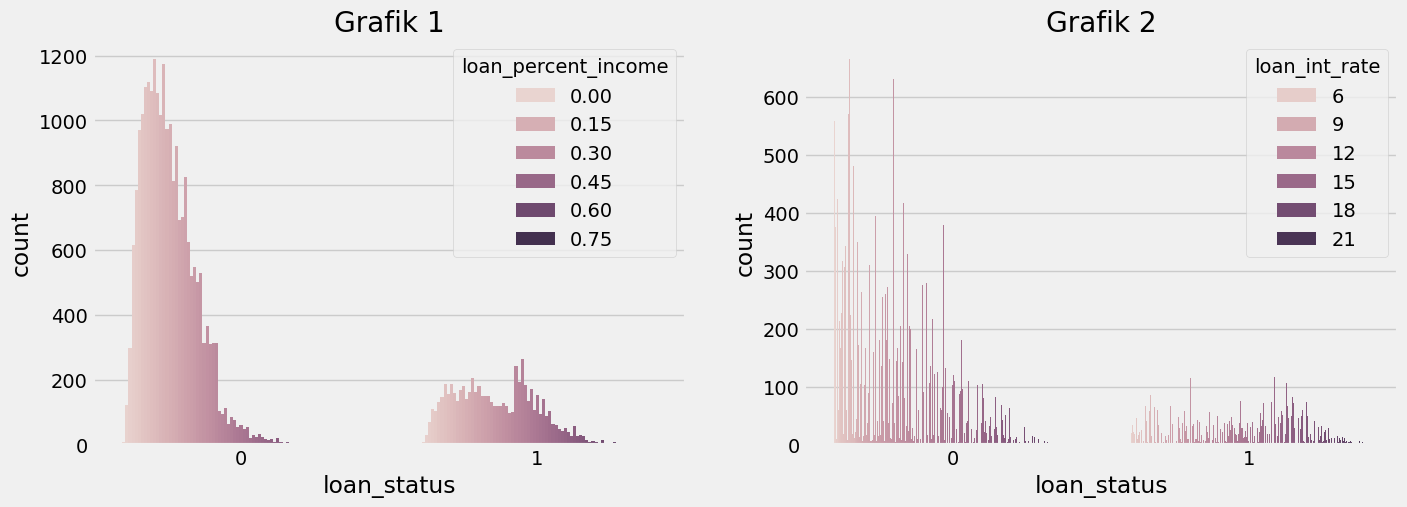
Bireylerin büyük bir çoğunluğu kredi ödemeleri için gelirlerinin %10-30 arasını ayırmaktadır. Bu dağılım, düşük loan\_percent\_income oranlarına sahip bireylerin daha düşük kredi riski taşıyabileceğini düşündürmektedir.

**Figür 6**



**Grafik 1** ‘de *loan\_status* (temerütte düşme, temerrüte düşmeme) ve *loan\_percent\_income* (kredinin gelire oranı) arasındaki ilişki incelendiğinde koyu olan renkler kredinin gelire oranın fazla olduğunu yani olumsuz olan durumları temsil eder. Ve temerrüte düşen kişilerde bu olumsuz durum yoğunluğunun fazla olduğunu gözlemleriz. Yani kredi/gelir oranı arttıkça temerrüte düşme ihtimali artmaktadır.

**Grafik 2** ‘de *loan\_status* ile *loan\_int\_rate*(kredi faiz oranı) arasındaki ilişki incelenmiştir. Ve temerrüte düşen kişiler (1) incelendiğinde kredi faiz oranı yüksek olan kısmın yoğunlaştığını görürüz. Bu durum kredideki faiz oranı artıkça kişilerin krediyi ödeyemediği şeklinde yorumlanabilir.



**Grafik 3 ‘** de *loan\_status* ile *person\_home\_ownership*(ev sahipliği) arasındaki ilişki incelendiğinde kendi evine sahip olan kişilerin temerrüte düşme oranın az olduğu ve kirada olan kişilerin temerrüte daha fazla düştüğü gözlemlenir. Temerrüte düşmeyen kişilerde rent ve mortgage neredeyse aynı sayıya sahiptir.

**Grafik 4** ‘de *loan\_status* ile *loan\_grade(kredi notu)* arasındaki ilişki incelenmiştir. Ve kredi nou yüksek olan kişilerin(A,B) yüksek yoğunluk ile temerrüte düşmediği gözlemlenmiştir. Temerrüte düşenlerin büyük yoğunluğu D yani düşük kredi notuna sahip kişilerdir.

metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Grafik 5’** de *loan\_status* ile *loan\_intent*(kişinin kredi için niyet) arasındaki ilişki incelenmiştir. Ve temerrüte düşen kişiler ve düşmeyen kişler arasında kredi alma sebepleri arasında doğrudan bir rişki gözlemlenmemiştir. Örneğin dağılım olarak bakıldığında Venture oranı 0’da 1‘den daha fazladır ve Education için kredi alanlar da aynı şekilde 0’da 1’den daha fazla yoğunluğa sahiptir. Medical için kredi alan kişilerde 1’deki yoğunluk 0 dan fazladır.

**Grafik 6’** da *loan\_status* ile *cb\_person\_default\_on\_file* (kişinin te temerrüt geçmişi olması*)* arasındaki ilişki incelenmiştir. Ve 0’daki dağılıma bakıldığında daha öne temerrüte düşmeyen kişilerin yoğunluğunun düşenlere oranla çok daha fazla olduğu saptanmıştır. 1’deki dağılım incelendiğinde ise bu oran çok daha yakındır yani önceden temerrütte düşen ve düşmeyenleler neredeyse aynı dağılıma sahiptir.

metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Boxplot 1 ‘de** *loan\_grade* ile *cb\_person\_default\_on\_file* (temerrüt geçmişi olması) arasındaki ilişki incelenmiştir.

Daha öncesinde temerrüte düşen kişilerin(Y) kredi notu A-B-C de yoğunlaşmıştır ve medyanı da yine bu değerler arasındandır ve G ye doğru olan notları aykırı değer olarak değerlendirmiştir. Daha önce temerrüte düşmemiş kişilerde(N) kredi notu daha düşük olacak şekilde bir yoğunlaşma vardır.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Boxplot 2** ‘deloan\_amnt ile loan\_intent arasındaki ilişki incelenmiştir. Ve grafik incelendiğinde Homeimprovement için alınan kredinin 5000-15000 arasında yoğunlaştığı ve aykırı değer bulunmadığı gözlemlenmektedir. Genel olarak baktığımızda tüm kredi gerekçelerinin 5000- 10000 arasında medyan aldığı gözlemlenmektedir.

metin, diyagram, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

* 1. **Veri Modelleme**

Keşifsel Veri Analizi adımından sonra, veri kümesini 80/20 oranında iki parçaya ayırdık: eğitim seti ve test seti. Eğitim setinde 23448 satır, test setinde ise 5863 satır bulunmaktadır. Eğitim seti, bir makine öğrenimi modelini eğitmek için kullanılan büyük bir veri setidir. Bu, modelin öğreneceği ve hatırlanması gereken önemli özellikleri çıkaracağı veri kümesidir. Test seti, eğitimden sonra model tarafından elde edilen sonuçların gerçekten etkili olup olmadığını doğrulamak için kullanılan bir veri setidir.

Modelin performansını daha güvenilir şekilde değerlendirmek ve aşırı öğrenme (verittin) gibi sorunları önlemek amacıyla **k-fold çapraz doğrulama** (k-fold cross validation) yöntemi kullanılmıştır. Veri kümesi, her birinde eşit veya benzer sayıda veri bulunan 5 alt gruba (katmana) ayrılmıştır. Bu yöntemde model, 5 farklı deneyde eğitim ve test süreçlerinden geçirilir. Her deneyde, bir katman test seti olarak kullanılırken geri kalan 4 katman eğitim için kullanılır. Böylelikle modelin farklı veri alt kümelerindeki başarımı gözlemlenmiş olur.

K-fold çapraz doğrulamanın tercih edilme nedeni, modelin yalnızca belirli bir eğitim-test ayrımına bağlı performans göstermesini önlemek ve genelleme kapasitesini artırmaktır. Ayrıca bu yöntem, modelin farklı veri örneklerinde tutarlı sonuçlar verip vermediğini değerlendirerek daha gerçekçi bir başarı ölçümü sunar. 5 katmanlı (5-fold) bir yapı, veri kümesinin yeterli sayıda gözlem içermesi durumunda hem hesaplama maliyeti hem de değerlendirme doğruluğu açısından uygun bir denge sağlar.

* + 1. **Logistic Regression ve LightGBM Classification Modeli**

**3.4.3.1 Lojistik regresyon:**

Lojistik regresyon, isminde “regresyon” geçmesine rağmen bir sınıflandırma algoritmasıdır. Yani görseldeki hayvanın kedi mi, köpek mi olduğu veya verilmiş olan bilgilerin bir erkeğe mi yoksa bir kadına mı ait olduğunu tahmin etme gibi iki sınıflı sınıflandırma problemlerinde sıkça kullanılır.

Bir dizi girdi değişkeni ile iki olası değerden (örneğin evet/hayır, doğru/yanlış, 1/0) birini alabilen bir çıktı değişkeni arasındaki ilişkiyi anlamamıza yardımcı olan bir istatistiksel analiz türüdür. Bunu, girdi değişkenlerinin değerleri göz önüne alındığında çıktı değişkeninin belirli bir değer alma olasılığını tahmin ederek yapar.

Lojistik regresyonun, lineer regresyon ile arasındaki en büyük farkı iki sınıfı birbirinden ayıracak çizgiyi nasıl uyguladığıdır (fit). Lineer regresyon, optimum çizgiyi çizmek için “En Küçük Kareler Yöntemi” (Least Squares) kullanırken, lojistik regresyon “Maksimum Olabilirlik” (Maximum Likelihood) kullanır.

Lojistik regresyon, sınıflandırma yapmak için ***Sigmoid***(Lojistik) Fonksiyonu kullanır.

**Avantajları:**

1. Lojistik regresyonun uygulanması, yorumlanması kolaydır.
2. Veri seti doğrusal olarak ayrılabiliyorsa oldukça iyi performans gösterir.
3. Overfitting’e daha az meyillidir ama büyük veri setlerinde overfit olabilir.

**Dezavantajları:**

1. Gözlem sayısı özellik sayısından azsa, Lojistik Regresyon kullanılmamalıdır, aksi takdirde overfit olabilir.
2. Lojistik regresyonun ayrım yapabilmesi için veri setinin doğrusal olarak ayrılabiliyor olması lazım.

Logistic Regression modelini kredi risk analizi için şu şekilde oluşturdum:

# 2. Modelleri tanımlama

models = {

    'Logistic Regression': LogisticRegression(max\_iter=1000, random\_state=42),

    'LightGBM': lgb.LGBMClassifier(random\_state=0)

}

Logistic Regression Model Değerlendirme Sonuçlarında **recision** 0.7217, **recall**: 0.4445,

**F1-Score**: 0.5502, **accuracy** 0.8402 olarak çıkmıştır. Bu sonuçlara logistic regression modelinin kredi risk analizi için uygun bir model olmadığını gözlemledik bu yöntem yerine daha karmaşık modeller uygulanabilir, random forest Gradient Boosting gibi.

Modelin recall, F1-Score değerlerinin düşük çıkmasının temel nedeni, pozitif sınıfın (örneğin, "yüksek riskli kredi başvuruları") doğru şekilde tahmin edilememesidir. Bu durumun temel sebebi Logistic Regression modelinin verideki sınıf dengesizliğinden çok fazla etkilenmesidir çünkü modelin çoğunluk sınıfa odaklanmasına ve azınlık sınıfını göz ardı etmesine neden olur. Ayrıca Logistic Regression, lineer bir modeldir ve karmaşık, doğrusal olmayan ilişkileri öğrenemez. Eğer pozitif sınıflar arasındaki ilişkiler doğrusal değilse, model bu yapıyı yeterince iyi yakalayamayabilir.

Bir veri modelini geliştirirken önemli olan geliştirdiğimiz projede modelleme değerlendirme sonuçlarının yüksek çıkması değildir çünkü her model her proje için uygun olmayabilir bunun çok çeşitli sebepleri vardır. Örneğin şu an kendi geliştirdiğimiz projede aynı kodları ile iki farklı modeli çalıştırdığımda Logistic Regression da üstteki değerleri verirken LightGBM Classification modelinde başarı elde edilmektedir.

**3.4.3.2 LightGBM Classification**

LightGBM, Gradyan Tabanlı Tek Taraflı Örnekleme (GOSS) ve Özel Özellik Birleştirme (EFB) gibi önemli yenilikleri içerir. GOSS, daha büyük gradyan örneklerini hedef alarak daha verimli bir şekilde örnekleme yaparken, EFB otomatik özellik seçimi gerçekleştirir ve artırma sürecini geliştirir. Bu yenilikler, modelin daha hızlı eğitilmesine olanak tanırken, tahmin doğruluğunu korur veya artırır. Bu özellikleri sayesinde, LightGBM, sınıflandırma ve regresyon görevlerinde büyük tabular verileri işlemek için ideal bir araçtır.

LightGBM'in model yapısı, güçlü bir tahminci model oluşturmak için karar ağaçları kullanan bir gradyan artırma yöntemine dayanır. Verimliliğe olan vurgusu, dal büyüme yaklaşımındaki yapısal tercihlerden ve histogram tabanlı algoritmaların kullanımından açıkça görülmektedir. Dal büyüme yaklaşımında, ağaçlar maksimum kayıp azaltımı sağlayan düğüm seçilerek daha derin ağaçlar oluşturulur, bu da doğruluğu artırırken, küçük veri setlerinde aşırı öğrenmeye (overfitting) yol açabilir. Histogram tabanlı yöntem ise sürekli özellikleri histogramlara dönüştürerek hesaplama karmaşıklığını ve bellek kullanımını azaltır, bu da özellikle büyük veri setlerini işlerken avantaj sağlar.

Bu stratejilerin kombinasyonu, paralel ve dağıtık hesaplama yetenekleriyle birlikte, LightGBM'in hızlı eğitim süreleri, düşük bellek tüketimi ve yüksek ölçeklenebilirlik sunmasına olanak tanır. Bu özellikler, farklı makine öğrenmesi görevleri için LightGBM'i uygun bir seçenek haline getirir.

Bu şekilde modeli oluşturdum.

# 2. Modelleri tanımlama

models = {

    'Logistic Regression': LogisticRegression(max\_iter=1000, random\_state=42),

    'LightGBM': lgb.LGBMClassifier(random\_state=0)

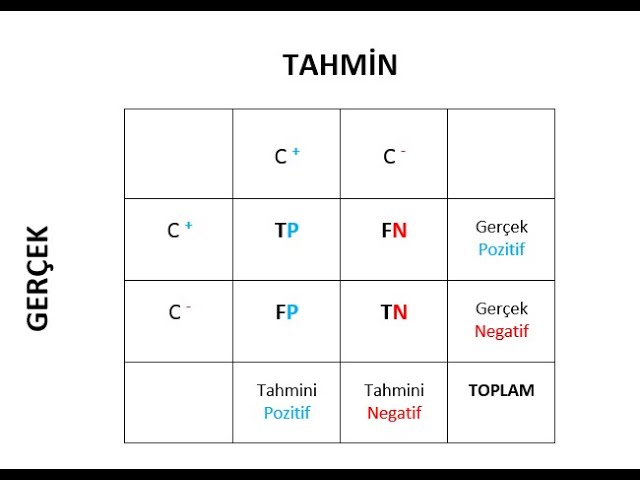
}

LightGBM’in kredi risk analizi projemde vermiş olduğu değerlendirme metrikleri şu şekildedir. accuracy: 0.9319, precision: 0.9665, recall: 0.7153, F1-Score: 0.8221 şeklindedir. Bu sonuçların aynı kodda daha iyi sonuç vermesinin sebebi tamamen modelin matematiksel yapısı ev mimarisiyle ilgilidir. Gradient boosting algoritmalarının verideki dengesizlikten çok etkilenmeyip Logistic regression modelinin dengesizliğe hassasiyetinin fazla olmasıdır.

* 1. **Model Değerlendirme**

Bir modelin performansını ölçmek için precision, recall, F1 skor, AUC değeri ve accuracy gibi çeşitli ölçütler kullanılabilir. Bu çalışmadaki ana değerlendirme metrikleri AUC değeri, F1 skor, precision, accuracy ve recall içermektedir. Bu metrikleri tanımlamak için öncelikle dört değişken oluşturuyoruz: (i) TP, sınıflandırma modeli tarafından doğru tahmin edilen pozitif örneklerin sayısını, (ii) FN, model tarafından negatif bir sınıf olarak yanlış tahmin edilen pozitif örneklerin sayısını (Tip II hata olarak da bilinir), (iii) FP, model tarafından pozitif bir sınıf olarak yanlış tahmin edilen negatif örneklerin sayısını (Tip I hata olarak da bilinir) ve (iv) TN, model tarafından doğru tahmin edilen negatif örneklerin sayısını temsil eder.

**Tablo 3. Confusion Matrisi**



Karışıklık Matrisinden, sınıflandırıcının tahmin performansını değerlendirmek için çeşitli ölçütler hesaplanabilir. En yaygın kullanılan metriklerden biri, doğru sınıflandırılmış örneklerin tüm örnek sayısına oranını ölçen doğruluktur. Doğruluk aşağıdaki formülle hesaplanır:

Accuracy =

Temerrüt riski analizinde, temerrüt kullanıcılarının belirlenmesi bankalar için büyük önem taşımaktadır. Bu bağlamda, recall ve precision önemli değere sahip iki değerlendirme göstergesidir. Precision, sınıflandırıcı tarafından tahmin edilen gerçek pozitif örneklerin oranını belirleyerek hesaplanır. Öte yandan, recall, sınıflandırıcı tarafından doğru tahmin edilen pozitif örneklerin oranı olarak tanımlanır:

Recall

Accuracy ve recall, bir sınıflandırıcının güvenilir veriler üzerinde doğru tahmin yapma yeteneğini değerlendirmek için kullanılan ölçütlerdir. Ancak, accuracy genel performansı ölçerken, recall yalnızca pozitif sınıfların doğru tahmin edilme oranını değerlendirir. Bu iki metrik arasında tutarsızlıklar olabilir ve bu da her ikisini aynı anda iyileştirmeyi zorlaştırabilir. Bu nedenle, bir modelin tahmin performansını daha kapsamlı değerlendirmek için, hem precision hem de recall’u dengeleyen F1 skoru kullanılır. F1 skoru hesaplama yöntemi aşağıdaki denklemde gösterilmektedir:

ROC eğrisi sınıflandırma problemleri için çok önemli bir performans ölçümüdür. ROC bir olasılık eğrisidir ve altında kalan alan olan AUC ayrılabilirliğin derecesini veya ölçüsünü temsil eder.

ROC eğrisinde X ekseninde FPR(Yanlış Pozitif Oran) ve Y ekseninde ise TPR (Gerçek Pozitif Oranı) bulunmaktadır.

**Tablo 4.**

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

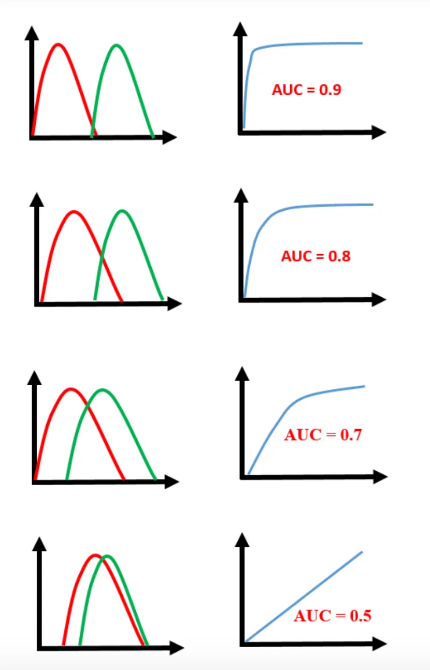
Açıklama otomatik olarak oluşturuldu

Eğrinin altında kalan alan arttıkça sınıflar arasında ayırt etme performansı artmaktadır.

AUC modelin sınıfları ne kadar başarılı ayırt edebildiğini anlatır. AUC arttıkça, model 0'ları 0 ve 1'leri 1 olarak tahmin etmede daha iyidir.

**Tablo 5**’de görüldüğü gibi AUC değeri yükseldikçe yeşil ve kırmızı olan sınıflar birbirinden iyi bir oranda ayrılmaktadır. Son grafikte görülen AUC değeri 0.5 olan modelin performansının kötü olduğunu ve rastgele tahminleme yaptığı söylenebilir. (Öğündür)

**Tablo 5.**



Dengeli dağılmayan veri setlerinde model başarısını sadece Accuracy metriği ile ölçmek yeterli değerlidir ve Precision-Recall metriklerini mutlaka kontrol etmemiz gereklidir. Ancak Precision değeri arttıkça Recall değerinin azaldığı görülmektedir. Bu nedenle bu iki değerin harmonik ortalaması olan F1 metriği bize sınıflandırma modelimizin gerçek başarısını göstermektedir. ROC eğrisi de bize F1 değeri arttıkça AUC altında kalan alanın arttığını göstermiştir.

* + 1. **Logistic Regression**

**Performans Metrikleri (Test Seti)**

Çapraz doğrulama aşamasında elde edilen metrikler, modelin genelleme performansını ortaya koymaktadır. Ortalama sonuçlar şu şekildedir:

* **Doğruluk Skoru (Accuracy):** 0.8420

Accuracy, modelin doğru sınıflandırdığı örneklerin toplam örnek sayısına oranıdır. Modelin doğru tahmin yapma oranı oldukça yüksek olup, genel performansın tatmin edici olduğunu göstermektedir.

* **Kesinlik (Precision):** 0.7217

Precision, modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma oranıdır. Model, pozitif sınıf (1) için yaptığı tahminlerde orta düzeyde bir isabet oranına sahiptir.

* **Duyarlılık (Recall):** 0.4445

Duyarlılık metriği, modelin pozitif sınıfları tespit etmede bazı zorluklarla karşılaştığını göstermektedir. Yanlış negatif tahminlerin oranı bu metriği olumsuz etkileyen bir faktör olarak değerlendirilebilir.

* **F1 Skoru:** 0.5502

F1 skoru, precision ve recall arasındaki dengeyi yansıtmaktadır. Bu metrik, modelin genel başarımını dengeli bir şekilde değerlendirmektedir. Fakat bu modelde düşük çıkmıştır.

|  |  |
| --- | --- |
| **Metrik** | **Değer** |
| Accuracy | 0.8420 |
| Precision | 0.7217 |
| Recall | 0.4445 |
| F1-Score | 0.5502 |

**Detaylı Sınıflandırma Raporu**

**Negatif Sınıf (0):** Model negatif sınıfta %90 F1-Skoru ile oldukça başarılıdır. Negatif sınıfın hem yüksek recall (0.95) hem de yüksek precision (0.86) skorlarıyla iyi bir şekilde tahmin edildiği gözlemlenmiştir.

**Pozitif Sınıf (1):** Pozitif sınıf tahminlerinde orta seviyede precision (%72) oranı olmasına rağmen recall %44 seviyesinde kalmıştır. Bu durum, modelin pozitif sınıfı belirlemede kötü olduğunu ve eksik tahminler yaptığını göstermektedir.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sınıf (Label)** | **Precision** | **Recall** | **F1-Skor** | **Support** |
| 0 (Negatif) | 0.86 | 0.95 | 0.90 | 4573 |
| 1 (Pozitif) | 0.72 | 0.44 | 0.55 | 1289 |

**Karışıklık Matrisi Analizi**

**TP** (Doğru Pozitif): Modelin "temerrüt" olarak tahmin ettiği ve gerçekte de temerrüt olan örnekler. 573 temerrüt değeri bu modeldin performans düşüklüğünü göstermektedir.

**TN** (Doğru Negatif): Modelin "ödenmiş" olarak tahmin ettiği ve gerçekte de ödenmiş olan örnekler. Modelde 4352 değerine karşılık gelmektedir.

**FP** (Yanlış Pozitif): Modelin "temerrüt" olarak tahmin ettiği, ancak gerçekte ödenmiş olan örnekler. Bu modelde 221 gibi çok az bir değere karşılık gelmektedir.

**FN** (Yanlış Negatif): Modelin "ödenmiş" olarak tahmin ettiği, ancak gerçekte temerrüt olan örnekler. Modelde 716 ya denk gelmektedir. Bu değer test verisine oranladığımızda yüksek çıkmıştır bu da modelin iyi çalışmadığını gösterir.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

* + 1. **LightGBM Classification**

Çapraz doğrulama aşamasında elde edilen metrikler, modelin genelleme performansını ortaya koymaktadır. Ortalama sonuçlar şu şekildedir:

* **Doğruluk Skoru (Accuracy):** 0.9319

Accuracy, modelin doğru sınıflandırdığı örneklerin toplam örnek sayısına oranıdır. Modelin doğru tahmin yapma oranı oldukça yüksek olup, genel performansın tatmin edici olduğunu göstermektedir.

* **Kesinlik (Precision):** 0.9665

Precision, modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma oranıdır. Model, pozitif sınıf (1) için yaptığı tahminlerde yüksek bir isabet oranına sahiptir. Yanlış pozitif tahminlerin oranının düşük olması, modelin güvenilirliğini artırmaktadır.

* **Duyarlılık (Recall):** 0.8221

Duyarlılık metriği, modelin pozitif sınıfı doğru tespit etme oranını ifade etmektedir. Ve bu modelde yüksek çıkmıştır.

* **F1 Skoru:** 0.9429

F1 skoru, precision ve recall arasındaki dengeyi yansıtmaktadır. Bu metrik, modelin genel başarımını dengeli bir şekilde değerlendirmektedir.

**Performans Metrikleri (Test Seti)**

|  |  |
| --- | --- |
| **Metrik** | **Değer** |
| Accuracy | 0.9319 |
| Precision | 0.9665 |
| Recall | 0.8221 |
| F1-Score | 0.9429 |

**Detaylı Sınıflandırma Raporu**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sınıf (Label)** | **Precision** | **Recall** | **F1-Skor** | **Support** |
| 0 (Negatif) | 0.93 | 0.99 | 0.96 | 4573 |
| 1 (Pozitif) | 0.97 | 0.72 | 0.82 | 1289 |

**Negatif Sınıf (0):** Model negatif sınıfta %96 F1-Skoru ile oldukça başarılıdır. Negatif sınıfın hem yüksek recall (0.99) hem de yüksek precision (0.93) skorlarıyla iyi bir şekilde tahmin edildiği gözlemlenmiştir.

**Pozitif Sınıf (1):** Pozitif sınıf tahminlerinde, yüksek precision (%97) oranı olmasına rağmen recall %72 seviyesinde kalmıştır. Bu durum, modelin pozitif sınıfı belirlemede hassas fakat eksik tahminler yaptığını göstermektedir.

**Karışıklık Matrisi Analizi**

**TP** (Doğru Pozitif): Modelin "temerrüt" olarak tahmin ettiği ve gerçekte de temerrüt olan örnekler. 922 Temerrüt değeri veride az olduğu için bu şekilde çıkmıştır yani aslında yüksek bir performans göstermektedir.

**TN** (Doğru Negatif): Modelin "ödenmiş" olarak tahmin ettiği ve gerçekte de ödenmiş olan örnekler. Modelde 4541 değerine karşılık gelmektedir. Toplam test verisine oranladığımızda yüksek bir performans elde edildiği gözlemlenir.

**FP** (Yanlış Pozitif): Modelin "temerrüt" olarak tahmin ettiği, ancak gerçekte ödenmiş olan örnekler. Bu modelde 32 gibi çok az bir değere karşılık gelmektedir. Bu modelin performansının iyi olduğunu gösterir.

**FN** (Yanlış Negatif): Modelin "ödenmiş" olarak tahmin ettiği, ancak gerçekte temerrüt olan örnekler. Modelde 367 ye denk gelmektedir.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

**ROC Eğrisi ve AUC Analizi**

LightGBM (AUC = 0.94):

* LightGBM modeli, eğrisiyle daha geniş bir alan kapladığı için daha yüksek bir AUC değerine sahiptir.
* Bu, modelin sınıflandırma performansının oldukça iyi olduğunu ve pozitif sınıfı negatif sınıftan ayırt etmede başarılı olduğunu gösterir.

Logistic Regression (AUC = 0.84):

* Lojistik regresyonun AUC değeri LightGBM'e göre daha düşüktür.
* Bu model, pozitif ve negatif sınıfları ayırt etmekte LightGBM kadar etkili değildir ancak yine de iyi bir performans sergilemektedir.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

1. **Sonuçlar ve Tartışmalar**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-skor | AUC |
| Random Forest | 0.9318 | 0.9744 | 0.7085 | 0.8205 | 0.93 |
| Logistic Regresyon | 0.8420 | 0.7217 | 0.4445 | 0.5502 | 0.84 |
| LightGBM | 0.9319 | 0.9665 | 0.8221 | 0.9429 | 0.94 |
|  |  |  |  |  |  |

Bu proje kapsamında, kredi risk analizi için çeşitli makine öğrenimi modelleri karşılaştırılmıştır. Logistic Regression ve LightGBM modelleri detaylı şekilde değerlendirilmiştir. Sonuçlar, özellikle LightGBM modelinin diğer modellere göre üstün bir performans sergilediğini göstermektedir.

**4.1 Performans Karşılaştırması**

* **LightGBM**: En yüksek doğruluk (Accuracy: %93.19), precision (%96.65), ve F1-Score (%94.29) değerlerine ulaşmıştır. Aynı zamanda modelin ROC eğrisi altında kalan alan (AUC: 0.94) oldukça yüksektir, bu da sınıflandırma başarısının güçlü olduğunu göstermektedir.
* **Logistic Regression**: Diğer modellere kıyasla en düşük performansı sergilemiştir. Özellikle recall (%44.45) ve F1-Score (%55.02) değerlerinin düşük olması, bu modelin sınıf dengesizliğinden etkilenme potansiyelini ortaya koymaktadır.

**4.2 Önemli Bulgular**

1. **En Önemli Değişkenler**:
   1. **loan\_percent\_income (kredi/gelir oranı)**: Temerrüt olasılığına en çok katkıda bulunan değişkenlerden biri olarak belirlenmiştir. Bu oran arttıkça, temerrüde düşme ihtimali de artmaktadır.
   2. **loan\_grade (kredi notu)**: Yüksek kredi notuna sahip bireylerin temerrüde düşme ihtimali düşük bulunmuştur.
   3. **person\_income (kişinin geliri)**: Daha yüksek gelir düzeyi, kredi ödeme kapasitesini artırmaktadır.
2. **Etkisiz Değişkenler**:
   1. **loan\_intent (kredi amacı)**: Kredi alma sebebinin temerrüt olasılığı üzerinde anlamlı bir etkisi gözlenmemiştir.
   2. **person\_age (kişinin yaşı)**: Yaş değişkeninin kredi temerrüdü üzerinde dolaylı bir etkisi olduğu görülmüş, ancak kredi geçmişi uzunluğu bu konuda daha belirleyici bir faktör olarak kabul edilmiştir.
3. **LightGBM’nin Başarısı**:
   1. LightGBM modeli, özellikle verideki dengesizliklere karşı daha az hassas olması ve matematiksel yapısının güçlü olması nedeniyle diğer modellerden daha iyi performans göstermiştir.

**4.3 Öneriler**

* Finansal kuruluşlar, kredi riskini yönetmek için LightGBM gibi ileri makine öğrenimi tekniklerini kullanarak daha doğru ve etkili tahminler yapabilir.
* Model performansını artırmak için, veri setindeki sınıf dengesizliğini ele almak üzere SMOTE veya benzeri teknikler kullanılabilir.
* Daha geniş bir veri setiyle modelin genelleme kapasitesini artırmak mümkündür.

**Referanslar**

Abellán, J., & Castellano, F.J. (2017). A comparative study on base classifiers in ensemble

methods for credit scoring. Expert Syst. Appl., 73, 1-10.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate

bankruptcy. The Journal of Finance, 23(4), 589–609.

Altman, E.I., and Sabato, G. (2007). Modelling Credit Risk for SMEs: Evidence from The

US Market. ABACUS, 43 (3), 332-357.

Breiman, Leo. 2000. Some Infinity Theory for Predictors Ensembles. Technical Report;

Berkeley: UC Berkeley, vol. 577.

Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, and Akhtar

Siddique. 2016. Risk and risk management in the credit card industry. Journal of Banking

and Finance 72: 218–39.

Chen, T.; Guestrin, C., (2016), "XGBoost: A Scalable Tree Boosting System". 22nd ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16.

Gahlaut, A., Tushar, & Singh, P.K. (2017). Prediction analysis of risky credit using Data

mining classification models. 2017 8th International Conference on Computing,

Communication and Networking Technologies (ICCCNT), 1-7.

Galindo, Jorge, and Pablo Tamayo. 2000. Credit risk assessment using statistical and

machine learning: Basic methodology and risk modeling applications. Computational

Economics 15: 107–43.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017).

Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural

information processing systems, 30.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. Consumer credit-risk models

via machine-learning algorithms. Journal of Banking and Finance 34: 2767–87

Maruma, C., Tu, C., Nawej, C. (2022). Banking Credit Risk Analysis using Artificial

Neural Network. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of

Seventh International Congress on Information and Communication Technology. Lecture

Notes in Networks and Systems, vol 447. Springer, Singapore.

https://doi.org/10.1007/978-981-19-1607-6\_76

Pandey, T. N., Jagadev, A. K., Mohapatra, S. K., & Dehuri, S. (2017, August). Credit risk

analysis using machine learning classifiers. In 2017 International Conference on Energy,

Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1850-1854). IEEE.

Satchidananda, S. S., & Simha, J. B. (2006). Comparing decision trees with logistic

regression for credit risk analysis. International Institute of Information Technology,

Bangalore, India.

Siddiqi, N. (2006). Credit Risk Scorecards, Developing and Implementing Intelligent

Credit Scoring. Hoboken, NJ: John Wiley & Sons, Inc.

Yazdanfar, D., and Nilsson, M. (2008). The Bankruptcy Determinants of Swedish SMEs.

Institute for Small Business and Entrepreneurship, Belfast, Ireland.

Zdravevski, E., Lameski, P., Kulakov, A., & Gjorgjevikj, D. (2014). Feature selection and

allocation to diverse subsets for multi-label learning problems with large datasets. 2014

Federated Conference on Computer Science and Information Systems, 387-394.