FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

odb. as. Ing. Kamil Ekštein, Ph.D.
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd Západočeské univerzity
Univerzitní 22, CZ-30614 Plzeň
Tel.: (+420) 377 63 2406
E-mail: kekstein@kiv.zcu.cz

Posudek oponenta bakalářské práce

Max Nonfried
High Frequency Currency and Cryptocurrency Trading

Anglicky psaná bakalářská práce Maxe Nonfrieda se zabývá problematikou aplikace jedné z klasických statistických metod analýzy dat, jmenovitě predikčního modelu ARIMA, na vysokofrekvenční data obchodování s měnovými páry. Autor si jako cíl pak mj. vytkl kvalitativně porovnat základní používané investiční strategie "Mean-reversion" (tj. návrat ceny k běžné dlouhodobé hodnotě) a "Momentum investing" (tj. setrvačnostní investování do aktiv, které měly dlouhodobě vysoké výnosy) se strategií založenou na predikcích modelu ARIMA. Vytčený cíl (a požadavky zadání práce) se mu v zásadě podařilo velmi dobře splnit.

Hned na úvod je ovšem třeba říci, že autorem použité datové sady nejsou zcela typickým zástupcem dat tzv. high-frequency tradingu (HFT), jelikož jsou stále ještě dost „pomalé" a premisy, na kterých jsou vystavěny komerční HFT algotradery (tedy zejména kvazistacionarita trendu ceny v periodě obchodování), zde sice mohou, ale nemusejí být nutně splněny. Nicméně s ohledem na metodologii autorem provedených testů a porovnání to zřejmě nehraje roli.

Text práce je rozumně rozvržen a pokrývá jak potřebné prerekvizitní teoretické poznatky, tak následně metodologii a na ní vystavěnou implementaci kódu k provedení konkrétních experimentů. Za poněkud neobvyklé u kvalifikační práce odevzdávané na Katedře informatiky a výpočetní techniky považuji absenci samostatné kapitoly věnující se implementaci programového vybavení. Tuto část autor neopominul, ale nestandardně ji vřadil do kapitoly "Methodology", což nepovažuji za úplně šťastné řešení, přinejmenším proto, že metodologie by jistě měla být nezávislá na použitém vývojovém prostředku a technologii implementace.

Předloženou práci lze jen těžko popsat jinak, než jako téměř profesionální dílo. Autor velmi pečlivě a srozumitelně popsal výchozí teoretické poznatky, vysvětil jak použité statistické postupy a predikční techniky, tak také testy a rozebral dostupná data, na kterých jsou testy provedeny. Konkrétně v případě rozboru dat bych ale uvítal více informací, případně grafů (histogramů, apod.). Autor sice uvádí, že se mu nepodařilo prokázat, že by např. distribuce operací v čase odpovídala kterémukoliv z testovaných rozdělení, nicméně nějaký např. scatter plot alespoň části těchto dat by jistě čtenáři lépe umožnil udělat si nějakou konkrétnější představu o podobě zpracovávaných dat.

Kvalitu zdrojového kódu v jazyce Python v podstatě není možné hodnotit, neboť se jedná o krátké tzv. snippety kódu do frameworku Backtrader, z nich nejdelší má zhruba 4 KB a 123 řádek, a tedy o nějaké např. časové nebo paměťové optimalitě není vlastně ani jak debatovat (tím spíše, že autor kód v Pythonu používá pouze jako skript volající knihovní funkce). Zdrojový text je zapsán ve shodě se zvyklostmi, je rozumně okomentovaný a dobře čitelný. Práce zjevně není typickou programátorskou kvalifikační prací, nicméně bod 4. zadání předložené dílo jistě splňuje.

Nezpochybnitelný je ovšem rozsah práce analytické, ať už při studiu statistických metod a technik HFT (což jsou oblasti nepokryté studijním plánem v autorově studijním programu) nebo při zkoumání a zprovozňování použitých nástrojů, ale zejména pak při vyhodnocování dosažených výsledků. To je také nejdůležitější a nejzásadnější přínos práce: Autor velice přehledně ukazuje, jakých výsledků (a to přímo v pojmech zisku či ztráty na burze, nikoliv prostřednictvím nějaké těžko uchopitelné chybové metriky) dosahují při obchodování analyzované strategie klasické v porovnání se strategií založenou na predikci hodnoty aktiva pomocí ARIMA modelu. Tyto výsledky jsou velmi kvalitní, mimořádně zajímavé a velmi dobře dále využitelné.

Po formální stránce je práce na vynikající úrovni: Text je napsaný výbornou odbornou angličtinou s naprostým minimem chyb (jen občas někde chybí nebo naopak přebývá člen), čte se velmi dobře. Autorův styl neunavuje, nenudí, je přiměřeně popisný a není zbytečně epický. Důležité části textu, např. definice pojmů, sdělení zásadní pro pochopení, apod. by ovšem mohly být více (míněno častěji) zvýrazňovány.

Sazba je provedena v LaTeXu a výsledný dokument působí velice harmonickým dojmem. Text je vhodně doplněn množstvím vzorců, které jsou vysázeny ve shodě se zvyklostmi, až na vzorce (2.13) a (2.14), kde na levé straně rovnice stojí technicky vzato součin proměnných A, I a C, resp. B, I a C, což ale autor nepochybně nezamýšlel (všude jinde jsou ovšem vícepísmenné identifikátory členů vysázeny správně stojatým písmem).

Obrázky jsou pěkné a v akceptovatelné technické kvalitě, ale je jich s ohledem na charakter práce málo (přičemž o účelnosti některý pak lze úspěšně pochybovat, např. obr. 3.2). Jediný obrázek, který technickou kvalitou nestačí a neodpovídá úrovni práce, je obr. 3.3, který je bitmapový a v nepřijatelně nízkém rozlišení.

Výhrady lze mít k použitému způsobu provedení citací. Autor zvolil subformát normy ISO 690 obvykle označovaný "author-year", kdy identifikátor citovaného díla tvoří příjmení autorů oddělená středníkem následovaná čárkou a rokem vydání publikace. To by samo o sobě bylo zcela v pořádku, jedná se o legitimní možnost. Problém spatřuji v tom, že tento způsob identifikace díla se obvykle uzavírá do kulatých závorek v případě, že identifikátor není podmětem nebo předmětem věty, což ale autor neučinil, takže v některých pasážích není zcela patrné, kdy se jedná o součást textu a kdy o identifikátor citovaného díla (typicky např. posl. odst. na str. 5, dále velmi nepřehledné úvodní odstavce části 2.1, atd.).

Drobnou technickou vadou sazby je velikost fontu URL adres v poznámkách pod čarou, která není snížena na stupeň odpovídající velikost textu této poznámky (tedy `\footnotesize`).

Autor využívá v práci přiměřené množství titulů odborné literatury (celkem 14), jejíž výběr dobře odpovídá tématickému zaměření práce. Jde z větší části o knihy, v menší míře pak o odborné konferenční články a webové tutoriály. Zdroje korektně cituje, až na výše uvedenou nepříliš podstatnou technikálii. Také na str. 7 uvedený identifikátor zdroje "Office of the Federal Register; Records, 2010, p. 281" není zcela v souladu s ISO 690 author-year, neboť ani "Records", ani "Office of the Federal Register" nejsou **autoři** citovaného díla.

Všechny body zadání práce byly splněny. Splnění bodu 2. („Zpracovat podrobnou rešerši metod detekce a predikce různých režimů v časových řadách vysokofrekvenčních finančních dat.") je ale poněkud diskutabilní. Z pozice oponenta práce není dost dobře možné posoudit, zda autorem předložená rešerše je dostatečně „podrobná". Subjektivně bych očekával poněkud vyšší míru podrobnosti popisu zmíněných metod a v práci mi to trochu chybí – ovšem pokud je autorovo zpracování dostatečné z pohledu zadavatele práce, lze souhlasit, že bod 2. zadání byl také splněn.

Práce je nepochybně vynikajícím dílem spíše vědecko-výzkumného charakteru, samozřejmě s přihlédnutím k očekávané odborné erudici autora na bakalářské úrovni. Autor prokázal schopnost zpracovat poměrně náročné teoretické partie, navíc z oblasti, která nebyla zcela pokryta jeho studijním plánem. Zcela jistě disponuje i potenciálem k další vědecko-výzkumné práci.

Práci každopádně **doporučuji k obhajobě** a i přes výše uvedené výhrady hodnotím klasifikačním stupněm

„výborně".

Ing. Kamil Ekštein, Ph.D.
KIV FAV ZČU

V Plzni dne 30. května 2023

Otázky k práci:

1. V sekci 4.2 uvádíte výsledky testu, zda frekvence příchodu požadavků na obchodování odpovídá nějaké distribuci. V metodologické části zmiňujete, že balíky `fitter` a `distfit` dokážou otestovat shodu s 89 různými rozděleními, ovšem v sekci 4.2 pak jsou vyhodnoceny jen 4: Znamená to, že jsou to ty, které byly ve smyslu zmíněné metriky SSE nejvíce podobné? Nebo z jakého důvodu byla vybrána právě ta uvedená rozdělení?

2. Nedávalo by více smysl ukázat v práci graf shody použitých dat s některou z distribucí, které se běžně pro modelování času mezi příchody požadavků používají, tedy např. exponenciální a Poissonovskou? Vypovídal by v takovém případě výsledek analýzy více o charakteru dat?