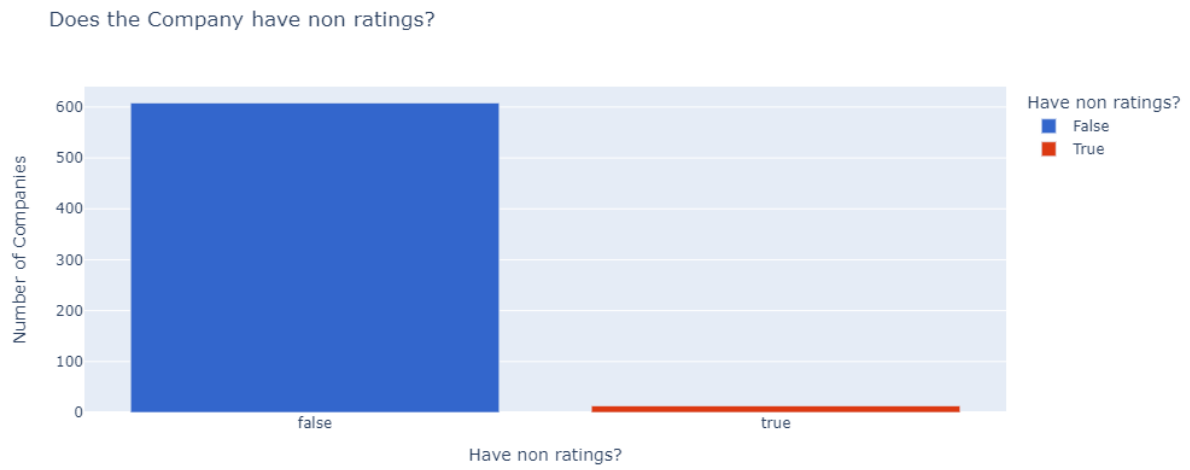
Pobieranie danych

Dane zostały pobrane w sposób automatyczny z wykorzystaniem biblioteki yfinance. Dla każdego z okresów, w których dla danych spółek zostały opracowane oceny eksperckie obliczono stopy zwrotu korzystając z wartości akcji na ten dzień i na dzień za 1,3,6, oraz 12 miesięcy. W przypadku, gdy przyszłe dane nie były dostępne (np. ze względu na dzień, w którym giełda jest zamknięta) to zwroty liczono wykorzystując dane najbliższe dacie, na którym wartość akcji była potrzebna.

Analiza EDA

Czy istnieją spółki, które nie mają żadnych ocen?

Naszą analizę rozpoczęliśmy od sprawdzenie czy istnieją spółki, które ani razu nie zostały ocenione przez ekspertów.

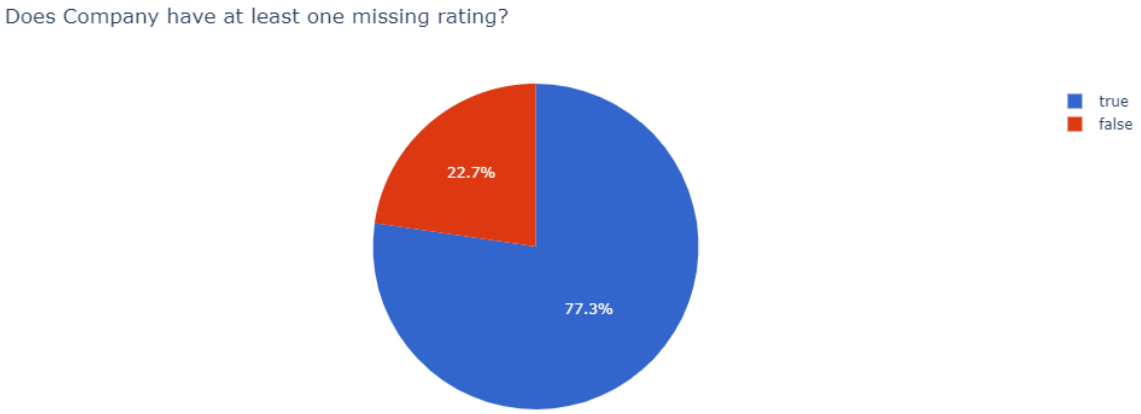


Okazuje się, że istnieje 13 spółek dla których nie ma żadnych ocen. Zostały one przez nas usunięte.

Czy istnieje duplikaty w danych?

Zidentyfikowaliśmy 115 duplikatów w wierszach i w każdym przypadku postanowiliśmy zostawić pierwszy z nich. Liczba ta może wynikać nie tylko z własności danych pierwotnych, ale także z zaimplementowanych procedur pobierania danych giełdowych.

Jaki procent spółek posiada braki w ocenach?

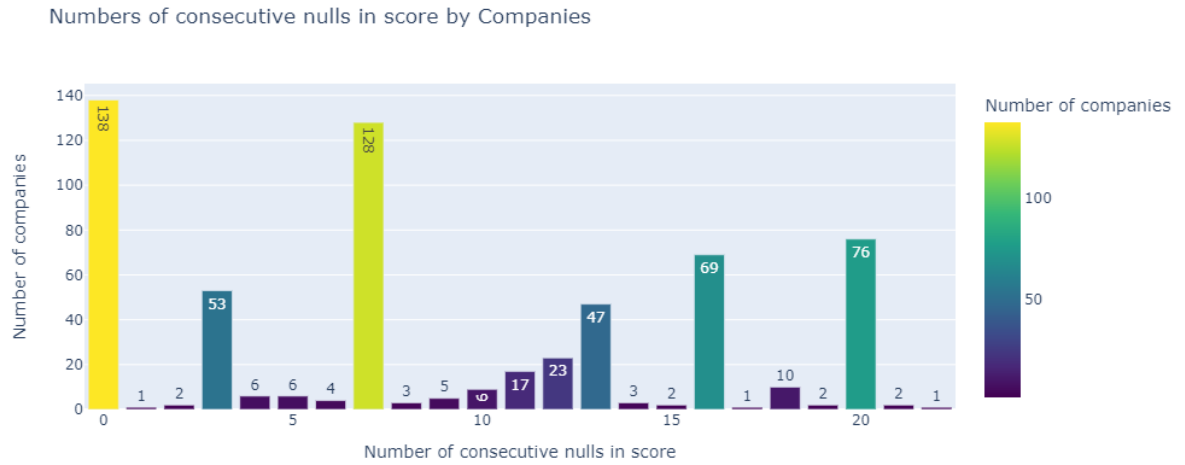


Okazuje się, że zdecydowana większość ze spółek posiada braki w danych. W związku z tym decyzja o usunięciu lub zostawieniu a także możliwa strategia wypełnienia braków danych stanowi istotny element projektu.

Natomiast po usunięciu spółek, które ani razu nie zostały ocenione przez ekspertów okazało się, że wszystkie pozostałych dostępne są przynajmniej częściowe informacje o zwrotach.

Rozkład następujących po sobie wartości null

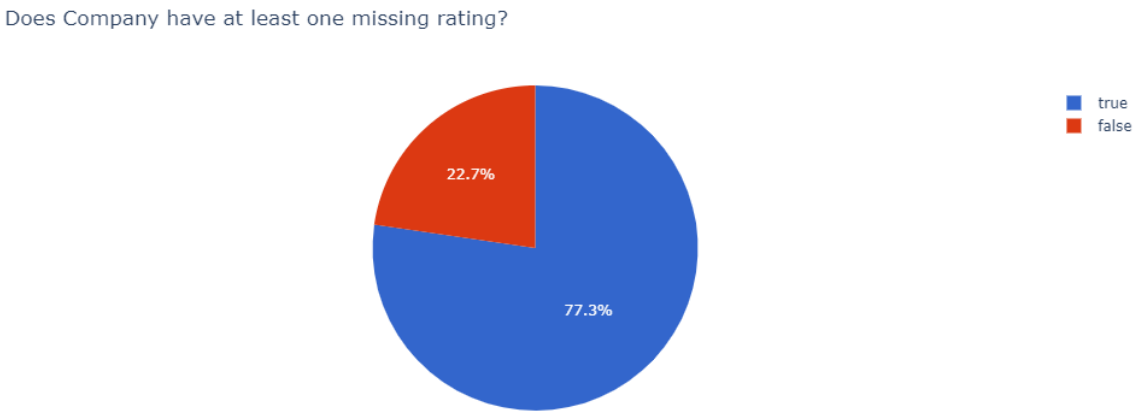
Uznaliśmy, że najbardziej kłopotliwe w dalszej analizie są następujące po sobie braki ocen. W związku z tym dla każdej spółki obliczyliśmy maksymalną liczbę następujących po sobie wartości typu null.



Postanowiliśmy usunąć wszystkie spółki, które nie posiadały ocen 17 lub więcej razy z rzędu.

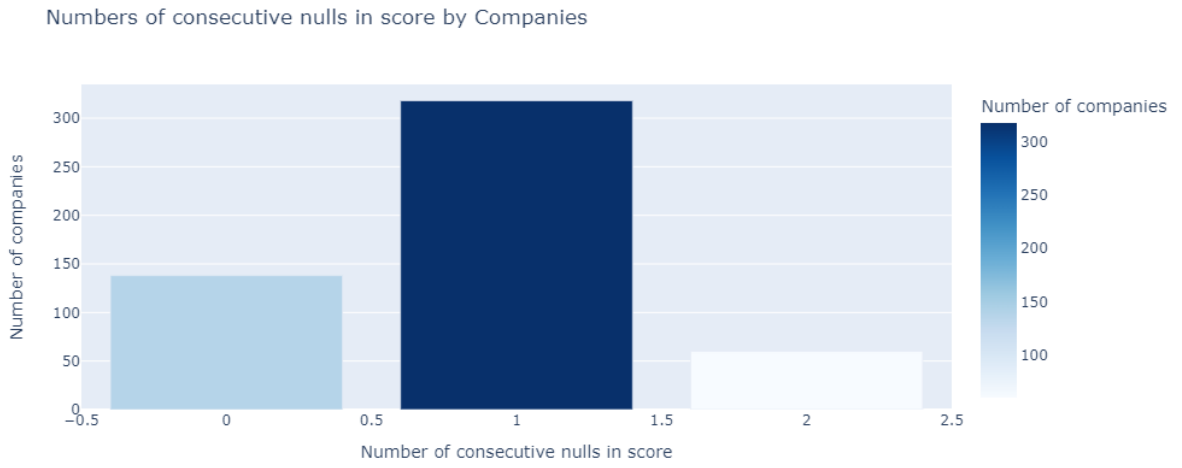
Jaki procent braków ocen dotyczy wyłącznie jednego przedziału czasowego?

Następnie zweryfikowaliśmy jaki procent wszystkich braków ocen dla każdej spółki dotyczy tylko jednego okresu. Może to świadczyć nie tylko o brakach danych, ale także o tym, że w danym okresie czasu spółka nie była publicznie notowana.



Zdecydowana większość braków ocen dotyczyła jednego konkretnego okresu. Uznaliśmy w takim wypadku za niezasadne uzupełnia tych braków.

Ostatecznie postanowiliśmy usunąć wszystkie sekwencje trzech lub więcej następujących po sobie braków ocen dla każdej ze spółki. Decyzja ta skutkowała zmniejszenie zbioru dostępnych obserwacji o 23 %.



Zdecydowana większość pozostałych braków dotyczyły pojedynczych ocen. W związku z tym można było dokonać wypełniania brakujących wartości z wykorzystaniem metody forward-fill.

Braki danych w zwrotach

Liczba braków w danych w zwrotach pozyskanych z danych giełdowych okazała się być relatywnie niewielka i związku z tym, wiersze, które zawierały wartości typu Null zostały wyłączone z dalszej analizy.

	Column	Nan_number
7	1MReturn	12
8	3MReturn	8
9	6MReturn	11
10	12MReturn	341

Sposób wyboru modelu i wykorzystywane miary

Mean residua deviance

$$D(y,\hat{\mu}) = 2\Big(\log\left(p(y \mid \hat{\theta}_s)\right) - \log\left(p(y \mid \hat{\theta}_0)\right)\Big).$$

Dewiancja reszt opisuje jak dobrze zmienna objaśniana może zostać zaprognozwana przez model z określoną liczbą zmiennych objaśniających. Im niższa wartość tym lepiej model przewiduję wartość zmiennej objaśnianej. Zmienna ta porównuje logarytm funkcji prawdopodobieństwa naszego modelu do modelu teoretycznego (np. Modelu Poissona).

MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Jedna z najczęściej wykorzystywanych funkcji kosztu. MSE oznacza “mean square error”, a więc średni błąd kwardatowy. Powodem dla którego metoda ta jest tak często wykorzystywana jest fakt iż daje nam nieobciążony i efektywny estymator. W żadnym razie nie oznacza to jednak, że jest to "najlepsza" funkcja celu. Jedną z jej podstawowych i najważniejszych wad jest fakt iż "przywiązuje zbyt dużą wagę" wartościom odstającym (ang. outliers).

RMSE

Jest to pierwiastek z MSE (r pochodzi od angielskiego “root”). Reprezentuje pierwiastek z drugiego momentu próbkowania różnic między wartościami przewidywanymi a wartościami obserwowanymi. Z racji na pierwiastek jest zawsze dodatni. Wartość 0 oznacza idealne dopasowanie do danych (w praktyce niespotykana). Im mniejsze wartości tym lepiej dla modelu. Miara jest zależna od względnej skali użytych liczb, więc musimy porównywać te same zestawy danych.

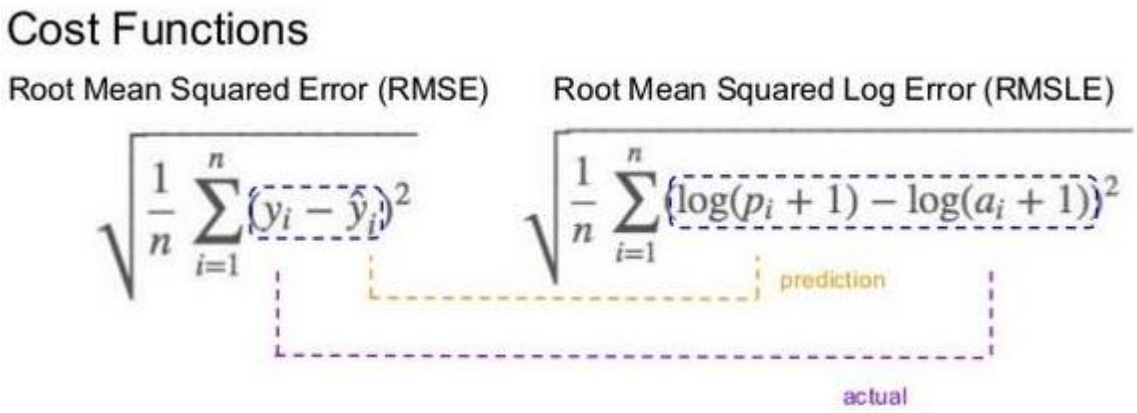
MAE

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}.$$

Jest to średni błąd bezwzględny (“mean absolute error”). Wartość 0 oznacza idealne dopasowanie do danych (w praktyce niespotykana).W porównaniu do wartości błędu średniokwadratowego, ta miara dopasowania jest mniej czuła na wartości odstające, to znaczy wyjątkowo duże wartości błędu będą

wpływać na wartość MAE w mniejszym stopniu niż na wartość MSE. Jest popularny w praktyce prognozowania biznesowego z racji na intuicyjną, prostą interpretację.

RMSE



Miara podobna do RMSE, przy czym zakłada transformację logarytmiczną wartości prognozowanych i wartości zaobserwowanych w danych. Dodanie jedynki (do predykcji i danych właściwych) jest uwzględnione by uniknąć logarytmów naturalnych z zera. Jeżeli wartości są ujemne to korzystanie z tej funkcji nie ma sensu (z racji na naturę logarytmów). Funkcja dobrze sprawdza się jeżeli cel rośnie wykładniczo, zwracamy uwagę na wzrosty procentowe, a nie absolutne, posiadamy duży rozstęp w zmiennej celu, nie chcemy karać za duże różnice, gdy dane i predykcje są dużymi liczbami, chcemy bardziej karać wartości niedoszacowane (w stosunku do przeszacowanych).

R^2

Współczynnik determinacji, informuję o tym jaka część zmienności (wariancji) zmiennej objaśnianej w próbie pokrywa się z korelacjami ze zmiennymi zawartymi w modelu. Używany w modelach regresji. Ustandaryzowany, przyjmuję wartości z przedziału [0, 1], im większy tym model lepiej dopasowany do danych. Jego wartości najczęściej są wyrażane w procentach.

Benchmark model

Jako model do przeprowadzenie benchmarku wybrano najprostszy wariat z metod wchodzących w skład uogólnionych modeli liniowych (ang. Generalized regressions models), mianowicie regresję liniową (gausowską). Rezultaty dla wszystkich horyzontów okazały się niezwykle słabe. W niektórych przypadkach R^2 przyjmuje wartości ujemne, co oznacza, że mniejsze wartości błędów można by uzyskać zastępując model najzwyklejszą poziomą linią.

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	Residual deviance	AIC	model
0	0.008129	0.090163	0.065222	0.097584	0.001513	0.008129	47.670116	-11571.913671	Benchmark linear regression 1m train
1	0.019651	0.140183	0.106700	0.143964	0.003801	0.019651	115.234407	-6395.973929	Benchmark linear regression 3m train
2	0.048958	0.221264	0.165151	0.211789	0.008019	0.048958	287.087181	-1043.211748	Benchmark linear regression 6m train
3	0.120654	0.347353	0.244971	0.288756	0.009602	0.120654	707.516602	4245.970580	Benchmark linear regression 12m train

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	Residual deviance	AIC	model
0	0.008385	0.091569	0.067875	0.096225	-0.082290	0.008385	19.645671	-4547.511283	Benchmark linear regression 1m test
1	0.023115	0.152035	0.110979	0.149934	-0.003807	0.023115	54.157831	-2171.602445	Benchmark linear regression 3m test
2	0.057266	0.239304	0.171791	0.204001	-0.113019	0.057266	134.175165	-45.930784	Benchmark linear regression 6m test
3	0.135841	0.368566	0.260128	0.297425	0.009684	0.135841	318.275929	1977.888330	Benchmark linear regression 12m test

Wykorzystane metody

Generalized Linear Models (GLM).

Uogólnione modele liniowe (GLM) szacują modele regresji dla zmiennych podążających za rozkładem wykładniczym. Oprócz rozkładu normalnego, obejmują one rozkłady Poissona, dwumianowy i gamma. Każdy z nich służy innym celom i w zależności od wyboru funkcji dystrybucji i łącza może być używany zarówno do przewidywania lub klasyfikacji.

GLM obsługuje zarówno klasyfikację binarną, jak i wielomianową. W przypadku klasyfikacji binarnej kolumna odpowiedzi może mieć tylko dwa poziomy; w przypadku klasyfikacji wielomianowej kolumna odpowiedzi będzie miała więcej niż dwa poziomy.

Gdy GLM wykonuje regresję (z kolumnami czynników), jedną kategorię można pominąć, aby uniknąć wielowspółliniowości. Jeśli regularyzacja jest wyłączona (lambda = 0), to jedna kategoria jest pomijana. Jednak w przypadku korzystania z domyślnego parametru lambda uwzględniane są wszystkie kategorie.

Najprostszym przykładem GLM jest regresja liniowa. Ma wiele zastosowań i kilka zalet w porównaniu z innymi rodzinami. W szczególności jest szybszy i wymaga bardziej stabilnych obliczeń. Funkcja łączenia g jest tożsamością, a gęstość f odpowiada rozkładowi normalnemu. Rodzina Gaussowska modeluje zależność między odpowiedzią y a wektorem towarzyszącym x jako funkcję liniową:

$$\tilde{y} = x^T \beta + \beta_0$$

Model jest dopasowywany przez rozwiązanie problemu najmniejszych kwadratów, co jest równoważne maksymalizacji prawdopodobieństwa dla rodziny Gaussa.

$$\max_{\beta_1 \beta_0} -\frac{1}{2N} \sum_{i-1}^{N} \left(x_i^T \beta + \beta_0 - y_i\right)^2 - \lambda \left(\alpha ||\beta|_1 + \frac{1}{2}(1-\alpha)||\beta|_2^2\right)$$

Suma kwadratów błędów przewidywania:

$$D = \sum_{N}^{i=1} (y_i - \tilde{y}_i)^2$$

Generalized Additive Models (GAM)

Ogólny model addytywny jest ogólnym modelem liniowym, w którym predykator liniowy zależy liniowo od zmiennych predykcyjnych i gładkich funkcji zmiennych predykcyjnych.

Prosty model liniowy. Zakładając n obserwacji, xi ze zmienną odpowiedzi yi, gdzie yi jest obserwacją zmiennej Yi, niech ui≡E(Yi). Przyjmując liniową zależność pomiędzy zmiennymi predykcyjnymi a odpowiedzią istnieje następująca zależność między xi i Yi:

$$Y_i = u_i + \epsilon_i \text{ gdzie } u_i = \beta_i x_i + \beta_0$$

gdzie βi i β0 są nieznanymi parametrami, ϵi jest i.i.d zerową zmienną z wariancjami δ2.

$$Y_i = f(x_i) + \epsilon_i \text{ gdzie } f(x_i) = \sum_{j=1}^{k} b_j(x_i)\beta_j + \beta_0$$

Distributed Random Forest (DRF)

Rozproszony losowy las generuje las klasyfikacji lub regresji. Każde drzewo jest słabym uczniem zbudowanym na podzbiorze wierszy i kolumn. Większa ilość drzew zmniejsza wariancję. Zarówno klasyfikacja, jak i regresja biorą pod uwagę średnią prognozę dla wszystkich swoich drzew, aby uzyskać końcową prognozę, niezależnie od tego, czy przewiduje się klasę, czy wartość liczbową.

Przypisanie liścia węzła. Drzewa grupują obserwacje w węzły liści, te informacje mogą być przydatne do inżynierii funkcji lub interpretacji modelu.

Stacked Ensambles

Stacked Ensambles jest metodą nadzorowania algorytmów uczenia maszynowego w zespole, który znajduje optymalną kombinację zbioru algorytmów przewidywania przy użyciu procesu zwanego układaniem w stos.

Stacking, znany również jako Super Learning lub Stacked Regression jest klasą algorytmów, która polega na szkoleniu "metaucznia" ("metalearner") drugiego poziomu w celu odnalezienia optymalnej kombinacji podstawowych uczniów. Celem układania w stosy jest zgromadzenie razem silnych i zróżnicowanych grup uczniów.

Extremely Randomized Trees (w ramach DRF)

W lasach losowych podzbiór cech branych pod uwagę jest używany do określenia najbardziej dyskryminujących progów, które będą wybrane jako reguła podziału. W ekstremalnie losowych drzewach (XRT) używany jest losowy podzbiór cech kandydujących, lecz nie są szukane najbardziej dyskryminujące progi, lecz są one losowane dla każdej cechy kandyduącej, a najlepszy z nich jest wybierany jako reguła podziału. Pozwala to zredukować wariancję modelu, kosztem większego wzrostu biasu.

Automatyczny wybór modelu

Zdecydowaliśmy się skorzystać z biblioteki H20, która posiada wbudowaną automatyczną procedurę wyboru najlepszego modelu. W procesie przeszukiwania kluczowe są wartości mean residual variance. Wykonania pełnego przeszukiwania nie było jednak możliwe ze względu na ograniczenia sprzętowe.

Wyniki

Prognozy dla 1 miesięcznych zwrotów

Wyniki na danych treningowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.007382	0.085917	0.062309	0.093536	0.093339	0.007382	StackedEnsemble_BestOfFamily_1_AutoML_1_20220610_171527
1	0.007382	0.085917	0.062309	0.093536	0.093339	0.007382	StackedEnsemble_AllModels_1_AutoML_1_20220610_171527
2	0.008129	0.090163	0.065222	0.097584	0.001509	0.008129	GLM_1_AutoML_1_20220610_171527
3	0.007905	0.088909	0.064471	0.096414	0.029085	0.007905	GBM_1_AutoML_1_20220610_171527
4	0.007636	0.087384	0.063426	0.094914	0.062105	0.007636	GBM_2_AutoML_1_20220610_171527

Wyniki dla danych testowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.008455	0.091949	0.068209	0.096593	-0.091299	0.008455	StackedEnsemble_BestOfFamily_1_AutoML_1_20220610_171527
1	0.008455	0.091949	0.068209	0.096593	-0.091299	0.008455	StackedEnsemble_AllModels_1_AutoML_1_20220610_171527
2	0.008385	0.091571	0.067875	0.096228	-0.082347	0.008385	GLM_1_AutoML_1_20220610_171527
3	0.008530	0.092357	0.068389	0.097026	-0.101002	0.008530	GBM_1_AutoML_1_20220610_171527
4	0.008637	0.092936	0.068928	0.097561	-0.114856	0.008637	GBM_2_AutoML_1_20220610_171527

Prognozy dla 3 miesięcznych zwrotów

Wyniki na danych treningowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.016326	0.127772	0.097540	0.131841	0.172389	0.016326	StackedEnsemble_BestOfFamily_1_AutoML_2_20220610_174632
1	0.016318	0.127741	0.097535	0.131796	0.172781	0.016318	StackedEnsemble_AllModels_1_AutoML_2_20220610_174632
2	0.018101	0.134542	0.102573	0.138379	0.082364	0.018101	GBM_2_AutoML_2_20220610_174632
3	0.017791	0.133383	0.101758	0.137240	0.098103	0.017791	GBM_3_AutoML_2_20220610_174632
4	0.018833	0.137233	0.104765	0.140970	0.045285	0.018833	GBM_1_AutoML_2_20220610_174632

Wyniki dla danych testowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.023527	0.153387	0.112507	0.151372	-0.021730	0.023527	StackedEnsemble_BestOfFamily_1_AutoML_2_20220610_174632
1	0.023515	0.153346	0.112488	0.151338	-0.021185	0.023515	StackedEnsemble_AllModels_1_AutoML_2_20220610_174632
2	0.023625	0.153705	0.112886	0.151707	-0.025979	0.023625	GBM_2_AutoML_2_20220610_174632
3	0.023690	0.153917	0.113114	0.151931	-0.028807	0.023690	GBM_3_AutoML_2_20220610_174632
4	0.023356	0.152828	0.112263	0.150883	-0.014299	0.023356	GBM_1_AutoML_2_20220610_174632

Prognozy dla 6 miesięcznych zwrotów

Wyniki na danych treningowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.041525	0.203776	0.152482	0.195987	0.158622	0.041525	StackedEnsemble_BestOfFamily_1_AutoML_3_20220610_175421
1	0.041705	0.204219	0.152884	0.196421	0.154966	0.041705	StackedEnsemble_AllModels_1_AutoML_3_20220610_175421
2	0.043753	0.209173	0.156543	0.200613	0.113470	0.043753	GBM_2_AutoML_3_20220610_175421
3	0.043205	0.207858	0.155689	0.199525	0.124577	0.043205	GBM_3_AutoML_3_20220610_175421
4	0.046694	0.216088	0.162202	0.206971	0.053888	0.046694	GBM_1_AutoML_3_20220610_175421

Wyniki dla danych testowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.057004	0.238755	0.172200	0.204074	-0.107922	0.057004	StackedEnsemble_BestOfFamily_1_AutoML_3_20220610_175421
1	0.056997	0.238740	0.172528	0.204093	-0.107782	0.056997	StackedEnsemble_AllModels_1_AutoML_3_20220610_175421
2	0.057371	0.239523	0.173037	0.204896	-0.115055	0.057371	GBM_2_AutoML_3_20220610_175421
3	0.057593	0.239986	0.174312	0.205531	-0.119377	0.057593	GBM_3_AutoML_3_20220610_175421
4	0.057047	0.238845	0.172858	0.204151	-0.108750	0.057047	GBM_1_AutoML_3_20220610_175421

Prognozy dla 12 miesięcznych zwrotów

Wyniki dla danych treningowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.041525	0.203776	0.152482	0.195987	0.158622	0.041525	StackedEnsemble_BestOfFamily_1_AutoML_3_20220610_175421
1	0.041705	0.204219	0.152884	0.196421	0.154966	0.041705	StackedEnsemble_AllModels_1_AutoML_3_20220610_175421
2	0.043753	0.209173	0.156543	0.200613	0.113470	0.043753	GBM_2_AutoML_3_20220610_175421
3	0.043205	0.207858	0.155689	0.199525	0.124577	0.043205	GBM_3_AutoML_3_20220610_175421
4	0.046694	0.216088	0.162202	0.206971	0.053888	0.046694	GBM_1_AutoML_3_20220610_175421

Wyniki dla danych testowych

	MSE	RMSE	MAE	RMSLE	R^2	Mean Residual Deviance	model
0	0.057004	0.238755	0.172200	0.204074	-0.107922	0.057004	StackedEnsemble_BestOfFamily_1_AutoML_3_20220610_175421
1	0.056997	0.238740	0.172528	0.204093	-0.107782	0.056997	StackedEnsemble_AllModels_1_AutoML_3_20220610_175421
2	0.057371	0.239523	0.173037	0.204896	-0.115055	0.057371	GBM_2_AutoML_3_20220610_175421
3	0.057593	0.239986	0.174312	0.205531	-0.119377	0.057593	GBM_3_AutoML_3_20220610_175421
4	0.057047	0.238845	0.172858	0.204151	-0.108750	0.057047	GBM_1_AutoML_3_20220610_175421

Komentarz do wyników

Rezultaty modelów są generalnie rzecz biorąc na bardzo niskim poziomie. Wynika to przede wszystkim z niskiej korelacji pomiędzy ocenami ekspertów a zwrotami i może świadczyć o tym, że o wiele lepszego modelu nie da się opracować. Dodatkowo należy zauważyć, że prognozy dla okazywały się być tym lepsze im dłuższy był horyzont prognozy, przy czym wszystkie rezultaty na zbiorach testowych

wskazują, że równie dobrze można by zgadywać. Modele nie prezentują żadnej użytecznej mocy predykcyjnej.