

# Predviđanje cene akcija pomoću rekurentnih neuronskih mreža

**Student:** Stefan Djurica, SV35/2021

## 1 Eksplorativna analiza podataka

Za potrebe ovog projekta, podaci o cenama akcija su prikupljeni dinamički korišćenjem Python biblioteke yfinance.

Odabrana je akcija kompanije Apple (simbol: AAPL) zbog njene visoke likvidnosti i dostupnosti obimnih istorijskih podataka.

Prikupljen je vremenski period od **1. januara 2000. do 1. januara 2025. godine.**

### 1.1 Prikaz osnovnih statistika

Skup podataka sadrži dnevne vrednosti cena (otvaranje, najviša, najniža, zatvaranje) i obim trgovine.

Osnovne statističke metrike za prikupljene podatke su prikazane u tabeli ispod:

Metrika	Close (USD)	High (USD)	Low (USD)	Open (USD)	Volume
Broj unosa	6289	6289	6289	6289	6.29E+03
Srednja vr.	41.15	41.55	40.7	41.11	3.86E+08
Std. dev.	60.02	60.58	59.38	59.95	3.84E+08
Min	0.197	0.198	0.191	0.195	2.32E+07
25%	2.01	2.04	1.97	2.01	1.16E+08
50%	13.8	13.94	13.68	13.83	2.68E+08
75%	44.07	44.41	43.74	44.03	5.18E+08
Max	258.1	259.18	256.72	257.28	7.42E+09

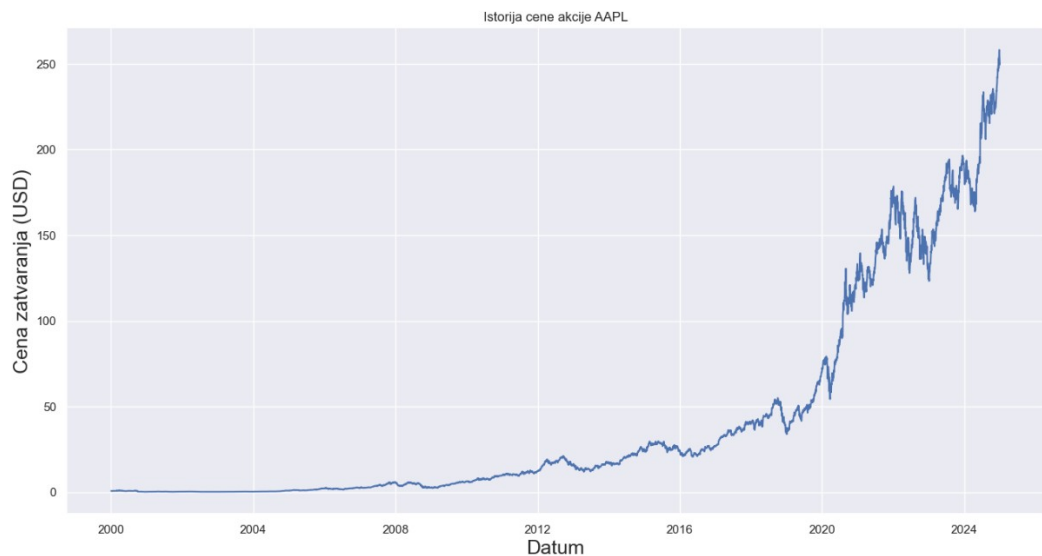
*Tabela 1. prikaz informacija dataset-a*

Iz tabele se može uočiti širok raspon vrednosti cena, što ukazuje na značajan rast vrednosti akcije tokom posmatranog perioda, ali i na visoku volatilnost.

### 1.2 Vizuelizacije

- **Istorijsko kretanje cene zatvaranja (Close)**

Grafik ispod prikazuje kretanje cene zatvaranja akcije AAPL. Jasan je dugoročni uzlazni trend, ispresecan periodima naglih padova i korekcija, što je karakteristično za tržište kapitala.



Slika 1. Istorija cene akcija AAPL-a

- **Korelaciona matrica**

Analiza korelacije između različitih obeležja je ključna za odabir ulaznih promenljivih za model. Toplotna mapa (heatmap) ispod vizuelizuje ove odnose.



Slika 2. Heatmap

Uočava se da su cene otvaranja (Open), najviša (High) i najniža (Low) cena u skoro savršenoj pozitivnoj korelaciji ( $\approx 1.00$ ) sa cenom zatvaranja (Close). Njihovo uključivanje u model bi unelo redundantnost i overfitting. S druge strane, obim trgovine (Volume) ima umerenu negativnu korelaciju (-0.46) sa cenom, što ga čini dobrim kandidatom za dodatno obeležje jer pruža drugačiji tip informacije koji nije direktno sadržan u kretanju cene.

### 1.3 Uočeni neobični obrasci i anomalije

Podaci preuzeti putem yfinance API-ja su se pokazali kao veoma čisti, bez nedostajućih vrednosti.

Glavna karakteristika koja otežava modelovanje je **nestacionarnost** vremenske serije, gde se statistička svojstva (kao što su srednja vrednost i varijansa) menjaju tokom vremena.

## 2 Preprocesiranje podataka

Faza preprocesiranja je bila ključna za pripremu podataka za LSTM mrežu.

- **Odabir obeležja (Feature Selection):** Na osnovu eksplorativne analize, za ulazna obeležja modela odabrane su kolone Close i Volume.
- **Skaliranje podataka (Data Scaling):** Vrednosti odabranih obeležja su skalirane na opseg [0, 1] korišćenjem MinMaxScaler iz scikit-learn biblioteke. Ovo je standardna procedura za LSTM mreže jer pomaže u stabilizaciji i ubrzavanju procesa treniranja.
- **Kreiranje sekvenci (Sequence Creation):** Podaci su transformisani u sekvencijalni format pogodan za LSTM. Korišćen je pristup "pokretnog prozora" (sliding window) dužine 60 dana. Za svaki dan  $t$ , model dobija na ulaz niz podataka (Close, Volume) od dana  $t-60$  do  $t-1$  i ima zadatak da predvidi Close cenu za dan  $t$ . Ovim je ulazni podatak za model postao trodimenzionalni niz oblika (broj\_uzoraka, 60, 2).

## 3 Podela skupa podataka

Za probleme sa vremenskim serijama, ključno je održati hronološki redosled podataka. Zbog toga je skup podeljen na **trening (prvih 80%)** i **test (poslednjih 20%)** skup.

Ovaj pristup verno simulira realnu situaciju gde se model trenira na prošlim podacima kako bi predvideo buduće vrednosti koje nikada nije video.

## 4 Isprobani algoritmi

## 4.1 Odabir algoritma

Odabrana je **LSTM (Long Short-Term Memory)** arhitektura, koja je specijalizovana varijanta rekurentnih neuronskih mreža (RNN). LSTM je idealan za ovaj problem jer je dizajniran da prepoznaje i uči dugoročne zavisnosti u sekvencijalnim podacima, što je osnovna karakteristika kretanja cena akcija. Osnovna arhitektura modela se sastojala od dva LSTM sloja, praćena Dropout slojevima za regularizaciju i jednim Dense izlaznim slojem.

## 4.2 Podešavanje hiperparametara

Da bi se pronašla optimalna konfiguracija modela, korišćena je biblioteka KerasTuner sa RandomSearch strategijom. Proces je tekao u dve faze:

1. **Prva faza:** Široka pretraga je izvršena za sledeće hiperparametre:

- Broj jedinica (units) u LSTM slojevima: [32, 128]
- Stopa odustajanja (dropout): [0.1, 0.4]
- Stopa učenja (learning\_rate): [0.01, 0.001, 0.0001]

2. **Druga faza:** Na osnovu rezultata prve faze, fiksirani su dropout na 0.1 i learning\_rate na 0.001, a opseg pretrage za broj jedinica je proširen:

- Broj jedinica (units): [128, 256]

Ova fokusirana pretraga je kao optimalnu vrednost identifikovala **224 jedinice**.

## 4.3 Ostvareni rezultati

Proces optimizacije je doneo značajno poboljšanje performansi, što se vidi iz uporednog prikaza **Root Mean Squared Error (RMSE)** metrike na test skupu:

Model	RMSE (USD)	Poboljšanje
Osnovni model (pre tjuniranja)	~\$33	-
Nakon 1. faze tjuniranja	\$17.73	~46%
Nakon 2. faze tjuniranja	**\$14.45**	~56%

Treba imati u vidu da je proces treniranja neuronskih mreža stohastičke prirode, pre svega zbog nasumične inicijalizacije težina i primene Dropout regularizacije. Usled toga, rezultati mogu neznatno varirati prilikom svakog novog pokretanja treninga. Tokom višestrukih testiranja konačnog modela, primećeno je da RMSE vrednost na test skupu **fluktuiru u opsegu od približno \$11 do \$15**. Stoga, navedeni rezultat od \$14.45 treba posmatrati kao reprezentativnu vrednost performansi, dok stvarna preciznost modela leži unutar pomenutog opsega."

## 5 Odabrano rešenje

### 5.1 Opis konačnog rešenja

Konačno rešenje je LSTM model sa hiperparametrima dobijenim nakon druge faze optimizacije:

- **Arhitektura:**
  - Ulazni sloj oblika (60, 2)
  - Prvi LSTM sloj sa **224 jedinice** (return\_sequences=True)
  - Dropout sloj sa stopom **0.1**
  - Drugi LSTM sloj sa **224 jedinice**
  - Dropout sloj sa stopom **0.1**
  - Izlazni Dense sloj sa 1 jedinicom
- **Optimizator:** Adam sa stopom učenja **0.001**
- **Funkcija gubitka:** mean\_squared\_error
- **Trening:** Za sprečavanje preprilagođavanja (overfitting), korišćen je EarlyStopping callback koji prati grešku na validacionom skupu (val\_loss) sa patience=6.

### 5.2 Obrazloženje odluke

Ovo rešenje je odabrano jer je, kroz sistematski proces podešavanja hiperparametara, pokazalo najbolje performanse na test skupu, postigavši najniži RMSE.

Finalni rezultati evaluacije na test skupu su sledeći:

- **RMSE:** \$14.45
- **MAE:** \$11.02 (prosečna apsolutna greška)

Vizuelizacija rezultata pokazuje da model uspešno prati opšti trend stvarne cene, ali, kao što se i očekivalo, teže predviđa nagle i nepredviđene skokove i padove.

Prikaz greške na poslednjih nekoliko dana trgovanja pokazuje da predviđanja odstupaju u rasponu od -12% do -16%, što ukazuje na to da model ima tendenciju da potcenjuje cenu tokom perioda snažnog rasta.

Dodatno je primetno da predikcija poprilično dobro prati pravi trend cena što je pogotovo vidno nad više od pola testnih podataka gde model ima mnogo manju grešku od -2% do -5%.

## Zaključak

Projekat je uspešno realizovan. Razvijen je i optimizovan LSTM model koji je u stanju da sa zadovoljavajućom tačnošću predvidi cenu akcije za naredni dan. Rezultati potvrđuju početnu pretpostavku da je prediktivna moć inherentno ograničena oslanjanjem isključivo na istorijske numeričke podatke, ali i demonstriraju snagu rekurentnih neuronskih mreža u analizi vremenskih serija.

## Reference

- Google Gemini
- Udemy
- Uvod u LSTM mreže : [link](#)
- TensorFlow/Keras: [link](#)
- Scikit-learn: [link](#)
- Pandas: [link](#)
- yfinance: [link](#)