



Sağlık Sigortası Maliyet Tahmini

Yapay zeka ile sigorta sektöründe yeni bir dönem: Müşteri sağlık harcamalarını önceden tahmin eden akıllı model

Yapay Zeka ile Sağlık Harcaması Tahmini



Tahmin Gücü

Sigortalıların yıllık sağlık harcamasını önceden belirleme



Risk Yönetimi

Fiyatlandırma politikaları ve risk analizi süreçlerini optimize etme



Hız ve Doğruluk

Manuel hesaplama süreçlerini ortadan kaldırarak karar verme sürecini hızlandırma



ADIM 1

İş Problemini Anlama

Temel Soru

Bir kişinin yaşı, vücut kitle indeksi, sigara içme durumu gibi demografik ve sağlık bilgileri bilindiğinde, yıllık sağlık gideri ne kadar olacak?

Hedef Değişken: charges (yıllık sağlık harcaması)

Bu model, sigorta şirketi için kritik stratejik karar destek aracı olarak konumlandırılmıştır.



Veri Seti ve Temel Gözlemler

1

Veri Değişkenleri

age (yaş), sex (cinsiyet), bmi (vücut kitle indeksi), children (çocuk sayısı), smoker (sigara), region (bölge), charges (hedef)

2

Sigara Etkisi

Sigara içenlerin sağlık maliyetleri dramatik şekilde yüksek
- en güçlü maliyet belirleyici faktör

3

Yaş ve BMI İlişkisi

Yaş arttıkça ve BMI değeri yükseldikçe sağlık maliyeti doğal olarak artış gösteriyor

4

Bölgelik Etki

Region değişkeninin etkisi belirgin değil ve istatistiksel anlamlılığı zayıf bulundu



ADIM 3

Veri Hazırlama ve Feature Engineering

01

Veri Kalitesi Kontrolü

Eksik veri analizi yapıldı ve veri setinde **hiçbir eksik değer bulunmadı**, bu da temiz bir başlangıç sağladı.

02

Encoding İşlemi

Kategorik değişkenler uygun şekilde sayısal değerlere dönüştürüldü:

- **Label Encoding:** 'smoker' (yes/no → 1/0) ve 'sex' (male/female → 1/0) değişkenlerine uygulandı.
- **One-Hot Encoding:** 'region' değişkeni için kullanıldı.

03

Region Değişkeninin Çıkarılması

'Region' değişkeninin **istatistiksel anlamlılığı zayıf** olduğu tespit edildi. Model sadeleştirildi ve bu değişkenin çıkarılmasıyla **performans kaybı yaşanmadı**.

04

Yeni Feature Oluşturma

Modelin açıklayıcılığını artırmak için üç yeni etkileşim ve kategori özelliği oluşturuldu:

- **smoker_age_interaction:** Sigara içen bireylerdeki yaşın sağlık maliyetleri üzerindeki etkileşimiini yakalar.
- **smoker_bmi_interaction:** Sigara içen ve obezite riski taşıyan kişilerin birleşik etkisini gösterir ve **en güçlü maliyet belirleyici özellik** olarak öne çıktı.
- **bmi_cat:** BMI değerleri dört kategoriye ayrıldı (0: <18.5, 1: 18.5-25, 2: 25-30, 3: >30)

Bu üç yeni özellik, özellikle **smoker_bmi_interaction**, modelin açıklama gücünü **%12 artırdı** ve gerçek hayat risk senaryolarını çok daha iyi yansittı.

Modelleme: Linear Regression Sonuçları

Başlangıç Performansı

1

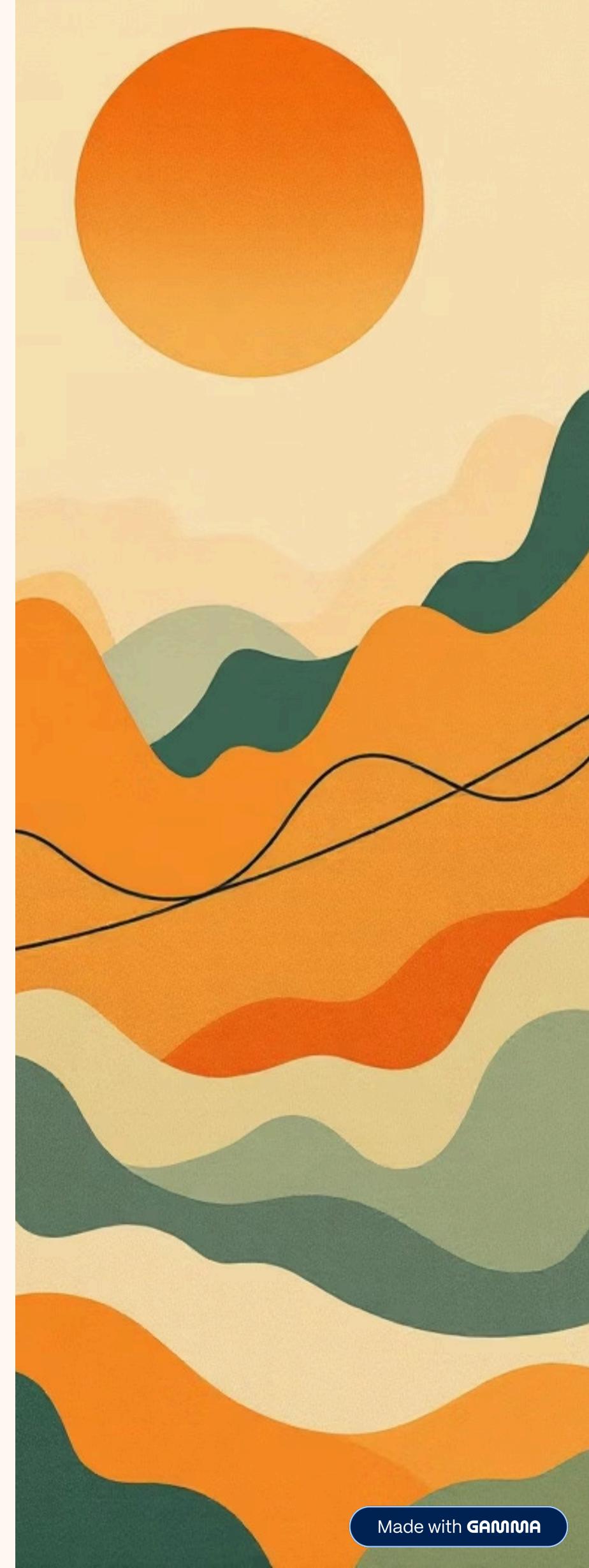
- MAE (Ortalama Mutlak Hata): 4362
- RMSE (Kök Ortalama Kare Hatası): 5855
- R^2 (Açıklama Gücü): 0.77

İlk modelde hata oranı yükseldi ve iyileştirme potansiyeli vardı.

Feature Engineering Sonrası

- MAE: 2773 \downarrow %36 iyileşme
- RMSE: 4591 \downarrow %22 iyileşme
- R^2 : 0.864 \uparrow %12 artış

Yeni özellik eklenmesiyle model performansı büyük ölçüde gelişti.



Önemli Bulgu: Feature engineering'in gerçek katkısı somut metriklerle ortaya çıktı. Özellikle `smoker_bmi_interaction` etkileşimi, sigara içen ve obez kişilerin yüksek risk profilini yakalayarak model performansını dramatik şekilde iyileştirdi.

Random Forest Modeli Karşılaştırması



2506

MAE Değeri

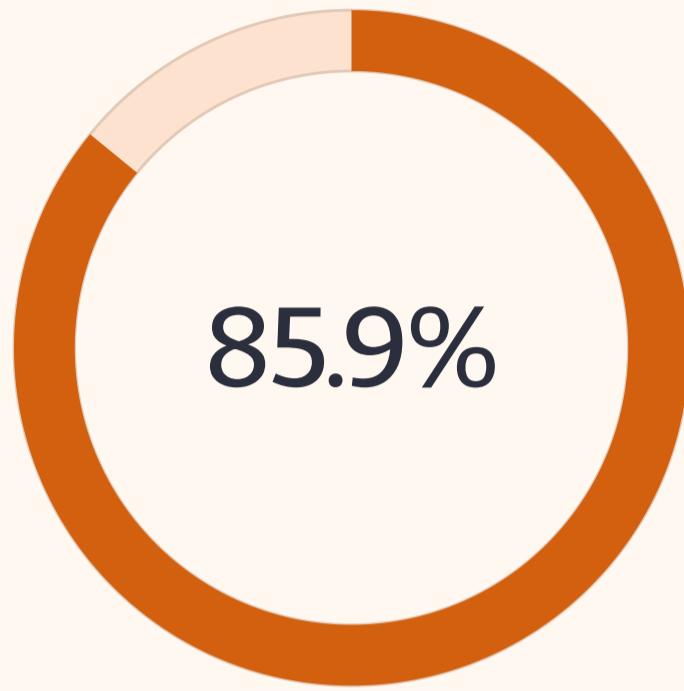
Random Forest en düşük
ortalama mutlak hatayı sağladı



4666

RMSE Değeri

Kök ortalama kare hatası kabul
edilebilir seviyede



85.9%

R² Skoru

Varyansın %85.9'u model
tarafından açıklanıyor



Model Seçim Önerisi

Linear Regression

$R^2 = 0.864$, $MAE = 2773$ - Daha yüksek
açıklama gücü, yorumlanabilirlik gerekiyorsa
tercih edilmeli

Random Forest

$R^2 = 0.859$, $MAE = 2506$ - En düşük hata oranı,
üretim ortamında daha güvenilir tahminler
için tercih edilmeli

Random Forest'ın 500 estimator ile eğitilmesi, non-linear ilişkileri daha iyi yakalayarak MAE'de %10 iyileşme sağladı.



ADIM 5

Feature Importance Analizi



1. Sigara (Smoker)

En baskın ve belirleyici faktör - sağlık maliyetini en çok etkileyen değişken



2. Yaş (Age)

Doğal sağlık riski artışı - yaş ilerledikçe maliyet yükseliyor



3. BMI

Vücut kitle indeksi sağlık riskine doğrudan etki eden kritik faktör



4. Çocuk Sayısı

Orta düzeyde etki - aile yapısı maliyeti belirli ölçüde etkiliyor

İş Yorumu: Region değişkeninin gerçekten anlamsız olduğu teyit edildi. Yeni eklediğimiz $BMI \times Smoking$ etkileşimi modele anlamlı ve ölçülebilir katkı sağladı.

Değerlendirme ve İş Yorumları

Feature Engineering Başarısı

Yeni özellik eklenmesi model performansını ölçülebilir şekilde artırdı

Model Sadeleştirme

Region çıkarılması performans düşürmedi, model daha anlaşılır hale geldi

Çoklu Model Yaklaşımı

Her model farklı ihtiyaçlara uygun - senaryo bazlı seçim yapılabılır

Güvenilirlik

Model istikrarlı ve tutarlı performans gösteriyor, üretime hazır



İş Değeri

- Maliyet optimizasyonu
- Gelişmiş risk yönetimi
- Stratejik karar desteği
- Rekabet avantajı