**Titanic Yolcu Hayatta Kalma Tahmini - Makine Öğrenmesi (Dataiku)**

**Proje Raporu**

Bu proje, **Titanic yolcularının hayatta kalma olasılığını tahmin etmek** için **makine öğrenmesi** yöntemlerini kullanmaktadır. Model, **Dataiku DSS** platformunda oluşturulmuş olup veri hazırlama, model eğitimi ve değerlendirme aşamalarını içermektedir.

Bu proje, Titanic yolcularının hayatta kalma olasılığını tahmin etmek için **Random Forest modeli** kullanılarak gerçekleştirilmiştir.

**Amacımız, veri ön işleme, model eğitimi ve değerlendirme süreçlerini uygulayarak Random Forest modelinin bu tür bir problemde nasıl performans gösterdiğini incelemektir.**

Model, **başarıyla eğitilmiş** ve tatmin edici sonuçlar elde edilmiştir.

**Veri Seti**

Projede **Kaggle Titanic veri seti** kullanılmıştır. Bu veri seti, yolcuların yaş, cinsiyet, bilet sınıfı, yanında bulunan aile üyeleri ve bilet ücreti gibi bilgilerini içermektedir.

Veri setinde şu dosyalar yer almaktadır:

* **train.csv** → Modelin eğitimi için kullanılan etiketli veriler (Survived: 0 = Hayatta kalmadı, 1 = Hayatta kaldı)
* **test.csv** → Tahmin yapılacak veri seti (etiketsiz)
* **predictions.csv** → Model tarafından üretilen tahmin sonuçları

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Doğruluk (Accuracy)** |  |  | | --- | |  | | |  | | --- | | 0.846 |  |  | | --- | |  | |
| |  | | --- | | **Kesinlik (Precision)** |  |  | | --- | |  | | |  | | --- | | 0.769 |  |  | | --- | |  | |
| |  | | --- | | **Duyarlılık (Recall)** |  |  | | --- | |  | | |  | | --- | | 0.877 |  |  | | --- | |  | |
| |  | | --- | | **F1-Skoru** |  |  | | --- | |  | | 0.820 |
| |  | | --- | | **Maliyet Matrisi Kazancı** |  |  | | --- | |  | | |  | | --- | | 0.318 |  |  | | --- | |  | |
| |  | | --- | | **ROC AUC Skoru** |  |  | | --- | |  | | |  | | --- | | 0.889 |  |  | | --- | |  | |
| |  | | --- | | **Lift Skoru** |  |  | | --- | |  | | |  | | --- | | 2.025 |  |  | | --- | |  | |
| |  | | --- | | **Ortalama Kesinlik** |  |  | | --- | |  | | |  | | --- | | 0.887 |  |  | | --- | |  | |
| |  | | --- | | **Log Kayıp (Log Loss)** |  |  | | --- | |  | | |  | | --- | | 0.391 |  |  | | --- | |  | |
| |  | | --- | | **Kalibrasyon Kaybı** |  |  | | --- | |  | | 0.128 |

* **Doğruluk (Accuracy):** 0.846  
  → Modelin tüm tahminlerinin %84.6'sının doğru olduğunu gösteriyor. Titanic veri seti gibi **dengesiz (imbalanced) olmayan** bir veri seti için iyi bir değer.
* **Kesinlik (Precision):** 0.769  
  → Modelin **hayatta kalma** sınıfına yaptığı pozitif tahminlerin %76.9’unun gerçekten doğru olduğunu gösteriyor. Yani model, **false positive (yanlış pozitif) tahminler yapmaya meyilli** olabilir.
* **Duyarlılık (Recall):** 0.877  
  → Gerçekten **hayatta kalan** yolcuların %87.7’sini doğru tahmin edebilmiş. Yüksek olması iyi bir şey çünkü **hayatta kalanları kaçırma olasılığı düşük.** Ama precision daha düşük olduğu için, bazı yanlış pozitifler olabilir.
* **F1-Skoru:** 0.820  
  → Precision ve Recall arasındaki dengeyi gösterir. Modelin genel performansı fena değil ama **precision daha düşük olduğu için, belki yanlış pozitifleri fazla veriyor.**
* **Maliyet Matrisi Kazancı (Cost Matrix Gain):** 0.318  
  → Bu, yanlış tahminlerin **ne kadar maliyetli olduğunu** gösterir. Çok yüksek bir değer değil ama **yanlış tahminlerin modelin başarısını azalttığını** gösteriyor.
* **ROC AUC Skoru:** 0.889  
  → Modelin **sınıfları ayırma yeteneğini** gösterir. 0.889 gibi bir değer, **gayet başarılı** olduğunu gösteriyor. **Mükemmel bir model 1.0 olurdu, yani model çok iyi bir ayrım yapıyor.**
* **Lift:** 2.025  
  → Modelin **rastgele tahminden kaç kat daha iyi olduğunu** gösterir. 2.025 değeri, modelin **iki kat daha iyi tahminde bulunduğunu** gösterir. Yani tamamen rastgele bir modelden **çok daha iyi.**
* **Ortalama Kesinlik (Average Precision):** 0.887  
  → Bu metrik, modelin tahmin yaparken **genel doğruluğunu** ölçer. 0.887 gayet iyi bir değer.
* **Log Loss:** 0.391  
  → Düşük olması iyidir, çünkü **modelin tahminlerinde ne kadar belirsizlik olduğunu** gösterir. 0.391 çok yüksek değil, model iyi çalışıyor.
* **Kalibrasyon Kaybı (Calibration Loss):** 0.128  
  → Modelin **tahmin edilen olasılıklar ile gerçek sınıfların ne kadar uyumlu olduğunu** gösterir. 0.128 düşük bir değer olduğu için modelin olasılıkları **gerçek sınıflarla büyük ölçüde uyumlu.**

### ****Genel Yorum:****

* Modelin **duyarlılığı (recall) yüksek**, yani **hayatta kalanları kaçırma ihtimali düşük.**
* **Precision biraz düşük** (0.769), yani **yanlış pozitif oranı biraz yüksek olabilir.** Model **hayatta kalmayanları yanlışlıkla hayatta kaldı olarak tahmin edebilir.**
* **ROC AUC Skoru (0.889) ve Lift Skoru (2.025) çok iyi**, yani **model sınıfları iyi ayırabiliyor.**
* **Log Loss düşük (0.391)**, yani **tahminlerde fazla belirsizlik yok.**
* **Kalibrasyon kaybı (0.128) düşük**, yani **modelin olasılık tahminleri güvenilir.**

**Özetle:**  
Model **genel olarak başarılı**, ancak **precision biraz düşük** olduğu için **yanlış pozitifleri fazla verebilir.** Eğer **yanlış pozitifler kritikse (örneğin, birine yanlışlıkla hayatta kalmış demek büyük bir hata yaratıyorsa),** modeli **daha yüksek precision verecek şekilde ayarlayabiliriz.** Ama şu anki haliyle **gayet iyi çalışıyor.**