Analiza i predikcija kratkoročnih logaritamskih prinosa S&P 500 indeksa primenom modela vremenskih serija

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, koja 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 bolje procenjuju rizik i donose informisane odluke.

2. Ciljevi projekta

Osnovni cilj ovog projekta je ispitati u kojoj meri i na koji način se savremeni modeli vremenskih serija mogu iskoristiti za analizu i predikciju kratkoročnih logaritamskih prinosa S&P 500 indeksa, uz identifikaciju njihovih prednosti, mana i ograničenja.

- Analiza i identifikacija osnovnih karakteristika vremenske serije u svrhu razumevanja nejene strukture i ponašanja :
  1. Trend – dugoročno kretanje serije (tendencija rasta ili pada)
  2. Sezonalnost – periodični obrasci koji se ponavljaju u fiksnim intervalima (npr. mesečno, godišnje)
  3. Cikličnost - obrasci koji se ponavljaju u neregularnim intervalima (krize)
  4. Šum – nasumične i nepredvidive promene u seriji
- Predikcija logaritamskih prinosa pomoću ARIMA i Facebook Prophet modela
- Modelovanje volatilnosti pomoću GARCH modela
- Deskriptivna PCA analiza u svrhu identifikacije kompanija i sektora koji najviše doprinose varijansi tržišta

Konkretna istraživačka pitanja :

1. Da li vremenska serija S&P 500 indeksa pokazuje stabilne trendove i obrasce u posmatranom periodu?
2. Kako se ti pokazatelji ponašaju u različitim tržišnim periodima (krizni period i period stabilnog rasta)?
3. Koji model vremenskih serija (ARIMA, Prophet) preciznije predviđa log_returns?
4. Na koji način GARCH model opisuje volatilnost (rizik)?
5. Koje kompanije i sektori najviše doprinose varijansi tržišta i kako se taj doprinos menja tokom vremena?

3. Definisanje cilja predikcije

Umesto direktne predikcije vrednosti S&P 500 indeksa, u ovom radu se modeluju logaritamski prinosi (log returns), jer poseduju povoljnija statistička svojstva i omogućavaju primenu standardnih modela vremenskih serija. Predikciju vršimo za kratkoročni vremenski interval od 1-2 trgovinska dana što odgovara praktičnim potrebama analize tržišnog ponašanja. Ovaj izbor horizonta usklađen je sa ciljevima projekta, jer modeli poput ARIMA i Prophet pokazuju najbolje performanse u kratkoročnim predikcijama dok se za duže intervale neizvesnost značajno povećava.

4. Skup podataka

Skup podataka koje ćemo analizirati je javno dostupan dataset „S&P 500 Stocks (daily updated)“ preuzet sa platforme Kaggle. Skup podataka sastoji se iz 3 pojedinačne tabele čiji nazivi kolona su specificirani u nastavku. Podatke iz tabele „sp500_index.csv“ koristićemo za predikciju vrednosti indeksa, odnosno logaritamskih prinosa.

Podaci tabele sp500_index.csv obuhvataju sledeće atribute :

- Date – datum trgovine (primetiti da nisu svi pokriveni svi kalendarski datumi)
- S&P 500 – vrednost indeksa na određen datum trgovanja

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

Podaci tabele sp500_companies.csv obuhvataju sledeće atribute : Exchange, Symbol, Shortname, Longname, Sector, Industry, Currentprice, Marketcap, Ebitda, Revenuegrowth, City, State, Country, Fulltimeemployees, Longbusinesssummary, Weight

5. Metodologija

Projekat će obuhvatiti sledeće korake :

- Pretprocesiranje podataka – podela na trening, validacioni i test skup po vremenskom redosledu, otklanjanje nedostajućih vrednosti (u tabelama u kojima postoje), izračunavanje logaritamskih prinosa (*log-returns*) indeksa.
- Vizuelizacija i analiza podataka (*Exploratory Data Analysis - EDA*) – vizuelizacija log returns kroz vreme, dekompozicija serije
- Testiranje uslova stacionarnosti (ADF test)
- Primena ARIMA modela za predikciju logaritamskog prinosa (odabir parametara kroz ACF/PACF)
- Primena Prophet modela za predikciju logaritamskih prinosa (koristi nelinearne obrasce)
- Primena GARCH modela za predikciju volatilnosti logaritamskog prinosa
- Deskriptivna PCA analiza glavnih komponenti varijanse (u ovom radu se ne koristi za poboljšavanje performansi prediktivnih modela, već kao alat za interpretaciju tržišne strukture)
- Analiza i interpretacija rezultata

6. Način evaluacije

Evaluacija modela sprovodi se korišćenjem konzistentne podele podataka (time-based split), a kao metrike koristićemo srednju apsolutnu grešku (MAE), koren srednje kvadratne greške (RMSE) i srednju apsolutnu procentualnu grešku (MAPE). Takođe, prikazaćemo i interval predikcije za predikcije ARIMA, Prophet i GARCH modela.

7. Tehnologije

Python, Jupyter Notebook

Biblioteke : Pandas, NumPy, Scikit-learn, Statsmodels, Matplotlib, Seaborn, Prophet, arch

8. Primeri gotovih rešenja i korišćeni materijali pri izradi rada

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

Facebook's Prophet - https://facebook.github.io/prophet/docs/quick_start.html