VERİ MADENCİLİĞİ (FET445)

MÜŞTERİ ŞİKAYETLERİNDEN İTİRAZ EDİLİP EDİLMEYECEĞİ TAHMİNİ

AVENGERS

Barış Yasin Şahin	22040301029
Musa Uluğ	21040301044
Yusuf Yenigün	21040301052
Salih İmran Büker	22040301062

Youtube Linki: https://youtu.be/WvYJvjAiAwA?si=1rGs9oLxjtRuOpd6

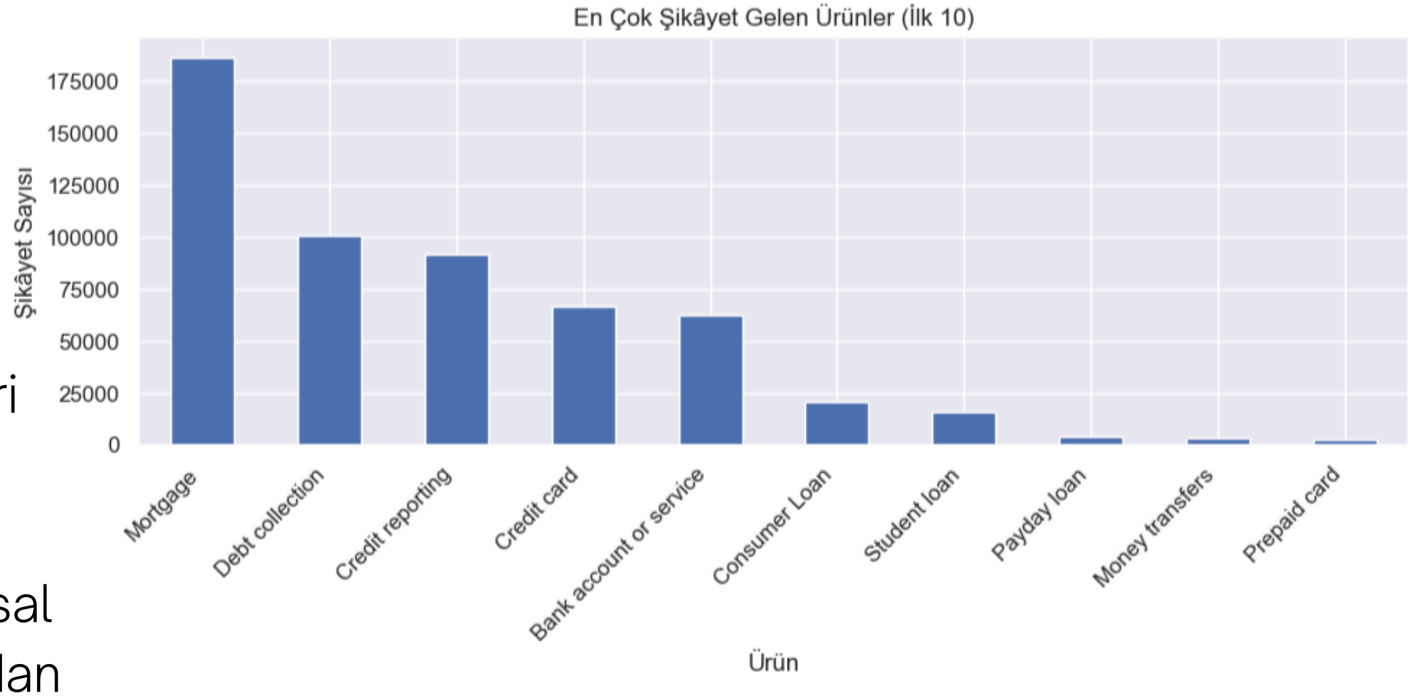
08.01.2026

PROJE AMACI

Günümüzde finansal kurumlar ve kamu kuruluşları, müşterilerden gelen çok sayıda şikâyeti manuel olarak incelemekte ve bu şikâyetlerin itirazla sonuçlanıp sonuçlanmayacağını önceden öngörememektedir. Bu durum, hem müşteri memnuniyetinin azalmasına hem de operasyonel maliyetlerin artmasına yol açmaktadır. Bu projede amaç, **Consumer Complaint Database** veri seti kullanılarak, **bir müşteri şikâyetinin itiraz edilip edilmeyeceğini** önceden tahmin edebilen bir makine öğrenmesi modeli geliştirmektir. Problem, hedef değişkenin iki sınıftan oluşması nedeniyle **ikili sınıflandırma (binary classification)** problemi olarak ele alınmıştır. Geliştirilen model ile, özellikle itiraz edilme riski yüksek olan şikâyetlerin erken aşamada tespit edilmesi ve kurumların bu şikâyetlere daha hızlı ve etkin şekilde müdahale edebilmesi hedeflenmektedir.

VERİ SETİMİZ HAKKINDA

Bu çalışmada kullanılan veri seti, Amerika Birleşik Devletleri'ndeki finansal kurumlara iletilen müşteri şikâyetlerinden oluşan **Consumer Complaint Database'tir**. Veri setinin amacı, bir müşteri şikâyetinin süreç sonunda **itiraz edilip edilmeyeceğini tahmin etmektir.** Toplamda **555.957 gözlem** ve **21 özellikten** oluşan veri seti, hem kategorik hem de sayısal değişkenleri içeren karmaşık bir yapıya sahiptir. Kategorik özellikler arasında şikâyet konusu, ürün türü, kurum bilgisi ve başvuru kanalı yer alırken; sayısal özellikler zamanla ilişkili bazı özetleyici alanlardan oluşmaktadır. Gerçek hayat verisi olması nedeniyle veri seti **sınıf dengesizliği** içermekte olup, itiraz edilmeyen şikâyetler çoğunluğu oluşturmaktadır.



Veri Seti Linki: https://www.kaggle.com/datasets/kaggle/us-consumer-finance-complaints

DATA SETİNDE TRAIN / TEST AYRIMI

Data Seti Toplam: 555.957

Train seti: 444.765 (%80)

Test seti: 111.192 (%20)

Stratified split kullandık...

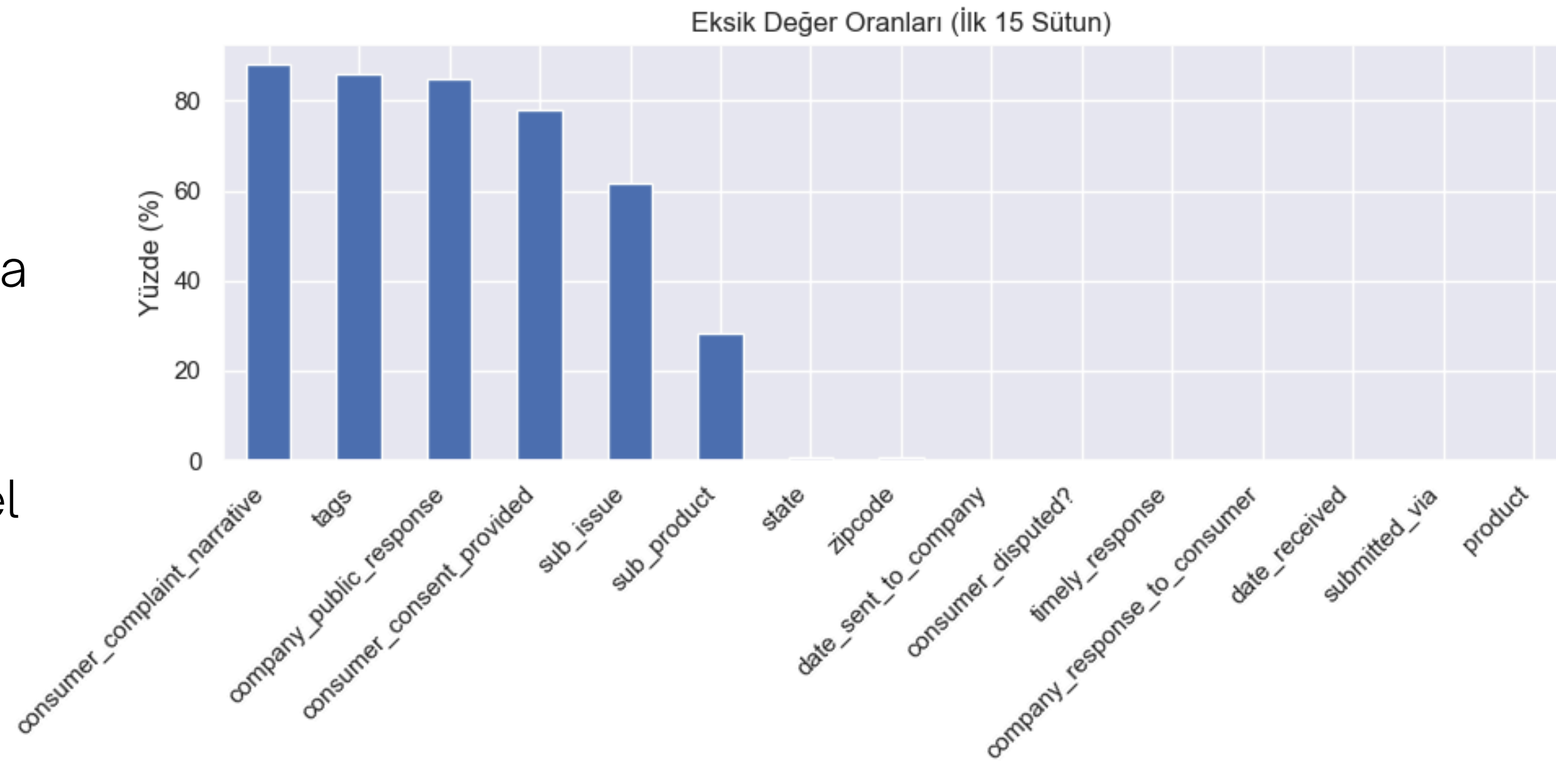
Veriyi %80 eğitim, %20 test olacak şekilde ayırarak modelin genelleme performansını ölçtük

SINIF DAĞILIMI

Hedef Değişken Dağılımı

- Veri seti dengesiz
- İtiraz edilmeyen şikayetler çoğunlukta

Verinin dengesiz olması nedeniyle model seçiminde ve değerlendirmede özel kararlar aldık.



YAKLAŞIM VE TASARIM KARARLARI

Bu çalışmada, müşteri şikâyetlerine ait veriler kullanılarak itiraz edilme durumunun tahmin edilmesi amacıyla denetimli öğrenme temelli bir sınıflandırma yaklaşımı benimsenmiştir. Veri setinin hem kategorik hem de sayısal değişkenler içermesi nedeniyle, modelleme sürecinde veri ön işleme adımları büyük önem taşımıştır. Bu kapsamda eksik veriler uygun yöntemlerle ele alınmış, kategorik değişkenler sayısal forma dönüştürülmüş ve tüm bu işlemler **Pipeline** yapısı içerisinde birleştirilerek veri sızıntısının (data leakage) önüne geçilmiştir. Gerçek hayat verisi olması nedeniyle hedef değişkende belirgin bir **sınıf dengesizliği** bulunduğundan, model eğitiminde itiraz sınıfını kaçırmamak amacıyla yaklaşımı tercih edilmiştir. Veriler, modelin genelleme performansını objektif biçimde değerlendirebilmek için %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Model performansının değerlendirilmesinde yalnızca doğruluk (accuracy) metriğine bağlı kalınmamış; dengesiz sınıf yapısı göz önünde bulundurularak **precision, recall, F1 skoru ve ROC AUC** gibi metrikler birlikte kullanılmıştır. Bu tasarım kararları sayesinde, geliştirilen modellerin hem daha adil karşılaştırılması hem de problem bağlamında daha anlamlı sonuçlar üretmesi hedeflenmiştir.

KULLANDIĞIMIZ MODELLER

- Dummy Classifier (Baseline karşılaştırma)
- Logistic Regression (class_weight = balanced)
- Decision Tree Classifier
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM / LinearSVC)

Her model, aynı eğitim–test bölünmesi kullanılarak adil biçimde değerlendirilmiştir.

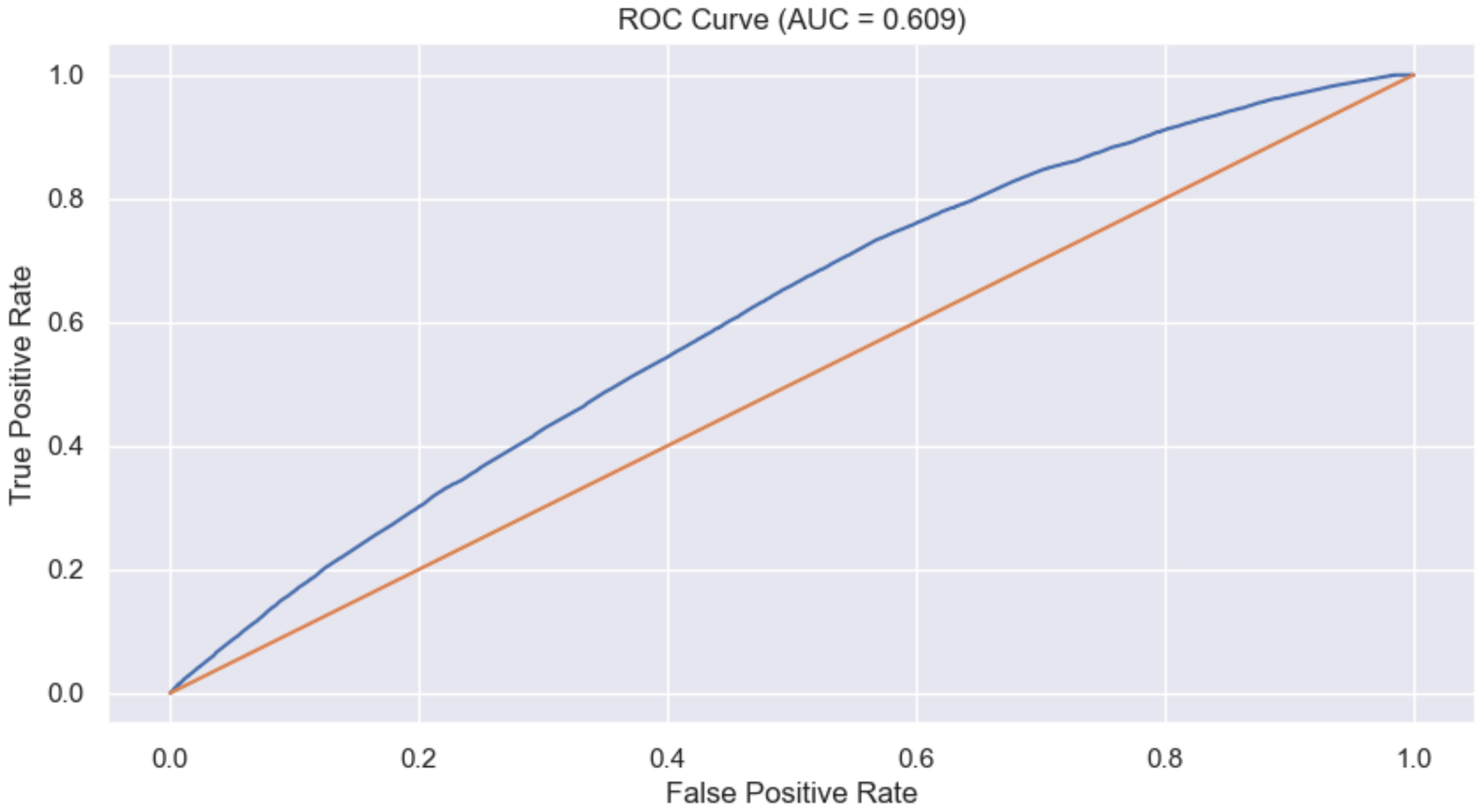
KULLANDIĞIMIZ MODELLERİN KARŞILAŞTIRMASI

Ekip Üyesi	Model	Accuracy	ROC AUC	Recall (Yes)	F1 Score
Barış	Dummy (most frequent)	0.7983	0.5	0	0
Barış	LogReg (balanced)	0.5144	0.6092	0.6927	0.365
Barış	Decision Tree (balanced)	0.4703	0.5992	0.7587	~0.48
Yusuf	**Decision Tree (balanced, extended)**	0.5577	**0.6253**	0.6444	**~0.46**
Yusuf	Decision Tree (extended + FS)	~0.55	~0.62	~0.63	~0.44
Yusuf	LogReg (balanced, extended)	~0.55	~0.62	~0.64	~0.45
Musa	ComplementNB – full	0.7406	0.597	0.1305	~0.22
Musa	ComplementNB + chi² (k=50)	0.7418	0.5917	0.1302	~0.22
Musa	KNN (k=5)	0.7526	–	0.1051	~0.19
Salih	LinearSVC (balanced)	0.514	–	0.6932	~0.36
Salih	LogReg (balanced, C=0.5)	0.5122	0.6092	0.697	~0.37
Salih	LogReg + mutual_info (k=200)	0.5122	0.6092	0.697	~0.37

ROC CURVE

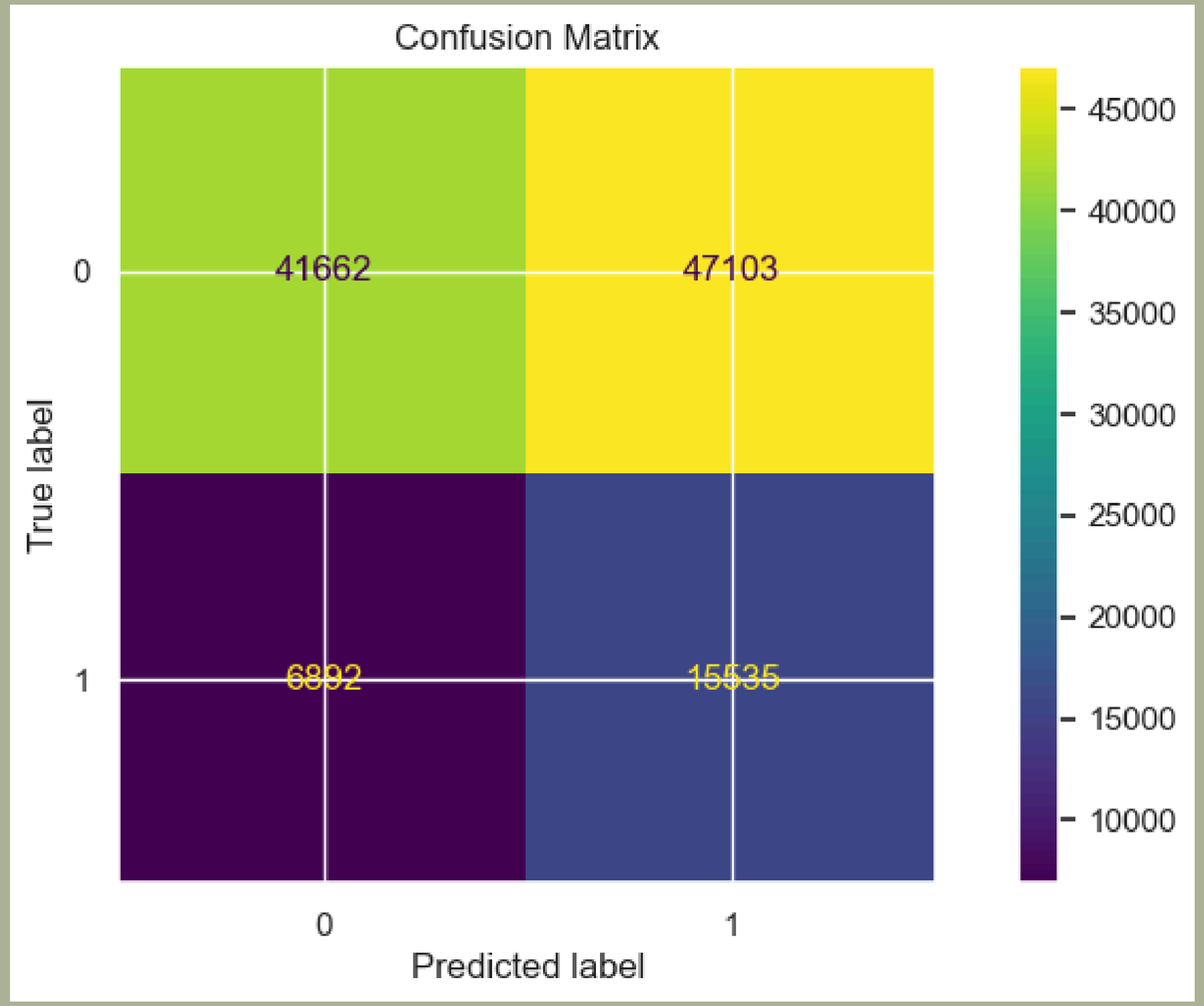
ROC CURVE – EN İYİ MODEL: DECISION TREE (BALANCED, EXTENDED) ROC AUC ≈ 0.6253

ROC EĞRISI, MODELIN RASTGELE TAHMINE KIYASLA SINIFLARI AYIRT EDEBILDIĞINI GÖSTERMEKTEDIR.



CONFUSION MATRIX

Confusion matrix, sınıflandırma modelinin doğru ve yanlış tahminlerini detaylı olarak analiz etmeyi sağlayan bir değerlendirme yöntemidir. Bu çalışmada elde edilen confusion matrix incelendiğinde, modelin **itiraz edilen şikâyetleri (pozitif sınıf)** büyük ölçüde doğru tahmin edebildiği, ancak bu durumun bazı **yanlış pozitif (false positive)** tahminlere yol açtığı görülmektedir. Bu davranış, veri setindeki sınıf dengesizliği nedeniyle modelin itiraz sınıfını kaçırmamaya öncelik vermesinden kaynaklanmaktadır. Dolayısıyla confusion matrix, modelin yüksek recall değerine karşılık precision'da yaşanan düşüşü görsel ve anlaşılır biçimde ortaya koymaktadır.



EN BAŞARILI MODEL VE GEREKÇE

DECISION TREE (BALANCED, EXTENDED)

NEYE GÖRE EN BAŞARILI?

ROC AUC: 0.6253 → TABLODA EN YÜKSEK

RECALL (YES): 0.6444 → ITIRAZ SINIFINI YAKALAMA GÜCÜ YÜKSEK

F1 ≈ 0.46 → PRECISION–RECALL DENGESI IYI

DENGESIZ VERI IÇIN DOĞRU METRIKLERDE ÖNDE

DEĞER VE SONUÇ

Bu çalışmada, gerçek hayat verisi içeren ve sınıf dengesizliği barındıran bir müşteri şikâyet veri seti üzerinde farklı makine öğrenmesi modelleri sistematik olarak karşılaştırılmıştır. Elde edilen sonuçlar, model başarısının yalnızca doğruluk (accuracy) metriği ile değerlendirilmesinin yetersiz olduğunu; özellikle **recall, F1 skoru ve ROC AUC** gibi metriklerin **dengesiz veri problemlerinde daha anlamlı sonuçlar** sunduğunu göstermiştir. Yapılan karşılaştırmalar sonucunda, **genişletilmiş özellik seti kullanan Decision Tree** modeli, itiraz edilen şikâyetleri yakalama başarısı ve ayırt edicilik gücü açısından en dengeli performansı sergilemiştir. Bu çalışma, müşteri itiraz riskinin erken aşamada tespit edilmesine yönelik karar destek sistemleri için temel bir makine öğrenmesi yaklaşımı sunmaktadır.