ANADOLU HAYAT EMEKLİLİK

ANADOLU HAYAT EMEKLİLİK DATATHON

FİNAL SUNUMU

Anıl Öztürk - İsmail Denizli - Ahmet Tarık Karakaş

ANADOLU HAYAT EMEKLİLİK

PROBLEM

Müşteri 2021 yılının ilk çeyreğinde katkı payı artışı talebinde bulunacak mı?

- 2020'ye ait vade ve ödeme dökümleri verilmiş
- Müşterilere ait kategorik veriler mevcut
- Başarı oranı **F1-Score** ile değerlendiriliyor

LİTERATÜR

MEVCUT YAKLAŞIMLAR

1. KULLANILAN ÖZGÜN VERİLER

- Kredi skoru
- Emeklilik anındaki katkı oranı
- Emeklilik anındaki borsa durumu
- Sağlık durumu
- İndirim oranları
- Eş bilgileri

2. TAHMİN MODELİ TÜRÜ

- Logistic Regression
- SVM
- Karar ağacı ailesi
- Yapay sinir ağları

- https://livrepository.liverpool.ac.uk/3074450/1/Anales%2019-rocha-boado.pdf
- http://laccei.org/LACCEI2019-MontegoBay/full_papers/FP343.pdf
- https://www.actuarios.org/wp-content/uploads/2019/12/Art6-Anales2019.pdf
- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3690990
- https://www.sciencedirect.com/science/article/pii/S2352827317302331

ANADOLU HAYAT EMEKLİLİK

ÇÖZÜM AŞAMALARI

VERİ ANALİZİ & ÖN İŞLEME

Verideki dağılımların incelenmesi ve formatın standardize edilmesi

MODEL GELİŞTİRME SÜRECİ

Yarışma boyunca veri ve model üzerinde yapılan geliştirme ve iyileştirmeler

DENENEN ALTERNATİFLER

Çözüm için denenen alternatif metotlar, veri manipülasyonları ve teknikler

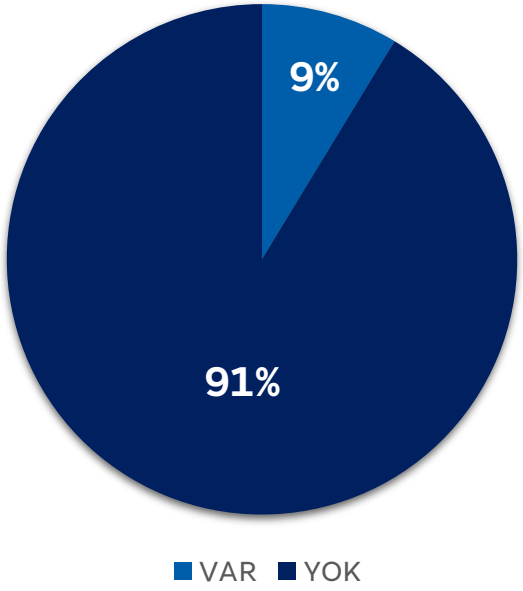
ÇIKARIMLAR

Veriden çıkarılan içgörü, problemin kurumsal bağlamda yorumlanması

VERİ ANALİZİ

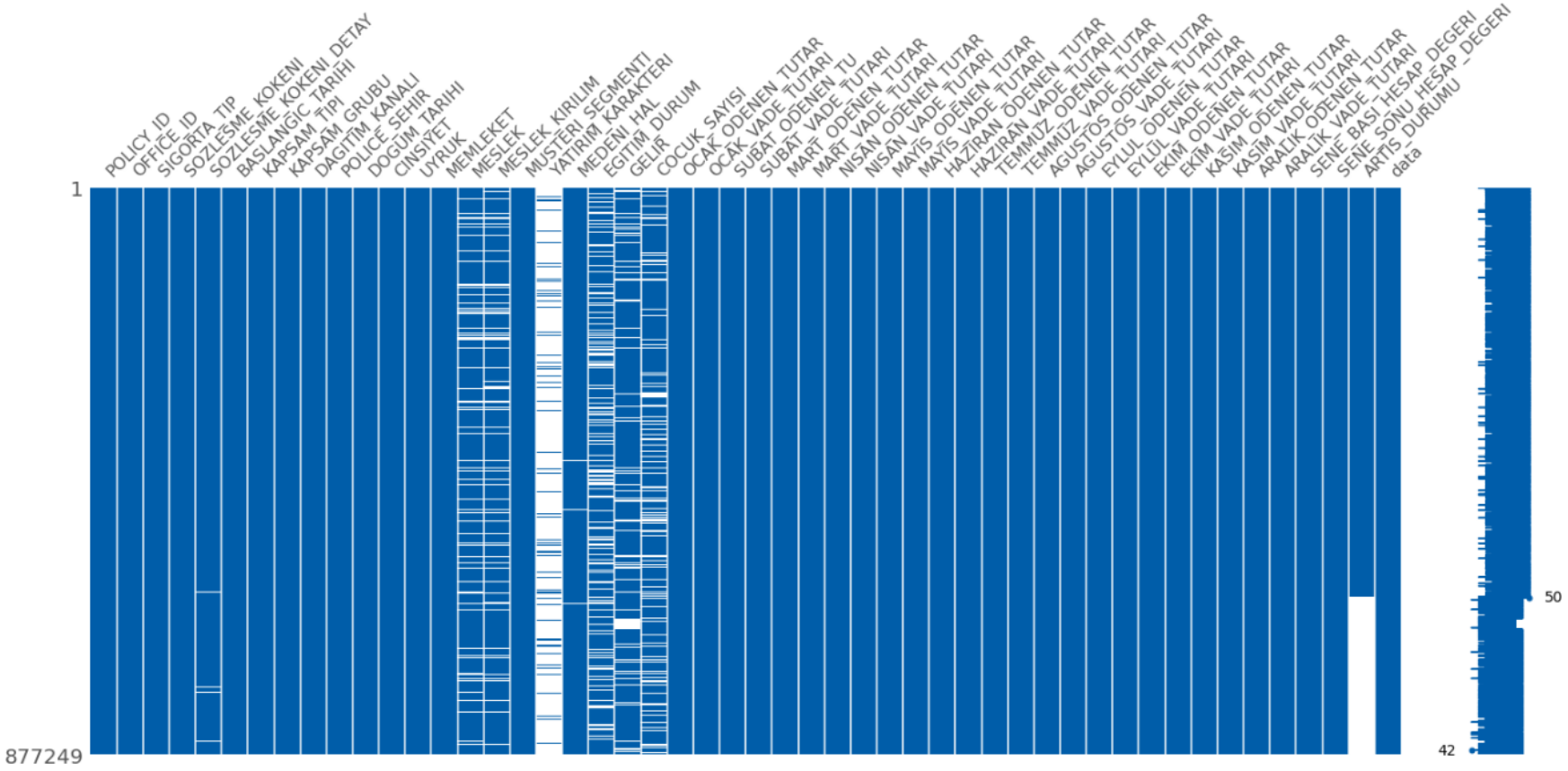
CLASS IMBALANCE

2021’in ilk çeyreğinde vade tutarında artış



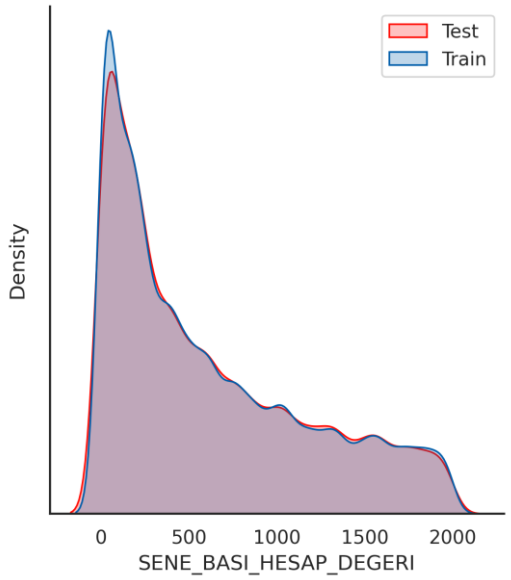
VERİ ANALİZİ

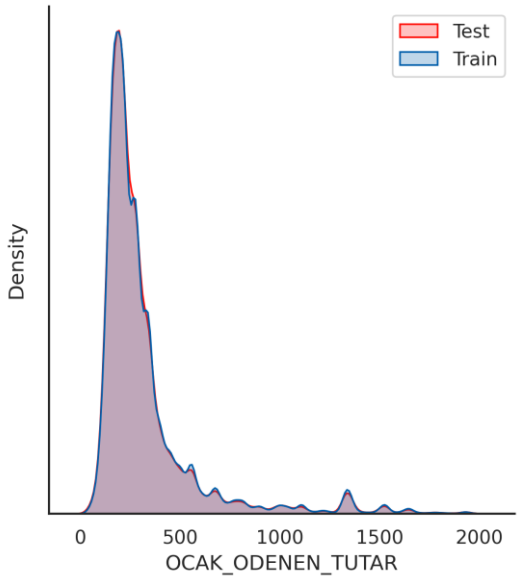
TANIMSIZ DEĞERLER

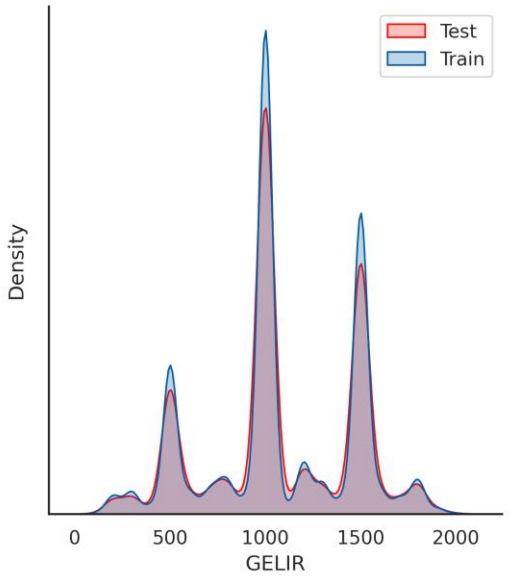


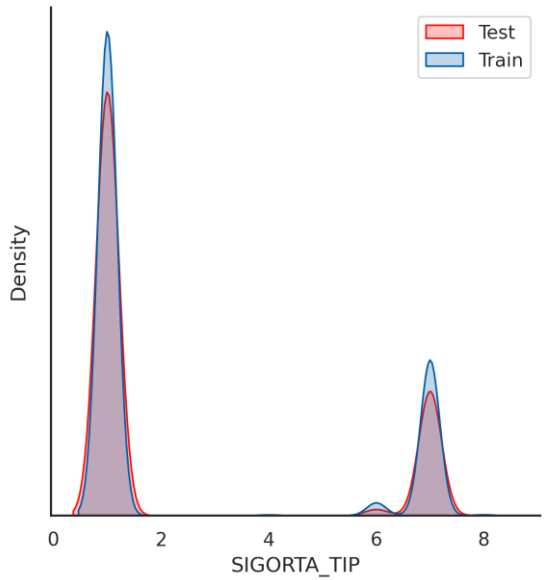
VERİ ANALİZİ

TRAIN-TEST GİRDİ DAĞILIMLARI



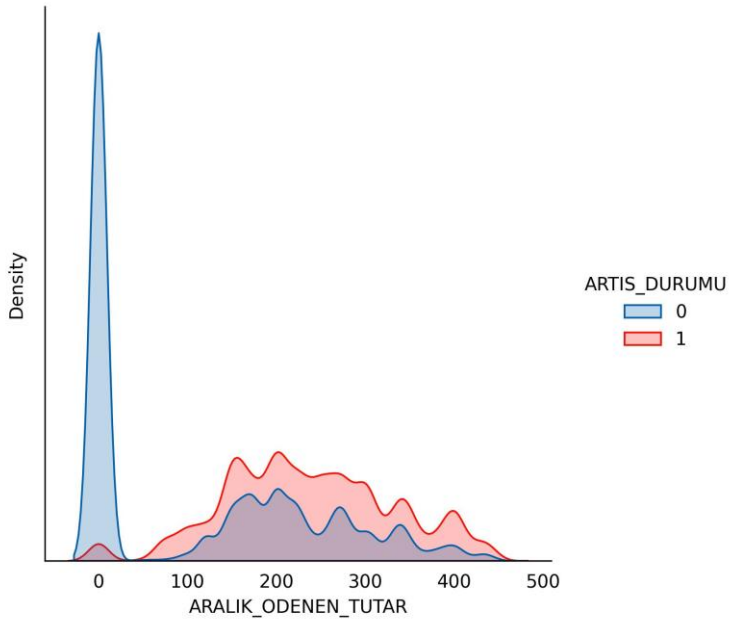


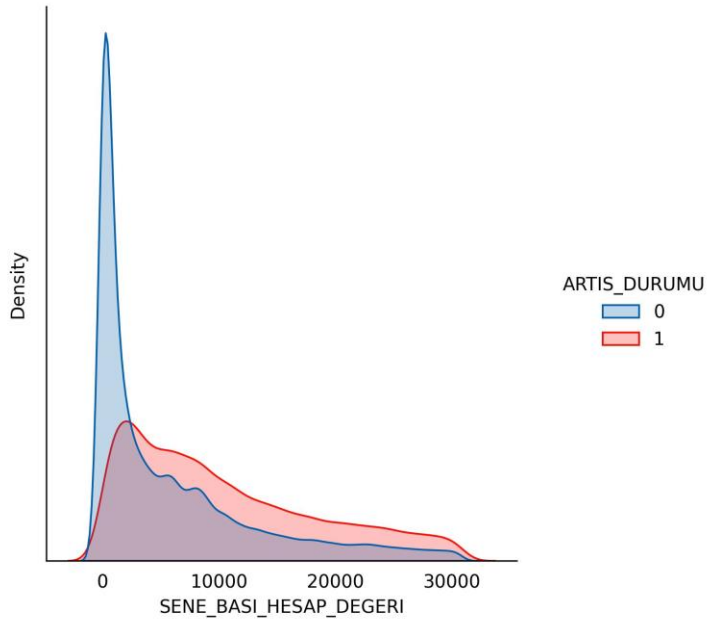


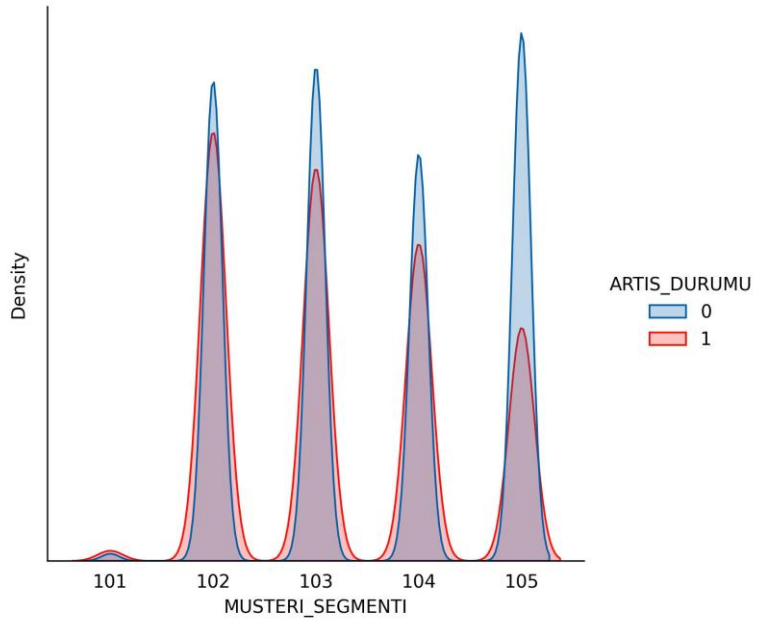


VERİ ANALİZİ

ETİKET DEĞERİNE GÖRE VERİ DAĞILIMLARI







VERİ ANALİZİ

ÖZGÜN DEĞERLER

Değişken	Özgün Değer Sayısı	
	Train	Test
OFFICE ID	1749	1700
POLİÇE ŞEHRİ	603	288
KAPSAM TİPİ	239	206
MESLEK KIRILIMI	87	87
UYRUK	80	64
MESLEK	32	32
DAĞITIM KANALI	20	20
SÖZLEŞME KÖKENİ DETAYI	10	8
KAPSAM GRUBU	10	10

+ ‘DİĞER’

ANADOLU HAYAT EMEKLİLİK

VERİ ANALİZİ

VERİSETLERİ ARASI ÖZGÜNLÜK

Değişken	Sadece	
	Train Verisinde Bulunan	Test Verisinde Bulunan
POLİÇE ŞEHRİ	378	63
OFFICE ID	57	8
KAPSAM TİPİ	36	3
UYRUK	21	5
SÖZLEŞME KÖKENİ DETAYI	2	
MÜŞTERİ SEGMENTİ	1	1
SÖZLEŞME KÖKENİ	1	

→ NaN

MODEL SEÇİMİ

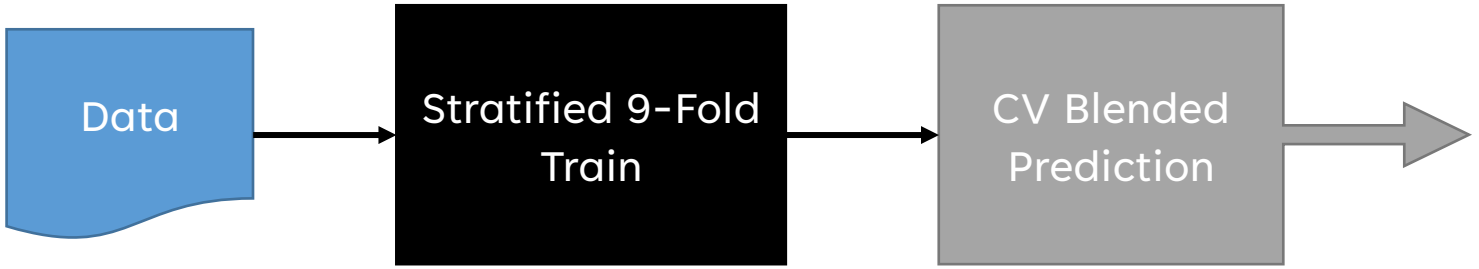
- CatBoost
- LGBM
- XGBoost
- TabNet
- HistGradientBoostingClassifier

MODEL SEÇİMİ

- **CatBoost**
- LGBM
- XGBoost
- TabNet
- HistGradientBoostingClassifier

CatBoost

MODEL GELİŞTİRME SÜRECİ



Public F1: 0,26039

Private F1: 0,25972

ANADOLU HAYAT EMEKLİLİK

MODEL GELİŞTİRME SÜRECİ

CLASS IMBALANCE + HYPERPARAMETER TUNING

- **Class Imbalance:** ‘scale_pos_weight’
- Grid-search ile ana model parametrelerinin belirlenmesi

Public F1: 0,40268

Private F1: 0,41039

MODEL GELİŞTİRME SÜRECİ

SAYISAL FEATURE EXTRACTION

- **Ödeme ve vade verileri üzerinden istatistiksel özetler**

(max, min, mean, std, var, quantile, skewness, kurtosis)

- **Vade ve ödeme için tutarda artış görülen ay sayısı**
- **Ödeme yapılmış ay sayısı**
- **Ödeme ve vadede en son artış gerçekleşen ay**
- **Sene sonu ve başı arasındaki hesap değeri farkı**
- **Poliçe başlangıcının ay ve yılı**
- **Poliçe başlangıcındaki müşteri yaşı**

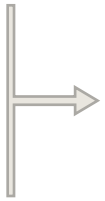
Public F1: 0,42183

Private F1: 0,42753

MODEL GELİŞTİRME SÜRECİ

KATEGORİK KOMBİNASYONLAR

- **Doğum tarihi**
- **Meslek**
- **Medeni Hal**
- **Eğitim Durumu**
- **Uyruk**
- **Memleket**



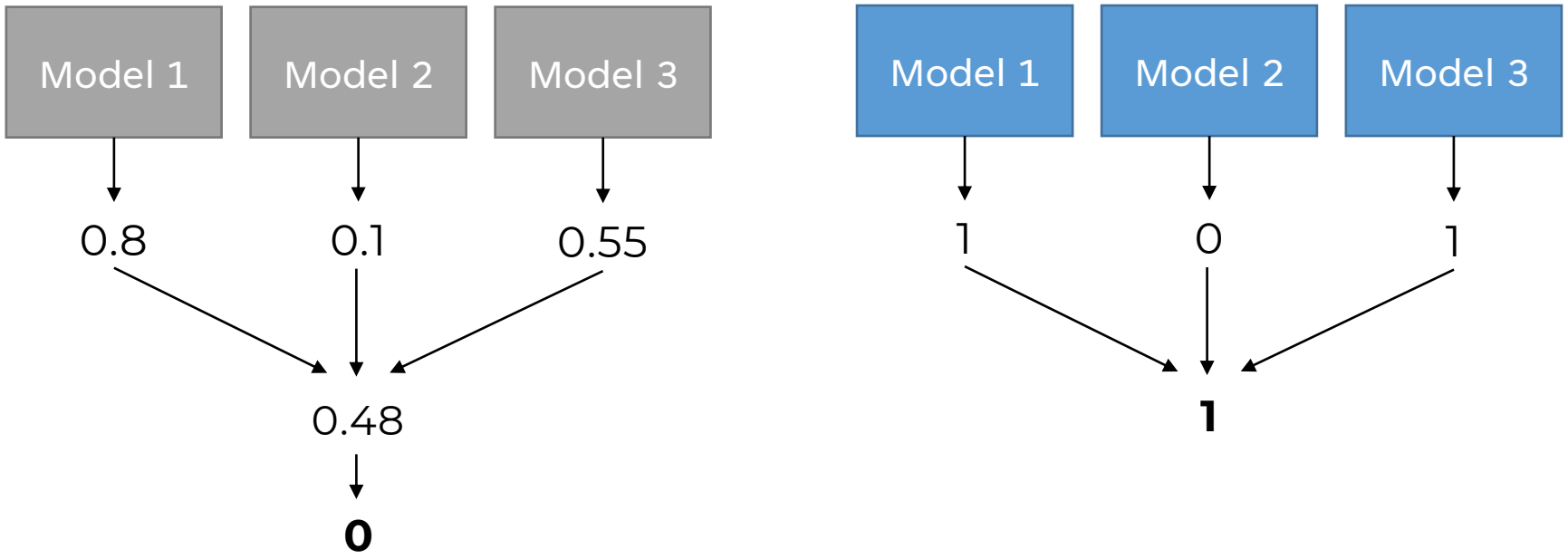
Doğum tarihi + Meslek + Medeni Hal

Public F1: 0,45324

Private F1: 0,45745

MODEL GELİŞTİRME SÜRECİ

VOTING

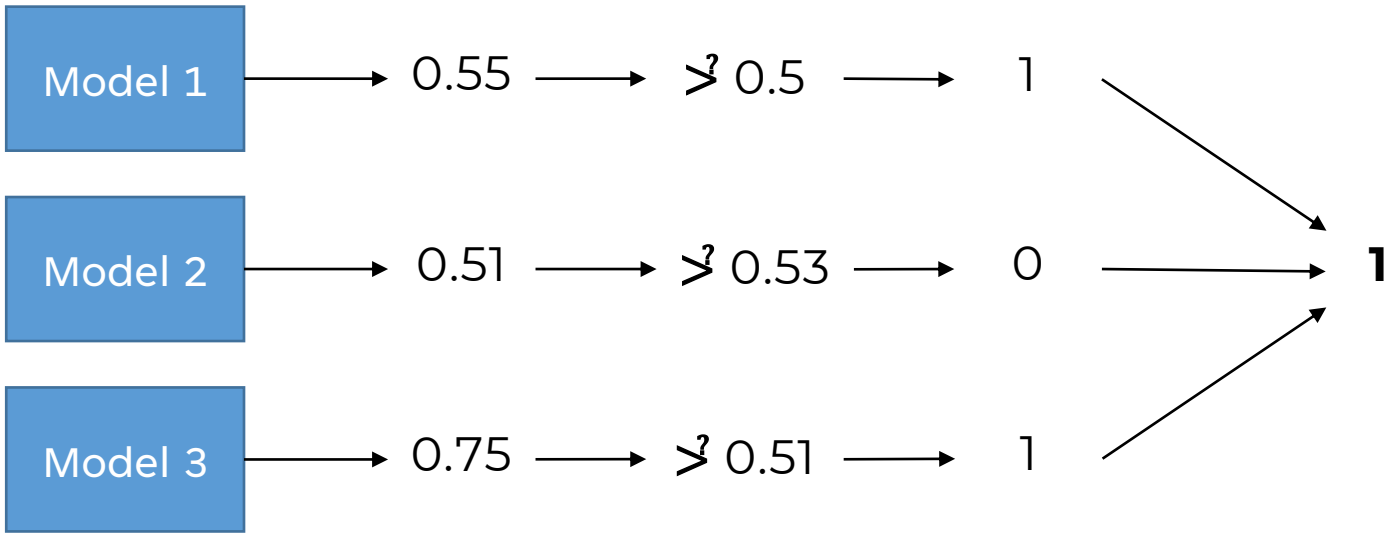


Public F1: 0,45599

Private F1: 0,45951

MODEL GELİŞTİRME SÜRECİ

THRESHOLDING



Public F1: 0,46880

Private F1: 0,46793

MODEL GELİŞTİRME SÜRECİ

GELİR BÖLÜMLEME

Gelir
1500
0
5700
18500
2000

→

Gelir Quantile-Cut
Q-01
Q-01
Q-04
Q-09
Q-02

→

- **Doğum tarihi**
- **Meslek**
- **Medeni Hal**
- **Eğitim Durumu**
- **Uyruk**
- **Memleket**
- **Gelir Quantile-Cut**

Public F1: 0,47147

Private F1: 0,47461

MODEL GELİŞTİRME SÜRECİ

CPU MODE

Training on GPU

CatBoost supports training on GPUs.

Training on GPU is non-deterministic, because the order of floating point summations is non-deterministic in this implementation.

https://catboost.ai/en/docs/features/training-on-gpu

Some types of Ctrs require target data in the training dataset. Such Ctrs are not calculated if this data is not available. In this, case only one-hot encoded categorical features are used if training is performed on GPU (and the default value of unique values threshold for a categorical feature to be considered one-hot is increased according to this condition) and all categorical features are ignored if training is performed on CPU.

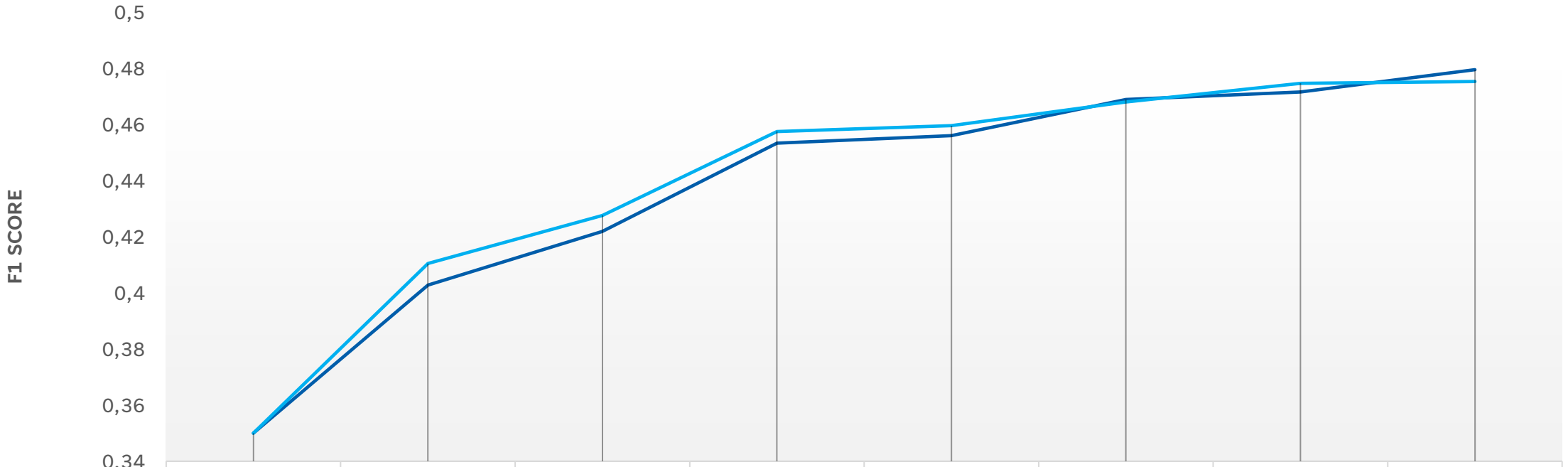
https://catboost.ai/en/docs/features/categorical-features

Public F1: 0,47944

Private F1: 0,47520

MODEL GELİŞTİRME SÜRECİ

ÖZET



	Baseline	Imbalance + Hyperparam Tuning	Sayısal FE	Kategorik Kombinasyonlar	Voting	Thresholding	Gelir bölümleme	CPU Mode
Public Set	0,35	0,40268	0,42183	0,45324	0,45599	0,4688	0,47147	0,47944
Private Set	0,35	0,41039	0,42753	0,45745	0,45951	0,46793	0,47461	0,4752

MODEL GELİŞTİRME SÜRECİ

KAÇINILAN YAKLAŞIMLAR

- **Target Mean Encoding**
- **Model Override**

DENENEN ALTERNATİFLER

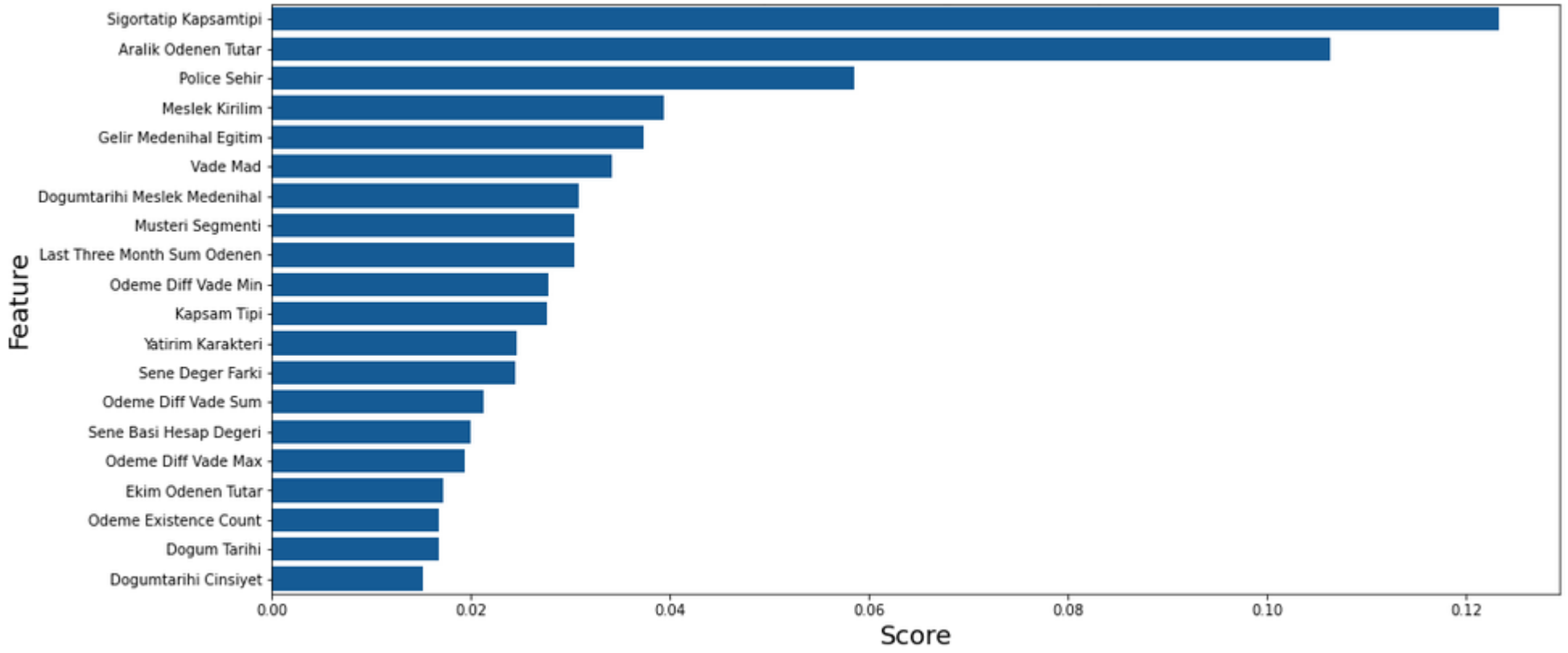
- Sayısal değerler üzerinde transformation
- NaN değerler üzerinde model tabanlı imputation
- Stacking
- Meta Level Model Blending

Ekstra veri kullanımı:

- TEFE-TÜFE verisi aylık değişimleri ile vade-ödeme miktarlarının aylık değişimlerinin ilişkisi
- Her ayın ödeme-vade tutarlarının o ayki dolar cinsinden karşılığı

ÇIKARIMLAR

ÖNEMLİ PARAMETRELER



- Gelir verisinin kategorize edilmiş hali önemli. Anketlerde bu yapıya geçilebilir.

ANADOLU HAYAT EMEKLİLİK

ÇIKARIMLAR

KURUMSAL BAĞLAM

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

PRECISION = Başarılı arama oranı *(Teklif aramalarının yüzde kaçı hedef kitleye yapıldı?)*

RECALL = Kapsayıcılık *(Hedef kitlenin yüzde kaçı arandı?)*

Metrik	Rastgele	Model
Precision	8.7265 %	49.2764 %
Recall	49.8839 %	49.5974 %

TEŞEKKÜRLER