T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BAYES TOPLAMSAL REGRESYON AĞAÇLARININ OTOMOBİL FİYATLARI ÜZERİNE BİR UYGULAMASI

Aleyna ALTUNTAŞ ARABOĞLU

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

Şubat, 2024

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BAYES TOPLAMSAL REGRESYON AĞAÇLARININ OTOMOBİL FİYATLARI ÜZERİNE BİR UYGULAMASI

Aleyna ALTUNTAŞ ARABOĞLU tarafından hazırlanan tez çalışması 09.02.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **YÜKSEK LİSANS** olarak kabul edilmiştir.

Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Fatma NOYAN TEKELİ, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Seher ARIKAN TEZERGİL, Üye
Marmara Üniversitesi _____

Danışmanım Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK sorumluluğunda tarafımca hazırlanan "Bayes Toplamsal Regresyon Ağaçlarının Otomobil Fiyatları Üzerine Bir Uygulaması" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Aleyna ALTUNTAŞ ARABOĞLU

İmza

Aileme

ve

eşime

TEŞEKKÜR

Tez çalışmam boyunca, beni yönlendiren, cesaretlendiren ve desteğini her zaman hissettiren tez danışmanım Prof. Dr. Gülhayat Gölbaşı Şimşek'e, özverili rehberliği, bilgisi ve önerileri için içtenlikle teşekkür ederim.

Ayrıca, bu çalışma süresince beni maddi ve manevi olarak destekleyen her zaman yanımda olan canım aileme ve sevgili eşime sonsuz teşekkür ederim.

Aleyna ALTUNTAŞ ARABOĞLU

İÇİNDEKİLER

TEŞEKKÜR iv
SİMGE LİSTESİ vii
KISALTMA LİSTESİ xii
ŞEKİL LİSTESİ xiv
TABLO LİSTESİ xvi
ÖZET xvii
ABSTRACT xix
1 GİRİŞ 1
1.1 Literatür Özeti 1
1.2 Tezin Amacı 3
1.3 Hipotez 3
2 BAYESYEN ÇIKARSAMANIN GELİŞİMİ 4
2.1 Tarihsel Gelişim Süreci 4
2.2 Bayes Teoremi 5
2.3 Bayes Tahmin Edicisi 6
2.4 Bayesci Yaklaşımın Temelleri ve İstatistiksel Kavramlara İlişkin Yorumları 8
2.4.1 Belirsizliğin Değerlendirilmesi 8
2.4.2 Olasılık 9
2.4.3 Parametre 9
2.4.4 Nokta Tahmini 9
2.4.5 Güven Aralığı 10
2.4.6 Hipotez Testi 11
2.5 Önsel Dağılım 12
2.5.1 Bilgi Vermeyen Önsel Dağılım 13
2.5.2 Eşlenik Önsel Dağılımlar 15
2.5.3 Sübjektif Önsel Dağılım 16
2.6 Sonsal Dağılım 16
2.7 Markov Zinciri Monte Carlo Yöntemi 17
2.7.1 Markov Zincirleri 17
2.7.2 Monte Carlo İntegrasyonu 19
2.7.3 Metropolis-Hastings Algoritması 19
2.7.4 Gibbs Örneklemesi 20

3 BAYES TOPLAMSAL REGRESYON AĞAÇLARI **22**
3.1 Bayes Regresyon Ağaçlarının Toplamı 22
3.2 BART Algoritması 26
4 SINIFLANDIRMA VE REGRESYON AĞAÇLARI **29**
5 UYGULAMA **31**
5.1 Kullanılan Programlar 31
5.2 Veri Seti 31
5.3 Bulgular 33
5.3.1 Tanımlayıcı İstatistikler 33
5.3.2 İlişkisel Bulgular 38
5.3.3 Regresyon Analiz Sonuçları 56
6 SONUÇ **76**
KAYNAKÇA **78**
TEZDEN ÜRETİLMİŞ YAYINLAR **83**

SİMGE LİSTESİ

ID_1	7 Koltuk
P(A)	A olayının olma olasılığı
TD_1	ABS-Anti Blokaj Fren Sistemi
DD_1	Adaptif far
DD_2	Alaşımlı jant
α	Alfa
$\subseteq$	Alt küme
H_1	Alternatif hipotez
ID_2	Anahtarsız çalıştırma
ID_3	Arka kol dayama
TD_2	ASR-Elektronik Patinaj Önleme Sistemi
P(B)	B olayının olma olasılığı
R^2	Belirlilik katsayısı
β	Beta
ID_4	Bluetooth
MM_1	BMW 116d
MM_2	BMW 318i
DD_3	Cam tavan
MM_3	Citroen C4
$\prod$	Çarpım
MM_4	Dacia Duster
MM_5	Dacia LODGY
MM_6	Dacia LOGAN
MM_7	Dacia SANDERO
ID_5	Deri koltuk
TD_3	Diferansiyel kilidi
ID_6	Dijital klima
TD_4	Diz hava yastığı
π^*	Durağan dağılım

Θ	Durum uzayı
TD_5	Elektrik takviyeli direksiyon
DD_4	Elektrikli aynalar
ID_7	Elektrikli sürücü koltuğu
TD_6	ESP-Elektronik Stabilizasyon Programı
$=$	Eşittir
$\neq$	Eşit değildir (Farklıdır)
F	F test istatistiği
DD_5	Far sensörü
DD_6	Far yıkama
$\setminus$	Fark
MM_8	Fiat EGEA
MM_9	Fiat LINEA
$J(\theta)$	Fisher'in bilgi matrisi
MM_{10}	Ford FIESTA
MM_{11}	Ford FOCUS
ID_8	Fonksiyonel direksiyon
$K(x, y)$	Geçiş çekirdeği
$P^{(n)}$	Geçiş matrisi
DD_7	Geri dönüş kamerası
N	Gözlem sayısı
Y	Gözlemler vektörü
ID_9	Hafızalı sürücü koltuğu
ϵ_i	Hata terimi
ID_{10}	Hız sabitleyici
TD_7	Hıza duyarlı direksiyon
MM_{12}	Honda CR-V
MM_{13}	Hyundai ACCENT
MM_{14}	Hyundai ELANTRA
MM_{15}	Hyundai GETZ
MM_{16}	Hyundai i10
MM_{17}	Hyundai i20

MM_{18}	Hyundai i30
MM_{19}	Hyundai ix35
MM_{20}	Hyundai TUCSON
ID_{11}	Isıtmalı arka koltuklar
DD_{8}	Isıtmalı aynalar
ID_{12}	Isıtımalı ön koltuklar
TD_{8}	Isofix
P	İkinci el otomobil fiyatları
$\int$	İntegral
μ_{kj}	j. regresyon ağacı için k. düğümün ortalama parametresi
DD_{9}	Katlanır aynalar
$\cap$	Kesişim
MM_{21}	Kia CEED
MM_{22}	Kia RİO
MM_{23}	Kia SPORTAGE
$\in$	Küme elemanı
Λ	Lamda
TD_{9}	Lastik basınç göstergesi
TD_{10}	LDW-Şeritten Ayrılma İkazı
DD_{10}	LED Gündüz sürüş farları
MM_{24}	Mercedes A180
TD_{11}	Mesafe takip sistemi
MM_{25}	Mini COOPER
ID_{13}	Navigasyon
MM_{26}	Nissan QASHQAI
F	Olasılık dağılımı
p	Olasılık yoğunluk fonksiyonu
MM_{27}	Opel ASTRA
MM_{28}	Opel CORSA
MM_{29}	Opel INSIGNIA
MM_{30}	Opel MOKKA
$\propto$	Oransallık

$\bar{\mu}$	Ortalama
ID_{14}	Otomatik kararan dikiz aynası
TD_{12}	Otomatik park sistemi
ID_{15}	Ön kol dayama
π	Önsel dağılımın olasılık yoğunluk fonksiyonu
θ	Parametre vektörü
DD_{11}	Park sensörü
χ^2	Pearson Ki-Kare istatistiği
MM_{31}	Peugeot 208
MM_{32}	Peugeot 308
MM_{33}	Peugeot 508
TD_{13}	Perde hava yastığı
R	R Studio Programı
g	Raslantı değişkeni fonksiyonu
MM_{34}	Renault CLIO
MM_{35}	Renault GRAND SCENIC
MM_{36}	Renault KADJAR
MM_{37}	Renault MEGANE
MM_{38}	Renault SYMBOL
$\bar{\sigma}$	Standart sapma
MM_{39}	Seat IBIZA
MM_{40}	Seat LEON
σ	Sigma
DD_{12}	Sis farı
MM_{41}	Skoda FABIA
MM_{42}	Skoda OCTAVIA
MM_{43}	Skoda RAPID
ID_{16}	Start/Stop
DD_{13}	Sunroof
q(y|x)	Teklif dağılımı
$\sum$	Toplam
MM_{44}	Toyota COROLLA

ID_{17}	USB Bağlantısı
$\overline{\sigma}^2$	Varyans
MM_{45}	Volkswagen BORA
MM_{46}	Volkswagen CC
MM_{47}	Volkswagen GOLF
MM_{48}	Volkswagen JETTA
MM_{49}	Volkswagen PASSAT
MM_{50}	Volkswagen POLO
MM_{51}	Volkswagen TIGUAN
MM_{52}	Volvo S60
$\cong$	Yaklaşık eşittir
DD_{15}	Yağmur sensörü
H_0	Yokluk hipotezi
TD_{14}	Yokuş kalkış desteği
ID_{18}	Yol bilgisayarı dış
$\hat{\theta}_B$	θ ‘nın Bayes tahmin edici
DD_{14}	Xenon far

KISALTMA LİSTESİ

A	Yaş
AH	Azami hız
AIC	Akaike Bilgi Kriteri (Akaike Information Criterion)
ANOVA	Analysis of Variance
BART	Bayes Toplamsal Regresyon Ağaçları (Bayesian Additive Regression Trees)
BK	Bagaj kapasitesi
C	Çekiş
CART	Sınıflama ve Regresyon Ağaçları (Classification and Regression Trees)
Değ.	Değişken
G	Genişlik
H	Hızlanma
kg	Kilogram
km	Kilometre
KT	Kasa tipi
MCMC	Markov zinciri Monte Carlo
MG	Motor gücü
MH	Metropolis-Hastings
MH	Motor hacmi
mm	Milimetre
Nm	Newton metre
Ort.	Ortalama
OYT	Ortalama yakıt tüketimi
P	Fiyat
Prob.	Probability
R	Renk
sn	Saniye
SS	Silindir sayısı
St. Hata	Standart Hata

St. Sapma	Standart Sapma
T	Tork
TL	Türk lirası
TU	Turbo
U	Uzunluk
V	Vites
VIF	Variance Inflationary Factor
W	Ağırlık
Y	Yükseklik
YT	Yakıt türü

ŞEKİL LİSTESİ

Şekil 3.1 Bir regresyon ağacı örneği $gjx, Tj, \mu j$ **22**

Şekil 3.2 Basit iki regresyon ağacı örneği kullanarak regresyon ağaçlarının toplamını gösterme..... **23**

Şekil 3.3 BART'ın m = 4 ile BART içindeki MCMC adımlarının iterasyon 3'üne göre sezgisi..... **23**

Şekil 4.1 Basit regresyon ağacı örneği **30**

Şekil 5.1 Satış fiyatı ile marka model değişkeni arasındaki ilişkiyi gösteren kutu grafiği..... **39**

Şekil 5.2 Satış fiyatı ile renk değişkeni arasındaki ilişkiyi gösteren kutu grafiği . **40**

Şekil 5.3 Satış fiyatı ile vites, silindir sayısı, yakıt türü ve turbo değişkenleri arasındaki ilişkiyi gösteren kutu grafiği..... **41**

Şekil 5.4 Satış fiyatı ile kasa tipi ve çekiş değişkenleri arasındaki ilişkiyi gösteren kutu grafiği..... **42**

Şekil 5.5 Satış fiyatı ile teknik donanımlar arasındaki ilişkiyi gösteren kutu grafiği **43**

Şekil 5.6 Satış fiyatı ile iç donanımlar arasındaki ilişkiyi gösteren kutu grafiği... **44**

Şekil 5.7 Satış fiyatı ile dış donanımlar arasındaki ilişkiyi gösteren kutu grafiği . **45**

Şekil 5.8 Satış fiyatı ile otomobil yaşı, kilometre, hızlanma, ortalama yakıt tüketimi, bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği..... **46**

Şekil 5.9 Satış fiyatı ile uzunluk, genişlik, yükseklik ve ağırlık değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği **47**

Şekil 5.10 Test veri setindeki satış fiyatı ile marka model değişkeni arasındaki ilişkiyi gösteren kutu grafiği **48**

Şekil 5.11 Test veri setindeki satış fiyatı ile renk değişkeni arasındaki ilişkiyi gösteren kutu grafiği **49**

Şekil 5.12 Test veri setindeki satış fiyatı ile vites, silindir sayısı, yakıt türü ve turbo değişkenleri arasındaki ilişkiyi gösteren kutu grafiği **50**

Şekil 5.13 Test veri setindeki satış fiyatı ile kasa tipi ve çekiş değişkeni arasındaki ilişkiyi gösteren kutu grafiği **51**

Şekil 5.14 Test veri setindeki satış fiyatı ile teknik donanımlar arasındaki ilişkiyi gösteren kutu grafiği **52**

Şekil 5.15 Test veri setindeki satış fiyatı ile iç donanımlar arasındaki ilişkiyi gösteren kutu grafiği **53**

Şekil 5.16 Test veri setindeki satış fiyatı ile dış donanımlar arasındaki ilişkiyi gösteren kutu grafiği **54**

Şekil 5.17 Test veri setindeki satış fiyatı ile otomobil yaşı, kilometre, hızlanma, ortalama yakıt tüketimi, bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği **55**

Şekil 5.18 Test veri setindeki satış fiyatı ile uzunluk, genişlik, yükseklik ve ağırlık değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği **56**

Şekil 5.19 Doğrusal ve yarı logaritmik modelin kalıntı grafiği **64**

Şekil 5.20 Veri setine ait regresyon ağacı grafiği **66**

Şekil 5.21 Veri setine ait rastgele orman grafiği **68**

Şekil 5.22 Rastgele orman modeline göre otomobil özelliklerinin önem sırası grafiği **70**

Şekil 5.23 Eğitim veri setine ait ağaç sayısı grafiği **71**

Şekil 5.24 Test veri setine ait ağaç sayısı grafiği **71**

Şekil 5.25 Eğitim veri setine ait yakınsama teşhis grafikleri **72**

Şekil 5.26 Eğitim veri setine ait artıkların dağılım grafiği **72**

Şekil 5.27 Eğitim veri setine ait tahmin edilen ve gerçek fiyat karşılaştırma grafiği **73**

Şekil 5.28 Test veri setine ait tahmin edilen ve gerçek fiyat karşılaştırma grafiği **73**

Şekil 5.29 BART modeline göre otomobil özelliklerinin önem sırası grafiği **75**

TABLO LİSTESİ

Tablo 2.1 Klasik yaklaşım ve bayesyen yaklaşıma göre işleyen süreçler **9**
Tablo 2.2 Farklı dağılımların Jeffrey önsel dağılımları **15**
Tablo 2.3 Olasılık fonksiyonlarının eşlenik ön dağılımları ve son dağılımları **16**
Tablo 5.1 Kategorik değişkenlere ait frekans değerleri **33**
Tablo 5.2 Sürekli değişkenlere ait tanımlayıcı istatistikler **37**
Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları **56**
Tablo 5.4 Doğrusal ve yarı logaritmik regresyon modeli performans ölçüm sonuçları **64**
Tablo 5.5 VIF test sonuçları **65**
Tablo 5.6 Regresyon ağacı modeli performans ölçüm sonuçları **67**
Tablo 5.7 Rastgele orman değişken önem değerleri **68**
Tablo 5.8 Rastgele orman modeli performans ölçüm sonuçları **70**
Tablo 5.9 BART modeli ağaç modeli performans ölçüm sonuçları **72**
Tablo 5.10 BART modeli önemli ilk 30 değişkenin değerleri **74**

ÖZET

Bayes Toplamsal Regresyon Ağaçlarının Otomobil Fiyatları Üzerine Bir Uygulaması

Aleyna ALTUNTAŞ ARABOĞLU

İstatistik Anabilim Dalı

İstatistik Programı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

Ağaç topluluğu algoritmaları, çok çeşitli regresyon ve sınıflandırma problemleri için yüksek tahmin doğruluğu üretme yeteneklerinden dolayı makine öğreniminin ön saflarında ortaya çıkmıştır. Rastgele ormanlar gibi sınıflandırmaya dayalı algoritmalar, verilere uyum sağlamak için algoritmik prosedürlere dayanır. Buna karşılık, Bayes Toplamsal Regresyon Ağaçları (Bayesian Additive Regression Trees (BART)) gibi daha yeni algoritmalar, uyumları oluşturmak için temel bir Bayesci olasılık modeline odaklanır. Bu tezin amacı BART modeli ile otomobillerin birtakım özelliklerinin fiyat üzerindeki etkisinin tahmin edilmesidir. Çalışmanın ilk bölümünde, Bayesci yaklaşım ayrıntılı olarak ele alınmaktadır. Önsel dağılım ve türleri tanıtılıp sonsal dağılımdan ve yöntemlerinden bahsedilmektedir. Bayesci yaklaşımlara vurgu yapılarak BART algoritmaları için metodolojik gelişmelere odaklanılmaktadır. Çalışmanın ikinci bölümünde, sınıflandırmaya dayalı regresyon ağaçlarından Sınıflama ve Regresyon Ağaçları (Classification and Regression Trees (CART)) modeline odaklanılmaktadır. Bu iki model hakkında bilgi verildikten sonra otomobil verisi üzerinde regresyon ağacı (regression tree), rastgele orman (random forest) ve BART algoritmaları

uygulanarak otomobillerin fiyatlarının belirlenmesinde hangi özelliklerin etkili olduğu tahmin edilmektedir. Bu üç algoritmanın performans ölçüm sonuçlarına göre ikinci el otomobil fiyatı üzerinde önemli etkiye sahip değişkenleri belirlemede en iyi performansı BART algoritması göstermektedir. Otomobillerin marka kilometre, yaş, genişlik, ağırlık, yükseklik, uzunluk, hızlanma, azami hız, ortalama yakıt tüketimi, bagaj kapasitesi, tork, vites, renk, kasa tipi, motor hacmi özelliklerinin yanı sıra iç-dış ve teknik donanımlarından park sensörü, sis farı, start-stop, usb bağlantısı, mesafe takip sistemi, Bluetooth, gündüz sürüş farları, hıza duyarlı direksiyon, adaptif far, elektrikli sürücü koltuğu ve yokuş kalkış desteği değişkenlerinin de fiyat üzerinde etkisi olduğu sonucuna varılmaktadır.

Anahtar Kelimeler: bayes toplamsal regresyon ağaçları, BART, CART, regresyon ağacı, rastgele orman

ABSTRACT

An Application of Bayesian Additive Regression Trees on Automobile Prices

Aleyna ALTUNTAŞ ARABOĞLU

Department of Statistics

Master of Science

Supervisor: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

Tree ensemble algorithms have emerged at the forefront of machine learning due to their ability to generate high prediction accuracy for a wide range of regression and classification problems. Classification-based algorithms such as Random Forests rely on algorithmic procedures to fit the data. In contrast, newer algorithms like Bayesian Additive Regression Trees (BART) focus on a foundational Bayesian probability model to build fits. The aim of this thesis is to predict the impact of various features of automobiles on their prices using the BART model. The first section of the study extensively discusses the Bayesian approach, introducing prior distributions and their types, and delving into posterior distributions and methods, with an emphasis on Bayesian approaches and methodological developments for BART algorithms. The second section of the study focuses on the Classification and Regression Trees (CART) model, a regression tree based on classification. After providing information about these two models, regression tree, random forest, and BART algorithms are applied to automobile data to predict which features are influential in determining the prices of cars. According to the performance measurement results of these three algorithms, the BART algorithm demonstrates the best performance in determining variables that have a significant impact on the

price of used cars. Variables such as brand, mileage, age, width, weight, height, length, acceleration, maximum speed, average fuel consumption, trunk capacity, torque, transmission, color, body type, and engine volume of cars, as well as interior-exterior and technical features such as parking sensors, fog lights, start-stop, USB connection, distance tracking system, Bluetooth, daytime running lights, speed-sensitive steering, adaptive headlights, electric driver's seat, and hill start assist, are found to have an impact on prices.

Keywords: bayesian additive regression trees, BART, CART, regression tree, random forest

1
GİRİŞ

1.1 Literatür Özeti

Ağaç topluluk yöntemleri, regresyon ve sınıflandırma problemlerinde tahmin yapmak için popüler tercihler haline gelmiştir. Rastgele ormanlar (random forests) Breiman (2001) [1] ve stokastik gradyan artırma Friedman (2002) [2] gibi algoritmalar, iki iyi kurulmuş ve geniş bir şekilde kullanılan prosedürdür. Son gelişmeler arasında, dinamik ağaçlar Taddy vd. (2011) [3] ve Bayes toplamsal regresyon ağaçları (BART) Chipman vd. (2010) [4] bulunmaktadır; bunlar, öncekilerden farklı olarak, temelde saf bir algoritma yerine bir altta yatan Bayesci olasılık modeline dayanmaktadır. BART, bir dizi simülasyon ve gerçek dünya uygulamasında önemli bir potansiyel göstermiştir, örneğin, dağ yollarında çığ tahmin etme Blattenberger ve Fowles (2014) [5], transkripsiyon faktörlerinin DNA ile etkileşimini tahmin etme Zhou ve Liu (2008) [6] ve film gişe gelirlerini tahmin etme Eliashberg (2010) [7] [8].

Başarısı nedeniyle, BART Chipman vd. (2010) [4], biomarker keşfi Hernandez vd. (2015) [9], iç mekan radon konsantrasyonlarının tahmini Kropat vd. (2015) [10], nedensel etkilerin tahmini Leonti vd. (2010) [11] ve Hill (2011) [12], genomik çalışmalar Liu vd. (2010) [13], hastane performans değerlendirmesi Liu vd. (2015) [14], kredi riski tahmini Zhang ve Hardle (2010) [15], kasırga olayları sırasında enerji kesintilerinin tahmini Nateghi vd. (2011) [16], ulaşımda seyahat sürelerinin tahmini Chipman vd. (2010) [17] ve tümör deneylerinde somatik tahmin Ding vd. (2012) [18] dahil olmak üzere birçok uygulama ile araştırma topluluğu arasında popülerlik kazanmıştır. BART ayrıca sağkalım sonuçlarına Bonato vd. (2011) [19] ve Sparapani vd. (2016) [20], çoknomial sonuçlara Kindo vd. (2016) [21] ve Agarwal vd. (2013) [22] ve yarı sürekli sonuçlara Linero vd. (2018) [23] genişletilmiştir. Nedensel çıkarım literatüründe, BART'ın kullanımını teşvik eden dikkate değer makaleler arasında Hill (2011) [12], Green ve Kern (2012) [24] bulunmaktadır. Ayrıca BART, eksik veri literatüründe de önemli bir rol oynamaktadır. Xu vd. (2016) [25], eksik kovaryatların ardışık imputasyonu için BART'ı kullanmanın bir yolunu önerdi, Kapelner ve Bleich (2015) [26] ise

kovaryatlardaki eksikliği bir kategori olarak ele almayı ve sonuçta Metropolis-Hasting (MH) adımında olasılık maksimize edilecek şekilde bölme kriterlerini kurmayı önerdi. Eksik çıktıların imputasyonu için, Tan vd. (2019) [27] BART'ın, hem ortalama hem de eğilim modellerinin yanlış belirtilebileceği durumlarda mevcut çift yönlü sağlam yöntemlerin direncini artırabileceğini inceledi. BART'ı kullanma veya genişletme üzerine diğer daha yeni girişimler arasında BART'ın quantile regresyonuna uygulanması Kindo vd. (2016) [21], BART'ın sayısal yanıtlara genişletilmesi Murray (2017) [28], BART'ın fonksiyonel verilere uygulanması Starling vd. (2018) [29], BART'ın tekrarlanan olaylara uygulanması Sparapani vd. (2018) [30], BART kullanılarak alt grupların belirlenmesi Sivaganesan vd. (2017) [31], Schnell vd. (2016) [32], Schnell vd. (2018) [33] ve BART'ın ölüme bağlı seçim yanlılığını hesaba katmak için eksik ana strataları eksiksiz hale getirmek için güçlü bir model olarak kullanılması Tan vd. (2018) [34] bulunmaktadır [35].

Literatürde ikinci el otomobil fiyatının ve fiyatı etkileyen değişkenlerin belirlenmesine ilişkin birçok çalışma bulunmaktadır. Söz konusu çalışmaların bazıları aşağıda kısaca özetlenmiştir.

Asilkan ve Irmak (2009) [36] yapay sinir ağları ile 2008-2009 yılları arasındaki fiyatların tahminini 2005-2007 yılları arasındaki verileri kullanarak yapmışlardır. Aynı zamanda bu verileri kullanarak zaman serileri analiz yöntemleri ile de tahminler yaptılar ve bu sonuçları karşılaştırdılar.

Daştan (2016) [37] Türkiye'de ikinci el otomobil fiyatlarını etkileyen faktörleri hedonik fiyat modellemesi ile belirledi. Çalışmasında elde ettiği bulgulara bakıldığında, marka ve model, yaş, çekiş, kilometre, vites, yakıt türü, tork, genişlik, yakıt tankı hacmi, ABS, hardtop, panoramik cam tavan, arka cam buz çözücü, hidrolik direksiyon, start/stop, renk, net ağırlık, yükseklik, boyalı parça sayısı, aux, şerit değiştirme yardımcısı, sürücü hava yastığı, tavan hava yastığı, bi Xenon far, elektrikli ön camlar, otomatik kararan dikiz aynası, ısıtmalı direksiyon, geri görüş kamerası, hızlanma, uzunluk, değişen parça sayısı, ABC, immobilizer, xenon far, sis farı, far yıkama, sunroof, soğutmalı torpido ve ön görüş kamerası değişkenlerinin otomobil fiyatları üzerinde etkisi olduğu görülmektedir.

Yayar ve Yılmaz (2018) [38] çalışmalarında ikinci el otomobillerin özelliklerinin talep fiyatı üzerindeki etkisini hedonik regresyon modeli oluşturarak tespit ettiler. Çalışma sonucuna göre ikinci el otomobil talep fiyatını etkileyen fiyat artırıcı özellikleri aracın Çorum kent merkezinde yer alması, markasının Volkswagen, Toyota, Mercedes olması, dizel yakıt kullanması, lastik arıza göstergesine sahip olması, motor silindir hacminin büyük olması, start/stop özelliğinin olması, arka kameraya sahip olması; ikinci el otomobil talep fiyatını düşürücü özelliklerini ise; otomobil modelinin eski olması, renginin gri olması ve servis imkânına sahip olması olarak tespit ettiler.

1.2 Tezin Amacı

Bu tezde doğrusal regresyon, yarı logaritmik regresyon ve regresyon ağaçları gözden geçirilip BART metodolojisi tanıtılacak, CART ve BART algoritmaları kullanılarak fiyatı en iyi tahminleyen model oluşturularak otomobillerin birtakım özelliklerinin fiyat üzerindeki etkisi tahmin edilecektir.

1.3 Hipotez

Otomobillerin özellikleri ile otomobil fiyatlarını modellemek ve satış fiyatlarını tahmin etmek için BART gibi ağaç temelli makine öğrenmesi yöntemleri kullanılabilir.

2
BAYESYEN ÇIKARSAMANIN GELİŞİMİ

2.1 Tarihsel Gelişim Süreci

İstatistiğin gelişim sürecinde temelde Klasik (Frekansçı, Berkeley istatistiği) ve Bayesci yaklaşımları belirginleşmiştir. Başlangıç varsayımlarının yorumlanması ve pek çok konu ile kavramın ele alınışı, alternatif yaklaşımların birbirine tercih edildiği bir bağlamda incelenmiştir. Ancak zaman içinde, istatistik alanında klasik yaklaşımın hakimiyeti artmıştır [39].

İstatistik tarihini incelediğimizde, istatistiksel çıkarımın 18. yüzyıl ile 20. yüzyıl tarihleri arasında Bayesci yaklaşımın önemli bir etki yarattığı gözlemlenmektedir. Bu dönemde, Laplace, Thomas Bayes'in çalışmalarını geliştirerek olasılık teorisine büyük katkılarda bulunmuştur. Teorik istatistikte daha detaylı ve başarılı çalışmaların ortaya çıkmasında, Bayes ve Laplace'ın çıkarım yöntemlerini ve kavramlarını kullanarak çalışan Fisher, Neyman, Pearson, Wald gibi bilim insanları etkili olmuştur. Klasik yaklaşım ve Bayesyen yaklaşım arasında kavramlar ve çıkarım yöntemleri açısından farklılıklar bulunmaktadır. Ancak, Pearson, Neyman ve Fisher gibi isimler, Bayesci yaklaşım ile "ters olasılık" kavramına eleştirel bir tutum sergileyerek, 1920 ile 1950 yılları arasında Bayesyen yaklaşımın gelişmesini olumsuz yönde etkilemişlerdir [39].

Klasik yaklaşımın etkisi altında popularitesini kaybetmeye başlayan Bayesyen yaklaşıma 1900'lü yıllarda yapılan çalışmalarla tekrar ilgi duyulmaya başlanmıştır. Bu çalışmalar arasında Ramsey tarafından "Gerçeklik ve Olasılık" adlı çalışmanın yapılması, Bayesyen hipotez testini geliştiren H. Jeffreys tarafından "Ters olasılık"ın neden olduğu mantıksal eksikliklerin giderilmesi üzerine bir çalışmanın yapılması, objektif yaklaşımlı Bayesyen analizi geliştirmesi, I. J. Good, L. J. Savage ve B.de Finetti'nin Bayesyen analizi subjektif yaklaşımlı olarak geliştirmesi, D.V.Lindley ve R.Schlaifer'in Bayesyen yaklaşım doğrultusunda, işletme ve endüstriyel sorunlara yeni fikirler geliştirmesi yer almaktadır. Bazı matematiksel yapıların çözümünün sağlanamaması teorik altyapıyı daha ayrıntılı oluşturmaya engel olmasıyla birlikte model seçimindeki bazı son yapıların integral hesaplamalarının sonucuna ulaşılması da imkansız hale gelmiştir. Ancak

Metropolis tarafından geliştirilen çalışmalar Markov Zinciri Monte Carlo (MCMC) tekniğine temel oluşturmuştur. Hastings, MCMC yönteminin istatistik alanındaki çalışmalarını geliştirmesiyle, P.H. Peskun, ve S. Geman'ın yapmış oldukları çalışmaları ile Gelfand ve Smith'in ise yardımlarıyla bu engel kaldırılmıştır. Böylece Bayes teoremi bilim insanları tarafından geliştirilmiş ya da tekrar yorumlanmış ve istatistik alanında devrimsel bir gelişme niteliği taşımıştır [40].

2.2 Bayes Teoremi

Bayes Teoremi, koşullu olasılığa dayanır ve bir olayın olasılığını tanımlamak için olayla ilgili olabilecek koşulların ön bilgisini kullanır. A ve B herhangi iki olay ve A ve B olaylarının olma olasılıkları 0'dan farklı olmak üzere, P(A|B) = P(A∩B)/P(B), B olayının bilindiği durumda A olayının koşullu olasılığını ifade eder. Bu eşitlik, P(A∩B) = P(B) P(A|B) olarak ifade edilebilir. P(A∩B) olasılığı, A ve B olaylarının birlikte gerçekleşme olasılığıdır. Bu ilişki "Bayes Teoremi" olarak bilinir ve P(B|A) = [P(B) P(A|B)] P(A) şeklinde ifade edilir [41].

Eğer B olayı n tane ayrık olayın bir kümesinden oluşuyorsa (B_1, B_2, …, B_n) A ve B olayları arasındaki ilişki A=(B_1∩A)∪(B_2∩A)∪…∪(B_n∩A) şeklinde ifade edilebilir. P(B_i∩A) ayrık olaylar ise P(A) = $\sum_{i=1}^{n} P(Bi \cap A)$ geçerlidir. Yazılan koşullu olasılığa göre P(B_i∩A) = P(B_i)P(A|B_i)'dir. Bu durumda A olayının olasılığı P(A)=$\sum_{i=1}^{n} P(Bi)P(A|Bi)$ şeklinde ifade edilebilir. Bu durumda Bayes teoremine göre A olayı biliniyorken B_i olayının gerçekleşme olasılığı Eşitlik 2.1'de verildiği gibi gösterilir [41].

$$P(B_i \setminus A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i)P(A \setminus B_i)}{P(A)} = \frac{P(B_i)P(A \setminus B_i)}{\sum_{i=1}^{n} P(B_i)P(A \setminus B_i)} \quad (2.1)$$

Burada, P(B_i) olasılıklarına önsel olasılık denir. Bu olasılıklar yapacağımız denemenin verebileceği herhangi bir bilginin elde edilmesinden önceki olasılıklardır. Olabilirlik fonksiyonu ise P(A|B_i)'dir. P(B_i|A) koşullu olasılığı ise bu iki bilgi birleştirilerek oluşturulmuştur ve sonsal olasılığı ifade etmektedir. Sonsal olasılıklar, denemenin verebileceği herhangi bir bilginin elde edilmesinden sonraki olasılıklardır [42]. Bayesyen çıkarsama kapsamında P(B|A) olasılığındaki B, analizdeki parametrenin belli bir değerine karşılık gelmektedir. Genel olarak Bayes teoremi Eşitlik 2.2'deki gibi gösterilebilir [41].

$$P(A \setminus B) = \frac{P(B \setminus A)P(A)}{P(B)} \propto P(B \setminus A)P(A) \quad (2.2)$$

Teorem, veriden elde edilen bilgiler kullanılarak ön olasılıkların güncellendiğini ifade eder. Bu teoremi uygulayabilmek için $P(B_i)$ değerleri gibi ön olasılık değerlerini bilmek gereklidir [39].

Parametrenin, olasılık dağılımı olan bir rastlantı değişkeni gibi düşünüldüğü Bayesyen yaklaşımda, çıkarsama için, Bayes teoremine göre;

Eşitlik 2.3'te p olasılık yoğunluk fonksiyonunu (o.y.f.) temsil ederken θ parametre vektörünü, y ise gözlemler vektörünü, $p(\ y,\theta\ )$ birleşik o.y.f.'yi ifade eder. Buna göre:

$$p(y \setminus \theta).p(\theta) = p(y,\theta) = p(\theta \setminus y).p(y) \quad (2.3)$$

Bu eşitliği kullanarak elde edilen:

$$p(\theta \setminus y) = \frac{p(\theta).p(y \setminus \theta)}{p(y)} \quad (2.4)$$

Paydadaki $p(y)$, belirli bir sabit (örneğin c^{-1} gibi) olup $p(y)$'nin açılımı şu şekilde ifade edilebilir.

$p(y) = \int p(y \setminus \theta)\,.p(\theta)d\theta$ parametrenin dağılımı sürekli ise,

$p(y) = \sum p(y \setminus \theta).p(\theta)$ parametrenin dağılımı kesikli ise,

Söz konusu c sabiti, son dağılımın integralini ya da toplamını 1'e eşitleyen bir normalleştirme sabitidir. O.y.f, sabit terim olmadan yazıldığında;

$$p(\theta \setminus y) \propto p(\theta).p(y \setminus \theta)$$
$$son\ oyf\ \propto \text{ön}\ oyf\ \times benzerlik\ fonksiyonu \quad (2.5)$$

Yukarıda Eşitlik 2.5'de yer alan "$\propto$ " işareti, oransallığı ifade eder.

2.3 Bayes Tahmin Edicisi

Klasik yaklaşıma göre olasılık dağılımları $F = \{f(.;\theta{:}\,\theta\epsilon\Theta\}$ ailesindeki olasılık yoğunluk fonksiyonlarını indisleyen θ parametresi, parametre kümesinde bilinmeyen bir değer olarak kabul edilmektedir. X_1, X_2, …, X_n örneklemindeki X_i'ler bağımsız ve aynı $f(.;\theta)$ olasılık yoğunluk fonksiyonuna sahip rasgele değişkenler olarak ele alınır. Bayes yaklaşımında ise olasılık dağılımına sahip bir rasgele değişken olan θ'nın dağılımı önsel (prior) dağılım olarak adlandırılır.

Genellikle bu öznel (subjective) bir dağılım olmak üzere X_1, X_2, …, X_n örneklemi, θ rasgele değişkeni ile koşullandırılmış dağılımdan alınmış gibi düşünülmektedir. Yani $i = 1,2,..n$ için X_i'ler bağımsız ve her biri $f(.;\theta)$ olasılık (yoğunluk) fonksiyonuna sahip rasgele değişkenler olarak ele alınmaktadır. Bu durumda, örneklemin olasılık (yoğunluk) fonksiyonu Eşitlik 2.6'da verildiği gibi gösterilir [41]:

$$f(x_1, x_2, \dots, x_n) = f(x_1\backslash\theta) f(x_2\backslash\theta) \dots f(x_n\backslash\theta) = \prod_{i=1}^{n} f(x_i\backslash\theta) \tag{2.6}$$

ve θ 'nın önsel dağılımının olasılık (yoğunluk) fonksiyonu π olmak üzere:

$$f_{x_1,x_2,\dots,x_n,\theta}(x_1, x_2, \dots, x_n, \theta) = f(x_1\backslash\theta) f(x_2\backslash\theta) \dots f(x_n\backslash\theta)\pi(\theta)$$
$$= \prod_{i=1}^{n} f(x_i\backslash\theta)\, \pi(\theta) \tag{2.7}$$

Eşitlik 2.7 elde edilir. $X = (X_1, X_2, \dots, X_n)'$ ve $x = (x_1, x_2, \dots, x_n)'$ vektör gösterimleri ile aşağıdaki eşitlikler yazılabilir.

$$f(x\backslash\theta) = f(x_1\backslash\theta) f(x_2\backslash\theta) \dots f(x_n\backslash\theta) = \prod_{i=1}^{n} f(x_i\backslash\theta)\, \pi \tag{2.8}$$

$$f_{X,\theta}(x\backslash\theta) = f(x, \theta)\pi(\theta) \tag{2.9}$$

Eşitlik 2.10'da X'in marjinal dağılımının olasılık fonksiyonu verilmiştir.

$$f_X(x, \theta) = \int_{\Theta} f(x\backslash\theta)\pi(\theta)\, d\theta \tag{2.10}$$

Bu eşitliğe göre örneklem değerleri verildiğinde θ'nın sonsal (posterior) dağılımının olasılık fonksiyonu Eşitlik 2.11'de verilmiştir.

$$\pi(\theta\backslash x) = \frac{f_{X,\theta}(x\backslash\theta)}{f_X(x)} = \frac{f(x,\theta)\pi(\theta)}{f_X(x)} \tag{2.11}$$

Önsel dağılımın gözlemlenen verilerle güncellenmesi yoluyla sonsal dağılım elde edilebilir. Bu güncellenmiş dağılımın beklenen değeri, verileri oluşturan dağılımda yer alan θ için tahmin edici olarak kullanılabilir [41].

$$\hat{\theta}_B = \int_{\Theta} \theta\pi(\theta\backslash x)\, d\theta \tag{2.12}$$

Eşitlik 2.12'de belirtilen $\hat{\theta}_B$ istatistiğine θ 'nın Bayes tahmin edicisi denir.

Bayes tahmin edicileri, bir istatistik olan T(X) ile belirli bir $k(\theta, T)$ kayıp fonksiyonuna dayalı olarak tanımlanan risk fonksiyonu Eşitlik 2.13'de verilmiştir.

$$R(\theta, T) = \int k(\theta, T(X)) f(x\backslash\theta) dx \tag{2.13}$$

Eşitlik 2.14 ile gösterilen integralin beklenen değerini minimuma eşitleyen tahmin edicidir.

$$R(T) = \int_{\Theta} R(\theta, T)\pi(\theta)\, d\theta \tag{2.14}$$

Kısaca $T = \{T: R(T) < \infty\}$ olarak tanımlanan bir kümede $T^* \in T$ için Eşitlik 2.15'e göre:

$$R(T^*) \leq R(T) \quad , \; T \in T \tag{2.15}$$

T^* tahmin edicisi Bayes tahmin edicisi olarak adlandırılır. Buna göre T^* tahmin edicisi, Eşitlik 2.16'da yer alan kayıp fonksiyonuna dayalıdır [41].

$$k(\theta, T) = (T - \theta)^2 \tag{2.16}$$

2.4 Bayesci Yaklaşımın Temelleri ve İstatistiksel Kavramlara İlişkin Yorumları

İstatistik, bilimdeki tümdengelim ve tümevarım yaklaşımlarının ötesinde, belirli ilkelerin değerlendirilmesine dayalı olarak farklı yorumlar ve uygulamalar geliştirmiştir. Bu bağlamda, istatistikteki düşünce tarzları zaman içinde netlik kazanarak, klasik ve Bayesyen yaklaşım olmak üzere iki ayrı yaklaşımı ortaya çıkarmıştır. Tümdengelim mantığına dayanan Klasik yaklaşım nedensellik ilkesinin deterministik yorumunu benimserken, tümevarım mantığına dayanan Bayesyen yaklaşım ise olasılık yorumunu benimsemektedir [43].

İstatistikte bir olasılığın nesnelliği incelenen süreçle ilgili olarak önceden edinilmiş hiçbir bilginin ihmal edilmemesi ilkesine dayanır. Bayesci yaklaşım süreçle ilgili önceden edinilmiş bilgilerin mevcut veri setiyle birleştirilerek sonsal dağılımın olasılığını hesaplamaya olanak tanır. Bu özellik, Bayes yaklaşımını klasik yaklaşıma tercih etmek için yeterli sebepler arasındadır [41].

2.4.1 Belirsizliğin Değerlendirilmesi

Bayesci yaklaşımın temel fikri teorileri, önermeleri veya nedensellik ilişkilerini, her kapsamdaki belirsizlikleri olasılıklarla ifade etmektir. Klasik yaklaşım ise varsayımlara dayanarak belirsizliği, sıklıklarına göre değerlendirip deterministik bir yaklaşımla kabul etme veya reddetme eğilimindedir. Bu bağlamda Bayesci yaklaşımı benimseyenler, belirsizliği ifade eden bir teoriyi sadece kabul etmenin

veya reddetmenin bilgi oluşturma sürecine katkıda bulunmayacağını ve aksine önemli yanılsamalara neden olabileceğini savunabilirler [39].

2.4.2 Olasılık

Bayesyen yaklaşım olasılığın ortaya konan birbirinden farklı tanımlarından Subjektif Teori'yi benimsemektedir. Bu teoriye göre Bayesci istatistikte bir olayın olasılığı, o olaya ait önceden bilinen inanç düzeyi ile gerçekleşen deney sonuçlarının (verilerin) birleştirilmiş halidir. Bu birleştirme işlemi, Bayes teoremi ve koşullu olasılık prensiplerine dayanır. Bayesyen istatistikte amaç en kesin bilgiye ulaşarak "1" olasılığına yaklaşmaktır [44]. Öte yandan klasik yaklaşımda tümdengelim yöntemi kullanıldığından amaç yanlışlama yöntemi ile "0" olasılığını elde etmektir [39].

Tablo 2.1 Klasik yaklaşım ve bayesyen yaklaşıma göre işleyen süreçler

Klasik Yaklaşım	**Bayesyen Yaklaşım**
Varsayımlarla	Varsayımlarsız
Deneme	Deneme
Yanlışlama	Doğrulama
$\mathrm{p(t,g)} = 0$	$\frac{1}{2} < \mathrm{p(t,g)} < 1$

"t" bir teori ve "g" onunla ilişkili elde edilen gözlemler olmak üzere farklı olasılık yaklaşımlarına göre süreç Tablo 2.1'de ifade edildiği gibidir [43].

2.4.3 Parametre

Bayesyen yaklaşım, parametreyi bir olasılık dağılımı olarak ele alır ve parametre tahmini için bir ön olasılık dağılımı belirleyerek bu dağılıma dayalı olarak parametre tahmincisinin son olasılık dağılımını elde eder. Öte yandan, klasik yaklaşım parametreyi bir bilinmeyen sabit olarak görüp, parametre tahminini sadece eldeki verilere dayanarak gerçekleştirir. Klasik yaklaşımda parametre, tekrarlanan gerçek denemelerin sonucu olmadığından dolayı olasılık dağılımıyla ilişkilendirilmez [39].

2.4.4 Nokta Tahmini

Ana kitleyi doğrudan gözlemlemenin mümkün olmadığı durumlarda parametre değerlerine ulaşmak da zordur. Bu nedenle ana kitle parametreleri hakkında tahminler yapmak için genellikle ana kitleden alınan örnekler kullanılır. Bir parametre tahmininin temelinde genellikle nokta tahmin yer almaktadır. Parametre

değerlerini iyi şekilde yansıtması için nokta tahminleri yansızlık, yeterlilik, tutarlılık ve etkinlik gibi özelliklere sahip olmalıdır [43].

2.4.4.1 Yansızlık

Eğer $X_1, X_2, \ldots, X_n$, parametresi θ olan kitleden bir örneklemi temsil ediyorsa ve T de θ'nın herhangi bir tahmin edicisi ise her $\theta \epsilon \Theta$ için $E_\Theta(T) = \theta$ oluyorsa, tahmin edici (T) θ parametresi için yansız bir tahmin edicidir. Bu ifadede yan, $E_\Theta(T) = -\theta$ şeklinde ifade edilir [41].

2.4.4.2 Yeterlilik

Bilinmeyen θ parametresi için yeterli bir tahmin edici, örneklem içindeki bilgilerin tamamını içerir, dolayısıyla herhangi bir bilgi kaybı olmadan özetleyen bir tahmin edicidir. Örneğin, $X_1, X_2, \ldots, X_n$, olasılık yoğunluk fonksiyonu $f(x, \theta)$ olan bir kitleden örneklem alalım. $T(X_1, X_2, \ldots, X_n) = T(\underline{X})$ tahmin edicisi θ için yeterli ise, θ hakkında $X_1, X_2, \ldots, X_n$ örneklemine dayalı olarak hesaplanan bir sonuç yalnızca $T(\underline{X})$ tahmin edicisinin değerine bağlıdır. Bu durumda $T(\underline{X})$ bilindiğinde $\underline{X}'in$ koşullu dağılımı θ parametresine bağlı değilse, $T(\underline{X})$ için θ yeterli bir tahmin edici denir [41].

2.4.4.3 Tutarlılık

$X_1, X_2, \ldots, X_n$ olasılık yoğunluk fonksiyonu $f(x, \theta)$ olan kitleden bir örneklem üzerinde θ'nın tahmin edicisi $T_n = T_n(\underline{X})$ için n sonsuza yaklaştıkça T_n'in θ'yi güvenilir bir şekilde tahmin etme özelliği yani $n \to \infty$ iken $T_n \xrightarrow{p} \theta$ durumu T_n tahmin edicilerinin dizisinin θ için tutarlı olduğu anlamına gelir [41].

2.4.4.4 Etkinlik

Rastgele bir θ parametresi için T_n ve $T_n *$ tahmin ediciler olarak kabul edildiğinde, bütün θ'lar için $Var_\theta(T_n) \leq Var_\theta(T_n *)$ koşulu sağlanıyorsa T_n tahmin edicisi $T_n *$ tahmin edicisine göre daha etkindir olarak nitelendirilir [41].

2.4.5 Güven Aralığı

Klasik yaklaşımda güven aralığı parametrelerin belirli değer aralığında yer almasını ifade ederken, Bayesyen yaklaşımda ise bu durum belirli değerler aralığında yer alma olasılığını ifade eder. Klasik yaklaşımda "güven aralığı" kavramı kullanılırken

Bayesyen yaklaşımda bu kavram "güvenilir aralık", "Bayesyen aralık" ya da "En Yüksek Son Yoğunluk Aralığı" olarak ifade edilmektedir [39].

Bayes yaklaşımında hesaplanan son dağılım $p(\theta|y)'e$ göre, veri y olduğunda parametre θ'nın, parametre uzayının belirli $\overline{R}$ altbölgesinde bulunması olasılığı;

$$P\{\theta \in \overline{R}\ |y\} = \int_{\overline{R}}\ P(\theta|y)\ d\theta\text{'dır.} \tag{2.17}$$

Bu ifade de, sınırlandırılacak olasılık miktarı 0,95 olması halinde;

$$P\{\theta \in \overline{R}\ |y\} = \int_{\overline{R}}\ P(\theta|y)\ d\theta = (1-\alpha) = 0{,}95 \tag{2.18}$$

aranılan bölgenin sınırlarına veya aralığın değerlerine ulaşmayı ifade eder [12]. Bölge içindeki her bir noktanın olasılık yoğunluğu, dışındaki noktaların her bir noktadan daha büyük olduğu için bu bölge "En Yüksek Son Yoğunluk" bölgesi olarak bilinir. Ayrıca bu bölge, veri gözlemlendiğinde içerdiği olasılık miktarı açısından parametre uzayında olası minimum hacimli bölgedir; aynı zamanda olası en dar güven aralığıdır [43].

2.4.6 Hipotez Testi

Bayesyen yaklaşım olasılık kavramına daha geniş bir açıdan bakmasından dolayı hipotezlere olasılık ataması yapar. Klasik Aristo yaklaşımında, hipotezlere atanan olasılık değeri 0 veya 1 değil, 0-1 aralığında olması gerekmektedir. Sürekli bir dağılımda, parametrenin tam olarak sıfıra eşit olma olasılığı 0 olduğundan, Bayes yaklaşımının çift taraflı hipotez testinde güçsüz olduğu durumlar ortaya çıkabilir. Bu durumda çözüm, regresyon katsayısı için 0'a yakın bir değer alarak daha esnek bir yaklaşım benimsemek olabilir [39].

Bayesyen yaklaşımda hipotezlerin ön olasılıkları belirlenerek son olasılıkları elde edilir. En iyi hipotez, son olasılığı en fazla olan hipotezdir [39]. H_0 ve H_1 hipotezleri, y gözlemi, θ ve ϕ parametreleri, $\theta_0$ve ϕ_0 bu parametrelerin belirli değerlerini gösterirken şu denklem geçerlidir;

$$\frac{P(H_0|y)}{P(H_1|y)} = \frac{P(H_0)}{P(H_1)} * \frac{P(y|\theta=\theta_0)}{P(y|\theta=\phi_0)} \tag{2.19}$$

Bu ifade şu şekilde açıklanabilir;

$son\ bahis\ oranı = ön\ bahis\ oranı\ \times bayes\ faktör$ü
$(posterior\ odds = prior\ odds\ \times bayes\ factor)$

Eşitlik (2.19)'da $P(H_0) = P(H_1)$ ise veya bilgi vermeyen ön dağılım için $P(H_0) = P(H_1) = \frac{1}{2}$ ise son bahis oranı bayes faktörüne eşit çıkar. Bu durumda bayes faktör, benzerlik oranı ile aynı olur [39].

Klasik yaklaşımda hipotez testi süreci Fisher'in çalışmalarıyla başlamış ve gelişmiştir. Fisher'a göre araştırmacının reddetmesi gereken hipotez H_0 hipotezidir. Test istatistiği, eldeki örnek bilgileri kullanılarak hesaplanır. Belirlenen anlamlılık düzeyine dayanarak, kurulan H_0 hipotezi ya kabul edilir ya da reddedilir. Ancak, daha sonraki dönemlerde Neyman ile Pearson, H_0 hipotezinin, bir alternatif hipotez (H_a) ile karşılaştırılması gerektiğini ve hipotezlerin kabul ya da reddedileceğini düşünmüşlerdir. Bu karar sürecinde Tip I hata (H_0 hipotezinin gerçekte doğru olduğu durumda, bu hipotezi yanlışlıkla reddetme durumu) ve Tip II hata (H_0 hipotezinin gerçekte yanlış olduğu durumda, bu hipotezi yanlışlıkla kabul etme durumu) olarak iki tip hata ortaya çıkar. Karar verme aşamasında testin gücünü maksimumda tutmak önemlidir. Fakat aynı anda H_0'ı reddederken, H_0'ın gerçekte doğru olma olasılığını da en aza indirmek önemlidir. Benzerlik oranı ise hipotezlerin karşılaştırarak H_0 hipotezinde, parametrenin belirli bir değere eşit olma olasılığının maksimum olduğu benzerlik tahmininin, parametrenin bütün olası değerleri için maksimum benzerlik tahmini değerine oranlanmasıyla hesaplanır [39].

Hipotezlerle ilgili karar alma süreçlerine göre, Bayesyen ve Klasik yaklaşımlar farklılıklar göstermektedir. Hipotezlere olasılık atayarak, tekrarlanan denemelerin sonuçlarına dayanarak en yüksek olasılık değerini seçme yöntemi tümevarım yöntemiyle uyumluyken; belirli bir hipotezi yanlışlamak üzere seçme ve ardından tekrarlanan denemelerle bu hipotezi reddetme yaklaşımı tümdengelimci yöntem ile uyumludur [39].

2.5 Önsel Dağılım

Önsel bilgi, bayesci istatistikte kullanılır ve araştırmacının mevcut bilgisine göre eşlenik (conjugated), bilgi vermeyen (noninformative) ve sübjektif (bilgi içeren) önsel dağılım olmak üzere üç farklı kategoride sınıflandırılabilir. Bilgi vermeyen önsel dağılımdaki analizler klasik yaklaşım ile yapılanlara kıyasla eşdeğer çözümler

sunmaktadır. Ancak sübjektif ve eşlenik önsel dağılımda yapılan analizlerde elde edilen sonuçlar klasik yaklaşımdan farklılık göstermektedir [41].

Bayesci yaklaşımda önsel dağılımın olmaması durumunda Bayesci çıkarım ve istatistiksel modelleme gerçekleştirilemez. Çünkü sonsal dağılım fonksiyonu önsel dağılımın yardımı ile belirlenmektedir. Önsel olasılıklar için kesikli ya da sürekli dağılımlarda integral ya da toplam olasılık değerinin 1 değerine ulaşması tam (proper) önsel olasılık olarak adlandırılır. Tam olmayan (improper) önsel olasılık terimi ise bu koşulu sağlamayan durumları ifade eder [41].

Bayesci çıkarımda örneklem arttıkça önsel dağılımın sonsal dağılım üzerindeki etkisi azalmaktadır. Bu durumu kısaca özetlemek gerekirse, örneklem sayısının değişmesi olabilirlik fonksiyonunun basıklık veya sivrilik özelliğini etkileyebilir [41].

2.5.1 Bilgi Vermeyen Önsel Dağılım

Bu dağılım, kullanılan parametre için ön bilginin az olduğu veya önsel bilginin sonsal dağılıma etkisinin en az olduğu durumlarda verilerden elde edilen bilgi dışında bilgiye ihtiyaç duyulmadığında tercih edilir [45]. Bu yöntem kullanıldığında, analizler sonucunda elde edilen Bayesyen tahminlerle, Klasik yaklaşımla elde edinilen tahminler benzer sonuçlar verebilmektedir. Bilgi vermeyen önsel dağılımlar aşağıdaki gibidir [41]:

- Düzgün (üniform) önsel dağılımlar
- Belirsiz (diffuse/vague) önsel dağılımlar
- Jeffrey'in önsel dağılımı

Bilgi vermeyen önsel dağılımların çoğu tanım bölgesindeki integralin 1'e eşit olması koşulunu sağlamamaktadır. Yapılan çalışmalarda önsel dağılım eksik olsa dahi sonsal dağılımlar hep eksiksiz elde edilmektedir [41].

2.5.1.1 Düzgün (Üniform) Önsel Dağılımlar

Bu dağılımlarda belirlenen aralıkta θ parametresine eşit olasılık değerleri verilir [46]. Bu prensip, yetersiz neden ilkesine (principle of insufficient reason) dayanmaktadır. Bir olayın olma olasılığı başka bir olayın olma olasılığına göre daha mümkün değilse aynı olasılık değerleri atanır [39].

Yapılan bir araştırmada Laplace-Bayes önermesi θ parametresine ait az veya hiçbir önsel bilgi bulunmadığı durumlarda tercih edilir. Önerme Eşitlik 2.20'de gösterilmektedir [41]:

P (θ\Y, M) ∝ L ve θ∈R olmak üzere:

P (θ\Y, M) ∝ P (Y\θ, M) P(θ) (2.20)

2.5.1.2 Belirsiz Önsel Dağılımlar

Bayesyen yaklaşımda rastlantı değişkeni olarak kabul edilen parametrenin dağılımının cinsine bağlı olarak, bu parametrenin dağılım şeklini belirleyen parametre değişmektedir. Bu dağılımların şekil parametrelerine büyük varyansın belirsizlikle özdeş olması fikrine dayanarak geniş bir aralıkta yer alan ve hemen hemen düzgün, bilgi vermeyen bir ön dağılım oluşturulur [39].

$p(\theta) \sim N(\bar{\mu}, \bar{\sigma})$ ise $\bar{\sigma}^2$'ye büyük değer atanır.

$p(\sigma) \sim G(\alpha, \beta)$ ise α ve β'ya oldukça küçük değerler atanır.

Bu tarz belirsiz ön dağılım uygun bir dağılımdır ve bu özelliği belirsiz ön dağılımın güçlü yanını oluşturmaktadır. Belirsiz ön dağılımı düzgün ön dağılıma tercih etmek, söz konusu özelliğinden dolayı hesaplamaları olabildiğince kolaylaştırır [39].

Belirsiz ön dağılımın güçsüz yanlarından biri, son dağılımın uygun olmadığı ve gözlemlerin az olduğu durumlarda ortaya çıkabilir. Örneğin, benzerlik fonksiyonunun çok yayvan bir zirve yaptığı durumu düşünelim. Varsayalım ki belirsiz ön dağılım kullanılarak (örneğin büyük varyanslı normal dağılım) tekrar hesaplansa dahi sonuçlar iyi elde edilemez [47]. Çünkü bu durumda uygunluğu ile yeni ön dağılımın varyansına aşırı duyarlı bir son dağılım bulunmuş olabilir [39].

2.5.1.3 Jeffrey'in Önsel Dağılımı

Bilgi vermeyen önsel dağılım oluşturulabilmesi amacıyla Jeffrey'in geliştirdiği bu teorem, $p(y\backslash\theta)$ olabilirlik fonksiyonuna ilişkin Fisher'in bilgi matrisinin determinantının karekökü alınarak hesaplanmaktadır. Bu teoreme göre Eşitlik 2.21 aşağıdaki gibidir [41].

$$p(\theta) \propto |J(\theta)|^{1/2} \quad (2.21)$$

Fisher bilgi matrisi ise Eşitlik 2.22'de verildiği gibidir. Bu eşitlik yardımıyla herhangi bir bilgi vermeyen önsel dağılım hesaplanabilir [49].

$$J(\theta) = -\int \left|\frac{d^2 logp(y\backslash\theta)}{d\theta^2}\right| p(y\backslash\theta)dy \tag{2.22}$$

Jeffreys, ön dağılımı belirlerken başlangıçta Fisher tarafından geliştirilen bilgi matrisinden yararlanmıştır. Son dağılımın logaritmasını alarak dağılımın üslü değerini hesaplar ve katlı diferansiyelini alarak parametreleri yalnız bırakır. Ardından hesaplanan bu parametre fonksiyon ya da değerlerinin beklenen değerini alır. Nihayetinde varolan parametre bilgisine dayalı olarak parametre olasılık bölgesi parametre uzayında belirlenmiş olur [39].

Jeffrey önsel dağılımlarının farklı dağılımlar için elde edilmiş gösterimleri Tablo 2.2'de verilmiştir.

Tablo 2.2 Farklı dağılımların Jeffrey önsel dağılımları [48]

Olabilirlik Fonksiyonu	**Parametre(ler)**	**Önsel Dağılım**
Bernoulli	$\theta = p$	$(p(1-p))^{-1/2}$
Binom	$\theta = p$	$(p(1-p))^{-1/2}$
Poisson	$\theta = \lambda$	$\lambda^{-1/2}$
Normal (σ^2 biliniyor)	$\theta = \mu$	1
Normal (μ biliniyor)	$\theta = \sigma^2$	σ^{-1}
Normal	$\theta = \pi, \sigma^2$	σ^{-1}

2.5.2 Eşlenik Önsel Dağılımlar

Eşlenik ön dağılımının temelini benzerlik fonksiyonunun yeterli istatistiklerine göre çarpanlarına ayrılması ve bu yeterli istatistiklerin dağılımları oluşturmaktadır. Eşlenik ön dağılımda belirlenen ön dağılımla benzerlik fonksiyonundan son dağılım hesaplanabilir. Eşlenik ön dağılımın amacı önsel olasılık dağılımıyla ilk olasılık dağılımının aynı dağılım ailesine ait olmasını sağlamaktır. Bu sayede ilgilenilen parametreye ilişkin ön bilgiler yeniden gözden geçirilebilmekte ve daha tutarlı sonuçlara ulaşılabilmektedir [39]. Eşlenik dağılımlar için Karadağ (2011) [49] tarafından oluşturulan bilgiler Tablo 2.3'de verilmiştir.

Tablo 2.3 Olasılık fonksiyonlarının eşlenik ön dağılımları ve son dağılımları [48]

Olabilirlik Fonksiyonu	Parametre(ler)	Önsel Dağılım	Sonsal Dağılım
Bernoulli	P	Beta	Beta
Binom	P	Beta	Beta
Poisson	λ	Gamma	Gamma
Üstel	λ	Gamma	Gamma
Normal (σ^2 biliniyor)	μ	Normal	Normal
Normal (μ biliniyor)	σ^2	Ters Gamma	Ters Gamma
Normal	μ, σ^2	Normal-Gamma	Normal-Gamma
Tek Biçimli	a,b	Pareto	Pareto
Gamma	α, β	Gamma	Gamma

2.5.3 Sübjektif Önsel Dağılım

Sübjektif (bilgi veren) önsel dağılım olabilirlik fonksiyonunu etkilemeyen ancak sonsal dağılımı etkileyen bir dağılımdır. Bu dağılım, araştırmacının önceki çalışmalardan elde ettiği bilgileri içerir [50]. Bilgi veren önsel dağılımın oluşturulmasında yararlanılan bilgiler, önceden edinilmiş verilerden, araştırmacının önceki bilgilerinden ve uzman görüşlerinden elde edilen analizlerle birleştirilerek güncellenir [51]. Bu sayede mevcut verilerle entegre edilen önceki bilgiler, sonuçları güçlendirmek ve daha bilgili bir çıkarsama yapmak için kullanılır [41].

2.6 Sonsal Dağılım

Bayesci istatistiksel çıkarsamada önsel dağılımın belirlenmesinin önemi gibi sonsal dağılımın belirlenmesi de önemlidir. Araştırmalarda karşılaşılan problemlerden biri sonsal dağılımı belirleme sürecinde gerekli olan integrallerin çözümlerine ulaşılamaması ya da seçilen önsel dağılımların eşlenik aileye ait olmamasıdır [52]. Bu zorluklar, analizlerin karmaşıklığını artırabilir. Bu tür problemleri çözmek ve analizleri daha yönetilebilir hale getirmek için MCMC (Markov Chain Monte Carlo) metodu geliştirilmiştir. MCMC, Bayesci istatistikçiler tarafından sıklıkla kullanılan bir yöntemdir [41].

2.7 Markov Zinciri Monte Carlo Yöntemi

MCMC yöntemi, belirli bir olasılık dağılımından örneklem çıkarmak için kullanılan bir simülasyon tekniğidir [41]. Bayesci istatistikte sıkça karşılaşılan zorluk, çözümü kolay olmayan yüksek boyutlu integrallerden yararlanılmasıdır. Bu integraller sonsal dağılımları elde etmek için kullanılır. Monte Carlo yaklaşımı, Markov zinciri ve örnekleme yöntemine dayanarak bu yüksek boyutlu ve karmaşık integral problemlerine çözüm getiren stokastik bir yöntemdir. Oluşturulan örneklerle gerçekleştirilen benzetim yöntemi, dağılımdaki parametrelerin birbirlerine göre koşullu olasılıklarını elde etmek için kullanılır [45].

MCMC yöntemlerinde, Monte Carlo integrasyonu kullanılarak Markov zincirleri üzerinden işlem yapılır. Metropolis-Hastings algoritması ve Gibbs örnekleme algoritması en yaygın kullanılan MCMC yöntemlerindendir. Bu yöntemler, ilgili dağılımlardan örnekler çeker ve ardından örneklem ortalamaları kullanılarak beklenen değerler yaklaşık hesaplanır. MCMC yöntemlerinde, Markov zincirleri kullanılarak bu örneklemler çekilir. Bu zincirleri oluşturmak için çeşitli yollar mevcuttur ancak bu yöntemlerin tümü Metropolis ve Hastings tarafından geliştirilen algoritmanın özel durumlarıdır [45].

2.7.1 Markov Zincirleri

MCMC'de üretilen rasgele sayılar birbirine oldukça bağımlıdır ve her bir benzetim değeri bir önceki değere bağlı bir zincir değeri üretir. Eğer bu zincir yeterince uzun süre çalıştırılırsa, ilgilenilen son dağılımın istenilen şekle yakınsadığı kabul edilir. MCMC yönteminin bir özelliği, en uygun benzetim algoritması seçildiğinde, zincirin başlangıç değeri göz ardı edilerek sonsal dağılımın yakınsamasını mümkün kılmasıdır [53]. Zincir uzun bir benzetim sürecinde çalıştırılırsa elde edilen sonsal dağılım daha yüksek doğruluk seviyelerine ulaşabilir [54].

Stokastik süreç, bilinen bir durum uzayında (Θ) tanımlanan tesadüfi büyüklüklerin ardışık bir kümesidir. Θ, bir sıralama içeren $\{\theta^{[t]} : t \in T\}$ dizisi olduğundan sürecin sırası bilinmektedir. Durum uzayı (state space), bu sürecin tüm olası değerler kümesi olarak adlandırılır [55].

Markov zinciri boyutu $t = (1,2, \dots, n)$ olan θ^{t-1} önceki rasgele değişkenine bağlı olarak oluşan θ^t rasgele değişkenlerinin oluşturduğu stokastik bir süreçtir [41].

$$p\left(\theta^{|t|} \epsilon\, A \backslash \theta^{|0|}, \theta^{|1|}, \ldots, \theta^{|t-2|}, \theta^{|t-1|}\right) - p\left(\theta^{|t|} \epsilon\, A \backslash \theta^{|t-1|}\right) \quad (2.23)$$

Burada A, tüm durum uzayında tanımlanmış herhangi bir kümedir. Markov zinciri, sadece bir önceki duruma bağlı olan sürecin herhangi bir durumunun tesadüfi bir sürecidir ve bu özelliğe "Markov özelliği" denir [55].

Kesikli durum uzayına sahip Markov sürecine "Markov zinciri" denir ve bu zincirler kesikli veya sürekli parametreli olabilir [45].

Kesikli parametreli bir Markov zinciri $\{X_t, t = 0,1,2, \ldots\}$ için, $p_{ij} = P\left(X_{m+1} = \frac{j}{X_m} = i\right)$ durum uzayında yer alan i durumundan j durumuna bir adımda geçiş olasılığını, $p_{ij}^{(n)} = P\left(X_{m+n} = \frac{j}{X_m} = i\right)$ i durumundan j durumuna n adımda geçiş olasılığını, $p_j^{(n)} = P(X_n = j)$ ise sürecin n'inci adımda j durumunda bulunma olasılığını gösterir. İlk olasılık olarak adlandırılan sürecin başlangıcındaki bu olasılıkların oluşturduğu vektör π_0'dır. Bir π_0 başlangıç vektörünün belirlenmesiyle zincir başlar. Başlangıç vektörünün bir elemanı haricinde diğer elemanların hepsi 0 değerini alır. π_n ise n'inci adımda, durum uzayında yer alan durumların olasılıklarının oluşturduğu vektördür [45].

Durum uzayı $S = \{0,1,2, \ldots\}$ olan bir Markov zincirinde, geçiş matrisi olarak adlandırılan bir adım geçiş olasılıklarının oluşturduğu matris $P = \left[p_{ij}\right]$ ile gösterilir. n adım geçiş olasılıklarına sahip olan matris ise $P^{(n)} = P^n = \left[p_{ij}^{(n)}\right]$ şeklinde tanımlanır [45].

Zincirin i durumundan j durumuna n+r adımda geçiş olasılığı aşağıdaki "Chapman-Kolmogorov" ifadesi kullanılarak bulunur [45].

$$p_{ij}^{(n+r)} = \sum_k p_{ik}^{(n)} p_{kj}^{(r)} \quad (2.24)$$

Chapman-Kolmogorov eşitliği matris biçiminde $P^{(n+r)} = \left[P_{ij}^{(n+r)}\right] = P^{(n)} P^{(r)}$ olarak ifade edilebilir [45].

Eğer durum uzayındaki tüm durumlar arasında geçiş varsa yani $p_{ij} > 0$ tüm i, j'ler için geçerliyse Markov zinciri durağan bir dağılıma sahiptir. Durağan dağılım $\pi^* = \pi^* P$ eşitliğini sağlar [45].

2.7.2 Monte Carlo İntegrasyonu

Monte Carlo yönteminde $h(x)$ biçimindeki karmaşık bir fonksiyonun integralini rasgele sayı üretilerek çözmeyi amaçlar. Bu integral aşağıdaki şekilde ifade edilebilir [45].

$$\int_a^b h(x)dx = \int_a^b g(x)p(x)dx = E_{p(x)}[g(x)] \quad (2.25)$$

Burada $g(x)$ raslantı değişkeninin bir fonksiyonu ve $p(x)$ ise x'in olasılık yoğunluk fonksiyonudur. Eğer $p(x)$ olasılık yoğunluk fonksiyonundan $x_1, x_2, \dots, x_n$ biçiminde bir örneklem çekilirse, Eşitlik (2.25) ile verilen integral yaklaşık olarak aşağıdaki biçimde yazılabilir [45].

$$\int_a^b h(x)dx = E_{p(x)}[g(x)] \cong \frac{1}{n}\sum_{i=1}^{n} g(x_i) \quad (2.26)$$

Sonuçta MCMC yöntemiyle ana kitleden çekilen örneklem değerlerinin ortalamasının hesaplanmasıyla karmaşık hedef fonksiyonunun beklenen değeri yaklaşık olarak bulunur. Eşitlik (2.26) Monte Carlo integrasyonunu göstermektedir ve bu yöntemle Bayesci çıkarsamada sonsal dağılımın değeri yaklaşık olarak elde edilebilir [43].

2.7.3 Metropolis-Hastings Algoritması

MCMC algoritmalarının en yaygın kullanılanı Metropolis-Hastings (MH) algoritmasıdır [57]. $x \in E$ ile bir E durum uzayı üzerinde bulunan $f(x)$ dağılımından örneklem çekmekle ilgilenildiği varsayılsın. x noktasından y noktasına geçmek için bir $K(x,y)$ geçiş çekirdeği oluşturmak üzere, MH algoritması iki adımlı bir süreç kullanır [57].

- $q(y|x)$ ifadesi bir teklif dağılımı belirtir.
- Kabul oranı $a(x,y) = min\left[\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right]$ ile $q(y|x)$ dağılımından gelen çekilişleri kabul eder.

Böylece geçiş çekirdeği $K(x,y) = q(y|x)a(x,y)$ ile verilir. Metropolis-Hastings algoritmasında aşağıdaki adımlar takip edilir [58].

1. $q(y|x_t)$ dağılımından yeni bir y örneği alır.
2. $U(0,1)$ dağılımından bir u üretilir.
3. $x_{t+1} = \begin{cases} y & olasılıkla\ a(x_t,y) \\ x_t & olasılıkla\ 1-a(x_t,y) \end{cases}$

Burada;

$a(x,y) = min\left[\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right]$ [57]. (2.27)

2.7.4 Gibbs Örneklemesi

Gibbs örneklemesi, Bayesci hesaplamaları yapmak için yaygın olarak kullanılan bir yöntemdir ve Geman (1984) [59] tarafından tanıtılmıştır. $x \in \chi \subseteq R^d$'de $f(x)$ dağılımdan örneklem çekmek istediğimizi düşünelim. Gibbs örneklemesi, geçiş çekirdeğini $K(x,y)$ birden çok adıma bölerek kullanır. Her adımda, diğer koordinatlara göre koşullu yoğunluğuna dayalı olarak bir koordinat ilerletilir. Gibbs örneklemesi, hedef dağılımın bir değişkeninin koşullu dağılımından tekrar tekrar örnekleme yapan bir MCMC algoritmasıdır. Algoritma aşağıdaki gibidir [57]:

1. Başlangıç noktası $x^0 = (x_1^{(0)}, x_2^{(0)}, \dots, x_p^{(0)})$ olarak alınır. Burada t=0'dır.
2. t=1, 2, ... döngü sayısı olarak tanımlanır ve her bir x bileşeni aşağıdaki gibi oluşturulur.

   $f(x^1|x_{(t)}^2, \dots, x_{(t)}^d, y)$'den $x_{(t+1)}^1$

   $f(x^2|x_{(t)}^1, x_{(t)}^3 \dots, x_{(t)}^d, y)$'den $x_{(t+1)}^2$

   ⋮

   $f(x^d|x_{(t)}^1, \dots, x_{(t+1)}^{d-1}, y)$'den $x_{(t+1)}^d$

   elde edilir.
3. $t = t + 1$ alınır ve $t < T$ ise 2. adıma tekrar gidilir. Bu koşul sağlanmıyorsa döngü bitirilir [41].

Tam geçiş çekirdeği aşağıdaki gibi yazılır:

$K_{1\to d}(x_{t+1}|x_t) = \prod_{i=1}^{d} f(x_{t+1}^i|x_{t+1}^1, \dots, x_{t+1}^{i-1}, x_t^{i+1}, \dots, x_{t+1}^d)$. (2.28)

Eşitlik (2.28)'de verilen ifade basitleştirilirse,

$f(x_t)K_{1\to d}(x_{t+1}|x_t) = f(x_{t+1})K_{d\to 1}(x_t|x_{t+1})$. (2.29)

Eşitlik (2.29)'un her iki tarafının integrali alınırsa,

$\int f(x_t)K_{1\to d}(x_{t+1}|x_t)dx = f(y)$. (2.30)

Böylece f, geçiş çekirdeği $K_{1\to d}(x_{t+1}|x_t)$ tarafından oluşturulan Markov zincirinin durağan dağılımını ifade eder. Yukarıda verilen Gibbs örnekleyici tersine çevrilemez. Ancak her yinelemede güncellenecek rastgele bir bileşen seçilerek (rastgele tarama) tersine çevrilebilir olanlar üretilebilir. Rastgele taramalı Gibbs örnekleyici, kabul oranı $min\left[\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right] = 1$ olan MH örnekleyicinin özel bir durumu olarak görülebilir [57].

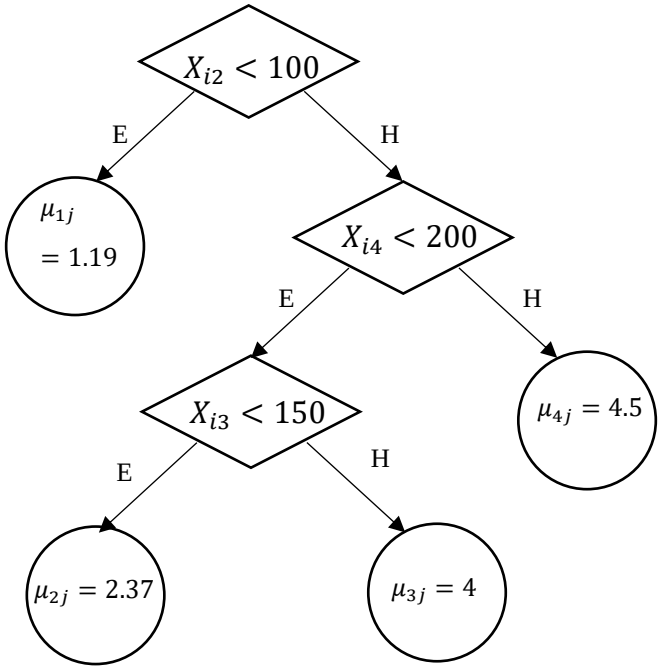
3
BAYES TOPLAMSAL REGRESYON AĞAÇLARI

3.1 Bayes Regresyon Ağaçlarının Toplamı

Ağaç temelli regresyon yöntemleri (CART), Rastgele ormanlar (Random Forests) ve gradyan artırma oldukça başarılıdır ve denetimli öğrenme için yaygın olarak kullanılır. Bayes toplamsal regresyon ağaçları (BART) genellikle üstün tahmin/tahmin doğruluğu sağlayan ancak daha az bilinen bir yöntemdir. BART modeli ilk olarak Chipman ve arkadaşları tarafından (2010)'de [4] ortaya çıkmıştır [60].

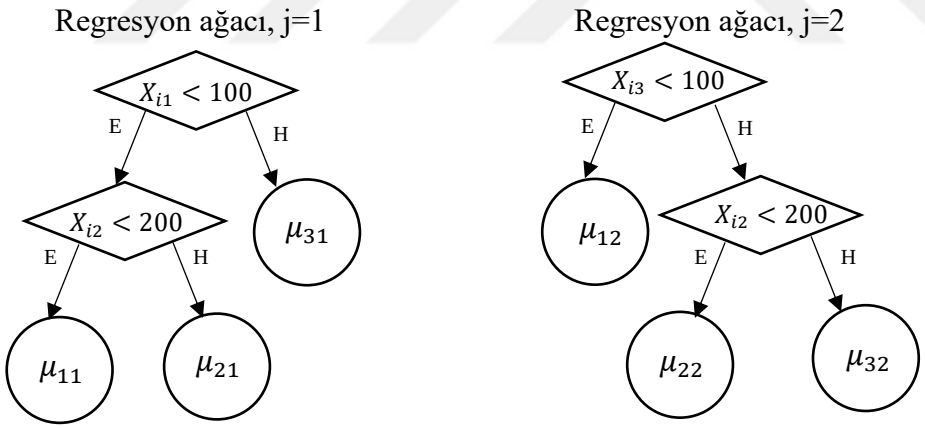
ϵ_i'nin sıfır ortalamalı ve σ^2 varyanslı Normal dağılımlı olduğu ve $f(x)$'nin bilinmeyen bir fonksiyon olduğu varsayıldığında $y_i = f(x_i) + \epsilon_i$ ifadesi toplamsal hata ortalama regresyon modelini ifade eder. BART modeli, bu ifadedeki $f(x)$'i tahmin etmeye çalışır ve $f(x)$'i tahmin etmek için ise regresyon ağaçlarının toplamını şu şekilde belirtir [60]:

$$f(x) = \sum_{j=1}^{m} g_j(x, T_j, \mu_j) \tag{3.1}$$



Şekil 3.1 Bir regresyon ağacı örneği $g_j(x, T_j, \mu_j)$ [35]

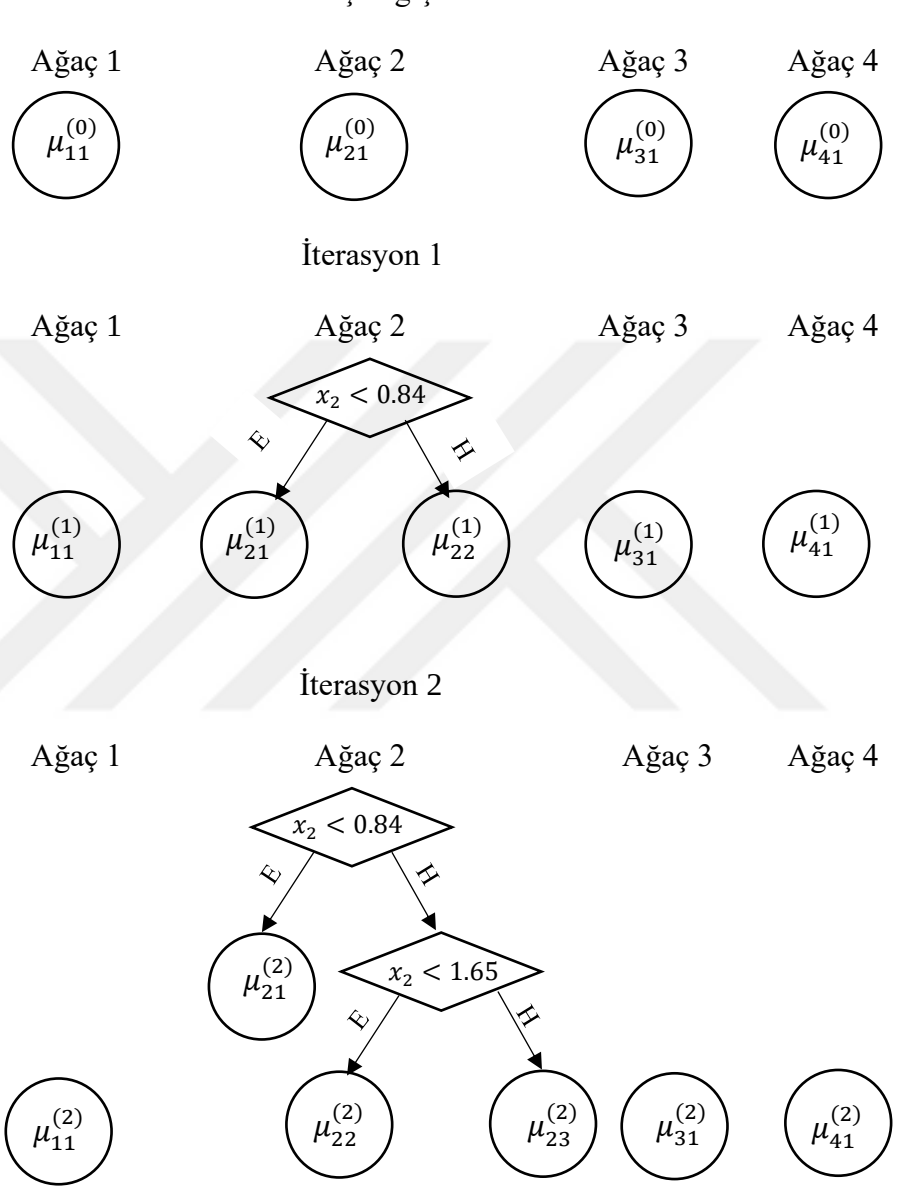
Şekil 3.1'de verilen $g_j(x, T_j, \mu_j)$ regresyon ağacında $X_i = (X_{i1}, \ldots, X_{i4})$ ortak değişkenler olduğu varsayılsın ve i konusu için $E(Y_i|X_i)$ bilinmek istensin. Bölünmenin olduğu her yere düğüm denir. Üst düğümde (kök), $X_{i2} < 100$ koşulu vardır. $X_{i2} < 100$ doğruysa, yol sola, aksi takdirde sağa izlenir. $X_{i2} < 100$'ün doğru olduğu varsayılırsa, bölünmemiş bir düğüme ulaşıldığı görülür. Buna terminal düğüm denir ve μ_{1j} = 1.19 parametresi Y_i'nin tahmin edilen değeri olarak kullanılır. Bunun yerine $X_{i2} < 100$'ün doğru olmadığı varsayılsın. Ardından, sağ taraf boyunca hareket ederken, $X_{i4} < 200$ koşuluna sahip başka bir dahili düğümle karşılaşılır. Bu koşul kontrol edilir ve eğer bu koşul doğru ise sola, yanlış ise sağa giden yol takip edilir. Bu işlem bir uçbirim düğümüne ulaşana kadar devam eder ve o uçbirim düğümündeki μ_{kj} parametresi Y_i 'nin tahmin edilen değeri olarak atanır. μ_{kj}, j. regresyon ağacı için k. düğümün ortalama parametresidir. Bu nedenle, örneğin, $X_{i1} = 30, X_{i2} = 120, X_{i3} = 115$ ve $X_{i4} = 191$ olan bir i öznesine, μ_{2j} = 2.37 tahmini sonucu atanacaktır. Tahmin, bunun yerine $X_{i'1} = 130, X_{i'2} = 135, X_{i'3} = 192$ ve $X_{i'4} = 183$ ortak değişkenleri olan başka bir i' konusu için μ_{3j} = 4 olacaktır [35].



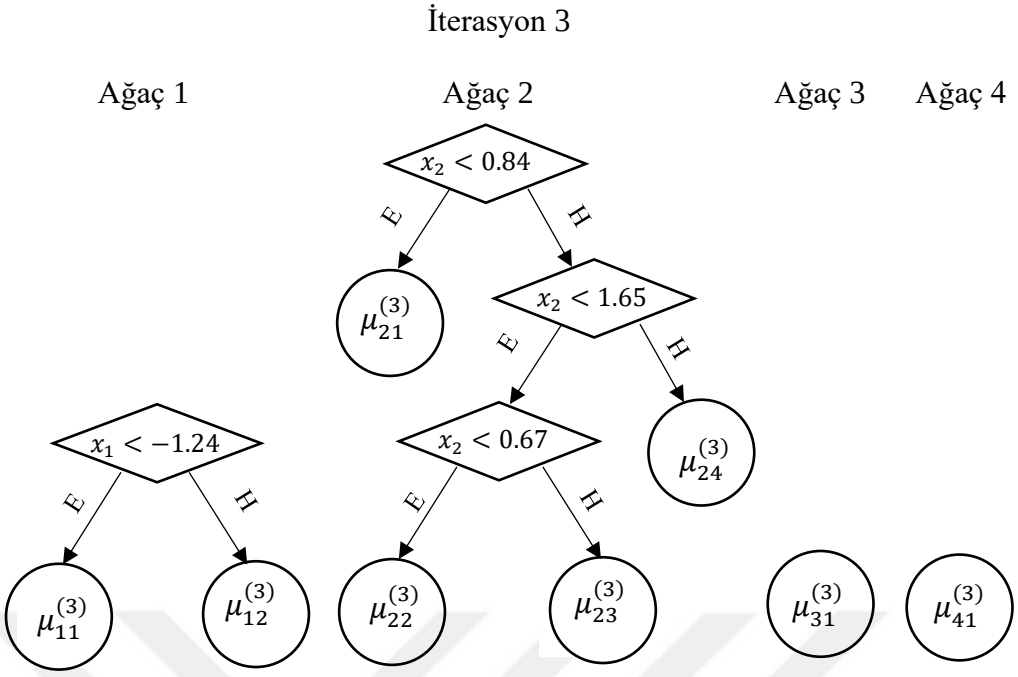
Şekil 3.2 Basit iki regresyon ağacı örneği kullanarak regresyon ağaçlarının toplamını gösterme [35]

Bir regresyon ağaçları toplamı kavramını görmenin yolu, şekildeki regresyon ağaçlarını ANOVA modelleri olarak düşünmektir [4]. Ağaçların toplamının ANOVA modeli:

$$y = g(X{:}\,T_1, M_1) + \; g(X{:}\,T_2, M_2) + \; \varepsilon_i$$
$$= \; \mu_{11} I\{X_{i1} < 100\} I\{X_{i2} < 200\} + \mu_{21} I\{X_{i1} < 100\} I\{X_{i2} \geq 200\}$$
$$+\mu_{31} I\{X_{i1} \geq 100\} + \mu_{12} I\{X_{i3} < 100\} + \mu_{22} I\{X_{i3} \geq 100\} I\{X_{i2} < 200\}$$
$$+\mu_{32} I\{X_{i3} \geq 100\} I\{X_{i2} \geq 200\} + \; \varepsilon_i \qquad (3.2)$$



Şekil 3.3 BART'ın m = 4 ile BART içindeki MCMC adımlarının iterasyon 3'üne göre sezgisi [35]



Şekil 3.3 BART'ın m = 4 ile BART içindeki MCMC adımlarının iterasyon 3'üne göre sezgisi (devamı) [35]

Algoritmanın ilk adımı (başlangıç), dört regresyon ağacının tek köklü düğümlerle başlatılmasıdır. Dört regresyon ağacının tümü tek köklü düğümler olduğundan, bu düğümler için başlatılan parametreler $\mu_{ij}^{(0)} = \frac{\bar{y}}{m} = \frac{\bar{y}}{4}$ şeklinde olacaktır. BART, ilk başlangıçtan sonra MCMC yinelemesinde her bir regresyon ağacı için ağaç yapılarını çizmeye başlar. Genelliği kaybetmeden ilk regresyon ağacını (T_1, M_1) belirleyerek başlar. İlk önce $R_1 = Y - \sum_{j\neq 1} g\left(X, T_j, M_j\right) = Y - [g(X, T_2, M_2) + g(X, T_3, M_3) + g(X, T_4, M_4)] = Y - 3 \times \frac{\hat{Y}}{4}$ hesaplanır. Daha sonra, bu yineleme için T_1 ağaç yapısının arka çizimini belirleyen bir MH algoritması kullanılır. MH'nin temel fikri, T_1 'den yeni bir ağaç yapısı önermek, buna T_1^* adını vermek ve ardından aşağıdakileri dikkate alarak T_1^*'in kabul edilip edilmemesinin olasılığını hesaplamaktır: $R_1|T_1^*$ (yeni ağaç yapısı verilen kalıntı olasılığı), $R_1|T_1$ (önceki ağaç yapısı verilen kalıntı olasılığı), T_1^* 'i gözlemleme olasılığı, T_1 'i gözlemleme olasılığı ve T_1^* 'den T_1 'e hareket olasılığı ve T_1 'den T_1^* 'e hareket etme olasılığı. T_1^* kabul edilirse, T_1 , T_1^* olacak şekilde güncellenir. Şekil 3.3'ten, ilk iterasyonda T_1^* 'in kabul edilmediği görülebilir, bu nedenle ağaç yapısı tek bir kök düğüm olarak kalır.

Algoritma daha sonra M_1 'i T_1 için yeni güncellenmiş regresyon yapısına dayalı olarak günceller ve (T_2, M_2) belirlemeye devam eder [35].

(T_2, M_2)'yi belirlemek için, algoritma yine $R_2 = Y - \sum_{j \neq 2} g\left(X, T_j, M_j\right) = Y - [g(X, T_1, M_1) + g(X, T_3, M_3) + g(X, T_4, M_4)] = Y - [\hat{\mu}_{11}^{(1)} + 2 \times \frac{\hat{Y}}{4}]$, burada $\hat{\mu}_{11}^{(1)}$, regresyon ağacı 1 için güncellenmiş parametredir. Benzer şekilde, MH yeni bir T_2^* önermek için kullanılır ve R_2, T_2^* 'nin kabul edilip edilmeyeceğine karar vermek için kabul olasılığını hesaplamak için kullanılır. Şekil 3.3'te yeni önerilen T_2^* 'nin kabul edildiği ve dolayısıyla yineleme 1'de T_2 için yeni bir ağaç yapısının kullanıldığı görülebilir [35].

(T_3, M_3) için ise yine R_3, yeni önerilen T_3^*'ü kabul etmediği dolayısıyla $g(X, T_3, M_3)$ için çizilen tek bir $\hat{\mu}_{31}^{(1)}$ parametresinin kullanıldığı görülmektedir [35].

Sonuç olarak, yineleme 1 için R_4 hesaplanırken, bu $R_4 = Y - \left[\hat{\mu}_{11}^{(1)} + \hat{\mu}_{21}^{(1)} I\{X_2 < 0.84\} + \hat{\mu}_{22}^{(1)} I\{X_2 \geq 0.84\} + \hat{\mu}_{31}^{(1)} + \hat{\mu}_{41}^{(1)}\right]$ olur. T_4^* kabul edilmedi ve (T_4, M_4) için ağaç yapısı olarak tek bir T_4 düğümü güncellendi. Regresyon ağacı çizimleri tamamlandıktan sonra BART, BART modelindeki parametrelerin geri kalanını çizmeye devam eder [35].

3.2 BART Algoritması

İlk olarak, BART ön dağılımlarla başlar. BART modeli için ön dağılım $P(T_1, M_1, \ldots, T_m, M_m, \sigma)$ şeklindedir. Her zamanki ön belirti şudur: $\{(T_1, M_1), \ldots, (T_m, M_m)\}$ ve σ bağımsızdır ve (T_1, M_1), ..., (T_m, M_m) birbirinden bağımsızdır.

Daha sonra önsel dağılım şu şekilde yazılabilir [61]:

$$
\begin{aligned}
P(T_1, M_1, \ldots, T_m, M_m, \sigma) &= P(T_1, M_1, \ldots, T_m, M_m) P(\sigma) \\
&= \left[\prod_j^m P\left(T_j, M_j\right)\right] P(\sigma) \\
&= \left[\prod_j^m P\left(M_j \backslash T_j\right) P\left(T_j\right)\right] P(\sigma) \\
&= \left[\prod_j^m \left\{\prod_k^{b_j} P(\mu_{kj} \backslash T_j)\right\} P\left(T_j\right)\right] P(\sigma) \qquad (3.3)
\end{aligned}
$$

Önsel dağılım, sonsal dağılıma neden olur.

$$[P(T_1, M_1), \dots, (T_m, M_m), \sigma \backslash Y] \propto P(Y \backslash (T_1, M_1), \dots, (T_m, M_m), \sigma) \times P\big((T_1, M_1), \dots, (T_m, M_m), \sigma\big) \quad (3.4)$$

Gibbs örneklemesi kullanılarak;

$$P[\big(T_j, M_j\big) \backslash T_{(j)}, M_{(j)}, Y, \sigma] \quad (3.5)$$

j. ağaç hariç uygun ağaçların m − 1 regresyon toplamının artıkları

$$R_{ij} = g\big(X_i, T_j, M_j\big) + \varepsilon_i \quad (3.6)$$

Böylece Gibbs örneklemesinden elde edilen ifade, tek bir regresyon ağacından elde edilen aşağıdaki sonsal çizimlere eşdeğerdir.

$$R_j = Y - \sum_{u \neq j} g(X, T_w, M_w) \quad \text{veya} \quad P[\big(T_j, M_j\big) \backslash R_j, \sigma] \quad (3.7)$$

$P(T_j | R_j, \sigma)$ elde etmek için önce M_j'nin integralini alarak bir önceki eşitlikten bir beraberlik elde edilir. Bir MH algoritması kullanarak $P(T_j | R_j, \sigma)$ çizilir, burada ilk önce j. ağaç için $q(T_j, T_j^*)$ olasılık dağılımına sahip T_j^* aday ağacı üretilir ve sonra T_j^* 'yi olasılık ile kabul eder [35].

$$\alpha\big(T_j, T_j^*\big) = min\left\{1, \frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} \frac{P(R_j | X, T_j^*, M_j)}{P(R_j | X, T_j, M_j)} \frac{P(T_j^*)}{P(T_j)}\right\} \quad (3.8)$$

$\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)}$, önceki ağacın yeni ağaca nasıl hareket etme olasılığının, yeni ağacın önceki ağaca nasıl hareket etme olasılığına oranıdır, $\frac{P(R_j | X, T_j^*, M_j)}{P(R_j | X, T_j, M_j)}$ yeni ağacın önceki ağaca göre olabilirlik oranıdır ve $\frac{P(T_j^*)}{P(T_j)}$ önceki ağaca karşı yeni ağacı ifade etmektedir [35].

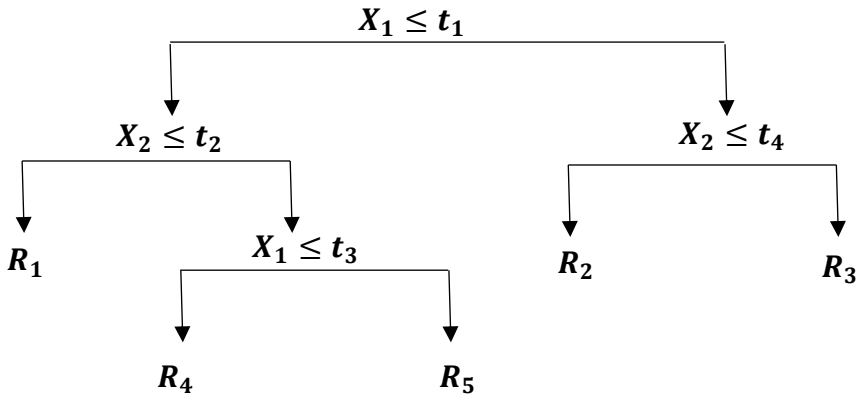
Dört yerel adım kullanılarak önceki T_j ağacına göre yeni bir T_j^* ağaç önerilebilir:
(i) büyüme; bir terminal düğümünün iki yeni alt düğüme bölündüğü yerde büyümek.
(ii) budama; aynı terminal olmayan düğümün hemen altındaki iki terminal alt düğümü, üst terminal olmayan düğümü bir terminal düğümü olacak şekilde birleştirilir.

(iii) takas; iki terminal olmayan düğümün bölünme kriterleri değiştirilir. (iv) değişiklik, tek bir terminal olmayan düğümün bölünme kriterleri değiştirilir. $P(T_j, R_j, \sigma)$ çizimini elde ettikten sonra, $P(\mu_{kj}|T_j, R_j, \sigma) \sim N\left(\frac{\sigma_\mu^2 \sum_{k=1}^{n_k} R_{kj}}{n_k\sigma_\mu^2+\sigma^2}, \frac{\sigma^2\sigma_\mu^2}{n_k\sigma_\mu^2+\sigma^2}\right)$ çizilir. Burada R_{kj}, μ_{kj} uçbirim düğüm parametresine tahsis edilen R_j'deki öğelerin alt kümesidir ve n_k, μ_{kj}'ye tahsis edilen R_{kj}'lerin sayısıdır [35].

4
SINIFLANDIRMA VE REGRESYON AĞAÇLARI

Yanıt değişkeni kesikli olduğu zaman ağaç tabanlı algoritmalar yöntemlerinden biri olan sınıflandırma; sürekli olduğu zaman ise regresyon ağacı olarak adlandırılır. Bu algoritmaya genel olarak sınıflandırma ve regresyon ağaçları yani CART denir. Sınıflandırma ya da regresyon ağacı, örneklemdeki her birim ile ilgili bir dizi soruların hiyerarşik gösterimidir. Bu sorular, her birimdeki sınıflama verilerinin değerleriyle ilgilidir. Bu sorular yanıtlandığı zaman her birimin en çok hangi sınıfa ait olduğu bilinir. Bu bilginin genel gösterimi ağaç olarak adlandırılır çünkü soruların yukarıdan aşağıya mantıksal gösterimi, düğümleri bağlayan bir dizi dallar ve altta da yapraklar şeklinde olan kökten uca ağaç biçimindedir. Her bir düğümde, sınıflama değişkenlerinden biri ile ilgili bir soru ortaya atılır ve düğümde yer alacak dal bu sorunun cevabına bağlıdır. Soruların sorulmasındaki sırasının tanımlanması önemlidir çünkü ağacın yapısı buna göre belirlenir. Bunu yapmanın birçok yolu varsa da genel ilke, her düğüm ayırma fırsatında düğüm saflığındaki kazancı maksimum edecek soruyu sormaktır; burada düğümün saflığı, düğümdeki yanıtın değişkenliğini minimum yaparak arttırılır. Böylece eğer yanıt kesikli bir sınıflandırma ise daha yüksek saflık, daha az sayıda sınıf ya da kategori anlamına gelir. Bir düğüm, yanıtın tek bir sınıf ya da kategorisini içerdiğinde tamamen saf olur. Eğer yanıt sürekli ise o zaman düğümdeki yanıtların değişkenliğin bir ölçüsü olarak örneğin standart sapma, hata kareler ortalaması ya da ortalama mutlak sapma kullanılır ki düğümün saflığının maksimum olabilmesi için bu ölçüler olabildiğince küçük olmalıdır [62].

CART, girdi alanını yinelemeli olarak bölümlere ayırarak ve elde edilen her bölgede yerel bir model tanımlayarak tanımlanan karar ağaçları olarak da bilinir. CART yaklaşımını açıklamak için aşağıdaki şekilde yer alan ağacı göz önünde bulundurabiliriz [63].



Şekil 4.1 Basit regresyon ağacı örneği [63]

Şekil 4.1'de verilen regresyon ağacında bölünmenin olduğu her yere düğüm denir. Üst düğümde (kök), x_1'in t_1 değerinden küçük veya eşit olup olmadığı koşulu vardır. Bu koşul doğruysa, yol sola, aksi takdirde sağa izlenir. Koşulun doğru olduğu varsayılırsa, x_2'nin başka bir t_2 değerinden küçük veya eşit olup olmadığı koşuluna ulaşılır. Aynı şekilde koşul doğruysa sola, yanlışsa sağa doğru bir yol izlenir. Bu koşulunda doğru olduğu varsayılırsa, bölünmemiş bir düğüme ulaşılır. Buna terminal düğüm denir ve R_1 parametresi tahmin edilen değer olarak kullanılır. Bu işlem bir uçbirim düğümüne ulaşana kadar devam eder ve o uçbirim düğümündeki R parametresi tahmin edilen değer olarak atanır.

5
UYGULAMA

5.1 Kullanılan Programlar

Analizler, R Studio programında yapılmıştır. R Studio programında yer alan bartMachine kütüphanesi kullanılarak ikinci el otomobil verisinin bayes toplamsal regresyon ağaçlarının modeli ve özet çıktıları elde edilmiştir.

5.2 Veri Seti

Anakütleyi ülkemizdeki ikinci el otomobillerin oluşturduğu bu çalışmada Türkiye'de ikinci el otomobillerin hangi özelliklerinin otomobil fiyatlarını etkilediği tespit edilmeye çalışılmıştır. İkinci el otomobil web sitesinden (http://www.sahibinden.com) 2022 Aralık ayı içerisinde satışta olan toplam otomobil sayısının 11701 olduğu belirlenmiştir. Veriyi %80-%20 oranlarında eğitim ve test verisi olarak ayırmış olup, eğitim veri seti 9363 ve test veri seti ise 2338 gözlemden oluşmaktadır.

Literatürde yer alan çalışmalardan yola çıkılarak BART modeli oluşturmak için ikinci el otomobil fiyatlarını (P) etkileyebilecek bağımsız değişkenlerinden marka ve model (MM_1: BMW 116d, MM_2: BMW 318i, MM_3: Citroen C4, MM_4: Dacia Duster, MM_5: Dacia LODGY, MM_6: Dacia LOGAN, MM_7: Dacia SANDERO, MM_8: Fiat EGEA, MM_9: Fiat LINEA, MM_{10}: Ford FIESTA, MM_{11}: Ford FOCUS, MM_{12}: Honda CR-V, MM_{13}: Hyundai ACCENT, MM_{14}: Hyundai ELANTRA, MM_{15}: Hyundai GETZ, MM_{16}: Hyundai i10, MM_{17}: Hyundai i20, MM_{18}: Hyundai i30, MM_{19}: Hyundai ix35, MM_{20}: Hyundai TUCSON, MM_{21}: Kia CEED, MM_{22}: Kia RİO, MM_{23}: Kia SPORTAGE, MM_{24}: Mercedes A180, MM_{25}: Mini COOPER, MM_{26}: Nissan QASHQAI, MM_{27}: Opel ASTRA, MM_{28}: Opel CORSA, MM_{29}: Opel INSIGNIA, MM_{30}: Opel MOKKA, MM_{31}: Peugeot 208, MM_{32}: Peugeot 308, MM_{33}: Peugeot 508, MM_{34}: Renault CLIO, MM_{35}: Renault GRAND SCENIC, MM_{36}: Renault KADJAR, MM_{37}: Renault MEGANE, MM_{38}: Renault SYMBOL, MM_{39}: Seat IBIZA, MM_{40}: Seat LEON, MM_{41}: Skoda FABIA, MM_{42}: Skoda OCTAVIA, MM_{43}: Skoda RAPID, MM_{44}: Toyota COROLLA, MM_{45}: Volkswagen BORA, MM_{46}: Volkswagen CC, MM_{47}: Volkswagen GOLF, MM_{48}:

Volkswagen JETTA, MM_{49}: Volkswagen PASSAT, MM_{50}: Volkswagen POLO, MM_{51}: Volkswagen TIGUAN, MM_{52}: Volvo S60), renk (R, gri:1, diğerleri:0), yaş (A), kasa tipi (KT, sedan:1, diğerleri:0), çekiş (C, önden-arkadan:1, diğerleri:0), kilometre (KM, km), vites (V : otomobillerin hız ve çekişini ayarlamaya yarayan dişliler düzeneği. Manuel:0, otomatik:1), hızlanma (H, sn), ortalama yakıt tüketimi (OYT, lt), ağırlık (W, kg), silindir sayısı (cylinder count, SS, 4:1, diğerleri:0), yakıt türü (YT: mekanik enerji oluşturmak için kullanılabilen yanıcı bir madde. Benzin:1, dizel:0), motor gücü (MG, beygir gücü, hp, 0-100:1, diğerleri:0), motor hacmi (MH: pistonların alt ölü nokta ile üst ölü nokta arasındaki hacmi, 0-1400:1, diğerleri:0), tork (T : motor tarafından üretilen dönme kuvveti, Nm), en yüksek hız (AH, km/h), uzunluk (U, mm.), yükseklik (Y, mm.), genişlik (G, mm.), bagaj kapasitesi (BK, lt.), turbo (TU: varsa:1, yoksa:0), iç donanım özellikleri (ID_1: 7 Koltuk, ID_2: Anahtarsız çalıştırma, ID_3: Arka kol dayama, ID_4: Bluetooth, ID_5: Deri koltuk, ID_6: Dijital klima, ID_7: Elektrikli sürücü koltuğu, ID_8: Fonksiyonel direksiyon, ID_9: Hafızalı sürücü koltuğu, ID_{10}: Hız sabitleyici, ID_{11}: Isıtmalı arka koltuklar, ID_{12}: Isıtmalı ön koltuklar, ID_{13}: Navigasyon, ID_{14}: Otomatik kararan dikiz aynası, ID_{15}: Ön kol dayama, ID_{16}: Start/Stop, ID_{17}: USB Bağlantısı, ID_{18}: Yol bilgisayarı), dış donanım özellikleri (DD_1: Adaptif far, DD_2: Alaşımlı jant, DD_3: Cam tavan, DD_4: Elektrikli aynalar, DD_5: Far sensörü, DD_6: Far yıkama, DD_7: Geri dönüş kamerası, DD_8: Isıtmalı aynalar, DD_9: Katlanır aynalar, DD_{10}: LED Gündüz sürüş farları, DD_{11}: Park sensörü, DD_{12}: Sis farı, DD_{13}: Sunroof, DD_{14}: Xenon far, DD_{15}: Yağmur sensörü), teknik donanım özellikleri (TD_1: ABS-Anti Blokaj Fren Sistemi, TD_2: ASR-Elektronik Patinaj Önleme Sistemi, TD_3: Diferansiyel kilidi, TD_4: Diz hava yastığı, TD_5: Elektrik takviyeli direksiyon, TD_6: ESP-Elektronik Stabilizasyon Programı, TD_7: Hıza duyarlı direksiyon, TD_8: Isofix, TD_9: Lastik basınç göstergesi, TD_{10}: LDW-Şeritten Ayrılma İkazı, TD_{11}: Mesafe takip sistemi, TD_{12}: Otomatik park sistemi, TD_{13}: Perde hava yastığı, TD_{14}: Yokuş kalkış desteği) ve satış fiyatı (P) oluşan toplam 120 değişken vardır. Modelde kullanılan kategorik değişkenlere ilişkin tanımlayıcı istatistikler Tablo 5.1'de ve sürekli değişkenlere ilişkin tanımlayıcı istatistikler ise Tablo 5.2'de görülmektedir.

5.3 Bulgular

5.3.1 Tanımlayıcı İstatistikler

Tablo 5.1 Kategorik değişkenlere ait frekans değerleri

Değ.	Eğitim	Test	Değ.	Eğitim	Test	Değ.	Eğitim	Test
MM_1	37	11	KT_1	1223	312	TD_{3-0}	9297	2319
MM_2	25	5	V_0	5372	1311	TD_{3-1}	66	19
MM_3	7	2	V_1	3991	1027	TD_{4-0}	8447	2093
MM_4	515	119	C_0	413	121	TD_{4-1}	916	245
MM_5	418	94	C_1	8950	2217	TD_{5-0}	2142	516
MM_6	667	170	SS_0	350	84	TD_{5-1}	7221	1822
MM_7	230	57	SS_1	9013	2254	TD_{6-0}	2437	588
MM_8	77	18	YT_0	7417	1854	TD_{6-1}	6924	1750
MM_9	4	1	YT_1	1946	484	TD_{7-0}	3123	773
MM_{10}	176	41	MG_0	625	175	TD_{7-1}	6240	1565
MM_{11}	288	57	MG_1	8738	2163	TD_{8-0}	29	7
MM_{12}	105	36	MH_0	8459	2122	TD_{8-1}	9334	2331
MM_{13}	644	166	MH_1	904	216	TD_{9-0}	9178	2284
MM_{14}	36	6	TU_0	758	190	TD_{9-1}	185	54
MM_{15}	143	34	TU_1	8605	2148	TD_{10-0}	9210	2304
MM_{16}	5	1	ID_{1-0}	8801	2192	TD_{10-1}	153	34
MM_{17}	17	3	ID_{1-1}	562	146	TD_{11-0}	6990	1751
MM_{18}	636	164	ID_{2-0}	8341	2056	TD_{11-1}	2373	587
MM_{19}	242	76	ID_{2-1}	1022	282	TD_{12-0}	9187	2286
MM_{20}	31	7	ID_{3-0}	5715	1431	TD_{12-1}	176	52
MM_{21}	264	77	ID_{3-1}	3648	907	TD_{13-0}	4462	1058
MM_{22}	73	21	ID_{4-0}	4605	1129	TD_{13-1}	4901	1280
MM_{23}	162	48	ID_{4-1}	4758	1209	TD_{14-0}	4324	1059
MM_{24}	12	2	ID_{5-0}	9023	2252	TD_{14-1}	5039	1279

Tablo 5.1 Kategorik değişkenlere ait frekans değerleri (devamı)

MM_{25}	1	-	ID_{5-1}	340	86	DD_{1-0}	9054	2256
MM_{26}	17	5	ID_{6-0}	5507	1364	DD_{1-1}	309	82
MM_{27}	31	8	ID_{6-1}	3856	974	DD_{2-0}	3367	790
MM_{28}	125	34	ID_{7-0}	7571	1884	DD_{2-1}	5996	1548
MM_{29}	39	5	ID_{7-1}	1792	454	DD_{3-0}	8816	2169
MM_{30}	19	10	ID_{8-0}	1979	470	DD_{3-1}	547	169
MM_{31}	71	19	ID_{8-1}	7384	1868	DD_{4-0}	1902	438
MM_{32}	42	14	ID_{9-0}	9022	2244	DD_{4-1}	7461	1900
MM_{33}	9	5	ID_{9-1}	341	94	DD_{5-0}	5629	1388
MM_{34}	42	5	ID_{10-0}	5190	1284	DD_{5-1}	3734	950
MM_{35}	23	9	ID_{10-1}	4173	1054	DD_{6-0}	8233	2061
MM_{36}	112	19	ID_{11-0}	9199	2295	DD_{6-1}	1130	277
MM_{37}	32	10	ID_{11-1}	164	43	DD_{7-0}	9181	2273
MM_{38}	74	19	ID_{12-0}	8978	2228	DD_{7-1}	182	65
MM_{39}	58	16	ID_{12-1}	385	110	DD_{8-0}	2302	529
MM_{40}	327	79	ID_{13-0}	8757	2181	DD_{8-1}	7061	1809
MM_{41}	15	-	ID_{13-1}	606	157	DD_{9-0}	6989	1702
MM_{42}	106	25	ID_{14-0}	5726	1411	DD_{9-1}	2374	636
MM_{43}	6	1	ID_{14-1}	3637	927	DD_{10-0}	7906	1945
MM_{44}	8	2	ID_{15-0}	3769	900	DD_{10-1}	1457	393
MM_{45}	16	4	ID_{15-1}	5594	1438	DD_{11-0}	5357	1299
MM_{46}	49	9	ID_{16-0}	6154	1529	DD_{11-1}	4006	1039
MM_{47}	749	191	ID_{16-1}	3209	809	DD_{12-0}	2865	715
MM_{48}	680	160	ID_{17-0}	3066	754	DD_{12-1}	6498	1623
MM_{49}	1182	302	ID_{17-1}	6297	1584	DD_{13-0}	8640	2130
MM_{50}	518	14	ID_{18-0}	1926	447	DD_{13-1}	723	208
MM_{51}	56	19	ID_{18-1}	7437	1891	DD_{14-0}	9335	2331

Tablo 5.1 Kategorik değişkenlere ait frekans değerleri (devamı)

MM_{52}	142	38	TD_{1-0}	469	96	DD_{14-1}	28	7
R_0	8311	2082	TD_{1-1}	8894	2242	DD_{15-0}	5617	1387
R_1	1052	256	TD_{2-0}	2223	534	DD_{15-1}	3746	951
KT_0	8140	2026	TD_{2-1}	7140	1804			

Tablo 5.1 veri setindeki kategorik değişkenlerin frekans değerlerini göstermektedir. Bu tabloya göre eğitim veri setinde ikinci el otomobillere ait marka modellerden en fazla gözlem Volkswagen PASSAT (MM_{49}) en az gözlem ise Mini COOPER (MM_{25})'da test veri setinde ise en fazla gözlem yine Volkswagen PASSAT (MM_{49}) en az gözlem Fiat LINEA (MM_9), Hyundai i10 (MM_{16}) ve Skoda RAPID (MM_{43}) marka modellerinde bulunmaktadır. Gri renkli otomobil sayısı eğitim veri setinde 8311 test veri setinde 2082, kasa tipi sedan olan otomobil sayısı eğitim veri setinde 1223 test veri setinde 312, otomatik vitese sahip otomobil sayısı eğitim veri setinde 3991 test veri setinde 1027, önden ve arkadan çekiş sistemine sahip otomobil sayısı eğitim veri setinde 8950 test veri setinde 2217, 4 silindir sayısına sahip otomobil sayısı eğitim veri setinde 9013 test veri setinde 2254, yakıt türü benzin olan otomobil sayısı eğitim veri setinde 1946 test veri setinde 484, beygir gücü 0-100 hp olan otomobil sayısı eğitim veri setinde 8738 test veri setinde 2163, motor hacmi 0-1400 arasında olan otomobil sayısı eğitim veri setinde 904 test veri setinde 216, turbosu olan otomobil sayısı eğitim veri setinde 8605 test veri setinde 2148, 7 koltuk (ID_1) özelliği olan otomobil sayısı eğitim veri setinde 562 test veri setinde 146, anahtarsız çalıştırma (ID_2) özelliği olan otomobil sayısı eğitim veri setinde 1022 test veri setinde 282, arka kol dayama (ID_3) özelliği olan otomobil sayısı eğitim veri setinde 3648 test veri setinde 907, Bluetooth (ID_4) özelliği olan otomobil sayısı eğitim veri setinde 4758 test veri setinde 1209, deri koltuk (ID_5) özelliği olan otomobil sayısı eğitim veri setinde 340 test veri setinde 86, dijital klima (ID_6) özelliği olan otomobil sayısı eğitim veri setinde 3856 test veri setinde 974, elektrikli sürücü koltuğu (ID_7) olan otomobil sayısı eğitim veri setinde 1792 test veri setinde 454, fonksiyonel direksiyon (ID_8) özelliği olan otomobil sayısı eğitim veri setinde 7384 test veri setinde 1868, hafızalı sürücü koltuğu (ID_9) olan otomobil sayısı eğitim veri setinde 341 test veri setinde 94, hız sabitleyici (ID_{10}) özelliği olan otomobil sayısı eğitim veri setinde 4173 test veri setinde 1054, ısıtmalı

arka koltuklar (ID_{11}) özelliği olan otomobil sayısı eğitim veri setinde 164 test veri setinde 43, ısıtmalı ön koltuklar (ID_{12}) özelliği olan otomobil sayısı eğitim veri setinde 385 test veri setinde 110, navigasyon (ID_{13}) özelliği olan otomobil sayısı eğitim veri setinde 606 test veri setinde 157, otomatik kararan dikiz aynası (ID_{14}) özelliği olan otomobil sayısı eğitim veri setinde 3637 test veri setinde 927, ön kol dayama (ID_{15}) özelliği olan otomobil sayısı eğitim veri setinde 5594 test veri setinde 1438, start/stop (ID_{16}) özelliği olan otomobil sayısı eğitim veri setinde 3209 test veri setinde 809, USB bağlantısı (ID_{17}) özelliği olan otomobil sayısı eğitim veri setinde 6297 test veri setinde 1584, yol bilgisayarı (ID_{18}) özelliği olan otomobil sayısı eğitim veri setinde 7437 test veri setinde 1891, ABS-Anti Blokaj Fren Sistemi (TD_1) özelliği olan otomobil sayısı eğitim veri setinde 8894 test veri setinde 2242, ASR-Elektronik Patinaj Önleme Sistemi (TD_2) özelliği olan otomobil sayısı eğitim veri setinde 7140 test veri setinde 1804, diferansiyel kilidi (TD_3) özelliği olan otomobil sayısı eğitim veri setinde 66 test veri setinde 19, diz hava yastığı (TD_4) özelliği olan otomobil sayısı eğitim veri setinde 916 test veri setinde 245, elektrik takviyeli direksiyon (TD_5) özelliği olan otomobil sayısı eğitim veri setinde 7221 test veri setinde 1822, ESP-Elektronik Stabilizasyon Programı (TD_6) özelliği olan otomobil sayısı eğitim veri setinde 6924 test veri setinde 1750, hıza duyarlı direksiyon (TD_7) özelliği olan otomobil sayısı eğitim veri setinde 6240 test veri setinde 1565, isofix (TD_8) özelliği olan otomobil sayısı eğitim veri setinde 9334 test veri setinde 2331, lastik basınç göstergesi (TD_9) özelliği olan otomobil sayısı eğitim veri setinde 185 test veri setinde 54, LDW-Şeritten Ayrılma İkazı (TD_{10}) özelliği olan otomobil sayısı eğitim veri setinde 153 test veri setinde 34, mesafe takip sistemi (TD_{11}) özelliği olan otomobil sayısı eğitim veri setinde 2373 test veri setinde 587, otomatik park sistemi (TD_{12}) özelliği olan otomobil sayısı eğitim veri setinde 176 test veri setinde 52, perde hava yastığı (TD_{13}) özelliği olan otomobil sayısı eğitim veri setinde 4901 test veri setinde 1280, yokuş kalkış desteği (TD_{14}) özelliği olan otomobil sayısı eğitim veri setinde 5039 test veri setinde 1279, adaptif far (DD_1) özelliği olan otomobil sayısı eğitim veri setinde 309 test veri setinde 82, alaşımlı jant (DD_2) özelliği olan otomobil sayısı eğitim veri setinde 5996 test veri setinde 1548, cam tavan (DD_3) özelliği olan otomobil sayısı eğitim veri setinde 547 test veri setinde 169, elektrikli aynalar (DD_4) özelliği olan otomobil sayısı eğitim veri setinde 7461 test veri setinde 1900, far sensörü (DD_5) özelliği olan otomobil

sayısı eğitim veri setinde 3734 test veri setinde 950, far yıkama (DD_6) özelliği olan otomobil sayısı eğitim veri setinde 1130 test veri setinde 277, geri dönüş kamerası (DD_7) özelliği olan otomobil sayısı eğitim veri setinde 182 test veri setinde 65, ısıtmalı aynalar (DD_8) özelliği olan otomobil sayısı eğitim veri setinde 7061 test veri setinde 1809, katlanır aynalar (DD_9) özelliği olan otomobil sayısı eğitim veri setinde 2374 test veri setinde 636, LED Gündüz sürüş farları (DD_{10}) özelliği olan otomobil sayısı eğitim veri setinde 1457 test veri setinde 393, park sensörü (DD_{11}) özelliği olan otomobil sayısı eğitim veri setinde 4006 test veri setinde 1039, sis farı (DD_{12}) özelliği olan otomobil sayısı eğitim veri setinde 6498 test veri setinde 1623, sunroof (DD_{13}) özelliği olan otomobil sayısı eğitim veri setinde 723 test veri setinde 208, xenon far (DD_{14}) özelliği olan otomobil sayısı eğitim veri setinde 28 test veri setinde 7, yağmur sensörü (DD_{15}) özelliği olan otomobil sayısı ise eğitim veri setinde 3746 test veri setinde 951 şeklindedir.

Tablo 5.2 Sürekli değişkenlere ait tanımlayıcı istatistikler

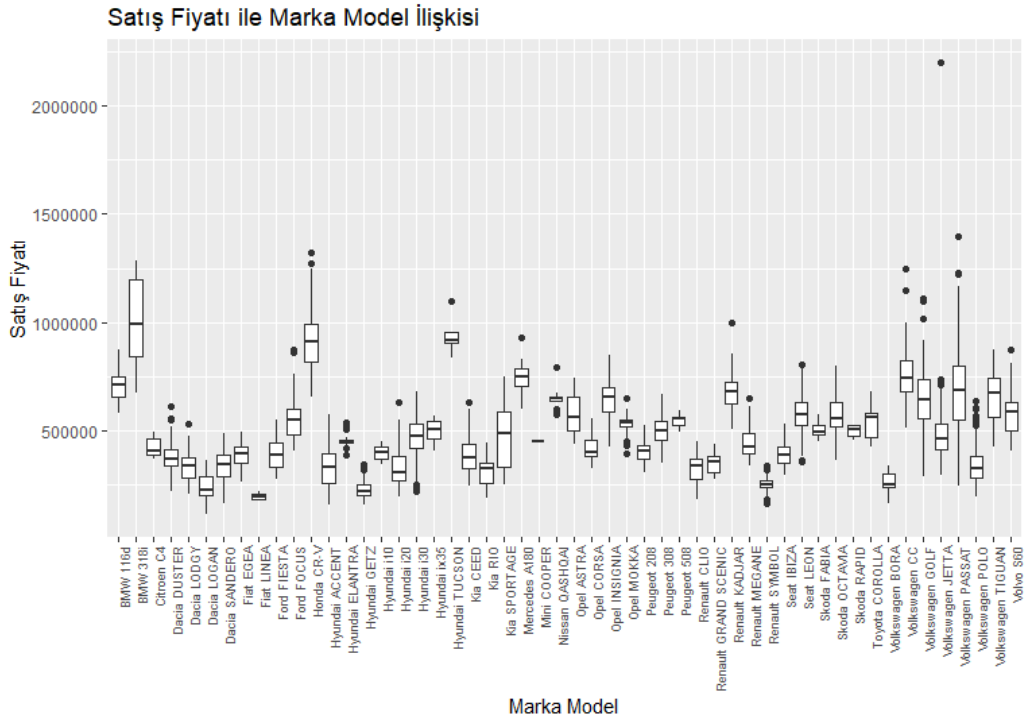
Değişken	**Eğitim**		**Test**	
	Ortalama	**St. Sapma**	**Ortalama**	**St. Sapma**
P	465668	187847,8	467278	187079,7
KM	161178	85875,8	159849	82963,08
W	1304	143,702	1310	144,054
U	4366	264,017	4369	260,471
G	1771	52,253	1772	51,565
Y	1513	80,555	1514	81,385
BK	426,9	140,201	427	138,621
T	226,2	45,544	227,5	46,124
AH	183,9	15,733	184,1	15,579
H	11,96	1,789	11,94	1,779
A	8,074	2,624	8,036	2,540
OYT	4,668	0,812	4,695	0,842

Tablo 5.2 sürekli değişkenlerin tanımlayıcı istatistiklerini göstermektedir. Bu tabloya göre otomobilin ortalama fiyatı eğitim veri setinde 465668 TL test veri setinde 467278 TL, ortalama kilometre eğitim veri setinde 161178 km test veri setinde 159849 km, ortalama ağırlık eğitim veri setinde 1304 kg test veri setinde 1310 kg, ortalama uzunluk eğitim veri setinde 4366 mm test veri setinde 4369 mm, ortalama genişlik eğitim veri setinde 1771 mm test veri setinde 1772 mm, ortalama yükseklik eğitim veri setinde 1513 mm test veri setinde 1514 mm, ortalama bagaj kapasitesi eğitim veri setinde 426,9 lt test veri setinde 427 lt, ortalama tork (motor tarafından üretilen dönme kuvveti) eğitim veri setinde 226,2 Nm (Newton metre) test veri setinde 227,5 Nm (Newton metre), ortalama azami hız eğitim veri setinde 183,9 km/saat test veri setinde 184,1 km/saat, ortalama 100 km'ye çıkma (hızlanma, ivme) eğitim veri setinde 11,96 sn test veri setinde 11,94 sn, ortalama yaş eğitim veri setinde ve test veri setinde 8, 100 km'de ortalama yakıt tüketimi ise eğitim veri setinde 4,668 lt test veri setinde 4,695 lt'dir.

5.3.2 İlişkisel Bulgular

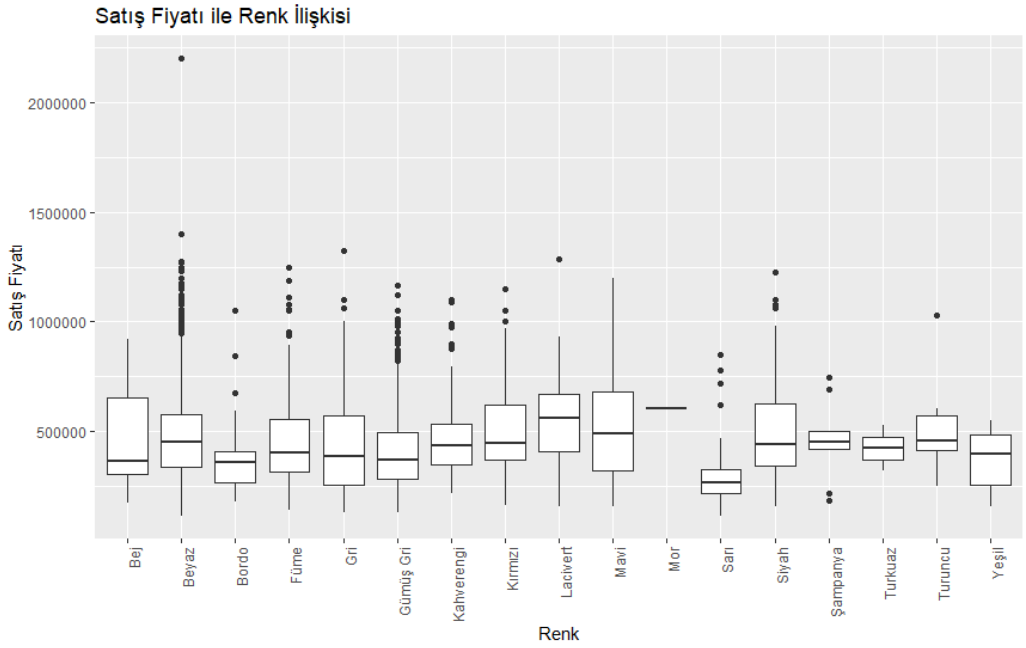
Bu bölümde, eğitim ve test veri setlerinde ikinci el otomobil fiyatlarını etkileyebileceği düşünülen bağımsız değişkenler arasından kategorik değişkenler ile otomobil fiyatı arasındaki ilişkiyi analiz etmek için kullanılan Box Plot (Kutu Grafiği) ve sürekli değişkenler ile otomobil fiyatı arasındaki ilişkiyi analiz etmek için kullanılan Scatter Plot (Saçılım Grafiği) sonuçlarına yer verilmiştir.

5.3.2.1 Eğitim Veri Setine Ait İlişkisel Bulgular



Şekil 5.1 Satış fiyatı ile marka model değişkeni arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.1'e göre farklı marka ve modellerin medyanlarının ve satış fiyatı dağılımlarının birbirinden farklı olduğu bu nedenle marka ve modellerin satış fiyatları üzerinde etkisi olduğu söylenebilir. Volkswagen marka otomobillerin diğer marka otomobillere göre daha yüksek satış fiyatlarına sahip olduğu söylenebilir. Benzer şekilde Dacia, Hyundai gibi marka otomobillerin ise daha düşük satış fiyatlarına sahip olduğu söylenebilir.



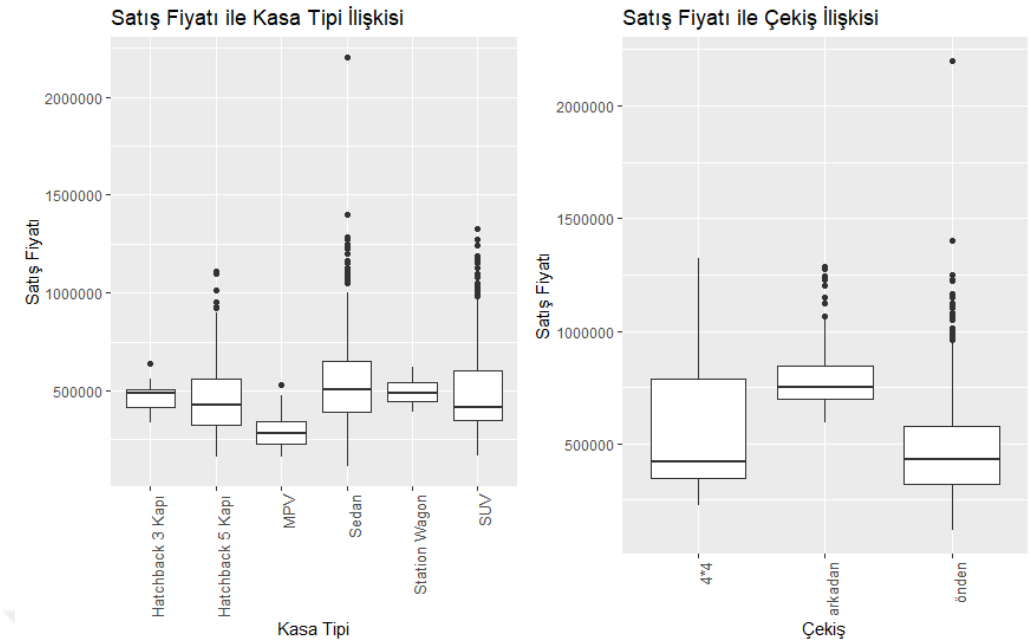
Şekil 5.2 Satış fiyatı ile renk değişkeni arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.2'e göre farklı otomobil renklerinin medyanlarının ve satış fiyatı dağılımlarının birbirinden farklı olduğu bu nedenle renk değişkeninin satış fiyatları üzerinde etkisi olduğu söylenebilir. Satış fiyatı en yüksek olan otomobillerin lacivert, turuncu ve mavi renk otomobiller olduğu benzer şekilde satış fiyatı en düşük olan otomobillerin sarı renk otomobiller olduğu söylenebilir.



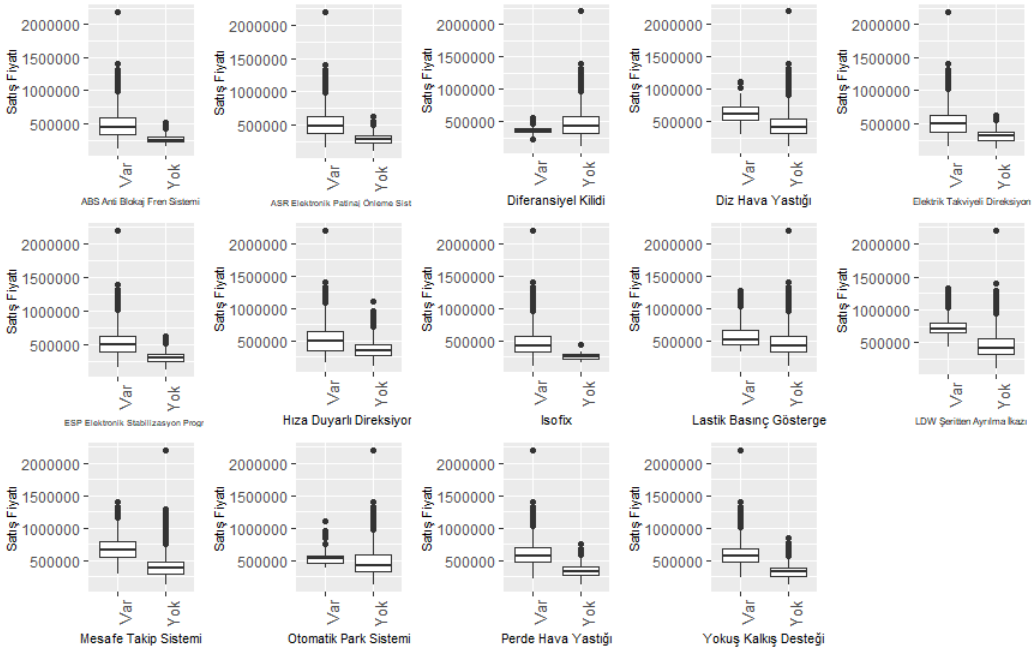
Şekil 5.3 Satış fiyatı ile vites, silindir sayısı, yakıt türü ve turbo değişkenleri arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.3'e göre vites, silindir sayısı, yakıt türü ve turbo değişkenlerinin satış fiyatları üzerinde etkisi olduğu söylenebilir. Otomatik vitese sahip otomobillerin satış fiyatının manuel vitese sahip otomobillerinkine göre; silindir sayısı 4 olan otomobillerin satış fiyatının silindir sayısı 3 olan otomobillere göre; yakıt türü benzin olan otomobillerin satış fiyatının yakıt türü dizel olan otomobillere göre ve turbosu olan otomobillerin satış fiyatının turbosu olmayanlara göre daha yüksek olduğu söylenebilir.



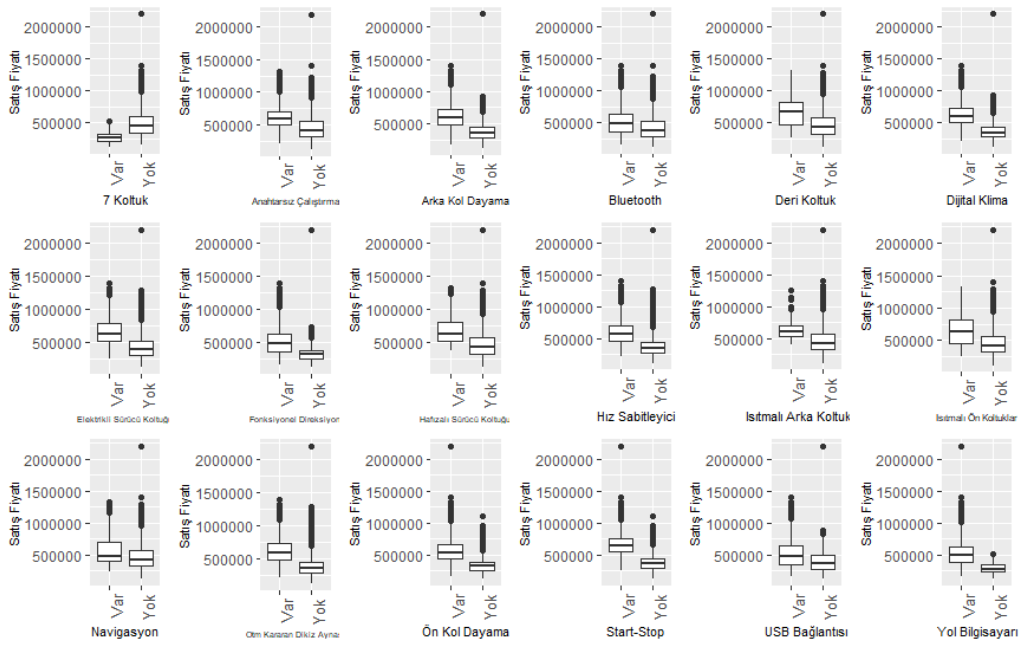
Şekil 5.4 Satış fiyatı ile kasa tipi ve çekiş değişkenleri arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.4'e göre kasa tipi ve çekiş değişkeninin satış fiyatları üzerindeki etkisi olduğu söylenebilir. Kasa tipi sedan olan otomobillerin diğer kasa tiplerine sahip otomobillere göre daha yüksek satış fiyatına sahip olduğu ve arkadan çekişli otomobillerin 4*4 ve önden çekişli otomobillere göre daha yüksek satış fiyatına sahip olduğu söylenebilir.



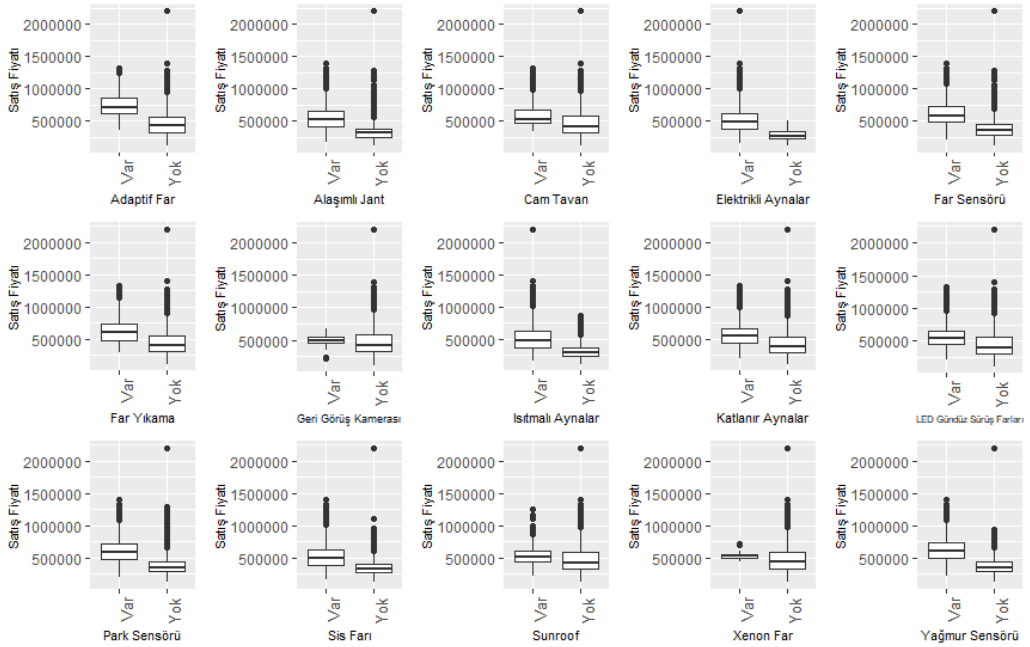
Şekil 5.5 Satış fiyatı ile teknik donanımlar arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.5'e göre teknik donanımların satış fiyatları üzerinde etkisi olduğu söylenebilir. Teknik donanımlar arasında ABS (Anti Blokaj Fren Sistemi), ASR (Elektronik Patinaj Önleme Sistemi), diz hava yastığı, elektrik takviyeli direksiyon, ESP (Elektronik Stabilizasyon Programı), hıza duyarlı direksiyon, Isofix, lastik basınç göstergesi, LDW (Şeritten Ayrılma İkazı), mesafe takip sistemi, otomatik park sistemi, hava yastığı ve yokuş kalkış desteği olan otomobillerin satış fiyatının daha yüksek olduğu söylenebilir.



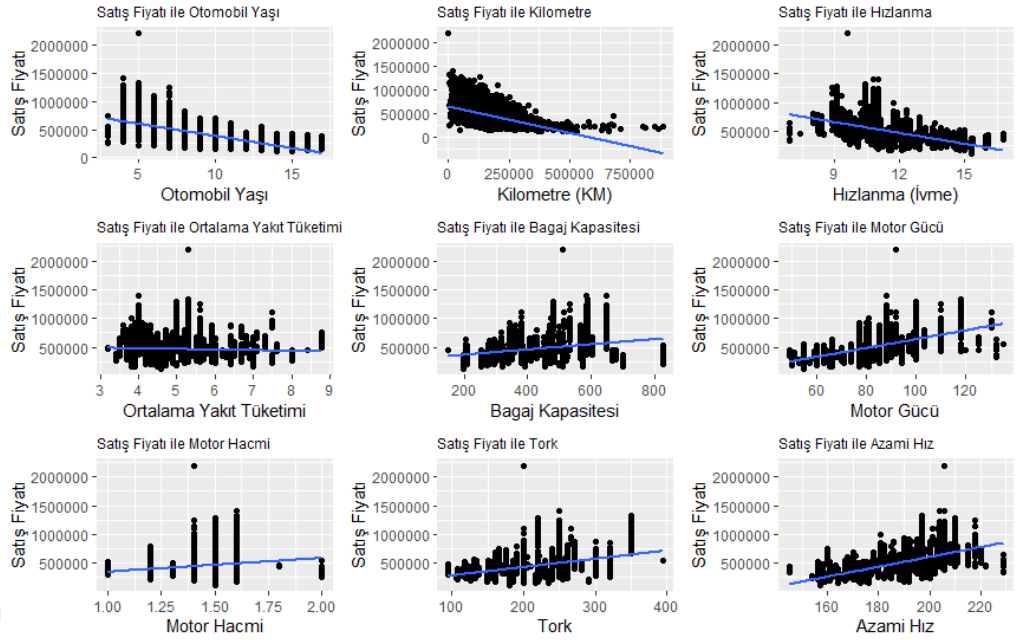
Şekil 5.6 Satış fiyatı ile iç donanımlar arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.6'a göre iç donanımların satış fiyatları üzerinde etkisi olduğu söylenebilir. İç donanımlar arasında anahtarsız çalıştırma, arka kol dayama, Bluetooth, deri koltuk, dijital klima, elektrikli sürücü koltuğu, fonksiyonel direksiyon, hafızalı sürücü koltuğu, hız sabitleyici, ısıtmalı arka ve ön koltuk, navigasyon, otomatik kararan dikiz aynası, ön kol dayama, start-stop, usb bağlantısı ve yol bilgisayarı olan otomobillerin satış fiyatının daha yüksek olduğu söylenebilir.



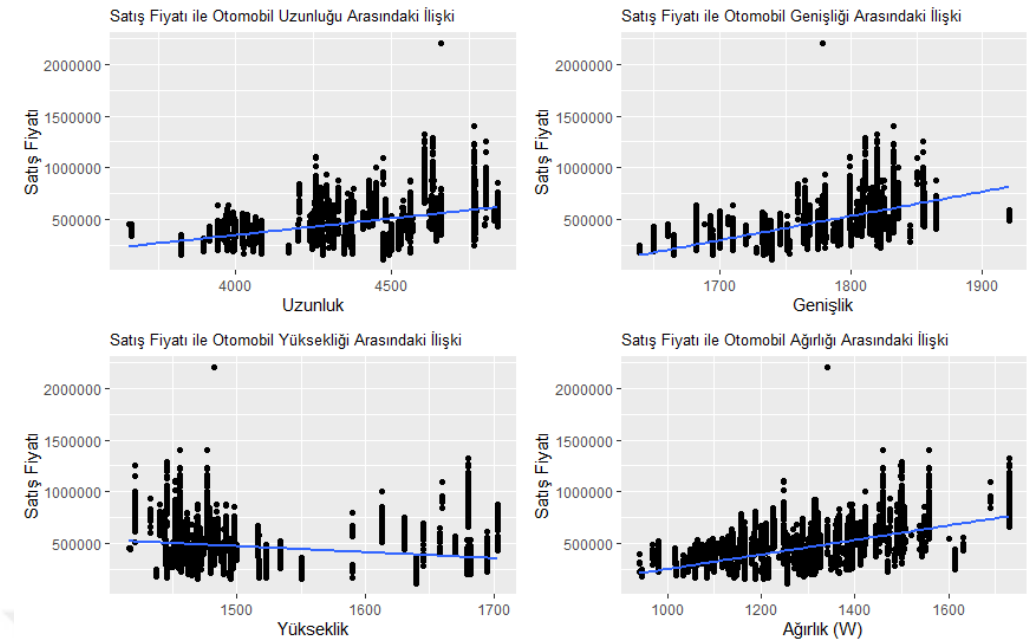
Şekil 5.7 Satış fiyatı ile dış donanımlar arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.7'e göre dış donanımların satış fiyatları üzerinde etkisi olduğu söylenebilir. Dış donanımlar arasında adaptif far, alaşımlı jant, cam tavan, elektrikli aynalar, far sensörü, far yıkama, geri görüş kamerası, ısıtmalı aynalar, katlanır aynalar, LED gündüz sürüş farları, park sensörü, sis farı, sunroof, xenon far ve yağmur sensörü olan otomobillerin satış fiyatının daha yüksek olduğu söylenebilir.



Şekil 5.8 Satış fiyatı ile otomobil yaşı, kilometre, hızlanma, ortalama yakıt tüketimi, bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği

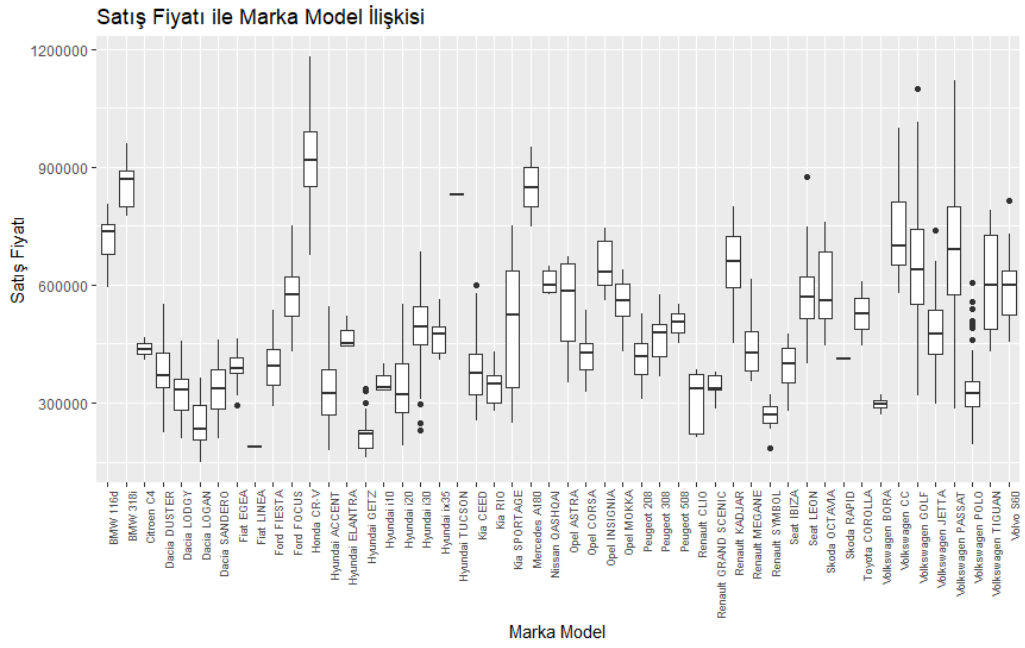
Şekil 5.8'e göre otomobil yaşı, kilometresi, hızlanması(ivme) ve ortalama yakıt tüketimi arttıkça otomobil fiyatının düşme eğiliminde olduğu söylenebilir. Benzer şekilde bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız arttıkça otomobil fiyatının artma eğiliminde olduğu söylenebilir. Böylece kilometre, hızlanma, ortalama yakıt tüketimi, bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız değişkenlerinin satış fiyatları üzerinde etkisi olduğu söylenebilir.



Şekil 5.9 Satış fiyatı ile uzunluk, genişlik, yükseklik ve ağırlık değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği

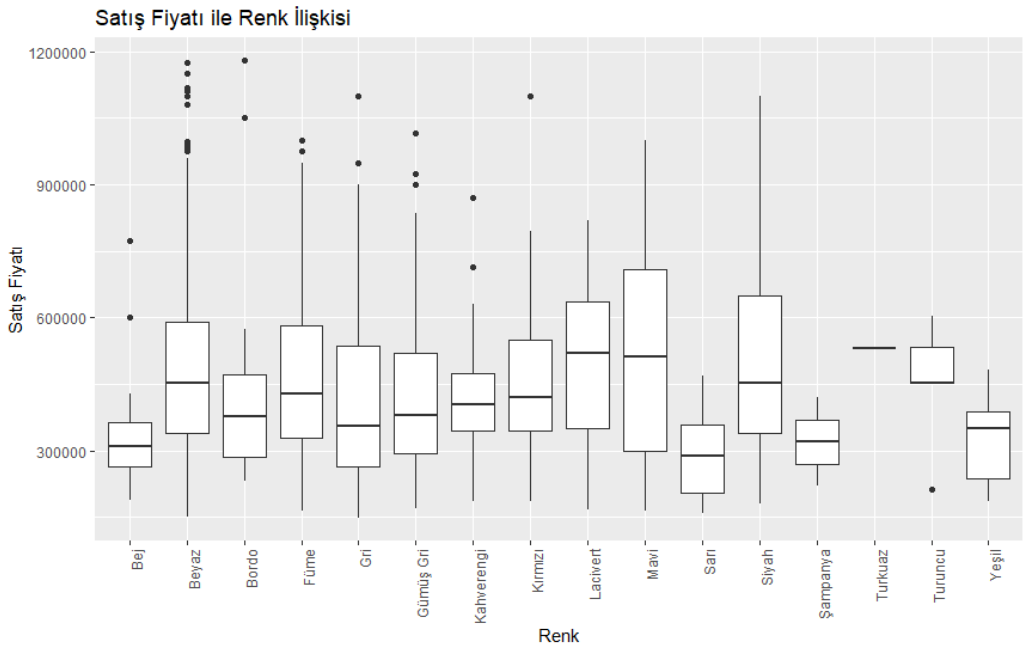
Şekil 5.9'a göre otomobil yüksekliği arttıkça otomobil fiyatının düşme eğiliminde olduğu; otomobil uzunluğu, genişliği ve ağırlığı arttıkça ise otomobil fiyatının artma eğiliminde olduğu söylenebilir. Böylece uzunluk, genişlik, yükseklik ve ağırlık değişkenlerinin de satış fiyatları üzerinde etkisi olduğu söylenebilir.

5.3.2.2 Test Veri Setine Ait İlişkisel Bulgular



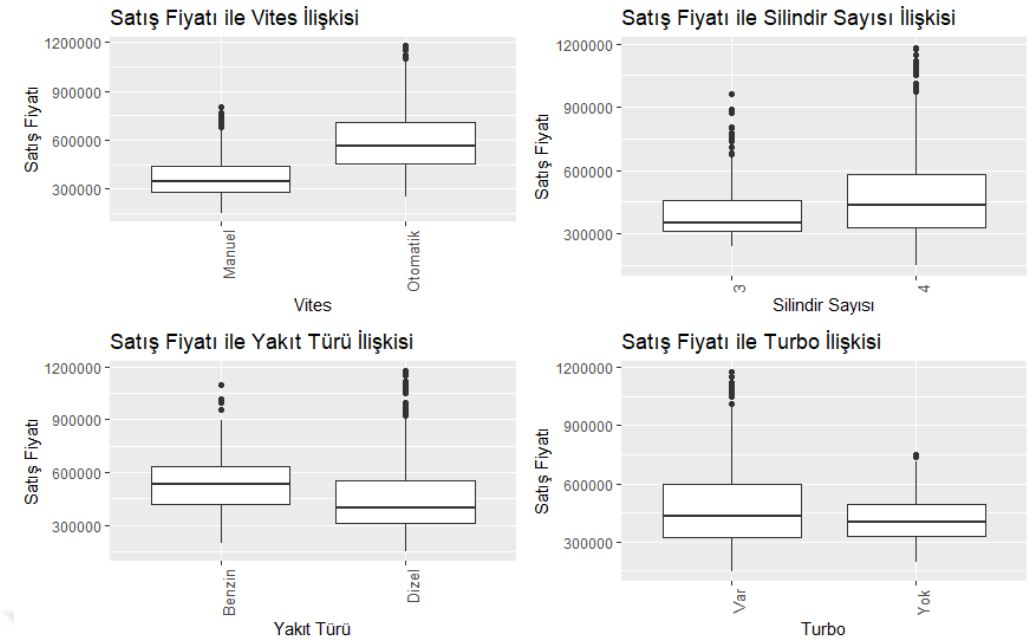
Şekil 5.10 Test veri setindeki satış fiyatı ile marka model değişkeni arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.10'a göre farklı marka ve modellerin medyanlarının ve satış fiyatı dağılımlarının birbirinden farklı olduğu bu nedenle marka ve modellerin satış fiyatları üzerinde etkisi olduğu söylenebilir. BMW, Honda ve Volkswagen marka otomobillerin diğer marka otomobillere göre daha yüksek satış fiyatlarına sahip olduğu söylenebilir. Benzer şekilde Dacia, Hyundai gibi marka otomobillerin ise daha düşük satış fiyatlarına sahip olduğu söylenebilir.



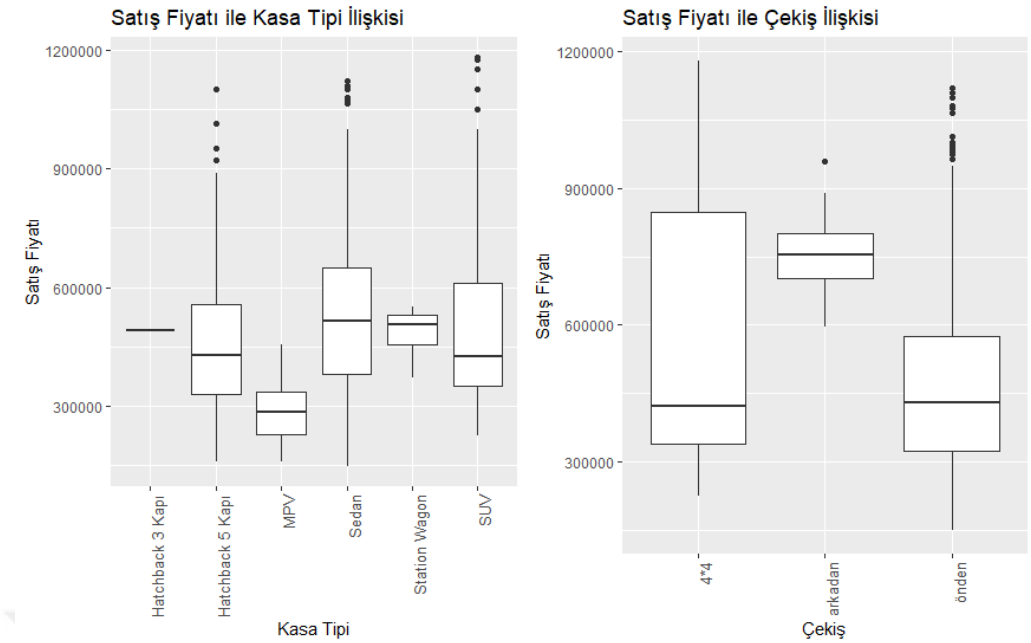
Şekil 5.11 Test veri setindeki satış fiyatı ile renk değişkeni arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.11'e göre farklı otomobil renklerinin medyanlarının ve satış fiyatı dağılımlarının birbirinden farklı olduğu bu nedenle renk değişkeninin satış fiyatları üzerinde etkisi olduğu söylenebilir. Satış fiyatı en yüksek olan otomobillerin lacivert, turuncu ve mavi renk otomobiller olduğu benzer şekilde satış fiyatı en düşük olan otomobillerin sarı renk otomobiller olduğu söylenebilir.



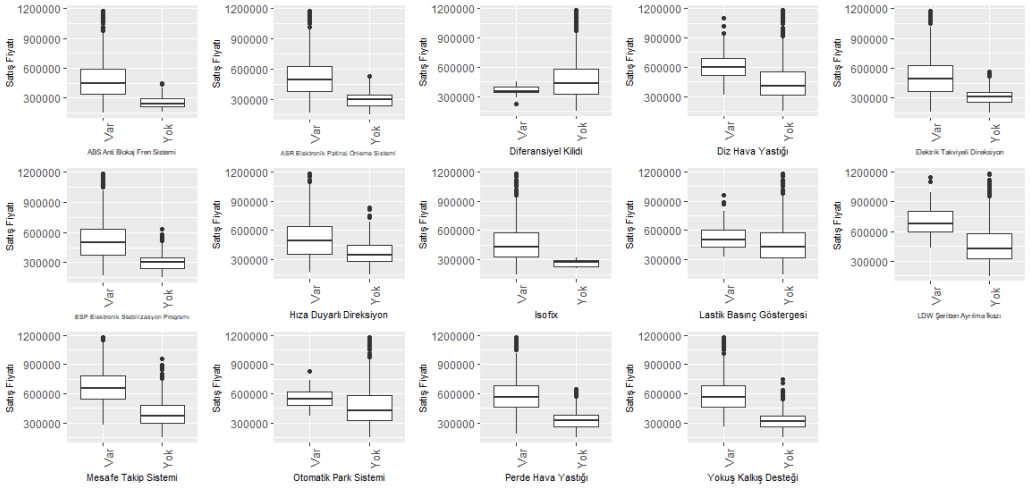
Şekil 5.12 Test veri setindeki satış fiyatı ile vites, silindir sayısı, yakıt türü ve turbo değişkenleri arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.12'e göre vites, silindir sayısı, yakıt türü ve turbo değişkenlerinin satış fiyatları üzerinde etkisi olduğu söylenebilir. Otomatik vitese sahip otomobillerin satış fiyatının manuel vitese sahip otomobillerinkine göre; silindir sayısı 4 olan otomobillerin satış fiyatının silindir sayısı 3 olan otomobillere göre; yakıt türü benzin olan otomobillerin dizel satış fiyatının yakıt türü dizel olan otomobillere göre ve turbosu olan otomobillerin satış fiyatının turbosu olmayanlara göre daha yüksek olduğu söylenebilir.



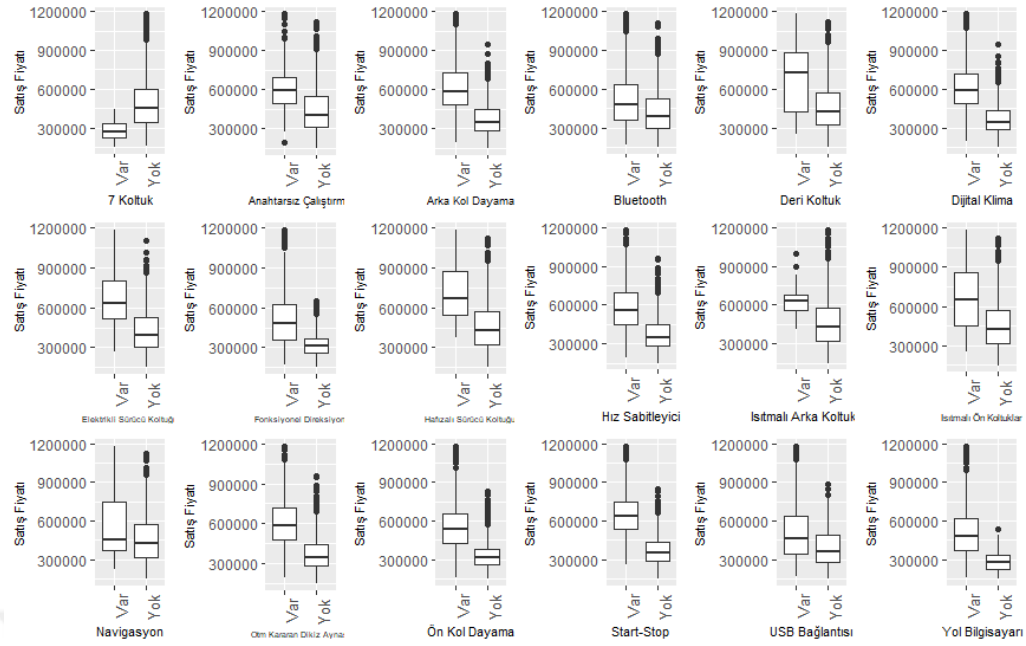
Şekil 5.13 Test veri setindeki satış fiyatı ile kasa tipi ve çekiş değişkeni arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.13'e göre kasa tipi ve çekiş değişkeninin satış fiyatları üzerinde etkisi olduğu söylenebilir. Kasa tipi sedan olan otomobillerin diğer kasa tiplerine sahip otomobillere göre daha yüksek satış fiyatına sahip olduğu ve arkadan çekişli otomobillerin 4*4 ve önden çekişli otomobillere göre daha yüksek satış fiyatına sahip olduğu söylenebilir.



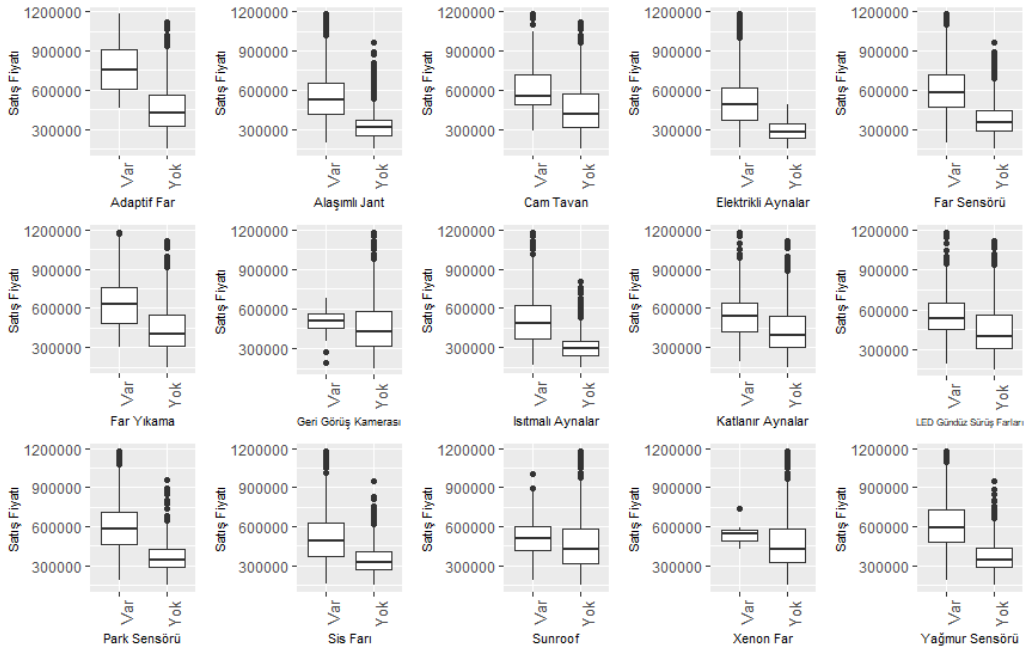
Şekil 5.14 Test veri setindeki satış fiyatı ile teknik donanımlar arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.14'e göre teknik donanımların satış fiyatları üzerinde etkisi olduğu söylenebilir. Teknik donanımlar arasında ABS (Anti Blokaj Fren Sistemi), ASR (Elektronik Patinaj Önleme Sistemi), diz hava yastığı, elektrik takviyeli direksiyon, ESP (Elektronik Stabilizasyon Programı), hıza duyarlı direksiyon, Isofix, lastik basınç göstergesi, LDW (Şeritten Ayrılma İkazı), mesafe takip sistemi, otomatik park sistemi, hava yastığı ve yokuş kalkış desteği olan otomobillerin satış fiyatının daha yüksek olduğu söylenebilir.



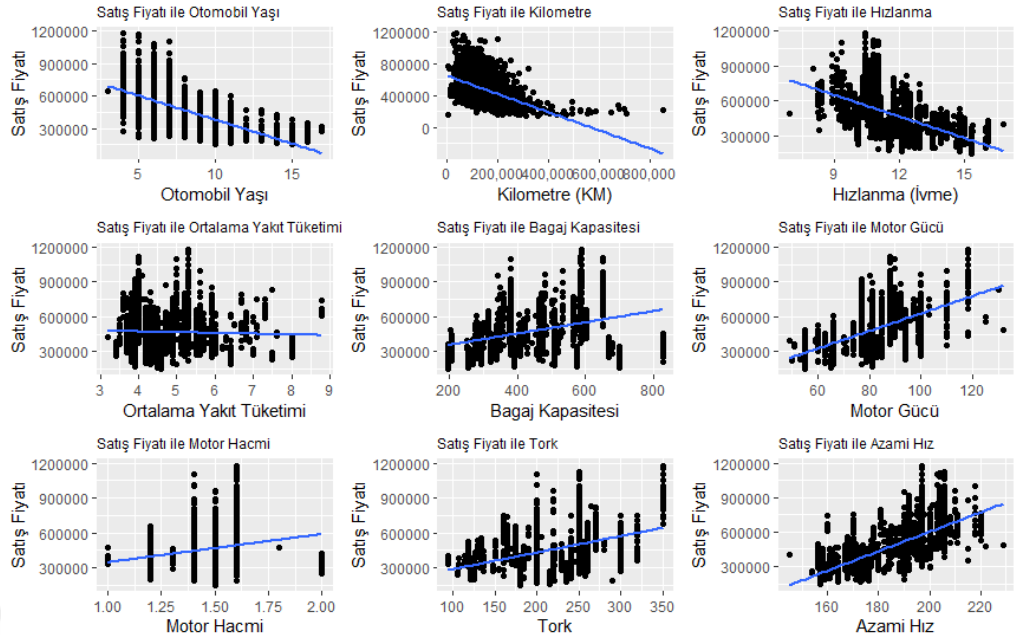
Şekil 5.15 Test veri setindeki satış fiyatı ile iç donanımlar arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.15'e göre iç donanımların satış fiyatları üzerinde etkisi olduğu söylenebilir. İç donanımlar arasında anahtarsız çalıştırma, arka kol dayama, Bluetooth, deri koltuk, dijital klima, elektrikli sürücü koltuğu, fonksiyonel direksiyon, hafızalı sürücü koltuğu, hız sabitleyici, ısıtmalı arka ve ön koltuk, navigasyon, otomatik kararan dikiz aynası, ön kol dayama, start-stop, usb bağlantısı ve yol bilgisayarı olan otomobillerin satış fiyatının daha yüksek olduğu söylenebilir.



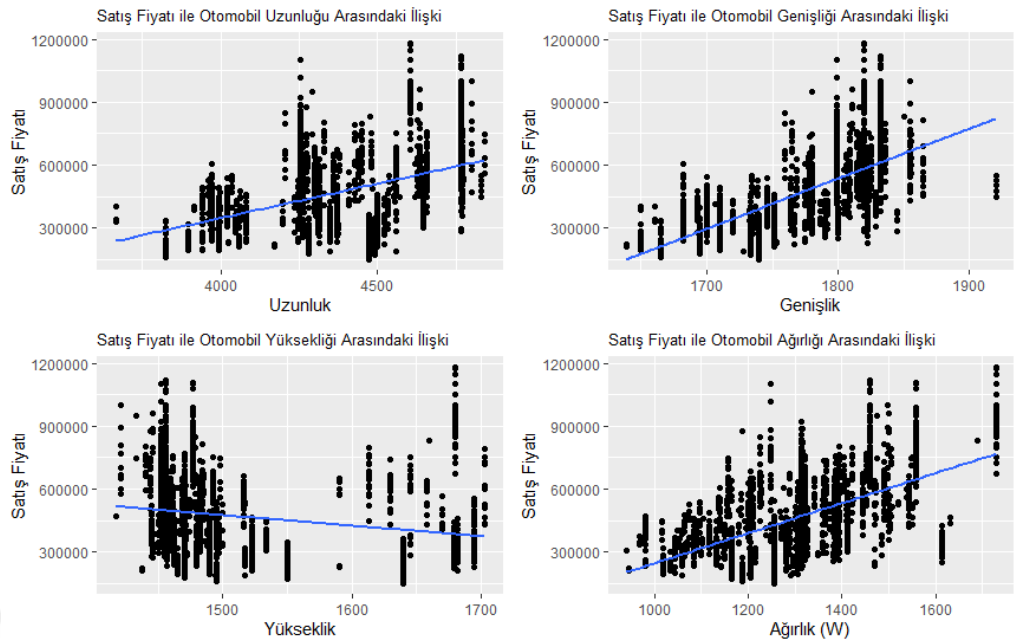
Şekil 5.16 Test veri setindeki satış fiyatı ile dış donanımlar arasındaki ilişkiyi gösteren kutu grafiği

Şekil 5.16'a göre dış donanımların satış fiyatları üzerinde etkisi olduğu söylenebilir. Dış donanımlar arasında adaptif far, alaşımlı jant, cam tavan, elektrikli aynalar, far sensörü, far yıkama, geri görüş kamerası, ısıtmalı aynalar, katlanır aynalar, LED gündüz sürüş farları, park sensörü, sis farı, sunroof, xenon far ve yağmur sensörü olan otomobillerin satış fiyatının daha yüksek olduğu söylenebilir.



Şekil 5.17 Test veri setindeki satış fiyatı ile otomobil yaşı, kilometre, hızlanma, ortalama yakıt tüketimi, bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği

Şekil 5.17'e göre otomobil yaşı, kilometresi, hızlanması (ivme) ve ortalama yakıt tüketimi arttıkça otomobil fiyatının düşme eğiliminde olduğu söylenebilir. Benzer şekilde bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız arttıkça otomobil fiyatının artma eğiliminde olduğu söylenebilir. Böylece kilometre, hızlanma, ortalama yakıt tüketimi, bagaj kapasitesi, motor gücü, motor hacmi, tork ve azami hız değişkenlerinin satış fiyatları üzerinde etkisi olduğu söylenebilir.



Şekil 5.18 Test veri setindeki satış fiyatı ile uzunluk, genişlik, yükseklik ve ağırlık değişkenleri arasındaki ilişkiyi gösteren saçılım grafiği

Şekil 5.18'e göre otomobil yüksekliği arttıkça otomobil fiyatının düşme eğiliminde olduğu; otomobil uzunluğu, genişliği ve ağırlığı arttıkça ise otomobil fiyatının artma eğiliminde olduğu söylenebilir. Böylece uzunluk, genişlik, yükseklik ve ağırlık değişkenlerinin de satış fiyatları üzerinde etkisi olduğu söylenebilir.

5.3.3 Regresyon Analiz Sonuçları

Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları

	Doğrusal Model			**Yarı Logaritmik Model**	
Değ.	**Katsayı**	**St. Hata**		**Katsayı**	**St. Hata**
Sabit	410800	296900		12,897***	4,047
MM_2	280800****	26420		0,312****	0,047
MM_3	-317900****	31160		-0,415****	0,056
MM_4	-165000****	33040		-0,319****	0,057
MM_5	-293400****	32760		-0,410****	0,057

Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları (devamı)

MM_6	-231700****	29440		-0,499****	0,051
MM_7	-226600****	29100		-0,465****	0,051
MM_8	-335800****	26270		-0,578****	0,049
MM_9	-185500****	40240		-0,652****	0,072
MM_{10}	-294100****	28520		-0,464****	0,051
MM_{11}	-204900****	24480		-0,273****	0,044
MM_{12}	71890**	36090		-0,087	0,065
MM_{13}	-249100****	24980		-0,394****	0,045
MM_{14}	-298100****	28000		-0,391****	0,050
MM_{15}	-250000****	25890		-0,490****	0,046
MM_{16}	-19280	53900		-0,361****	0,096
MM_{17}	-290200****	32780		-0,503****	0,060
MM_{18}	-317100****	26450		-0,482****	0,047
MM_{19}	-270200****	23730		-0,416****	0,043
MM_{20}	-143300***	47450		-0,272***	0,085
MM_{21}	-273300****	22940		-0,424****	0,041
MM_{22}	-235400****	25860		-0,435****	0,046
MM_{23}	-139000****	37290		-0,323****	0,066
MM_{24}	19330	34440		0,025	0,062
MM_{25}	-119700*	69470		0,035	0,125
MM_{26}	-23020	36870		-0,071	0,065
MM_{27}	-239100****	25590		-0,375****	0,046

Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları (devamı)

MM_{28}	-147900****	27280		-0,216****	0,048
MM_{29}	-139200****	37670		-0,116*	0,067
MM_{30}	-61840*	35050		-0,117*	0,063
MM_{31}	-340300****	28060		-0,482****	0,051
MM_{32}	-324800****	26790		-0,400****	0,048
MM_{33}	-309400****	35300		-0,355****	0,063
MM_{34}	-285800****	28740		-0,444****	0,051
MM_{35}	-202600****	30340		-0,319****	0,054
MM_{36}	-180400****	33020		-0,250****	0,059
MM_{37}	-323500****	30360		-0,394****	0,054
MM_{38}	-259100****	29170		-0,520****	0,051
MM_{39}	-289900****	30650		-0,454****	0,056
MM_{40}	-188500****	23210		-0,216****	0,041
MM_{41}	-308700****	32840		-0,444****	0,059
MM_{42}	-232200****	24640		-0,264****	0,044
MM_{43}	-248400****	35260		-0,347****	0,063
MM_{44}	-202000****	35310		-0,334****	0,063
MM_{45}	-171900****	36690		-0,273****	0,067
MM_{46}	-96940**	40010		-0,045	0,071
MM_{47}	-92760****	24550		-0,111**	0,044
MM_{48}	-174000****	23650		-0,184****	0,042
MM_{49}	-90070****	24950		-0,036	0,044

Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları (devamı)

MM_{50}	-152300****	27540		-0,215****	0,050
MM_{51}	-17410	30210		-0,101*	0,053
MM_{52}	-151400****	28480		-0,232****	0,052
A	-20250****	450,9		-0,324****	0,006
V	68810****	2445		0,108****	0,004
KM	-0,4781****	0,009		-0,186****	0,003
R	-365,8	1986		-0,018****	0,004
MG	6360	6532		0,005	0,011
MH	-38190****	4779		-0,052****	0,009
H	-3568	2728		-0,155**	0,063
OYT	-19950****	3239		-0,070**	0,031
C	-42550****	7982		-0,057****	0,015
W	22,03	34,07		0,057	0,077
SS	17680**	8236		0,001	0,014
YT	27910***	10610		-0,010	0,020
T	-252,8**	105,4		-0,064	0,045
U	-47,64	33,36		-0,007	0,307
G	464,5****	119,8		0,634	0,389
Y	-80,87	76,59		-0,257	0,209
BK	164,8****	40,03		0,042	0,041
KT	19220	13440		-0,091****	0,024
TU	8947	6293		0,019	0,015

Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları (devamı)

AH	892**	433,4		0,093	0,153
ID_1	88420****	22450		0,033	0,055
ID_2	5555	8398		0,013****	0,015
ID_3	37470****	5406		0,044****	0,010
ID_4	-19110****	3965		-0,024	0,007
ID_5	8013	19350		-0,038****	0,034
ID_6	18020****	5116		0,052	0,009
ID_7	4760	8202		-0,003**	0,015
ID_8	-37140****	4873		-0,020	0,009
ID_9	-2743	15360		0,018	0,028
ID_{10}	427,9	3586		-0,009	0,006
ID_{11}	7921	29610		0,003	0,054
ID_{12}	-33340**	16640		-0,017	0,030
ID_{13}	-5818	5674		0,028***	0,010
ID_{14}	6920	8863		0,030**	0,016
ID_{15}	-5148	4700		0,024**	0,008
ID_{16}	5781****	6058		0,014	0,011
ID_{17}	59070****	4744		0,060****	0,009
ID_{18}	22350****	6301		0,066****	0,011
TD_1	-60340**	10250		-0,012	0,018
TD_2	-21360	10170		-0,005	0,019
TD_3	2337****	12070		-0,021	0,022

Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları (devamı)

TD_4	-69530****	14520		-0,077***	0,026
TD_5	40450	9004		-0,014	0,016
TD_6	12390	11200		-0,0180	0,020
TD_7	-2015****	7015		-0,023*	0,013
TD_8	-80460****	24230		0,028	0,044
TD_9	114100****	12440		-0,171****	0,022
TD_{10}	46900	10140		0,024	0,018
TD_{11}	2631	7253		0,021	0,013
TD_{12}	-11380	8316		0,004	0,015
TD_{13}	6094***	7595		0,030**	0,014
TD_{14}	-23360****	7731		0,013	0,014
DD_1	-47660***	8530		-0,097****	0,015
DD_2	-15770***	5251		0,0001	0,009
DD_3	26060****	9559		0,074****	0,017
DD_4	-46190***	7815		0,004	0,014
DD_5	23660***	9120		-0,003	0,016
DD_6	28690****	6012		0,032***	0,011
DD_7	45840***	17200		0,064**	0,031
DD_8	21700***	6600		-0,026**	0,012
DD_9	14380****	4356		0,016**	0,008

Tablo 5.3 Eğitim veri setinin regresyon analizi sonuçları (devamı)

DD_{10}	18050****	5088		0,033****	0,009
DD_{11}	13620***	4644		0,014*	0,008
DD_{12}	14960***	4573		-0,019**	0,008
DD_{13}	7836	6390		-0,006	0,011
DD_{14}	-56070**	28000		0,030	0,050
DD_{15}	-43410****	9466		-0,027	0,017
N		9363	N		9363
R^2		0,9032	R^2		0,9321
Düz. R^2		0,902	Düz. R^2		0,9312
F İstatistiği (Prob.)		731 (0,000)	F İstatistiği (Prob.)		1076 (0,000)

****, ***, ** ve * sırasıyla %0,1, %1, %5 ve %10 önem düzeyinde istatistiki açıdan anlamlılığı göstermektedir. Kategorik değişkenlerin ilk kategorisi temel sınıf olarak alınmıştır.

Tablo 5.3'te eğitim veri setinde tahmin edilen doğrusal regresyon ve yarı logaritmik regresyon modelinde istatistiki açıdan bir bütün olarak anlamlı olduğu görülmektedir (Prob. (F-İstatistiği) = 0.000).

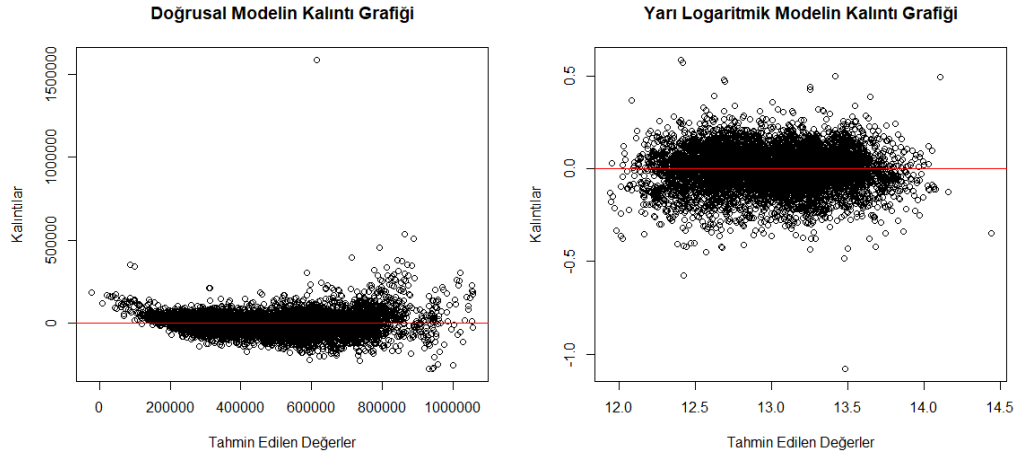
Tablo 5.3'e göre ikinci el otomobilin marka modelinin BMW 318i (MM_2) olması BMW 116d (MM_1)'e göre fiyatı %31,2 oranında artırdığı söylenebilir. Vitesin otomatik olması %10,8, anahtarsız çalıştırma (ID_2) özelliğinin olması %1,3, arka kol dayama (ID_3) özelliğinin olması %4,4, navigasyon (ID_{13}) olması %2,8, otomatik kararan dikiz aynası (ID_{14}) olması %3, usb bağlantısı (ID_{17}) olması %6, yol bilgisayarı (ID_{18}) olması %6,6, perde hava yastığı (TD_{13}) olması %3, cam tavan (DD_3) olması %7,4, far yıkama özelliğinin (DD_6) olması %3,2, geri görüş kamerası (DD_7) olması %6,4, katlanır aynalar (DD_9) olması %1,6, led gündüz sürüş farları (DD_{10}) olması %3,3, park sensörü (DD_{11}) olması %1,4 oranında fiyatı arttırmaktadır.

Ayrıca, ikinci el otomobilin marka modelinin Citroen C4 (MM_3) olması %41,5, Dacia DUSTER (MM_4) olması %31,9, Dacia LODGY (MM_5) olması %41, Dacia LOGAN (MM_6) olması %49,9, Dacia SANDERO (MM_7) olması %46,5, Fiat EGEA (MM_8) olması %57,8, Fiat LINEA (MM_9) olması %65,2, Ford FIESTA (MM_{10}) olması %46,4, Ford FOCUS (MM_{11}) olması %27,3, Hyundai ACCENT (MM_{13}) olması %39,4, Hyundai ELANTRA (MM_{14}) olması %39,1, Hyundai GETZ (MM_{15}) olması %49, Hyundai TUCSON (MM_{16}) olması %36,1, Hyundai i10 (MM_{17}) olması %50,3, Hyundai i20 (MM_{18}) olması %48,2, Hyundai i30 (MM_{19}) olması %41,6, Hyundai ix35 (MM_{20}) olması %27,2, Kia CEED (MM_{21}) olması %42,4, Kia RIO (MM_{22}) olması %43,5, Kia SPORTAGE (MM_{23}) olması %32,3, Opel ASTRA (MM_{27}) olması %37,5, Opel CORSA (MM_{28}) olması %21,6, Opel INSIGNIA (MM_{29}) olması %11,6, Opel MOKKA (MM_{30}) olması %11,7, Peugeot 208 (MM_{31}) olması %48,2, Peugeot 308 (MM_{32}) olması %40, Peugeot 508 (MM_{33}) olması %35,5, Renault CLIO (MM_{34}) olması %44,4, Renault GRAND SCENIC (MM_{35}) olması %31,9, Renault KADJAR (MM_{36}) olması %25, Renault MEGANE (MM_{37}) olması %39,4, Renault SYMBOL (MM_{38}) olması %52, Seat IBIZA (MM_{39}) olması %45,4, Seat LEON (MM_{40}) olması %21,6, Skoda FABIA (MM_{41}) olması %44,4, Skoda OCTAVIA (MM_{42}) olması %26,4, Skoda RAPID (MM_{43}) olması %34,7, Toyota COROLLA (MM_{44}) olması %33,4, Volkswagen BORA (MM_{45}) olması %27,3, Volkswagen GOLF (MM_{47}) olması %11,1, Volkswagen JETTA (MM_{48}) olması %18,4, Volkswagen POLO (MM_{50}) olması %21,5, Volkswagen TIGUAN (MM_{51}) olması %10,1, Volvo S60 (MM_{52}) olması %23,2, otomobil yaşının 1 yıl artması %32,4, kilometresinin 1 km artması %18,6 , otomobil renginin gri olması %1,8, motor hacminin 0-1400 arasında olması %5,2, otomobilin 1 sn. daha geç hızlanması %15,5, ortalama yakıt tüketiminin 1 artması %7, önden-arkadan çekişli olması %5,7, otomobilin kasa tipinin sedan olması %9,1, deri koltuk özelliğinin (ID_5) olması %3,8, elektrikli sürücü koltuğu (ID_7) olması %0,3, diz hava yastığı (TD_4) olması %7,7, hıza duyarlı direksiyon (TD_7) olması %2,3, lastik basınç göstergesi (TD_9) olması %17,1, adaptif far (DD_1) olması %9,7, ısıtmalı aynalar (DD_8) olması %2,6, sis farı (DD_{12}) olması %1,9 oranında fiyatı azaltmaktadır.

Tablo 5.4 Doğrusal ve yarı logaritmik regresyon modeli performans ölçüm sonuçları

Metrik	Eğitim Verisi Tahmin		Test Verisi Tahmin	
	Doğrusal Model	**Yarı Logaritmik Model**	**Doğrusal Model**	**Yarı Logaritmik Model**
RMSE	58440,4	0,105	54710,39	0,104
R^2	0,903	0,932	0,914	0,932

Tablo 5.4'e göre doğrusal model tarafından yapılan tahminlerin ortalama olarak gerçek değerlerden yaklaşık 54710,39 birim saptığı ve modelin hedef değişkendeki (fiyat) varyansın yaklaşık %91'ini, yarı logaritmik model tarafından yapılan tahminlerin ise ortalama gerçek değerlerden yaklaşık 0,104 birim saptığı ve modelin hedef değişkendeki varyansın yaklaşık %93'ünü açıkladığı söylenebilir.



Şekil 5.19 Doğrusal ve yarı logaritmik modelin kalıntı grafiği

Şekil 5.19'da doğrusal regresyon ve yarı logaritmik regresyon modellerinin kalıntı grafiği incelendiğinde yarı logaritmik regresyon modelinin doğrusal regresyon modeline göre kalıntılarının sıfır etrafında rasgele dağılmış ve iki paralel doğru arasında çok büyük sapmalar göstermeden normal bir dağılım gösterdiği görülmektedir. Bu nedenle yarı logaritmik regresyon modelinin daha uygun olduğu söylenebilir.

Çoklu doğrusal bağlantı sorununu test etmek için Variance Inflation Factors (VIF) testi gerçekleştirilmiştir. Tablo 5.5'te verilen sonuçlar incelendiğinde modellerde çoklu bağlantı sorunu olduğu görülmektedir.

Tablo 5.5 VIF test sonuçları

D	**1**	**2**	**D**	**1**	**2**	**D**	**1**	**2**
MM	6,869766e+28	8,802085e+28	ID_3	18,808	18,795	TD_8	4,905	5,025
A	3,787	3,284	ID_4	10,632	9,694	TD_9	8,117	8,145
V	3,957	3,909	ID_5	35,457	34,974	TD_{10}	4,473	4,460
KM	1,629	1,755	ID_6	17,160	17,227	TD_{11}	26,942	26,777
R	1,065	1,063	ID_7	28,182	28,666	TD_{12}	3,453	3,500
MG	7,195	6,275	ID_8	10,713	10,892	TD_{13}	38,952	39,199
MH	5,392	5,601	ID_9	22,408	22,816	TD_{14}	40,209	39,276
H	64,439	74,027	ID_{10}	8,598	8,645	DD_1	6,285	6,366
OYT	18,741	19,674	ID_{11}	40,835	41,856	DD_2	17,187	17,256
C	7,271	7,889	ID_{12}	29,549	30,061	DD_3	13,605	13,550
W	64,856	60,574	ID_{13}	5,275	5,158	DD_4	26,761	26,467
SS	6,607	6,272	ID_{14}	50,511	49,962	DD_5	53,979	54,591
YT	50,153	57,974	ID_{15}	14,381	14,501	DD_6	10,382	10,515
T	62,343	82,418	ID_{16}	22,378	22,905	DD_7	15,264	15,240
U	209,920	294,728	ID_{17}	13,417	13,754	DD_8	21,859	21,742
G	106,013	112,328	ID_{18}	17,558	17,500	DD_9	9,719	9,686
Y	103,032	97,647	TD_1	13,527	13,118	DD_{10}	9,208	9,234
BK	85,253	160,304	TD_2	50,714	52,756	DD_{11}	14,293	14,475

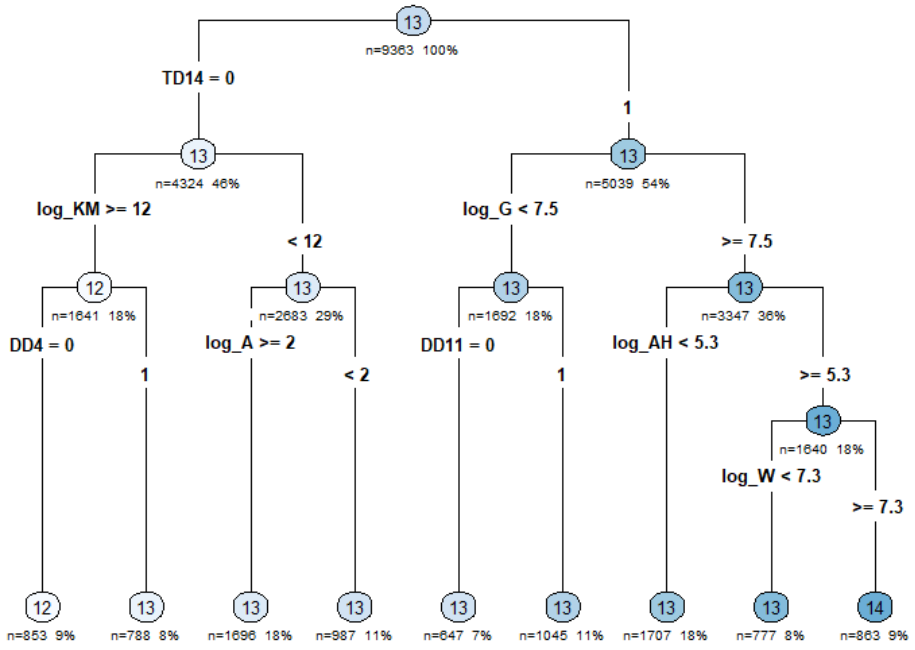
Tablo 5.5 VIF test sonuçları (devamı)

KT	55,489	54,612	TD_3	2,7605	2,820	DD_{12}	12,023	12,055
TU	7,974	13,561	TD_4	50,335	50,420	DD_{13}	7,874	7,920
AH	125,821	148,215	TD_5	38,712	38,631	DD_{14}	6,325	6,301
ID_1	76,953	142,062	TD_6	65,394	66,404	DD_{15}	58,209	57,964
ID_2	18,564	18,768	TD_7	29,606	30,961			

D: Değişkenleri, 1: Doğrusal modeli ve 2: Yarı logaritmik modeli ifade etmektedir.

5.3.3.1 Regresyon Ağacı

CART ve BART modellerinde yukarıdaki analizler doğrultusunda yarı logaritmik model kullanılarak ilerlenmiştir. Şekil 5.20'e göre yarı logaritmik model ile oluşturulan regresyon ağacında (regression tree) fiyatı logaritmik olarak 14 olan bir otomobilin ağırlığı 7,3'e eşit veya büyük, azami hızı 5,3'e eşit veya büyük, genişliği 7,5'e eşit veya büyük ve yokuş kalkış desteği olan otomobillerdir.



Şekil 5.20 Veri setine ait regresyon ağacı grafiği

Tablo 5.6 Regresyon ağacı modeli performans ölçüm sonuçları

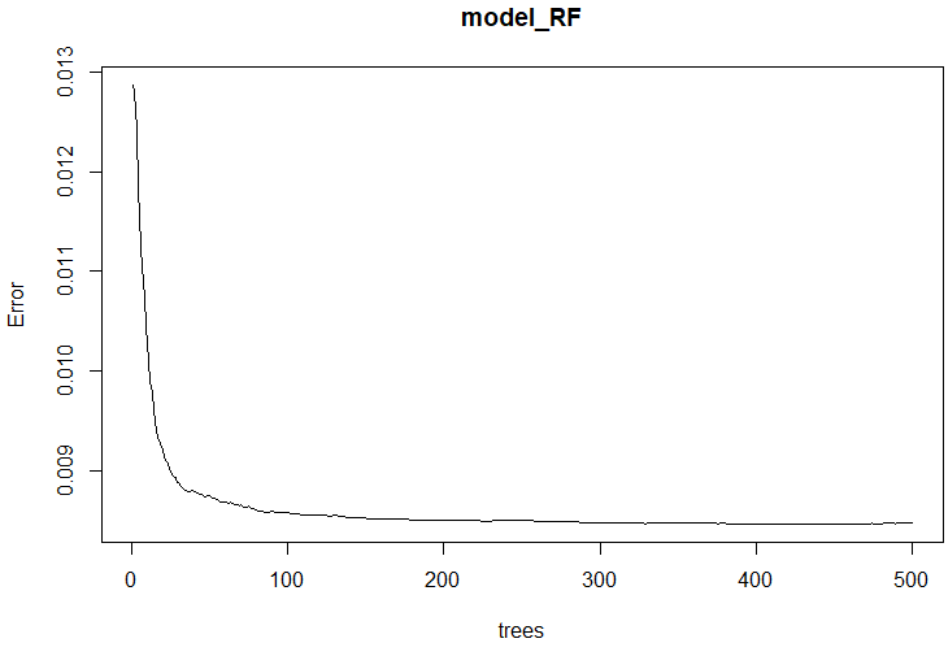
Metrik	Eğitim Verisi Tahmin	Test Verisi Tahmin
RMSE	0,186	0,185
R^2	0,785	0,785

RMSE değeri olan 0,185, regresyon ağacı modeli tarafından yapılan tahminlerin ortalama olarak gerçek değerlerden yaklaşık 0,185 birim saptığı anlamına gelmektedir. Daha düşük bir RMSE değeri, daha küçük tahmin hatalarını gösterdiğinden genellikle daha iyidir.

R-kare (R^2) değeri modelin hedef değişkendeki varyansı ne kadar iyi açıkladığını ölçer. 0 ile 1 arasında değişir ve daha yüksek değer daha iyi uyumu gösterir. Bu modelde R^2 0,785'dir; bu, modelin hedef değişkendeki (fiyat) varyansın yaklaşık %78,5'ini açıkladığı anlamına gelmektedir.

5.3.3.2 Rastgele Orman

Rastgele orman (random forest) algoritması kullanıldığında, bir dizi karar ağacı önyüklemeli (bootstrapped) eğitim örneklerinden oluşturulur. Ancak bu karar ağaçları oluşturulurken, her bölünmede yalnızca rastgele bir alt küme öngörücü kullanılır. Her bölünmede m tane öngörücüyü içeren yeni bir örnek alınır ve genellikle $m \approx sqrt(p)$ olarak seçilir; yani her bölünmede düşünülen öngörücülerin sayısı genellikle toplam öngörücülerin karekköküne eşittir. Bu, daha zayıf öngörücülerin bir bölünme sırasında daha fazla düşünülme şansına sahip olmalarını sağlar ve random forest modelinin daha iyi genelleme yapabilmesine olanak tanır. Son tahmin, tüm ağaçların ortalaması veya çoğunluk oyu (regresyon veya sınıflandırma probleminin türüne bağlı olarak) olacaktır [19].



Şekil 5.21 Veri setine ait rastgele orman grafiği

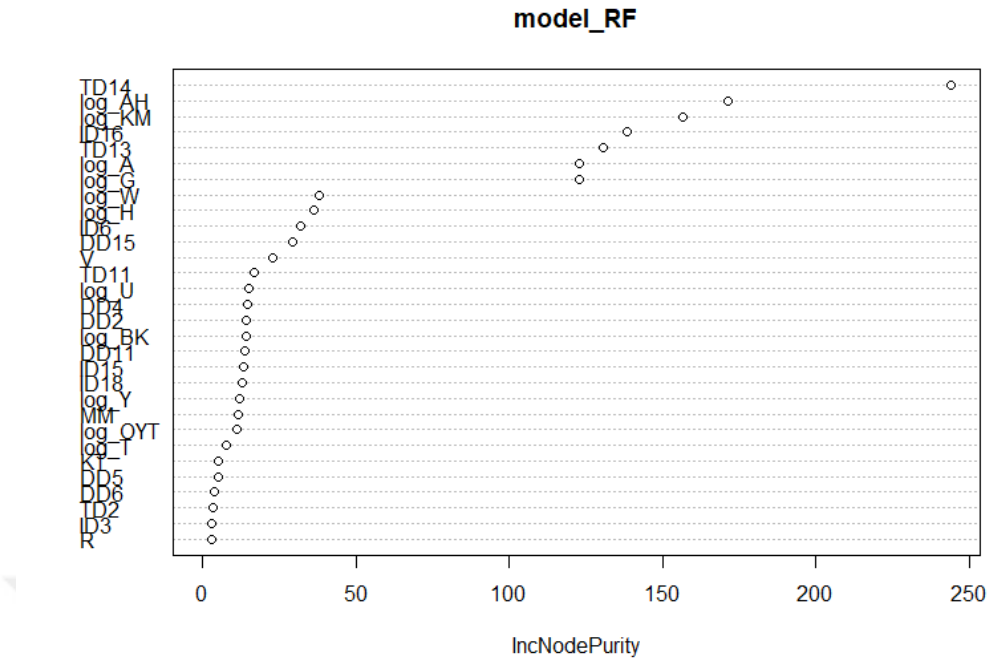
Şekil 5.21'de hata oranının yaklaşık 100 ağaçta sabitlendiği ancak yaklaşık 400 ağaca kadar yavaş yavaş azalmaya devam ettiği ve hangi ağaç sayısının en düşük hata oranını sağladığı görülmektedir. Yani 442 ağacın ortalama 0,092 otomobil satış fiyatı hatasını sağladığı söylenebilir.

Tablo 5.7 Rastgele orman değişken önem değerleri

Değişken	**Önem Değeri**	**Değişken**	**Önem Değeri**	**Değişken**	**Önem Değeri**
TD_{14}	243,819	T	7,546	DD_9	0,850
AH	171,207	KT	5,203	TD_1	0,777
KM	156,492	DD_5	5,162	ID_8	0,756
ID_{16}	138,375	DD_6	3,582	ID_{13}	0,737
TD_{13}	130,352	TD_2	3,123	TU	0,677
A	122,976	ID_3	3,082	TD_9	0,636
G	122,830	R	2,979	TD_4	0,619

Tablo 5.7 Rastgele orman değişken önem değerleri (devamı)

W	37,894	ID_7	2,629	ID_{11}	0,616
H	36,123	TD_6	2,488	DD_{13}	0,604
ID_6	31,664	ID_{17}	2,345	C	0,590
DD_{15}	29,177	DD_8	2,195	ID_9	0,448
V	22,736	ID_{14}	2,122	ID_5	0,431
TD_{11}	16,575	ID_{10}	2,111	TD_{10}	0,355
U	14,978	DD_3	2,080	MG	0,340
DD_4	14,528	DD_{12}	1,899	DD_1	0,332
DD_2	14,126	TD_5	1,610	ID_1	0,265
BK	13,941	MH	1,405	ID_{12}	0,263
DD_{11}	13,823	TD_7	1,305	TD_{12}	0,077
ID_{15}	13,354	ID_4	1,288	DD_7	0,061
ID_{18}	12,889	ID_2	1,201	TD_3	0,052
Y	12,184	DD_{10}	0,977	TD_8	0,027
MM	11,698	YT	0,971	DD_{14}	0,017
OYT	11,258	SS	0,965		



Şekil 5.22 Rastgele orman modeline göre otomobil özelliklerinin önem sırası grafiği

Tablo 5.7 ve Şekil 5.22 rastgele orman modelinin değişkenlerinin önem değerlerini göstermektedir. Rastgele orman modeline göre ikinci el otomobillerde fiyatı etkileyen en önemli 3 değişkeni yokuş kalkış desteği, azami hız ve kilometredir.

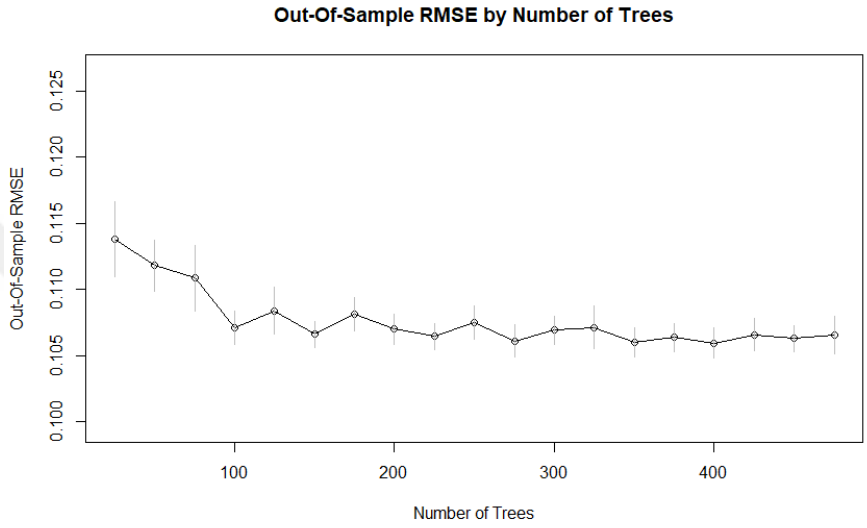
Tablo 5.8 Rastgele orman modeli performans ölçüm sonuçları

Metrik	Eğitim Verisi Tahmin	Test Verisi Tahmin
RMSE	0,069	0,089
R^2	0,970	0,95

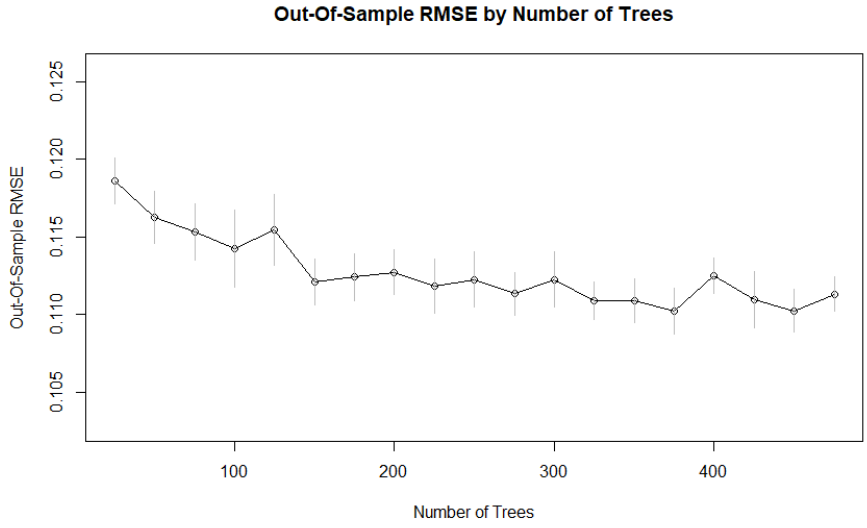
Tablo 5.8'e göre rastgele orman modeli tarafından yapılan tahminlerin ortalama olarak gerçek değerlerden yaklaşık 0,089 birim saptığı ve modelin hedef değişkendeki (fiyat) varyansın yaklaşık %95'ini açıkladığı söylenebilir.

5.3.3.3 Bayes Toplamsal Regresyon Ağaçları

R programında yer alan bartMachine kütüphanesinde ağaç sayısı için varsayılan 50 değerini kullanmak yerine rmse_by_num_trees fonksiyonu kullanılarak optimum ağaç sayısı 150 belirlenmiş olup BART algoritması bu ağaç sayısına göre çalıştırılmıştır. Şekil 5.23 ve Şekil 5.24 sırasıyla eğitim ve test veri setleri için ağaç sayılarına göre RMSE değerlerini göstermektedir.



Şekil 5.23 Eğitim veri setine ait ağaç sayısı grafiği

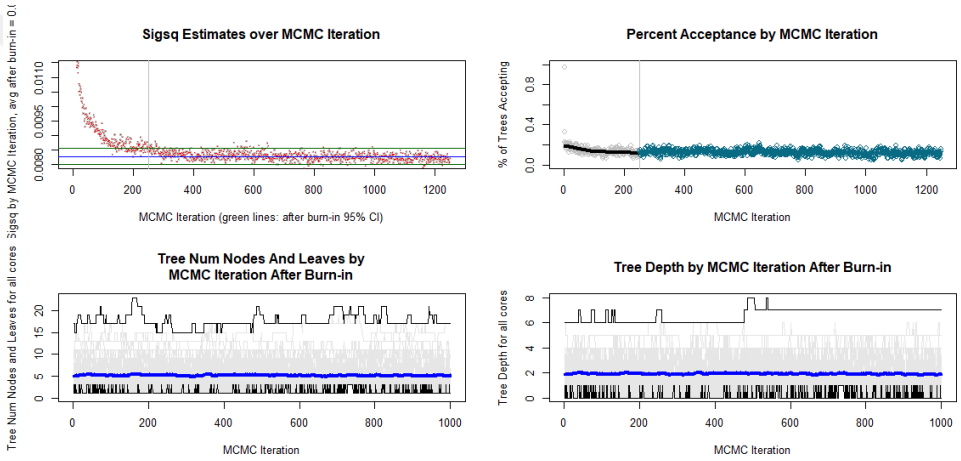


Şekil 5.24 Test veri setine ait ağaç sayısı grafiği

Tablo 5.9'a göre eğitim verisinde BART tarafından yapılan tahminlerin ortalama olarak gerçek değerlerden yaklaşık 0,103 birim saptığı ve modelin hedef değişkendeki (fiyat) varyansın yaklaşık %94'ünü açıkladığı söylenebilir.

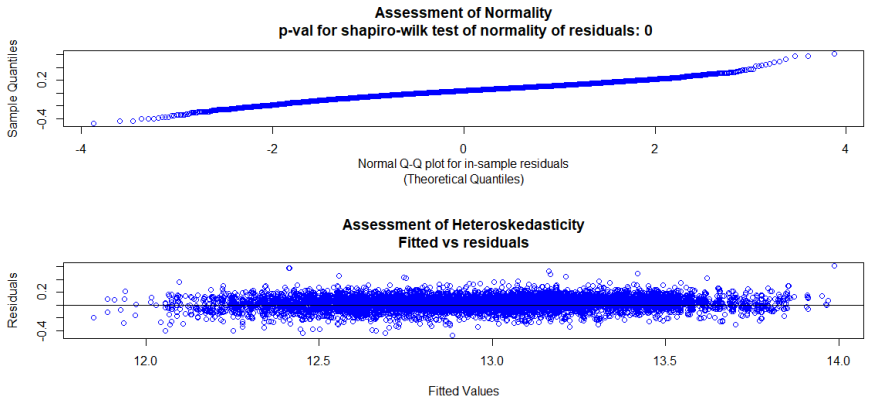
Tablo 5.9 BART modeli ağaç modeli performans ölçüm sonuçları

Metrik	Eğitim Verisi Tahmin	Test Verisi Tahmin
RMSE	0,1	0,103
R^2	0,939	0,938



Şekil 5.25 Eğitim veri setine ait yakınsama teşhis grafikleri

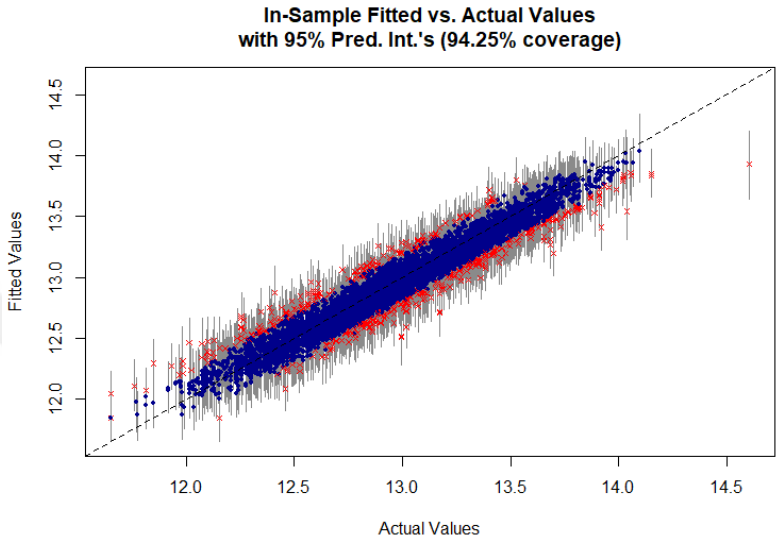
Şekil 5.25'e göre MCMC örneklemesinin yaklaşık 200 yinelemede yakınsamaya ulaştığı görülebilir.



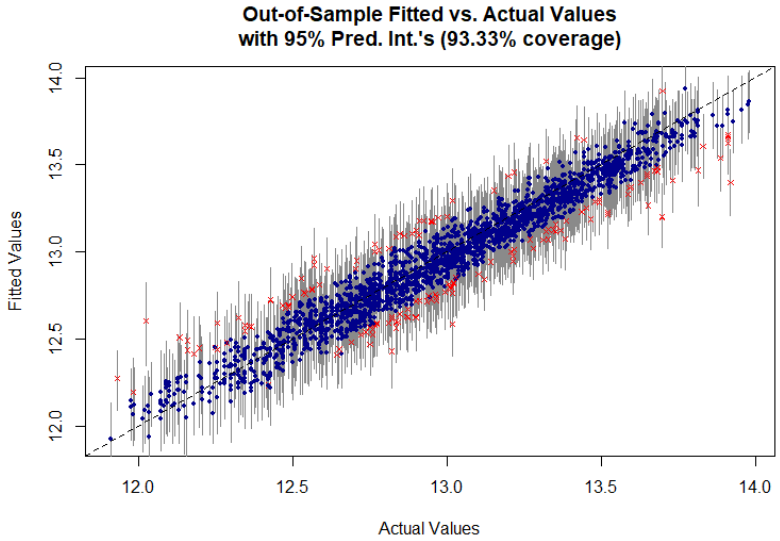
Şekil 5.26 Eğitim veri setine ait artıkların dağılım grafiği

Şekil 5.26'e göre artıkların sıfır etrafında rasgele dağılmış ve iki paralel doğru arasında çok büyük sapmalar göstermeden normal bir dağılım gösterdiği görülmektedir.

Şekil 5.27 ve Şekil 5.28 tahmin edilen medyan fiyatın gerçek medyan fiyata karşı karşılaştırma grafiklerini göstermektedir. Şekil 5.27 ve Şekil 5.28'ye bakılarak algoritmanın hem eğitim hem de test setindeki performansı incelenebilir.



Şekil 5.27 Eğitim veri setine ait tahmin edilen ve gerçek fiyat karşılaştırma grafiği

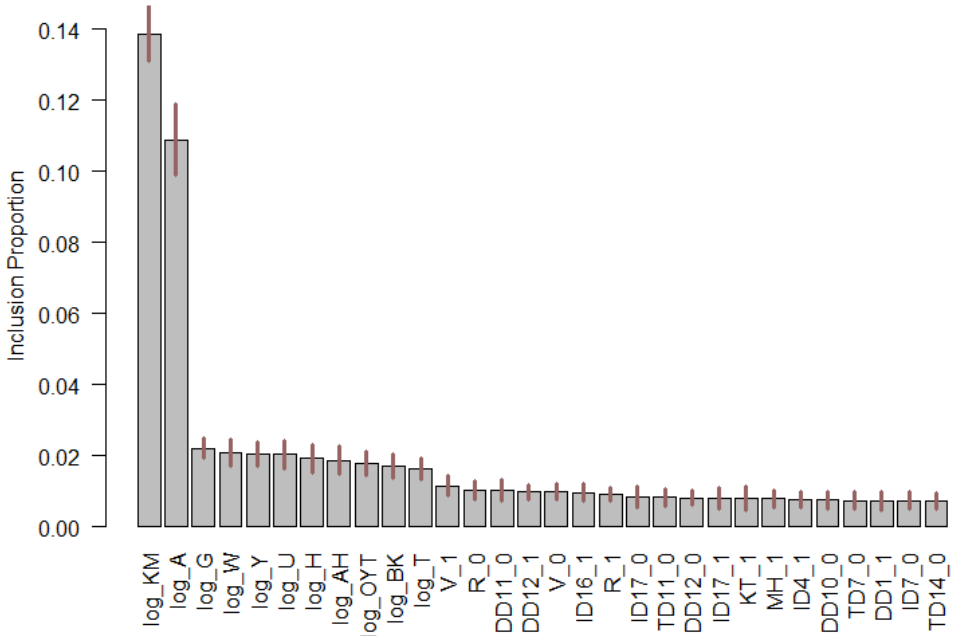


Şekil 5.28 Test veri setine ait tahmin edilen ve gerçek fiyat karşılaştırma grafiği

Karşılaştırma grafiklerine göre her veri noktası için yaklaşık %93 doğruluğunda (%95 tahmin aralığı) bir tahmin aralığı verilebildiği görülmektedir.

Tablo 5.10 BART modeli önemli ilk 30 değişkenin değerleri

Değişken	**Önem Değeri**	**Değişken**	**Önem Değeri**	**Değişken**	**Önem Değeri**
KM	0,1387	T	0,0163	DD_{12-0}	0,0083
A	0,1090	V_1	0,0116	ID_{17-1}	0,0082
G	0,0221	R_0	0,0104	KT_1	0,0082
W	0,0211	DD_{11-0}	0,0104	MH_1	0,0079
Y	0,0205	DD_{12-1}	0,0099	ID_{4-1}	0,0077
U	0,0204	V_0	0,0098	DD_{10-0}	0,0076
H	0,0193	ID_{16-1}	0,0098	TD_{7-0}	0,0074
AH	0,0188	R_1	0,0092	DD_{1-1}	0,0074
OYT	0,0181	ID_{17-0}	0,0086	ID_{7-0}	0,0074
BK	0,0171	TD_{11-0}	0,0084	TD_{14-0}	0,0073



Şekil 5.29 BART modeline göre otomobil özelliklerinin önem sırası grafiği

Tablo 5.10 ve Şekil 5.29'a göre ikinci el otomobillerde fiyatı etkileyen en önemli 30 özellikler arasında sırasıyla otomobilin kilometresi, yaşı, genişliği, ağırlığı, yüksekliği, uzunluğu, hızlanması, azami hızı, ortalama yakıt tüketimi, bagaj kapasitesi, torku, vites, renk, park sensörü, sis farı, start-stop, usb bağlantısı, mesafe takip sistemi, kasa tipi, motor hacmi, Bluetooth, gündüz sürüş farları, hıza duyarlı direksiyon, adaptif far, elektrikli sürücü koltuğu ve yokuş kalkış desteği yer almaktadır.

6
SONUÇ

Bu çalışmada Bayes toplamsal regresyon ağaçları (BART) ile ikinci el otomobillerin birtakım özelliklerinin fiyat üzerindeki etkisi araştırılmıştır. Öncelikle Bayesyen çıkarsamanın teorik kısmı anlatılmıştır. Önsel dağılım ve sonsal dağılım tanımlarına değinilmiş olup türlerinden bahsedilmiştir. Markov Zinciri Monte Carlo (MCMC) yöntemi ve algoritmaları anlatılmıştır. İlk olarak Chipman ve arkadaşları tarafından (2010)'de ortaya çıkan BART konusu detaylı bir şekilde ele alınmıştır. Önsel, sonsal dağılım ve MCMC yöntemine ait algoritmaların detaylarına yer vererek BART algoritması anlatılmıştır. Son olarak sınıflandırma ve regresyon ağaçları (CART) konusuna değinilmiştir.

Yapılan uygulamada aşamalı olarak ikinci el otomobillere ait özelliklerden oluşan değişkenlerin tanımlayıcı istatistiklerine yer verilmiştir. 2022 Aralık ayı içerisinde satışta olan ikinci el otomobillerin marka model, renk, vites, silindir sayısı, yakıt türü, turbo, kasa tipi, çekiş, teknik donanımlar, iç-dış donanımlar gibi kategorik özelliklerinin fiyat ile ilişkisi kutu grafikleri ile, otomobil yaşı, kilometre, hızlanma, ortalama yakıt tüketimi, bagaj kapasitesi, motor gücü, motor hacmi, tork, azami hız, uzunluk, genişlik, yükseklik, ağırlık gibi sürekli değere sahip özelliklerinin fiyat ile ilişkisi ise saçılma grafikleri ile araştırılmıştır. Daha sonra ise CART ve BART karşılaştırması için temel alınacak olan modeli belirlemek amacıyla doğrusal ve yarı logaritmik regresyon modeli oluşturulup kalıntı grafikleri incelenmiştir ve en iyi modelin yarı logaritmik model olduğuna karar verilmiştir.

Uygulamanın son aşamasında yarı logaritmik model ile regresyon ağacı, rastgele orman ve BART algoritması çalıştırılmıştır. Bu üç algoritmanın performansı incelendiğinde klasik regresyon için R^2 değeri 0,93, BART için 0,95 ve rastgele orman için 0,95 elde edilmiş olup modellerin tahmin performansı oldukça yüksektir. Ağaç temelli yöntemlerin kullanılması klasik regresyon modeline göre tahmin performansını artırmıştır.

Elde edilen sonuçlara göre yokuş kalkış desteği, azami hız, kilometre, start-stop, perde hava yastığı, yaş, genişlik, ağırlık, hızlanma, dijital klima, yağmur sensörü, vites, mesafe takip sistemi, uzunluk, elektrikli aynalar, alaşımlı jant, bagaj

kapasitesi, park sensörü, ön kol dayama, yol bilgisayarı, yükseklik, tork, kasa tipi, far sensörü, far yıkama, elektronik patinaj önleme sistemi (ASR), arka kol dayama, renk, elektrikli sürücü koltuğu, ortalama yakıt tüketimi, sis farı, usb bağlantısı, motor hacmi, bluetooth, gündüz sürüş farları, hıza duyarlı direksiyon, adaptif far ve değişkenlerinin otomobil fiyatını önemli düzeyde etkiledikleri tespit edilmiştir.

Çalışmada elde edilen sonuçlar, araştırma döneminde elde edilebilen verinin büyüklüğü, kullanılan değişkenler ve analiz yöntemleri ile sınırlı olmaktadır. Farklı dönemlere ait elde edilebilen veriler, farklı değişkenler ve yöntemlerin kullanılması analiz sonuçlarının farklılaşmasına yol açabilir. Çalışmadan elde edilen bilgilerin ileride yapılacak bayes toplamsal regresyon ağaçları model analizlerine ışık tutması beklenmektedir.

KAYNAKÇA

[1] L. Breiman, «Random forests,» *Machine learning,* 2001, pp. 5-32.

[2] J. Friedman, «Greedy function approximation: A gradient boosting machine,» *Ann. Statist.*, 2001, pp. 1189-1232.

[3] M. Taddy, R. Gramacy ve N. Polson, «Dynamic Trees for Learning and Design,» *Journal of the,* 2011, p. 493.

[4] H. A. CHIPMAN, E. I. GEORGE ve R. E. MCCULLOCH, «BART: BAYESIAN ADDITIVE REGRESSION TREES,» *The Annals of Applied Statistics,* pp. 266-298, 2010.

[5] G. Blattenberger ve R. Fowles, «Avalanche Forecasting: Using Bayesian Additive Regression Trees (BART),» *Demand for Communications Services–Insights and Perspectives: Essays in Honor of Lester D. Taylor*, 2013, pp. 211-227.

[6] Q. Zhou ve J. S. Liu, «Extracting sequence features to predict protein-DNA binding: A comparative study,» *Nucleic Acids Research,* 2008, pp. 4137-4148.

[7] J. Eliashberg, «Green-Lighting Movie Scripts: Revenue Forecasting and Risk Management,» *Doktora Tezi*, University of Pennsylvania, 2010.

[8] A. Kapelner ve J. Bleich, «bartMachine: Machine Learning with Bayesian Additive Regression Trees,» arXiv preprint arXiv: 1312.2171v3, 2014.

[9] B. Hernández, S. Pennington ve A. Parnell, «Bayesian methods for proteomic biomarker development,» *EuPA Open Proteomics*, 2015, pp. 54-64.

[10] G. Kropat, F. Bochud ve M. Jaboyedoff, «Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units,» *J Environ Radioact*, 2015, pp. 51-62.

[11] M. Leonti, S. Cabras, C. Weckerle, M. Solinas ve L. Casu, «The causal dependence of present plant knowledge on herbals—contemporary medicinal plant use in Campania (Italy) compared to Matthioli (1568),» *J Ethnopharmacol*, 2010, pp. 379-391.

[12] J. Hill, «Bayesian nonparametric modeling for causal inference,» *J Comput Graph Stat*, 2011, pp. 217-240.

[13] Y. Liu, Z. Shao ve G.-C. Yuan, «Prediction of Polycomb target genes in mouse embryonic stem cells,» *Genomics*, 2010, pp. 17-26.

[14] Y. Liu, M. Traskin, S. Lorch, E. George ve D. Small, «Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance,» *Health Care Manag Sci,* 2015, pp. 58-66.

[15] J. Zhang ve W. Härdle, «The Bayesian additive classification tree applied to credit risk modelling,» *Comput Stat Data Anal,* 2010, pp. 1197-1205.

[16] R. Nateghi, S. Guikema ve S. Quiring, «Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes,» *Risk Analysis,* 2011, pp. 1897-1906.

[17] H. Chipman, E. George, L. Lemp ve R. McCulloch, «Bayesian flexible modeling of trip durations,» *Transp Res B Methodol,* 2010, pp. 686-698.

[18] J. Ding, A. Bashashati ve A. Roth, « Feature based classifiers for somatic mutation detection in tumour-normal paired sequencing data,» *Bioinformatics,* 2012, pp. 167-175.

[19] V. Bonato, V. Baladandayuthapani, B. Broom, E. Sulman, K. Aldape ve K.-A. Do, «Bayesian ensemble methods for survival prediction in gene expression data,» *Bioinformatics,* 2011, pp. 359-367.

[20] R. Sparapani, B. Logan, R. McCulloch ve P. Laud, «Nonparametric survival analysis using Bayesian Additive Regression Trees (BART),» *Statist Med,* 2016, pp. 2741-2753.

[21] B. Kindo, H. Wang ve E. Peña, «Multinomial probit Bayesian additive regression trees,» *Stat,* 2016, pp. 119-131.

[22] R. Agarwal, P. Ranjan ve H. Chipman, «A new Bayesian ensemble of trees approach for land cover classification of satellite imagery,» *Can J Remote Sens,* 2013, pp. 507-520.

[23] A. Linero, D. Sinhay ve S. Lipsitz, «Semiparametric mixed-scale models using shared Bayesian forests,» arXiv preprint: 1809.08521, 2018.

[24] D. Green ve H. Kern, «Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees,» *Public Opin Q,* 2012, pp. 491-511.

[25] D. Xu, M. Daniels ve A. Winterstein, «Sequential BART for imputation of missing covariates,» *Biostatistics,* 2016, pp. 589-602.

[26] A. Kapelner ve J. Bleich, «Prediction with missing data via Bayesian Additive Regression Trees,» *Can J Stat,* 2015, pp. 224-239.

[27] Y. Tan, A. Flannagan ve M. Elliott, «"Robust-squared" imputation models using BART,» *J Surv Stat Methodol,* 2019.

[28] J. Murray, «Log-linear Bayesian additive regression trees for categorical and count responses,» arXiv preprint: 1701.01503, 2017.

[29] J. Starling, J. Murray, C. Carvalho, R. Bukowski ve J. Scott, «BART with targeted smoothing: an analysis of patient-specific stillbirth risk,» arXiv preprint: 1805.07656, 2018.

[30] R. Sparapani, L. Rein, S. Tarima, T. Jackson ve J. Meurer, «Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes,» *Biostatistics,* 2018.

[31] S. Sivaganesan, P. Müller ve B. Huang, «Subgroup finding via Bayesian additive regression trees,» *Statist Med,* 2017, pp. 2391-2403.

[32] P. Schnell, Q. Tang, W. Offen ve B. Carlin, «A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects,» *Biometrics,* 2016, pp. 1026-1036.

[33] P. Schnell, P. Müller, Q. Tang ve B. Carlin, «Multiplicity-adjusted semiparametric benefiting subgroup identification in clinical trials,» *Clinical Trials,* 2018, pp. 75-86.

[34] Y. Tan, C. Flannagan, L. Pool ve M. Elliott, «Accounting for selection bias due to death in estimating the effect of wealth shock on cognition for the Health and Retirement Study,» arXiv preprint: 1812.08855, 2018.

[35] Y. V. Tan ve J. Roy, «Bayesian Additive Regression Trees and the General BART model.,» *Statistics in medicine 38(25),* pp. 5048-5069, 2019.

[36] Ö. ASİLKAN ve S. IRMAK, «İKİNCİ EL OTOMOBİLLERİN GELECEKTEKİ FİYATLARININ YAPAY SİNİR AĞLARI İLE TAHMİN EDİLMESİ,» *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* pp. 375-391, 2009.

[37] H. DAŞTAN, «TÜRKİYE'DE İKİNCİ EL OTOMOBİL FİYATLARINI ETKİLEYEN FAKTÖRLERİN HEDONİK FİYAT MODELİ İLE BELİRLENMESİ,» *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* pp. 303-327, 2016.

[38] R. Yayar ve E. Yılmaz, «İkinci El Otomobil Talep Fiyatının Regresyon Analizi:TR83 Bölgesi Örneği*,» *Eskişehir Osmangazi Üniversitesi İİBF Dergisi,* pp. 39-52, 2018.

[39] O. Ekici, «Bayesyen Regresyon ve WinBUGS ile Bir Uygulama İstanbul Üniversitesi Sosyal Bilimler Enstitüsü,» *Doctoral dissertation, Yüksek Lisans Tezi, Ekonometri Anabilim Dalı,* p. 115, 2005.

[40] F. J. Anscombe, «Bayesian Statistics,» *The American Statistician 15(1),* pp. 21-24, 1961.

[41] E. Özgören, «Bayes Açıortay Regresyon Tekniği ve Bir Uygulama,» *Afyon Kocatepe Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi,* 2019.

[42] H. C. İNAL ve S. DÜNAY, Olasılık ve Matematiksel İstatistik, Hacettepe Üniversitesi Yayınları, 2013.

[43] N. Gasım, «Bayesyen Model ile Doğrusal Regresyon Modellerinin Karşılaştırılması Üzerine Bir Uygulama,» *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Yüksek Lisans Tezi,* 2013.

[44] G. E. P. Box ve G. C. T. Tiao, «Nuisance Parameters,» *Bayesian Inference in Statistical Analysis,* John Wiley & Sons, 2011, p. 70.

[45] D. Tektaş, «İki Düzeyli Lojit ve Probit Modellerde Parametre Tahminlerine Bayesci Bir Yaklaşım,» *Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi,* 2006.

[46] S. Gündoğdu, «Balıklarda Büyüme Parametrelerinin Bayesyen İstatistiksel Yöntemle Tahmini,» *Doktora Tezi, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü,* 2016.

[47] Kass, R. E., Wasserman ve Larry, «The Selection of Prior Distributions by Formal,» *Journal of the American Statistical*, 1996, pp. 1343-1362.

[48] J. Gill, «Bayesian Method: A Social and Behavioral Sciences Approach.,» *Chapman&Hall/CRC,* p. 2. edition, 2008.

[49] Ö. Karadağ, Bayesci Hiyerarşik Modeller, Ankara, 2011.

[50] Z. Öztürk, «Genelleştirilmiş Lineer Karma Modellerde Bayesci Yaklaşımın Kullanımı,» *Doktora Tezi*, Samsun, Ondokuz Mayıs Üniversitesi, 2014.

[51] G. Zerey, «Klasik Bayesci Portföy Seçim Modellerinin Bist 30 Üzerinde,» *Doktora Tezi*, Samsun, Ondokuzmayıs Üniversitesi, 2018.

[52] M. A. Cengiz, E. Terzi, T. Şenel ve N. Murat, «Lojistik Regresyonda Parametre Tahmininde Bayesci Bir Yaklaşım,» *Afyon Kocatepe Üniversitesi Fen Bilimleri Dergisi,* pp. 15-22, 2012.

[53] E. Avcı, «Bayesci Sağkalım Analizi ve Meme Kanseri Verileri Üzerine Bir Uygulama,» *Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü Doktora Tezi,* 2012.

[54] İ. Kurt, «Bayesgil Yaşam Analizi ve Cox Regresyon Yaşam Analizi'nin Türetilmiş ve Gerçek Veri Setlerinde Uygulanması,» Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü, 2008.

[55] İ. Saçaklı Saçıldı, «Gelişmiş ve Gelişmekte Olan Piyasalarda Hisse Senedi Getiri Volatilitelerinin Bayes Yaklaşımı İle Analizi,» *Marmara Üniversitesi Sosyal Bilimler Enstitüsü Doktora Tezi,* 2011.

[56] N. Metropolis, «Equation of state calculations by fast computing machines.,» *The journal of chemical physics,* 1953, pp. 1087-1092.

[57] S. Sharma, «Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy,» *Annu. Rev. Astron. Astrophys.,* pp. 1-49, 2017.

[58] W. K. Hastings, «Monte Carlo Sampling Methods Using Markov Chains and Their Applications,» *Biometrika,* 1970, pp. 97-109.

[59] S. Geman ve D. Geman, «Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,» *IEEE Transactions on pattern analysis and machine intelligence,* 1984, pp. 721-741.

[60] J. He, S. Yalov ve P. Hahn, «XBART: Accelerates Bayesian Additive Regression Trees,» *In The 22nd International Conference on Artificial Intelligence and Statistics,* pp. 1130-1138, 2019.

[61] J. Hill, A. Linero ve J. Murray, «Bayesian Additive Regression Trees: A Review and Look Forward,» *Annual Review of Statistics and Its Application,* no. 7, pp. 251-278, 2020.

[62] D. C. Montgomery, E. A. Peck ve G. G. Vining, «Sınıflama ve Regresyon Ağaçları (CART),» *Introduction to Linear Regression Analysis,* 2013, pp. 524-526.

[63] S. Maiti ve N. Wattanasirichaigoon, MODEL COMPARISON BETWEEN BART AND CART, Chicago: ILLINOIS INSTITUTE OF TECHNOLOGY, 2020.

[64] S. Stigler, «Who discovered Bayes's theorem?,» *The American Statistician,* pp. 290-296, 1983.

[65] O. Mamun, «A Primer to Bayesian Additive Regression Tree with R,» 3 5 2021. [Çevrimiçi]. Available: https://towardsdatascience.com/a-primer-to-bayesian-additive-regression-tree-with-r-b9d0dbf704d.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. A. Altuntaş, G. G. Şimşek, «Bayes Toplamsal Regresyon Ağaçları ve Otomobil Fiyatları Üzerine Bir Uygulama,» *3rd International Conference on Engineering and Applied Natural Sciences,* January 14-17, 2023, Konya.