Podstawy Przetwarzania Danych
Raport końcowy

Paweł Młyniec
Monika Zielińska
Urszula Żukowska

28 lutego 2021

1 Wstęp

Cel projektu to przewidywanie mocy wydarzenia medialnego wyrażonej dodatnią liczbą rzeczywistą na podstawie danych o aktywności zaobserwowanej w mediach społecznościowych. W tym celu została dokonana analiza danych, przetworzenie ich, a następnie przetestowanie różnorodnych modeli z zakresu uczenia maszynowego przeznaczonych do rozwiązywania zadania regresji.

2 Opis danych

Zbiór danych zawiera 140706 rekordów opisujących wydarzenia medialne. Każdy z rekordów opisany jest przez 78 pól, w tym 77 pól opisujących obserwacje aktywności społeczności dookoła tego tematu, zaś ostatnie pole zawiera liczbę określającą moc tego wydarzenia, określoną jako średnią liczbę dotyczących go aktywnych dyskusji.
Pola opisujące obserwacje zawierają obserwacje 11 czynników zarejestrowane na przestrzeni 7 dni. Z tego powodu każde kolejne 7 kolumn opisuje tygodniowy przebieg zmian danego parametru.

Obserwowane parametry to:

- Liczba utworzonych dyskusji zawierających określony temat
- Liczba nowych autorów wykazujących aktywność w sieciach społecznościowych dotyczącą tematu
- Całkowita liczba autorów wykazujących aktywność odnośnie tematu
- Stosunek liczby utworzonych dyskusji do liczby istniejących dyskusji dotyczących tematu
- Całkowita liczba kontenerów zawierających rozmowy na dany temat

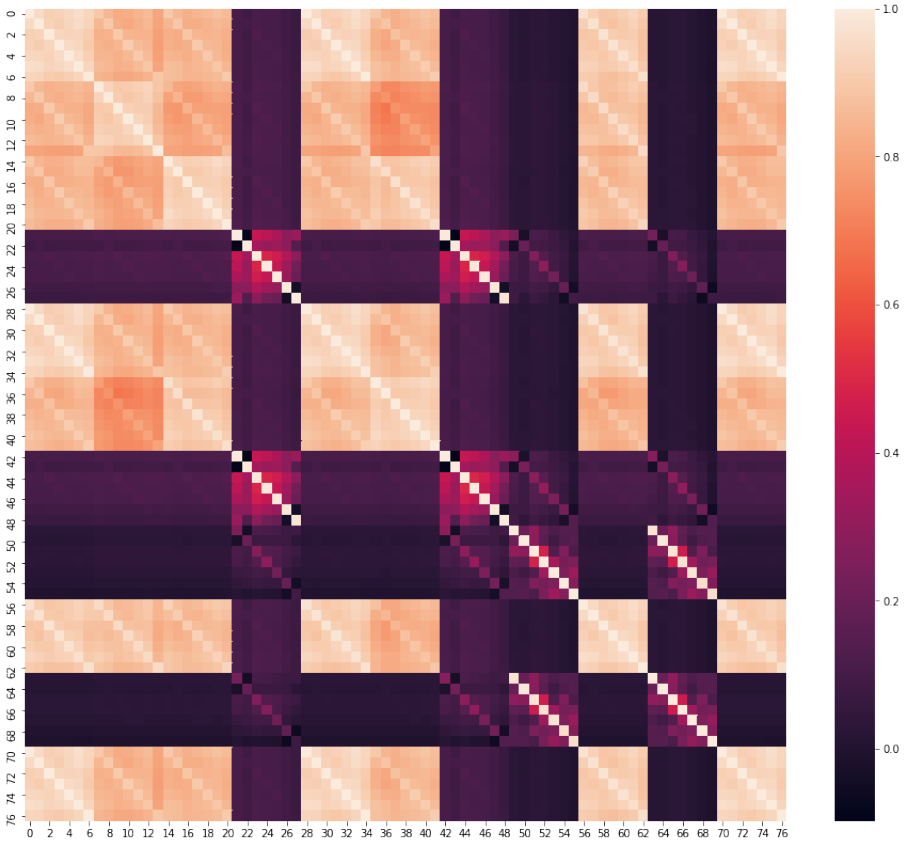
- Liczba wpisów dotyczących danego tematu
- Gęstość wpisów w dyskusji dotyczącej danego tematu
- Średnia liczba autorów wykazujących aktywność dookoła tego tematu w jednej dyskusji
- Liczba autorów wykazujących aktywność dookoła tematu
- Średnia długość dyskusji
- Liczba dyskusji dotyczących tematu

Przewidywana wartość - moc wydarzenia medialnego w zbiorze uczącym charakteryzuje się następującymi właściwościami:

- średnia: 193.234
- odchylenie standardowe: 653.167
- minimalna wartość: 0
- dolny kwartyl (25%): 2
- mediana (50%): 19
- górny kwartyl (75%): 130
- maksymalna wartość: 75724.500

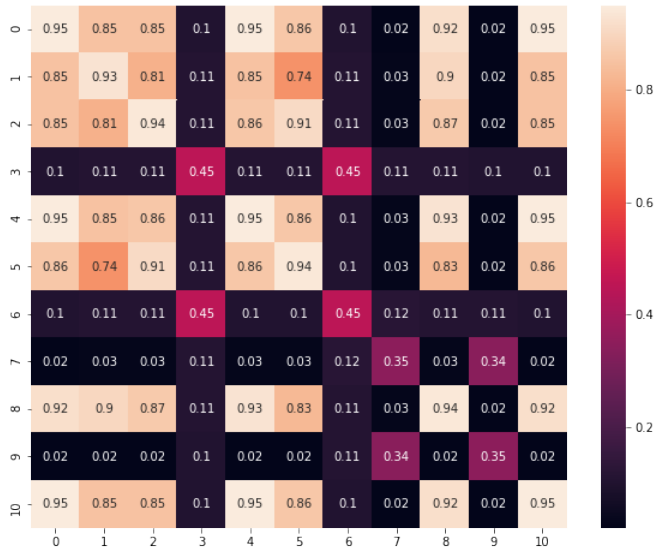
3 Wstępna analiza i przetwarzanie

Analizę danych rozpoczęto od zbadania kowariancji wszystkich danych. Wartości danych zostały znormalizowane, tak aby mieściły się w przedziale $[0, 1]$. Macierz kowariancji dla tak przygotowanych danych wygląda następująco.



Rysunek 1: Heatmapa prezentująca kowariancję dla wszystkich danych

Przeprowadzono także analizę danych traktując tygodniowe obserwacje jako jedną zmienną. Analizowano zatem 11 zmiennych zamiast 77. Macierz kowariancji takiego zbioru danych wygląda następująco.



Rysunek 2: Heatmapa prezentująca kowariancję 11 atrybutów

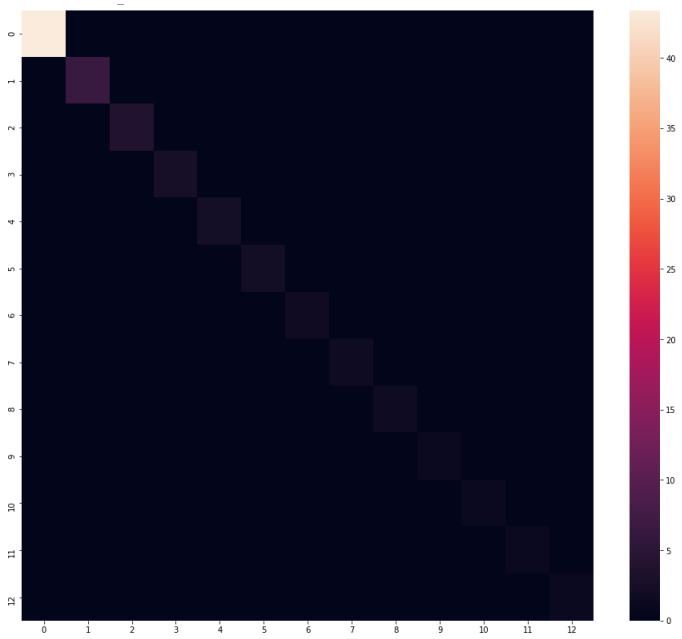
Zbadano także ich korelację. Poniżej znajduje się macierz przedstawiająca wyniki analizy.

Rysunek 3: Heatmapa prezentująca korelację 11 atrybutów

Widać wyraźnie, że wartości niektórych atrybutów są dosyć mocno powią-

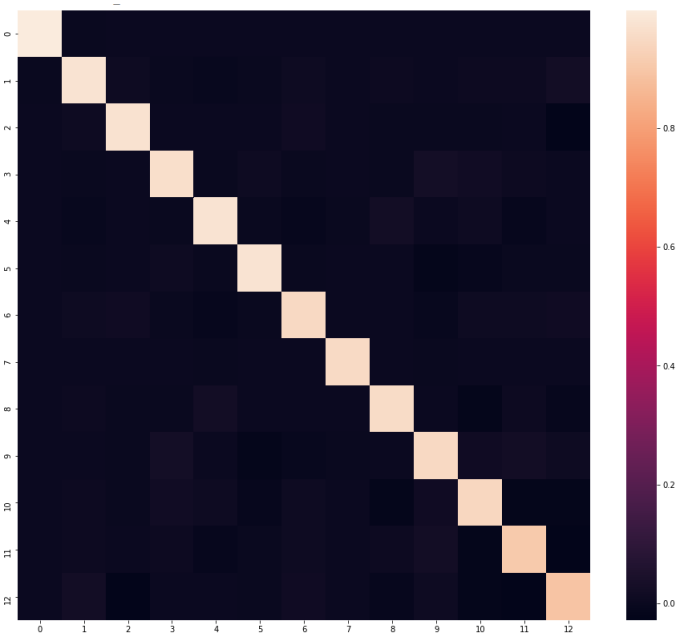
zane ze sobą. Utworzyły się grupy atrybutów, dla których wartości korelacji oraz kowariancji są wysokie. Grupy te pokrywają się na wszystkich wykresach i są widoczne zarówno w macierzach dla 77 atrybutów, jak i dla 11 atrybutów. Można zatem przeprowadzić dalszą analizę danych w celu zbadania możliwości redukcji liczby atrybutów.

Na danych została przeprowadzona analiza głównych składowych (ang. *Principal Component Analysis*, PCA) oraz analiza czynnikowa. W tym celu została najpierw przeprowadzona ich normalizacja do przedziału $[0, 1]$. Analiza głównych składowych z wykorzystanym kryterium wartości własnych większych od 1 daje rezultat 13 składowych. Ich kowariancja została przedstawiona za pomocą heatmapy poniżej.



Rysunek 4: Heatmapa prezentująca kowariancję głównych składowych

Z kolei analiza czynnikowa z identycznym kryterium dała rezultat 13 czynników, zaś ich macierz kowariancji prezentuje się następująco.



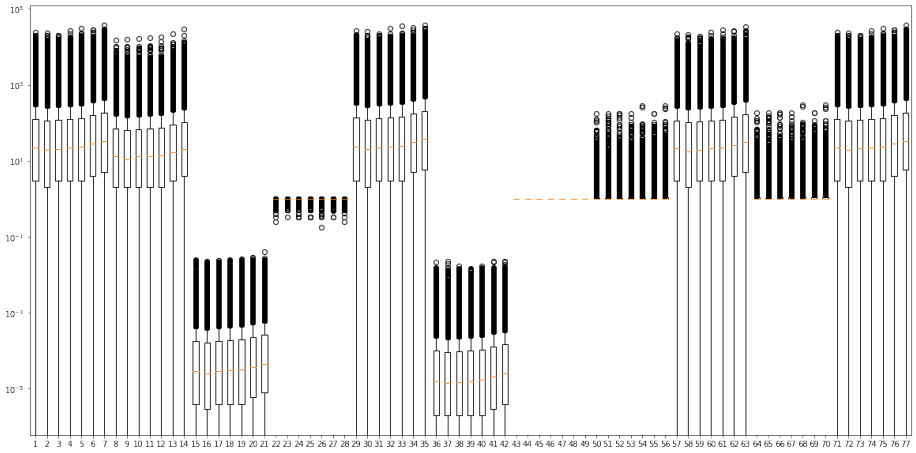
Rysunek 5: Heatmapa prezentująca kowariancję czynników

Można faktycznie zauważyć, że każda z metod zwraca czynniki, których kowariancja jest zerowa i istnieje zgodność pomiędzy nimi co do liczby wybranych czynników. Jednak istotną różnicą jest wariancja każdego z czynników. W głównych składowych otrzymanych za pomocą metody PCA wariancja wysoka występuje jedynie w pierwszej składowej, zaś pozostałe mają wariancję stosunkowo niewielką. W przypadku analizy czynnikowej każdy z czynników ma porównywalną wariancję.

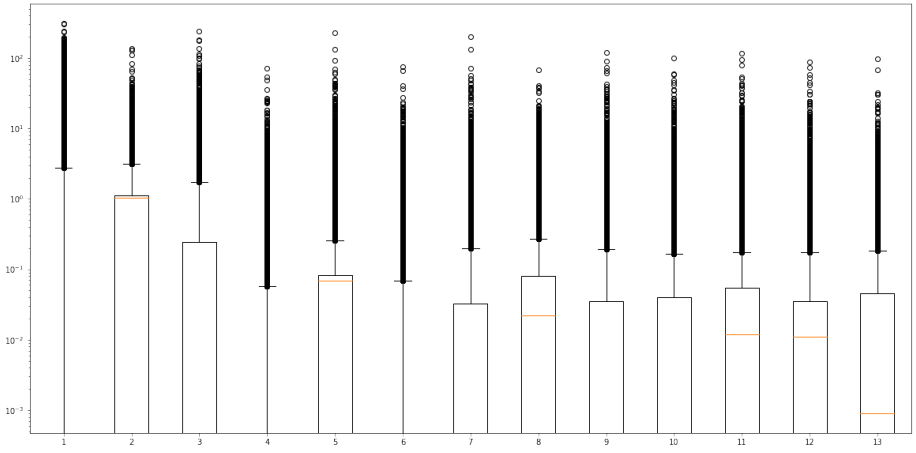
Jako, że koszt obliczeniowy nie jest istotnym kryterium w przypadku naszego badania, w dalszej części zadania nie zostały wykorzystane wygenerowane czynniki. Spowodowałby one przyspieszenie działania metod obliczeniowych, jednak wpłynęłyby negatywnie na wytłumaczalność modeli.

3.1 Dane odstające i wykresy pudełkowe

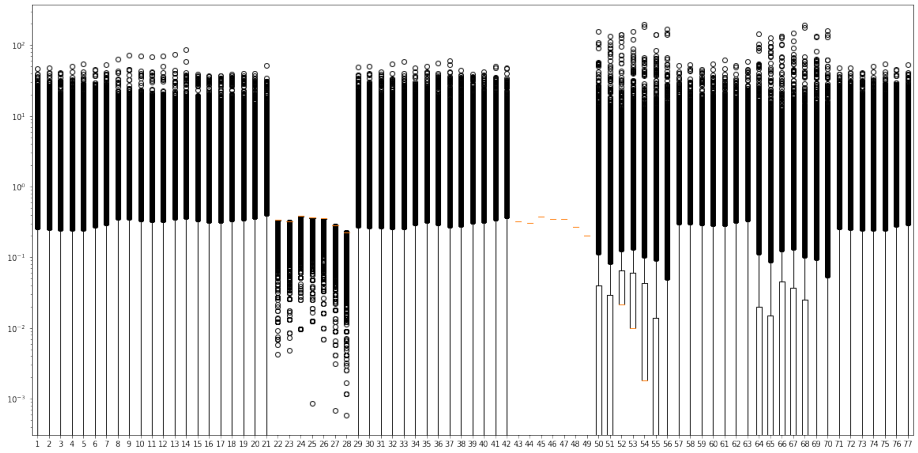
Wykres pudełkowy dla danych przedstawia się następująco:



Rysunek 6: Wykres pudełkowy dla wszystkich zmiennych, wartość zmiennej od jej typu. Kolejny indeks definiuje typ zmiennej jak w sekcji 2.



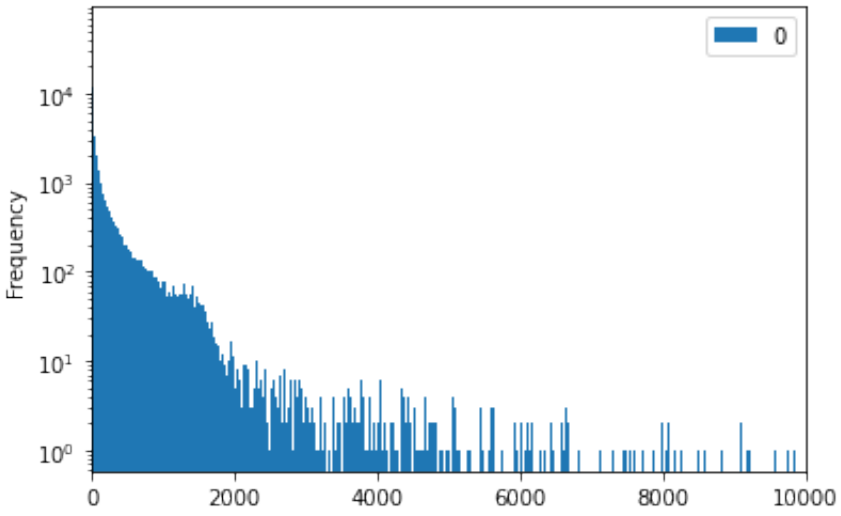
Rysunek 7: Wykres pudełkowy dla zmiennych po przetworzeniu PCA, wartość zmiennej od jej typu. Kolejny indeks definiuje kolejną zmienną PCA.



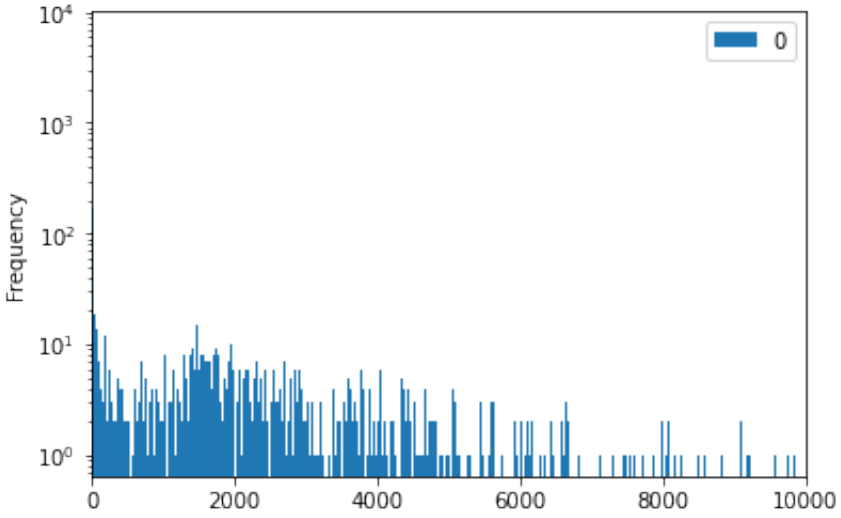
Rysunek 8: Wykres pudełkowy dla wszystkich zmiennych ustandaryzowanych, wartość zmiennej od jej typu. Kolejny indeks definiuje typ zmiennej jak w sekcji 2.

Dla wszystkich rysunków można zauważyć znaczną liczbę obserwacji odstających. Co jest ciekawe na rysunku 6 jak i 8 pierwsze kwantyle są rozciągniętne na dużej przestrzeni. Po normalizacji zgodnie z przewidywaniami drugi kwantyl znajduje się w okolicach zera natomiast znacząco powiększyły się trzecie kwantyle.

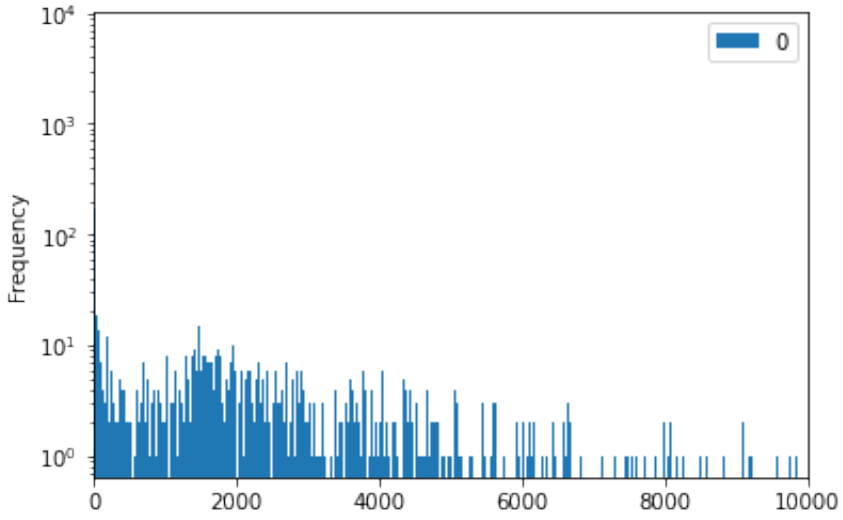
3.2 Kryterium Chauveneta



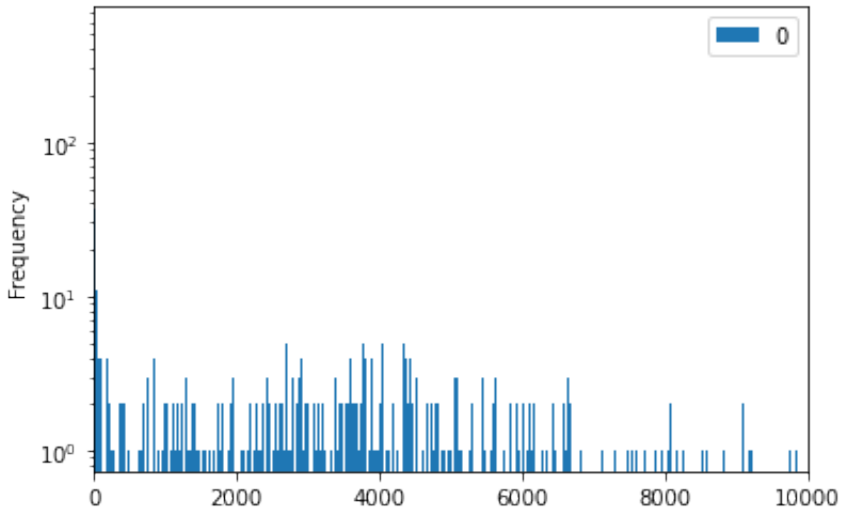
Rysunek 9: Histogram dla wszystkich etykiet przed odcięciem wartości odstających



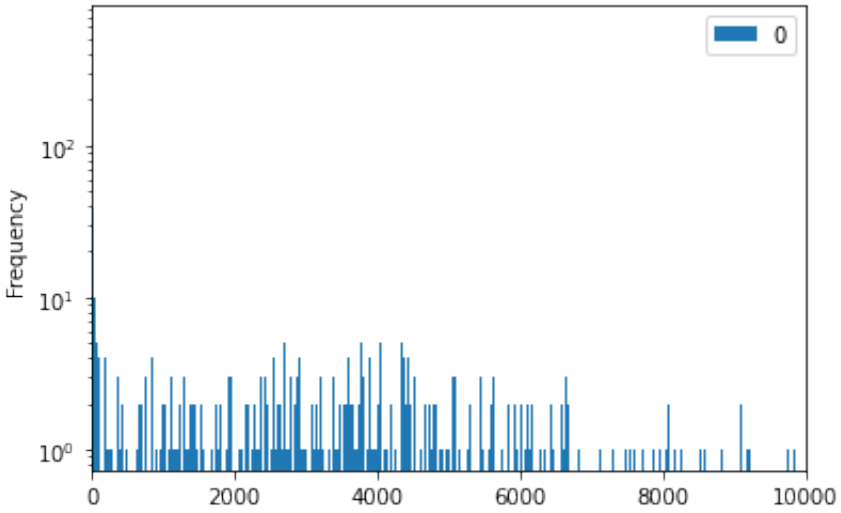
Rysunek 10: Histogram dla wszystkich etykiet po odcięciu wartości odstających



Rysunek 11: Histogram dla wszystkich etykiet po odcięciu wartości odstających po standaryzacji



Rysunek 12: Histogram dla wszystkich etykiet po odcięciu wartości odstających po analizie czynnikowej



Rysunek 13: Histogram dla wszystkich etykiet po odcięciu wartości odstających po PCA

	Wszystkie etykiety	Odrzucone etykiety Chauvenet	Odrzucone etykiety PCA	Odrzucone etykiety FA
Liczba	120707	1879	6388	1879
Średnia	193	2815	2367	804
Odchylenie standardowe	653	3990	2367	2152

Standaryzacja nie wpłynęła na liczbę odrzuconych obserwacji dlatego nie jest zawarta w tabeli.

Po zastosowaniu kryterium Chauveneta na nieprzetworzonych danych i sporządzeniu histogramów widać, że obserwacje odstające skupione były w kierunku małych wartości etykiet. Jednakże pod względem wartości średnich w wypadku danych nieprzetworzonych, po standaryzacji i po PCA ich wartość średnia oraz odchylenie standardowe są dużo wyższe niż wszystkich etykiet.

Można wyciągnąć z tego wnioski, że obserwacje były odrzucane proporcjonalnie do występujących wartości oraz dodatkowo były usunięte wartości o dużych wartościach etykiet.

Co ciekawe po zastosowaniu PCA liczba odrzuconych obserwacji wzrosła około 3,5 krotnie.

4 Metody regresji

Podczas trenowania wszystkich modeli dane zostały podzielone na zbiór treningowy oraz testowy o rozmiarze odpowiednio 90% oraz 10%.

Wyniki regresji były oceniane na podstawie następujących miar:

- Root Mean Squared Error
- Mean Absolute Error
- R^2

4.1 Prosta sieć neuronowa

Jako pierwszy model została zastosowana prosta sieć neuronowa. Składa się ona z pięciu warstw: warstwy wejściowej o rozmiarze 77, trzech warstw ukrytych o rozmiarze 100 i funkcji aktywacji ReLU oraz warstwy wyjściowej rozmiaru 1. Wytrenowana sieć dała następujące rezultaty:

Rodzaj danych	$\boldsymbol{RMSE}$	$\boldsymbol{MAE}$	$\boldsymbol{R^2}$
Dane nieprzetworzone	212.293	60.288	0.8416
Kryterium Chauveneta	196	50.99	0.86
PCA	195.83	61.47	0.865
Kryterium Chauveneta i PCA	205.88	64.449	0.851
Dane standaryzowane	174.313	55.1299	0.893
Dane standaryzowane i kryterium Chauveneta	154.105	51.273	0.9165

4.2 Rekurencyjna sieć neuronowa

Po wykonaniu testów z użyciem rekurencyjnej sieci neuronowej można stwierdzić, iż nie jest ona odpowiednim narzędziem do przeprowadzenia regresji na badanym zbiorze danych. Wyniki uzyskane za jej pomocą były dużo gorsze niż dla pozostałych użytych metod.

4.3 Regresja ridge

Kolejną przetestowaną metodą regresji jest regresja ridge. Pozwoliła ona na uzyskanie poniższych wyników:
$RMSE = 189.41$
$MAE = 63.64$
$R^2 = 0.91$

4.4 Regresja ElasticNet

$RMSE = 177.71$
$MAE = 62.37$
$R^2 = 0.90$

4.5 Model XGBoost

Do zadania predykcji został również użyty popularny model XGBoost. Test z wykorzystaniem tej metody został wykonany dla wielu różnych wartości parametrów *learning_rate*, *colsample_bytree*, *subsample*, *max_depth* oraz *n_estimators*. Najlepsze rezultaty uzyskano przy pomocy modeli o następujących parametrach:

1. *learning_rate*: 0.15, *max_depth*: 9, *subsample*: 0.9, *colsample_bytree*: 0.6, *n_estimators*: 25
2. *learning_rate*: 0.1, *max_depth*: 9, *subsample*: 0.9, *colsample_bytree*: 0.6, *n_estimators*: 25
3. *learning_rate*: 0.15, *max_depth*: 8, *subsample*: 0.9, *colsample_bytree*: 0.6, *n_estimators*: 25
4. *learning_rate*: 0.15, *max_depth*: 10, *subsample*: 0.8, *colsample_bytree*: 1, *n_estimators*: 25
5. *learning_rate*: 0.15, *max_depth*: 9, *subsample*: 0.9, *colsample_bytree*: 0.7, *n_estimators*: 25
6. *learning_rate*: 0.15, *max_depth*: 10, *subsample*: 0.8, *colsample_bytree*: 0.5, *n_estimators*: 25

Poniżej przestawione zostały wyniki dla tych modeli:

Numer modelu	$RMSE$	MAE	R^2
1.	165.07	52.84	0.90
2.	165.59	54.29	0.90
3.	165.66	53.14	0.90
4.	168.92	52.40	0.90
5.	165.70	52.66	0.90
6.	171.41	52.75	0.90

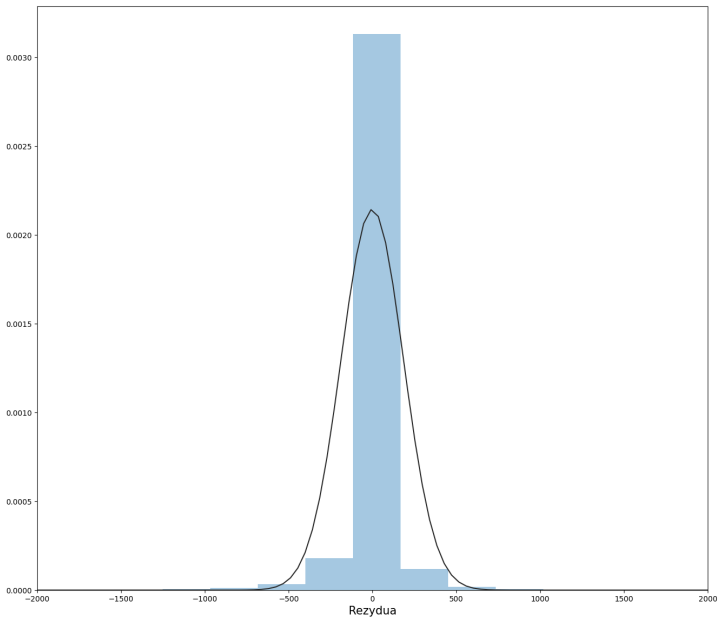
Tabela 1: Wyniki regresji uzyskane za pomocą modelu XGBoost

5 Wykres rezyduów

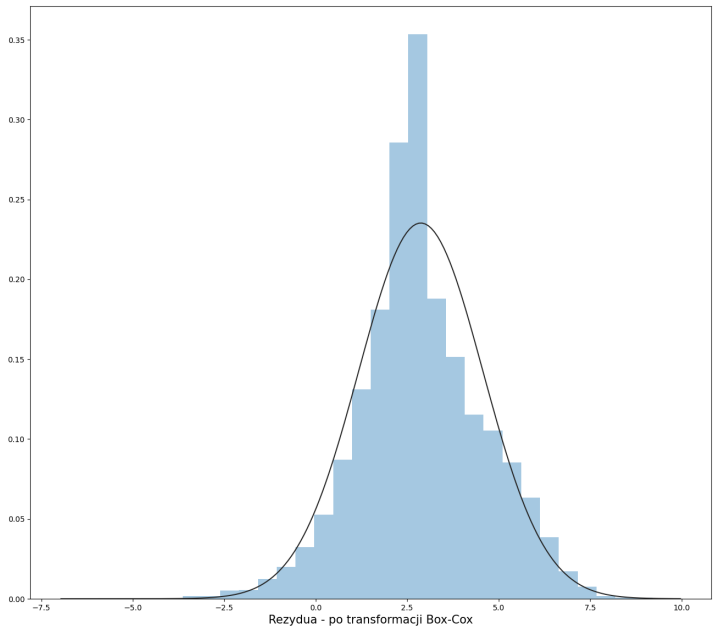
Poniżej przedstawiono wykresy rezyduów, czyli różnicy pomiędzy wynikami modelu regresji i wartościami rzeczywistymi.

Dla każdego modelu zostały przedstawione dwa wykresy - pierwszy dla oryginalnych danych, a drugi dla danych poddanych transformacji Box-Cox. Transformacja ta umożliwia takie dopasowanie oryginalnych danych, aby ich nowy rozkład jak najbardziej przypominał rozkład normalny.

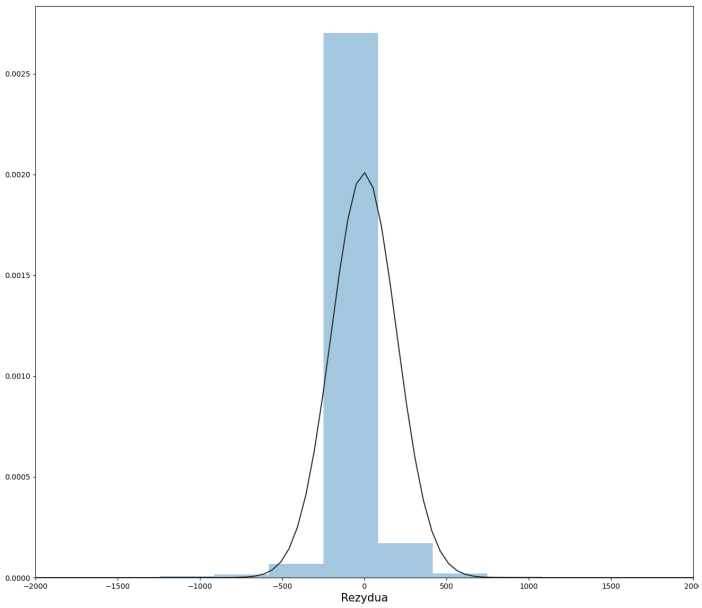
Na wszystkich wykresach rezyduów największa gęstość obserwacji znajduje się w centralnym punkcie. Oznacza to, że większość predykcji została wykonana poprawnie i ich wartości nie różnią się dużo od wartości rzeczywistych. Centralnym punktem wykresów dla danych oryginalnych jest oczywiście wartość 0. Natomiast dla danych po transformacji Box-Cox środek wykresów jest przesunięty. Jest to spowodowane faktem, iż tylko wartości nieujemne mogą być poddane tej transformacji. W związku z tym, zbiór danych poddany normalizacji Box-Cox składał się z modułów różnicy pomiędzy wynikami predykcji i wartościami rzeczywistymi.



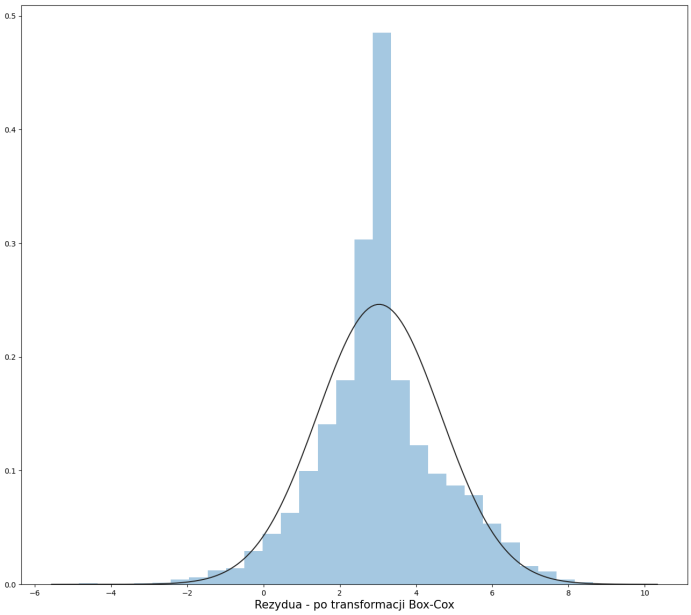
Rysunek 14: Wykres rezyduów - regresja Ridge



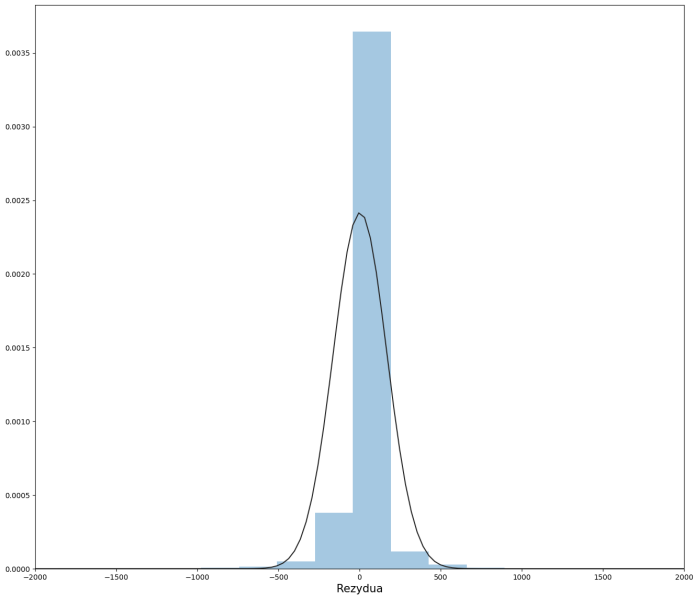
Rysunek 15: Wykres rezyduów po normalizacji metodą Box-Cox - regresja Ridge



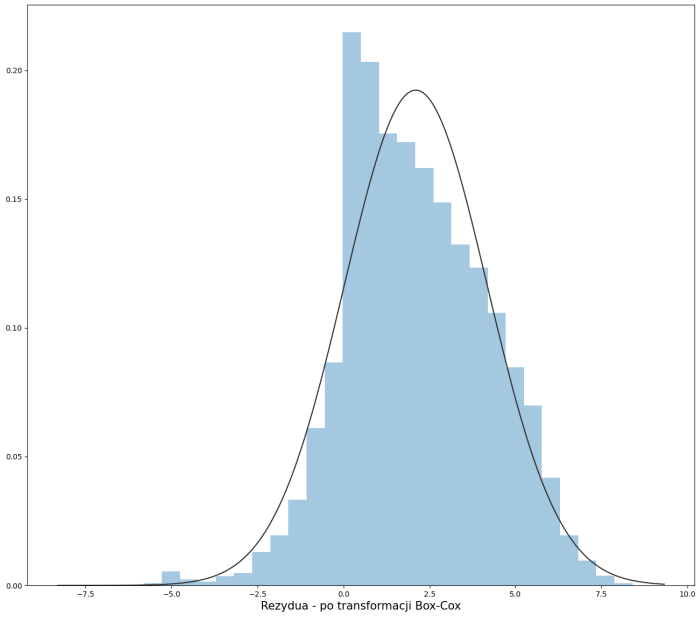
Rysunek 16: Wykres rezyduów - regresja ElasticNet



Rysunek 17: Wykres rezyduów po normalizacji metodą Box-Cox - regresja ElasticNet



Rysunek 18: Wykres rezyduów - model XGBoost



Rysunek 19: Wykres rezyduów po normalizacji metodą Box-Cox - model XGBoost

6 Wnioski

Przeprowadzone eksperymenty ukazały przewagę modelu XGBoost ponad pozostałymi sprawdzonymi modelami - prostą siecią neuronową, siecią rekurencyjną, regresją ridge oraz regresją ElasticNet. Przy optymalnych parametrach pozwolił on na znaczące zmniejszenie miar błędów RMSE oraz MAE w porównaniu do pozostałych modeli. Jednakże jest to wynik, który wciąż pozostawia duże pole do pracy i poprawy, gdyż błąd bezwzględny rzędu wielkości 50 prawdopodobnie okazałby się zbyt duży do realnego zastososowania przy wartościach rzeczywistych o takich parametrach statystycznych jak moc wydarzenia medialnego w testowanym zbiorze (górny kwantyl przewidywanej wartości: 130).

W zależności od zastosowania można by także rozważyć standaryzację danych oraz zastosowanie kryterium Chauveneta, gdyż obie te metody wstępnego przetwarzania danych przyniosły pozytywne rezultaty. Jednakże usuwanie rekordów o wartościach odstających przy pomocy kryterium Chauveneta może oddalić model od rzeczywistości, nie pozwalając na naukę sytuacji rzadko występujących, nietypowych.