

TIME-SERIES ANALYSIS



VOLUME PREDICTION S&P 500

LUKÁŠ MÁLEK

✉ malek.luky@gmail.com

👤 linkedin.com/in/malek-luky/

💻 github.com/malek-luky

PRAHA

29.12.2022

ÚVOD

Cílem úkolu je predikce množství obchodovaných indexů S&P ($^{\wedge}\text{GSPC}$) v roce 2017 a 2018. Předpověď probíhá pomocí strojového učení v pythonu, přesněji s použitím XGBoost a scikit-learn. Pro naučení používáme data od roku 2000. Zdrojový kód je vypracován v Jupyter Notebooku.

Tento úkol se od běžných botů na algoritmické obchodování odlišuje predikcí objemu obchodovaných indexů namísto ceny. Tudíž klasické technické indikátory, práci s candlesticks, support and resistance detection a strategie pro buy and sell jako například stop-loss či ATR Trailing není vhodné použít.

Prioritou reportu bylo získání co nejvíce zkušeností v oblasti algoritmického obchodování, takže některé použité metody nemusí být vhodné pro predikci objemu, ale lze je uplatnit v dalších oborech algotradingu.

PREREKVIZITY

K vypracování modelů byly použity následující knihovny:

- pandas-market-calendars
- scikit-learn
- xgboost
- yfinance
- pandas
- pandas_ta
- seaborn
- hyperopt

TEORIE

VOLBA MODELU

Problematiku by bylo možné nekonvenčně řešit analýzou bez nutnosti použití strojového učení. Takto fungují například analýzy pomocí candlesticks či support and resistance strategií pro odhad ceny. V rámci úkolu jsem ale dával důraz na procvičení strojového učení, tudíž jsem se do odhadování ceny z pouhé analýzy předchozích hodnot nepouštěl.

Při rozhodovacím procesu jsem se přes ARIMU, Prophet, RNN a LSTM dostal až k Transformerům, které mi jak z důvodu aktuálnosti (GPT3), tak i principem implementace v PyTorch s možností paralelizace na GPU / TPU přijdou nejlepším řešením v případě složitých výpočetních úkolů.

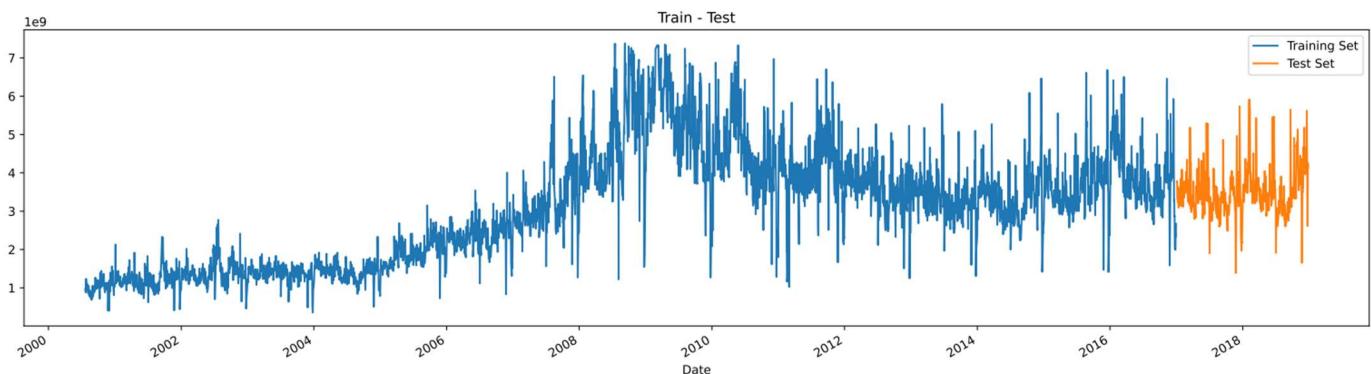
Jelikož naše problematika nevyžaduje tak velký výpočetní výkon a můj notebook nemá vhodné GPU pro plné využití paralelizace, uchýlil jsem se k XGBoost. Poskytuje velmi dobré výsledky a stal se jedním

z nejpopulárnějších algoritmů pro programovací soutěže. Spíše než na kvantitu modelů bylo dbáno na kvalitu a porozumění výsledků.

ANALÝZA DAT

HISTORICKÝ VÝVOJ

Při práci je nutno se podívat na historický vývoj, pohybuje se veličina okolo konstantní hodnoty, nebo naopak jeví lineární/kvadratický růst či pokles, je možné vyzpovídat sezónnost měřených hodnot jako například u spotřeby energie v domácnostech?

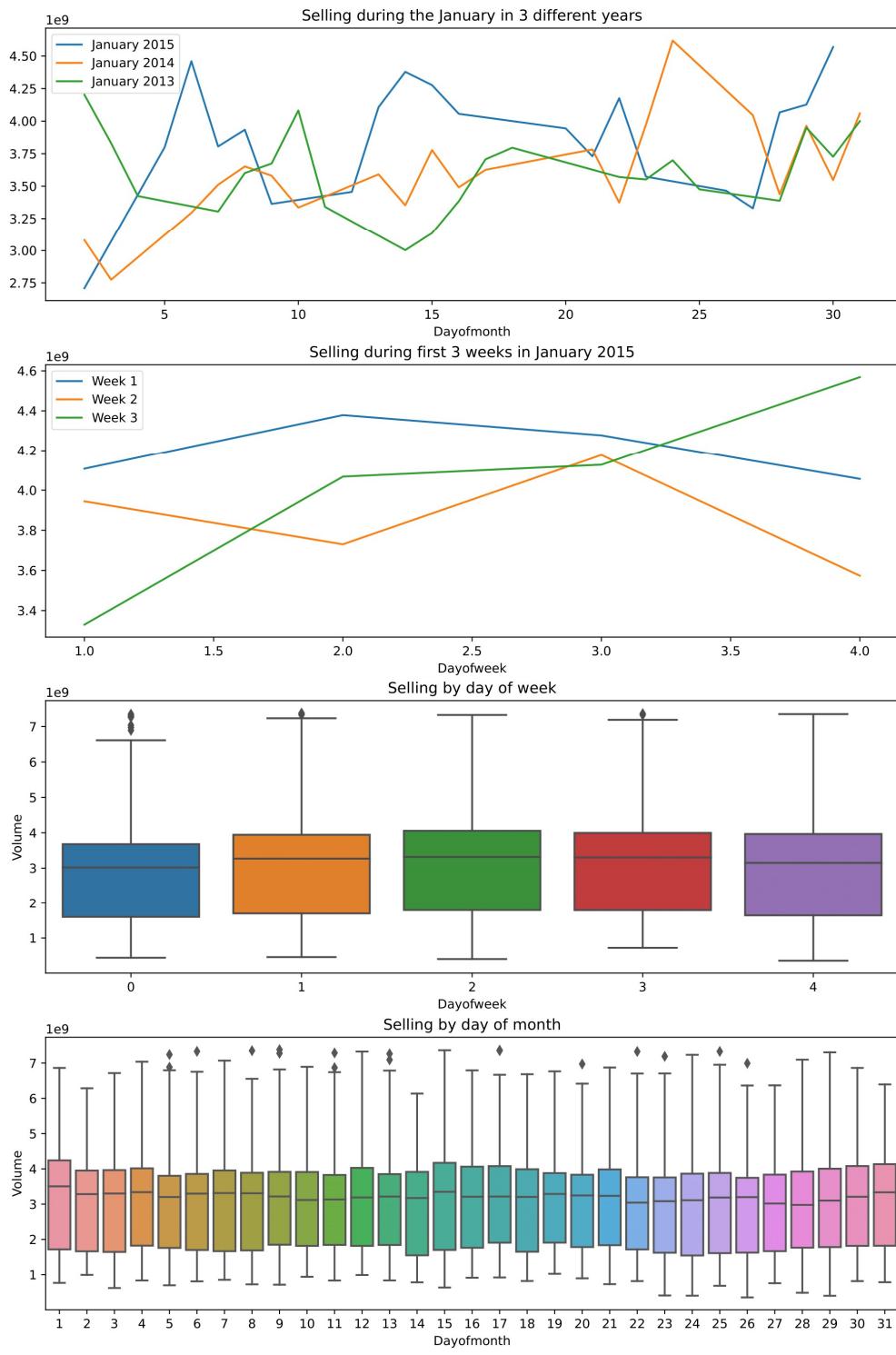


Graf 1: Historický vývoj objemu obchodů s indexem S&P500

Zde vidíme, že mezi lety 2000-2010 docházelo k postupnému růstu. Tudíž hrozí, že algoritmus bude predikovat přechozí hodnoty. Tento problém lze vyřešit, že namísto predikování celkového množství obchodovaných indexů budeme predikovat rozdíl oproti předchozímu dni.

OPAKUJÍCÍ SE VZOR

Pro ujištění, že obchodovaný objem je nezávislý na dnech, respektive že se například pravidelně v pondělí neobchoduje ve větších objemech než následující den, zobrazili jsme si vývoj v několika scénářích, které potvrdily naši domněnku.



Graf 2: Analýza opakujících se vzorů

ONE-HOT ENCODING

Jelikož data neobsahují žádné kategorie, one hot encoding není nutné použít. Stejně tak normalizace dat není z důvodu použití XGBoost nutná, jelikož pro algoritmy založené na Decision Trees normalizace nehraje roli na výsledné predikci.

SPLIT

Při rozdelení dat na training a testing dataset je potřeba dbát důraz na to, že nepoužijeme běžný k-fold či training-testing split. Naše trénovací data vždy musí být na konci. Pro použití K-Fold cross validace je možné využít knihovnu TimeSeriesSplit.

OUTLIERS

Historická data je nutné promazat o atypické hodnoty, které by mohly negativně ovlivnit trénování našeho modelu. Pro najití outliers byl použitý boxplot. Mezi hodnotami cen se vyskytoval pouze jeden outlier, který jsme ponechali v datasetu. Pro objem obchodovaných indexů se jich vyskytovalo více než 50. Všechny tyto hodnoty byly z datasetu odebrány.

N/A VALUES

Při využití některých technických indikátorů došlo ke vzniku N/A values na začátku roku 2000. Tyto hodnoty byly odebrány. Jelikož se vyskytovaly u features, které měly jednu z nejmenších důležitostí, lepším řešením bylo otestovat, zda se při odebrání celé feature s N/A hodnotami a ponechání většího objemu testovacích dat zvýší účinnost.

NORMALIZOVÁNÍ

Jak již bylo zmíněno dříve, pro XGBoost normalizace není nutná. Ovšem v případě LSTM by normalizace dat byla nutná. V tomto případě by nejlepší metodou byl MinMaxScaler, který převeze data na hodnoty mezi 0 a 1. MiMaxScaler neovlivňuje na tvar distribuce dané veličiny, takže žádná informace v datech nebude ztracena. Při zobrazení distribuce dat se ne vždy jedná o normálové rozdělení, tudíž StandardScaler, který data převeze do normálové distribuce, by nemuselo být vhodné použít. Zároveň jsou data rozdělena rovnoměrně, tudíž nehrozí, že by se velké množství dat zobrazilo pouze v okolí jedné hodnoty.

FEATURES

Data byla obohaceno o několik hodnot pro větší šanci najítí vhodných korelací mezi features a obchodovaným objemem. Přidané hodnoty jsou detailně popsány níže.

TECHNICAL INDICATORS

Z ceny a obchodovaného množství byly data obohaceny o následující technické indikátory.

1. RSI – Relative Strength Indicator
2. CCI – Commodity Channel index
3. AO – Awesome Oscilator
4. MOM – Momentum Indicator
5. MACD – Moving Average Convergence Divergence

6. ATR – Average True Range
7. BOP – Balance of Power
8. RVI – Relative Vigor Index
9. DM – Directional Movement
10. STOCH – Stochastic Oscilator
11. STOCHRSI – Stochastic RSI
12. SILLR – Williams %R

Některé z technických indikátorů vrátí více než jednu hodnotu. Nicméně po analýze výsledků vidíme, že žádný z technických indikátorů nemá na předpovídání hodnotu vliv. Není se čemu divit, jelikož tyto indikátory jsou využívány pro predikci ceny, nikoli množství obchodů.

VOLUME DIFERENCE

Jak bylo zmíněno v sekci ohledně historického vývoje, v reportu se podíváme nejen na predikci celkového Volume, ale také na predikci rozdílu objemu mezi dnešním a zítřejším dnem. Což je přesně to, co sloupec VolumeDiff obsahuje.

DATES

Přestože po původní analýze jsme vliv sezónnosti zavrhl, ukázalo se, že vliv dnů, měsíců a roků se podílí na výsledném odhadu. Data jsme doplnili o následující:

1. Den v týdnu
2. Měsíc
3. Rok
4. Den v měsíci
5. Kvartál

V případě většího množství času na vypracování by bylo zajímavé se detailněji podívat, zda přidání týdnu v roce ovlivní výsledky. Pro zajištění validního testování by bylo nutné udělat robustnější testování s vyšším množstvím iterací a k-fold cross validací.

MARKET CALENDAR

Tento indikátor se ukázal jako ten úplně nejdůležitější. Některé dny jsou otevírací hodiny zkráceny, což mělo velký vliv na výsledný objem obchodů. Běžná otevírací doba na NYSE a Nasdaq, kde jsou instrumenty z S&P500 obchodovány, byla 6.5. Ovšem v den kratší otevírací doby bylo otevřeno pouze 3.5h.

ADD LAGS

U tohoto indikátoru byl nízký vliv potvrzen. Jedná se o přidání sloupců s obchodovaným objemem den, týden, měsíc a půl rok zpátky.

TARGETS

Hodnoty následujícího dne, které chceme predikovat, respektive y_{test} a y_{train} našeho modelu.

NÁVRHY NA ZLEPŠENÍ

Jedná se o složitou problematiku, která v horizontu 14 dní je velmi těžká detailně vyřešit. Je tedy vhodné se podívat i na příležitosti, které by mohly vést k vyšší úspěšnosti modelu. Některé z nápadů níže jsou v GitHub reposítáři rozpracovány v *Todo List.ipynb*, ovšem většina z nich nebyla dostatečně odladěná pro využití v našem výsledném modelu.

ROBUSTNOST

Testování robustnosti modelu je vždy stěžejním aspektem. Jedná se o největší problém aktuálního modelu. Pokud bychom chtěli model použít pro investici vlastních peněz, bylo by vhodné robustnost modelu detailněji otestovat kombinací těchto možností:

1. K-Fold Cross validace: Otestuje model na více vzorcích
2. Přidání validačního setu: Určí, zda naměřené přesnosti jsou konstantní, či dochází k častým výkyvům
3. Testování na kratším časovém úseku: Před 18 lety se mohl trh chovat zcela odlišně, než jak je tomu teď. Hrozí, že se model natrénuje na zcela odlišných vzorech, které již nyní nelze uplatnit. Zároveň s rozšířením počítačů je stále větší vliv výpočetní techniky na provedených obchodech.
4. Ozkoušet model na nových údajích: Momentálně máme z burzy stažená pouze data celodenního vývoje .Byl by model stále funkční při použití hodinových záZNAMŮ?
5. Použít model na jiném indexu: Robustní model by měl být schopný predikovat i jiné indexy, u kterých se očekává podobné chování

HYPERPARAMETRY

Přestože model používá hyperopt knihovnu spolu s Bayesian Search, bylo by možné si více pohrát s různými testovanými intervaly i hloubkou testování. Z časových důvodů jsou parametry testovány pouze na 100 iterací. Při navýšení množství iterací bychom mohli dosáhnout lepších výsledků.

INVESTIČNÍ PŘÍLEŽITOST

Nabízí se možnost použití různých investičních strategií a modelů k predikci vhodných situací pro investování. Bude existovat korelace mezi investiční příležitostí a obchodovaným objemem?

SLEDOVÁNÍ NEOČEKÁVANÝCH UDÁLOSTÍ

Na obchodování budou mít vliv i velké světové novinky z finančního trhu. Jejich aktivní sledování a přidání závažnosti dané situace do našeho modelu může vést k lepší predikci. Bohužel se nepodařilo nalézt kalendář, kde by tyto události spolu se závažností byly snadno importovatelné.

PŘETRÉNOVÁNÍ MODELU

Postupné přetrénování našeho modelu na aktuálních datech by lépe odráželo budoucí trh. Nyní používáme nás nátrénovaný model z let 2000-2016 pro predikci konce roku 2018. Data od začátku roku 2017 již v našem modelu neuvažujeme.

VOLBA FEATURES

Při použití pouze šesti nejdůležitějších features má nový model lehce lepší výsledky. Bylo by vhodné vyzkoušet, zda některé kombinace vstupních argumentů dokážou najít korelace, které nám nyní unikají.

PŘIDÁNÍ VÍCE INDIKÁTORŮ

Přestože se cenové technické indikátory neosvědčili, bylo by vhodné přidat více indikátorů pro vývoj objemu jako například moving average. Na druhou stranu graf zobrazuje vysokou volatilitu. Zda sledování trendu může vést k lepším výsledkům se bohužel nedá bez podrobnější analýzy určit.

TESTOVÁNÍ EXTRÉMNÍCH VSTUPŮ

Model nebyl podroben zatěžkávací zkoušce extrémních vstupů. Jelikož hodnoty z trhu bývají přesné, spíše než extrémní a chybné vstupy může docházet k extrémním událostem na trhu, které nejdou predikovat. V neočekávaných situacích je lepší obchodování pozastavit, protože na tyto situace algoritmus není připraven.

VÝSLEDKY

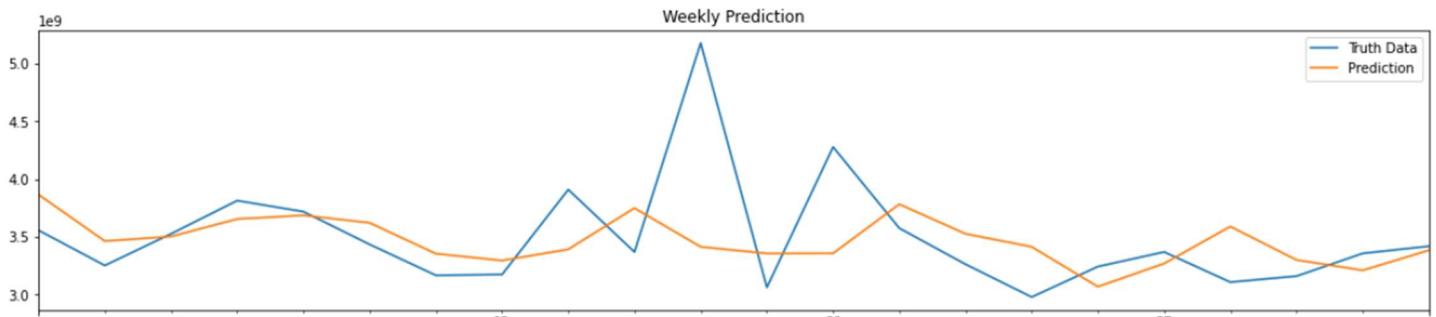
Přestože používáme pouze XGBoost, máme 4 různé modely dle predikovaných dat:

- Baseline: Predikce hodnoty z předchozího dne
- Model1: Predikce celkového objemu
- Model2: Predikce rozdílu obchodovaného objemu
- Model3: Predikce poklesu / růstu nadcházející den (
- Model4: Stejně jako Model2, používáme ale pouze 6 nejlepších features namísto všech

Všechny modely využívají XGBRegressor kromě Modelu3, který je naimplementovaný pomocí XGBClassifier.

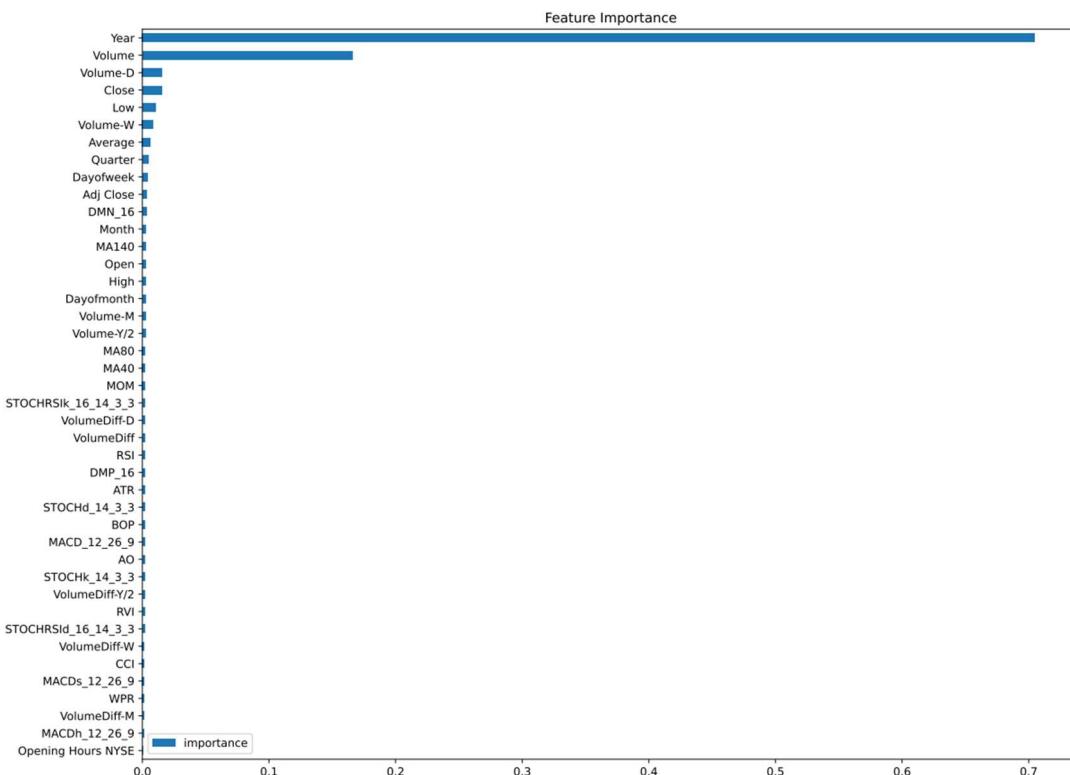
POPIS MODELŮ[◦]

MODEL1



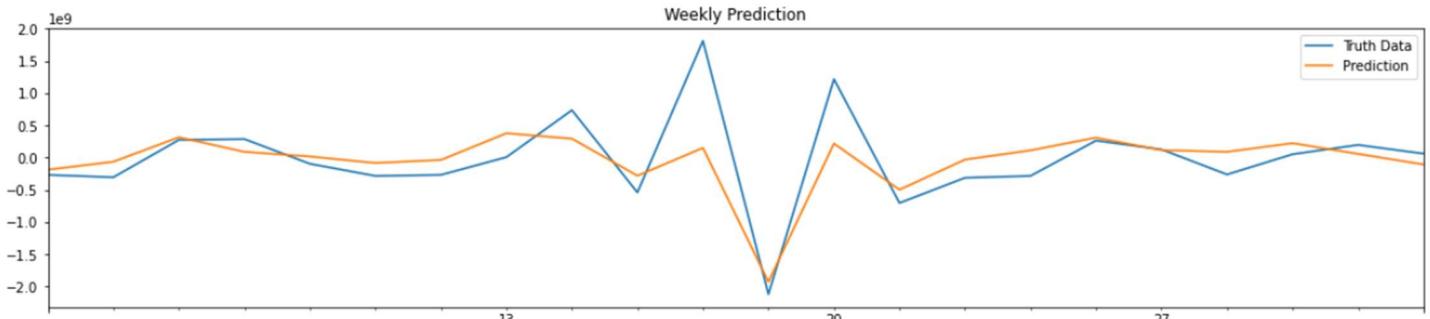
Graf 3: Měsíční predikce

Na první pohled se může zdát, že predikce vychází dobře. Při bližším pohledu na jednotlivé denní odhady ale vidíme, že model je silně ovlivněn predikcí v předchozích dnech. Při pohledu na features vidíme, že model predikuje primárně na základě roku. Přestože dává lepší výsledek než baseline, je těžké mu důvěrovat a považovat ho za dostatečně robustní.



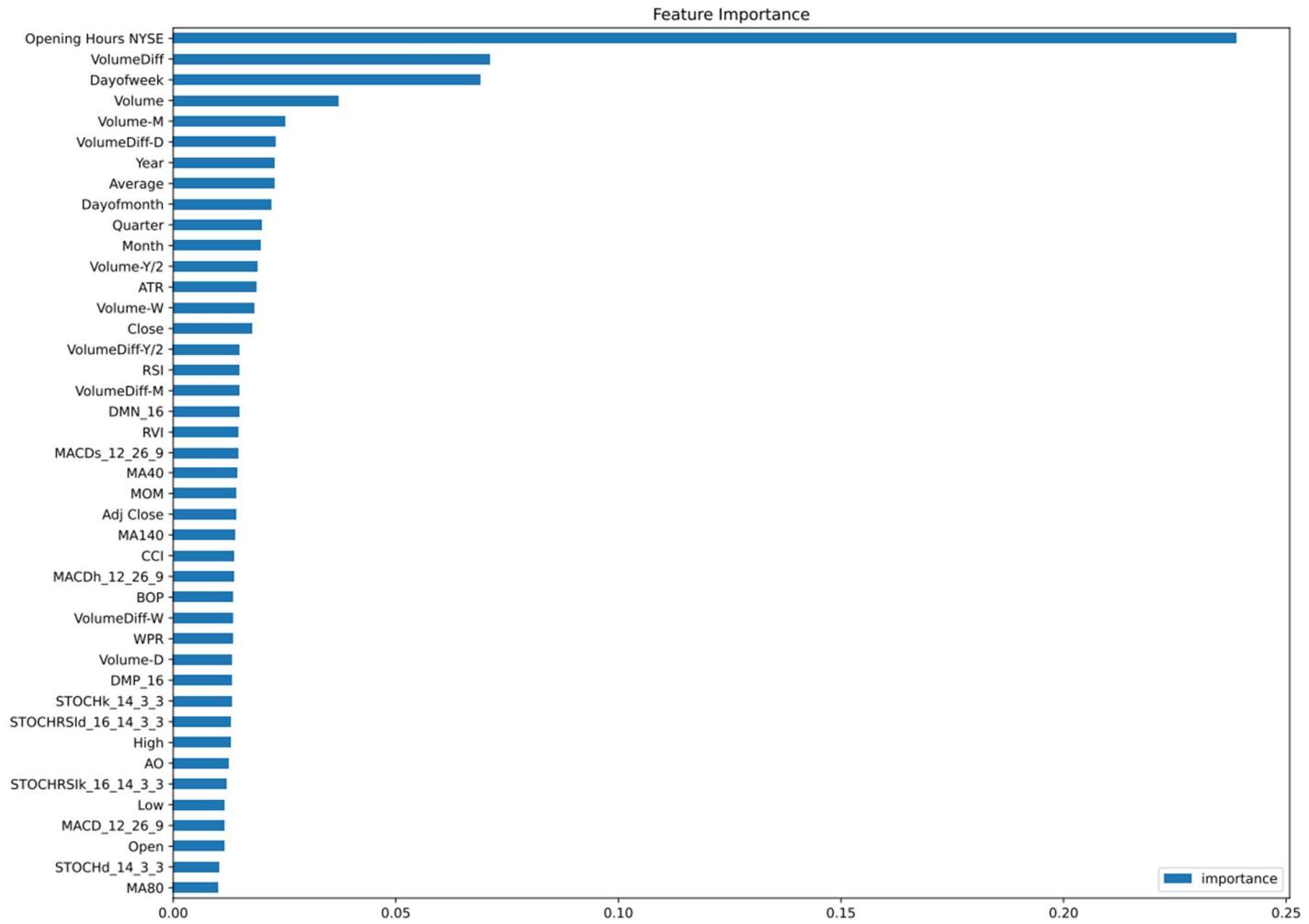
Graf 4: Důležitost jednotlivých Features

MODEL2



Graf 5: Měsíční predikce

Predikce vychází lépe a důležitost features je očekávání. Největší důležitost má množství hodin, které je burza otevřena, dále množství obchodovaných indexů v předchozích dnech, a také konkrétní dny v týdnu, kdy se obchoduje. Tuto naši domněnku nám potvrzují i výsledky, které řadí přesnost modelu na testovacích datech na nejlepší příčku.



Graf 6: Důležitost jednotlivých Features

MODEL3

Tento model má hlavní úkol lépe nastínit získané výsledky. Jelikož neobchodusujeme, je těžké validovat, zda naše predikce znamená, že jsme vydělali nebo nikoli.

Tento model predikuje růst či pokles obchodovaného objemu s přesností 67%, což v případě růstu či poklesu akcií se správně zvolenou strategií obchodování by mohlo vést k velmi hezkým výsledkům. Přestože to není perfektní a toto číslo by mohlo být navýšeno, víme, že jsme na správné cestě. Pro kontrolu chybovosti ve výsledných hodnotách se používá výpočet celkového objemu pomocí průměrné hodnoty změny, která byla dopisána pouze pro úplnost.

MODEL4

Tento model byl vytvořený spíše na vyzkoušení, že obrovské množství vstupních argumentů negativně ovlivňuje výsledek. Přestože se jedná o nepatrny rozdíl, tento model vyšel nejlépe ze všech, přestože měl menší množství vstupních argumentů.

VÝSLEDNÉ HODNOTY

MEAN ABSOLUTE ERROR

- Baseline: 367406593
- Model1: 297394606
- Model2: 289656753
- Model3: 411263513
- Model4: 284199543..

R2 SCORE

- Baseline: 0.058
- Model1: 0.406
- Model2: 0.453
- Model3: 0.000
- Model4: 0.457

SUM OF SQUARED ERRORS (SSE)

- Baseline: 165358263999299977216
- Model1: 104273341689698451456
- Model2: 95999114274289844224
- Model3: 175507204463577399296
- Model4: 95259395231091343360

ROOT MEAN SQUARED ERROR

- Baseline: 575655527
- Model1: 457126473
- Model2: 438614859
- Model3: 593058044
- Model4: 436921722

ZÁVĚR

Tento úkol byl po QMiners Hackathonu dalším skvělým příkladem, že udělat model je snadné, ale udělat dobrý model s očekávanými a smysluplnými výsledky zabere velké množství času. V rámci práce jsem chtěl problematiku co nejvíce pochopit a netrávit hodiny laděním drobných parametrů. Tudíž určitě je stále prostor pro zlepšení.

Nejdůležitější pro získání lepších výsledků je otestování robustnosti, přidání kalendáře s novinkami z finančního světa a přetrénování modelu na inkrementálních datech.

Modely dosahují lepší predikce než baseline, mají 67% úspěšnost predikce trendu a zvolené features v modelu2 a modelu4 jsou přesně podle očekávání. Po úspěšném otestování robustnosti bych vytvořeným modelům věřil.

ZDROJE

KURZY

- TensorFlow: <https://www.deeplearning.ai/courses/tensorflow-data-and-deployment-specialization/>
- AlgoTrading: <https://www.youtube.com/playlist?list=PLwEOixRFAUxZmM26EYI1uYtJG39HDW1zm>

ČLÁNKY

- R2 vs SSE: https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/#Difference_between_Mean_Square_Error_R-Squared
- Interactive Brokers API: <https://algotrading101.com/learn/interactive-brokers-python-api-native-guide/>
- XGBoost Hyperparameters: <https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook>
- Standardization: <https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02>
- HyperParameters: <https://kevinvecmanis.io/machine%20learning/hyperparameter%20tuning/dataviz/python/2019/05/11/XGBoost-Tuning-Visual-Guide.html>
- PyTorch Tutorial: <https://medium.com/edureka/pytorch-tutorial-9971d66f6893>

- Transformers Stock Predictions: <https://towardsdatascience.com/stock-predictions-with-state-of-the-art-transformer-and-time-embeddings-3a4485237de6>
- Optuna Hyperparameters Tuning: https://optuna.org/#code_examples
-

VIDEA

- CodeTrading Playlist:
<https://www.youtube.com/playlist?list=PLwEOixRFAUxZmM26EYI1uYtJG39HDW1zm>
- FB Prophet Model: https://www.youtube.com/watch?v=j0eioK5edqg&ab_channel=RobMulla
- Hyperparameters: PyTorch:
https://www.youtube.com/watch?v=5Xh9FusE8iE&list=LL&index=3&ab_channel=M%C4%B1sraTurp
- Optuna: https://www.youtube.com/watch?v=P6NwZVl8ttc&list=LL&index=5&ab_channel=PyTorch
- TensorFlow vs PyTorch:
https://www.youtube.com/watch?v=ay1E1f8VqP8&list=LL&index=6&ab_channel=PatrickLoeber
- XGBoost: StatQuest:
https://www.youtube.com/watch?v=vV12dGe_Fho&list=LL&index=2&ab_channel=RobMulla
- LSTM: https://www.youtube.com/watch?v=YCzL96nL7j0&ab_channel=StatQuestwithJoshStarmer
- RNNs: StatQuest:
https://www.youtube.com/watch?v=AsNTP8Kwu80&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=85&ab_channel=StatQuestwithJoshStarmer
- XGBoost StatQuest:
https://www.youtube.com/watch?v=OtD8wVaFm6E&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=65&ab_channel=StatQuestwithJoshStarmer