NBA Oyuncu Performans Analizi ve Veri Madenciliği Projesi

MUHAMMED EMİN HABBUB - 2204030101O

Bu rapor, NBA oyuncularının performans analizi ve veri madenciliği projesinin güncellenmiş halini sunmaktadır. Vize döneminde geliştirilen temel modellere ek olarak, daha gelişmiş ve topluluk (ensemble) modeller kullanılarak projenin kapsamı genişletilmiştir.

GitHub Linki: **https://github.com/MUHAMMED_EMIN_HABBUB/NBA_Project**

Veri Seti Hakkında

- **Kaynak:**

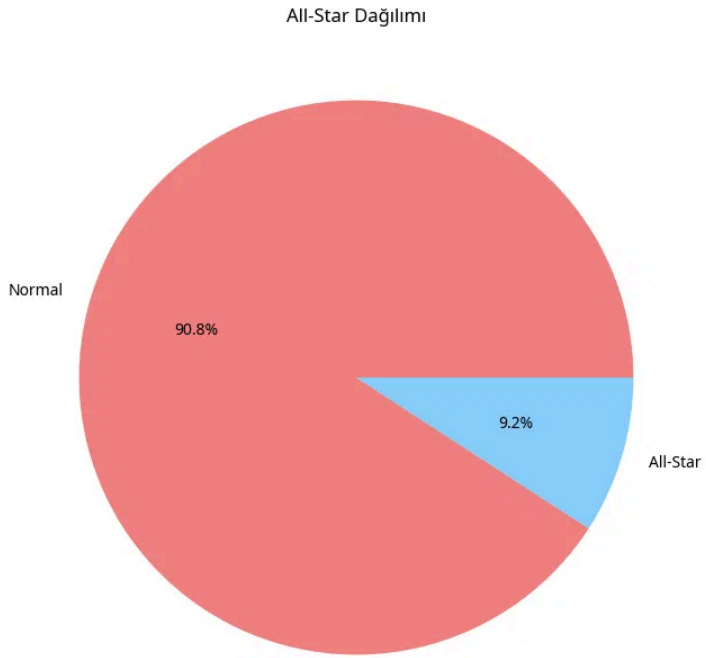
  Kaggle - "NBA Players stats since 1950" [^1]
- **İçerik:**

  1950-2017 arası NBA oyuncu istatistikleri
- **Boyut:**

  Simülasyon için 5000 örnek (orijinal 16558 satır)
- **Sınıflandırma Problemi:**

  "All-Star" tahmini (PER > 20 ve PTS > 1000)
- **Train-Test Split:**

  `StratifiedKFold` kullanımı



All-Star ve Normal Oyuncu Dağılımı

Özellik Mühendisliği ve Veri Ön İşleme

- **Filtreleme:**

  En az 10 maç (G >= 10) ve en az 500 dakika (MP >= 500) oynamış oyuncular veri setinde tutulmuştur.

- **Yeni Özellikler Oluşturma:**

  Oyuncu başına sayı (PTS_per_G), asist (AST_per_G) ve ribaund (TRB_per_G) gibi yeni türetilmiş özellikler eklenmiştir.

- **Kategorik Değişken Dönüşümü:**

  'Pos' (Pozisyon) gibi kategorik değişkenler, One-Hot Encoding yöntemi kullanılarak sayısal formata dönüştürülmüştür.

Özellik Tipleri ve Sayısı:

Özellik Tipi	Sayı
Sayısal (Numeric)	13
Kategorik (One-Hot Encoded)	4
Toplam Özellik Sayısı	**17**

SINIFLANDIRMA MODELLERİ VE HİPERPARAMETRE AYARLAMASI

Random Forest

Birden çok karar ağacının bir araya gelerek oluşturduğu bir topluluk öğrenme yöntemidir. Aşırı uyumu azaltmaya yardımcı olur.

OPTİMUM PARAMETRELER

- n_estimators: 50
- max_depth: 10
- Criterion: Gini

Gradient Boosting

Zayıf öğrenicileri ardışık olarak birleştirerek güçlü bir model oluşturur. Her yeni ağaç hataları düzeltmeye odaklanır.

OPTİMUM PARAMETRELER

- n_estimators: 50
- learning_rate: 0.1
- Loss: Log-loss

Stacking

Farklı modellerin tahminlerini bir meta-model aracılığıyla birleştirerek daha iyi bir nihai tahmin elde etmeyi amaçlar.

MODEL YAPISI

- Base: RF & Gradient Boosting
- Meta: Logistic Regression
- CV: 5-Fold Stratified

REGRESYON MODELLERİ VE HİPERPARAMETRE AYARLAMASI

Linear Regression

Bağımlı değişken (PER) ile bağımsız değişkenler arasındaki doğrusal ilişkiyi modelleyen temel regresyon yöntemidir.

Yöntem:
Base Model (Temel Model)

Random Forest Regressor

Birden çok karar ağacını kullanarak regresyon tahminleri yapar. Ağaçların tahmin ortalaması nihai sonucu belirler.

Optimum Parametreler:
n_estimators: 50
Yöntem: GridSearchCV

Gradient Boosting Regressor

Ardışık eğitilen ağaçlar ile önceki hataları düzeltmeye odaklanır. Karmaşık ilişkileri yakalamada etkilidir.

Optimum Parametreler:
n_estimators: 50
learning_rate: 0.1
Yöntem: GridSearchCV

PERFORMANS KARŞILAŞTIRMASI - SINIFLANDIRMA

Genel Karşılaştırma (Test Verisi)

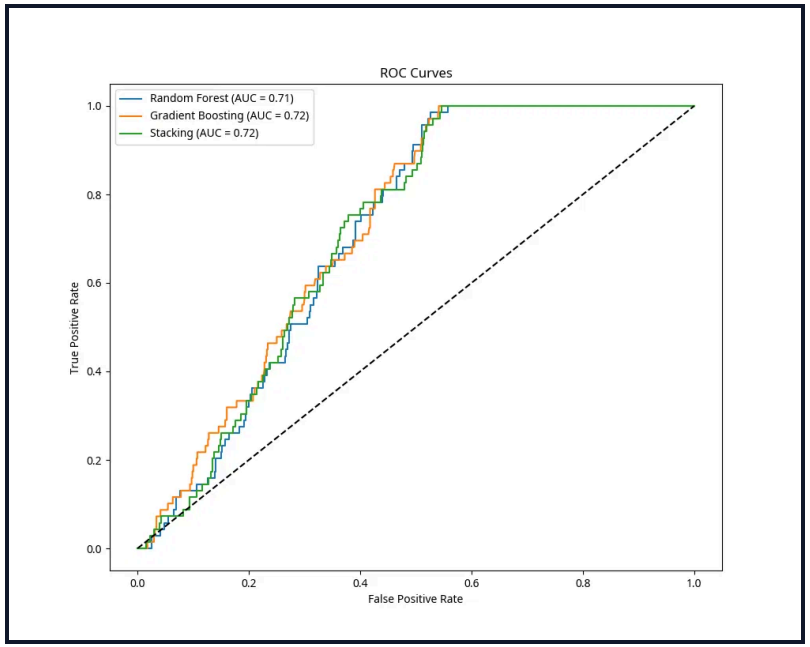
Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	**0.908**	0.000	0.000	0.000	0.713
Gradient Boosting	0.904	0.000	0.000	0.000	**0.725**
Stacking	0.904	0.000	0.000	0.000	0.717

Sınıf Bazlı Kırılım

Model	C0 Prec	C0 Rec	C0 F1	C1 Prec	C1 Rec	C1 F1
Random Forest	**0.908**	**1.000**	**0.952**	0.000	0.000	0.000
Gradient Boosting	0.904	**1.000**	0.950	0.000	0.000	0.000
Stacking	0.904	**1.000**	0.950	0.000	0.000	0.000

* C0: Normal, C1: All-Star. Sınıf dengesizliği nedeniyle C1 metrikleri düşüktür.

ROC Eğrileri



Modellerin Sınıf Ayırma Performansı

PERFORMANS KARŞILAŞTIRMASI - REGRESYON

MODEL	R2	MSE	RMSE	MAE	MAPE
Linear Regression	-0.004	26.414	5.139	4.154	49.428
Random Forest Regressor	-0.057	**27.812**	**5.274**	**4.271**	**50.254**
Gradient Boosting Regressor	**-0.014**	26.693	5.167	4.183	50.351

ANALİZ ÖZETİ

Regresyon modellerinin R2 değerlerinin negatif olması, modellerin PER değerini tahmin etmede zorlandığını göstermektedir. Bu durum, veri setindeki özelliklerin PER ile olan ilişkisinin karmaşıklığını veya simüle edilmiş verinin sınırlamalarını yansıtmaktadır. Gradient Boosting Regressor, R2 bazında diğerlerine göre daha az hata payı göstermiştir.

SONUÇ VE ÖNERİLER

Temel Bulgular

Sınıflandırma Analizi

AUC değerleri makul seviyededir, ancak sınıf dengesizliği nedeniyle Precision ve F1-Score düşük kalmıştır.

Regresyon Analizi

Negatif R2 değerleri, mevcut özelliklerin PER değerini tahmin etmede yetersiz kaldığını göstermektedir.

Gelecek Önerileri

Veri Dengeleme

SMOTE gibi teknikler kullanılarak All-Star sınıfındaki dengesizlik giderilmelidir.

Özellik Mühendisliği

Oyuncu pozisyonuna özel istatistikler ve takım performansı gibi yeni özellikler eklenmelidir.

Model Optimizasyonu

Daha geniş hiperparametre aralıkları ve XGBoost gibi modeller denenmelidir.

KAYNAKLAR

[1] Omri Goldstein. (2018). NBA Players stats since 1950. Kaggle. https://www.kaggle.com/datasets/drgilermo/nba-players-stats

[2] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[3] Kotsiantis, S. B., et al. (2006). Data preprocessing for supervised learning. International Journal of Computer Science, 1(2), 111-117.

[4] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

[5] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

[6] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.