



VERİ BİLİMİ BİTİRME PROJESİ RAPORU

Rossmann Mağaza Satışları Tahmini

Yasin Kahraman

Ağustos, 2026

softITo Yazılım Bilişim Akademisi, İstanbul

İÇİNDEKİLER

1. GİRİŞ

- 1.1. Projenin Amacı
- 1.2. Proje Kısıtları

2. MATERYAL VE YÖNTEM

- 2.1. Veri Seti
- 2.2. Sızıntısız Zaman Serisi Modelleme
- 2.3. Hedef Değişken Dönüşümü (Log1p)
- 2.4. Kapalı Mağaza Stratejisi
- 2.5. Kayan Pencere Çapraz Doğrulama (Walk-Forward)
- 2.6. Kullanılan Modeller ve Regularizasyon
- 2.7. Gecikme (Lag) Özellikleri ve Rekursif Tahmin
- 2.8. Topluluk (Ensemble) Yöntemi

3. UYGULAMA

- 3.1. Geliştirme Ortamı ve Not Defterleri
- 3.2. Veri Hattı
- 3.3. Keşifçi Veri Analizi (EDA) Bulguları
- 3.4. Özellik Mühendisliği
- 3.5. Doğrulama Stratejisinin Kurulması
- 3.6. Model Geliştirme ve Karşılaştırma
- 3.7. Optuna Hiperparametre Optimizasyonu
- 3.8. Gecikme Özellikleri ile Rekursif Simülasyon
- 3.9. Topluluk Modeli ve Nihai Sonuçlar
- 3.10. Test Tahmini ve Submission Üretimi

4. SONUÇ

5. KAYNAKÇA

1. GİRİŞ

Bu bölümde projenin amacı ve çalışma sırasında belirlenen kapsam kısıtlarına yer verilmiştir.

1.1. Projenin Amacı

Bu projenin amacı, Avrupa genelinde faaliyet gösteren **Rossmann** mağaza zincirine ait günlük satış miktarlarını, sızıntısız (data leakage-free) ve kronolojik bir makine öğrenmesi hattı ile tahmin etmektir. Çalışma, resmi Rossmann Mağaza Satışları yarışma verisi temel alınarak kurulmuş; mağaza, promosyon, tatil, rakip mağaza, hava durumu ve Google Trend sinyalleri tek bir modelleme tablosunda birleştirilmiştir.

Projenin operasyonel sorusu şudur: *Kronolojik olarak kilitli, hiçbir şekilde test dönemine sızıntı yapmayan bir öznitelik ve doğrulama tasarımıyla, mağaza bazlı satışlar resmi RMSPE (Root Mean Squared Percentage Error) metriğinde kabul edilebilir bir hata payıyla 6 haftalık bir ufuk için tahmin edilebilir mi?* Bu amaçla XGBoost ve LightGBM tabanlı modeller eğitilmiş, Optuna ile hiperparametre araması yapılmış ve modeller ağırlıklı bir topluluk (ensemble) yapısında birleştirilerek nihai tahmin üretilmiştir.

1.2. Proje Kısıtları

Kısıt 1: Modelleme yalnızca yarışmanın temel veri kümesi (train.csv, test.csv, store.csv) ve iki harici veri kaynağı (weather.csv, googletrend.csv) ile sınırlıdır. Üçüncü parti gerçek zamanlı veri beslemesi kullanılmamıştır.

Kısıt 2: Test seti (41.088 satır, 2015-08-01 → 2015-09-17 dönemi) model seçimi, hiperparametre ayarı veya eşik belirleme amacıyla **kesinlikle** kullanılmamıştır; yalnızca nihai submission.csv üretiminde bir kez skorlanmıştır.

Kısıt 3: Kapalı mağaza günleri (Open == 0), veri setinin yaklaşık %17'sini oluşturmaktadır. Bu günler kural bazlı olarak eğitim ve doğrulama kümelerinden izole edilmiş; tahmin anında bu günlere doğrudan 0 satış atanmıştır.

Kısıt 4: Doğrulama, rastgele bölme (train_test_split) yerine **takvim tarihine dayalı kronolojik bölme** ile yapılmıştır; böylece doğrulama kümesinde eğitim dönemiyle çakışan tarihler bulunmaz.

Kısıt 5: Gecikme (lag) özellikleriyle kurulan model, test döneminde gerçek geçmiş satışa erişemediği için yalnızca **rekursif simülasyon** ile değerlendirilmiştir; bu modelin doğrulama skoru da aynı rekursif kısıt altında ölçülmüştür.

2. MATERYAL VE YÖNTEM

Bu bölümde veri seti, sızıntı denetimi, hedef dönüşümü, doğrulama stratejisi, model ailesi ve topluluk yöntemine ilişkin kavramsal çerçeve özetlenmiştir.

2.1. Veri Seti

Proje beş ham dosya üzerine kuruludur:

Dosya	Satır	Kolon	Rol
train.csv	1.017.209	9	Mağaza bazlı günlük satış (etiketli)
test.csv	41.088	8	Tahmin edilecek dönem (2015-08-01 → 2015-09-17)
store.csv	1.115	10	Mağaza meta verisi (tip, rakip, promosyon)
weather.csv	15.840	24	Eyalet bazlı günlük hava durumu
googletrend.csv	2.072	3	Haftalık "Rossmann" arama trendi (ülke ve eyalet)

İşleme sonunda üretilen birleşik modelleme tabloları:

Tablo	Boyut	Tarih Aralığı
train_final.parquet	$1.017.209 \times 49$	2013-01-01 → 2015-07-31
test_final.parquet	41.088×48	2015-08-01 → 2015-09-17

2.2. Sızıntısız Zaman Serisi Modelleme

Projenin ilk sürümünde veri Store + Date sırasına göre dizilip satır bazında %85/%15 olarak bölünmüştü. Bu yaklaşım iki temel sorun barındırıyordu:

- Doğrulama kümesinde eğitim sırasında hiç görülmemiş mağazalar bulunuyordu,
- Doğrulama ve eğitim dönemleri takvimsel olarak **çakışıyordu**, dolayısıyla ölçülen skor gerçek görevi (geleceği tahmin etmeyi) yansıtmıyordu.

Bu nedenle bölme stratejisi **takvim tarihine göre** yeniden tasarlanmıştır: son 6 hafta doğrulama kümesi, geri kalan tüm geçmiş eğitim kümesi olarak ayrılmıştır. Doğrulama penceresinin uzunluğu (6 hafta), gerçek test dönemiyle birebir aynı tutularak doğrulama skorunun test performansını temsil kabiliyeti artırılmıştır.

2.3. Hedef Değişken Dönüşümü (Log1p)

Keşifçi veri analizinde günlük satış dağılımının belirgin biçimde sağa çarpık olduğu görülmüştür (çarpıklık katsayısı ham veride **1,59**). Bu çarpıklığı gidermek ve yarışmanın resmi metriği olan **RMSPE** (görelî/yüzdesel bir hata metriği) ile öğrenme hedefini uyumlu hale getirmek amacıyla hedef değişkene **$\log(1+\text{Satış})$** dönüşümü uygulanmıştır. Dönüşüm

sonrası çarpıklık **-0,11**'e inerek dağılım büyük ölçüde simetrikleşmiştir. Model çıktıları expm1 ile tekrar gerçek satış birimine çevrilmiştir.

2.4. Kapalı Mağaza Stratejisi

Verinin %17,0'ı (172.817 satır) kapalı mağaza günlerinden oluşmaktadır; bu günlerde satış tanım gereği 0'dır ve modele gürültü olarak katılmaması gerekir. Ayrıca açık olup satışı 0 olan 54 kayıt (%0,01) tespit edilmiş; bunlar muhtemel tadilat/stok sayımı günleri olarak değerlendirilip aykırı değer kabul edilmiştir. Bu doğrultuda model yalnızca **açık ve satış > 0** olan günlerde eğitilmiş; kapalı günler için nihai tahminde doğrudan **0** atanmıştır.

2.5. Kayan Pencere Çapraz Doğrulama (Walk-Forward)

Tek bir doğrulama penceresinin şansa bağlı olabileceği göz önünde bulundurularak, model **3 üst üste binmeyen 6 haftalık genişleyen pencerede (expanding window)** yeniden değerlendirilmiştir. Her kat için eğitim verisi, doğrulama penceresinin başlangıcından önceki **tüm** geçmiş veridir; bu yapı Noel dönemi gibi güçlü mevsimselliğin skoru nasıl etkilediğini de gözlemlenebilir kılmıştır.

2.6. Kullanılan Modeller ve Regularizasyon

Ana model ailesi gradyan artırmalı ağaçlardır: **XGBoost** ve **LightGBM**. Aşırı öğrenmeyi önlemek amacıyla:

- L1 ve L2 ceza terimleri (reg_alpha, reg_lambda) uygulanmış,
- Ağaç derinliği (max_depth) ve yaprak sayısı (num_leaves) sınırlandırılmış,
- Satır ve kolon alt-örnekleme (subsample, colsample_bytree) yapılmış,
- Erken durdurma (early_stopping_rounds) ile doğrulama setinde iyileşme durduğunda eğitim kesilmiştir.

Bu hiperparametrelerin bir kısmı **Optuna** kütüphanesi ile TPE (Tree-structured Parzen Estimator) algoritması kullanılarak sistematik biçimde aranmıştır.

2.7. Gecikme (Lag) Özellikleri ve Rekursif Tahmin

Bir mağazanın kendi geçmiş satışı, en güçlü sinyallerden biridir. Bu amaçla Sales_Lag7, Sales_Lag14 ve Sales_Roll28 (son 28 günlük hareketli ortalama) özellikleri türetilmiştir. Ancak bu yaklaşım iki kritik riski beraberinde getirir:

1. **Sızıntı riski:** Gecikme özellikleri yalnızca kesinlikle geçmişte kalan satışlardan (shift(7) ve üzeri) üretilmelidir. İlk 28 günde gecikme değeri bulunmadığından bu satırlar eğitimden düşülmüş, medyan gibi bir değerle doldurulmamıştır (bu tür bir doldurma sızıntı sayılmıştır).
2. **Gelecekte gecikme yokluğu:** Test döneminde gerçek satış henüz mevcut değildir. Bu nedenle **rekursif tahmin** simülasyonu kurulmuştur: gün gün ilerlenir, modelin bir önceki günkü tahmini bugünün gecikme özelliği olarak geri beslenir. Aynı simülasyon adilliği sağlamak amacıyla doğrulama penceresinde de birebir koşturulmuştur. Kapalı günlerin satışı takvimden önceden bilindiğinden simülasyona doğrudan 0 olarak yazılmıştır.

2.8. Topluluk (Ensemble) Yöntemi

Farklı hata örüntülerine sahip modellerin ağırlıklı ortalaması, çoğunlukla tekil modelden daha güçlü ve daha kararlı sonuç verir. Bu projede dört adayın (temel XGBoost + log hedef, LightGBM, Optuna-XGBoost, Lag-XGBoost) ağırlıkları, doğrulama penceresinde RMSPE'yi minimize edecek şekilde 0,05 adımlı konveks bir kombinasyon ızgarası taranarak bulunmuştur.

3. UYGULAMA

Bu bölümde ortam kurulumu, veri hattı, EDA bulguları, özellik mühendisliği, model geliştirme adımları ve nihai sonuçların uygulama detaylarına yer verilmiştir.

3.1. Geliştirme Ortamı ve Not Defterleri

Proje Python 3 ile geliştirilmiş ve üç ayrı Jupyter not defterine bölünmüştür:

Not Defteri	İşlev
RosmannVeriYukleme.ipynb	Ham verilerin yüklenmesi, birleştirilmesi, ön işlenmiş parquet çıktısının üretilmesi
RosmannEDA.ipynb	Keşifçi veri analizi; dağılım, mevsimsellik ve etki analizleri
Rosmann_Model.ipynb	Özellik mühendisliği, doğrulama, model eğitimi, Optuna araması, ensemble ve submission.csv üretimi

Kullanılan başlıca kütüphaneler: pandas, numpy, xgboost, lightgbm, optuna, scikit-learn, matplotlib, seaborn.

3.2. Veri Hattı

3.2.1. Ham Verilerin Yüklenmesi

train.csv, test.csv, store.csv ile harici store_states.csv, state_names.csv, googletrend.csv ve weather.csv dosyaları okunmuştur. StateHoliday kolonu karışık veri tipi uyarısını (DtypeWarning) önlemek amacıyla açıkça str olarak okutulmuştur.

3.2.2. Google Trend Verisinin İşlenmesi

Haftalık googletrend tablosu aylık periyoda indirgenmiş ve iki ayrı özellik türetilmiştir:

- **Trend_DE:** Almanya geneli aylık ortalama arama trendi,
- **Trend_State:** Mağazanın bağlı olduğu eyalete özgü aylık ortalama arama trendi.

Google Trend verisinde bulunmayan eyaletler (HB, NI) için ülke geneli **Trend_DE** yedek değer olarak kullanılmıştır.

3.2.3. Birleştirme (Merge) Stratejisi

Tüm tablolar, sessiz veri çoğalmasını (duplicate satır üretimini) önlemek amacıyla validate='many_to_one' parametresiyle birleştirilmiştir. Bu, her train/test satırının store, weather ve googletrend tablolarında **en fazla bir** eşleşen kayıtlı birleşmesini garanti eder;

birleştirme sonrası satır sayısının orijinal train/test satır sayısı ile birebir tutarlı kalması doğrulanmıştır.

3.2.4. Eksik Değerlerin Doldurulması ve Türetilmiş Özellikler

Mağaza ve hava durumu tablolarındaki eksik değerler kural bazlı doldurulmuş; rakip mağaza açılış bilgisi bilinmeyen kayıtlarda 0 kabul edilmiştir. Ayrıca mağazanın uzun vadeli **Promo2** kampanyasının ilgili ayı kapsayıp kapsamadığını gösteren **IsPromo2Month** ikili (binary) özelliği üretilmiştir. İşlem sonunda modelleme tablosu **49 kolona** (train) ulaşmıştır.

3.3. Keşifçi Veri Analizi (EDA) Bulguları

EDA çalışması, açık ve satış > 0 olan **844.338 satırlık** ana analiz tablosu üzerinde yürütülmüştür. Öne çıkan bulgular:

- **Kapalı günler:** Verinin %17,0'ı kapalı mağaza günüdür; açık olup 0 satış yapan 54 kayıt aykırı değer olarak işaretlenmiştir.
- **Çarpıklık:** Ham satış dağılımının çarpıklığı 1,59 iken $\log_{10}$ sonrası $-0,11$ 'e düşmüştür.
- **Promosyon etkisi:** Promo aktifken ortalama satış 5.930 €'dan 8.229 €'ya çıkarak yaklaşık %39 artış göstermiştir; uzun vadeli Promo2/IsPromo2Month etkisi buna kıyasla daha zayıf kalmıştır.
- **Haftanın günleri:** Pazar günleri mağazaların neredeyse tamamı kapalıdır (haftalık kayıtların yalnızca 3.593'ü Pazar gününe aittir); açık kalan az sayıdaki mağaza yüksek ciro üretmektedir.
- **Tatil etkisi:** Okul tatili yokken ortalama satış 6.897 €, varken 7.201 €'dur; hafif pozitif bir etki gözlenmiştir.
- **Rakip mesafesi:** CompetitionDistance ile satış arasındaki doğrusal korelasyon zayıftır ($-0,046$); rakip etkisinin doğrusal olmayan/çarpık bir davranış sergilediği değerlendirilmiştir.
- **Mevsimsellik:** Aralık ayı (Noel dönemi) belirgin bir satış zirvesidir; yıl $\times$ ay ısı haritasında güçlü mevsimsel örüntü gözlenmiştir.
- **Mağaza bazlı farklılık:** Mağaza ortalamaları 2.704 € ile 21.757 € arasında değişmekte olup en iyi/en kötü mağaza oranı **8,0 kata** ulaşmaktadır; bu da Store kimliğinin güçlü bir öznelik olduğunu göstermektedir.

3.4. Özellik Mühendisliği

Modelleme öncesi aşağıdaki türetilmiş özellikler oluşturulmuştur:

- **CompOpenAy:** Rakip mağazanın açılışından ilgili güne kadar geçen ay sayısı; bilinmeyen durumlarda (yıl bilgisi 0 doldurulmuş kayıtlar) 0 kabul edilip negatif değerler kırılmıştır ($\text{clip}(\text{lower}=0)$).
- **Döngüsel (cyclical) kodlama:** Ay ve haftanın günü bilgisi $\sin/\cos$ çiftleriyle kodlanarak "Aralık ile Ocak birbirine uzak" gibi yapay bir mesafe algısının önüne geçilmiştir.

- **Kategorik kodlama:** Kategorik kolonlar LabelEncoder ile sayısallaştırılmış; encoder **train + test birleşik** kümesiyle fit edilerek test setinde görülüp eğitimde görülmeyen bir kategori nedeniyle hata oluşması engellenmiştir.

Bu adımların sonunda hedef hariç **50 özellik** modelleme matrisine dahil edilmiştir.

3.5. Doğrulama Stratejisinin Kurulması

Bölme takvim tarihine göre yapılmış; son 6 hafta doğrulama kümesi olarak ayrılmıştır:

Küme	Dönem	Satır Sayısı	Mağaza Sayısı
Eğitim	2013-01-01 → 2015-06-18	802.942	1.115
Doğrulama	2015-06-19 → 2015-07-31	41.396	1.115

Doğrulama penceresinin uzunluğu (6 hafta), gerçek test döneminin uzunluğuyla birebir aynıdır; böylece ölçülen skor gerçek görevi doğru biçimde taklit eder. Resmi yarışma metriği olan **RMSPE** tüm karşılaştırmalarda esas alınmıştır.

3.6. Model Geliştirme ve Karşılaştırma

Sırasıyla aşağıdaki modeller kurulmuş ve tek doğrulama penceresinde karşılaştırılmıştır:

Model	RMSE	MAE	R ²	RMSPE
XGBoost Baseline (ham hedef)	960,1	681,9	0,9005	%14,91
XGBoost + Log Hedef	873,9	597,9	0,9176	%12,32
LightGBM + Log Hedef	895,6	614,3	0,9134	%12,51

Log dönüşümünün etkisi belirgindir: yalnızca hedef değişkeni log1p ile dönüştürmek RMSPE'yi %14,91'den %12,32'ye indirmiştir (yaklaşık 2,6 puanlık iyileşme).

Kayan pencere (walk-forward) doğrulaması ile 3 katta ölçülen sonuçlar:

Kat	Doğrulama Penceresi	RMSPE
1	2015-03-27 → 2015-05-08	%13,07
2	2015-05-08 → 2015-06-19	%12,08
3	2015-06-19 → 2015-07-31	%12,65

CV RMSPE ortalaması: %12,60 (std: 0,40). Katlar arasındaki düşük standart sapma, modelin farklı dönemlerde tutarlı bir performans sergilediğini göstermektedir.

3.7. Optuna Hiperparametre Optimizasyonu

Elle deneme yerine Optuna'nın TPE algoritmasıyla sistematik bir arama yapılmıştır. Arama maliyetini sınırlamak için ağaç bütçesi kısılmış ($n_estimators=700$, erken durdurma 40) ve arama tek doğrulama penceresi üzerinde koşturulmuştur.

En iyi bulunan parametreler ve doğrulama sonucu:

Parametre	Değer
learning_rate	0,0504
max_depth	10
min_child_weight	12,0571
subsample	0,8395
colsample_bytree	0,5702
reg_lambda	0,7979
Optuna arama skoru (RMSPE)	%12,48

Bulunan parametreler daha uzun bir eğitim bütçesiyle (3.000 ağaç, erken durdurma 100 tur; öğrenme oranı uzun eğitime uyum için 0,7 katsayısıyla kırpılmıştır) yeniden eğitilmiş ve nihai **XGBoost Optuna** modeli elde edilmiştir:

Model	RMSE	MAE	R ²	RMSPE
XGBoost Optuna (final)	869,3	595,7	0,9184	%12,26

Bu sonuç, kaba parametrelerle eğitilen log-hedefli XGBoost'a (%12,32) kıyasla marjinal ama istikrarlı bir iyileşme sağlamıştır.

3.8. Gecikme Özellikleri ile Rekursif Simülasyon

Sales_Lag7, Sales_Lag14 ve Sales_Roll128 özellikleriyle kurulan model, 773.431 satırlık gecikmeli eğitim kümesi üzerinde eğitilmiştir. Doğrulama ve test tahminleri, gün gün ilerleyen rekursif simülasyon ile üretilmiştir:

Model	RMSE	MAE	R ²	RMSPE
XGBoost Lag (rekursif)	1.018,5	723,1	0,8880	%15,38

Bu modelin diğer modellere kıyasla daha yüksek hata vermesi beklenen bir sonuçtur: model, kendi ürettiği önceki tahminler üzerinden ilerlediği için hata birikimi (error accumulation) etkisine maruz kalmaktadır. Buna rağmen model, ensemble aşamasında değerlendirmeye alınmıştır.

3.9. Topluluk Modeli ve Nihai Sonuçlar

Dört adayın (XGBoost + Log Hedef, LightGBM, XGBoost Optuna, XGBoost Lag) tahminleri, doğrulama penceresinde RMSPE'yi minimize eden ağırlıklarla birleştirilmiştir. Ağırlıklar 0,05 adımlı bir konveks kombinasyon ızgarası taranarak bulunmuştur:

Model	Ensemble Ağırlığı
XGBoost Optuna	0,50
XGBoost + Log Hedef	0,31
LightGBM	0,19
XGBoost Lag (rekursif)	0,00

Lag modelinin ağırlığının 0 çıkması, rekursif hata birikiminin ensemble için katkısını ortadan kaldırdığını göstermektedir; buna rağmen modelin sürece dahil edilmesi metodolojik olarak yanlış-negatif bir varsayımı önlemiştir.

Nihai model karşılaştırması (doğrulama kümesi, sıralı):

Model	RMSE	MAE	R ²	RMSPE
Ensemble (Topluluk)	861,2	588,5	0,9200	%12,14
XGBoost Optuna	869,3	595,7	0,9184	%12,26
XGBoost + Log Hedef	873,9	597,9	0,9176	%12,32
LightGBM + Log Hedef	895,6	614,3	0,9134	%12,51
XGBoost Baseline	960,1	681,9	0,9005	%14,91
XGBoost Lag (rekursif)	1.018,5	723,1	0,8880	%15,38

Topluluk modeli, tüm tekil modelleri geride bırakarak **%12,14 RMSPE** ile en iyi doğrulama skorunu elde etmiştir. Bu sonuç, XGBoost baseline'a kıyasla RMSPE'de yaklaşık **2,8 puanlık (~%18,6 görelî)** iyileşmeye karşılık gelmektedir.

3.10. Test Tahmini ve Submission Üretimi

Doğrulamada en iyi sonucu veren yaklaşım (Ensemble) esas alınarak test dönemi (2015-08-01 → 2015-09-17, 41.088 satır) için tahmin üretilmiştir. Süreç:

1. Açık günler (Open == 1) için dört aday modelin test tahminleri hesaplanmış, negatif değerler 0'a kırılmıştır (clip(0, None)).
2. Doğrulamada en iyi skoru veren yaklaşım (Ensemble) test tahmininde de kullanılmıştır.
3. Kapalı mağaza günlerine (**5.984 satır**) doğrudan **0** satış atanmıştır.
4. Sonuçlar Id sırasına göre sıralanarak submission.csv dosyasına yazılmıştır.

Sonuç dosyası özeti:

Metrik	Değer
Submission için seçilen yaklaşım	ENSEMBLE
Toplam tahmin satırı	41.088
Kapalı gün (0 atanan) satır	5.984
Tahmin değer aralığı	0 – 29.052 €

4. SONUÇ

Bu bitirme çalışmasında, Rossmann mağaza zincirine ait günlük satışlar; sızıntısız, kronolojik olarak kilitli ve regülasyona sahip bir makine öğrenmesi hattı ile tahmin edilmiştir. Ham veri yükleme aşamasından nihai `submission.csv` üretimine kadar tüm adımlarda test dönemi model seçimi veya hiperparametre ayarı için kullanılmamış; doğrulama tamamen takvim bazlı ve test dönemiyle aynı uzunlukta (6 hafta) kurulmuştur.

Çalışmanın öne çıkan teknik katkıları şunlardır:

- **Sızıntı denetimi:** many_to_one birleştirme, kronolojik bölme ve kapalı gün izolasyonu ile veri sızıntısı riskleri sistematik olarak kapatılmıştır.
- **Hedef dönüşümü:** log1p dönüşümü, sağa çarpık satış dağılımını düzeltmiş ve RMSPE ile öğrenme hedefi arasındaki uyumu artırarak RMSPE'yi tek başına ~2,6 puan iyileştirmiştir.
- **Sağlam doğrulama:** 3 katlı kayan pencere çapraz doğrulama (ortalama RMSPE %12,60, std 0,40), tek pencereye aşırı uyum riskini azaltmıştır.
- **Regülasyonlu hiperparametre araması:** Optuna ile yapılan sistematik arama, elle ayarlanan parametrelere kıyasla istikrarlı bir iyileşme (RMSPE %12,26) sağlamıştır.
- **Disiplinli gecikme mühendisliği:** Lag özellikleri yalnızca kesin geçmiş veriden üretilmiş, test dönemi için rekursif simülasyon ile adil biçimde değerlendirilmiştir.
- **Topluluk modeli:** Ağırlıklı ensemble (%50 Optuna-XGBoost, %31 XGBoost, %19 LightGBM), tüm tekil modelleri geride bırakarak **%12,14 RMSPE** ile en iyi sonucu vermiştir.

Sonuç olarak proje; sızıntı kontrollü, çapraz doğrulama ile seçilmiş, kilitli test dönemi üzerinde bir kez skorlanmış ve gerçek bir `submission.csv` çıktısı üreten uçtan uca bir teslimattır. Elde edilen yapı, kavramsal olarak zaman serisi sızıntı denetimini, pratik olarak da endüstri standardında bir satış tahmin hattını belgelemektedir. İleri çalışma olarak CatBoost denemesi, katmanlı istifleme (stacking), hiperparametre aramasının çapraz doğrulama ortalamasına dayandırılması ve resmi Alman tatil takvimi ile mağaza bazlı hedef kodlaması gibi ek özellikler önerilmektedir.

5. KAYNAKÇA

- [1] <https://www.kaggle.com/c/rossmann-store-sales>
- [2] <https://xgboost.readthedocs.io/>
- [3] <https://lightgbm.readthedocs.io/>
- [4] <https://optuna.readthedocs.io/>
- [5] <https://scikit-learn.org/stable/>
- [6] <https://pandas.pydata.org/docs/>
- [7] Hyndman, R. J. ve Athanasopoulos, G. *Forecasting: Principles and Practice*.
- [8] <https://seaborn.pydata.org/>