Wágner Bánk

Hugauf Dániel

Frecska Hajnalka

|  |  |
| --- | --- |
| Your Guide To Successful Forex Trading - Austinforex  Haladó adatelemzés és döntéstámogatás  Házi feladat - Dokumentáció | **Trading at the close**    **AIvengers:** |

# Feladat leírása

A tőzsde nyilvános és szabályozott piac, ahol különböző áruk és értékpapírok adásvétele zajlik. A piacok minden munkanapon nyitva vannak, általában 9:30 és 16 óra között (magyar idő szerint 15:30 és 22:00 óra között. A nyitástól a zárásig a kereskedők aktívan részt vesznek a részvények eladásában és vételében, azonban a legnagyobb forgalom a kereskedés utolsó 10 percében történik, amikor aukciót tartanak. A piaci résztvevők megbízásokat adnak be, ajánlatokat, illetve kéréseket tehetnek olyan árral, amennyiért részvényt szeretnének venni vagy eladni. Az eladási árak általában magasabbak, mint az ajánlati árak, mivel az eladni szándékozók magasabb árat szeretnének kapni a részvények eladásáért, mint a vásárolni szándékozók. Ez azt jelenti, hogy az eladási és az ajánlati árak összehangolása (ami egy befejezett ügyletnek felel meg) egy optimalizálási probléma: olyan végső záróárat kell találni (a piacnak az aukció végére egyetlen árról kell döntenie), amely maximalizálja a megvalósuló ügyletek számát. Az optimalizálásba beletartozik. hogy az aukció során ezen keresztár mozgásait előre jelezzék. Ezalatt a 10 perc alatt a megbízások folyamatosan érkeznek, az árjegyzők pedig azon dolgoznak, hogy megtalálják a legjobb árat, és eldöntsék, hogy belépjenek-e az aukcióba, illetve hogyan lépjenek be.

A feladat a megfelelő modell létrehozása a záróár megjóslására. Az adathalmaz egyedi kihívást jelent, abból a szempontból, hogy nem egy következő időszakra, hanem az egy perccel későbbi ármozgás előrejelzésére összpontosít.

# Adatok

Az adathalmaz 200 különböző NASDAQ-on jegyzett részvény historikus adatait tartalmazza 480 napra visszamenőleg, minden kereskedési nap záró aukciós idejéből (utolsó 10 perc).

|  |  |  |
| --- | --- | --- |
| stock\_id | A részvény egyedi azonosítója |  |
| date\_id | A dátum egyedi azonosítója |  |
| imbalance\_size | A nem megvalósuló ügyletek száma a referencia áron. |  |
| imbalance\_buy\_sell\_flag | Indikátor, megvalósult és nem megvalósult ügyletek számának arányának megfelelően. | vételi oldal túlsúly= 1  eladási oldal túlsúly = -1  egyensúly = 0 |
| reference\_price | Az az ár, amely maximalizálja a megvalósult ügyletek számát és minimalizálja az egyensúlytalanságot adott pillanatban. |  |
| matched\_size | A megvalósuló ügyletek száma a referencia áron. |  |
| far\_price | keresztár, amely maximalizálja a megvalósuló ügyletek számát (csak az aukciós ügyleteket figyelembe véve) |  |
| near\_price | keresztár, amely maximalizálja a megvalósuló ügyletek számát (aukciós és folyamatos piaci megbízásokat figyelembevéve) |  |
| bid/ask\_price | A legjobb vételi/eladási ár a nem aukciós könyvben |  |
| bid/ask\_size | A legjobb vételi/eladási dollárban kifejezett árának a mennyisége. |  |
| wap | súlyozott átlagár a nem aukciós könyvben. [2. ábra] |  |
| secconds\_in\_bucket | A napi záró aukció kezdete óta eltelt másodpercek száma. | Mindig nulláról indul |
| target | 60 másodperces jövőbeli mozgása a wap-nak, mínusz a szintetikus index 60 másodperces jövőbeni mozgása. [1. ábra] | 1 bázispontos árváltozás 0,01%-os árváltozásnak felel meg. |

A képen szöveg, Betűtípus, sor, fehér látható

Automatikusan generált leírás

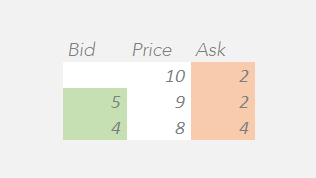
1. ábra célváltozó számítása

A képen szöveg, Betűtípus, fehér, sor látható

Automatikusan generált leírás

2. ábra WAP számítása

A kombinált aukciós könyv szemléltetése és ez alapján a számítások:



10-es áron 0 tétel fog megvalósulni, mivel nincsenek vételi kérések (bid=0)

9-es áron 5 tétel fog megvalósulni, mivel 5 vételi kérés van 9-es áron és 6 eladási ajánlat 9 vagy annál kevesebb áron. (5 bids >=9 && 6asks <=9)

8-as áron 4 kérés fog megvalósulni, mivel 9 darab vételi ajánlat van 8 vagy nagyobb áron, illetve 4 eladási ajánlat van 8 vagy kevesebb áron. (9 bids>=8 && 4asks<=8)

Tehát az ár, ami maximalizálja a kéréseket a 9. A mennyisége (matched\_size) 5. Az egyenlőtlenség 1 az eladási oldalon. A kombinált aukciós könyv szerinti ár lesz a near\_price, amit 5 perccel a zárás előtt hoznak nyilvánosságra.

# Adatelőkészítés

Adatgyűjtés

Az adathalmaz nyilvánosan elérhető a [Optiver - Trading at the Close | Kaggle](https://www.kaggle.com/competitions/optiver-trading-at-the-close/data) oldalon. ~ 640MB

Profiling

Az adathalmazt a ydata\_profiling könyvtár segítségével végeztük.

Az adathalmaz 17 változót tartalmaz a célváltozóval együtt. Ebből 15 numerikus, 1 kategorikus és 1 szöveg típusú változó.

több mint 5 millió sora tartalmazza a 200 különböző részvény, 480 napon át ívelő záró aukción belüli változásokat. A záró aukció utolsó 1 perce nem áll rendelkezésre egyik napon sem, a feladat ezen 1 perces intervallumon belüli mozgások prediktálása (200\*480\*54 = 5184000).

A változók között a far\_price és near\_price tartalmaz hiányzó sorokat, mindkettő esetében az esetek felében. Ezek az adatok csak 5 perccel a záróár, tehát a piac zárása előtt kerül nyilvánosságra.

Emellett a wap változó esetén figyelhető meg 220 hiányzó érték, mindegyik a 438. napon a 19, 101, 131 és 158-as részvény esetén. Ezen részvények esetén a 438. nap teljesen hiányzik.

Az adatok között számos korreláció előfordul

ask\_price ,bid\_price, reference\_price, wap, far\_price, near\_price

date\_id , time\_id

imbalance\_buy\_sell\_flag , near\_price

A képen szöveg, képernyőkép, diagram, tervezés látható

Automatikusan generált leírás

3. ábra korrelációk

Adattisztítás

A 438. napon megfigyelhető 4 részvény adatainak hiányzása miatt, ezeket a sorokat eltávolítottuk. Az eltávolított sorok aránya 0.004% az egész adathalmazhoz képest.

Kvantilis transzformáció

Az alapötlet, hogy számos gépi tanulási algoritmus előnyben részesíti vagy jobban teljesít, ha a numerikus bemeneti változók standard valószínűségi eloszlásúak, például Gauss (normál) vagy egyenletes eloszlásúak. A kvantilis transzformáció olyan módszer, amely a jellemzőket egyenletes vagy normál eloszlásúvá alakítja. Az adathalmazban 4 változó esetén volt megfigyelhető nem szabványos eloszlás, „ferdeség”.

A képen szöveg, Betűtípus, képernyőkép, szám látható

Automatikusan generált leírás

A kvantilis transzformáció után, a négy változó a fentivel megegyező sorrendben:

A képen szöveg, Betűtípus, képernyőkép, fekete látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, Téglalap, sor látható

Automatikusan generált leírás

4. ábra Az imbalance\_size eloszlása a transzformáció előtt

A képen Diagram, képernyőkép, diagram, sor látható

Automatikusan generált leírás

5. ábra Az imbalance\_size eloszlása a transzformáció után

Jellemzők kinyerése

Az *imbalance\_buy\_sell\_flag* változó 3 értéket vehet fel, ezeknek dummy változókat hoztunk létre.

A float64 típusú változókon típuskonverziót hajtottunk végre, helymegtakarítás céljából(float32).

Jellemzők kiválasztása

A magasan korreláló változókat elvetettük, majd a megmaradt változókon az RFECV (Recursive Feature Elimination with Cross-Validation) segítségével jellemzőszelekciót végeztünk, hogy kiválasszuk a legfontosabb változókat. A model úgy lett beállítva, hogy 50 körön keresztül, 0.3-as learning rate-el az abszolút errort figyelembe véve válassza ki a legmegfelelőbb változókat. A kiértékeléshez az átlagos abszolút eltérést vette figyelembe (MAE).

A képen szöveg, képernyőkép, Színesség, Párhuzamos látható

Automatikusan generált leírás

6. ábra Az RFECV eredménye

# Vizualizáció, temporális struktúra

A képen szöveg, diagram, Tervrajz, sor látható

Automatikusan generált leírás

7. ábra Egy részvény jellemzőinek változása a záró aukción belül

Az 7. ábra egy véletlenszerűen kiválasztott részvény, utolsó 9 perces időintervallumon belüli jellemzőérték változásait szemlélteti. Megfigyelhető, hogy a far\_price és near\_price nyilvánosságra hozatala után (300 másodperc – 5 perc) egyes változókban drasztikus változás figyelhető meg.

A képen szöveg, képernyőkép, Diagram, diagram látható

Automatikusan generált leírás

8. ábra A teljes időtartam alatt (480 nap) növekvő tendencia figyelhető meg a kereskedések számában

A képen szöveg, képernyőkép, diagram, Diagram látható

Automatikusan generált leírás

9. ábra Általános szabály, hogy az eladási ár minden esetben nagyobb mint a vételi ár, wap ezeknek a súlyozott átlaga

A képen kézírás, szöveg, sor látható

Automatikusan generált leírás

10. ábra A célváltozó alakulása különböző részvények esetén, ugyanazon a napon

A képen szöveg, kézírás, sor, diagram látható

Automatikusan generált leírás

11. ábra Ugyanazon részvény célváltozójának alakulása a különböző napokon

A 10. ábra és 11. ábra – n megfigyelhető a célváltozó alakulása. Ugyanazon részvény esetén több napot vizsgálva, illetve különböző részvényeket ugyanazon a napon vizsgálva látható, hogy nem figyelhető meg tendencia, az értékek első ránézésre kiszámíthatatlanul viselkednek és nagymértékben volatilisek.

# Új változók létrehozása

|  |  |
| --- | --- |
| ask\_ref\_ratio | Az eladási ár és a referenciaár közötti arány. Ha az eladási ár magasabb, mint a referenciaár, az bika hangulatra utal. Ha az eladási ár alacsonyabb, mint a referenciaár, az eladási hullámra utal. |
| ask\_wap\_diff | A különbség a vételi és eladási ár között |
| spread | Az eladási ár és a wap közötti különbség, mellyel a piaci hatékonyságot lehet jelezni. |
| price\_pressure | Az imbalance\_size és spread szorzata |
| total\_volume | ask + bid méretek összege |
| vwap\_reference\_price | lsd. egyenlet |
| vwap\_matched\_size | lsd. egyenlet |
| rolling\_std\_reference\_price | lsd. egyenlet |

df['ask\_ref\_ratio'] = df['ask\_price']/df['reference\_price']

df['ask\_wap\_diff'] = df['ask\_price'] - df['wap']

df['spread'] = df['ask\_price'] - df['bid\_price']

df["price\_pressure"] = df["Quantile\_imbalance\_size"] \* (df["spread"])

df["total\_volume"] = df['Quantile\_ask\_size'] + df['Quantile\_bid\_size']

df['vwap\_matched\_size'] = (df['wap'] \* df['matched\_size']).cumsum() / df['total\_volume'].cumsum()

df['vwap\_reference\_price'] = (df['wap'] \* df['reference\_price']).cumsum() / df['total\_volume'].cumsum()

df\_train['rolling\_std\_reference\_price'] = df\_train.groupby('stock\_id')['reference\_price'].transform(lambda x: x.rolling(window=54).std())

# Az új változók temporális struktúrája

A képen szöveg, sor, Diagram, diagram látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, Betűtípus, sor látható

Automatikusan generált leírásA képen szöveg, képernyőkép, Betűtípus, sor látható

Automatikusan generált leírás

A képen szöveg, sor, diagram, Diagram látható

Automatikusan generált leírás

# Modellek

1. VERZIÓ

Az kiválasztott oszlopok RFECV-vel

* imbalance\_size
* imbalance\_buy\_sell\_flag
* reference\_price
* matched\_size
* bid\_price
* bid\_size
* ask\_price
* ask\_size
* wap
* target

Modellek tanítása

**XGBoost**

learning\_rate = 0.1

n\_estimators = 200

colsample\_bynode = 0.8

max\_depth = 9

A képen szöveg, sor, Diagram, diagram látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, képernyő, szám látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, szám, sor látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, diagram, sor látható

Automatikusan generált leírás

**LightGBM**

learning\_rate = 0.1

n\_estimators = 100

num\_leaves = 75

colsample\_bytree = 0.8

A képen szöveg, sor, Diagram, diagram látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, diagram, sor látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, szám, sor látható

Automatikusan generált leírás

Eredmények

|  |  |
| --- | --- |
| Model | Best MAE |
| XGBoost | 6.322 |
| LightGBM | 6.324 |

Kiértékelés

A képen szöveg, diagram, Diagram, sor látható

Automatikusan generált leírás

A képen szöveg, sor, diagram, Diagram látható

Automatikusan generált leírás

12. ábra Kiértékelés eredménye az első 100 adatpont esetén

1. VERZIÓ

Paraméter optimalizálás

Az kiválasztott oszlopok RFECV-vel

* stock\_id
* date\_id
* seconds\_in\_bucket
* Quantile\_bid\_size
* Quantile\_ask\_size
* spread
* ask\_ref\_ratio
* ask\_wap\_diff
* imbalance\_buy\_sell\_flag\_1
* target

**XGBoost**

learning\_rate = 0.01

n\_estimators = 200

A képen szöveg, képernyőkép, szám, sor látható

Automatikusan generált leírás

coolsample\_bynode = 0.8

max\_depth = 6

A képen szöveg, képernyőkép, sor, szám látható

Automatikusan generált leírás

**LightGBM**

learning\_rate = 0.05

n\_estimators = 50

A képen szöveg, képernyőkép, szám, sor látható

Automatikusan generált leírás

coolsample\_bynode = 0.8

num\_leaves = 50

A képen szöveg, képernyőkép, szám, sor látható

Automatikusan generált leírás

Modellek tanítása

* Train\_test\_split arány 80%-20%

**XGBoost**

A képen szöveg, sor, Diagram, diagram látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, szám, diagram látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, diagram, szám látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, diagram, sor látható

Automatikusan generált leírás

**LightGBM**

A képen szöveg, sor, Diagram, képernyőkép látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, diagram, sor látható

Automatikusan generált leírás

A képen szöveg, képernyőkép, diagram, sor látható

Automatikusan generált leírás

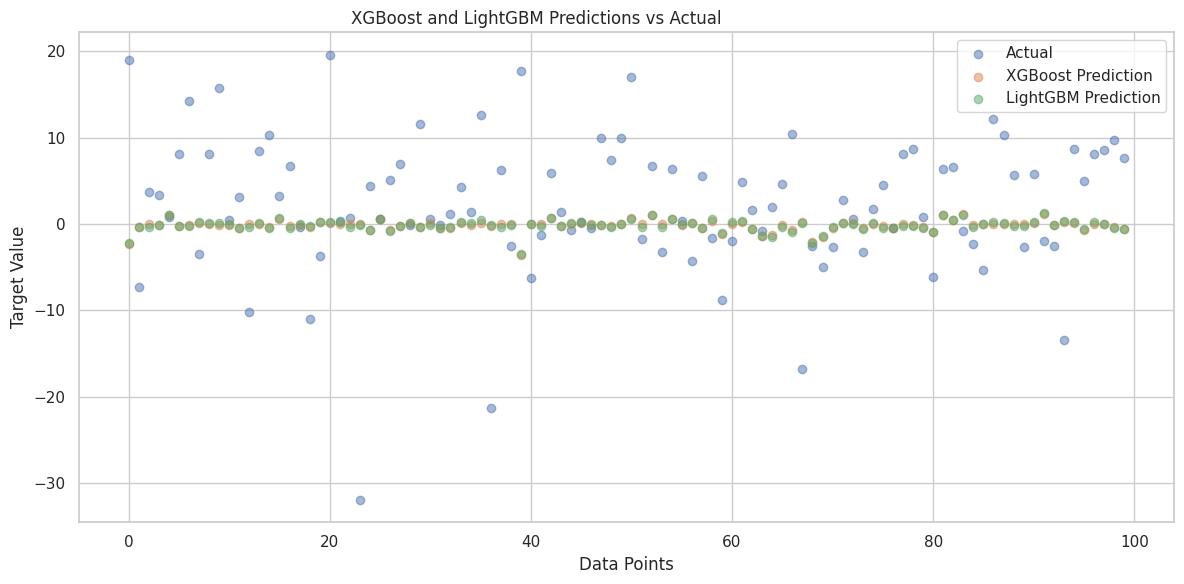
Eredmények

|  |  |  |
| --- | --- | --- |
| Model | Best MAE train | Best MAE eval |
| XGBoost | 6.626 | 5.977 |
| LightGBM | 6.627 | 5.975 |

Kiértékelés

A képen szöveg, diagram, Diagram, sor látható

Automatikusan generált leírás



13. ábra Predikciók az első 100 adatpontra

1. VERZIÓ

Az eddig előkészített adathalmazzal paraméter optimalizációt végeztünk különböző modellekre az **OPTUNA** segítségével. Emellett a **CatBoost** modellt is kipróbáltuk.

**A képen szöveg, képernyőkép, Betűtípus, szám látható

Automatikusan generált leírás**

14. ábra XGB paraméterei

**A képen szöveg, képernyőkép, Betűtípus látható

Automatikusan generált leírás**

15. ábra LightGBM paraméterei

A tanítás során keresztvalidációt alkalmaztunk az eredmények javítása és a túltanulás elkerülése érdekében. Az tanító adathalmazt 3 részre osztottuk és minden egyes iterációban egy másik részt használtunk a tesztelésre. Azt tapasztaltuk, hogy a modellek jelentős javulást tudtak elérni.

Eredmények a tanítás és a teszthalmazon végzett kiértékelés után

|  |  |  |
| --- | --- | --- |
| Model | Cross validation score (k=3) | Test MAE |
| Baseline CatBoost  CatBoost(k=4) | 6.33  6.4559 | 5.47  5.464 |
| XGBoost  **XGBoost after optuna** | 6.38  **6.32** | **5.408** |
| LightGBM  LightGBM after optuna | 6.31  6.306 | 5.439 |

Modellek átlagolása

A betanított modellekkel a teszt adathalmazon prediktáltunk, majd az eredményeiket átlagoltuk, az így kapott **MAE 5.44** lett. Ezek közül a legjobban teljesítő modell **az XGBoost** lett (**5.408**) a paraméter optimalizáció után.

A végső eredmények a *submission\_xgb.csv és submisssion\_avg.csv*  fájlokban található.

Az eredmények a feladat leírása szerinti formában:

16. ábra átlagolt eredmény

A képen szöveg, képernyőkép látható

Automatikusan generált leírás

17. ábra Átlagolt predikciók az első 100 pontra

A képen szöveg, Diagram, képernyőkép látható

Automatikusan generált leírás

18. ábra XGBoost predikciója az első 100 pontra