Analiza trendu dla cen Bitcoina na podstawie danych w latach 2023-2024

Apolonia Abramowicz, 272722

I. Wstęp

W ciągu ostatnich lat zaobserwowano olbrzymi wzrost zainteresoawania tzw. kryptowalutą, co jest zarówno skutkiem, jak i przyczyną szybkiego zwiększenia wartości najpopularniejszych jej rodzajów. Kryptowaluta jest szczególnym przypadkiem waluty wirtualnej, rozproszonym systemem księgowym, który bazuje na kryptografii i przechowuje informacje o stanie posiadania użytkownika. System korzysta z tzw. portfeli (będącymi węzłami systemu) posiadających specjalny klucz, a dostęp do niego powinien mieć tylko jego właściciel. W związku z wciąż niemalejącym ruchem na rynku kryptowalut oraz z jego tajemniczą dla wielu ludzi naturą, powstaje pytanie, czy inwestycja na nim jest bezpieczna? Odpowiedzenie na tak postawione pytanie byłoby oczywiście niemożliwe, można się jednak nad tym tematem zastanowić. Podstawowe ryzyko związane z kryptowalutą to brak jej fizycznego odzwierciedlenia, nie istenieje gwarancja jej materialnej wartości. Cena kryptowaluty podatna jest na duże wahania, w porównaniu np. z cenami akcji, wiąże się to przede wszystkim ze znacznie mniejszą kontrolą.

W tym projekcie skupiono się na najpopularniejszej kryptowalucie, Bitcoinie, w celu przeprowadzenia analizy związku jego ceny z woluemenem handlu oraz liczbą aktywnych portfeli danego dnia. Podjęta zostanie próba sprawdzenia, czy możliwe jest przewidzenie spadku lub wzrostu ceny na podstawie tych danych. Zależność przeanalizowana będzie dla lat 2023 oraz 2024, dane dotyczące cen oraz wolumenu przedstawione będą w dolarach amerykańskich.

II. Analiza danych

A. Zbiór danych

W celu doboru odpowiedniego zbioru danych wykorzystano API platformy Santiment, z którego pozyskano następujące dane:

1) Cena Bitcoina:

- **Opis:** Cena, czyli wartość jednego Bitcoina danego dnia w dolarach amerykańskich.
- **Analiza:** Jej wysokość jest wynikiem popytu i podaży na rynku kryptowalut. Ta dana jest kluczowa w celu zrozumienia trendów wzrostowych oraz spadkowych podczas analizy danych historycznych

2) Wolumen handlu Bitcoina:

- **Opis:** Wartość całkowita transakcji wykonanych w danym dniu, w dolarach amerykańskich.
- **Analiza:** Wolumen jest wskaźnikiem aktywności na rynku. Jego wysoka wartość sugeruje duże zainteresowanie inwestorów i może sugerować zmianę w sentymencie rynkowym.

3) Liczba aktywnych adresów:

- **Opis:** Liczba unikalnych adresów Bitcoina aktywnych w danym dniu, czyli takich które dokonały transakcji.
- **Analiza:** Wartość może służyć do okreslenia, czy cena Bitcoina ma związek z większą lub mniejszą aktywnością jego użytkowników.

B. Przetwarzanie wstępne

W celu poprawy jakości dalszej analizy, zastosowano normalizację rozumianą jako przekształcenie danych do wspólnej skali, bez deformowania różnic pomiędzy wartościami. Oznacza to, że dane zostały sprowadzone do postaci, w której ich średnia jest możliwie bliska 0, a odchylenie standardowe 1. Przedtem przeprowadzono również operację usunięcia wartości odstających zgodnie z regułą IQR (odstępu ćwiartkowego). Istotnym krokiem było również przesunięcie wolumenu oraz liczby aktywnych adresów o jeden w przód, gdyż chcemy zbadać trend dla dnia następnego na podstawie tych danych. Następnie usunięto powstałe niekompletne rekordy.

C. Analiza eksploracyjna

1) Analiza dla całego zbioru:

- Przed przetwarzaniem wstępnym

Tabela I: Statystyki dla całego zbioru

	Cena	Wolumen	Liczba adresów
średnia	36502.92286747071	22306079053.886	923212.954
std	14641.556959855554	12965013866.350721	107796.59305153813
min	16625.0805507091	5330878402.0	506836.0
25%	26612.200207398426	13307808178.75	863539.5
50%	29655.6854692676	19175915241.5	929520.0
75%	43036.9226306292	27107623717.0	992042.0
max	73079.3733787985	102802940877.0	1238270.0

Tabela II: Pierwsze 5 wierszy dla całego zbioru

Data	Cena USD	Wolumen USD	Liczba aktywnych adresów
2023-01-01	16625.0805507091	9244361700	717140
2023-01-02	16688.4713573932	12097775227	849628
2023-01-03	16679.8570798137	13903079207	982392
2023-01-04	16863.2382582833	18421743322	922045
2023-01-05	16836.7366452236	13692758566	934845

- Po przetwarzaniu wstępnym (przed przesunięciem wolumenu i liczby aktywnych adresów)

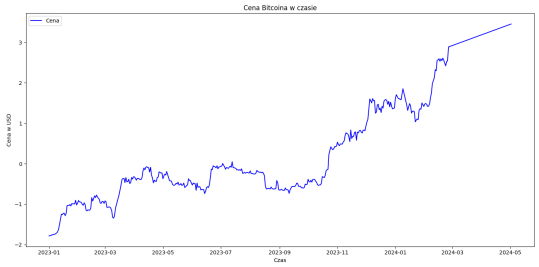
Tabela III: Statystki dla całego zbioru

	Cena	Wolumen	Liczba adresów
liczba	442.0	442.0	442.0
średnia	6.430251002353848e-17	0.0	-1.607562750588462e-17
std	1.001133144839459	1.0011331448394591	1.001133144839459
min	-1.4891158698619797	-1.611715345515214	-2.6425853263886894
25%	-0.6114292227472639	-0.769708607138204	-0.6584177281399394
50%	-0.35726911054533206	-0.20858126059199272	0.02775769923453661
75%	0.5758868497566988	0.630435252765362	0.6578998792494359
max	3.079623273337118	2.843544019819814	2.359165636945716

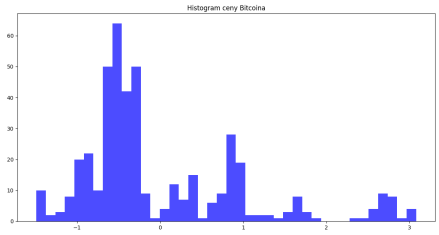
Tabela IV: Pierwsze 5 wierszy dla całego zbioru

Data	Cena USD	Wolumen USD	Liczba aktywnych adresów
2023-01-01	-1.4891158698619797	-1.1714181859588457	-2.197137848350136
2023-01-02	-1.4833932520985538	-0.8503870800844753	-0.8242792489649162
2023-01-03	-1.48417090773151	-0.6472764101255596	0.551439299800099
2023-01-04	-1.4676161351346289	-0.13889172484005954	-0.07388448723797127
2023-01-05	-1.4700085735588733	-0.6709391100215065	0.05875084631775313

2) Analiza dla ceny:

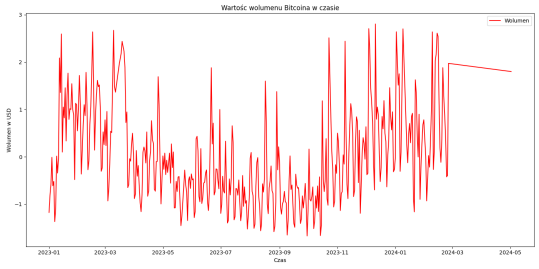


Rysunek 1: Wykres liniowy ceny Bitcoina

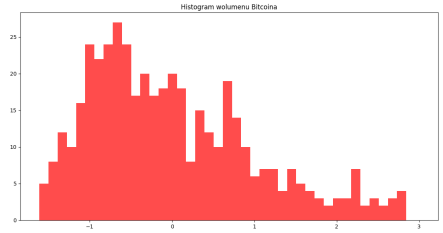


Rysunek 2: Histogram ceny Bitcoina

3) Analiza dla wolumenu:

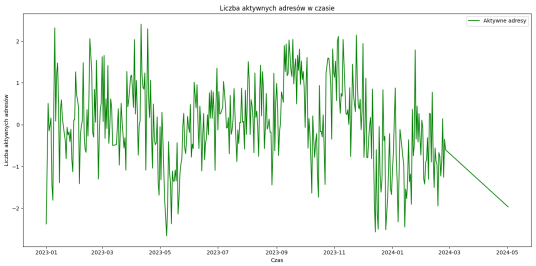


Rysunek 3: Wykres liniowy wolumenu Bitcoina

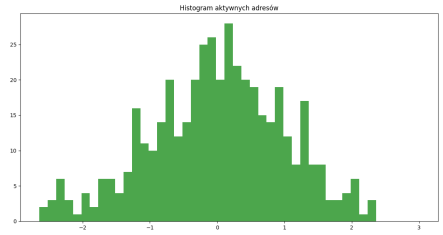


Rysunek 4: Histogram wolumenu Bitcoina

4) Analiza dla liczby aktywnych adresów:



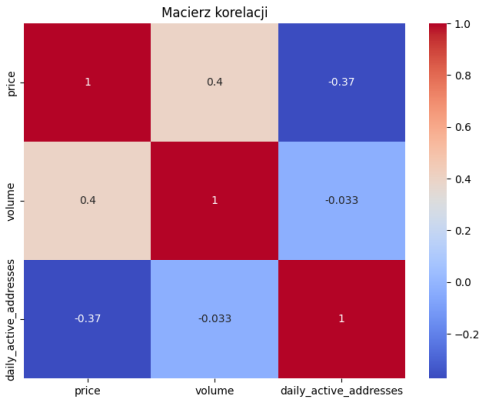
Rysunek 5: Wykres liniowy dla liczby aktywnych adresów



Rysunek 6: Histogram aktywnych adresów

D. Analiza cech

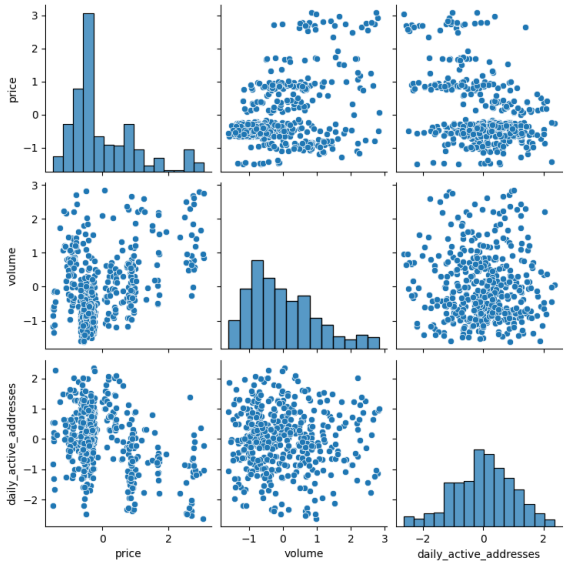
1) Analiza korelacji:



Rysunek 7: Macierz korelacji dla zbioru

Macierz korelacji pokazuje związek dodatni ceny z wolumenem, co wskazywałoby na to, że wraz z wzrostem woluemnu,

obserwuje się również zwiększenie wartości ceny. Współczynnik w przypaku związku ceny oraz liczby aktywnych adresów jest wartością ujemną, oznacza to że zazwyczaj w przypadku dużej aktywności użytkowników cena Bitcoina spada.



Rysunek 8: Wykresy par dla zbioru

Na wykresie rozrzutu dla ceny oraz wolumenu widoczna jest pewna zależność wskazująca, że wraz z wzrostem jednej danej obserwowany jest również wzrost drugiej. Potwierdza to wynik uzyskany za pomocą macierzy korelacji. Choć korelacja jest obecna, nie jest ona silną zależnością liniową. Podobnie jest w przypadku liczby aktywnych adresów oraz ceny (tutaj wzrost jednej cechy wiąże się zazwyczaj ze spadkiem drugiej)

2) Uzupełnienie danych o nowe cechy: Problem przewidywania trendu dla ceny Bitcoina jest problemem klasyfikacji, w tym celu należało wyłonić odpowiednią cechę. Co więcej, lby możliwie poprawić wydajność modeli, zbiór uzupełniono o nowe dane.

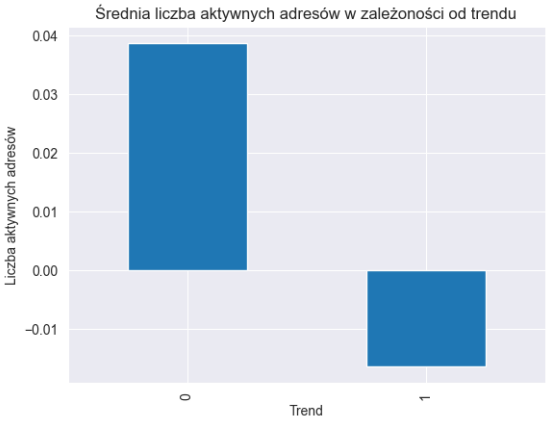
- **Trend:** Cecha trend została wyliczona za pomocą ceny. Najpierw ustalono różnicę między ceną z dnia obecnego oraz dnia poprzedniego, następnie ustalono wartość trendu wynoszącą:
  - **0** dla wyniku ujemngeo (spadek wartości)
  - **1** dla wyniku dodatniego (wzrost wartości)

  Ponieważ utrzymanie ceny jest zjawiskiem występującym bardzo rzadko, zdecydowano się na pominięcie tego przypadku. Trend jest cechą, której wartość będzie klasyfikowana w dalszej części badania.
- **Trend z poprzedniego dnia:** Cecha liczona w sposób podobny do cechy trend. W tym przypadku użyto różnicy cen z dwóch poprzednich dni.
- **Średnie kroczące:** Średnie wartości dla ceny z dni poprzednich, wyliczone dla trzech różnych przedziałów, trzech, siedmiu oraz czternastu dni wstecz.

3) Analiza korelacji po dodaniu nowych cech:

Rysunek 9: Wykres słupkowy dla wolumenu

Średnia wartość wolumenu jest dodatnia dla trendu wzrostowego oraz ujemna dla spadkowego. Oznacza to, że w dni wzrostowe wartość wolumenu będzie wyższa i niższa w dni spadkowe. Byłoby to zgodne z wcześniejszą analizą dla korelacji z ceną.



Rysunek 10: Wykres słupkowy dla liczby aktywnych adresów

Średnia liczby aktywnych adresów jest dodatnia dla trendu spadkoweg oraz ujemna dla wzrostowegoo. Oznacza to, że w dni wzrostowe wartość będzie niższa i wyższa w dni spadkowe. Byłoby to zgodne z wcześniejszą analizą dla korelacji z ceną.

Wykonano test na korelację punktowo-dwuseryjną dla średnich kroczących oraz trendu:

- dla 3 dni: statystyka=0.05, pvalue=0.29)
- dla 7 dni: statystyka=0.052, pvalue=0.2799))
- dla 14 dni: statystyka=0.054, pvalue=0.27)

Wskazuje to na bardzo niską korelację między tymi danymi.

Wykonano również test chi kwadrat dla trendu oraz trendu z poprzedniego dnia, gdzie chi = 1.38, a pvalue = 0,24, co również sugeruje bardzo niską zależność lub jej brak.

Tabela V: Statystki dla całego zbioru po procesie przetwarznia

	price	volume	daily_active_addresses	price_diff	trend	trend_3d	trend_7d	trend_14d	prev_trend
count	424.0	424.0	424.0	424.0	424.0	424.0	424.0	424.0	424.0
mean	0.050844675123642866	0.003380802237852037	-0.0001951171846799803 2	0.008717771297182316	0.5117924528301887	0.0333915478115077	0.0153092408639978	-0.018078592005658672	0.5094339622641509
std	0.9838416891065909	1.0011386236951887	0.9945637872395989	0.0844862631148142	0.5004514220711848	0.9694438881643347	0.9530538557076708	0.9241838157525862	0.5005015549460484
min	-1.1793149484551189	-1.6062483165255876	-2.649445107495756	-0.34640355054890903	0.0	-1.1603740008390218	-1.1846733094502462	-1.325598234423197	0.0
25%	-0.5895261434611986	-0.7666910336287698	-0.6614255645135552	-0.024110290217399043	0.0	-0.5858874780555083	-0.5871741174965194	-0.5920252339312647	0.0
50%	-0.34918626160858846	-0.21017645704020 73	0.0206465988545 6050775	0.0010533142038481047	1.0	-0.34762955346813873	-0.3603209583801487	-0.3841688554910375	1.0
75%	0.6340404381405282	0.6294279243776698	0.7072974201381206	0.03948156554767132	1.0	0.6373291258280128	0.6452642783456861	0.5302056314744101	1.0
max	3.1259659993916658	2.8422231299139336	2.359269413544465	0.4524712464013172	1.0	3.0700486034123045	2.926891659095685	2.8699826714811354	1.0

Tabela VI: Wybrany fragment zbioru danych po procesie przetwarznia

date	price	volume	daily_active_addresses	price_diff	trend	trend_3d	trend_7d	trend_14d	prev_trend
2023-01-18	-1.1334190411932898	0.6033066630393074	0.6301710300651725	-0.04326053379986328	0	-1.0984739581082836	-1.1846733094502462	-1.325598234423197	0.0
2023-01-19	-1.0969967129610378	1.165622099010943	0.2270104725993282	0.036422328232252	1	-1.104331193335719	-1.1486719803428687	-1.3005927137211835	0.0
2023-01-20	-0.9515171279852039	0.17113364160491232	0.04104173116616056	0.1454795849758339	1	-1.1068580871825846	-1.1196866156889336	-1.2728123713053425	1.0
2023-01-21	-0.9422677914938683	1.0300915687749304	-0.13161388141650188	0.009249336491335636	1	-1.0606442940465104	-1.083494683601618	-1.2353938454799052	1.0
2023-01-22	-0.9475031234988132	1.4393467450930004	-0.28160708292394737	-0.005253532004944859	0	-0.9969272108133701	-1.0599460782797503	-1.197334967559858	1.0

III. Eksperymenty

A. Wstęp

Przedstawiony problem jest problemem klasyfikacji, wartośc ceny może wzrosnąć, utrzymać się lub spaść. Cechą, której wartość będzie klasyfikowana, jest trend. Jako że ustalono, że trend stały jest zjawiskiem bardzo rzadko występującym, zdecydowano się na pominięcie tego przypadku.

Przed przeprowadzeniem eksperymentów dokanano podziału danych na dane testowe - 20% oraz treningowe - 80%.

Eksperymenty zostaną wykonane dla trzech modeli, następnie zostanie porównana ich skuteczność oraz nastąpi dobór najlepszego z nich. W niektórych eksperymentach zastosowano metodę grid search z biblioteki sklearn służącą do poszukiwania najlepszych parametrów. W celu zbalansowania zbioru danych użyto funkcji SMOTE z biblioteki imblearn.

B. Regresja logistyczna

1) ***Opis****:* Regresja logistyczna jest prostym modelem klasyfikacji, który szacuje prawdopodobieństwo przynależności próbki do klasy. Stosuje przy tym funkcję logistyczną do przekształcenia liniowej kombinacji cech na wartość prawdopobieństwa.

- związek między prawdopodobieństwem a zmiennymi niezależnymi wyrażamy wzorem:

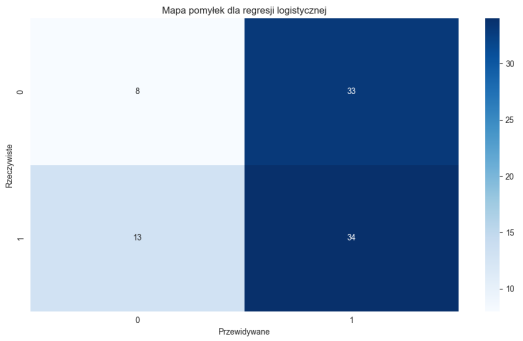
$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \qquad (1)$$

2) ***Uzasadnienie****:* Regresja logistyczna, pomimo swojej prostoty, często osiąga dobre wyniki w przypadku klasyfikacji z dwoma klasami. Postanowiono użyć jej jako modelu bazowego, z zatsrzeżeniem, że w związku z brakiem jednznacznie liniowej zależności między danymi, nie jest spodziewana wystarczająco zadowalająca skuteczność.

3) ***Eksperyment pierwszy****:* Na początku zdecydowano się na uwzględnienie tylko wolumenu oraz liczby aktywnych adresów.

Tabela VII: Raport dla regresji logistycznej

	precision	recall	f1-score	support
0	0.38095238095238093	0.1951219512195122	0.25806451612903225	41.0
1	0.5074626865671642	0.723404255319149	0.5964912280701754	47.0
accuracy			0.4772727272727273	
macro avg	0.44420753375977257	0.4592631032693306	0.42727787209960383	88.0
weighted avg	0.44852038508754927	0.4772727272727273	0.43881514637032465	88.0



Rysunek 11: Macierz pomyłek dla regresji logistycznej

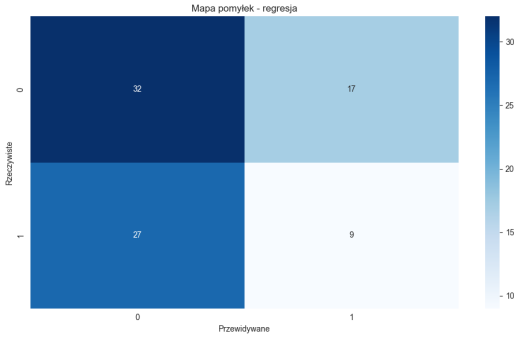
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,48

Obserwacje: Zgodnie z przewidywaniami, regresja logistyczna dla wolumenu oraz aktywnośći użytkowników wykazała się niewielką dokładnością. Model ten dokonał klasyfikacji prawidłowej dla mniej niż połowy danych. Zdecydowanie lepiej radził sobie jednak z przewidywaniem trendów wzrostowych.

4) ***Eksperyment drugi****:* W kolejnym eksperymencie uwzględniono średnie kroczące dla 3, 7 oraz 14 dni.

Tabela VIII: Raport dla regresji logistycznej

	precision	recall	f1-score	support
0	0.5423728813559322	0.6530612244897959	0.5925925925925926	49.0
1	0.34615384615384615	0.25	0.2903225806451613	36.0
accuracy			0.4823529411764706	
macro avg	0.4442633637548892	0.45153061224489793	0.4414575866188769	85.0
weighted avg	0.4592683487997546	0.4823529411764706	0.46457235223838644	85.0



Rysunek 12: Macierz pomyłek dla regresji logistycznej

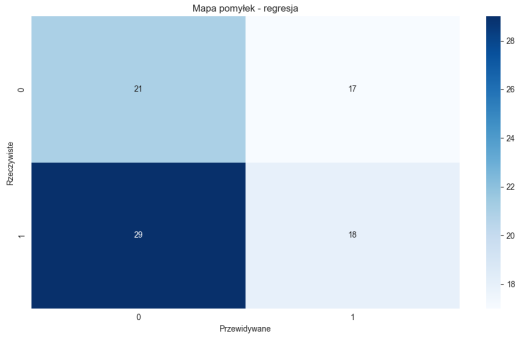
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,51

Obserwacje: Podobnie jak w przypadku poprzednim, wynik eksperymentu okazał się być niezadawalający. Widać przewgę prawidłowej klasyfikacji dla trendu wzrostowego. Dużo lepsze wyniki niż dla poprzedniego modelu otrzymano w walidacji krzyżowej.

*5) **Eksperyment trzeci**:* W tym eksperymencie uwzględniono trend dla dnia poprzedniego oraz wolumen handlu i aktywność użytkowników.

Tabela IX: Raport dla regresji logistycznej

	precision	recall	f1-score	support
0	0.42	0.5526315789473685	0.4772727272727273	38.0
1	0.5142857142857142	0.3829787234042553	0.43902439024390244	47.0
accuracy			0.4588235294117647	
macro avg	0.4671428571428571	0.4678051511758119	0.45814855875831484	85.0
weighted avg	0.4721344537815126	0.4588235294117647	0.45612364679796524	85.0



Rysunek 13: Macierz pomyłek dla regresji logistycznej

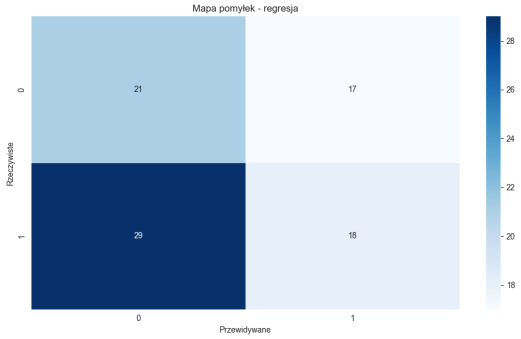
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,5

Obserwacje: Wynik tego ekperymentu okazał się być gorszy od dwóch poprzednich. Model wykazał lepsze wyniki podczas walidacji krzyżowej od pierwszego modelu, natomiast gorsze niż w przypadku drugiego eksperymentu.

*6) **Eksperyment czwarty**:* W tym eksperymencie uwzględniono wszystkie cechy.

Tabela X: Raport dla regresji logistycznej

	precision	recall	f1-score	support
0	0.42	0.5526315789473685	0.4772727272727273	38.0
1	0.5142857142857142	0.3829787234042553	0.43902439024390244	47.0
accuracy			0.4588235294117647	
macro avg	0.4671428571428571	0.4678051511758119	0.45814855875831484	85.0
weighted avg	0.4721344537815126	0.4588235294117647	0.45612364679796524	85.0



Rysunek 14: Macierz pomyłek dla regresji logistycznej

Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,5

Obserwacje: Wyniki ekperymentu są podobne do wyników dla poprzedniego modelu.

*7) **Wnioski dla regresji logistycznej**:* Regresja logistyczna nie przyniosła zadawalających rezultatów dla żadnego zbioru cech. Wykonano również eksperymanty dla różnych parametrów, jednakże nie zaobserwowano wyraźnej różnicy. Na podstawie wyników walidacji krzyżowej można stwierdzić, że najlepszym modelem używającym regresji logistycznej jest model, do którego stworzenia wykorzystano tylko średnie kroczące. Biorąc pod uwagę silnie liniową zależność tych cech z cenę, taki wynik może być uzasadniony.

C. Losowy las decyzyjny

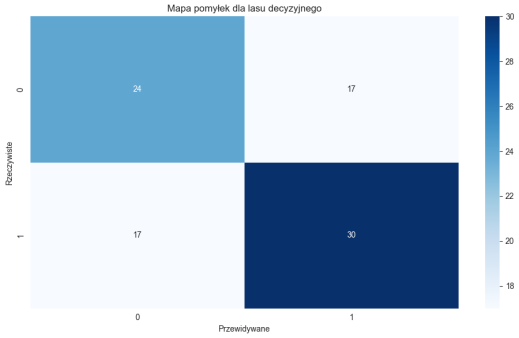
*1) **Opis**:* Losowy las decyzyjny jest modelem polegającym na tworzeniu wielu drzew decyzyjnych, z których każde trenowane jest na losowo wybranym podzespole danych (metoda losowanie z powtórzeniami). Wyniki poszczególnych drzew są łaczone poprzez głosowanie większościowe.

*2) **Uzasadnienie**:* Model ten dobrze radzi sobie z zależnościami nieliniowymi, gdyż jest w stanie odnaleźć bardziej złożone wzorce. Ta cecha może przydać się w przypadku analizowanego zbioru. Ponadto, model jest stosunkowo odporny na przetrenowanie.

*3) **Eksperyment piąty**:* W tym eksperymencie uwzględniono wolumen handlu oraz liczbę aktywnych adresów.

Tabela XI: Raport dla lasu losowego

	precision	recall	f1-score	support
0	0.5853658536585366	0.5853658536585366	0.5853658536585366	41.0
1	0.6382978723404256	0.6382978723404256	0.6382978723404256	47.0
accuracy			0.6136363636363636	
macro avg	0.611831862999481	0.611831862999481	0.611831862999481	88.0
weighted avg	0.6136363636363636	0.6136363636363636	0.6136363636363636	88.0



Rysunek 15: Macierz pomyłek dla lasu losowego

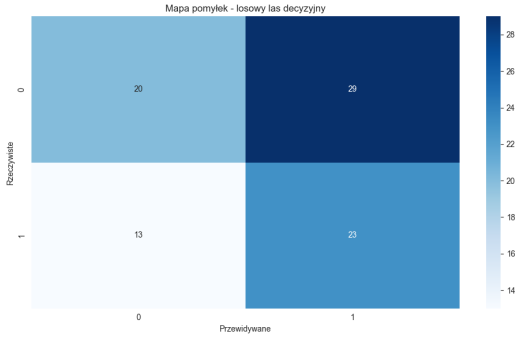
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,49

Obserwacje: Wynik eksperymentu okazał się być bardzo dobry. Widoczna jest nieznaczna przewaga prawidłowej klasyfikacji dla trendu wzrostowego, podobnie jak w przypadku poprzednich modeli. Walidacja krzyżowa dała jednak wynik o wiele niższy, co może sugerować niską skuteczność modelu dla nowych danych.

*4) **Eksperyment szósty**:* W kolejnym eksperymencie uwzględniono średnie kroczące dla 3, 7 oraz 14 dni.

Tabela XII: Raport dla lasu losowego

	precision	recall	f1-score	support
0	0.6060606060606061	0.40816326530612246	0.4878048780487805	49.0
1	0.4423076923076923	0.6388888888888888	0.5227272727272727	36.0
accuracy			0.5058823529411764	
macro avg	0.5241841491841492	0.5235260770975056	0.5052660753880266	85.0
weighted avg	0.5367064308240779	0.5058823529411764	0.5025955393243772	85.0



Rysunek 16: Macierz pomyłek dla lasu losowego

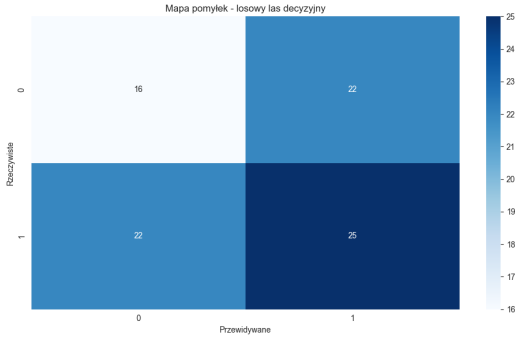
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,42

Obserwacje: Wynik znacząco gorszy niż dla poprzedniego eksperymentu. Wynik walidacji krzyżowej znacznie poniżej połowy.

5) ***Eksperyment siódmy****:* W tym eksperymencie uwzględniono trend dla dnia poprzedniego oraz wolumen handlu i aktywność użytkowników.

Tabela XIII: Raport dla lasu losowego

	precision	recall	f1-score	support
0	0.42105263157894735	0.42105263157894735	0.42105263157894735	38.0
1	0.5319148936170213	0.5319148936170213	0.5319148936170213	47.0
accuracy			0.4823529411764706	
macro avg	0.4764837625979843	0.4764837625979843	0.4764837625979843	85.0
weighted avg	0.4823529411764706	0.4823529411764706	0.4823529411764706	85.0



Rysunek 17: Macierz pomyłek dla lasu losowego

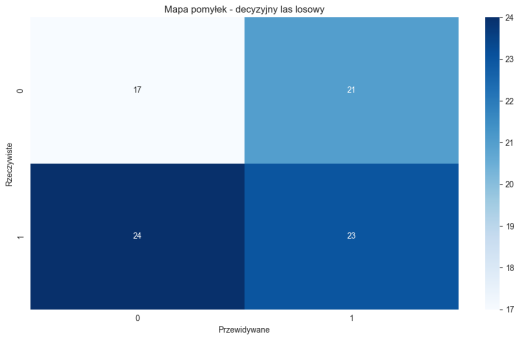
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,46

Obserwacje: Wynik znacząco gorszy niż dla poprzednich eksperymentów. Wynik walidacji krzyżowej poniżej połowy, jednak lepszy niż dla poprzedniego eksperymentu.

6) ***Eksperyment ósmy****:* W tym eksperymencie uwzględniono wszystkie cechy

Tabela XIV: Raport dla lasu losowego

	precision	recall	f1-score	support
0	0.4146341463414634	0.4473684210526316	0.4303797468354306	38.0
1	0.5227272727272727	0.4893617021276595	0.5054945054945055	47.0
accuracy			0.4705882352941764	
macro avg	0.468680709534368	0.46836506159014557	0.4679371261649743	85.0
weighted avg	0.47440328681361676	0.47058823529411764	0.47191378985868937	85.0



Rysunek 18: Macierz pomyłek dla lasu losowego

Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,44

Obserwacje: Niski wynik eksperymentu oraz niski wynik walidacji krzyżowej.

7) ***Wnioski dla losowego lasu decyzyjnego****:* Las losowy poradził sobie z problemem przewidzenia trendu znacznie lepiej od modelu regresji logistycznej, co byłoby zgodne ze spekulacjami. Najlepsze wyniki zaobserwowano dla eksperymentu piątego, najgorsze dla szóstego. Modele wykazywały na ogół niższą dokładność dla nowych danych, niż było to w przypadku regresji logistycznej. Modele, podobnie jak w poprzednim przypadku, wykazały większą dokładnośc dla trendu wzrostowego. W związku z losowością modelu, zaobserwowano różne wyniki. Przedstawione raporty oraz macierze pomyłek zostały wykonane dla losowo wybranego wyniku (najlepsza zaobserwowana dokładność była na poziomie 64% dla eksperymentu piątego).

D. Support Vector Machine

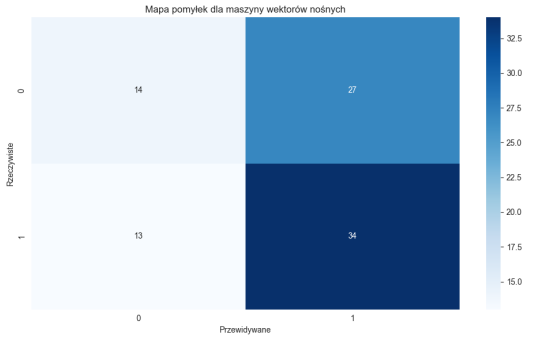
1) ***Opis****:* SVM to model, którego zadaniem jest odnalezienie takiej hiperpłaszczyzny, która rozdzielałaby klasy w sposób optymalny w przestrzeni wielowymiarowej.

2) ***Uzasadnienie****:* Podobnie jak w przypadku lasu losowego, SVM dobrze radzi sobie w odnajdywaniu nieliniowych zależności między danymi (np. dzięki wykorzystaniu radialnej funkcji bazowej jako jądra) . Ze względu na brak losowości, model może okazać się jednak być bardziej uniwersalny.

3) ***Eksperyment dziewiąty****:* W tym eksperymencie uwzględniono wolumen handlu oraz liczbę aktywnych adresów.

Tabela XV: Raport dla svm

	precision	recall	f1-score	support
0	0.5185185185185185	0.34146341463414637	0.4117647058823529	41.0
1	0.5573770491803278	0.723404255319149	0.6296296296296297	47.0
accuracy			0.5454545454545454	
macro avg	0.5379477838494231	0.5324338349766476	0.5206971677559913	88.0
weighted avg	0.5392725064856212	0.5454545454545454	0.5281243810655576	88.0



Rysunek 19: Macierz pomyłek dla svm

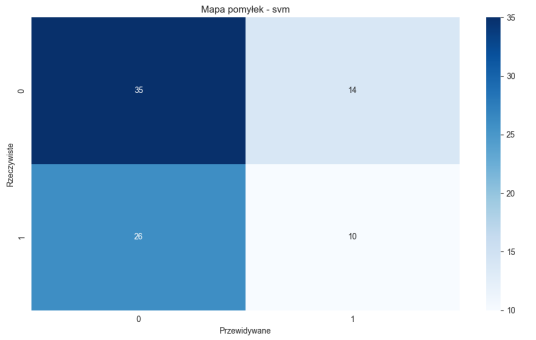
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,49

Obserwacje: Pomimo, że wynik eksperymentu wykazał dobrą dokładność, wynik walidacji krzyżowej był niższy od połowy. Model znacząco traci skuteczność dla nowych danych. Zaobserwowano wyższą precyzję dla trendów wzrostowych.

*4) **Eksperyment dziesiąty**:* W kolejnym eksperymencie uwzględniono średnie kroczące dla 3, 7 oraz 14 dni.

Tabela XVI: Raport dla svm

	precision	recall	f1-score	support
0	0.5737704918032787	0.7142857142857143	0.6363636363636364	49.0
1	0.4167	0.2777777777777778	0.3333333333333333	36.0
accuracy			0.5294117647058824	
macro avg	0.4952185792349727	0.4960317460317460 5	0.4848484848484848 6	85.0
weighted avg	0.5072324011571842	0.5294117647058824	0.5080213903743316	85.0



Rysunek 20: Macierz pomyłek dla svm

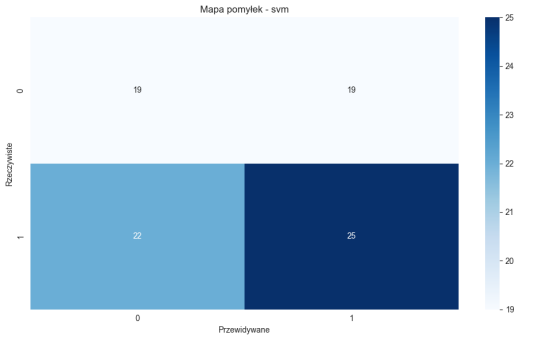
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,4

Obserwacje: Pomimo, że wynik eksperymentu wykazał dobrą dokładność, wynik walidacji krzyżowej był bardzo niski.

*5) **Eksperyment jedenasty**:* W tym eksperymencie uwzględniono trend dla dnia poprzedniego oraz wolumen handlu i aktywność użytkowników

Tabela XVII: Raport dla svm

	precision	recall	f1-score	support
0	0.4634146341463415	0.5	0.4810126582278481	38.0
1	0.5681818181818182	0.5319148936170213	0.5494505494505495	47.0
accuracy			0.5176470588235295	
macro avg	0.5157982261640799	0.5159574468085106	0.5152316038391989	85.0
weighted avg	0.5213447241424286	0.5176470588235295	0.5188547863156948	85.0



Rysunek 21: Macierz pomyłek dla svm

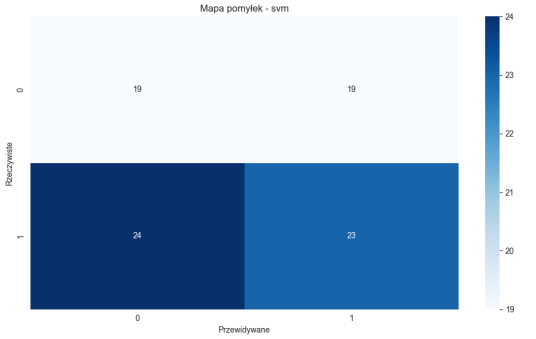
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,4

Obserwacje: Pomimo, że wynik eksperymentu wykazał dobrą dokładność, wynik walidacji krzyżowej był niski, choć wyższy niż w poprzednim eksperymencie.

*6) **Eksperyment dwunasty**:* W tym eksperymencie uwzględniono wszystkie cechy.

Tabela XVIII: Raport dla svm

	precision	recall	f1-score	support
0	0.4418604651162791	0.5	0.4691358024691358	38.0
1	0.5476190476190477	0.48936170212765956	0.5168539325842697	47.0
accuracy			0.4941176470588235 5	
macro avg	0.4947397563676634	0.49468085106382975	0.49299486752670274	85.0
weighted avg	0.5003387401472218	0.49411764705882355	0.49552112147397454	85.0



Rysunek 22: Macierz pomyłek dla svm

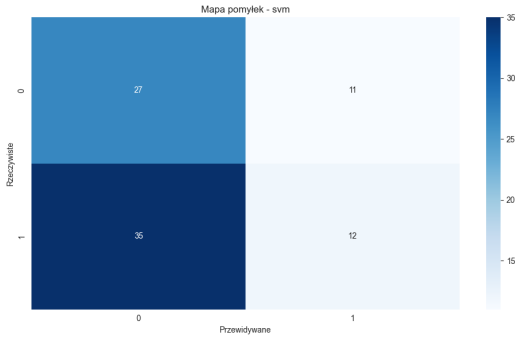
Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,47

Obserwacje: Zarówno wynik eksperyment, jak i walidacja krzyżowa były na niskim poziomie.

*7) **Eksperyment trzynasty**:* W związku z wyjątkowo niskimi wynikami dla eksperymentu dwunastego, podjęto próbe ponownego przeprowadzenia eksperymentu, nie użyto tym razem funkcji grid search.

Tabela XIX: Raport dla svm

	precision	recall	f1-score	support
0	0.43548387096774194	0.7105263157894737	0.54	38.0
1	0.5217391304347826	0.2553191489361702	0.34285714285714286	47.0
accuracy			0.4588235294117647	
macro avg	0.47861150070126224	0.4829227323628219	0.44142857142857145	85.0
weighted avg	0.4831779556142232	0.4588235294117647	0.4309915966386555	85.0



Rysunek 23: Macierz pomyłek dla svm

Walidacja krzyżowa: Średnia wartość dokładności dla walidacji krzyżowej to 0,48

Obserwacje: Wynik eksperymentu okazał się być jeszcze niższy od wyniku eksperymentu dwunastego. Model wykazał sie jednak większa dokładnośćią w przypadku walidacji krzyżowej.

*8) **Wnioski dla SVM**:* SVM średnio wykazał się lepszymi wynikami dla zbioru testowego od poprzednich modeli, jednakże znacznie gorszymi wynikami walidacji krzyżowej. Modele, podobnie jak w poprzednich eksperymentach, wykazały się lepszą precyzją w przypadku trendów wzrostowych.

IV. Wnioski

Przedstawione eksperymenty, zgodnie z oczekiwaniami, wykazały pewną zależnośc między ceną a wolumenem oraz ceną a aktywnością użytkowników. Zależnośc okazała się jednak nie być jednoznacznie liniowa, nie jest też ona silna. Wyniki eksperymentów dla średnich kroczących oraz trendu z dnia poprzedniego potwierdzają bardzo niską zależność dla tych cech. W tym przypadku większość modeli wykazała się bardzo niską dokładnością (zazwyczaj klasyfikacja prawidłowa dla mniej niż połowy danych testowych). Najwyższy wynik dla zbioru testowego otrzymano dla lasu losowego (do trenowania którego użyto dwóch cech, wolumenu oraz aktywności użytkowników), średnio najlepsze wyniki dla zaobserwowano SVM, a najgorsze dla regresji logistycznej. Regresja logistyczna wykazała najlepsze wyniki w walidacji krzyżowej (średnia dokładność to 0,4875), co może sugrować jej uniwersalność, a SVM najniższe (0,448). Średnie wartości dokładności dla walidacji krzyżowej dla każdego modelu nie przekraczają połowy. Wszystkie modele wykazały się wyższą precyzją klasyfikacji trendów wzrostowych.

Powodem, dla którego dokładność modeli nie jest wysoka, może być zbyt mała liczba danych, w ewentualnej przyszłej analizie warto byłoby skupić się również na czynnikach trudniej dostępnych, takich jak sentyment. Kolejnym powodem może być również nieprzewidywalność rynku kryptowalut. Jak było to wspomniane na wstępie, jest on rynkiem chwiejnym, zachowanie ceny kryptowaluty takiej jak Bitcoin jest trudniejsze do przewidzenia, niż w przypadku na przykład akcji giełdowych.

Pomimo podjętych prób zbalansowania zbioru danych oraz wyprowadzenia nowych cech, a także szukania najlepszych parametrów dla modeli, wyniki nie są zadawalające. Nawet dla wysokiej dokładności dla zbioru testowego, średnia dokładność walidacji krzyżowej wyniosła co najwyżej 51%. Oznacza to, że wolumen, cena oraz liczba aktywnych adresów nie są wystarczającą podstawą do wyliczenia trendu ceny dla dnia następnego.

Literatura

[1] Api. Api.Santiment. dostęp: 14.05.2024.
[2] Dane dotyczące cen, wolumenu handlu oraz liczby aktywnych adresów dla bitcoina w latach 2023/2024. https://app.santiment.net/charts. dostęp: 14.05.2024.
[3] Platforma. Santiment. dostęp: 14.05.2024.