Predviđanje cene akcija pomoću rekurentnih neuronskih mreža

Student: Stefan Djurica, SV35/2021

1 Eksplorativna analiza podataka

Podaci o cenama akcija za ovaj projekat prikupljeni su 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

Inicijalni skup podataka sadrži dnevne vrednosti cena (otvaranje, najviša, najniža, zatvaranje) i obim trgovine. Osnovne statistike prikazane su u Tabeli 1.

Metrika	Close (USD)	High (USD)	Low (USD)	Open (USD)	Volume
Broj unosa	6289	6289	6289	6289	6289
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
Max	258.1	259.18	256.72	257.28	7.42E+09

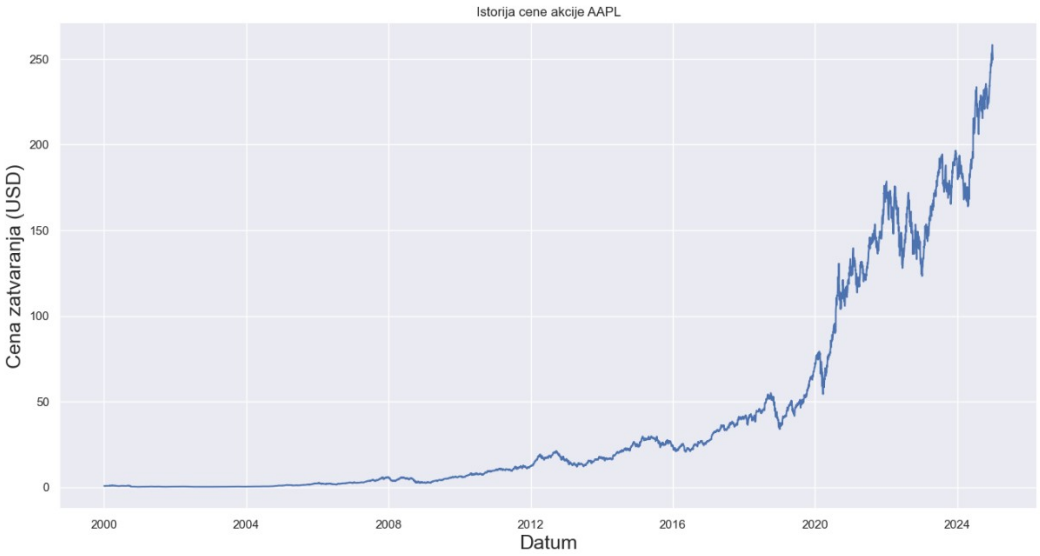
Tabela 1. Prikaz osnovnih statistika za sirovi skup podataka

Iz tabele je uočljiv širok raspon vrednosti, što ukazuje na značajan rast cene tokom posmatranog perioda.

1.2 Vizuelizacije

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

  Grafik (Slika 1) prikazuje jasan dugoročni uzlazni trend cene akcije, ispresecan periodima visoke volatilnosti, što je karakteristično za tržište kapitala.



Slika 1. Istorija cene akcija AAPL-a

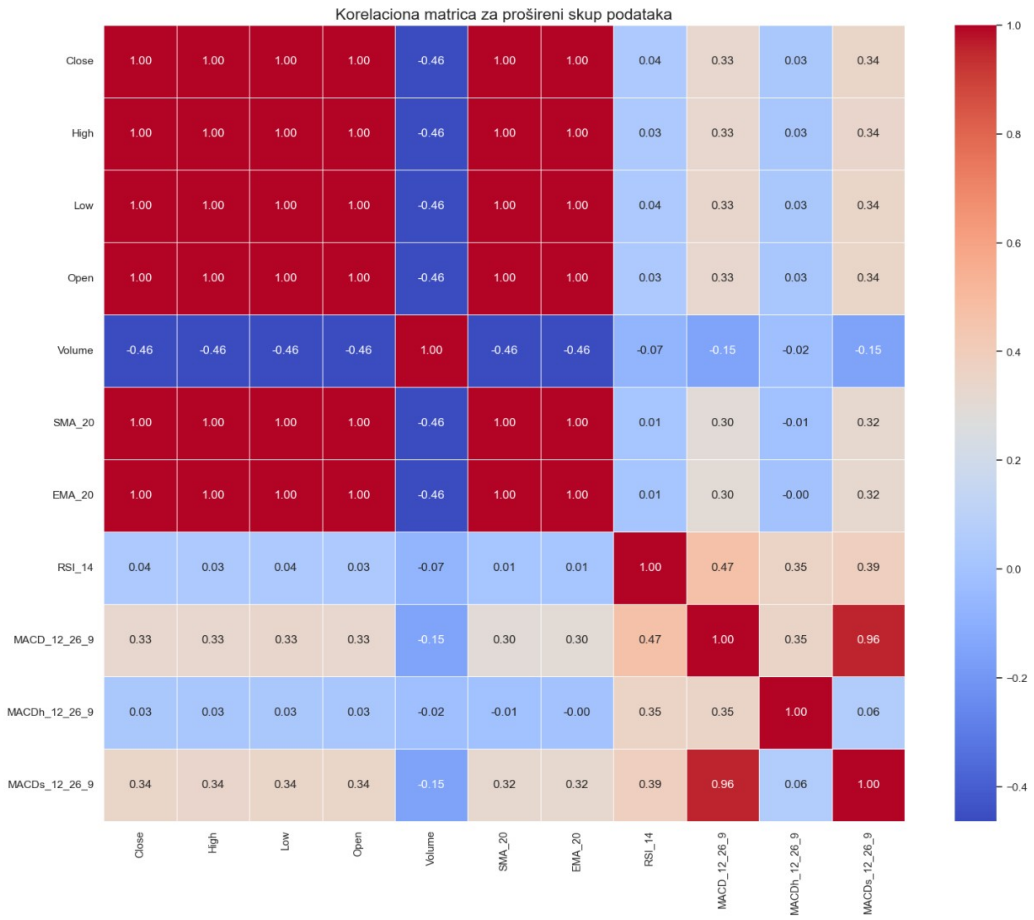
- **Analiza korelacije i odabir obeležja**

  Kako bi se obogatio skup podataka i modelu pružilo više konteksta, izvršen je inženjering obeležja (feature engineering). Pored osnovnih kolona, dodati su tehnički indikatori i kategorijska obeležja:

  - **Numerička obeležja:** SMA, EMA, RSI i MACD (sa histogramom i signalnom linijom).

  - **Kategorijska obeležja:** Dan u nedelji (Day_Name) i dnevni trend (Daily_Trend - Rast/Pad).

  Nakon toga, izvršena je analiza korelacije (Slika 2) kako bi se identifikovala redundantna obeležja.



Slika 2. Heatmap

Kao što se i očekivalo, cene Open, High, Low, kao i pokretni proseci SMA i EMA, pokazale su skoro savršenu korelaciju sa Close cenom. Da bi se smanjila redundantnost i potencijalni problemi sa multikolinearnošću, odlučeno je da se ova obeležja **ne koriste** u finalnom modelu.

1.3 Analiza korisnosti kategorijskih obeležja

Za procenu korisnosti kategorijskih obeležja korišćen je **Chi-Squared (χ^2) test nezavisnosti**. Ovi rezultati su vidni na slici 3.

- **Day_Name**: Test je pokazao **statistički značajnu vezu** (p-vrednost = 0.0003) između dana u nedelji i ishoda (Rast/Pad). Ovo ukazuje da je Day_Name korisno obeležje.

- **Daily_Trend**: Test zavisnosti između jučerašnjeg i današnjeg trenda **nije pokazao statističku značajnost** (p-vrednost = 0.1861). Ipak, odlučeno je da se ovo obeležje zadrži jer LSTM model, za razliku od prostog testa, može da uoči složenije sekvencijalne obrasce u nizu od 60 dana.

```
--- TEST #1: Chi-Squared Test Nezavisnosti za 'Daily_Trend' ---
Tabela kontingencije (uočene frekvencije):
Daily_Trend          Pad  Rast
Previous_Day_Trend
Pad                 1396  1587
Rast                1587  1685

Chi-Squared test za 'Daily_Trend':
Chi2 statistika: 1.7484
P-vrednost: 0.1861
Zaključak: P-vrednost je veća od 0.05. Nema dokaza o zavisnosti. Kolona možda nije korisna.

--- Chi-Squared Test Nezavisnosti za 'Day_Name' ---
Tabela kontingencije (uočene frekvencije):
Daily_Trend  Pad  Rast
Day_Name
Ponedeljak   502   671
Utorak       615   668
Sreda        600   683
Četvrtak     617   644
Petak        650   606

Chi-Squared test za 'Day_Name' vs 'Daily_Trend':
Chi2 statistika: 20.8128
P-vrednost: 0.0003
Zaključak: P-vrednost je manja od 0.05. Postoji statistički značajna veza između dana u nedelji i ishoda (Rast/Pad). Kolona 'Day_Name' je korisna.
```

Slika 3. Chi-Squared (χ^2) test nezavisnosti

2 Preprocesiranje podataka

Faza preprocesiranja je obuhvatila sledeće korake:

- **Inženjering i odabir obeležja:** Na osnovu analize, za finalni model odabrane su sledeće kolone: Close, Volume, sve tri MACD kolone, Day_Name i Daily_Trend.

- **Enkodiranje kategorijskih obeležja:** Tekstualne kolone Day_Name i Daily_Trend su pretvorene u numerički format pomoću **One-Hot enkodiranja** (pandas.get_dummies), čime je ukupan broj ulaznih obeležja porastao na 10.

- **Skaliranje podataka:** Sve vrednosti odabranih 10 obeležja su skalirane na opseg [0, 1] korišćenjem MinMaxScaler iz scikit-learn biblioteke. Skaler je fitovan isključivo na trening skupu.

- **Kreiranje sekvenci:** Podaci su transformisani u sekvencijalni format korišćenjem pristupa "pokretnog prozora" dužine 60 dana. Ulazni podatak za model je postao trodimenzionalni niz oblika (broj_uzoraka, 60, 10).

3 Podela skupa podataka

U skladu sa najboljim praksama za vremenske serije, skup podataka je hronološki podeljen na **trening (prvih 80%)** i **test (poslednjih 20%)** skup. Ovim se osigurava da se model evaluira na podacima koje nikada nije video.

4 Isprobani algoritmi

4.1 Odabir algoritma

Odabrana je **LSTM (Long Short-Term Memory)** arhitektura zbog njene dokazane sposobnosti da uči dugoročne zavisnosti u sekvencijalnim podacima, što je idealno za problem predviđanja cena akcija.

4.2 Podešavanje hiperparametara

Korišćena je biblioteka KerasTuner sa RandomSearch strategijom. Nakon inicijalne pretrage koja je ukazala da veći broj neurona daje bolje rezultate, izvršena je druga, fokusiranija runda pretrage. Konačni, optimizovani hiperparametri su:

- **Broj jedinica (units) u LSTM slojevima:** 384
- **Stopa odustajanja (dropout):** 0.1
- **Stopa učenja (learning_rate):** 0.0001
- **Trening:** EarlyStopping sa patience=10

4.3 Ostvareni rezultati

Dodavanjem novih obeležja i finim podešavanjem hiperparametara, model je postigao značajno bolje i, što je važnije, metodološki ispravnije rezultate u poređenju sa početnim, jednostavnijim modelom.

- **RMSE (Root Mean Squared Error):** Konačni model je na test skupu postigao RMSE vrednost od **$18.02**.

5 Odabrano rešenje

5.1 Opis konačnog rešenja

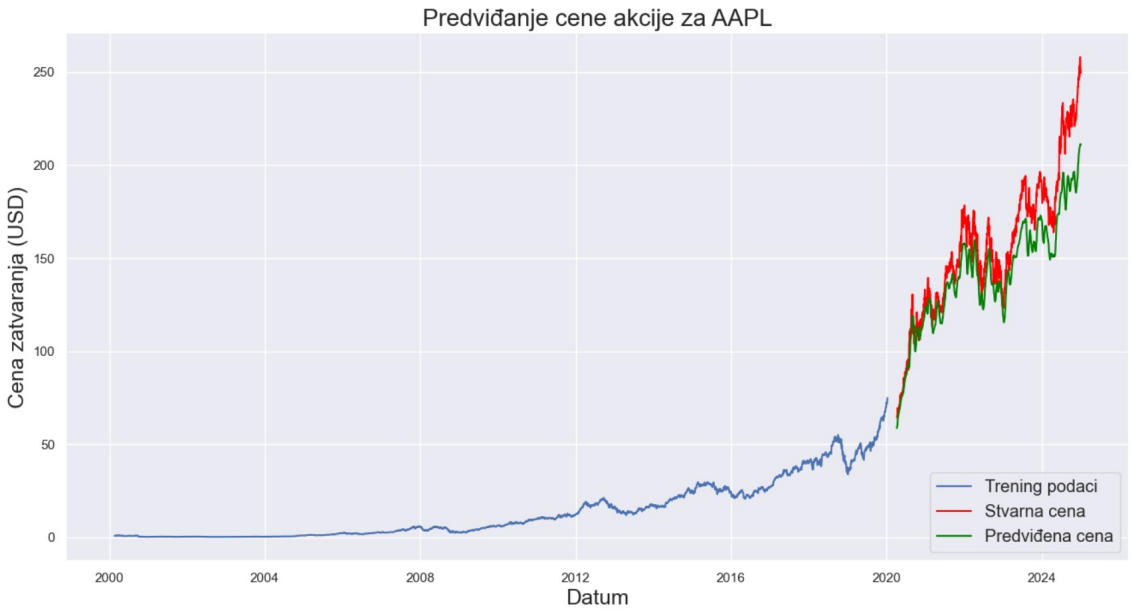
Konačno rešenje je LSTM model sa sledećom optimizovanom konfiguracijom:

- Arhitektura:
  - Ulazni sloj oblika (60, 10)
  - Prvi LSTM sloj sa **384 jedinice** (return_sequences=True)
  - Dropout sloj sa stopom **0.1**
  - Drugi LSTM sloj sa **384 jedinice**
  - Dropout sloj sa stopom **0.1**
  - Izlazni Dense sloj sa 1 jedinicom
- **Optimizator:** Adam sa stopom učenja **0.0001**
- **Funkcija gubitka:** mean_squared_error
- **Trening:** EarlyStopping callback sa patience=10.

5.2 Obrazloženje odluke

Ovo rešenje je odabrano jer predstavlja najbolju konfiguraciju pronađenu kroz sistematski proces inženjeringa obeležja i optimizacije hiperparametara.

Vizuelizacija rezultata (Slika 4) pokazuje da model uspešno prati opšti trend stvarne cene. Zelena linija (predikcije) dobro prati oblik crvene linije (stvarne cene), što potvrđuje da je model naučio opštu dinamiku tržišta.



Slika 4. Vizuelizacija rezultata

Međutim, analiza takođe otkriva i inherentna ograničenja modela:

- **Kašnjenje (Lag):** Model je reaktivan i njegova predviđanja blago kasne u odnosu na stvarne promene.
- **Potcenjivanje volatilnosti:** Predviđanja su "mirnija" i ne uspevaju da dostignu ekstremne vrhove i padove stvarne cene.
- **Greška pri snažnom trendu:** Tabela ispod pokazuje da tokom perioda snažnog rasta na kraju 2024. godine, model konstantno potcenjuje cenu, sa greškom koja se kreće od **-15% do -18%**. Ovo pokazuje da model teže predviđa kretanja koja značajno odstupaju od istorijskih proseka.

Datum	Stvarna cena	Predviđena cena	Procentualna razlika
2024-12-24	$257.29	$210.35	-18.24%
2024-12-26	$258.10	$210.79	-18.33%
2024-12-27	$254.69	$211.24	-17.06%
2024-12-30	$251.31	$211.32	-15.91%
2024-12-31	$249.53	$211.02	-15.44%

Zaključak

Projekat je uspešno realizovan. Razvijen je i optimizovan LSTM model koji je, dodavanjem relevantnih tehničkih i kategorijskih obeležja, u stanju da sa zadovoljavajućom tačnošću predvidi cenu akcije za naredni dan. Rezultati, sa finalnim RMSE od **$18.02**, potvrđuju početnu pretpostavku da je prediktivna moć modela ograničena kada se oslanja isključivo na istorijske numeričke podatke, ali istovremeno demonstriraju snagu i primenljivost rekurentnih neuronskih mreža u analizi kompleksnih vremenskih serija.

Reference

- Google Gemini
- Udemy
- Uvod u LSTM mreže : link
- TensorFlow/Keras: link
- Scikit-learn: link
- Pandas: link
- yfinance: link