Analiza i predviđanje vremenskih serija S&P 500 indeksa primenom metoda mašinskog učenja

Miloš Trišić RA39/2023

1. Opis problema

Jedan od najvažnijih izazova u finansijama i investicijama je razumevanje kretanja tržišta i identifikacija obrazaca u složenim vremenskim serijama. Problem ispravnog razumevanja tržišta leži u njegovoj složenosti i nestablinosti, koje nastaje usled velikog broja faktora koji istovremeno utiču na tržište. Razumevanje strukture, dinamike i mogućnosti predikcije ovakvih podataka je od velikog značaja kako u akademskom, tako i u industrijskom svetu.

S&P 500 indeks (Standard & Poor's 500) je berzanski indeks koji predstavlja zbirnu tržišnu vrednost 500 najvećih američkih kompanija po tržišnoj kapitalizaciji, a mnogi ekonomisti ga vide kao glavni pokazatelj stanja američkog tržišta. Indeks je kapitalizacijski ponderisan, što znači da kompanije sa većom tržišnom vrednošću (kao što su Apple, Microsoft, Google, NVIDIA..) imaju veći uticaj na promene koje se dešavaju nad indeksom. Pravilno razumevanje trendova i tržišta koje ovaj indeks opisuje pomaže investitorima i brokerima da vrše informisano ulaganje u akcije i da bolje procene rizik pre samog ulaganja.

2. Ciljevi projekta

Osnovni cilj ovog projekta je ispitati u kojoj meri se klasični modeli vremenskih serija i savremeni modeli mašinskog učenja mogu iskoristiti za analizu i predikciju vrednosti S&P 500 indeksa, kao i da se identifikuju njihove prednosti, mane i ograničenja.

- Analiza osnovnih karakteristika vremenske serije
- Predikcija budućih vrednosti indeksa pomoću nekoliko modela (Linearna Regresija, ARIMA, XGBoost) i poređenje njihovih performansi
- Proširenje analize uključivanjem dnevnih cena najvećih kompanija unutar indeksa (radi redukcije dimenzionalnosti i identifikacije glavnih komponenti koje objašnjavaju varijansu tržišta PCA) - opciono

3. Skup podataka

Skup podataka koje ćemo analizirati i nad kojim ćemo trenirati modele je javno dostupan dataset „S&P 500 Stocks (daily updated)" preuzet sa platforme Kaggle. Kao osnovni skup podataka koristimo tabelu sp500_index.csv koja sadrži istorijske dnevne vrednosti indeksa.

Podaci tabele sp500_index.csv obuhvataju sledeće atribute :

- Date – datum trgovine
- S&P 500 – vrednost indeksa na određen datum

Opciono – koristićemo dnevne cene akcija najvećih kompanija, što omogućava PCA analizu i dodatno objašnjenje varijanse tržišta (podaci iz tabele sp500_stocks.csv).

Podaci tabele sp500_ stocks.csv obuhvataju sledeće atribute :

- Date – datum trgovanja
- Symbol – oznaka kompanije
- Close – vrednost akcije na kraju dana
- Adj Close – prilagođena zatvarajuća akcija
- High – najviša vrednost akcije dana
- Low – najniža vrednost akcije dana
- Open – početna cena dana
- Volume – broj prodatih akcija

4. Metodologija

Projekat će obuhvatiti sledeće korake :

- Pretprocesiranje podataka – čišćenje i formatiranje podataka, otklanjanje nedostajućih vrednosti
- Vizuelizacija i analiza podataka (EDA)
- Dekompozicija vremenske serije
- PCA analiza (opciono)
- Modelovanje i predikcija (Linearna Regresija, ARIMA, XGBoost)

5. Način evaluacije

Evaluacija modela sprovodi se korišćenjem konzistentne podele podataka (time-based split), a kao metrike koristićemo srednju apsolutnu grešku (MAE) i koren srednje kvadratne greške (RMSE).

6. Tehnologije

Python, Jupyter Notebook
Biblioteke : Pandas, NumPy, scikit-learn, statsmodels, XGBoost, Matplotlib, Seaborn

7. Primeri gotovih rešenja i korišćeni materijali pri izradi rada

https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks