Универзитет „Св. Кирил и Методиј“ – Скопје
Факултет за информатички науки и компјутерско инженерство

Проект по предметот:

Управување со ИКТ проекти

Документација за Research проект:

Предвидување на цената на злато со сентимент анализа

Членови на тимот:

Тамара Јосифовска 213189

Вероника Коцева 213114

Ментор:

Милена Трајанова

1. Преглед на проектот

1.1 Цел на проектот

Главната цел на овој истражувачки проект е да се развие систем за предвидување на движењата на цената на златото со користење на комбинација од историски податоци за цени и сентимент анализа на новинарски статии. Проектот има за цел да создаде модел кој може да предвиди дали цената на златото ќе расте, опаѓа или ќе остане стабилна во одреден период.

1.2 Главни компоненти

- **Обработка на ценовни податоци**: Историски податоци за цените на златото (XAU/USD)
- **Сентимент анализа**: Анализа на тонот на новинарски статии
- **Технички индикатори**: RSI, MACD, momentum и други
- **Машинско учење**: XGBoost класификатор за предвидување
- **Евалуација**: Rolling window валидација со различни метрики

1.3 Бинарна vs Мултикласна класификација

Проектот развива два пристапа:

1. **Бинарна класификација**: Предвидување дали цената ќе расте за повеќе од 0.3% во наредните 3 дена
2. **Мултикласна класификација**: Предвидување на три категории - BUY, HOLD, SELL

2. Подготовка на податоци

2.1 Извори на податоци

- **Ценовни податоци**: XAU_USD Historical Data (цена, отворање, највисока, најниска, промена)
- **Новинарски податоци**: Два датасета со новинарски статии и нивниот тон
- **Временски период**: Податоците се мерџуваат по датум

2.2 Сентимент анализа

Од новинарските статии се извлекуваат следните карактеристики:

- **avg_tone**: Просечен тон на статиите по ден
- **max_tone**: Максимален тон
- **tone_std**: Стандардна девијација на тонот
- **article_count**: Број на статии по ден
- **has_gold_theme**: Дали статиите содржат тема за злато

3. Карактеристики (Features)

3.1 Ценовни карактеристики

- **Price, Open, High, Low**: Основни ценовни податоци
- **Change %**: Процентуална промена
- **Price_t-1**: Цена од претходниот ден
- **Price_change_t-1**: Промена од претходниот ден

3.2 Технички индикатори

- **Moving Averages (MA_3, MA_7)**: Пресметани просечни вредности на цената за последни 3 и 7 дена, кои служат за измазнување на трендот и детекција на потенцијални пресврти
- **Volatility (3-day)**: Мерка за нестабилност на цената во последните 3 дена, која укажува на нивото на ризик и можни ненадејни промени
- **RSI (Relative Strength Index)**: Индикатор што мери сила на ценовното движење и помага да се идентификуваат препродадени или преплатени состојби на пазарот
- **MACD и MACD сигнал**: Индикатори базирани на разлика помеѓу експоненцијални просеци, кои укажуваат на потенцијални тренд пресврти преку анализирање на моментумот
- **Momentum**: Мери брзина на промената на цената и укажува на јачината на тековниот тренд

3.3 Пресметани карактеристики

- **MA_3, MA_7**: Подвижни просеци за 3 и 7 дена
- **Volatility_3d**: Волатилност за 3 дена

- **avg_tone_ma3, avg_tone_ma7**: Подвижни просеци за тонот

4. Создавање на target променлива

Се предвидува раст на цената на златото 3 дена во иднина. Класификацијата е трокласна:

- **BUY (2)**: Очекиван пораст над праг
- **SELL (0)**: Очекиван пад под праг
- **HOLD (1)**: Во опсег на неактивност (фокусирано со мултипликатор)

5. Модел и методологија

5.1 Модел

XGBoost Classifier е избран модел со следните карактеристики:

- Користено е Rolling Window Evaluation со:
  - Прозорец: 200 податоци
  - Чекор: 50
- Балансирање со RandomOverSampler за зголемување на примероците од малцински класи
- Применето е GridSearchCV за избор на хиперпараметри
- Фокус на метрика: F1-score (macro) и точност по класа

6. Евалуација на моделот

6.1 Rolling Window валидација

- **Window size**: 200 примероци за тренирање
- **Step size**: 50 примероци за тестирање
- **Временски ред**: Почитување на временската последователност

6.2 Метрики за евалуација

- **F1-score**: Macro и по класи - F1-score е избран бидејќи обезбедува балансирана мерка во услови на неурамнотежени класи (особено важна за HOLD класа која доминира).
- **Accuracy**: Вкупна и по класи

- **Precision и Recall**: За секоја класа посебно

7. Резултати

7.1 Бинарна класификација

Моделот постигнува умерени резултати во предвидувањето на растот на цената за 0.3% во наредните 3 дена. Средните резултати се движат околу:

- **Средна Accuracy**: Варира во зависност од временскиот период
- **Средна F1-score**: Покажува стабилност низ различни временски периоди

7.2 Мултикласна класификација

Мултикласната класификација покажува различни перформанси за различни класи:

- **BUY класа**: Генерално добри резултати
- **HOLD класа**: Предизвици поради честа појава
- **SELL класа**: Умерени резултати

7.3 Feature importance

Анализата на важноста на карактеристиките покажува дека:

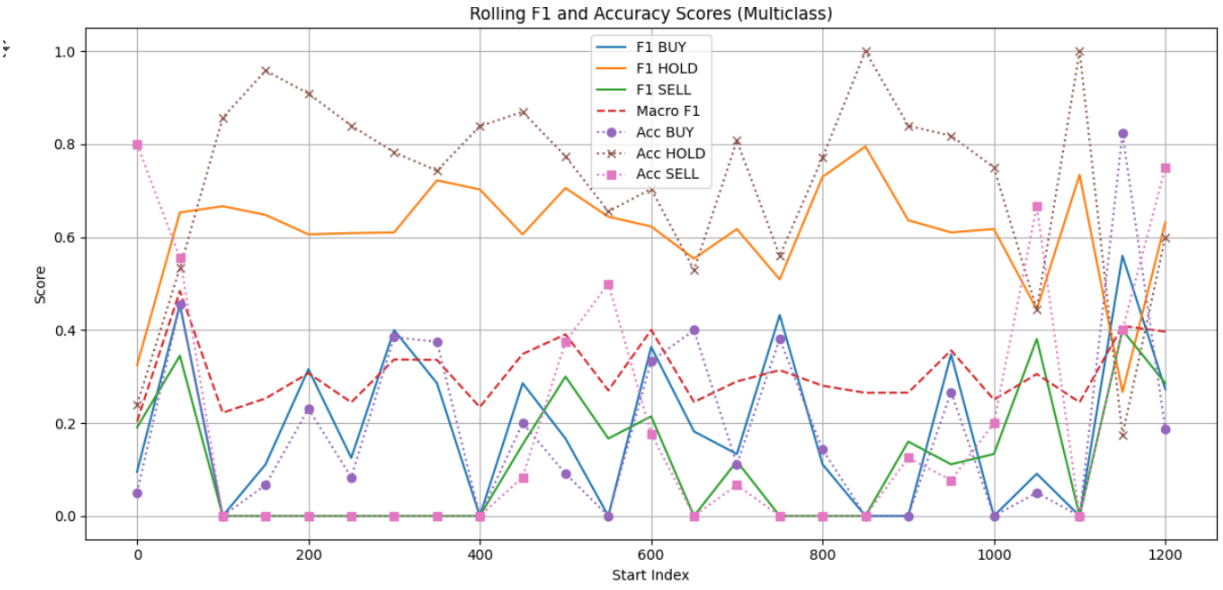
- Ценовните карактеристики имаат голема важност
- Техничките индикатори придонесуваат значително
- Сентимент карактеристиките имаат умерена важност

8. Визуализации

Графикот покажува тренд на цената на златото во долари од 2019 до 2025 година, каде што се гледа постепен пораст од околу 1,300 USD во 2019 до над 3,000 USD во 2025. Особено се забележува драматичен пораст во цената почнувајќи од 2024 година.

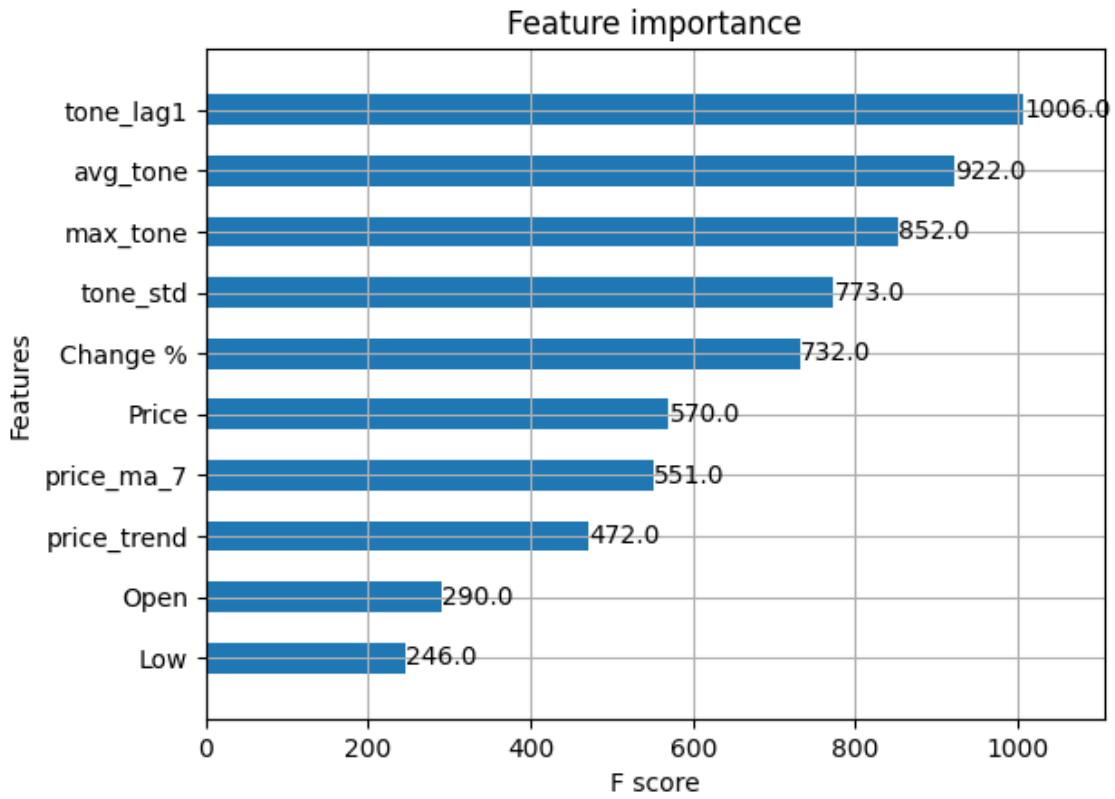
```
Вкупно денови со податоци: 1582
Највисока цена: 3059.58 USD
Просечна дневна промена: 0.06%
```

Следниот график покажува како се менуваат F1 резултатите и точноста за секоја класа (BUY, HOLD, SELL) преку временски прозорци, со цел да се оцени стабилноста и перформансите на моделот за различни пазарни услови




```
Средна Macro F1: 0.30612970068229284
```

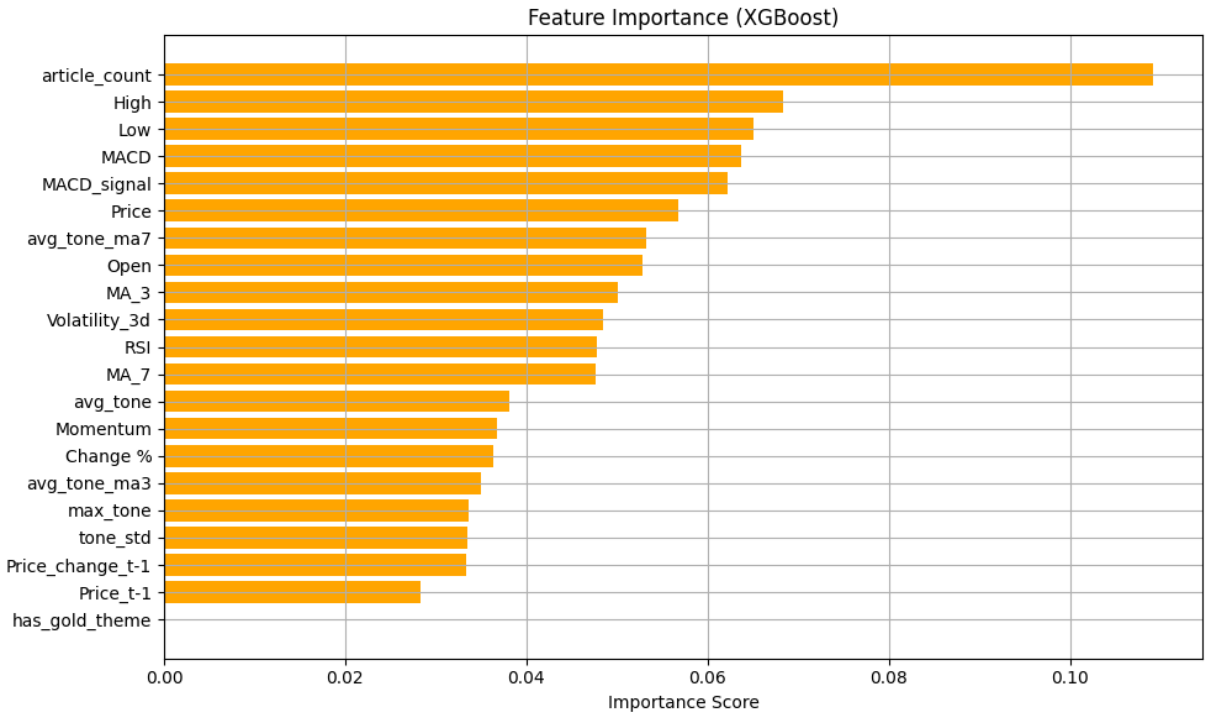
На следната слика може да се забележи кои влезни карактеристики (features) најмногу придонесуваат кон одлуката на моделот дали цената на златото ќе порасне во следните 3 дена.



- Највлијателна карактеристика е tone_lag1, што претставува сентимент на вестите од претходниот ден. Ова покажува дека тонот на вестите има силно влијание врз одлуката.
- Потоа следуваат avg_tone, max_tone, и tone_std, што значи дека тонот на вестите – односно дали тие се позитивни, негативни или неутрални – игра важна улога во тоа како моделот предвидува движење на цената. Од ова можеме да заклучиме дека реакциите на пазарот често зависат од начинот на кој се прикажуваат информациите во медиумите.
- Од техничките индикатори, најрелевантни се:
    - Change % – процентуалната дневна промена во цената,
    - Price, и
    - price_ma_7 – просек на цената во последните 7 дена.

- Карактеристиките како Open и Low имаат помал придонес, што укажува дека статичните дневни вредности не носат доволно информации во овој контекст.

Потоа следува графикот кој прикажува кои карактеристики (features) најмногу влијаеле врз одлуките на моделот за класификација на златото во една од трите категории: BUY, HOLD или SELL

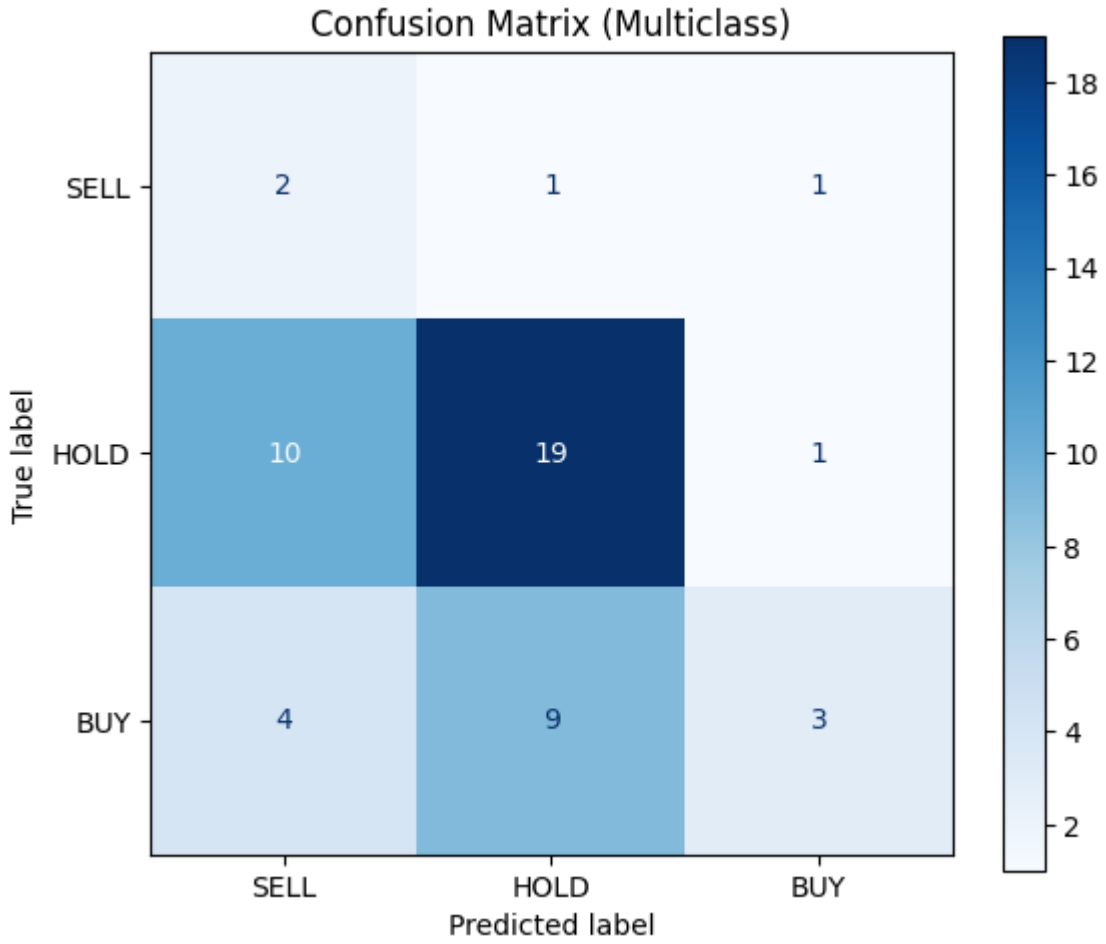


Најголемо влијание има:

- Бројот на статии (article_count) – бројот на објавени статии поврзани со златото, што значи дека самиот факт **колку се зборува за златото во медиумите** има голема улога при донесување одлука дали да се купи, продаде или задржи.
- Следат техничките индикатори High, Low, и MACD, кои ги отсликуваат пазарните движења.
- Исто така, важни се и тонските проседи од вестите (avg_tone_ma7, avg_tone_ma3, tone_std...), што покажува дека не само колку се пишувало, туку и како се зборувало за златото има улога во одлуките.

Може да се заклучи дека моделот користи комбинација од техничка анализа и тоналитет на вестите за да донесе поинформирана одлука дали да препорача купување, задржување или продавање.

На сликата е прикажана *Confusion Matrix* за последниот rolling прозорец во моделот. Оваа матрица прикажува колку примери од секоја класа (SELL, HOLD, BUY) моделот ги класифицирал точно и неточно.



На пример, од SELL примерите, само 2 се точно препознаени како SELL, додека 1 бил класифициран погрешно како HOLD, и 1 како BUY.

Kaj HOLD, моделот постигнува најдобра точност – 19 се правилно препознаени, а 10 биле класифицирани како SELL, што укажува дека моделот не може лесно да ги разликува.

Kaj BUY, 3 се точно, но 9 биле предвидени како HOLD и 4 како SELL, што покажува дека класата BUY е најтешка за моделот.

Заклучок

Овој проект овозможи да комбинираме податоци од различни извори – историски цени на златото и информации од вести – за да се изгради модел кој предвидува дали во наредните денови би било подобро да се купи, продаде или да се почека (BUY, SELL, HOLD). Главната идеја беше да се добие практичен систем кој не само што гледа на бројки, туку го зема во предвид и влијанието од медиумите – како зборовите и тонот на вестите можат да влијаат на пазарот.

Во текот на работата, користевме различни технички индикатори (како RSI, MACD, moving averages) за да се добие појасна слика за движењата на цената, а воедно се правеше и анализа на текстуалните информации од вестите. Еден од поголемите предизвици беше да се одржи баланс меѓу класите, бидејќи „HOLD“ често е најчеста состојба.

Резултатите покажаа дека моделот понекогаш има добри предвидувања, особено за HOLD, но сепак има простор за подобрување кај BUY и SELL. Сепак, преку rolling evaluation и визуелизации како F1-score графици и confusion matrix, се доби појасна претстава за тоа каде точно моделот греши и каде може да се подобри.