Tradiční modely, netradiční data: Predikce ceny TSLA

Václav Jež[1]

21. května 2025

Abstrakt: Tato práce zkoumá, jak tradiční nástroje časových řad obstojí v nevyzpytatelném prostředí světa predikcí cen akcií. V analýze jsou sesbírána unikátní data (např. tweety Elona Muska), která jsou upravena i méně tradičními způsoby (machine learning, PCA). Porovnává predikční výkonosti různých modelů a přístupů k jejich odhadu. Jako cíl si klade porovnat dle standartních metrik predikce těchto modelů jejich výkonost s anebo bez zahrnutí unikátních dat. Práce také porovnává jednorozměrný a vícerozměrný přístup modelování řad. Nachází, že zvolené modely porážejí naivní model a lze je tak z tohoto hlediska považovat za významné. Zahrnutí unikátních dat sentimentu a dalších v podobě exogenních proměnných nevede ke spolehlivějším výsledkům. Stejný závěr je nalezen pro složitější vícerozměrné modely časových řad oproti jednorozměrným.

Klíčová slova: ARIMA, VAR, TSLA, predikce, akcie

JEL klasifikace: G100, G170

1 Úvod

Svět akcií může být chaotickým a často zrádným místem. Jejich ceny jsou ovlivněny spoustu různými a složitými vztahy. Často se mohou odvíjet od nálady investorů na trhu či různými veřejnými prohlášeními firmy, stakeholderů a vlivných osobností. Spoustu investorů se v tomto často i iracionálním prostředí snaží o profitabilitu a jeden ze způsobů, jak ji dosáhnout může být vytvoření relativně spolehlivých predikcí ceny akcie. I nejlepší a nejsložitější modely mohou mít ale problém se spolehlivým predikováním jejich ceny. Tato práce tedy do tohoto světa vstupuje vybavená s poměrně unikátními daty a vstupuje vstříc testování právě více různých metod ke získání predikcí.

Motivace pro výběr takového úkolu byla záliba v tomto odvětví financí a zvědavost, zdali „základní“ modely časových řad mohou přinést úspěch. Autorův pocit, že na krátkodobé výkyvy ceny akcie může mít značný vliv právě sentiment na trhu, takzvaný „hype“ akcie či důležitá prohlášení vedly k stanovení cíle této práce. Tento cíl spočívá v porovnání výkonnosti predikcí tradičních modelů časových řad a to ARIMA, ARIMAX, VAR a VARX v případě sledované řady cen akcie firmy Tesla. Data využitá pro tyto modely k naplnění cíle práce jsou poměrně unikátní a mimo jiné zahrnují ať už náladu investorů na trhu, relativní vyhledávanost pojmu „tesla“ na Googlu nebo časovou řadu indikující sentiment vyplívající z tweetů Elona Muska. Právě takový výběr proměnných zároveň s vybranými modely, může poodhalit, zdali, zahrnutí nejen tradičních finančních dat vede k lepší výkonnosti predikcí v případě poměrně těžko modelovatelné veličiny jako je cena akcie. Práce používá ticker „TSLA“ hlavně z důvodu poměrně velkého vlivu jejího zakladatele Elona Muska na sociální síti X, kde mívají často jeho

[1] Masarykova univerzita, Ekonomicko–správní fakulta, obor: Ekonomie, 535432@mail.muni.cz

výroky o různých finančních aktivech velký dopad a reakce mezi investory.

2 Teoretický model a cíl

Na základě již zmíněného záměru přivést do analýzy ceny akcie TSLA méně obvyklá data v podobě nějakého sentimentu na trhu, prohlášení či zachycení „hypu“ akcie přichází práce s následující prvotní teoretickou specifikací modelů. Na základě této specifikace lze nejlépe porovnávat přínos těchto proměnných i porovnat různý přístup k modelování časových řad jako jednorozměrný a vícerozměrný.

Obr. 1: Prvotní návrh modelů

Jednorozměrné modely

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{K} \beta_k x_{kt} + \varepsilon_t$$

Vícerozměrné modely

$$y_t = \alpha + \beta(L) y_t + \epsilon_t$$

$$y_t = \alpha + \beta(L) y_t + \gamma x_t + \epsilon_t$$

Zdroj: vlastní zpracování

Na Obr. 1 lze vidět, jak bylo avizováno využití ARIMA, ARIMAX, VAR a VARX modelů. V případě jednorozměrných modelů označuje y_t cenu TSLA v čase t, a u ARIMAX ještě vektor x_t označuje vektor exogenních proměnných. Pro vícerozměrné modely označuje y_t vektor endogenních proměnných zahrnující cenu TSLA a dodatečné finanční veličiny. Verze každého modelu s exogenními proměnnými byla vybrána právě pro zahrnutí zmíněných proměnných sentimentu investorů, či externích šoků informací investorů a „hypu“ této firmy. Taková specifikace umožňuje porovnat přínos těchto proměnných oproti základním verzím modelů. Finanční data také umožňují vyzkoušet přístup s více časovými řadami současně, jelikož buď přímo z ceny takových aktiv či přímo z trhu lze získat spoustu technických finančních indikátorů, které by mohly prozrazovat něco o ceně do budoucna. Proto jsou modely rozdělené právě na jednodušší ARIMA a VAR přístup.

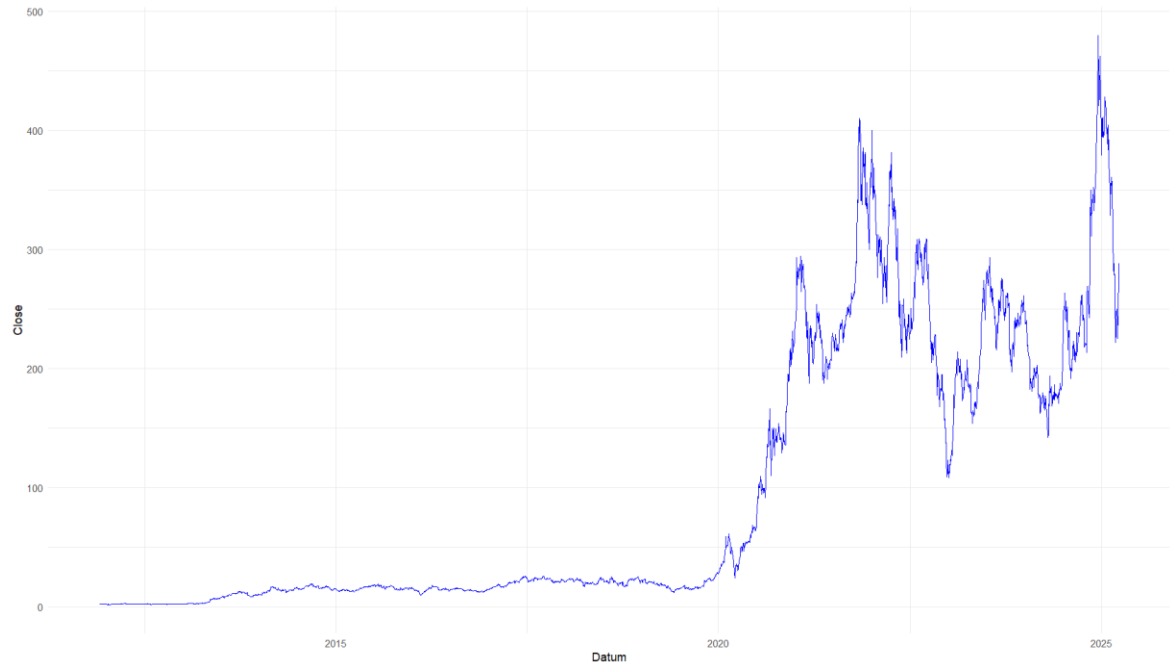
Tyto modely jsou tedy označené jako teoretické, protože se zatím jedná pouze o teorii, s kterou tato práce přichází a bude dále uvedena do konkrétních mezí dále v práci. Cíl práce tedy konkrétně spočívá v porovnání predikcí využitím různých metrik chyb predikce (MSE, RMSE, MAE, MASE). Porovnání predikcí bude hlavně mířit na výkonnost zavedení dodatečných exogenních proměnných v obou přístupech a také zdali zahrnutí více finančních dat (VAR modely) je výhodnější oproti základnímu přístupu pouze s cenou akcie (ARIMA modely).

3 Použitá data a transformace

Na základě prvotní teorii o různých specifikacích modelů představené v předchozí kapitoly byla dále získána data umožňující tuto analýzu. Data pro close cenu akcie TSLA a objemu obchodů na burze byla získána pomocí balíku tidyquant v R. Data pro volatilitu na trhu měřenou VIX indexem byla také získána za pomocí balíku tidyquant z R. Data o vyhledávanosti pojmu „tesla“ na Googlu byla získána z Google Trends (2025). Od Dada Lyndell (2025) byla získána data tweetů Elona Muska mezi roky 2010 až 2025. Data sentimentu investorů na burze byla získána z průzkumů od AAII (2025). Dále byly využité různé proměnné technických indikátorů, které byla však spočítány přímo z proměnné close. Pro data byla vybrána denní frekvence. Výsledný časový vzorek pro všechna data byl omezen na období 1. 12. 2011 až 25. 3. 2025. Toto je období po odstranění NA hodnot vzniklých při počítání některých indikátorů.

Cílovou proměnnou pro predikce je tedy close (uzavírací cena na burze daný den v dolarech). Vývoj v čase této proměnné může být vidět na Obr. 2.

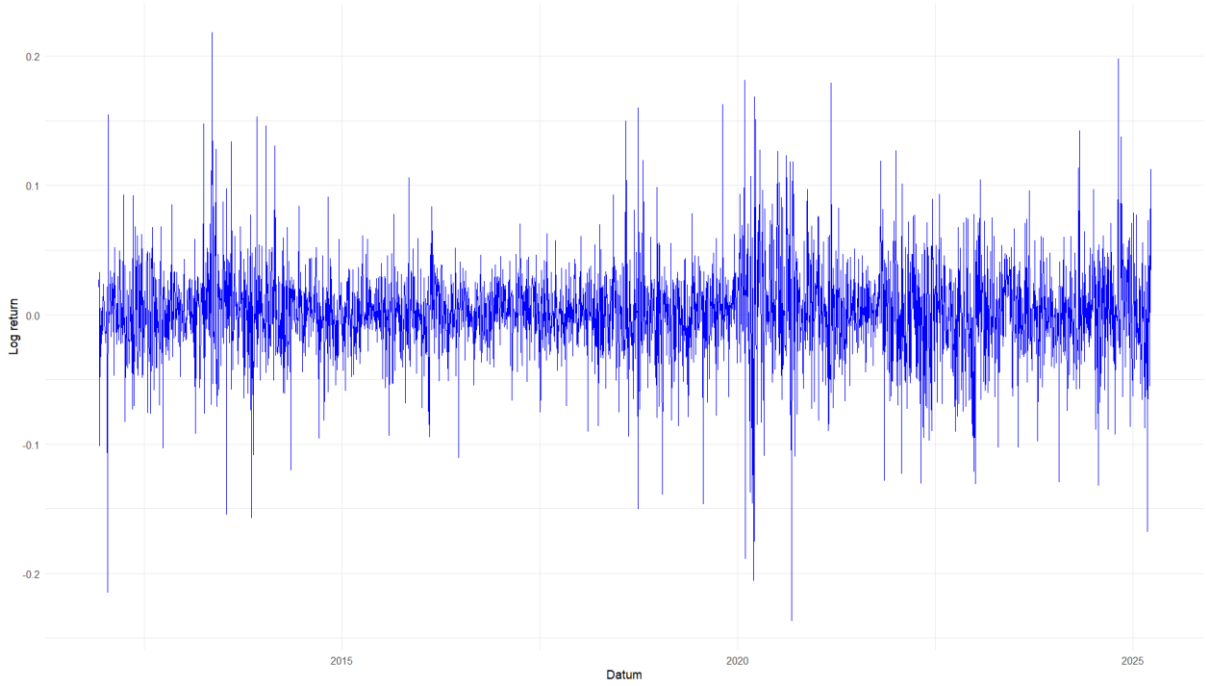
Obr. 2: Vývoj close na celém odbobí



Zdroj: vlastní zpracování v R

Tato proměnná je dále převedena na tzv. log return pomocí diference logaritmu. Tento krok je zvolen hlavně pro splnění podmínky stacionarity pro tuto řadu. Výsledná řada už na základě ADF testu zamítá nulovou hypotézu o jednotkovém kořenu, a tak je připravená pro vstup do modelů.

Obr. 3: Vývoj log return na celém období



Zdroj: vlastní zpracování v R

Následuje proměnná volume (objem obchodů akcie TSLA), která do modelů vstupuje ve své základní podobě.

Z dodatečných veličin ceny je odvozena proměnná basic volatility, vyjádřená jako high – low. Měří jednoduchým způsobem volatilitu v rámci každého dne.

Další proměnné lze označit společně do bloku technických indikátorů odvozených z ceny. Konkrétně se jedná o SMA 20, SMA 50, ATR, RSI, MACD, MACD signal, OBV, stochRSI a ADX. Veličiny SMA a MACD vyjadřují trend. MACD signal je pak nákupní / prodejní signál z hlediska MACD. ADX měří sálu trendu, bez směru. RSI a stochRSI měří momentum obchodů dané akcie. ATR značí volatilitu. OBV vykazuje kumulativní objem obchodů s ohledem na směr cenového pohybu. I spolu s ostatními proměnnými by hrozil problém s vysokým počtem proměnných v každém modelu, a tak analýza dále směřuje zmíněné technické proměnné do PCA analýzy. Tyto technické proměnné také bývají často korelované, což je solidní základ právě pro PCA. PCA analýza umožňuje snížit dimenze analýzy, a tak může dále vést k lepším odhadům modelů. Do PCA je celkem rozumné posílat stacionární proměnné, aby se např. přítomný trend dále nepromítal místo informacích v proměnných do výsledných PC. Z těchto proměnných jsou tedy SMA 20, SMA 50, ATR a OBV nejprve diferencovány, dále jsou všechny proměnné tohoto bloku normalizovány na průměr = 0 a sd = 1 (jako je běžné u PCA), a je na nich provedena samotná PCA. Na základě Scree plot na celém vzorku dat, jsou vybrány první 4 PC (vysvětlující 80,8 % variance) dále jako vstupy do modelů.

Závěrečné další proměnné lze uvést v bloku jako právě exogenní proměnné obsahující unikátnější data. Jedná se konkrétně o VIX close, což vyjadřuje uzavírací hodnotu indexu VIX měřící volatilitu na celém trhu. Tato proměnná je zvolena jako exogenní z důvodu relativně malého zastoupení firmy Tesla na celém americkém trhu, s praktickou nemožností index významně ovlivňovat a jedná se tak spíše o externí šok. Další proměnnou jsou 4 umělé proměnné vytvořené následujícím způsobem. Zmíněná data tweetů Elona Muska jsou nejprve filtrována na klíčová slova jako „tesla“, „tsla“, „tesla stock“, „earnings“, „model (označení

modelu)“, cybertruck“ eletric car“ a podobně. Jsou vybrána tak aby tweety obsahovali pouze výroky s možností ovlivnit investory a tím cenu TSLA na trhu. Dále jsou agregovány všechny tweety za jeden den právě do jednoho výroku na den. Pak je provedena sentiment analýza za pomocí machine learning opensource modelu FinBERT, která na základě těchto výroku vrací nejvíce pravděpodobný sentiment vyplívající z výroku, a to ve 3 úrovních (neutral, positive, negative). Tento sentiment tedy vyjadřuje celkový sentiment všech tweetů za daný den. Dále jsou označeny dny bez tweetu úrovní „none“. Úroveň none je využitá pouze jako referenční úroveň a do modelu vstupují pouze proměnné neutral, positive, negative. Další proměnou jsou již avizována data z Google trends. Tato data nabývají hodnot mezi 0–100 v závislosti na relativní vyhledávanosti v zvoleném časovém období. Vyhledávaný termín je přímo slovo „tesla“. Jako oblast jsou vybrány pouze Spojené státy, které by měly mít na US trh největší a nejrychlejší vliv. Google trends na delší období dodává pouze měsíční frekvenci dat, a tak jsou data převedena na denní způsobem replikování dané měsíční hodnoty každý den v daném měsíci. Jako poslední dvě proměnné tohoto bloku jsou proměnné sentimentu z průzkumu mezi investory. Tato data obsahují procentuální zastoupení investorů, co jsou bulish (cena poroste), bearish (cena klesne) anebo neutral. Zde se jedná o týdenní data, a tak jsou stejným přístupem jako pro Google trends převedena na denní. Z bulish a bearish je vytvořena proměnná jejich spreadu. Proměnné vstupující do modelu jsou tedy tento bull bear spread a zastoupeni neutral investorů. Všechny exogenní proměnné tohoto bloku byly dále zpožděny o jedno období (1 den), z důvodu předpokladu zpožděné reakce trhu na takové informace a také z důvodu následného predikování na 1 den do budoucna, který by jinak vyžadoval i predikci těchto proměnných. V příloze B jsou uvedeny grafy ukazující spojitost mezi vývojem bulish / bearish investorů a close ceny TSLA a také sentiment vyplívající z tweetů Elona Muska na celém vývoji close.

Všechny výsledné proměnné po úpravách a transformacích byly pomocí ADF testu testovány na přítomnost jednotkového kořene na hladině významnosti 5 %. V každém případě byla nulová hypotéza zamítnuta indikující stacionaritu na celém vzorku dat.

4 Odhady modelů

S nyní již kompletním setem proměnných jsou modely připravené pro odhady. Do ARIMA a ARIMAX modelů vstupuje vždy log return podoba close a v případě ARIMX ještě blok exogenních proměnných, a to tweets_sentiment – neutral, positive, negative a VIX index, Google trends data, bull-bear spread, a neutral survey sentiment. V případě VAR modelů je to vždy log return podoba close, volume, basic volatility a PC1, PC2, PC3, PC4 komponenty získané z PCA analýzy jako set endogenních proměnných. Pro VARX je to stejný blok endogenních s přidanými stejnými exogenními proměnnými jako v případě ARIMAX. Exogenní proměnné byly do modelu zavedeny se zpožděním, jak již bylo zmíněno, což umožňuje jejich reálné použití při predikci bez nutnosti jejich budoucí znalosti. Při každé predikci byly exogenní vstupy aktualizovány o nejnovější dostupné hodnoty, které nebyly použity při tréninku modelu, ale jsou v čase predikce realisticky dostupné, tj. dodržuje se reálná časová dostupnost dat. Když tedy model predikuje close na čas t + 1, a vyžaduje budoucí hodnoty exogenních proměnných v čase t + 1, z důvodu jejich zpoždění jsou tyto hodnoty dostupné a jsou to hodnoty fakticky v čase t.

Metoda odhadu v této práci je rolling window predikce. Práce sleduje pouze jednoikrokové predikce. V případě dat o akcii mohou být podmínky na trhu i pozice dané firmy často variabilní v čase, a tak by mohlo rolling window být více vystihující než např. expanding window, kde mohou zůstat „staré“ vztahy v datech. Velikost okna byla vybrána jako 1202 dní z důvodu

vysokého počtu parametrů (velký počet proměnných), které musí být odhadnuty. Toto okno (2 dny navíc) je vybráno i z důvodu různého řádu zpoždění mezi modely, a tak nutnosti uměle zarovnávat data v každém běhu tak, aby nedošlo k odhadům modelů na různém vzorku, a tak nemožnosti modely přímo porovnávat. Zpoždění byla nejdříve vybrána na základě funkce auto.arima a VARselect přímo v R nejprve na celém vzorku dat. Tato zpoždění byla dále otestována již na sledované rolling window metodě. Dále byla manuálně otestována další různá zpoždění. Výsledné vybrané řády jsou tedy (p = 1, d = 0, q = 1) pro ARIMA a ARIMAX a (p = 2) pro VAR a VARX. Tato zpoždění vedla k nejlepším predikčním výsledkům. Je také důležité zmínit, že v případě ARIMA a ARIMAX byl vybrán „CSS-ML“ estimátor, který zajistil lepší konvergenci a stabilitu parametrů než defaultní „CSS“. Také zmiňovaná PCA analýza proměnných byla provedena vždy odděleně v každém běhu každého okna odhadů z důvodu vyhnutí se problému data leakage.

5 Predikce modelů

S připravenými daty, popsaným způsobem odhadu modelů jsou tedy odhady provedeny. Dále jsou spočítány kritéria hodnocení predikcí získaných z rozdílu realizovaných a predikovaných hodnot. Mezi modely je také uveden naivní model jako kritérium, zdali mají modely vůbec smysl. Metriky ukazuje Tab. 1.

Tab. 1: Metriky odhadnutých predikcí

MODEL	MSE	RMSE	MAE	MASE
ARIMA	**0,001448**	**0,038052**	**0,026705**	**0,682347**
ARIMAX	**0,001465**	**0,03827**	**0,026838**	**0,685761**
VAR	**0,001463**	**0,03825**	**0,0271**	**0,692455**
VARX	**0,001478**	**0,038438**	**0,027245**	**0,69616**
NAIVE	**0,002915**	**0,053991**	**0,039136**	**1**

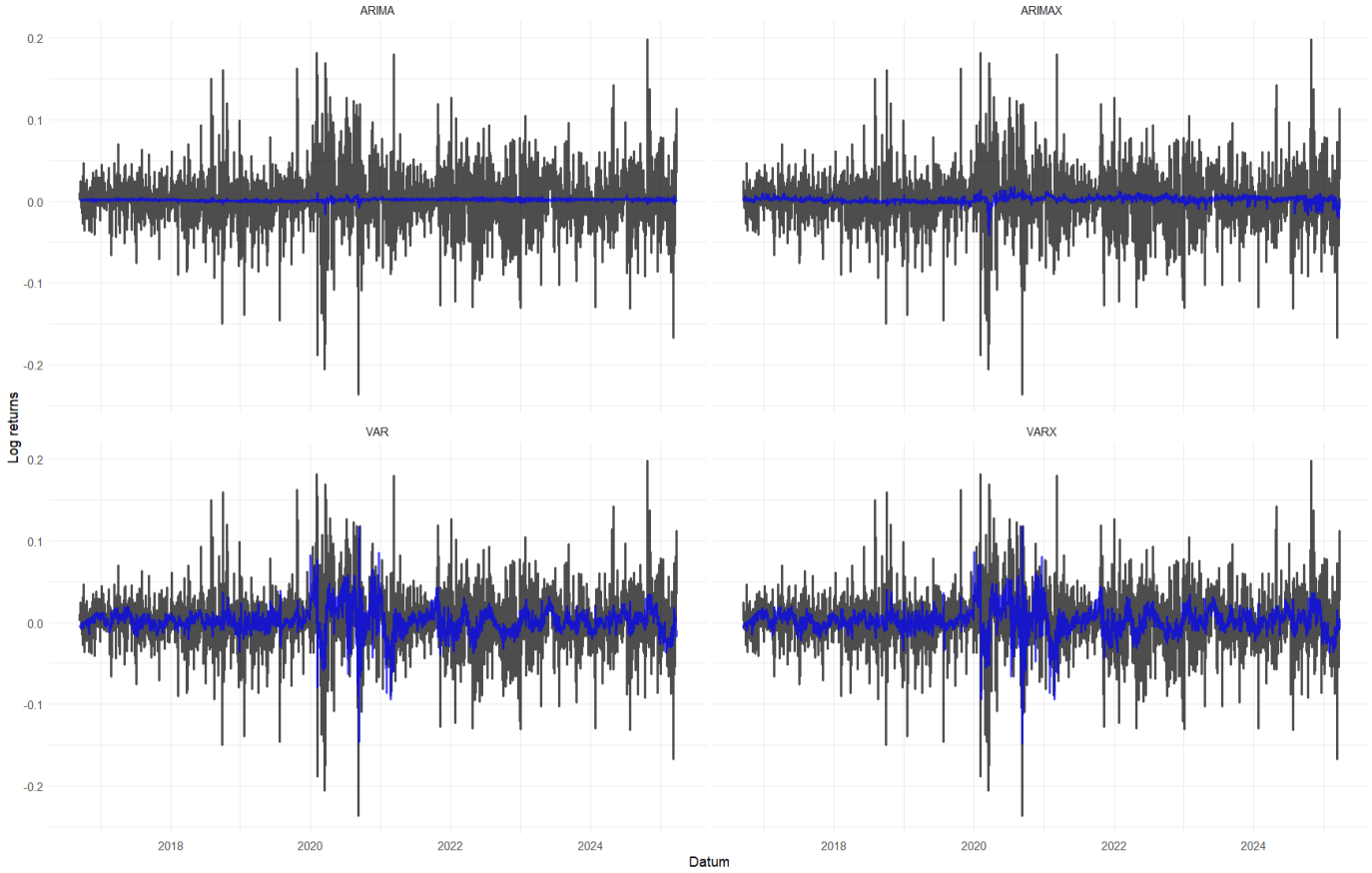
Zdroj: vlastní zpracování

Všechny modely úspěšně porážejí naivní model, jak je vidět dle hodnot MASE, které jsou v každém případě < 1. Tento fakt je vidět také na jednotlivých ostatních metrikách predikcí. Práce tedy shledává úspěch ve významnosti a využitelnosti všech odhadnutých modelů. Směrodatná odchylka log return akcie TSLA (tj. skutečných hodnot) má hodnotu 0.03601. V případě RMSE tedy je tedy chyba modelu přibližně stejně velká, což naznačuje, že predikční chyba modelu je srovnatelná s přirozeným kolísáním řady. U MAE je tento poměr nižší, což může svědčit o poměrně kvalitní predikční schopnosti modelů vzhledem k volatilitě řady. Predikce všech modelů graficky zobrazuje Obr. 4.

Výsledky implikují následující závěry. Zahrnutí proměnných sentimentu z průzkumů, indikovaných sentimentů týkajících se firmy Tesla zveřejněné Elonem Muskem, Google trends data a index VIX nepřináší lepší predikční výkonost na zvoleném přístupu modelování při zvoleném časovém vzorku (ARIMA vs ARIMAX, VAR vs VARX). Tyhle predikce dosahují dokonce o něco horší výsledky. Jsou to, ale poměrně velmi malé rozdíly, které nemusí implikovat špatnou metodologii nebo zvolená data. Každopádně zahrnutí těchto exogenních proměnných nepřináší žádné zlepšení, a tak je jednodušší modelování v případě této práce vhodnější.

Podobný případ práce nachází v případě porovnávání jedno a více rozměrného způsobu predikování log return akcie TSLA. Odhad modelu zahrnující navíc objem obchodů na burze, proměnnou basic volatility a PCA komponenty odhadnuté na různých technických indikátorech také přináší horší výsledky než zahrnutí pouze log return mezi vysvětlované proměnné (ARIMA vs VAR, ARIMAX vs VARX). Práce tedy nachází jako nejspolehlivější metodu (dle všech uvedených metrik) nejjednodušší ARIMA model, se vstupem log returnu akcie TSLA.

Obr. 4: Predikce všech modelů



Zdroj: vlastní zpracování v R

Vizuálně vykazují predikce obou VAR modelu lepší vzor, ale je také možné, že je model přeučeny na trenovacích datech tzv overfitting, jelikož predikce od ARIMA, které více oscilují okolo 0 vykazují lepší predikční metriky. Tuto hypotézu dále potvrzuje fakt, že při VAR (16) či VAR (4) byly výsledky horší, a tak model muže byt spíše přeparametrizován. VAR (1) byl ale stejný či lehce horší než VAR (2). A tak je zpoždění 2 nejspíše nejlepší přístup kde se však stále může jednat o nějaký problém, např. s počtem parametru modelu. Práce tedy naráží na možný limit overfittingu, který by si zasluhoval hlubší průzkum. ARIMA predikce vykazuji zhruba bílý šum kde střední hodnota predikcí je blízko 0 a vyskytuji se zde náhodně výkyvy. Toto se shoduje s hypotézou, že ceny akcií často následují model náhodné procházky. Práce také dokazuje základní ekonometrický odhad modelů a transformaci dat, analýza však zasluhuje další průzkum v praktické aplikaci a reálném tzv. backtestu s modelovým portfoliem, kde se zodpoví otázka, zdali se dají takto odhadnuté modely s poměrně dobrými výsledky i dobře využít i v reálném investičním prostředí.

6 Závěr

Práce nejdřív popsala teorii, kterou práce používá k naplnění cíle práce, spočívajícím v porovnání různých tradičních modelů časových řad s různými zahrnutými unikátními daty. Představila prvotní specifikace porovnávaných modelů a přístupu k nim. Dále práce popsala jak už základní data, tak unikátnější data sentimentu z průzkumů, či získaných machine learning sentiment analýzou tweetu Elona Muska. Tato kapitola popsala také všechny využité transformace a úpravy dat pro účel odhadů modelu. Práce dále specifikovala, že odhady budou sestrojeny způsobem rolling window, a jak každý běh probíhá. Nakonec práce vyhodnotila výsledky analýzy. Cíl práce je tedy hodnocen jako naplněný s následujícími závěry.

Práce nachází poměrně spolehlivé predikce všech modelů, které významně porážejí naivní model. Chyby predikce se pohybují zhruba na úrovní přirozené kolísavosti log returnu akcie TSLA. V případě porovnání zahrnutí exogenních proměnných do ARIMA a VAR modelu práce nachází skoro stejnou či méně dobrou výkonnosti predikcí těchto modelů. A tak unikátní data sentimentu na trhu, vyhledávanosti firmy Tesla na googlu, signál vyplývající z výroků Elona Muska na sociální síti X a VIX indexu nepřináší žádný zlepšení odhadnutých predikcí. Podobný výsledek je nalezen i při porovnání přístupu ARIMA vs VAR. Zahrnutí více endogenních proměnných jako objem obchodů na burze, základní volatilitou meřenou jako rozdíl high a low ceny TSLA, PCA komponenty získané z různých technických finančních indikátorů také nepřináší spolehlivější výsledky chyb predikce. Nejlepší výsledky analýza vykazuje pro jednoduchý ARIMA (1, 0, 1) model. Teoretický základ avizovaný na začátku práce, kde by měl sentiment investorů, „hype“ akcie či volatilita na celém trhu výrazně ovlivňovat chování investorů, a tak zahrnutí takových veličin do predikcí změn ceny akcie TSLA by mělo zlepšit predikční výkonnost se tedy nepotvrzuje. Práce také vyzývá k dalšímu průzkumu zvolených modelů na reálném testování profitability trading strategie postavené na jejich predikcích.

Reference

[1] AAII, 2025. The AAII Investor Sentiment Survey [online]. Dostupné z: https://www.aaii.com/sentimentsurvey

[2] DADA LYNDELL, 2025. Elon Musk Tweets 2010 to 2025 (March) [online]. Kaggle, Dostupné z: https://www.kaggle.com/datasets/dadalyndell/elon-musk-tweets-2010-to-2025-march

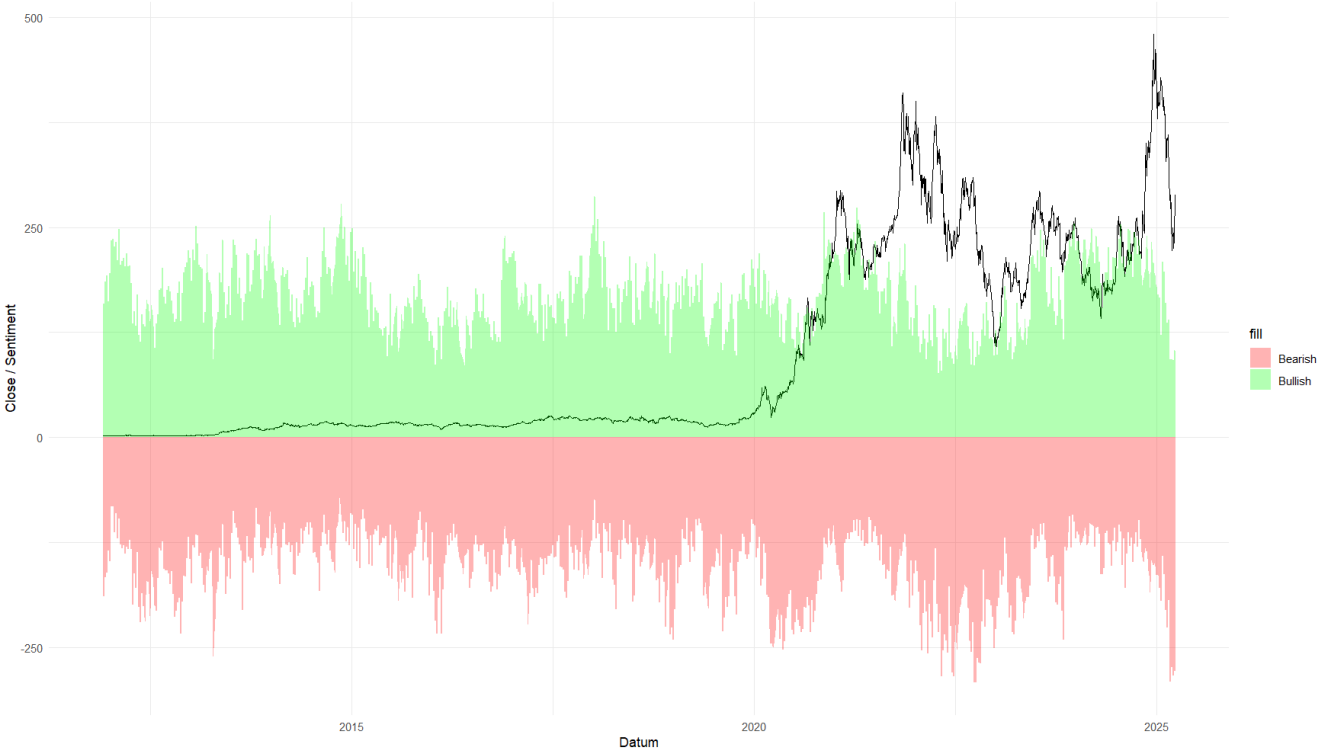
[3] GOOGLE TRENDS, 2025. Tesla [online]. Dostupné z: https://trends.google.com/trends/explore?date=2010-01-06%202025-05-01&geo=US&q=Tesla&hl=cs

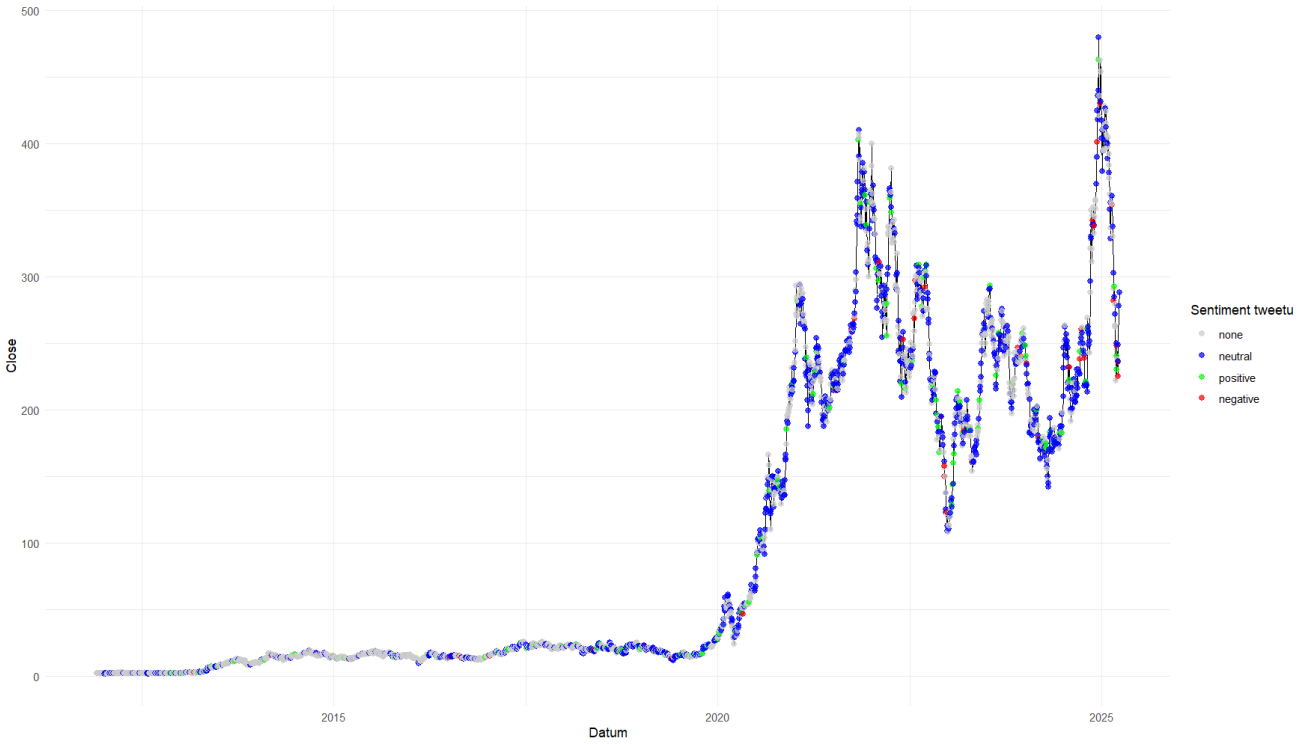
Přílohy

Příloha A

Všechna data, v podobě, ve které byla stažena i v transformované podobě, spolu se všemi skripty obsahující jejich úpravy a sentiment analýzu a také grafy jsou dostupné v archivu odevzdaném s tímto souborem.

Příloha B




500
400
300
200
100
0
Close
2015
2020
2025
Datum
Sentiment tweetu
none
neutral
positive
negative