

BTK AKADEMİ DATATHON 2024

FİNAL SUNUMU

CatBoost Baseline

Anıl Öztürk - Ahmet Tarık Karakaş

EKİP ÜYELERİ



ANIL ÖZTÜRK

Senior ML Engineer

Eatron Technologies



TARIK KARAKAŞ

Senior Data Scientist

Trendyol Group

PROBLEM

Girişimcilik Vakfı'na yapılan başvuruların değerlendirme sonucu kaç puan aldığı tahmin edilebilir mi?

- 2014 - 2022 yılları arasına ait eğitim verisi sunulmuş
- 2023 yılına ait başvurular tahmin edilecek
- Başvuranlara ait demografik bilgiler, eğitim seviyesi ve sosyal aktivitelere katılım gibi ek bilgiler de verilmiş
- Model performansı **RMSE** ile değerlendiriliyor

LİTERATÜR

MEVCUT YAKLAŞIMLAR

KULLANILAN ÖZNİTELİKLER

- Fakülte
- Öğrencinin lise ve üniversiteye giriş sıralamaları
- Ailenin gelir durumu
- Engellilik durumu
- Milliyet

- [A Data-Driven Approach in Predicting Scholarship Grants of a Local Government Unit in the Philippines Using Machine Learning](#)
- [A Study on Machine Learning Based Student Scholarship Prediction](#)
- [Increasing Enrollment by Optimizing Scholarship Allocations Using Machine Learning and Genetic Algorithms](#)

ÇÖZÜM AŞAMALARI

VERİ ANALİZİ & ÖN İŞLEME

Verinin incelenmesi, zenginleştirilmesi ve formatın standardize edilmesi

ÇÖZÜM MİMARİSİ

Feature, model, eğitim ve tahminleme detayları

ÇIKARIMLAR

Verinin problem özelinde önemi, yapılabilecek değişiklikler

VERİ ÖNİŞLEME

HİYERARŞİK VERİLERİN STANDARDİZE EDİLMESİ

```
# Eğitim durumunun ordinal encode edilmesi
```

```
education_map = {  
    '0': 0.0,  
    'doktora': 6.0,  
    'egitim yok': 0.0,  
    'egitimi yok': 0.0,  
    'ilkokul': 1.0,  
    'ilkokul mezunu': 1.0,  
    'lise': 3.0,  
    'lise mezunu': 3.0,  
    'nan': np.nan,  
    'ortaokul': 2.0,  
    'ortaokul mezunu': 2.0,  
    'universite': 4.0,  
    'universite mezunu': 4.0,  
    'yuksek lisans': 5.0,  
    'yuksek lisans / doktora': 5.0,  
    'yuksek lisans / doktora': 5.0  
}
```

```
# Lise notlarının ordinal encode edilmesi
```

```
if mode == "train":  
    lise_not_map = {  
        '3.50-3': 2.,  
        '3.00 - 4.00': 2.,  
        '3.00-2.50': 1.,  
        '2.50 ve altı': 1.,  
        '4.00-3.50': 2.,  
        '3.50-3.00': 2.,  
        '69-55': 1.,  
        '100-85': 2.,  
        '84-70': 2.,  
        '54-45': 0.,  
        '44-0': 0.,  
        '75 - 100': 2.,  
        '50 - 75': 1.,  
        'Not ortalaması yok': np.nan,  
        '25 - 50': 0.,  
        '0 - 25': 0.,  
        '50 - 74': 1.,  
        '25 - 49': 0.,  
        '0 - 24': 0.  
    }
```

VERİ ÖNİŞLEME

METİN VERİLERİNİN STANDARDİZE EDİLMESİ



VERİ ÖNİŞLEME

METİN VERİLERİNİN STANDARDİZE EDİLMESİ

Lise adı için özgün değerler:

- 'Antalya Kepez Atatürk Anadolu Lisesi',
- 'BAHÇELİEVLER ŞEHİT OSMAN YILDIZ TİCARET MESLEK ',
- '(9-10-11)Yaşar Kınıl Fen Lisesi (12)Sınav Temel Lisesi',
- '(Söğütözü) Gazi Anadolu Lisesi',
- '- FATİH KIZ ANADOLU İMAM HATİP LİSESİ (İSTANBUL - FATİH)',
- '-----',
- '-Gaziantep Ömer Özmimar Anadolu İHL -Ankara Tefik İleri Anadolu İHL. -İstanbul Tuzla Mahir İz Anadolu İHL(Açıköğretim Lisesi) (Son yazdığım okuldan mezun oldum, bunun sebebi son sınıfta İstanbul da bir kampa katılıp üniversiteye hazırlanmamdır)',
- '!',
- '.....',
- '0',
- '1.Murat Anadolu Lisesi',
- '10 Aralık Anadolu lisesi',
- '10 Aralık mesleki ve teknik Anadolu lisesi',
- '10 aralık mesleki ve teknik anadolu lisesi KASTAMONU',
- '10 sayılı tam orta Mekteb'

...

VERİ ÖNİŞLEME

METİN VERİLERİNİN STANDARDİZE EDİLMESİ

STK isimleri için özgün değerler:

- 'Aiesec',
 - 'Genç engelliler Gençlik ve Spor Kulübü',
 - 'Kızılay Gönüllüsü !Türkiye Eğitim Gönüllüleri Vakfı Gönüllüsü (TEGV)',
 - '"Aynı Dili Konuşalım" adında maddi durumu yetersiz ve dil öğrenme konusunda motivasyonu yüksek öğrencilere birçok farklı yabancı dili interaktif, canlı ve ücretsiz olarak öğretmeyi amaçladığımız gönüllülük projesinin iki kurucu ortağından biriyim. Son 1-1,5 yıldır çok yoğun bir şekilde projemizin fikirsel, teknik, iletişimsel, yönetsel tüm süreçlerinde aktif olarak yer alıyorum. Bunun yanı sıra, "Düşlerim Gerçekleşiyor"da dezavantajlı çocukların gelişimi için gönüllü çalışmalarda bulunuyorum.'
 - '"Spor İstanbul" aktif gönüllü üyesiyim.'
- ...

VERİ ÖNİŞLEME

METİN VERİLERİNİN STANDARDİZE EDİLMESİ

Üniversite isimleri için özgün değerler:

- 'Doğuş Üniversitesi',
- 'Dumlupınar Üniversitesi',
- 'DÜZCE ÜNİVERSİTESİ',
- 'Düzce Üniversitesi',
- 'DİCLE ÜNİVERSİTESİ'

VERİ ÖNİŞLEME

METİN VERİLERİNİN STANDARDİZE EDİLMESİ

DÜZCE ÜNİVERSİTESİ
Düzce Üniversitesi



VERİ ZENGİNLEŞTİRME

YÖK-ATLAS VE COĞRAFİ VERİLERİN EKLENMESİ



	program_adi	puan_turu	fakulte_adi
0	Psikoloji	EA	İnsan ve Toplum Bilimleri Fakültesi
1	Siyaset Bilimi ve Uluslararası İlişkiler	EA	İnsan ve Toplum Bilimleri Fakültesi
2	Mimarlık	SAY	Mimarlık Fakültesi
3	Bilgisayar Mühendisliği	SAY	Mühendislik Fakültesi
4	Elektrik-Elektronik Mühendisliği	SAY	Mühendislik Fakültesi

Bölüm – Fakülte Bilgileri

	uni_adi	sehir_uni
0	Abdullah Gül Üniversitesi	Kayseri
1	Acıbadem Mehmet Ali Aydınlar Üniversitesi	İstanbul
2	Adana Alparslan Türkeş Bilim Ve Teknoloji Üniv...	Adana
3	Adıyaman Üniversitesi	Adıyaman
4	Afyon Kocatepe Üniversitesi	Afyonkarahisar

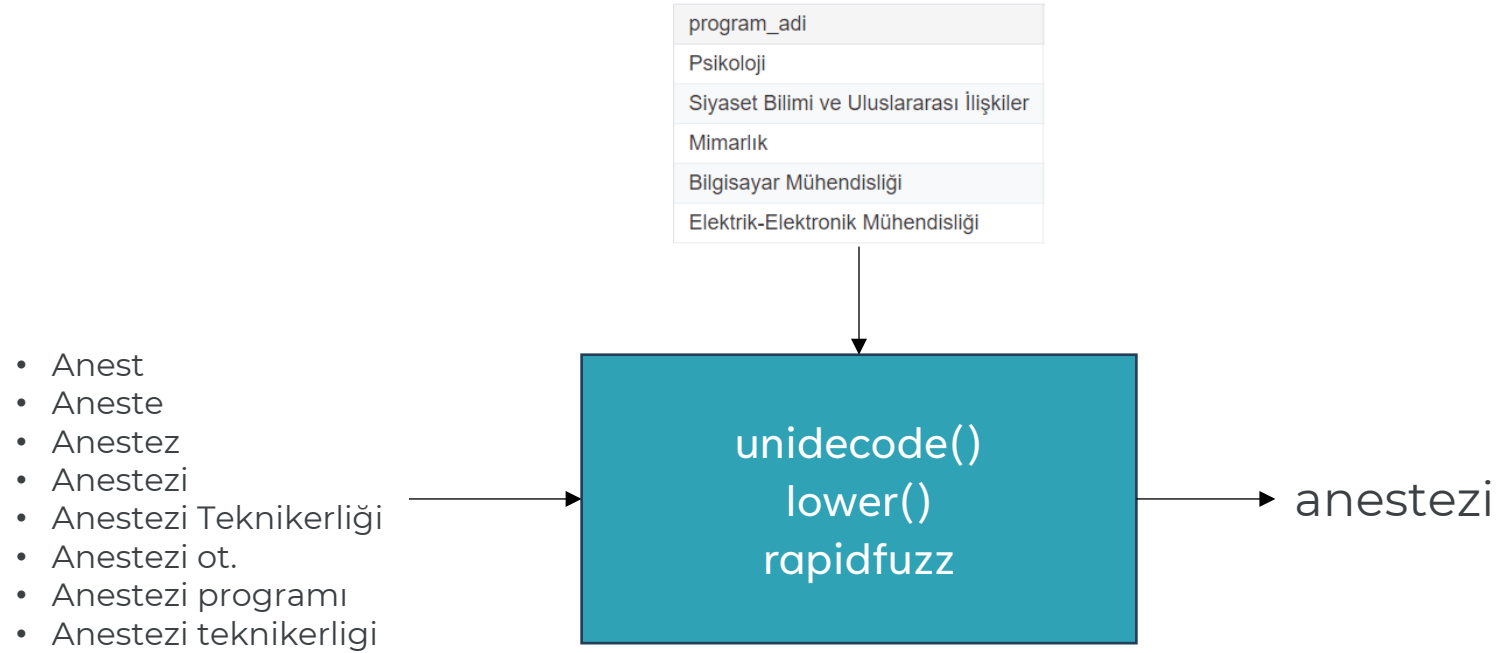
Üniversite – Şehir Bilgileri

	SehirAd	Nufus	BolgeAd
0	Adana	2201670	Akdeniz Bölgesi
1	Adıyaman	610484	Güneydoğu Anadolu Bölgesi
2	Afyonkarahisar	714523	Ege Bölgesi
3	Ağrı	542255	Doğu Anadolu Bölgesi
4	Amasya	326351	Karadeniz Bölgesi

Şehir – Bölge Bilgileri

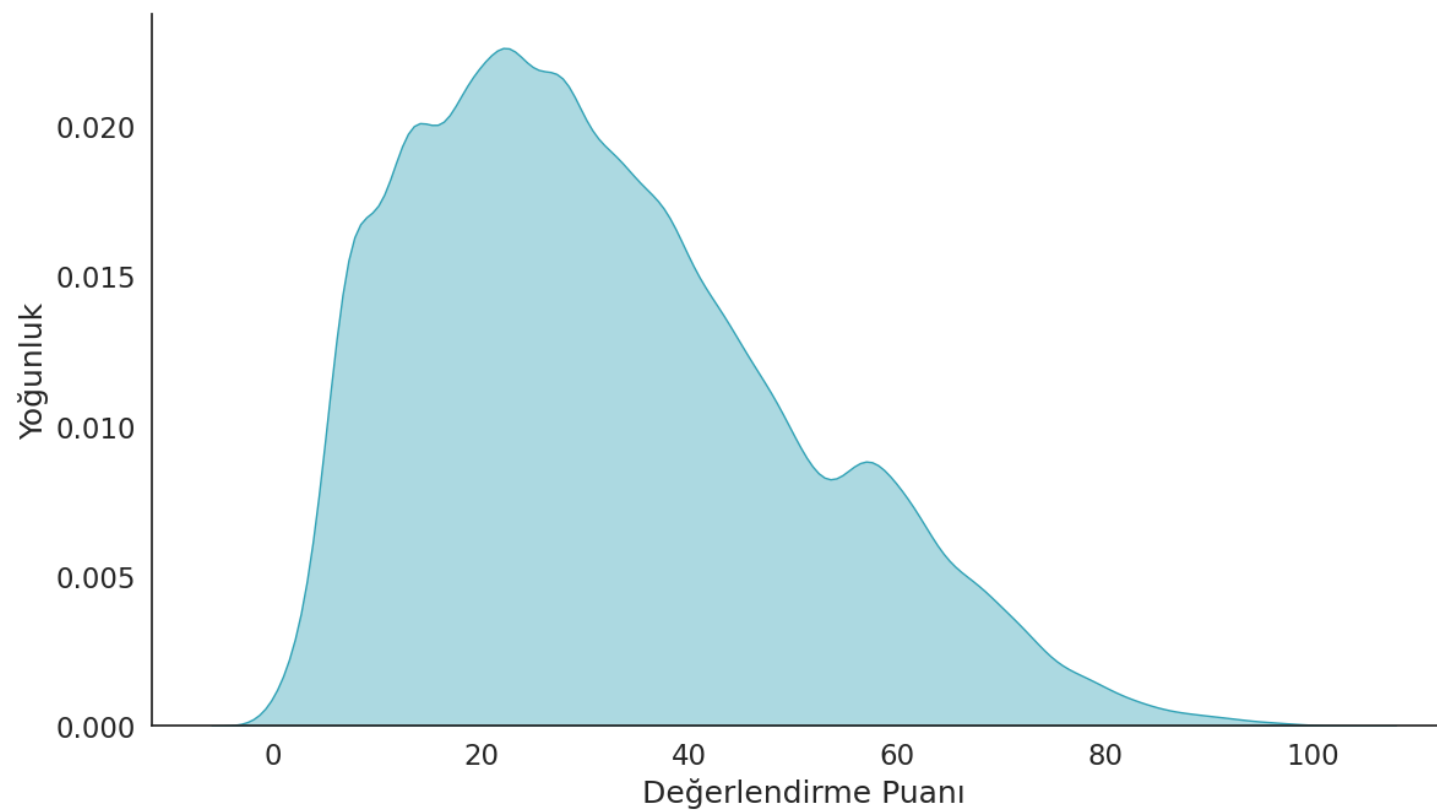
VERİ ÖNİŞLEME

METİN VERİLERİNİN STANDARDİZE EDİLMESİ



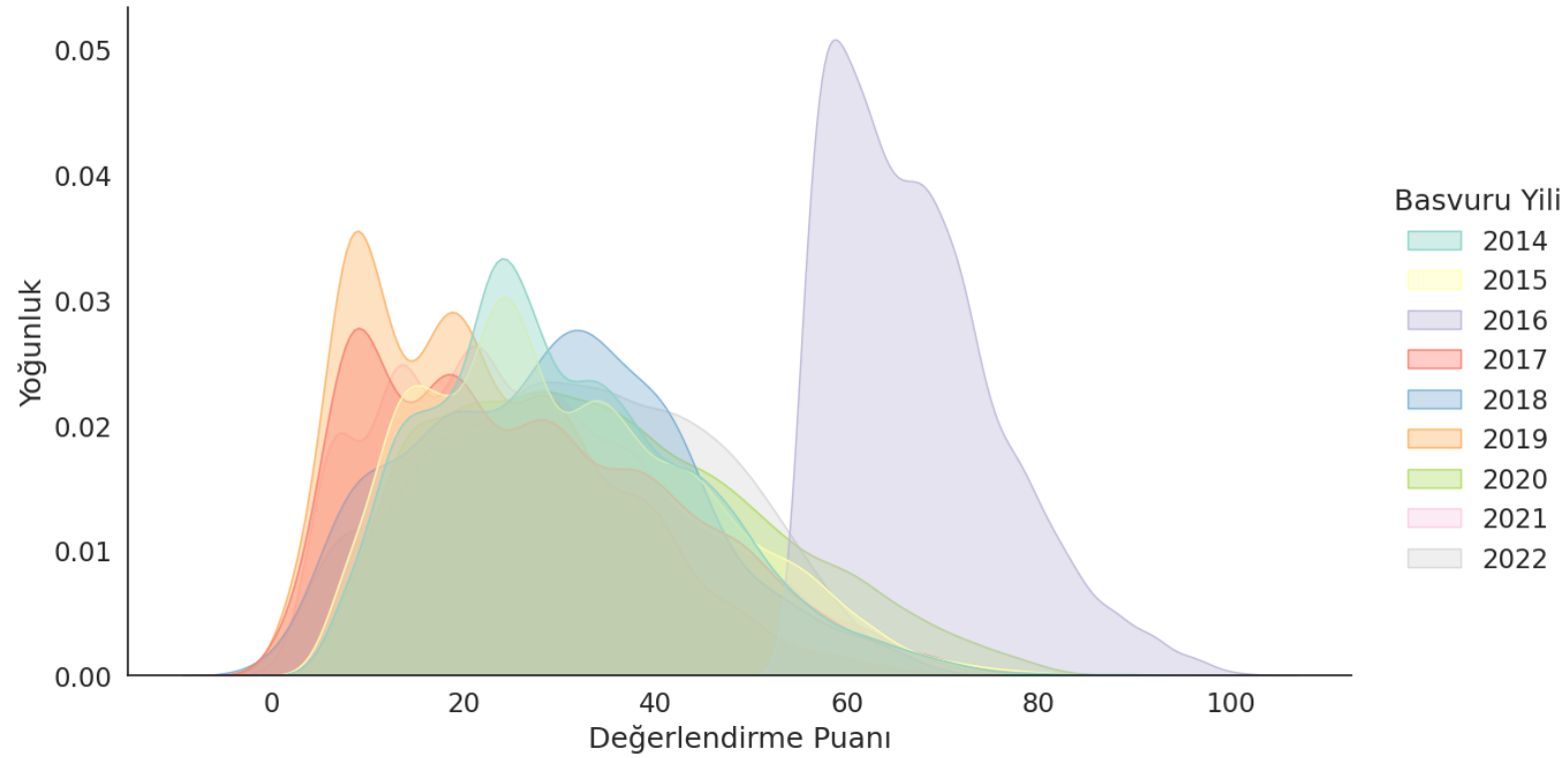
VERİ ANALİZİ

PUAN DAĞILIMI



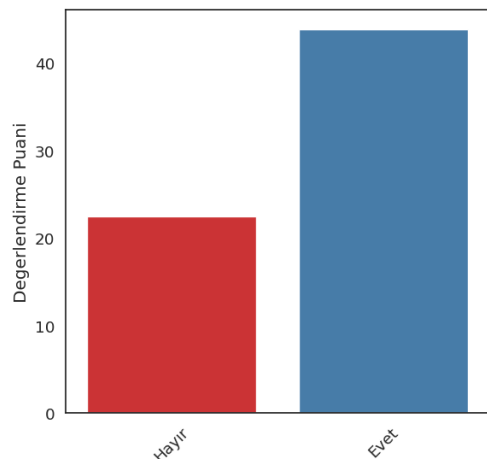
VERİ ANALİZİ

YILLARA GÖRE PUAN DAĞILIMI

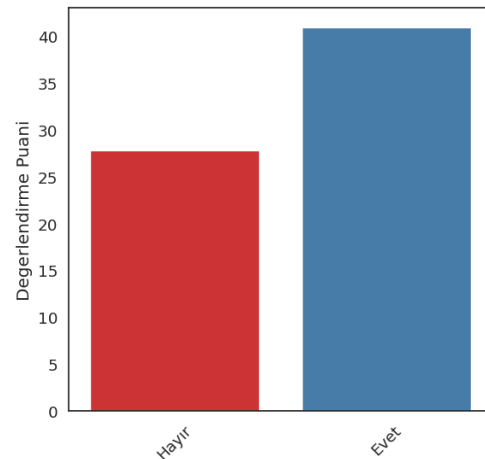


VERİ ANALİZİ

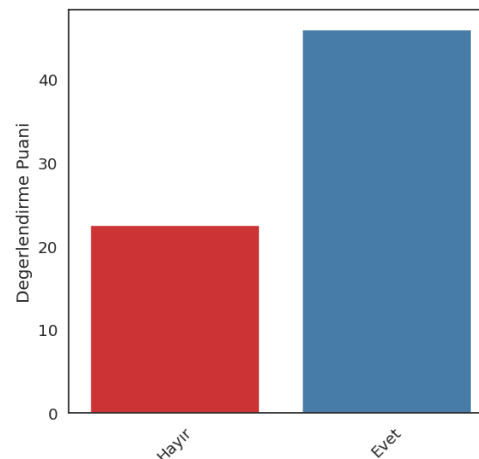
İKİLİ SORULARIN CEVAPLARINA GÖRE PUAN DAĞILIMI



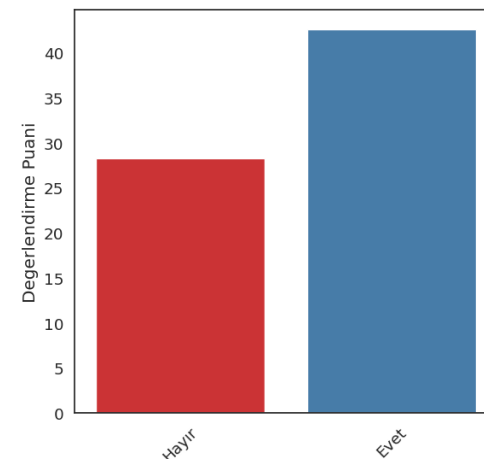
Girisimcilik Kulupleri Tarzi Bir Kulube Uye misiniz?



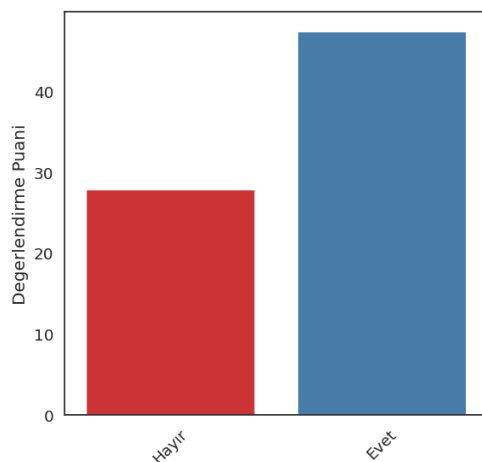
Profesyonel Bir Spor Daliyla Mesgul musunuz?



Stk Projesine Katildiniz Mi?

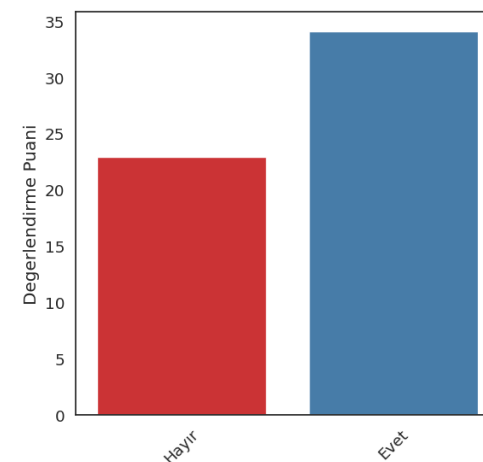


Girisimcilikle Ilgili Deneyiminiz Var Mi?



Aktif olarak bir STK üyesi misiniz?

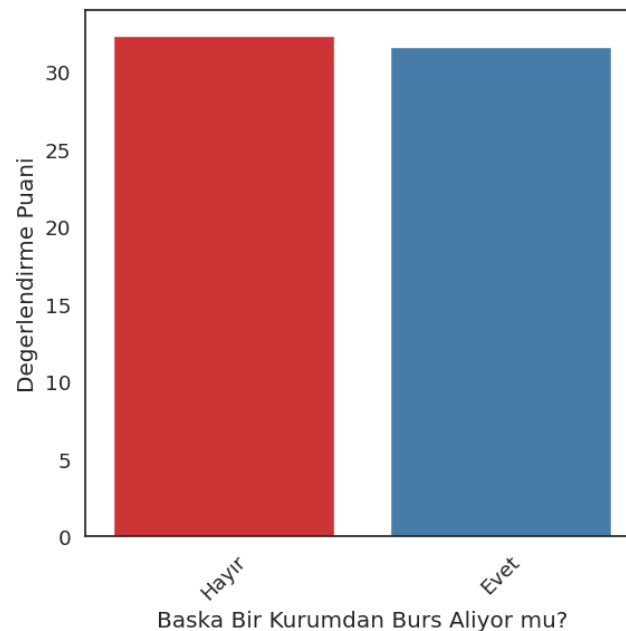
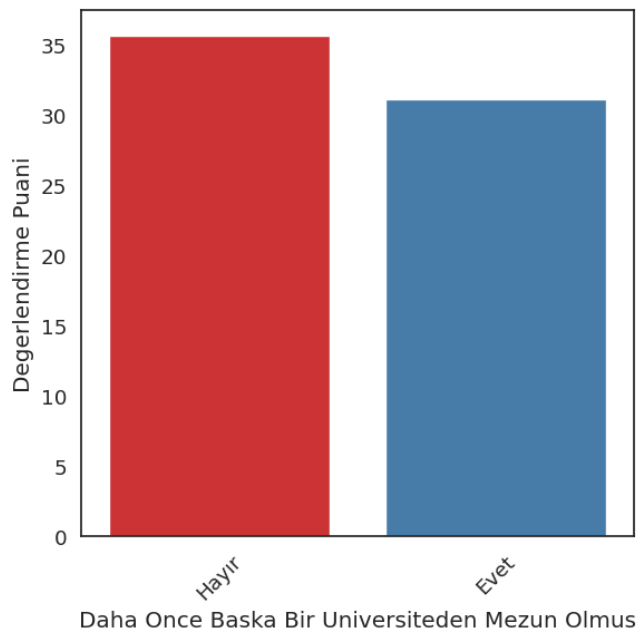
CatBoost Baseline - Final Sunumu



Ingilizce Biliyor musunuz?

VERİ ANALİZİ

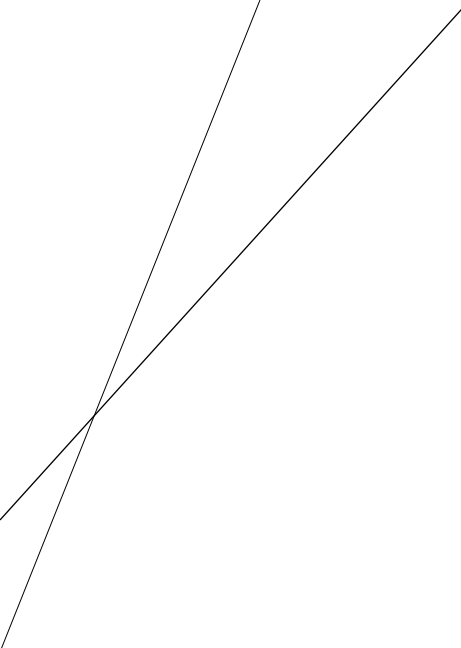
İKİLİ SORULARIN CEVAPLARINA GÖRE PUAN DAĞILIMI



VERI ANALIZI



ÜNİVERSİTE BÖLÜMLERİ ÜZERİNDE KELİME BULUTU

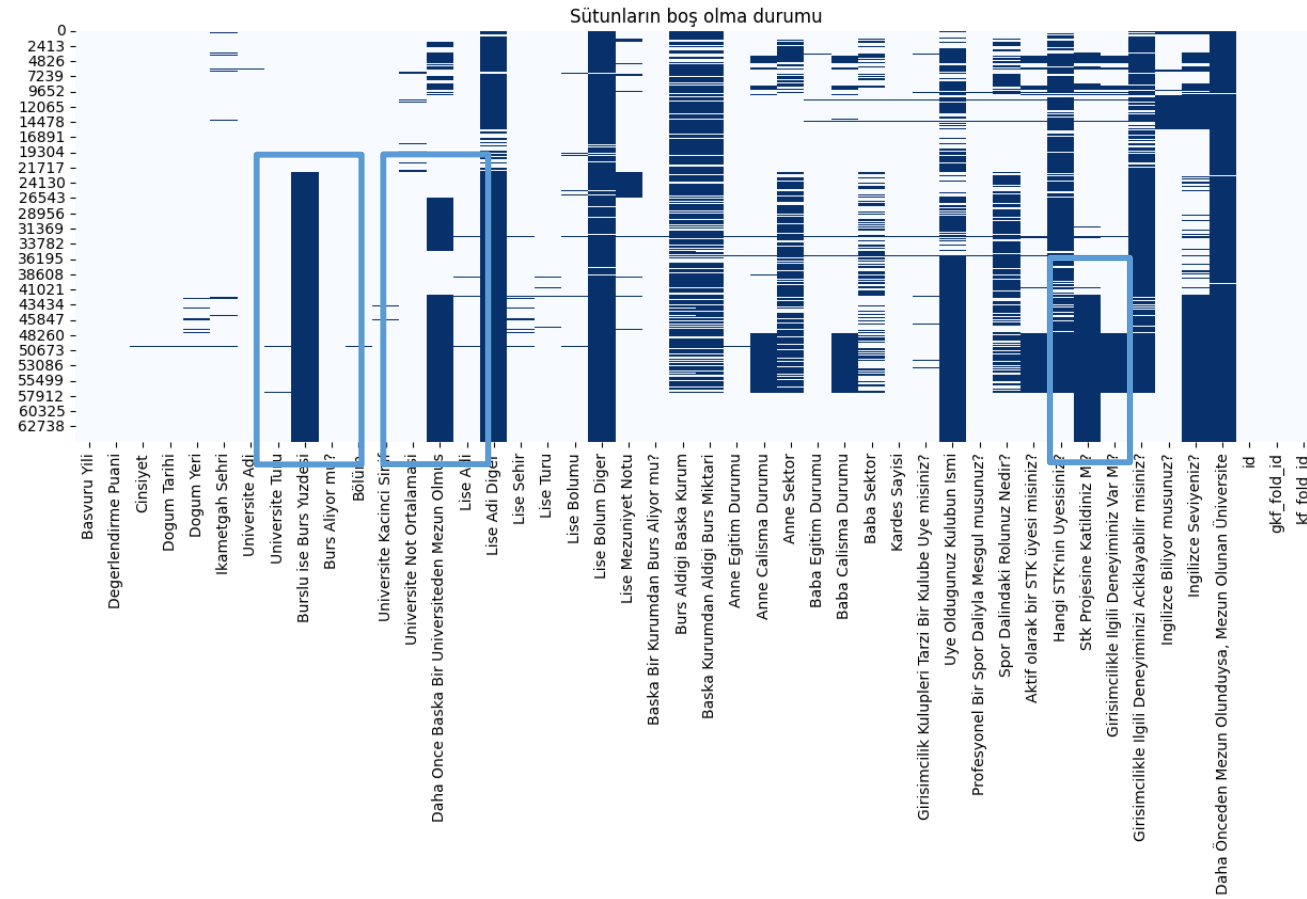


VERI ANALIZI



VERİ ANALİZİ

VERİDEKİ BOŞLUKLAR



VERİ ANALİZİ

VERİDEKİ BOŞLUKLARIN SIZINTI TESTİ (ADVERSARIAL VALIDATION)

Basvuru Yılı	Degerlendirme Puanı	Cinsiyet	Dogum Tarihi	Dogum Yeri	Ikametgah Şehri	Universite Adı	Universite Turu	Burslu İse Burs Yüzdesi	Burs Alıyor mu?	...	Hangi STK'nin Üyesisiniz?	Stk Projesine Katıldınız mı?	Girisimcilikle İlgili Deneyiminiz Var mı?	Girisimcilikle İlgili Deneyiminizi Açıklayabilir misiniz?	İngilizce Biliyor musunuz?	İngilizce Seviyeniz?	
7237	2016	74.0	Kadın	1994-03-06	Kayseri	Ankara	Melikşah Üniversitesi	Özel	50.0	Evet	...	İnovaTim	Evet	Hayır	NaN	Evet	Orta
50279	2021	17.0	Kadın	5/8/00 0:00	İstanbul	İstanbul	BOĞAZİÇİ ÜNİVERSİTESİ	Devlet	NaN	Hayır	...	NaN	NaN	NaN	NaN	Evet	NaN
23341	2016	78.0	Erkek	1994-05-07	Gümüşhane	Giresun	Giresun Üniversitesi	Devlet	NaN	Hayır	...	Günop,İnovatim	Hayır	Evet	NaN	Hayır	NaN
70187	2023	NaN	Erkek	12.03.2003 00:00	Kütahya	Kütahya	Uşak Üniversitesi	Devlet	NaN	Hayır	...	-	NaN	Hayır	-	Hayır	NaN
65304	2023	NaN	Kadın	3.02.2002 00:00	Bursa	Bursa	Sakarya Üniversitesi	Devlet	NaN	Hayır	...	Lösev'de gönüllü ablayım. Düzenli Kızılay kan ...	NaN	Evet	Amazon üzerinden ürün satma deneyimim oldu. Fa...	Evet	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
66141	2023	NaN	Kadın	5.02.2001 00:00	İstanbul	İstanbul	İstanbul Gelişim Üniversitesi	Özel	NaN	Evet	...	Afet çalışmaları	NaN	Hayır	-	Evet	NaN
40354	2019	44.0	Kadın	16/10/2000	İstanbul	İstanbul	YILDIZ TEKNİK ÜNİVERSİTESİ	DEVLET	NaN	Hayır	...	NaN	Evet	Hayır	NaN	Evet	NaN
69577	2023	NaN	Kadın	5.02.2002 00:00	Erzurum	Trabzon	Amasya Üniversitesi	Devlet	NaN	Hayır	...	-	NaN	Hayır	-	Hayır	NaN
32940	2017	37.0	ERKEK	8/15/1997	Mersin	Mersin	GAZİANTEP ÜNİVERSİTESİ	DEVLET	NaN	Hayır	...	NaN	Evet	Evet	NaN	Evet	Orta
42433	2020	15.0	Kadın	9/19/98 0:00	Bitlis	Antalya	VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ	Devlet	NaN	Hayır	...	NaN	NaN	Hayır	NaN	Hayır	NaN

Model

Satır hangi yıla ait?

AUC: ~0.98

VERİ ANALİZİ

VERİDEKİ BOŞLUKLAR (YIL BOYU BOŞ OLMA DURUMU)

Sütun İsmi	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Burslu ise Burs Yuzdesi	x	x	x	x	x	x	x	x		
Daha Once Baska Bir Universiteden Mezun Olmus	x	x	x		x	x				
Lise Adi Diger		x								
Lise Bolum Diger		x	x	x	x	x				
Lise Mezuniyet Notu	x	x		x	x	x	x	x	x	x
Anne Calisma Durumu	x	x	x	x	x	x	x		x	x
Baba Calisma Durumu	x	x	x	x	x	x	x		x	x
Uye Oldugunuz Kulubun Ismi	x	x	x	x	x					
Aktif olarak bir STK üyesi misiniz?	x	x	x	x	x	x	x		x	x
Hangi STK'nin Uyesisiniz?	x	x	x		x	x	x		x	x
Stk Projesine Katildiniz Mi?	x	x	x	x	x	x				
Girisimcilikle Ilgili Deneyiminiz Var Mi?	x	x	x	x	x	x	x		x	x
Girisimcilikle Ilgili Deneyiminizi Aciklayabilir misiniz?	x	x					x		x	x
Ingilizce Biliyor musunuz?		x	x	x	x	x	x	x	x	x
Ingilizce Seviyeniz?		x	x	x	x	x				
Daha Önceden Mezun Olunduysa, Mezun Olunan Üniversite			x		x	x				

ÇÖZÜM MİMARİSİ

VALİDASYON ŞEMASI

- Yıllar üzerinde Group-Kfold
- Yıllar üzerinde Timeseries split
- KFold

ÇÖZÜM MİMARİSİ

VALİDASYON ŞEMASI

- Yıllar üzerinde Group-Kfold
- Yıllar üzerinde Timeseries split
- **KFold**

ÇÖZÜM MİMARİSİ

FEATURE SELECTION

- CatBoost
- SHAP Values
- 5-Fold CV
- 5-step

ayarları kullanılarak feature-selection yapıldı.

Ekstra veri olarak eklenen:

- Program puan türü
- Fakülte adı

değişkenleri de algoritma tarafından seçildi.



ÇÖZÜM MİMARİSİ

FEATURE ENGINEERING

- Üniversite adı
- Girişimcilik deneyimi
- Burs aldığı kurum
- Üniversite bölümü

üzerinde TFIDF matrixleri çıkarıldı.

Seçilen girişim TFIDF değerleri:

- devam
- aldım
- sosyal
- için

Seçilen bölüm TFIDF değerleri:

- mühendisliği
- işletme
- bilgisayar mühendisliği

Seçilen üniversite TFIDF değerleri:

- istanbul
- üniversitesi

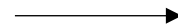
ÇÖZÜM MİMARİSİ

FEATURE ENGINEERING

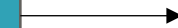


Hugging Face

Girişimcilik deneyimi



bert-base-turkish-128k-uncased



Text Embeddings

Final çözümde algoritma tarafından seçilmedi

ÇÖZÜM MİMARİSİ

FEATURE ENGINEERING

- Yaş (*Başvuru yılı – Doğum yılı*)
- Üniversiteye başlangıç yaşı (*Yaş – Sınıf*)
- Bölüm içinde:
 - Çap
 - Burs
 - İngilizce
 - Mühendis

geçme durumu

- Lise içinde:
 - Anadolu
 - Fen
 - Meslek

geçme durumu

- Burs kurumu içinde «kyk» geçme durumu

- Doğum şehrinin ikamet şehrine eşit olma durumu
- Üniversite şehrinin ikamet şehrine eşit olma durumu
- Girişimcilik deneyiminin karakter sayısı
- STK isminin karakter sayısı
- Girişimcilik metni üzerinde:
 - Metindeki cümle sayısı
 - Cümle başına ortalama kelime sayısı
 - Kelime başına ortalama karakter sayısı
 - Okunabilirlik endeksi (*Ateşman*) - <http://okunabilirlikindeksi.com/>

ÇÖZÜM MİMARİSİ

FEATURE ENGINEERING

- **Yaş (Başvuru yılı – Doğum yılı)**
- Üniversiteye başlangıç yaşı (Yaş – Sınıf)
- **Bölüm içinde:**
 - Çap
 - Burs
 - İngilizce
 - **Mühendis**

geçme durumu

- **Lise içinde:**
 - **Anadolu**
 - **Fen**
 - **Meslek**

geçme durumu

- **Burs kurumu içinde «kyk» geçme durumu**

- **Doğum şehrinin ikamet şehrine eşit olma durumu**
- **Üniversite şehrinin ikamet şehrine eşit olma durumu**
- **Girişimcilik deneyiminin karakter sayısı**
- **STK isminin karakter sayısı**
- Girişimcilik metni üzerinde:
 - Metindeki cümle sayısı
 - Cümle başına ortalama kelime sayısı
 - Kelime başına ortalama karakter sayısı
 - Okunabilirlik endeksi (Ateşman) - <http://okunabilirlikindeksi.com/>

ÇÖZÜM MİMARİSİ

MODEL SEÇİMİ

- CatBoost
- LGBM
- XGBoost
- Keras – Shallow NN
- AutoGluon
- TabNet

ÇÖZÜM MİMARİSİ

MODEL SEÇİMİ

- **CatBoost**
- **LGBM**
- XGBoost
- Keras – Shallow NN
- **AutoGluon**
- TabNet



Yandex
CatBoost



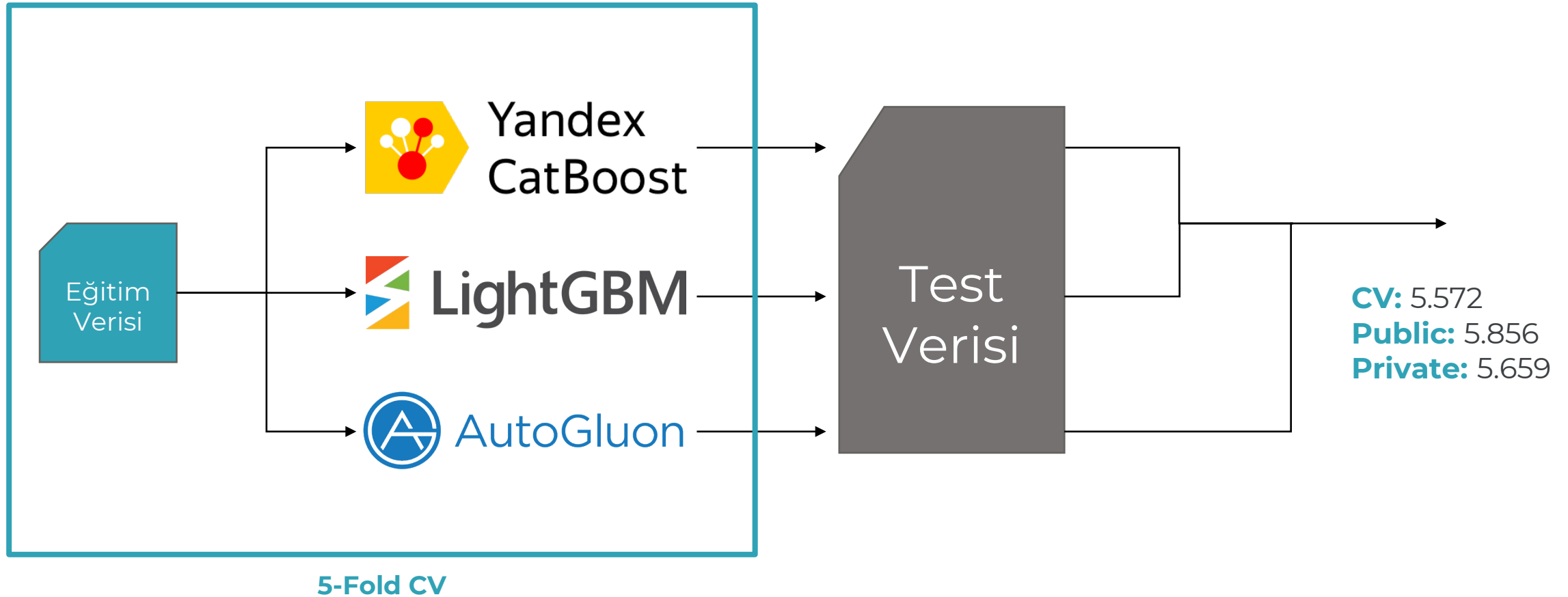
LightGBM



AutoGluon

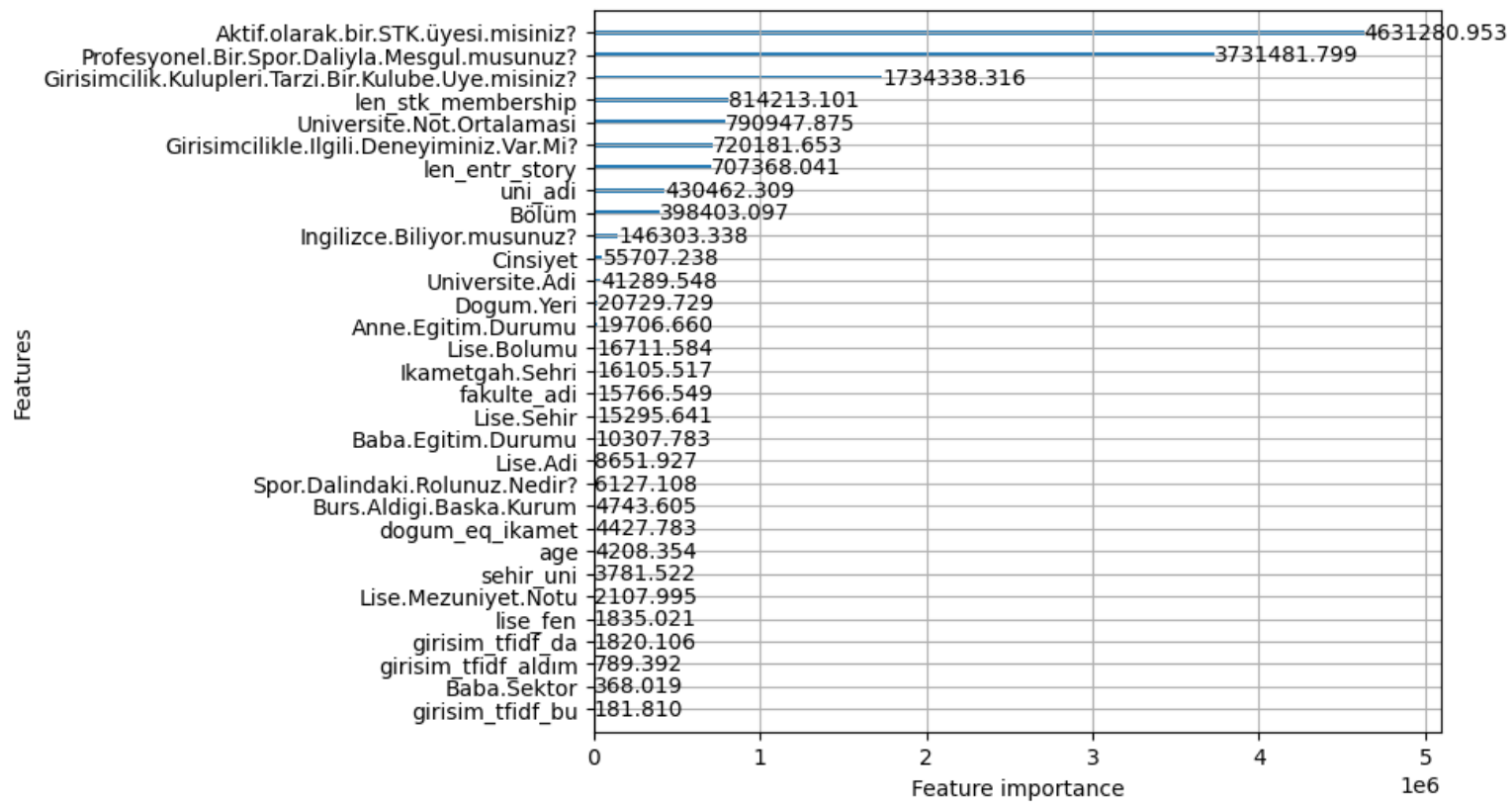
ÇÖZÜM MİMARİSİ

EĞİTİM VE TAHMİNLEME



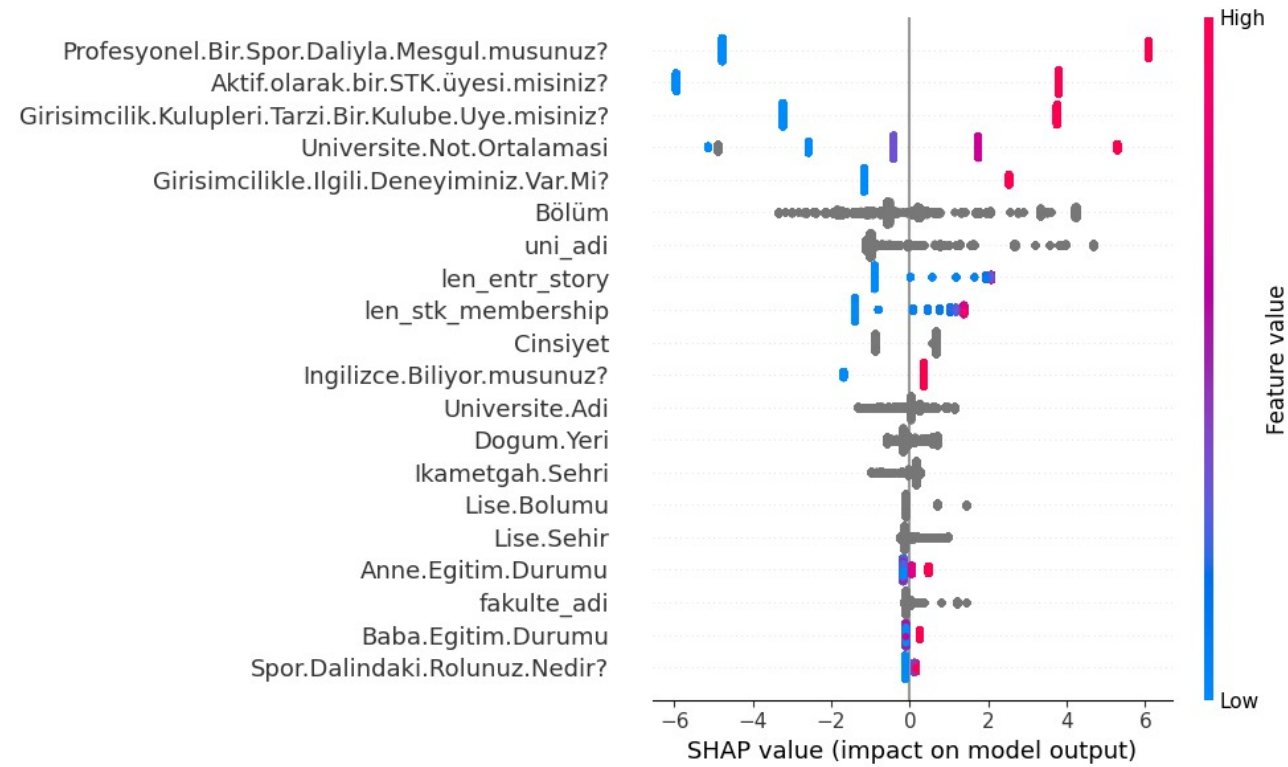
ÇIKARIMLAR

FEATURE IMPORTANCE



ÇIKARIMLAR

DEĞİŞKENLERİN KARARLARA ETKİLERİ



ÇIKARIMLAR

YAPILABİLECEK DEĞİŞİKLİKLER

- Sütunların doluluk istikrarı sağlanarak daha fazla yıl eklenebilir
- Jüride sabitleşmeye gidilebilir
- Kaggle tarafında 80-20 yerine 20-80 split yapılabilir
- Literatürdeki featurelar eklenebilir (*gelir durumu, milliyet, engellilik durumu*)



TEŞEKKÜRLER