Аналитическая платформа биржевых событий на основе текстовых данных

Команда 29

17.03.2025

Цель проекта

- Исследовать влияние новостных статей на динамику цены акций компании.
- Решить две задачи:
  - Регрессионное предсказание точного значения цены.
  - Классификация направления изменения цены (повышение/понижение).

Источник данных и базовая предобработка

- **Данные:** `Apple_data.csv`
  Содержат: `Date, company, headline, abstract, url, section, Open Price, Close Price`
  Для получения новостных статей был написан парсер NYTimes, а для цен - yahoo finance.
- **Предобработка:**
  - Слияние новостных данных и цен на основе поля `Date`.
  - Преобразование поля `Date` в формат даты и сортировка по возрастанию.
  - Заполнение пропусков:
    - Если отсутствует `Open Price` — заполняем предыдущей `Close Price`.
    - Если отсутствует `Close Price` — заполняем следующей `Open Price`.
  - Удаление оставшихся пропусков.

Очистка текстовых данных и извлечение признаков

- **Очистка текста:**
  - Приведение к нижнему регистру, удаление знаков препинания и чисел.
  - Лемматизация и удаление стоп-слов (NLTK).
- **Извлечение признаков:**
  - Создание очищенных полей: `cleaned_headline` и `cleaned_abstract`.
  - Сентимент-анализ:
    - TextBlob для оценки полярности.
    - VADER для получения compound оценки.
  - Вычисление относительного изменения цены:

$$\texttt{price\_change} = \frac{\texttt{Close Price} - \texttt{Open Price}}{\texttt{Open Price}}$$

Базовые модели (прошлые эксперименты)

Регрессия (Линейная регрессия):

- Признаки: `headline_sentiment, abstract_sentiment`; целевая переменная: `Close Price`.
- Результаты:
  - MAE: 23.18
  - MSE: 820.59
  - R^2: 0.0003
  - MAPE: 12.64%

Классификация:

- Бинарный признак: `price_increase` (1, если `Close Price` > `Open Price`).
- Агрегация данных по датам с вычислением среднего, максимума и количества новостей по сентиментам.
- Результаты (Логистическая регрессия): Accuracy: 55.7%.

Новая модель: Расширенное признаковое пространство

- **Текстовые эмбеддинги:**
  - Использование pre-trained модели BERT через SentenceTransformer для получения эмбеддингов заголовков.
  - Агрегация эмбеддингов по датам (среднее значение).
- **Технические индикаторы:**
  - 5-дневная и 10-дневная скользящие средние по цене закрытия.
  - Дневной процентный прирост цены (return).
- **Объединение признаков:**
  - Объединяются агрегированные текстовые эмбеддинги и технические индикаторы по датам.

Результаты новой модели

- Ensemble Stacking Classifier показал accuracy около 79.3%.
- Отчёт по классификации (на тестовой выборке):

```
Accuracy: 0.7931
              precision    recall  f1-score   support
           0       0.75      0.82      0.79        40
           1       0.84      0.77      0.80        47
    accuracy                           0.79        87
```

Модель	Accuracy	Macro (P / R / F1)	Confusion Matrix
XGBoost Classifier	79.5%	0.79 / 0.79 / 0.79	[31, 9; 9, 39]
CatBoost Classifier	84.1%	0.84 / 0.84 / 0.84	[33, 7; 7, 41]
Ensemble Stacking	79.3%	0.79 / 0.80 / 0.79	[33, 7; 11, 36]

Ансамблевый классификатор (Stacking)

- **Базовые модели:**
  - XGBoost Classifier
  - CatBoost Classifier
  - RandomForest Classifier
- **Финальная модель:** Logistic Regression.
- Используется `StackingClassifier` для объединения предсказаний базовых моделей.

Заключение

- Объединение эмбеддингов текстовых данных и технических индикаторов позволило создать более хорошую модель предсказания.
- Ensemble Stacking Classifier показал стабильную точность (около 79%).
- В дальнейшем планируется использовать данную модель в телеграмм-боте для прогнозирования направления изменения цены.