Borusan Otomotiv

BORUSAN DATATHON

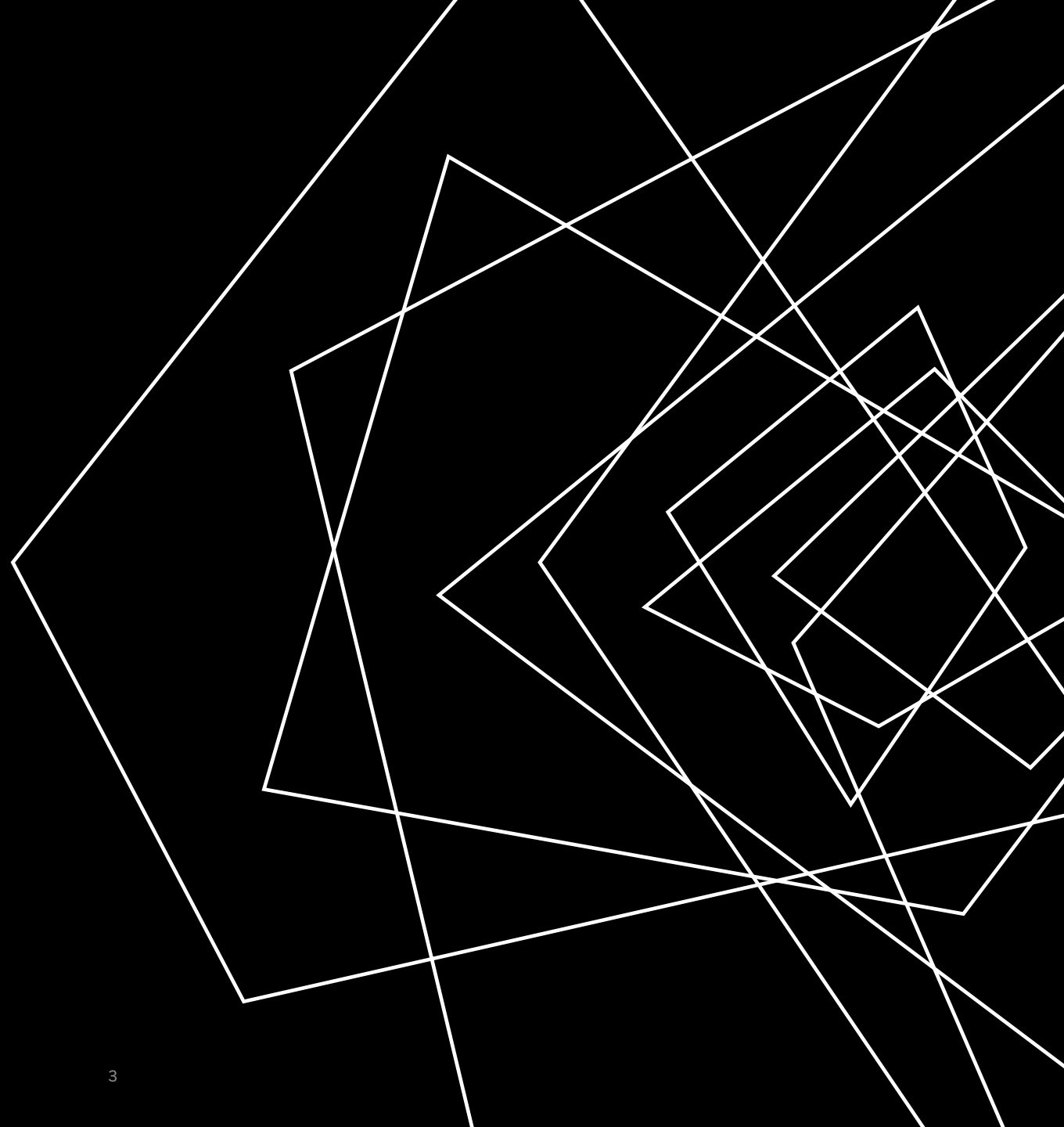
FİNAL SUNUMU

Anıl Öztürk

PROBLEM

Verilen 6 aylık dönem içerisinde Borusan altındaki markalara ait müşteri araçlarının parçaları arızalanacak mı?

- Servis kayıtları verilmiş
- Kayıtların parça detayları verilmiş
- Başarı oranı **F1-Score** ile değerlendiriliyor



LİTERATÜR

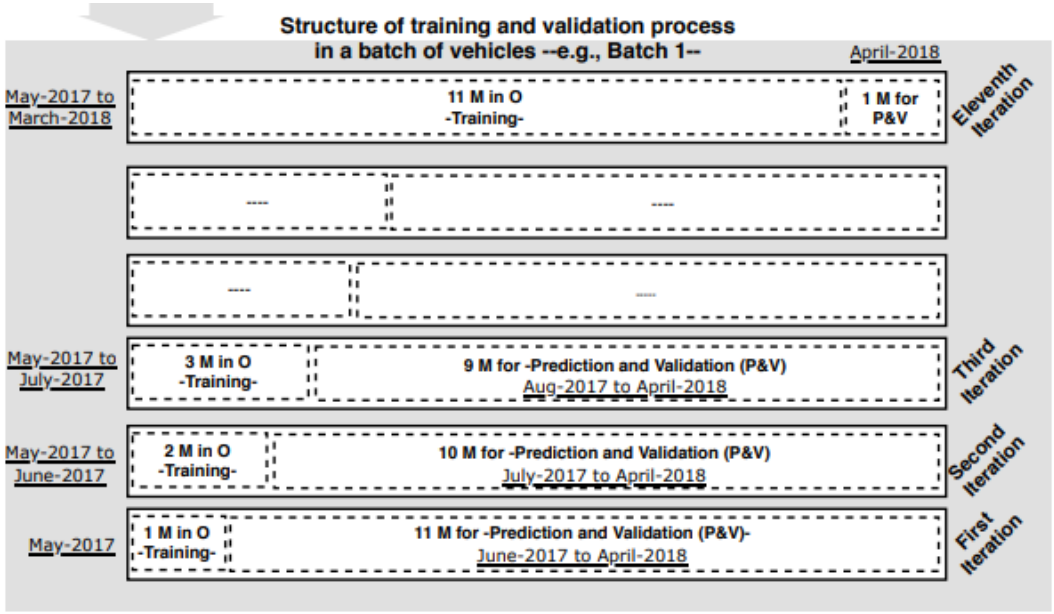
MEVCUT YAKLAŞIMLAR

1. KULLANILAN VERİLER

- Servis kayıtları
- Araç detayları
- Lokasyon ve sensör verileri
- Sürücü karakteristiği

2. TAHMİN MODELİ TÜRÜ

- Karar ağacı ailesi
- Yapay sinir ağları



- https://www.researchgate.net/publication/342724999_Early_Prediction_of_Quality_Issues_in_Automotive_Modern_Industry
- https://patents.google.com/patent/US20160035150A1/en
- https://www.proquest.com/openview/c0692c028ba5f251a4c1dc6732ccc162/1?pq-origsite=gscholar&cbl=4998670
- https://www.just-auto.com/interview/how-machine-learning-can-forecast-parts-failures/

ÇÖZÜM AŞAMALARI

VERİ ANALİZİ & ÖN İŞLEME

Verideki dağılımların incelenmesi ve formatın standardize edilmesi

MODEL TASARIM SÜRECİ

Kullanılan model ve tahminleme mimarisi

DENENEN ALTERNATİFLER

Çözüm için denenen alternatif veri manipülasyonları ve teknikler

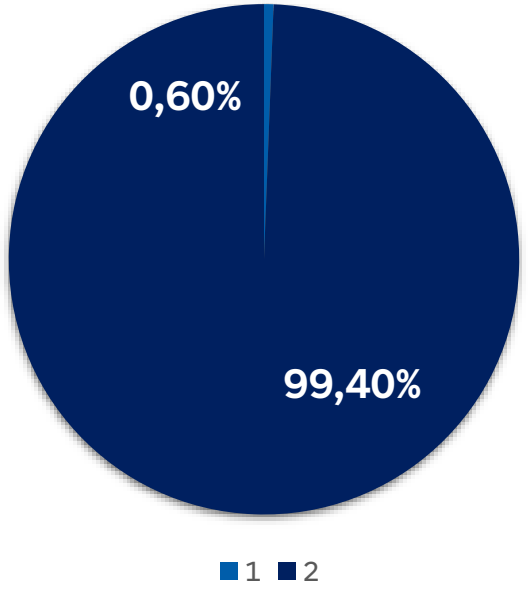
ÇIKARIMLAR

Problemden çıkarılan içgörü ve kurumsal bakış ile yorumlama

VERİ ANALİZİ

CLASS IMBALANCE

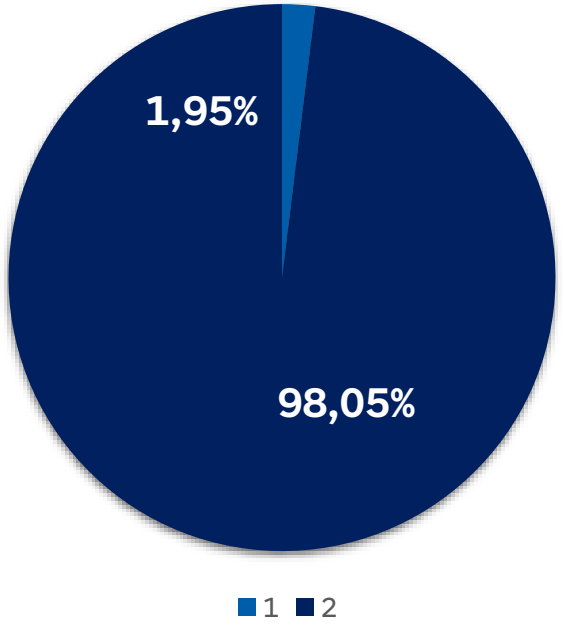
Servis kaydının arıza tespiti içerme durumu



VERİ ANALİZİ

CLASS IMBALANCE

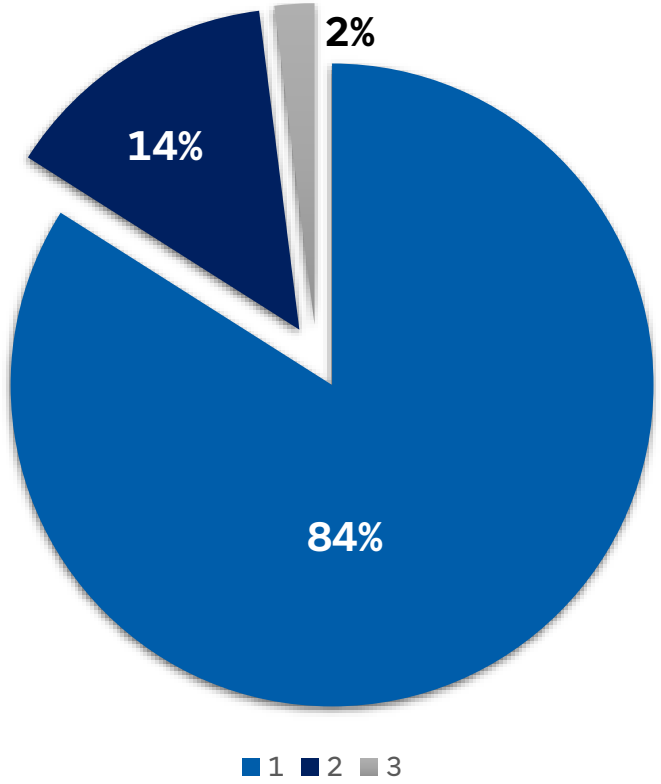
Araca özgü servis kaydının arıza tespiti içerme durumu



VERİ ANALİZİ

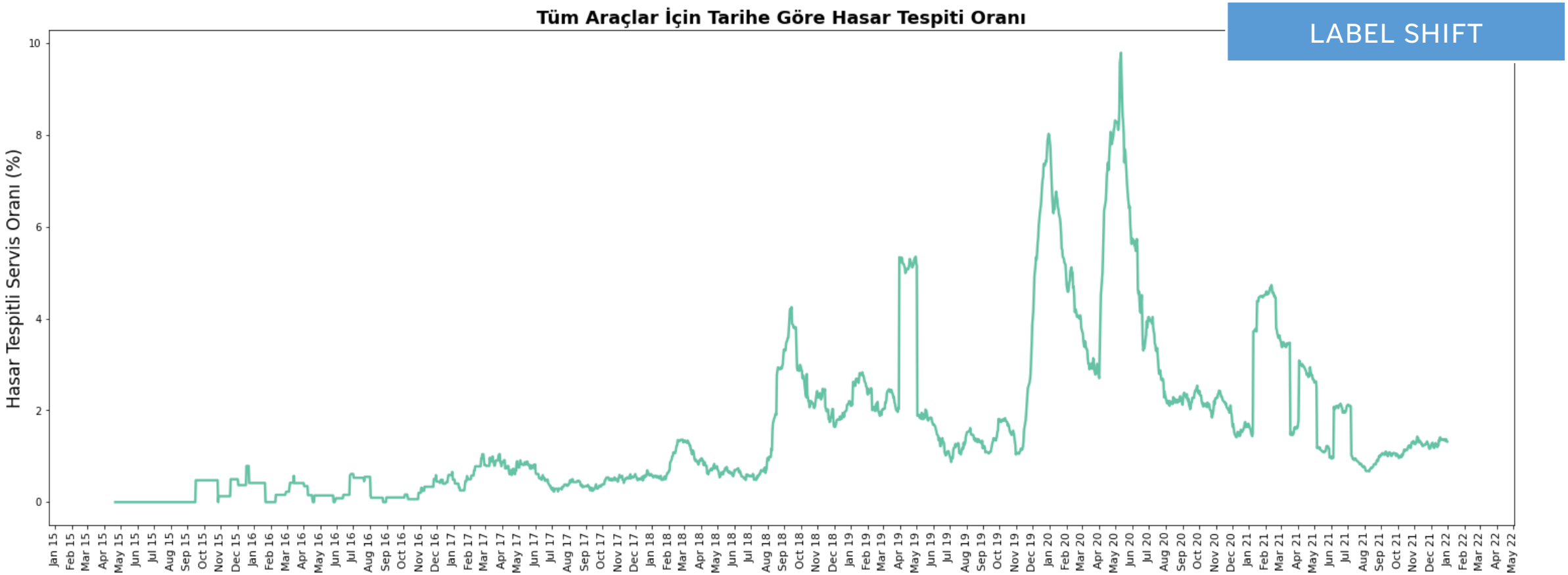
ARIZA ANALİZİ

Araç bazında arıza görülme sayısı



VERİ ANALİZİ

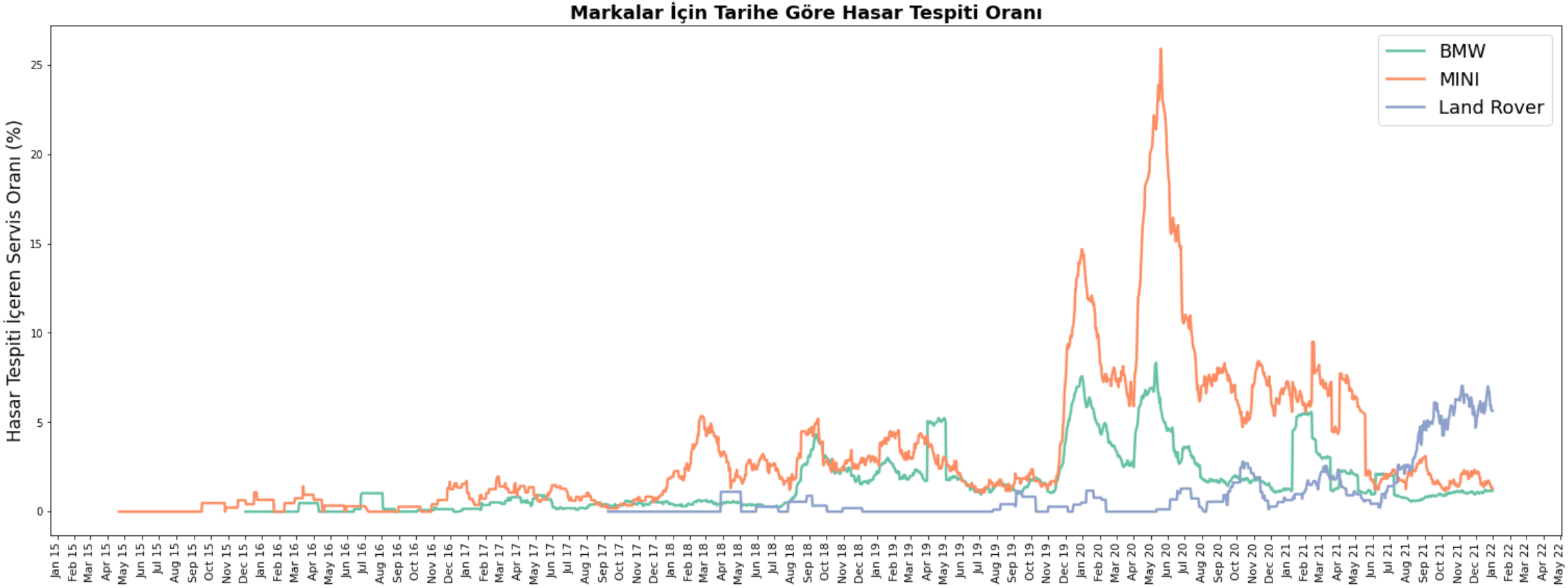
ARIZA ANALİZİ



LABEL SHIFT

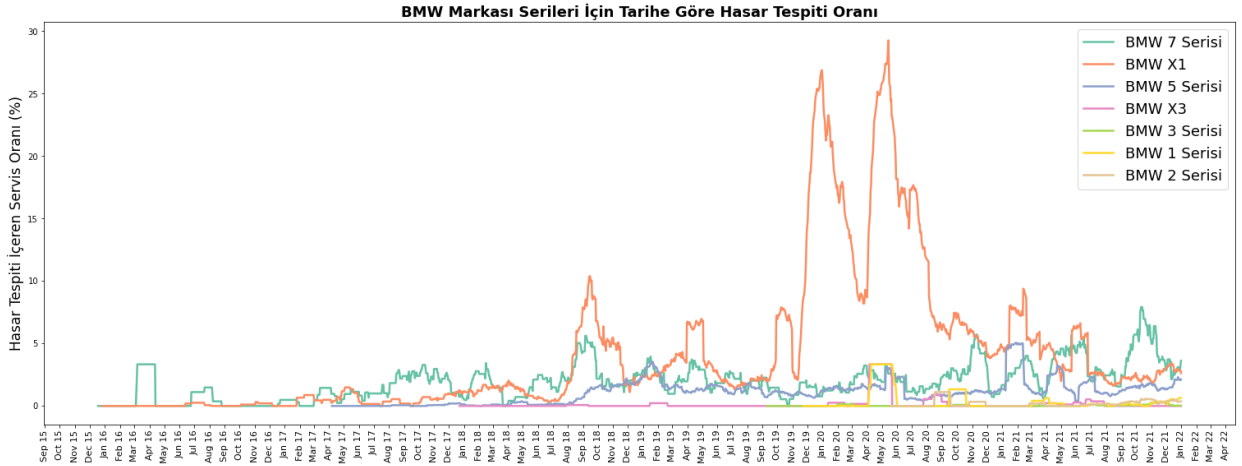
VERİ ANALİZİ

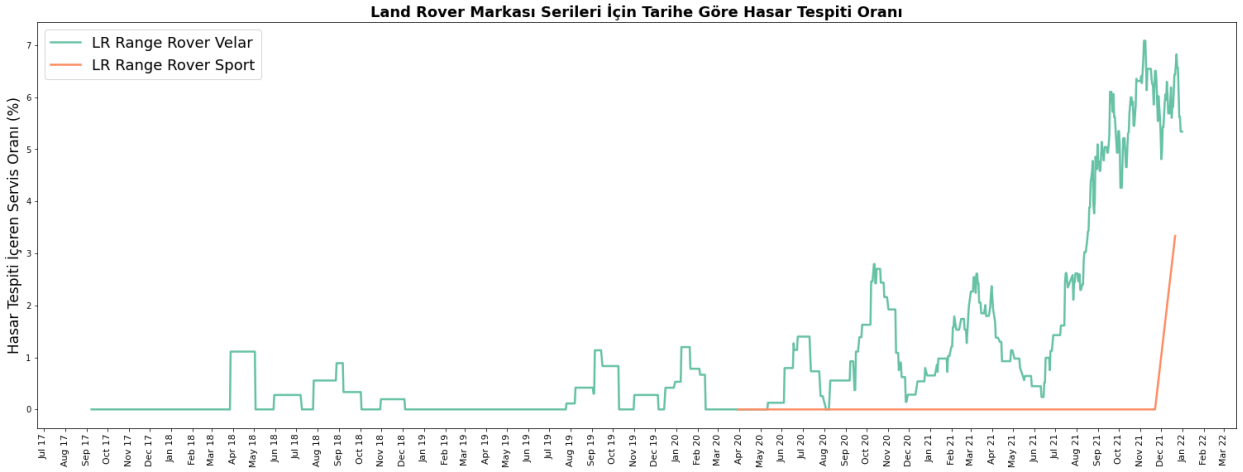
ARIZA ANALİZİ

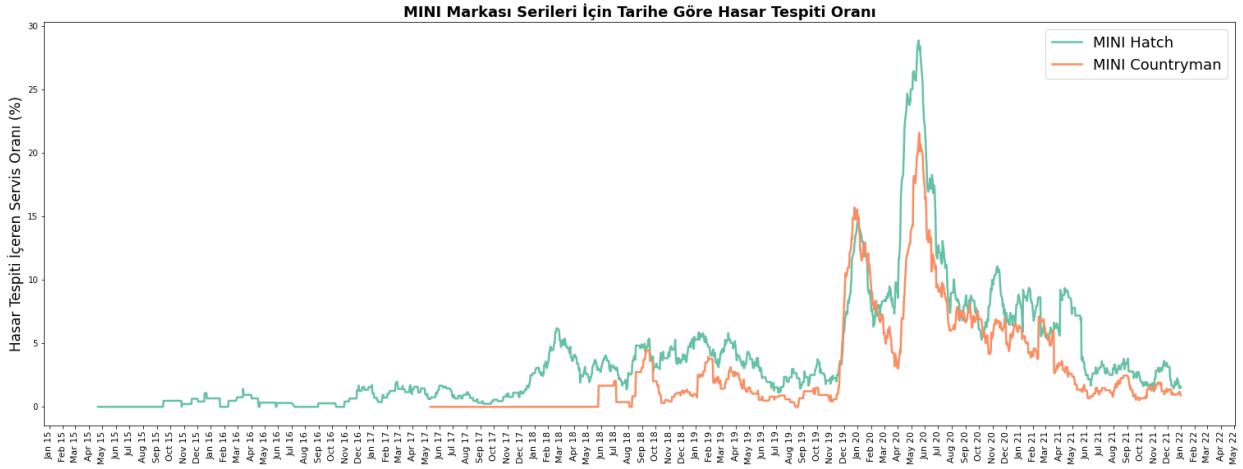


VERİ ANALİZİ

ARIZA ANALİZİ

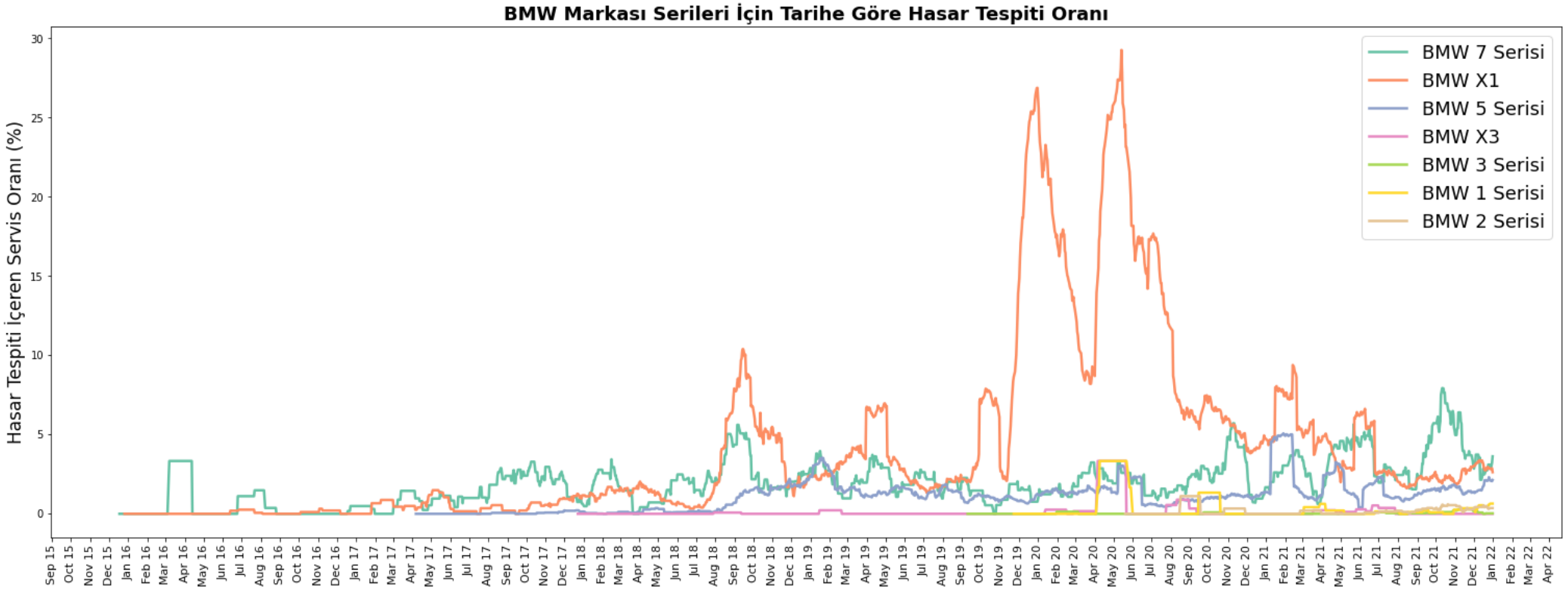






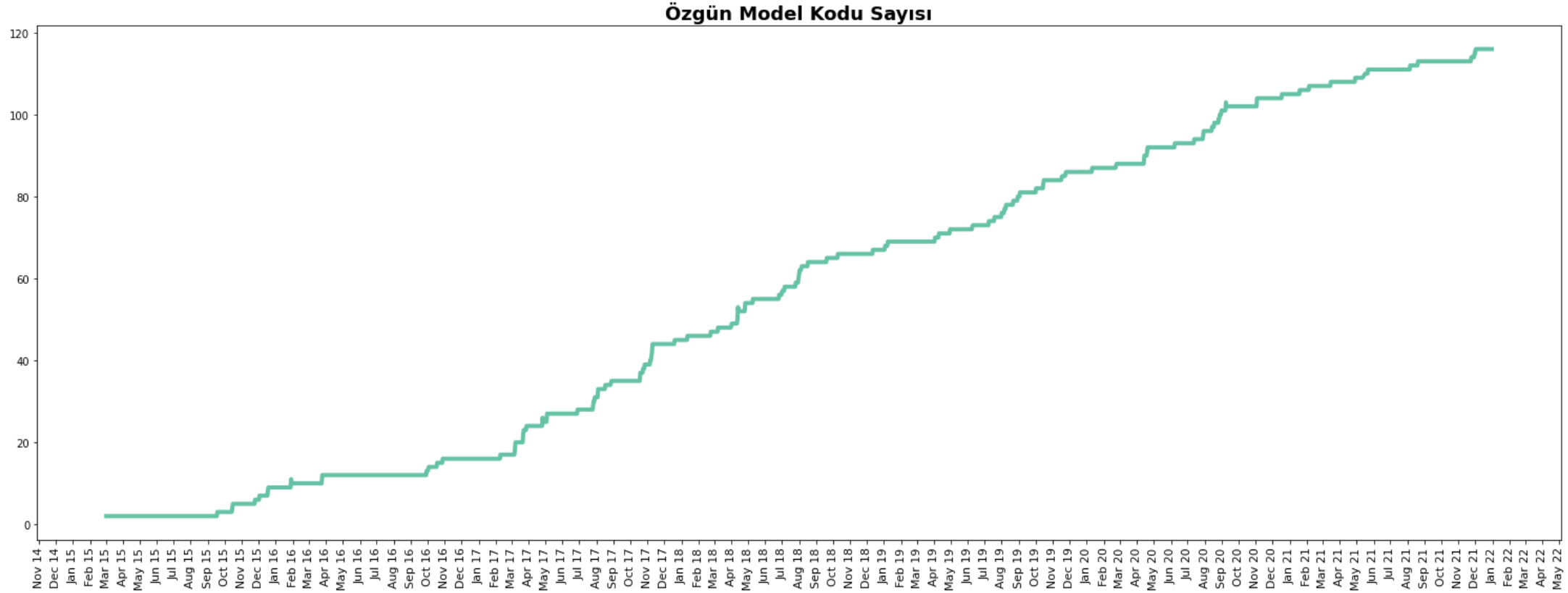
VERİ ANALİZİ

ARIZA ANALİZİ



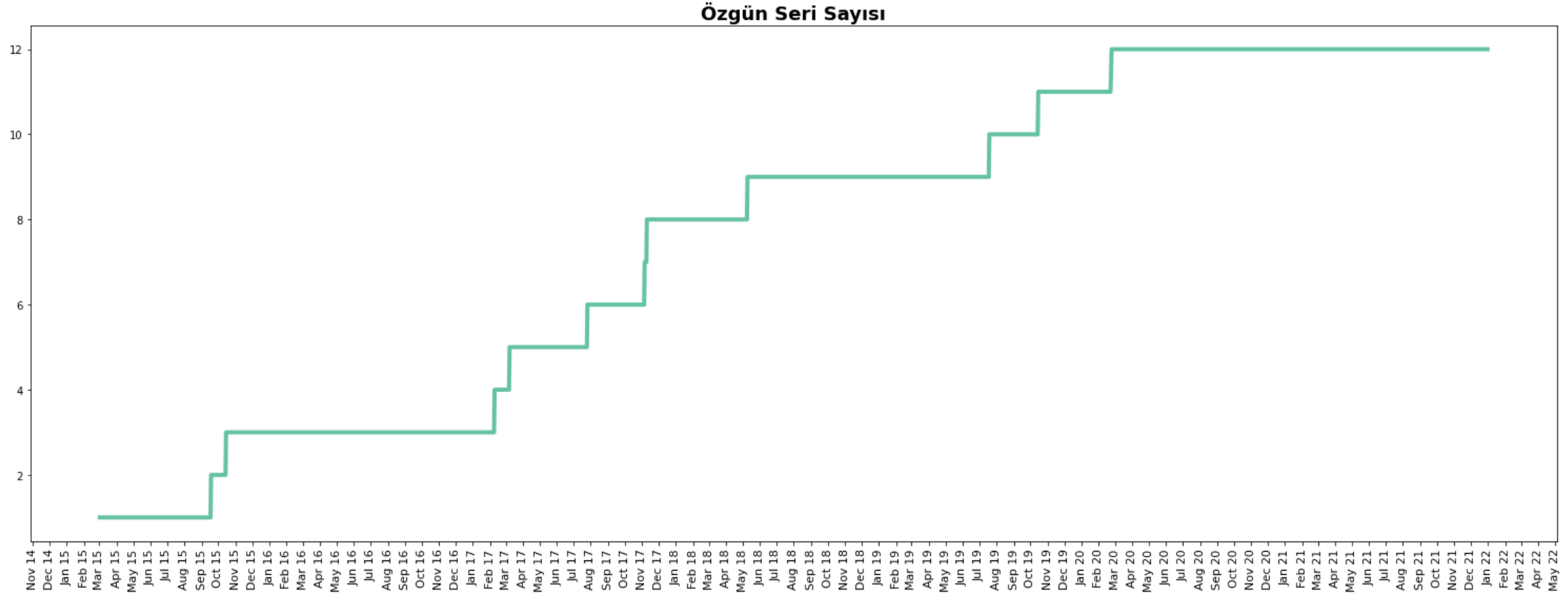
VERİ ANALİZİ

FEATURE SHIFT



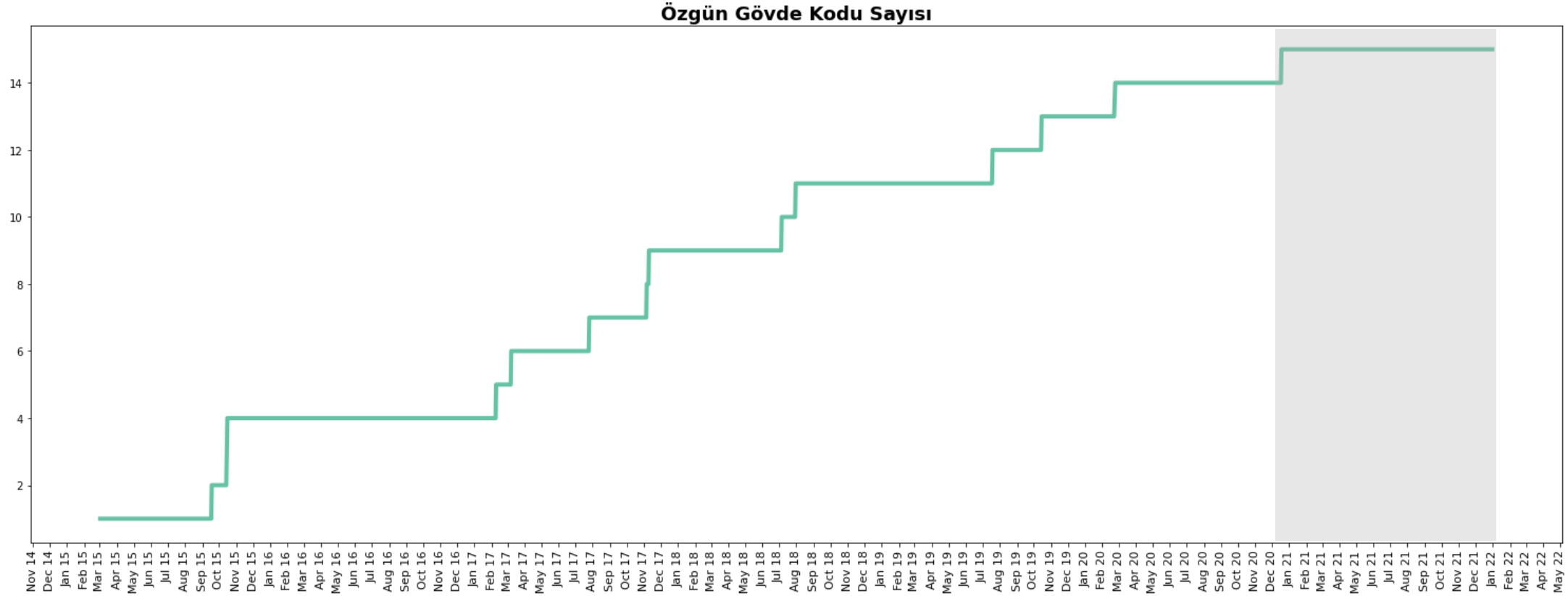
VERİ ANALİZİ

FEATURE SHIFT



VERİ ANALİZİ

FEATURE SHIFT



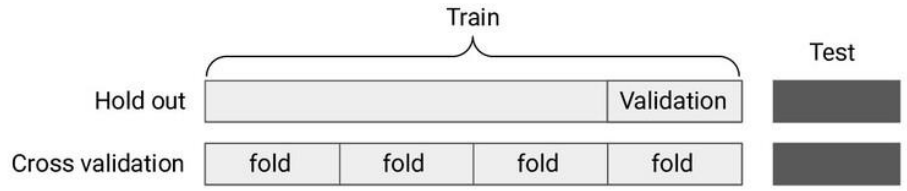
VERİ ANALİZİ

ADVERSARIAL VALIDATION

Cross Validation & Adversarial Validation

Cross validation: The data is divided into k folds; k-1 folds are used for training and the other fold is used for validation, which is done for all combinations.

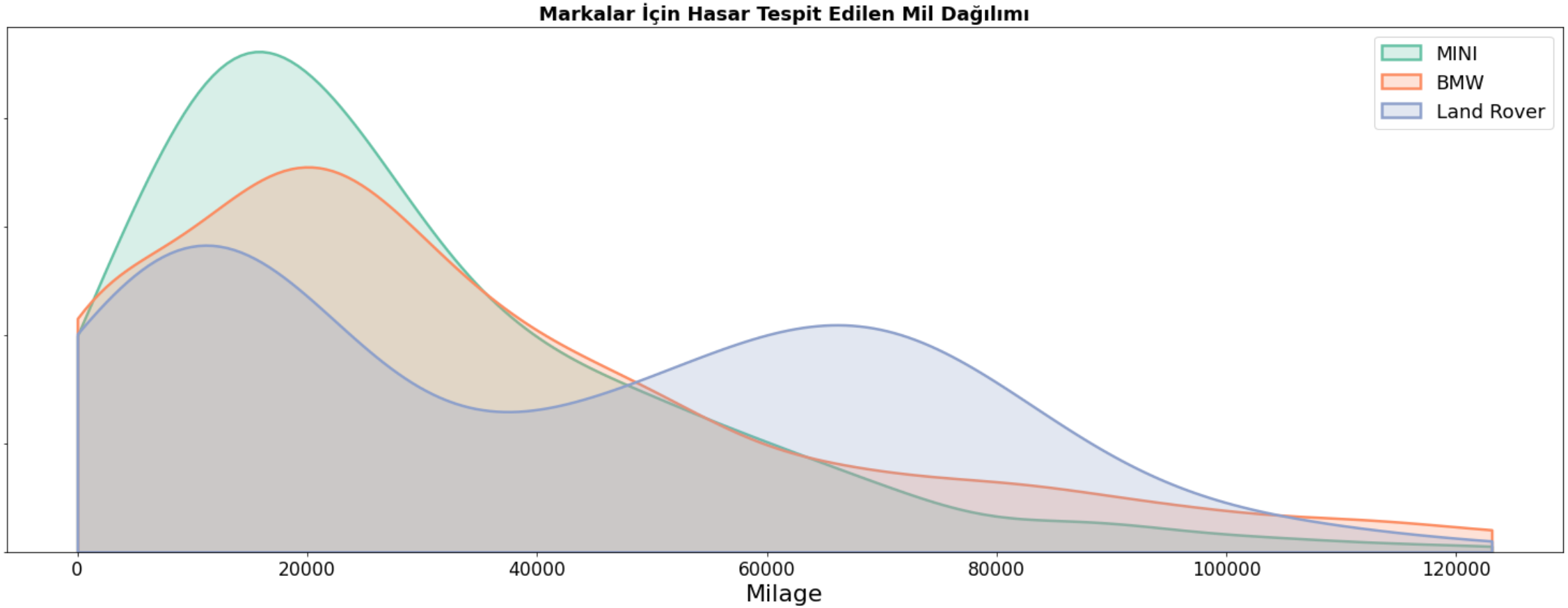
Adversarial validation: A binary classifier is trained to predict whether a sample belongs to test data or not. Training data highly similar to test data is sampled.



https://speakerdeck.com/upura/adversarial-validation-to-select-validation-data-for-evaluating-performance-in-e-commerce-purchase-intent-prediction

VERİ ANALİZİ

ARIZA ANALİZİ



VERİ ANALİZİ

«MAINGROUPDESCRIPTION»

Değer	Görülme Sayısı
Belirsiz	12238
Atölye El Kitabı Genel Bilgiler	693
Bakım	302
Motor	119

VERİ ANALİZİ

«SUBGROUPDESCRIPTION»

Değer	Görülme Sayısı
Belirsiz	13233
Termostat ve Bağlantılar	119

VERİ ANALİZİ

«ITEMDESCRIPTION»

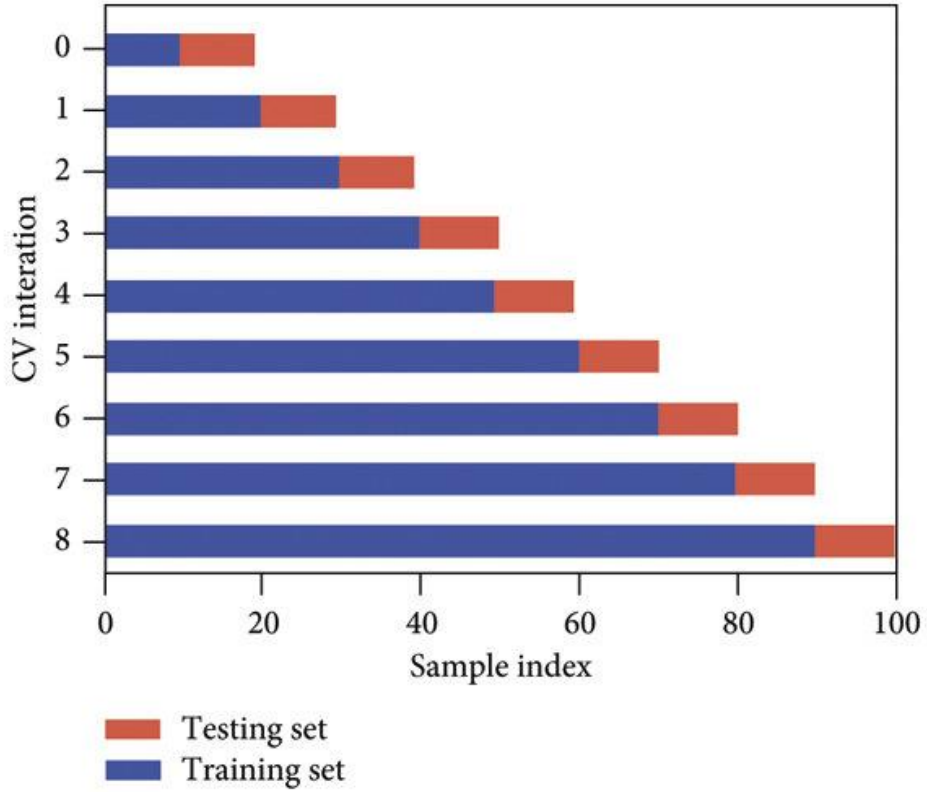
Değer	Görülme Sayısı
V kayışı	2728
Gergi Bilyası	2349
Vibrasyon Damper	2312
Dönel Titreşim Damperini Değiştirme	1864
Su Pompası	1064
Motor Askı Rotu	816
Krank Sensörü	627
Turbo	320
Genleşme Tankı	314
Turbo Radyatörü	302
Dönel Titreşim Damperinin Değiştirilmesi	301
Titreşim Damperi	122
Egzoz Gazı Turbo Ünitesi	120
Gergi	62
Su Pompa Kayışı	47
V Kayış	3
Gergi Bilyas	1

VERİ ANALİZİ

«ITEMDESCRIPTION»

Değer	Görülme Sayısı
V kayışı	2728
Gergi Bilyası	2349
Vibrasyon Damper	2312
Dönel Titreşim Damperini Değiştirme	1864
Su Pompası	1064
Motor Askı Rotu	816
Krank Sensörü	627
Turbo	320
Genleşme Tankı	314
Turbo Radyatörü	302
Dönel Titreşim Damperinin Değiştirilmesi	301
Titreşim Damperi	122
Egzoz Gazı Turbo Ünitesi	120
Gergi	62
Su Pompa Kayışı	47
V Kayış	3
Gergi Bilyas	1

VERİ OLUŞTURMA



MODEL SEÇİMİ

- CatBoost
- LGBM
- XGBoost

MODEL SEÇİMİ

- **CatBoost**
- LGBM
- XGBoost

MODEL GELİŞTİRME SÜRECİ

CLASS IMBALANCE

‘scale_pos_weight’

MODEL GELİŞTİRME SÜRECİ

FEATURE EXTRACTION

- **Son servisten itibaren geçen gün**
- **Son arızadan itibaren geçen gün**
- **Kayıt tarihinden itibaren geçen gün**
- **Toplam servis sayısı**
- **Toplam servis günü sayısı**
- **Toplam arıza sayısı**

- **Aracın teslim edilme sayısı**
- **Aracın özgün (unique) sahip sayısı**
- **Aracın bulunduğu özgün (unique) şehir sayısı**
- **Aracın son görülen kilometresi**
- **Tahminin yapıldığı ay**

- **Marka**
- **Seri**
- **Gövde Kodu**
- **Model Tanımı**
- **Model Kodu**
- **Model Yılı**
- **Model Kodu + Model Yılı**

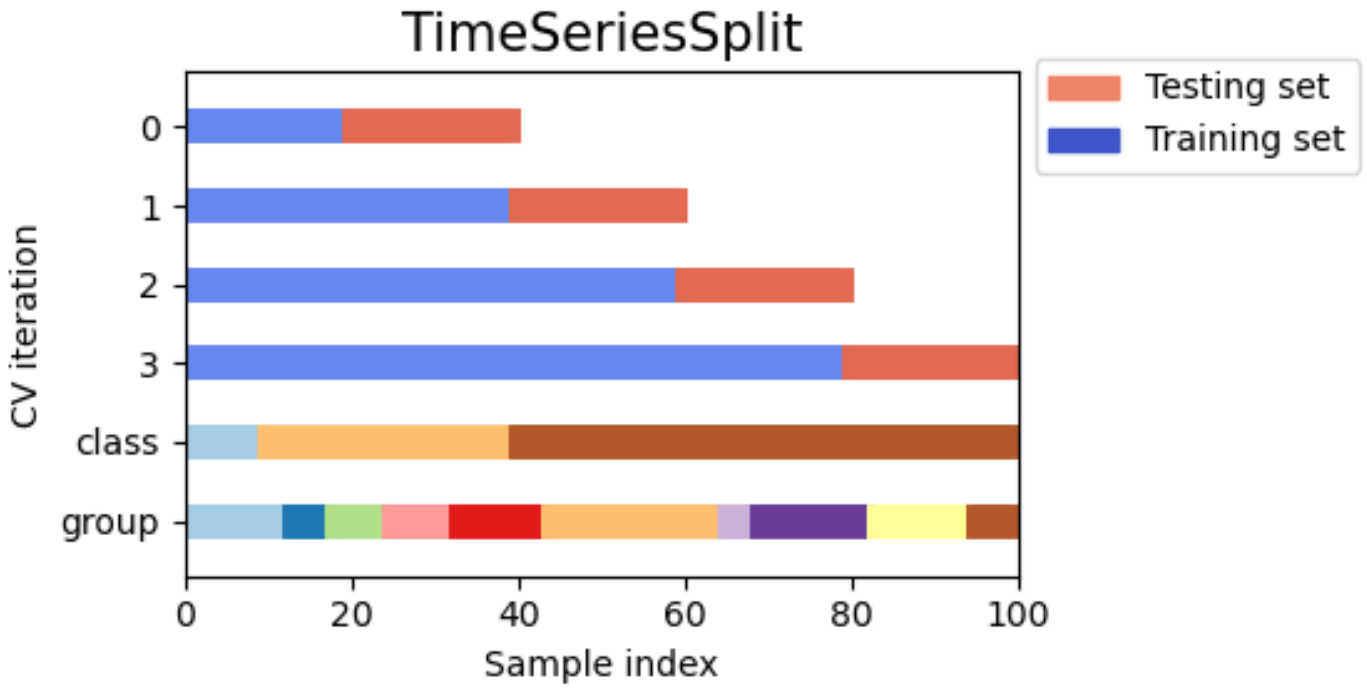
- **Zaman kısıtlı Mean Encoding**
  - Seri
  - Model Kodu + Model Yılı
  - Gövde Kodu
  - Vehicle ID

MODEL GELİŞTİRME SÜRECİ

FEATURE GENERATION

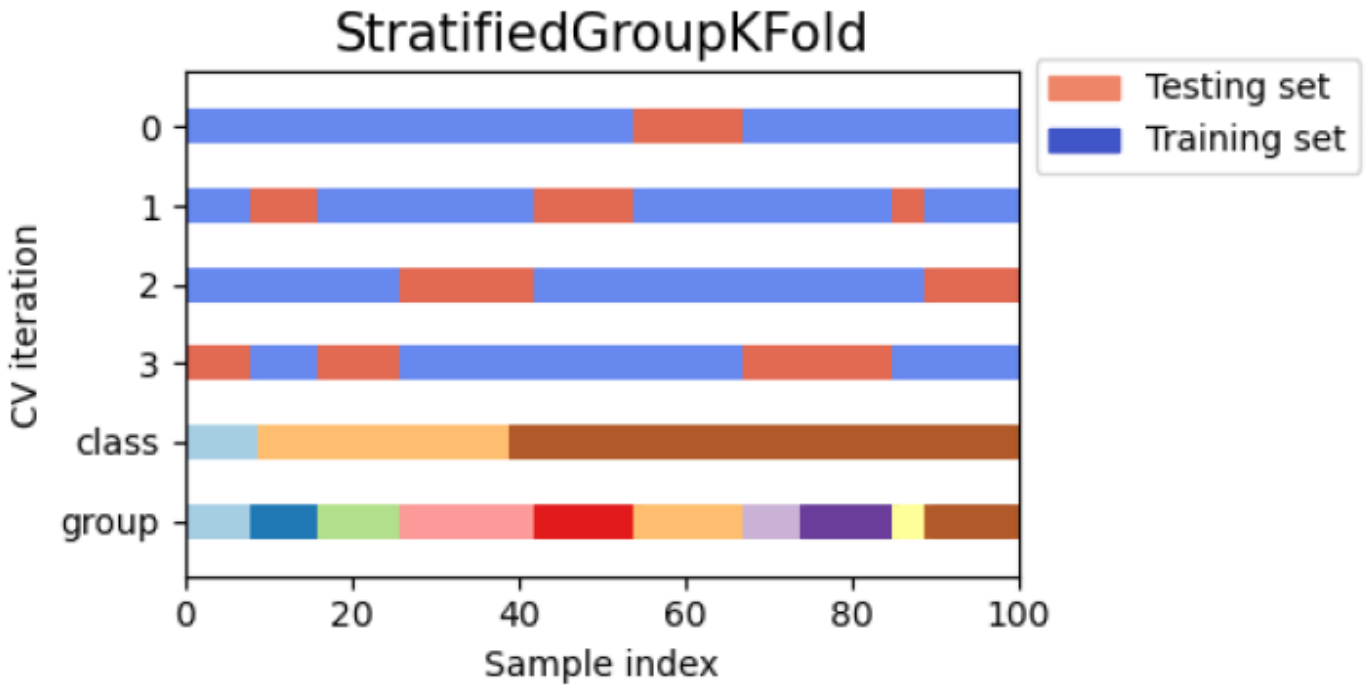
- **Tahmin dönemindeki yaklaşık kilometre** *(Son 2 servis kaydına bakarak approximation)*
- **Servis başına açılan ortalama parça kaydı**
- **Aracın servise gitme periyodu**
- **Aracın arızalanma periyodu**
- **Binned categories**

MODEL DOĞRULAMA SÜRECİ



MODEL DOĞRULAMA SÜRECİ

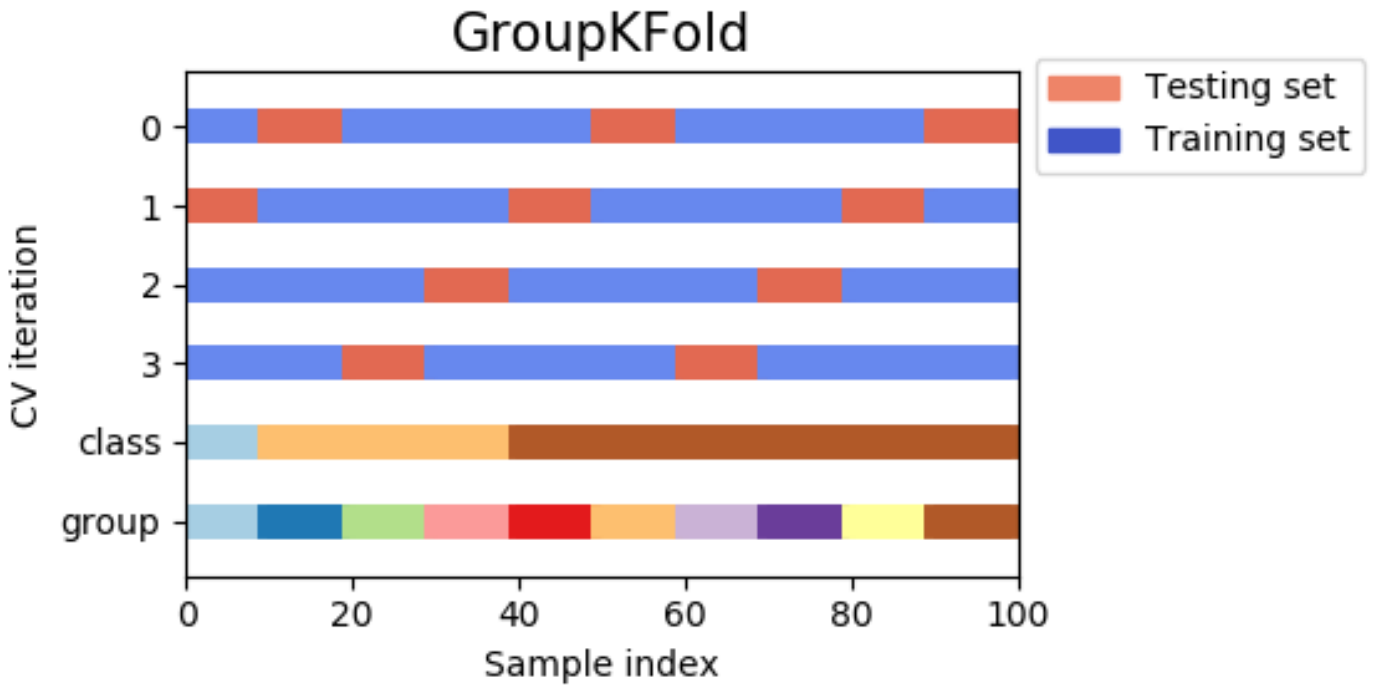
Vehicle ID



MODEL DOĞRULAMA SÜRECİ

Vehicle ID

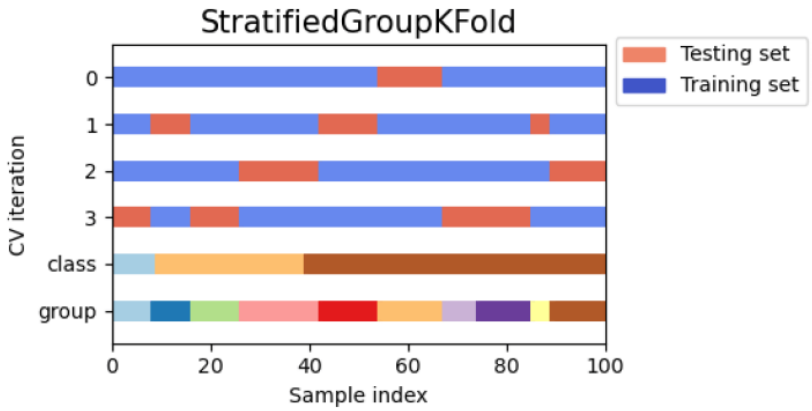
Prediction Season



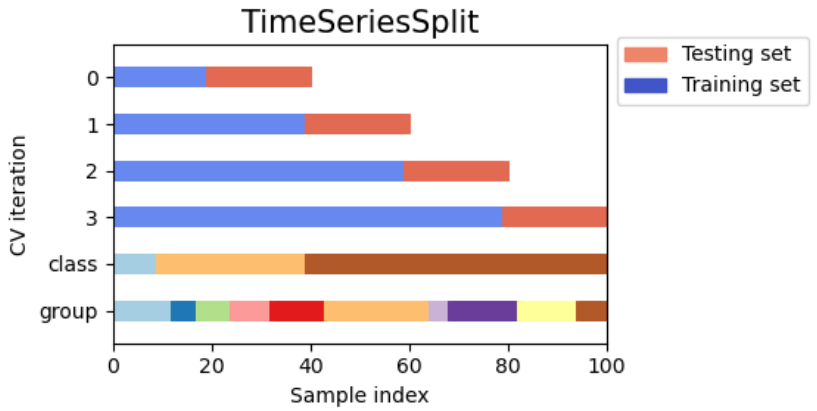
MODEL DOĞRULAMA SÜRECİ

Training

Vehicle ID



Sanity Check

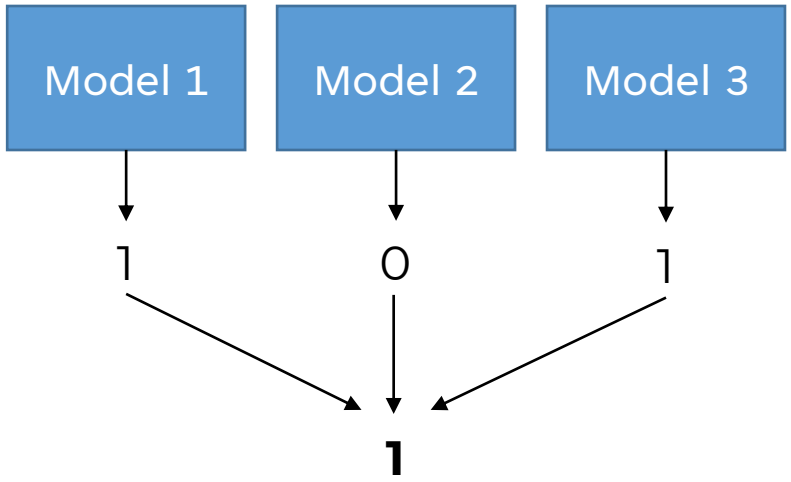


MODEL DOĞRULAMA SÜRECİ

	Private Score	Public Score	Use for Final Score
SGroupKFold – Vehicle	0.30416	0.28418	☐
GroupKFold - Time	0.30198	0.28423	☐
GroupKFold – Vehicle	0.30097	0.26927	☐
SGroupKFold – Vehicle	0.29919	0.27640	☐
GroupKFold – Time	0.29521	0.27994	☐
Seçilen Sub SGroupKFold – Vehicle	0.29455	0.28888	☐

TAHMİNLEME SÜRECİ

VOTING



DENENEN ALTERNATİFLER

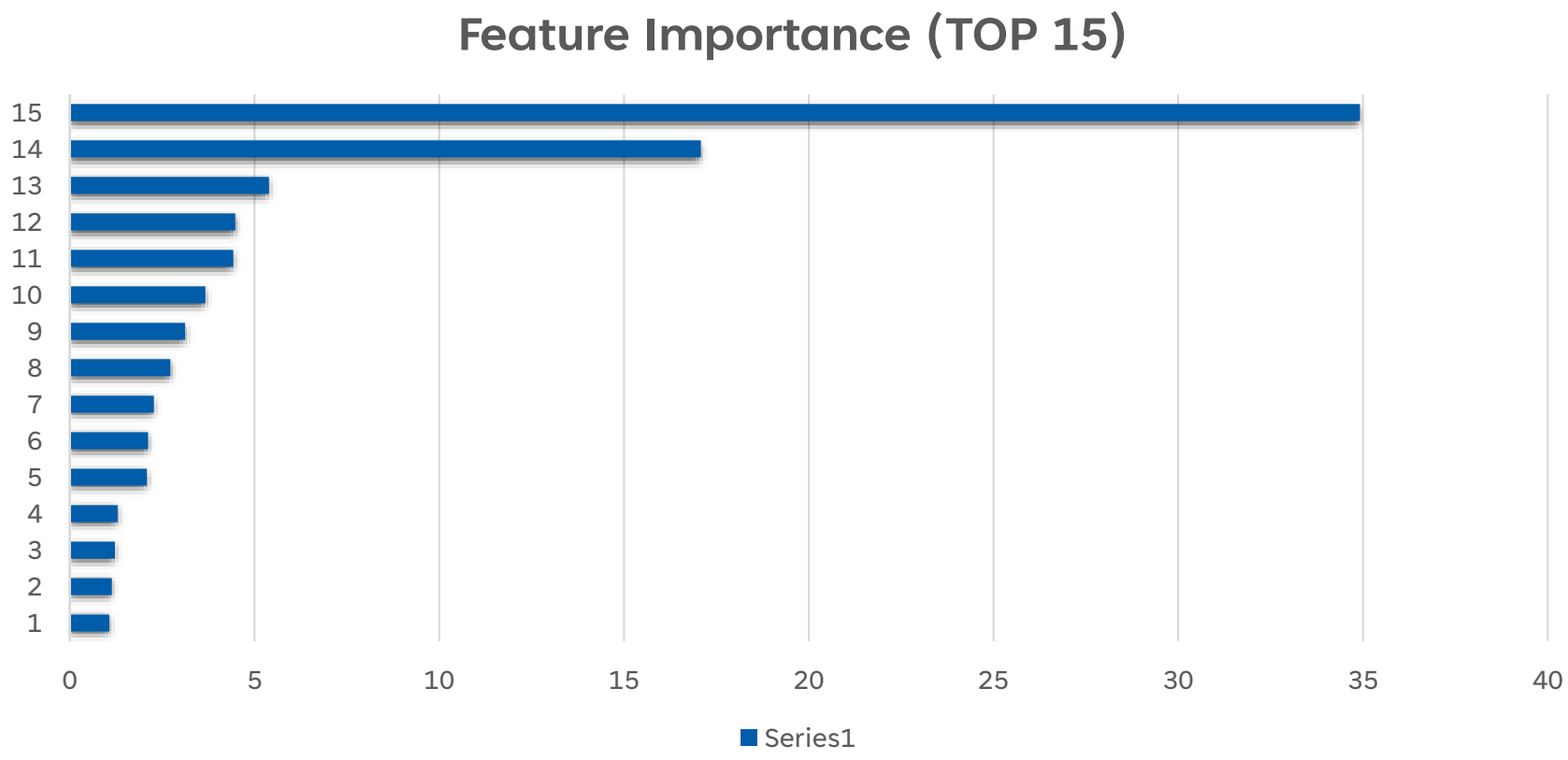
- Sayısal değerler üzerinde transformation
- Müşteri ve şube bazlı istatistikler
- Model Stacking

Ekstra veri kullanımı:

- USD-TRY aylık ve 3 aylık değişim oranları
- ÖTV aylık ve 3 aylık değişim oranları

ÇIKARIMLAR

ÖNEMLİ PARAMETRELER



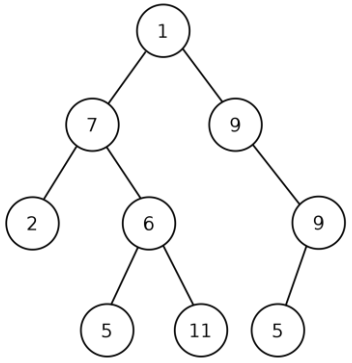
ÇIKARIMLAR

KURUMSAL BAĞLAM

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

PRECISION = *Arızalanacak denen kaç araç arızalandı?*

RECALL = *Arızalanacak araçların kaçı tespit edilebildi?*



Model Complexity ↑

Positive Count ↓

TEŞEKKÜRLER