Veri Madenciliği Proje Raporu

Öğrenci Adı: Hüseyin Canık-21360859064

Özet

Bu raporda, S&P / US,Canada(TSX),UK endekslerinin veri madenciliği teknikleri kullanılarak analiz edilmesi ve tahmin edilmesi hedeflenmiştir. Sınıflandırma ve kümeleme için oluşturulan modeller accuracy değerlerinde yetersiz kalmıştır. Veri seti 1990-2016 yılları arasındaki günlük endeks verilerini ve teknik göstergeleri içermektedir. Veri ön işleme adımları, özellik mühendisliği, tanımlayıcı analiz, modelleme (regresyon, zaman serisi modelleri) ve performans değerlendirme konuları ele alınmıştır. Sonuçlar, modellerin doğruluğu ve geleceğe yönelik tahminlerin geçerliliği üzerine tartışılmıştır.

İçindekiler

1. Giriş

Bu proje kapsamında, veri madenciliği teknikleri kullanılarak S&P/ US,Canada(TSX),UK endekslerinin geçmiş verileri üzerinde analizler yapılmış ve geleceğe yönelik tahmin modelleri geliştirilmiştir.

2. Veri Seti ve Ön İşleme

Kullanılan veri seti, 1990-2016 yılları arasındaki günlük S&P/US,Canada(TSX),UK endeks değerlerini içermektedir. Veri sütunları: Date, Price, Open, High, Low, Change %, SMA_10, EMA_10, RSI_14, MACD, MACD_Signal, MACD_Diff.
Ön işleme adımları şunlardır:

- Tarihsel eksik verilerin kontrolü ve gerekirse interpolasyon ile tamamlanması.
- Tarih kolonu datetime formatına dönüştürülmesi
-Teknik göstergelerin pandas kütüphanesi ile eklenmesi
- Teknik göstergelerin normalize edilmiş haliyle kullanılması.
- Eğitim ve test setlerinin oluşturulması (70/30)

3. Özellik Mühendisliği

Özellik mühendisliği aşamasında, teknik göstergelerden yeni öznitelikler türetilmiştir:
- Hareketli ortalamaların (SMA, EMA) farklı periyotlarda hesaplanması.
- RSI, MACD ve diğer momentum göstergelerinin belirli eşik değerlerine göre sınıflandırılması.
- Geriye dönük gecikmeli fiyat (lag) özelliklerinin oluşturulması.
- Özelliklerin normalize edilmesi (min-max, z-score vb.).

Us: Eğitim Seti Metrikleri:	Test Seti Metrikleri:
MAE: 0.0441	MAE: 0.0418
MSE: 0.0455	MSE: 0.0538
RMSE: 0.2133	RMSE: 0.2320
R2: 0.9557	R2: 0.9247
Adjusted R2: 0.9557	Adjusted R2: 0.9246

Canada: Eğitim Seti Metrikleri:	Test Seti Metrikleri:
MAE: 0.0373	MAE: 0.0341
MSE: 0.0027	MSE: 0.0016

RMSE: 0.0517

R2: 0.9860

Adjusted R2: 0.9860

RMSE: 0.0406

R2: 0.9977

Adjusted R2: 0.9977

UK: Eğitim Seti Metrikleri:

MAE: 0.1420

MSE: 0.1500

RMSE: 0.3874

R2: 0.8547

Adjusted R2: 0.8546

Test Seti Metrikleri:

MAE: 0.0163

MSE: 0.0003

RMSE: 0.0163

R2: -61970.3362

Adjusted R2: -62089.7415

4. Keşifsel Veri Analizi

Veri setinin tanımlayıcı istatistikleri incelenmiştir:
- Ortalama, medyan, standart sapma, en düşük ve en yüksek değerler.
- Zaman serisi grafikleri ve mevsimsellik analizleri.
- Aykırı değerlerin tespiti ve analizi
-Nan değerlere yerine göre ortalama ve ilk değer ataması
-Eklenen sinyallerin ve diğer sütunların StandardScaler yöntemi ile normalize edilmesi

5. Modelleme Yaklaşımları

Aşağıda kullanılan model özetlenmiştir:
5.1 Derin Öğrenme Modelleri (LSTM) - Zaman serisi verisinin LSTM ağlarına uygun hale getirilmesi için kaydırılmış pencere (windowing) yöntemi uygulanmıştır.

6. Sonuçlar ve Tartışma

Modellerin performans karşılaştırması aşağıdaki tabloda gösterilmiştir:
Model	MAE	MSE	RMSE	R^2
LSTM	...	...	...	...

Performans açısından LSTM tabanlı(regression) modellerin daha iyi sonuç verdiği gözlemlenmiştir.

7. Sonuç ve Öneriler

Bu çalışmalarda S&P / US,Canada(TSX),UK endeksleri üzerinde veri madenciliği teknikleri uygulanmıştır. Elde edilen sonuçlar, çeşitli modellerin tahmin başarımlarını karşılaştırmaya imkan tanımıştır. Gelecek çalışmalarda veri setine ek ekonomik göstergelerin dahil edilmesi ve derin öğrenme modellerinin derinleştirilmesi önerilmektedir.

Youtube Link:

- https://youtu.be/TrSYM-fVQUI

GitHub Link:

- https://github.com/Huseyincanik/DataMining-Daily-Stock-Index

8. Kaynakça

1. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Elsevier.
2. Chollet, F. (2017). Deep Learning with Python. Manning Publications.
3. Granger, C.W.J., & Newbold, P. (1974). Spurious regressions in econometrics. Journal of Econometrics.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.
5. Brownlee, J. (2018). Long Short-Term Memory Networks With Python. Machine Learning Mastery.