Raport z projektu zaliczeniowego przedmiotu Uczenie Maszynowe w finansach

Wstęp

Autorzy:

- Maciej Miniszewski,
- Daniel Struzik,
- Zuzanna Szymańska.

Link do repozytorium: https://github.com/Daniel-Struzik/UMwFinansach/

Opis założeń i celu projektu

Projekt ma na celu przygotowanie trzech modeli przewidujących stopę zwrotu spółki biorąc pod uwagę m.in. miarę `conviction`, która określa jak dobrze spółka poradzi sobie w notowaniach, gdzie 0 oznacza źle, a 1 oznacza dobrze.

Projekt przygotowano w języku `Python` z wykorzystaniem pakietów takich jak `numpy`, `pandas`, `scikit-learn`, `matplotlib`, `seaborn`, `keras`, `xgboost` oraz `lightgbm`.

Do realizacji projektu wybrano trzy modele porównywane na tle regresji liniowej: XGBoosting, LSTM oraz LightGBM.

Wybrane modele przewidują stopę zwrotu spółki na podstawie szeregu zmiennych, w pięciu przedziałach czasowych - cenę akcji w danym dniu (`RoR_date`), cenę akcji za miesiąc (`RoR_mtd`), cenę akcji za kwartał (`RoR_qtd`), cenę akcji za pół roku(`RoR_htd`) oraz za rok (`RoR_ytd`).

Do oceny trafności predykcji została wybrana miara `RMSE` oraz walidacja krzyżowa z wykorzystaniem tej miary.

Użytkownik korzystając ze skryptu `main.py` wprowadza interesujący go okres prognozy, otrzymując na wyjściu miary `RMSE` oraz wynik walidacji krzyżowej dla wszystkich modeli. Prognozowanie odbywa się na podstawie pliku `convictions_returns.csv` przygotowanego w kolejnym rozdziale. W skrypcie `main.py` nie zawarto procesu pobierania danych z `yfinance` w celu szybszego działania skryptu - pobieranie danych z wykorzystaniem API trwa ok. 20-30 minut. Jeżeli użytkownik chce wykorzystać inne dane, powinien najpierw przygotować dane z użyciem pliku `UMF_project_data_prep.ipynb` zamieszczonego na repozytorium GitHub, i następnie umieść gotowy plik nazwany `convictions_returns.csv` w folderze z plikiem `main.py`.

Przygotowanie i eksploracja zbioru danych

Przygotowanie zbioru danych

Przygotowanie danych do prognozy składało się z dwóch etapów:

- Wyczyszczenia bazowego zbioru danych,
- Pobraniu cen akcji spółek oraz feature enginnering,
- Połączenie zbiorów danych.

Szczegółowy kod wraz z opisami zawarty jest w repozytorium w notatniku pod nazwą `UMF_project_data_prep.ipynb`.

Pierwszy etap

Bazowy zbiór danych zamieszczony na platformie moodle - `CONVICTIONLISTTOPN_BSLD-408.csv` zawiera miarę `conviction` spółek z portfela inwestycyjnego, które opisane w następujący sposób:

- Kolumna `0` - zawiera wyciąg logu z systemu,
- Kolumna `1` - zawiera datę określenia miary `conviction`,
- Kolumny `2` oraz `3` - zawierają symbol spółki,
- Kolumna `4` - zawiera rodzaj przemysłu, w którym działą dana spółka,
- Kolumna `5` - zawiera identyfikator spółki,
- Kolumna `6` - zawiera miarę `conviction`.

Nieprzetworzony zbiór danych zawiera 7 kolumn oraz 37360 wierszy.

	0	1	2	3	4	5	6
0	10:01:54.481 77425 [77425-thread-2] INFO a.s....	2004-02-11	SU	SU	Energy Minerals	GN63J3-R	0.953727
1	10:01:54.481 77425 [77425-thread-2] INFO a.s....	2004-02-11	GGG	GGG	Producer Manufacturing	H5490W-R	0.952753
2	10:01:54.481 77425 [77425-thread-2] INFO a.s....	2004-02-11	WGR	WGR	Energy Minerals	V0622Q-R	0.947634
3	10:01:54.481 77425 [77425-thread-2] INFO a.s....	2004-02-11	CWT	CWT	Utilities	GSWXLY-R	0.934181
4	10:01:54.481 77425 [77425-thread-2] INFO a.s....	2004-02-11	BLL	BLL	Process Industries	VFT0VQ-R	0.922862
...	...	...	...	...	...	...	...
37355	10:27:03.049 77425 [77425-thread-2] INFO a.s....	2022-02-09	PEP	PEP	Consumer Non-Durables	PPCTFP-R	0.701507
37356	10:27:03.049 77425 [77425-thread-2] INFO a.s....	2022-02-09	SSNC	SSNC	Technology Services	G92RX2-R	0.701123
37357	10:27:03.049 77425 [77425-thread-2] INFO a.s....	2022-02-09	GEF	GEF	Process Industries	MPX0N4-R	0.697954
37358	10:27:03.049 77425 [77425-thread-2] INFO a.s....	2022-02-09	DPZ	DPZ	Consumer Services	F05QG0-R	0.697741
37359	10:27:03.049 77425 [77425-thread-2] INFO a.s....	2022-02-09	LIFZF	LIFZF	Non-Energy Minerals	Q404Y1-R	0.695644

37360 rows × 7 columns

W celu wyczyszczenia oryginalnego zbioru danych, usunięto następujące kolumny:

- `0` - nie zawiera istotnych informacji,
- `3` - jest identyczna z kolumna `2`,
- `5` - kolumna również nie zawiera istotnych infromacji.

Następnie zamieniono typ kolumny `1` na obiekt `datetime`. Symbole pewnych spółek zawierają końcówkę `.XX` - usunięto końcówkę z symboli.

W trakcie testów zauważono, że niektóre symbole spółek w zbiorze opisane są w formacie `XYZ.A`, natomiast na Yahoo Finance spółki te opisane są w formacie `XYZ-A`. Zamieniono więc `.` zawarte w symbolach spółek na `-`.

W kolejnym kroku zmieniono nazwy kolumn w następujący sposób:

- Kolumna `1 => date`,
- Kolumna `2 => symbol`,
- Kolumna `4 => sector`,
- Kolumna `6 => conviction`.

Ostatecznie, zbiór danych po zakończeniu pierwszego etapu wygląda następująco:

	date	symbol	sector	conviction
0	2004-02-11	SU	Energy Minerals	0.953727
1	2004-02-11	GGG	Producer Manufacturing	0.952753
2	2004-02-11	WGR	Energy Minerals	0.947634
3	2004-02-11	CWT	Utilities	0.934181
4	2004-02-11	BLL	Process Industries	0.922862
...	...	...	...	...
37355	2022-02-09	PEP	Consumer Non-Durables	0.701507
37356	2022-02-09	SSNC	Technology Services	0.701123
37357	2022-02-09	GEF	Process Industries	0.697954
37358	2022-02-09	DPZ	Consumer Services	0.697741
37359	2022-02-09	LIFZF	Non-Energy Minerals	0.695644

37360 rows × 4 columns

Drugi etap

Drugi etap przygotowania zbioru danych obejmował pobranie danych spółek z Yahoo Finance z wykorzystaniem pakietu `yfinance` oraz feature engineering w oparciu o pakiet do analizy technicznej `ta` (https://towardsdatascience.com/technical-analysis-library-to-financial-datasets-with-pandas-python-4b2b390d3543).

Dane zostały pobrane z użyciem pętli, która zawierała następujące kroki:

- ustalono przedział czasowego (wykraczający poza dostępne miary `, pobranie listy unikalnych symboli spółek z wyczyszczonego w pierwszym etapie zbioru, oraz ustalono przedział czasowy do wyliczenia wskaźników używanych w pakiecie ta`
- z wykorzystaniem pakietu `yfinance` pobrano wszystkie możliwe dni zawierające ceny `open`, `high`, `close`, `low`, `volume`, `adj_close` dla danej spółki
- z wykorzystaniem pakietu `ta` obliczono szereg wskaźników, odnoszących się m.in. do volatility, trendu czy momentum ceny danej spółki
- następnie przygotowano tymczasowy zbiór danych zawierający dodatkowe kolumny odnoszące się do prognozowanych przedziałów czasowych, tj. ceny `adj_close` danego dnia, ceny `adj_close` za miesiąc, ceny `adj_close` za kwartał, ceny `adj_close` za pół roku oraz ceny `adj_close` za rok,
- z wykorzystaniem tymczasowego zbioru z kolumnami prognozowanych przedziałów czasowych, dołączno dla każej daty prognozowanego okresu odpowiednią dla niej cenę

`adj_close` - jeżeli cena na wyznaczoną datę nie była dostępna, funkcja `merege_asof` brała kolejną nabliższą cenę `adj_close` .

Tak przygotowany zbiór ceny i miar dla danej spółki został dodany do poprzedniego przygotowanego zbioru dla spółki z wykorzystaniem funkcji `concat` . Pętla działa aż do sprawdzenia wszystkich, uniklanych symboli spółek. Symbole nie znalezione w bazie Yahoo Finance był pomijane.

Ostatecznie, zbiór danych przygotowany z wykorzystaniem opisanej pętli zawiera 26 kolumn i ponad 5.5 mln wierszy opisujących wszystkie dostępne ceny dla każdej pobranej spółki.

W celu przygotowania dodatkowych cech, kolumnę `date` rozbito na kolumny opisujące rok (`year`), miesiąc (`month`), dzień miesiąca (`day`) oraz dzień tygodnia (`weekday`). Lata w kolumnie `year` otrzymał liczbę porządkową, tj. rok 2004, określono jako 1., natomiast 2022 jako 19.

Trzeci etap

Ostatecznie, zbiór danych przygotowany w pierwszym etapie, został połączony ze zbiorem danych przygotowanym w drugim etapie z wykorzystaniem funkcji merge.

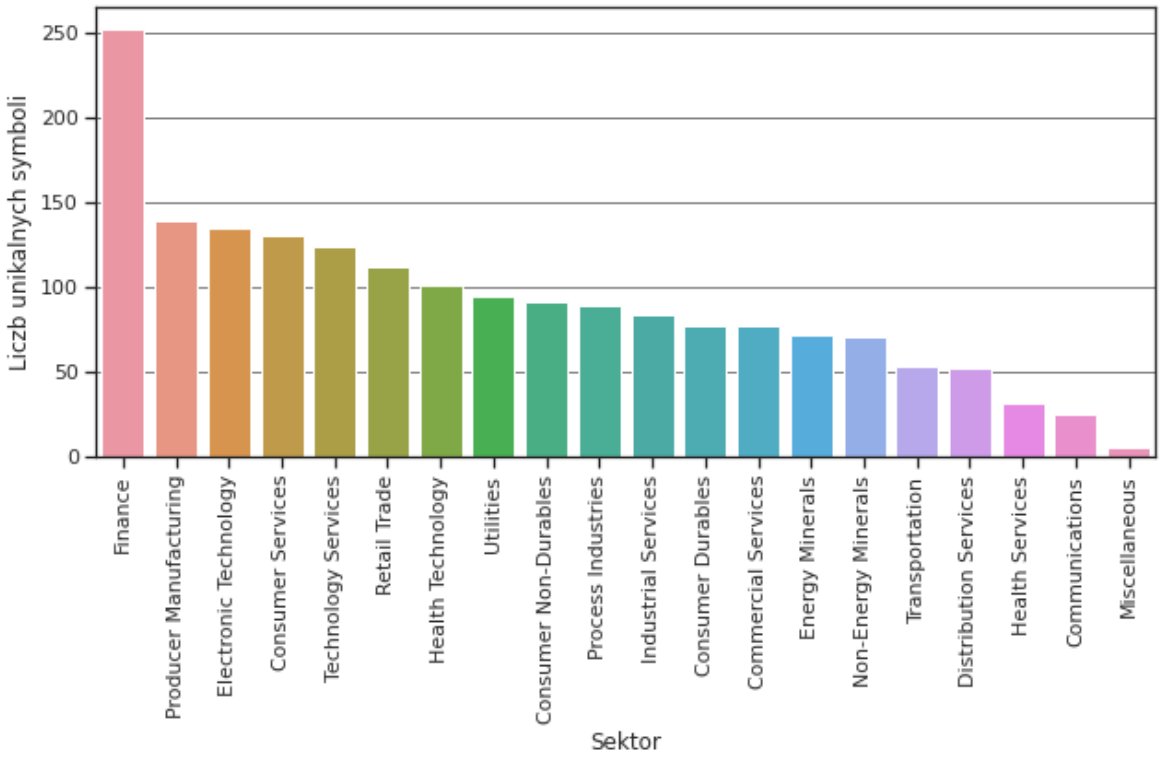
Usunięto także braki danych oraz zakodowano z wykorzystaniem `Label Encoder` nazwę sektora oraz symbol spółki.

Finalny zbiór danych zawiera 26596 wierszy oraz 27 kolumn.

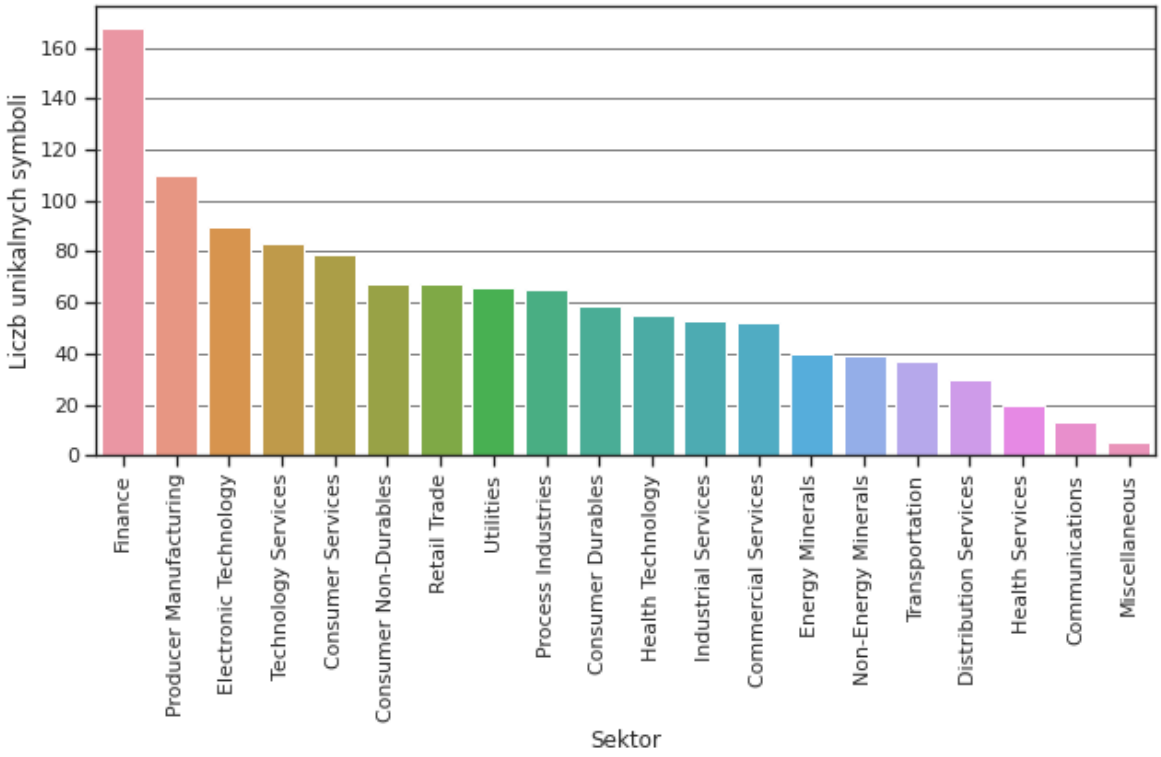
	date	year	month	day	weekday	symbol	sector	conviction	open	high	...	ema	kama	ppo	pvo	roc	adj_close_date	adj_close_mtd	adj_close_qtd	adj_close_htd	adj_close_ytd
0	2004-02-11	1	2	11	2	1009	7	0.953727	13.300000	13.365000	...	13.074966	13.294546	0.351834	-1.130415	0.113032	9.249799	9.475924	8.478457	9.726552	12.538750
1	2004-02-11	1	2	11	2	456	15	0.952753	9.282222	9.406667	...	9.184128	9.052690	0.866377	5.573530	6.557378	6.536161	6.573896	6.845714	7.591326	9.304582
2	2004-02-11	1	2	11	2	281	19	0.934181	14.650000	14.720000	...	14.445180	14.432016	0.562174	-1.852853	1.868515	9.002892	8.874455	8.466928	8.612760	10.808844
3	2004-02-11	1	2	11	2	138	14	0.922862	8.047500	8.127500	...	7.908240	7.969739	2.576472	9.683535	10.117328	6.898208	7.013620	6.748834	7.818182	9.416622
4	2004-02-11	1	2	11	2	64	7	0.912117	39.549999	39.980000	...	39.802467	39.843193	-1.645296	4.436061	-5.211801	31.821821	33.755253	32.688286	35.229458	45.768177

Podstawowa eksploracja danych

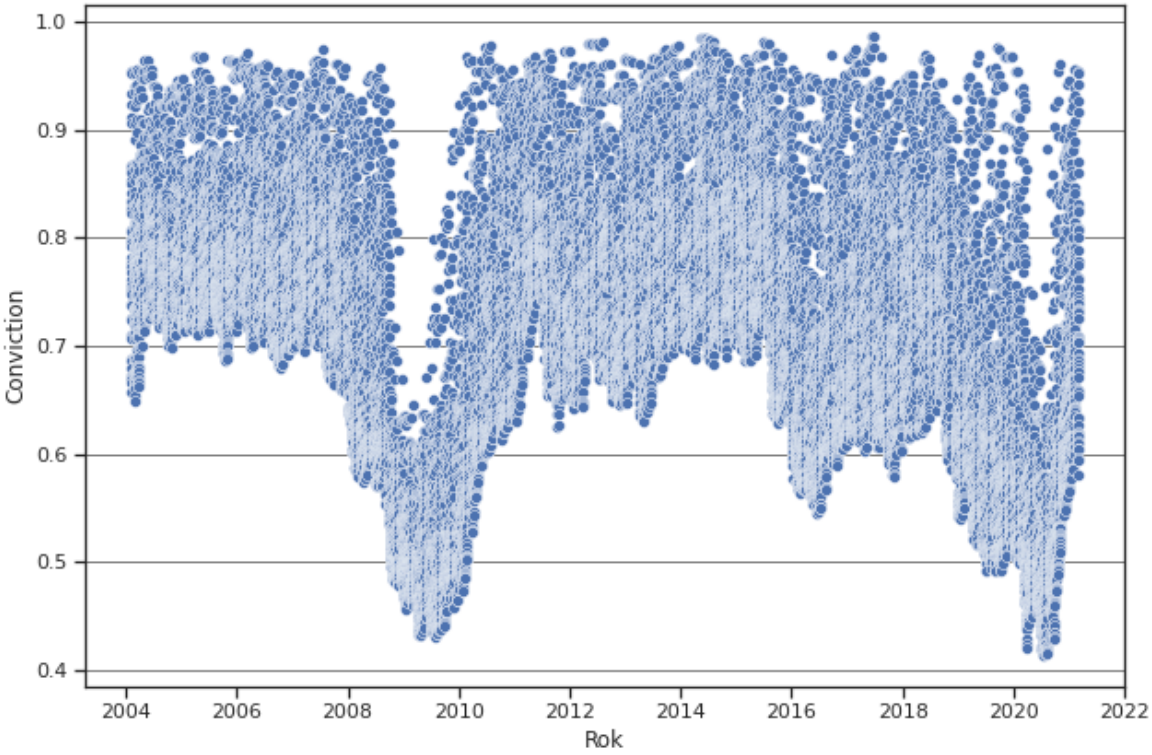
Oryginalny zbiór danych zawierał 1815 unikalnych spółek, skoncentrowanych w następujących sektorach:



Ostateczny zbiór danych zawiera 1192 spółki, zebrane w sektorach:



Sprawdzamy jak układała się miara `conviction` w badanym przedziale.



Przygotowanie i ocena modeli

LSTM - czyli long short-term memory jest jedną ze sztucznych sieci neuronowych. Jest w stanie procesować nie tylko pojedyncze punkty danych, ale także całe sekwencje.

XGBoost - algorytm wdrażający gradientowe wzmacnianie drzew, których bierze pod uwagę błędy resztkowe, a nie wagi, jak w klasycznym boostingu.

LightGMB - w odróżnieniu od XGBoost wykorzystuje technik takie jak GOSS i EFB. Zaletami tego modelu są zwiększona wydajność, skórcony czas uczenia, mniejsze zużycie pamięci, zwiększona dokładność itp.

Przetestowano wszystkie trzy modele dla wszystkich dostępnych szeregów czasowych. W przypadku LSTM nie została wykonana walidacja krzyżowa. Otrzymane wyniki należy rozpatrywać jako wartość bezwzględną. Dla prawie każdego przedziału czasowego najlepsze wyniki otrzymaliśmy z pomocą modelu LSTM. Pozostałe dwa modele (XGBoost oraz LightGBM) zwykle otrzymywały gorsze wyniki, jednak różnice były nieznaczne. Jedynie dla przedziału czasowego RoR_ytd wyniki RMSE były najlepsze w modelu LightGBM, następnie XGBoost, a ostatnie było LSTM. Przy RMSE dla walidacji krzyżowej model LightGBM za każdym razem dawał lepsze wyniki niż XGBoost. Wyniki wszystkich modeli zdecydowanie odbiegały od tych uzyskanych za pomocą regresji liniowej zarówno bez jak i z wykorzystaniem walidacji krzyżowej za każdym razem były w znacznym stopniu bardziej korzystne. W pierwotnym modelowaniu projektu RMSE wynosiło 37234890,465372, czyli znacznie więcej niż na nowym zbiorze danych.

Przedział czasowy	**RoR date**		**RoR mtd**		**RoR qtd**		**RoR htd**		**RoR yt**
Model	RMSE	RMSE dla walidacji krzyżowej	RMSE	RMSE dla walidacji krzyżowej	RMSE	RMSE dla walidacji krzyżowej	RMSE	RMSE dla walidacji krzyżowej	RMSE

Przedział czasowy	RoR date		RoR mtd		RoR qtd		RoR htd		RoR yt
Regresja liniowa	12.25232	-138.92938	12.25080	-19.97459	12.25091	-2.11846	12.25232	-20.02308	12.1471
XGBoost	0.00229	0.00825	0.00199	0.00840	0.01927	0.00840	0.00205	0.00840	0.00201
LSTM	0.00002		0.00092		0.00107		0.00005		0.00426
LightGBM	0.00186	-0.00477	0.00184	-0.00478	0.00184	-0.00478	0.00190	-0.00476	0.00184

Wnioski

Zrealizowany został cel projektu. Przygotowane zostały trzy modele, które przewidują stopę zwrotu spółki. Zostały zrealizowane wymagania funkcjonalne oraz niefunkcjonalne dotyczące działania programu. Z trzech modeli najlepiej działał LSTM bez walidacji krzyżowej, natomiast najlepszym modelem testowanym walidacją krzyżową okazał się LightGBM.