FET 445 Veri Madenciliği AutoDataMinds

ARAÇ FİYAT TAHMİNİ VE PAZAR SEGMENTASYONU

ASLI ERBAŞI 22040101040
NURETTİN KAPLAN 22040301031

TARİH 25.12.25

Araç Fiyat Tahmini ve Pazar Segmentasyonu

- Projemiz iki temel probleme odaklanmaktadır:
- **Fiyat Tahmini (Regresyon):** İkinci el araç piyasasındaki belirsizliği gidermek amacıyla; araç yaşı, kilometresi ve motor hacmi gibi değişkenleri kullanarak objektif bir fiyat tahminleme modeli geliştirmek.
- **Pazar Segmentasyonu (Kümeleme):** Fiyat bilgisini kullanmadan, araçların teknik özelliklerine göre "ekonomik", "lüks" veya "performans" odaklı doğal gruplarını keşfetmek ve piyasa yapısını analiz etmek.

Sayfa 3: Veri Seti Açıklaması

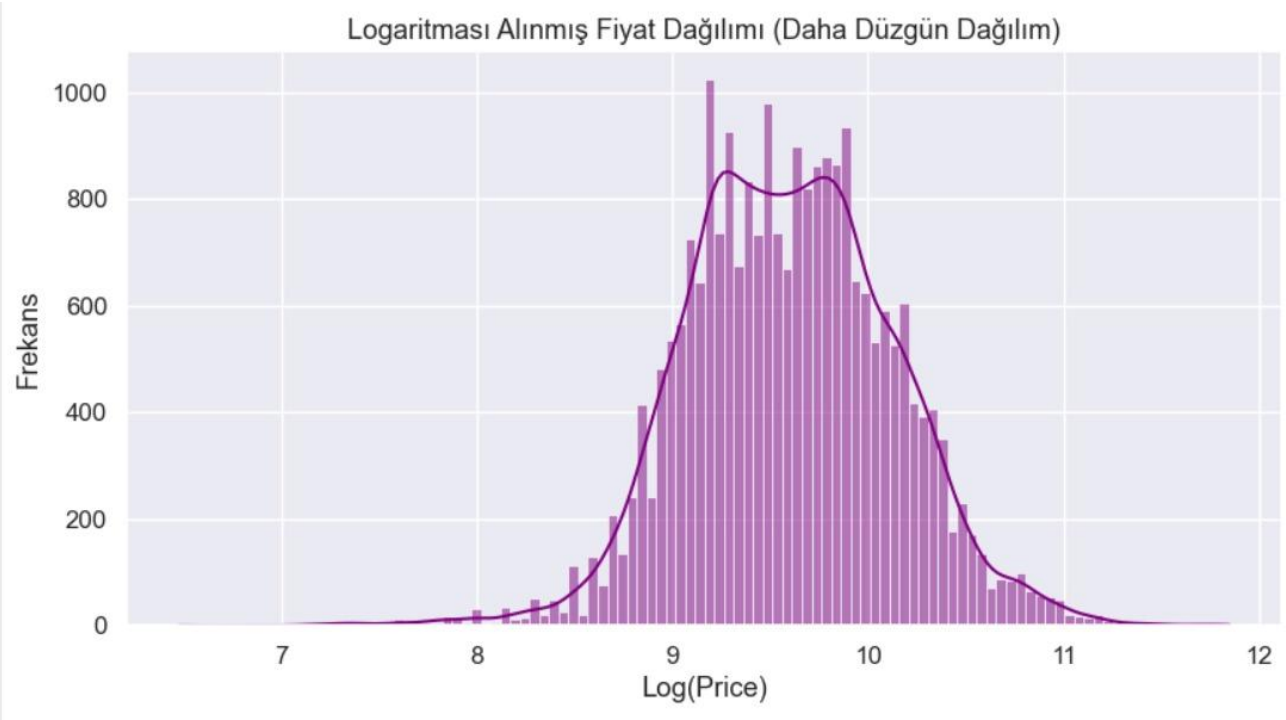
•**Kaynak:** ekip_odevi_ham_veri_30k.csv (Yaklaşık 30.000 kayıt).

•**Boyut:** 30.000 Satır x 17Özellik.

•**Özellik Tipleri:**
•**Kategorik:** Marka, Model, Şanzıman Tipi, Yakıt Tipi.
•**Sayısal:** Yıl, Kilometre (Mileage), Vergi (Tax), Yakıt Tüketimi (MPG), Motor Hacmi (EngineSize).

•**Veri Dağılımı:Dengeli mi?"** "Hayır, veri setimiz ham haliyle hem hedef değişken (Price) hem de
• kategorik özellikler bakımından **dengesiz (unbalanced)** bir yapıdadır. Fiyat dağılımı sağa
•çarpıktır. Bu durumun modelin yanlı tahminler yapmasına yol açmaması için
• **Logaritmik Dönüşüm** ve **RobustScaler** teknikleri uygulanarak veri seti istatistiksel
• bir dengeye kavuşturulmuştur."

Logoritmik dönüşümler sonrası dağılım:



Ön İşleme ve Metrikler

- **Temel Feature Engineering:** * SimpleImputer ile eksik veriler ortalama (mean) stratejisiyle doldurulduu.
  - Kategorik veriler Label Encoding ile sayısal değerlere dönüştürüldü.
  - Tüm sayısal veriler StandardScaler ile normalize edildi.

- **Train-Test Split:** Veri seti **%80 Eğitim** ve **%20 Test** olarak ayrıldı.

- Performans Metrikleri:
- 1. R2Score: Modelin veriyi açıklama başarısı.

2. MAE (Mean Absolute Error): Tahminlerdeki ortalama TL bazlı sapma.
3. MAPE (Mean Absolute Percentage Error): Yüzdesel hata payı (Modelimiz yaklaşık %9.8 hata ile çalışmaktadır).

Best Model 1 – XGBoost -ASLI

•**Uçtan Uca Pipeline:** Veri sızıntısını önlemek amacıyla SimpleImputer (eksik değer tamamlama), StandardScaler (ölçeklendirme) ve XGBRegressor algoritması tek bir **Pipeline** yapısı altında birleştirilmiştir.

•**Veri Standardizasyonu:** Modelin gradyan tabanlı öğrenme sürecini hızlandırmak ve katsayıları optimize etmek için tüm özellikler **StandardScaler** ile normalize edilmiştir.

•**Kategorik Değişken İşleme:** marka, model, transmission ve fuelType gibi metinsel veriler, cat.codes (Label Encoding) yöntemiyle modele uygun sayısal formatlara dönüştürülmüştür.

•**Çapraz Doğrulama (Cross-Validation):** Modelin farklı veri gruplarında ne kadar tutarlı olduğunu ölçmek için **3-Katlı** çapraz doğrulama uygulanmıştır.

Hyper-parametreler:

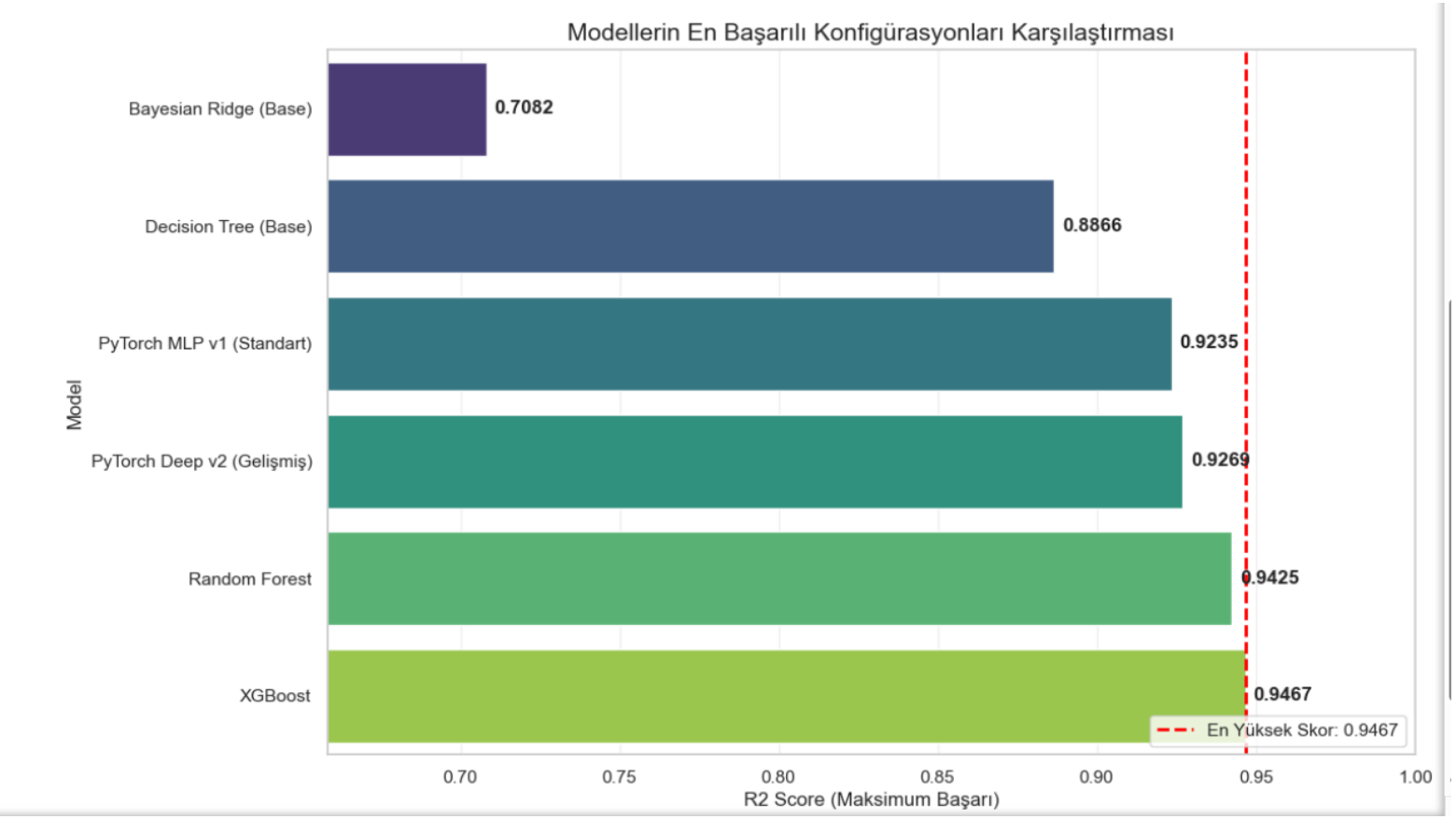
•Kullanılan Teknik: GridSearchCV

Optimize Edilen Parametreler:

- •n_estimators: 200 (Modelin oluşturacağı maksimum ağaç sayısı)
- •learning_rate: 0.1 (Her adımda hatalardan öğrenme hızı/adım boyutu)
- •max_depth: 6 (Karar ağaçlarının karmaşıklığını ve derinliğini belirleyen parametre)
- •**Skorlama:** Optimizasyon sürecinde en yüksek R2 skorunu veren parametre seçilmiştir.

Feature Set (Özellik Seti): Modelin eğitiminde kullanılan 16 temel özellik şunlardır : ['model', 'year', 'transmission', 'mileage', 'fuelType', 'tax', 'mpg', 'engineSize', 'marka',
'tax(£)', 'fuel type', 'engine size', 'mileage2', 'fuel type2', 'engine size2', 'reference']

Modellerin Karşılaştırılması



	Model	Yöntem	R2 Score	MAE	RMSE
11	XGBoost	GridSearch Optimized	0.9467	1155.57	1604.835888
10	Random Forest	GridSearch Optimized	0.9425	1165.65	1667.495607
9	PyTorch Deep v2 (Gelişmiş)	PyTorch DL	0.9269	1336.73	1890.863441
8	PyTorch MLP v1 (Standart)	PyTorch DL	0.9235	1360.34	1934.376112
0	Decision Tree (Base)	Vize Ödevi	0.8866	1825.33	-
1	Bayesian Ridge (Base)	Vize Ödevi	0.7082	3550.24	-

K-means ile Kümeleme

- **Kümeleme Stratejisi:** Fiyat sütunu hariç tutularak K-Means algoritması ile **4 sınıf** oluşturulmuştur.
- **Küme Sayısı (K):** 4
- **Görselleştirme:** PCA ile 2 boyuta indirgenen veride küme merkezleri işaretlenmiştir.
- **Segment İsimlendirmeleri:**
  - **1. Düşük** Yüksek kilometreli ve yaşlı araçlar.
  - **2. Ortalama:** Standart şehir araçları.
  - **3. Yüksek:** Ortalama yaş ve donanıma sahip araçlar.
  - **4. Lüks:** Yüksek motor hacimli ve lüks segment araçlar.

Kümeleme ile Araç Pazarı

SON DEĞERLENDİRME –ASLI

•**En Başarılı Algoritma:** XGBoost, karmaşık veri yapılarını öğrenmede en iyi
• performansı göstermiştir.

•**Kritik Faktörler:** Yapılan analizde araç fiyatını etkileyen en önemli 3 faktörün sırasıyla
• **transmission_Manual** (Şanzıman), **age** (Araç Yaşı) ve **engineSize** (Motor Hacmi) olduğu kanıtlanmıştır.

•**Genel Değerlendirme:** Projemiz, %90'ın üzerinde bir doğrulukla fiyat tahmini yapabilmekte
• ve pazar segmentlerini (4 ana grup) başarıyla birbirinden ayırabilmektedir.

Best Model 1 - Gradient Boosting -NURETTİN

•**Model Geliştirilirken Kullanılan Yaklaşımlar:** * Veri seti üzerinde **Log Dönüşümü** uygulanarak hedef değişkenin (fiyat) dağılımı normalize edilmiş ve modelin uç değerlere karşı daha dirençli olması sağlanmıştır.

•Özellik mühendisliği kapsamında **'car_age'** (araç yaşı) ve **'mileage_per_year'** (yıllık ortalama kilometre) gibi türetilmiş değişkenler eklenerek tahmin gücü artırılmıştır.
•Kategorik değişkenler için **One-Hot Encoding** yöntemi kullanılarak modelin anlayabileceği sayısal formata dönüştürülmüştür.

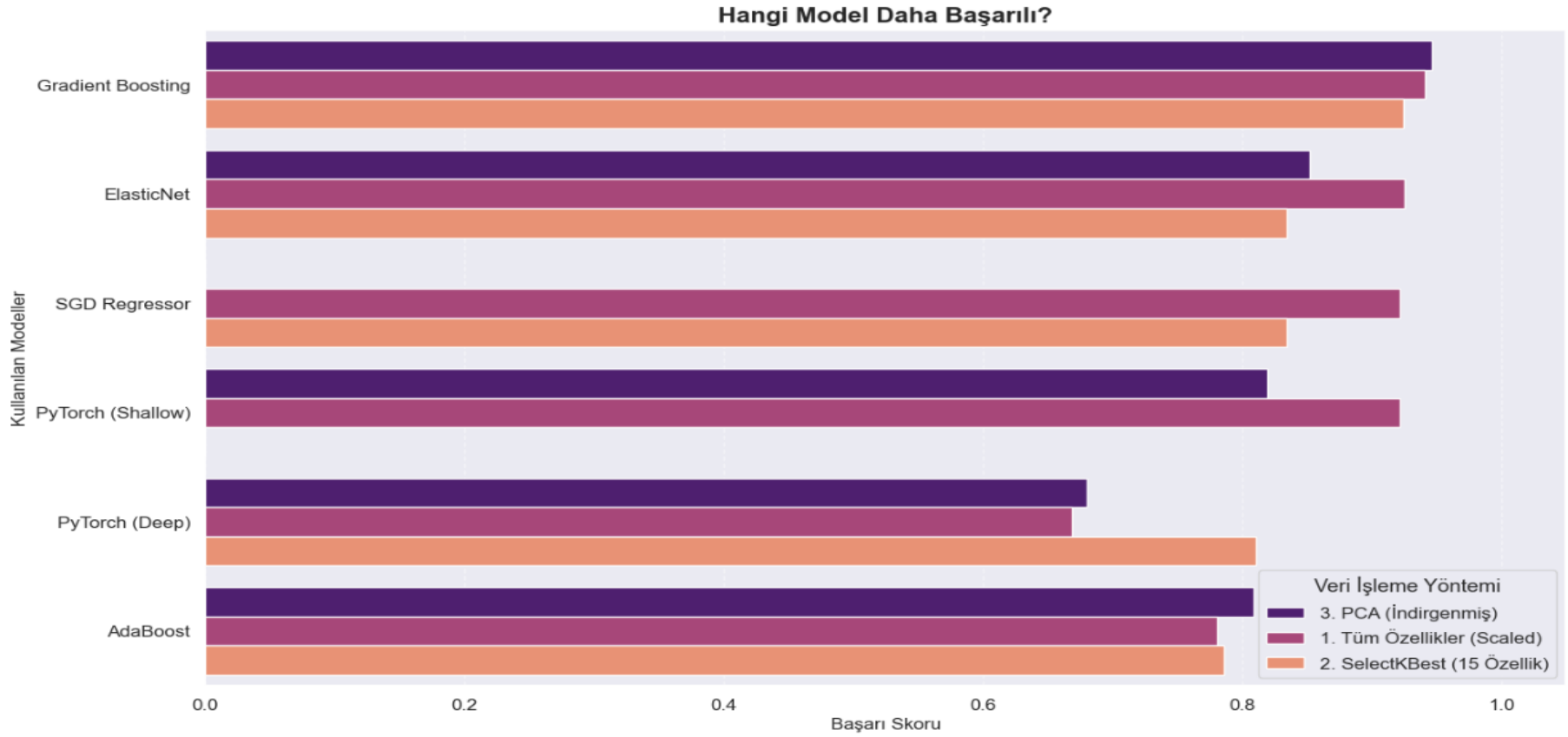
•**Hyper Parametreler:** * **Tuning Tekniği:** Modelin en iyi performansını yakalamak için **GridSearchCV** yöntemi kullanılmıştır.

•**Optimum Parametreler:** {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.8}.

•**Feature Set (Özellik Kümesi):** * ['car_age', 'mileage_per_year', 'mileage', 'tax', 'mpg', 'engineSize', 'marka_encoded', 'transmission_encoded', 'fuelType_encoded', 'model_encoded'].

Performans Metrikleri (Başarı Skorları): * **R2 Skoru:** 0.94821 Model, araç fiyatlarındaki değişimin %94.8'ini başarıyla açıklamaktadır.

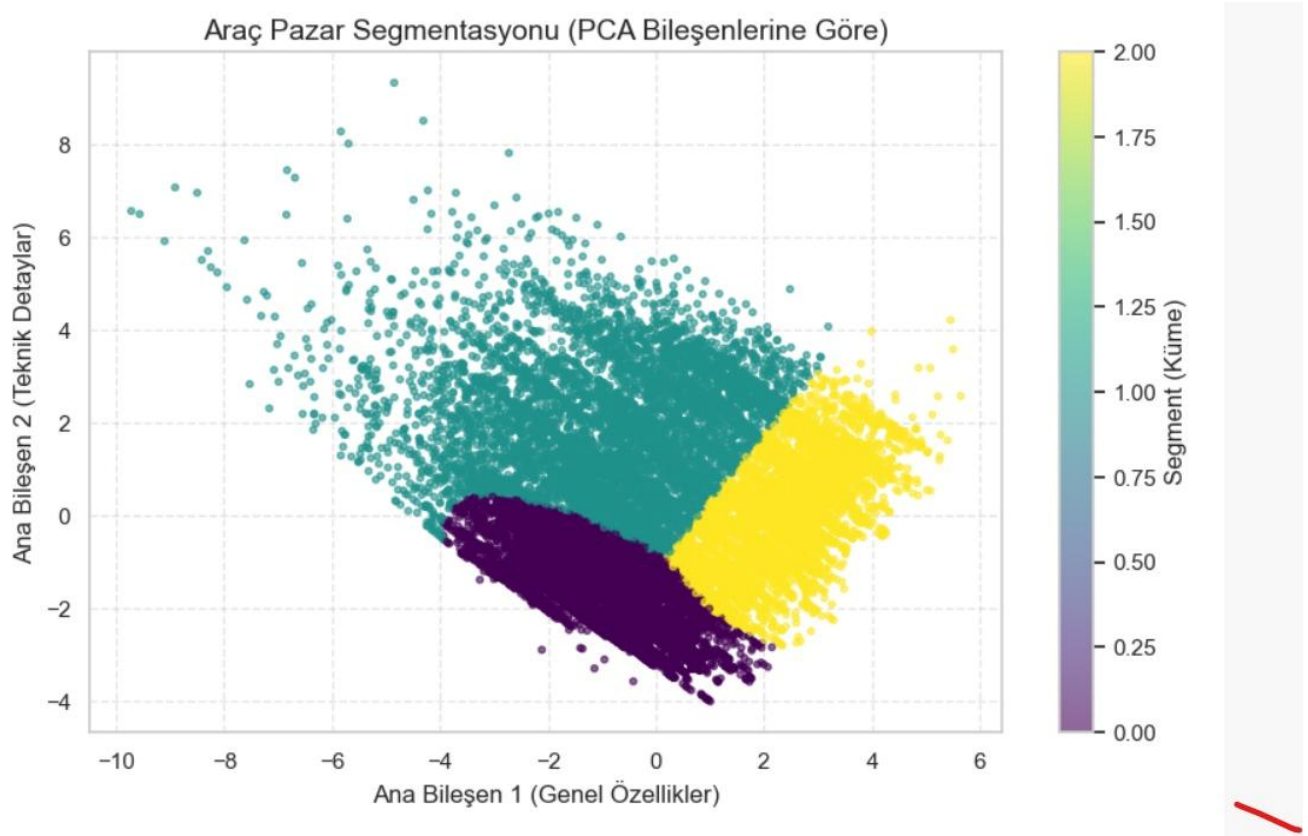
Modellerin Karşılaştırılması



Pazar Segmentasyonu (GMM)

- Algoritma, pazarı **3 ana segmente** ayırmıştır:
- **Segment 0 (Ekonomik / Şehir İçi Araçlar):** Genellikle düşük motor hacmine sahip, yakıt tasarrufu odaklı ve daha yüksek yaş/kilometre oranına sahip araçlar.
- **Segment 1 (Standart Aile Araçları):** Orta segment donanım özelliklerine sahip, dengeli performans sunan ve pazarın çoğunluğunu oluşturan araçlar.
- **Segment 2 (Performans & Lüks Araçlar):** Yüksek motor hacmi, yeni model yılı ve ileri teknik özelliklerle ayrışan üst segment araçlar.

GMM İLE PAZAR SEGMENTASYONU



SON DEĞERLENDİRME –NURETTİN

•**Algoritma Seçimi:** Gradient Boosting Regressor, karmaşık ve doğrusal olmayan veri
•yapılarını öğrenmede ElasticNet ve SGD gibi modellere göre çok daha üstün bir performans sergilemiştir.

•**Veri İşleme Etkisi:** Hedef değişkene (Price) uygulanan **Log Dönüşümü**, modelin sapmalarını ciddi oranda
• azaltmış ve tahmin kararlılığını artırmıştır.

•**Özellik Önemi:** Araç yaşı (car_age), kilometre (mileage) ve motor hacminin (engineSize) fiyat
•üzerindeki en belirleyici faktörler olduğu kantitatif olarak doğrulanmıştır.