Przewidywanie pogody: Analiza skuteczności modeli nauczania maszynowego

Oskar Gawryszewski

Maj 2024

1 Wstęp

Prognozowanie pogody odgrywa kluczową rolę w wielu dziedzinach życia, od rolnictwa i przemysłu do codziennych decyzji podejmowanych przez ludzi. Dostarczenie dokładnych prognoz pogody jest sporym wyzwaniem z powodu skomplikowanej natury atmosferycznych procesów, które mogą być trudne do przewidzenia z dużą dokładnością.

W naszym badaniu skupiamy się na analizie i porównaniu różnych metod prognozowania pogody. Naszym celem jest ocena wydajności różnych modeli prognostycznych w przewidywaniu temperatury maksymalnej (tmax), temperatury minimalnej (tmin) i opadów (prcp) na podstawie danych pogodowych z John F. Kennedy Airport (JFK), Nowy Jork, od stycznia 1970 roku do października 2022 roku, uzyskanych z Narodowych Ośrodków ds. Informacji Środowiskowej (National Centers For Environmental Information).

Metodyka, którą stosujemy do sprawdzania prognoz pogody, obejmuje analizę różnych modeli prognozowych, w tym: Ridge Regression, XGBoost, Random Forest, Naive Bayes, Sieć neuronowa z regularyzacją L1 i L2, K-najbliższych sąsiadów (KNN), Lasso Regression.

Dla każdego z tych modeli przeprowadzamy proces szkolenia na danych treningowych, ocenę za pomocą danych testowych oraz analizę wyników za pomocą metryk, takich jak średni błąd bezwzględny (MAE). Poprzez porównanie wydajności tych różnych modeli, dążymy do zidentyfikowania najbardziej efektywnych technik prognozowania pogody dla danych z JFK Airport.

2 Dane i pre-processing

2.1 Zbieranie danych

Dane pogodowe wykorzystane w tym badaniu zostały pozyskane z Narodowych Centrów Informacji Środowiskowej (NCEI) i obejmują szczegółowo dane pogodowe

z lotniska JFK. Zbiór danych obejmuje okres od 1 stycznia 1970 roku do 21 października 2022 roku.

2.2 Przetwarzanie i Oczyszczanie Danych

Przetwarzanie danych to kluczowy etap w eksploracji danych, mający na celu przekształcenie surowych danych w czysty i zrozumiały format odpowiedni do analizy i modelowania. W tej sekcji przedstawię kroki podjęte w celu przetworzenia i oczyszczenia surowych danych pogodowych:

2.2.1 Pozbycie sie kolumn z dużą ilością null wartości

Obecność brakujących wartości w zbiorze danych została oceniona poprzez obliczenie procentowego udziału wartości null dla każdej kolumny. Kolumny, których procent wartości null przekraczał 5%, zostały usunięte. Po usunięciu tych kolumn sprawdzamy ile wartości null zostało w naszym zbiorze.

Table 1: Dane pogodowe

station	0
name	0
prcp	0
snow	0
snwd	2
tmax	0
tmin	0

2.2.2 Uzupełnienie Brakujących Wartości

W celu zachowania ciągłości w zbiorze danych brakujące wartości zostały zastąpione metodą wypełniania w przód. Polega to na zastąpieniu brakujących wartości najnowszą niepustą wartością w tej samej kolumnie.

2.2.3 Standaryzacja Danych

Nazwy kolumn zostały znormalizowane do małych liter dla spójności ułatwienia odniesienia w dalszej analizie.

2.2.4 Analiza korelacji

Obliczono macierz korelacji w celu identyfikacji zależności między zmiennymi numerycznymi w zbiorze danych. Analiza ta dostarcza informacji na temat potencjalnych zależności i wieloliniowości między cechami. Możemy tutaj zauważyć sporą korelację między **tmin** (temperatura minimalna), a **tmax** (temperatura maksymalna).

Table 2: Korelacja między zmiennymi pogodowymi

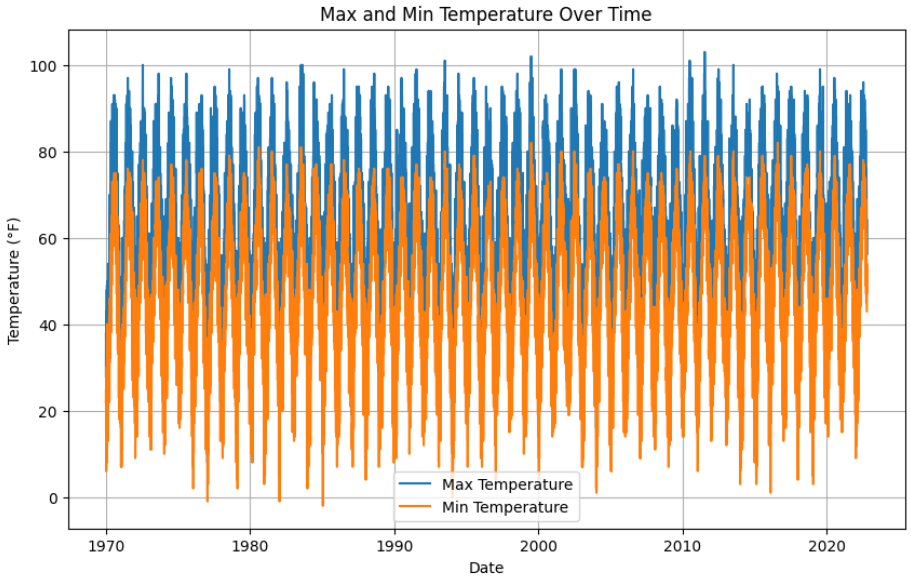
	prcp	snow	snwd	tmax	tmin
prcp	1.000	0.151	0.001	-0.008	0.052
snow	0.151	1.000	0.232	-0.175	-0.159
snwd	0.001	0.232	1.000	-0.260	-0.257
tmax	-0.008	-0.175	-0.260	1.000	0.955
tmin	0.052	-0.159	-0.257	0.955	1.000

3 Wizualizacja danych

3.1 Analiza temperatury i opadów

Wykres liniowy temperatury

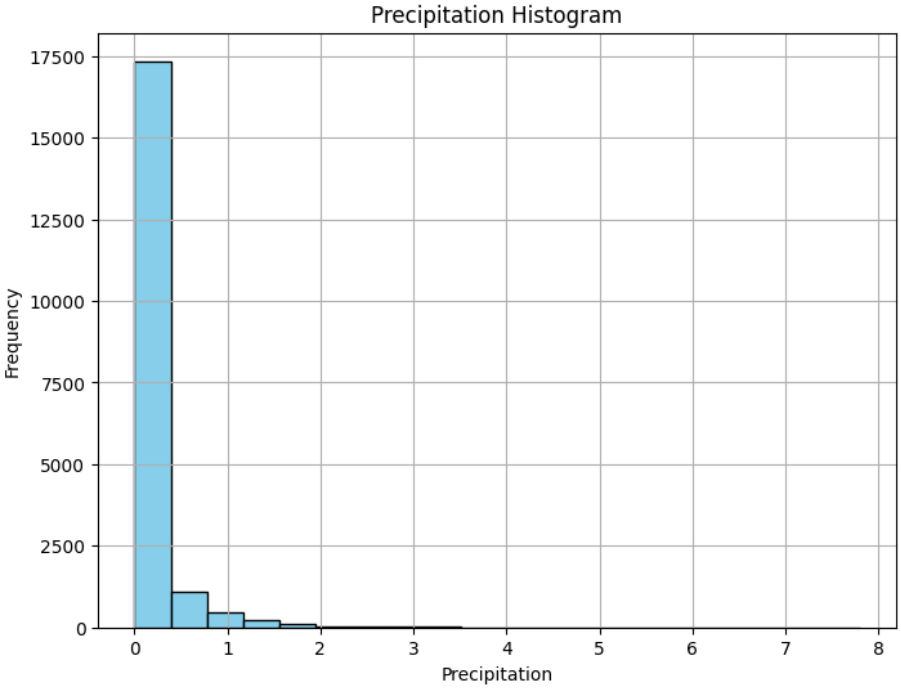
Na wykresie liniowym temperatury przedstawiono maksymalną i minimalną temperaturę w kolejnych okresach czasu. Pozwala to na zobrazowanie zmienności temperatury w czasie.



Wykres liniowy temperatury

Histogram opadów

Histogram opadów prezentuje częstość występowania różnych poziomów opadów. Jest to przydatne narzędzie do analizy rozkładu opadów w badanym okresie.



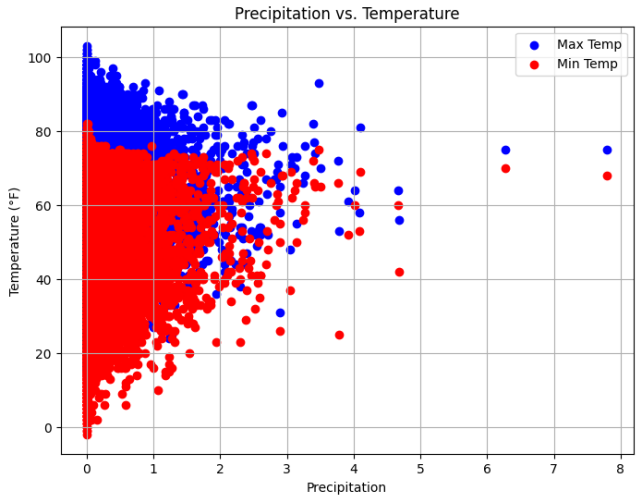
Histogram opadów

Figure 2: Analiza temperatury

3.2 Relacja między opadami a temperaturą

Wykres punktowy opadów vs. temperatury

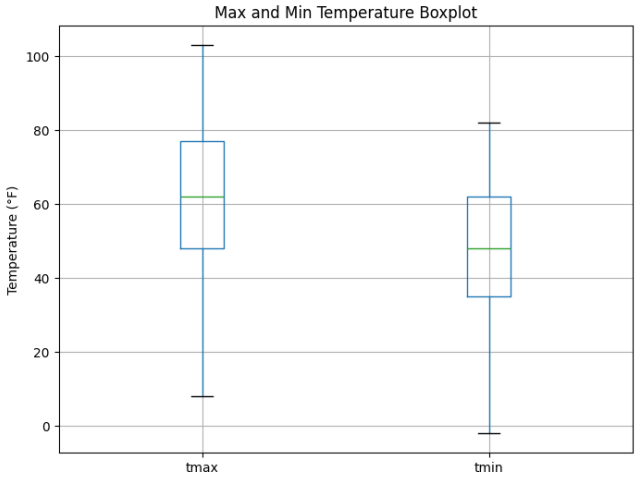
Wykres punktowy przedstawia zależność między poziomem opadów a temperaturą. Możemy zaobserwować, czy istnieje jakaś relacja między tymi dwoma zmiennymi.



Wykres punktowy opadów vs. temperatury

Wykres skrzypcowy temperatury

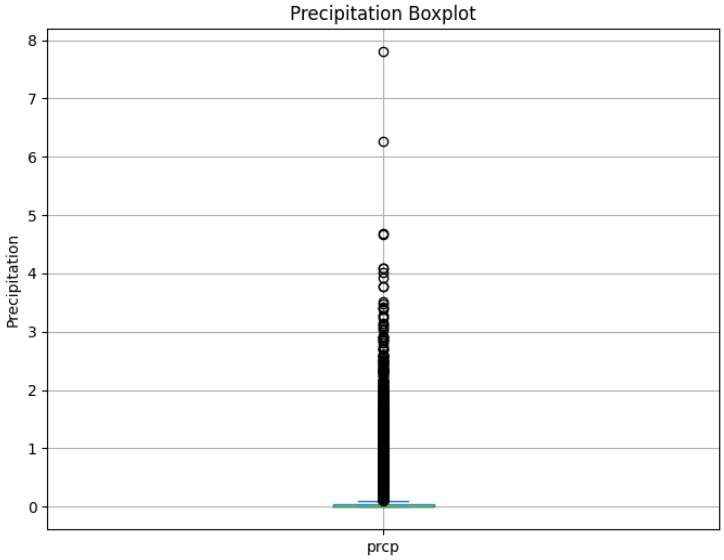
Na wykresie skrzypcowym temperatury wartości maksymalnej i minimalnej temperatury są prezentowane w formie pudełka, co pozwala na zobaczenie rozkładu tych temperatur w poszczególnych miesiącach lub okresach czasu. Linia wewnątrz pudełka reprezentuje medianę, a rozpiętość pudełka pokazuje kwartyle danych.



Wykres skrzypcowy temperatury

Wykres skrzypcowy opadów

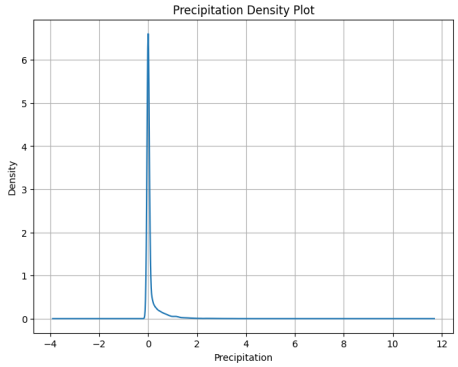
Wykres skrzypcowy opadów pozwala na zobrazowanie rozkładu poziomu opadów w poszczególnych miesiącach lub okresach czasu. Linia wewnątrz pudełka reprezentuje medianę, a rozpiętość pudełka pokazuje kwartyle danych.



Wykres skrzypcowy opadów

Wykres rozrzutu opadów

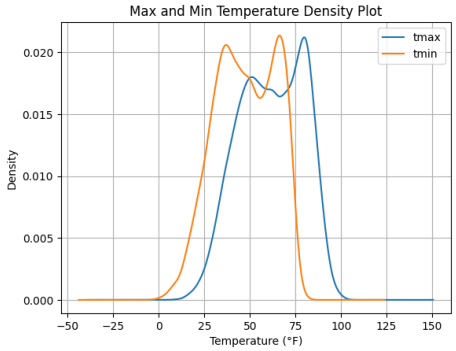
Wykres rozrzutu opadów prezentuje rozkład danych opadów w obrębie badanego okresu czasu. Pozwala to na analizę zmienności oraz ekstremalnych wartości opadów.



Wykres rozrzutu opadów

Wykres rozrzutu temperatury

Na wykresie rozrzutu temperatury przedstawione są gęstości rozkładu prawdopodobieństwa dla temperatur maksymalnych (tmax) i minimalnych (tmin). Ten rodzaj wykresu pozwala nam zobaczyć, jak często występują różne wartości temperatury i jak są one rozłożone wokół średniej. Obszary o większej gęstości sugerują większą koncentrację danych wokół określonych wartości, podczas gdy obszary o mniejszej gęstości sugerują większą zmienność lub rozproszenie danych.



Wykres rozrzutu temperatury

Dane związane ze śniegiem mają niską korelację względem pozostałych danych, więc ich wizualizację pominę.

4 Modele badawcze

W tej sekcji zostaną przedstawione różne modele, które będą używane do tworzenia predykcji temperatury. Każdy z modeli będzie analizowany pod kątem skuteczności na podstawie średniego błędu bezwzględnego (MAE), co pozwoli na porównanie ich efektywności.

4.1 Ridge Regression

Ridge Regression to metoda estymacji współczynników modeli regresji wielokrotnej w scenariuszach, w których zmienne niezależne są silnie skorelowane. . W naszym badaniu Ridge Regression została wykorzystana do przewidywania maksymalnej temperatury (tmax) na podstawie różnych parametrów pogodowych.

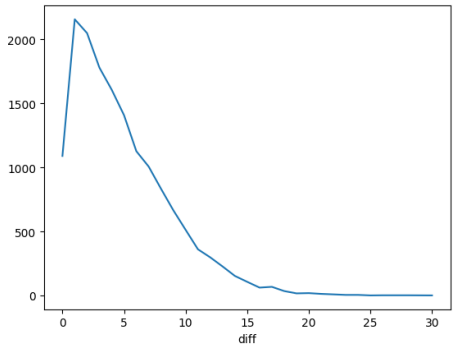
Przed dodaniem dodatkowych predyktorów, model Ridge Regression osiągnął średni błąd bezwzględny (MAE) na poziomie **5.1408364609200285**. Następnie, aby zwiększyć skuteczność modelu, zastosowaliśmy następujące dodatkowe predyktory:

- **Różnice procentowe (pct_diff)**: Obliczone jako różnica procentowa między wartościami tmax, tmin, a ich średnimi okresowymi, co pozwala na uwzględnienie zmienności temperatury w czasie.
- **Przesunięcie sekwencyjne (shift)**: Wykorzystane do stworzenia przesuniętej o jedną wartości temperatury, co umożliwia modelowi prognozowanie na podstawie poprzednich obserwacji.
- **Średnie okresowe (expanding_mean)**: Obliczone jako średnia ruchoma dla tmax, tmin oraz opadów (prcp) w określonych okresach czasu, co pozwala na uwzględnienie trendów długoterminowych w danych.

Po uwzględnieniu tych dodatkowych predyktorów, skuteczność modelu wzrosła, a średni błąd bezwzględny (MAE) spadł do poziomu **4.82072764061877**. Jak widać, dodane predyktory znacząco poprawiły jakość predykcji modelu, co pozwoliło na bardziej precyzyjne prognozowanie temperatury. Te dodatkowe predyktory zostaną również wykorzystane w analizie pozostałych modeli w celu oceny ich skuteczności.

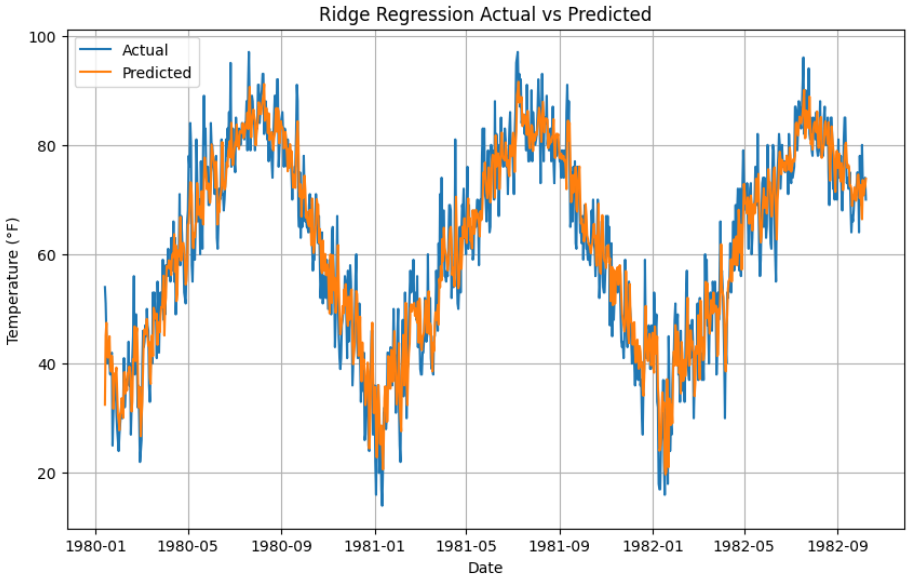
Tabela wyników prognoz od najgorszych do najlepszych

Data	Temperatura	Prognoza	Różnica
1990-03-12	85.0	55.057358	29.942642
2007-03-26	78.0	49.617807	28.382193
1998-03-26	80.0	51.670938	28.329062
⋮	⋮	⋮	⋮
2002-06-20	77.0	76.999629	0.000371
2006-03-23	52.0	52.000369	0.000369
1995-04-18	61.0	60.999997	0.000003



Wyniki ridge regression

Poniższy wykres pomoże nam zwizualizować jak model ridge regression pokrywa się z faktycznymi danymi.

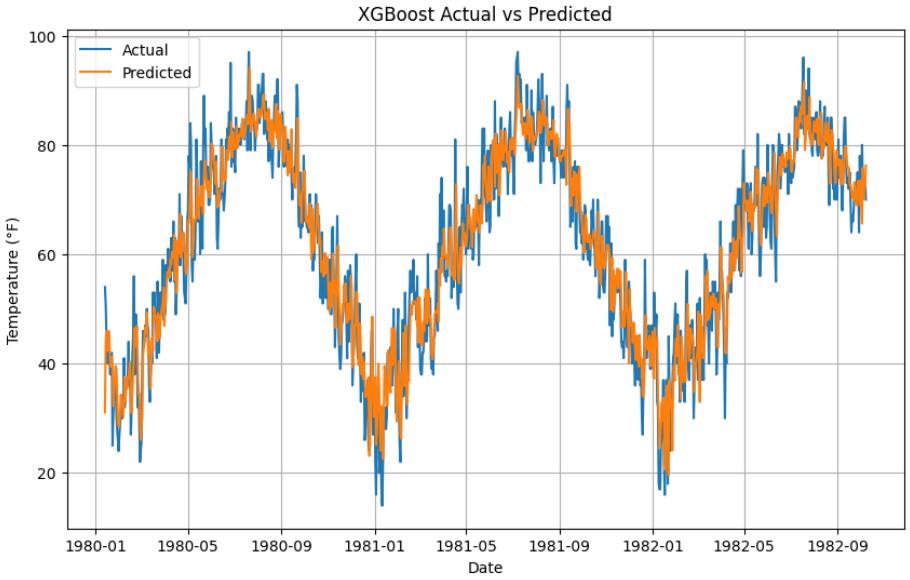


Wyniki ridge regression a faktyczne dane (1000 dni)

4.2 XGBoost

XGBoost (eXtreme Gradient Boosting) to wyjątkowo wydajna i skalowalna implementacja algorytmu gradientowego wzmocnienia. XGBoost oferuje także możliwość doboru najlepszych parametrów poprzez przeszukiwanie siatki (Grid Search), co pozwala na znalezienie optymalnych wartości hiperparametrów, takich jak liczba estymatorów (n_estimators), maksymalna głębokość drzewa (max_depth) i współczynnik uczenia (learning_rate). Wyniki dla tego modelu okazały się być lepsze niż dla Ridge Regression, ponieważ wskaźnik MAE wynosi: **4.788144296776628**.

DATE	actual	predicted	diff
$1980-01-13$	54.0	35.269497	18.730503
$1980-01-14$	51.0	45.255039	5.744961
$1980-01-15$	45.0	46.239704	1.239704
⋮	⋮	⋮	⋮
$2022-10-19$	61.0	59.270218	1.729782
$2022-10-20$	64.0	62.729061	1.270939
$2022-10-21$	64.0	64.562302	0.562302



Wyniki xgboost regression a faktyczne dane (1000 dni)

4.3 Sieci neuronowe

W tej sekcji wykorzystujemy modele sieci neuronowych do przewidywania temperatury maksymalnej. Pierwszym krokiem jest zbudowanie modelu. Model sieci neuronowej składa się z trzech warstw Dense, z których każda ma 64 neurony. Funkcją aktywacji dla warstw ukrytych jest ReLU, natomiast dla ostatniej warstwy wyjściowej nie ma funkcji aktywacji, ponieważ przewidujemy wartość liczbową.

Następnie przeprowadzamy walidację krzyżową w celu wyboru optymalnych parametrów modelu. Wykorzystujemy różne regularyzatory, takie jak regularyzacja L1 i L2, aby zapobiec nadmiernemu dopasowaniu i zwiększyć ogólną zdolność uogólniania modelu.

- **Regularyzacja L1**: Dodanie regularyzacji L1 do warstwy Dense pomaga w redukcji wartości wag, co prowadzi do rzadszego modelu. Wynik mean absolute error (MAE) po zastosowaniu regularyzacji L1 wynosi 5.03247799-15335025.

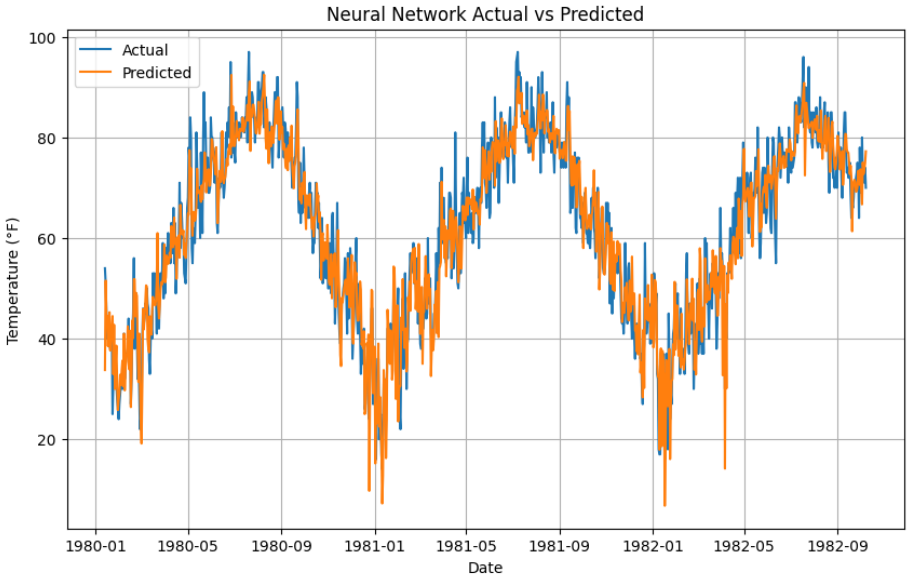
- **Regularyzacja L2**: Dodanie regularyzacji L2 do warstwy Dense pomaga w ograniczeniu wartości wag poprzez karanie dużych wartości wag. Wynik MAE po zastosowaniu regularyzacji L2 wynosi 4.988881605226816.

Jak widać, zastosowanie regularyzatorów L1 i L2 przyczyniło się do poprawy skuteczności modelu w porównaniu z modelem podstawowym, którego wynik

MAE wynosił 5.086059660045666. Wyniki analizy zostały przedstawione w tabeli poniżej.

DATE	actual	predicted	diff
1980-01-13	54.0	33.766613	20.233387
1980-01-14	51.0	51.532780	−0.532780
1980-01-15	45.0	44.200668	0.799332
⋮	⋮	⋮	⋮
2022-10-19	61.0	59.402153	1.597847
2022-10-20	64.0	64.771454	−0.771454
2022-10-21	64.0	66.644608	−2.644608

Wizualizacja wyników:



Wyniki sieci neuronowej a faktyczne dane (1000 dni)

4.4 Random Forest

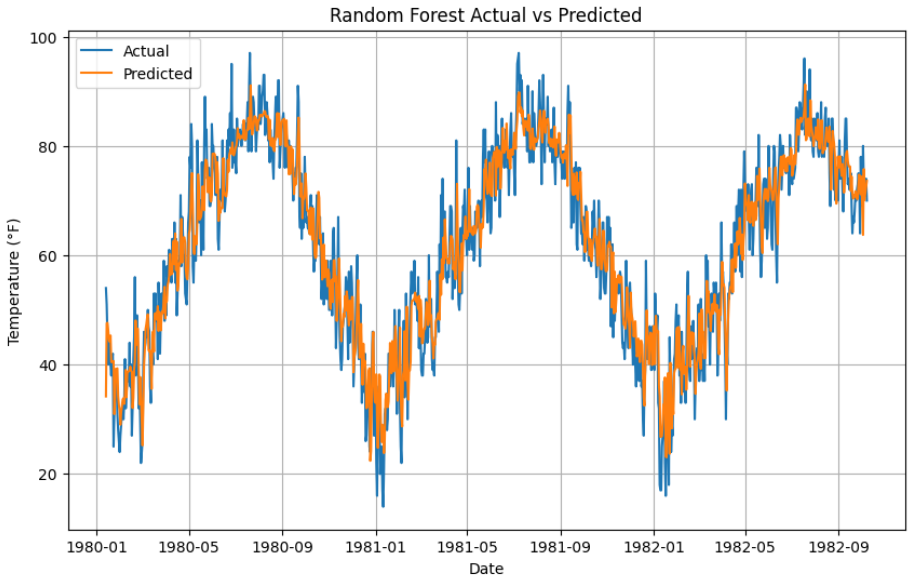
W tej sekcji używamy modelu Random Forest do prognozowania temperatury maksymalnej. Random Forest jest techniką zespołową, która wykorzystuje wiele drzew decyzyjnych do przewidywania wartości. Każde drzewo decyzyjne jest trenowane na losowych podzbiorach danych treningowych, a następnie głosowanie lub średnia ze wszystkich drzew jest używana do przewidywania wartości.

Pierwszym krokiem jest zbudowanie modelu Random Forest. Określamy liczbę drzew (n_estimators = 1000) oraz maksymalną głębokość drzewa (max_depth = 5).

Następnie przeprowadzamy walidację krzyżową w celu znalezienia optymalnych parametrów modelu. Wyniki przedstawione są poniżej.

DATE	actual	predicted	diff
1980-01-13	54.0	34.181652	19.818348
1980-01-14	51.0	47.548408	3.451592
1980-01-15	45.0	47.636482	2.636482
⋮	⋮	⋮	⋮
2022-10-19	61.0	58.383002	2.616998
2022-10-20	64.0	62.174943	1.825057
2022-10-21	64.0	63.142791	0.857209

- Wynik mean absolute error (MAE) dla modelu Random Forest wynosi 4.8779432797332225.



Wyniki Random Forest a faktyczne dane (1000 dni)

4.4.1 Algorytm Naive Bayes

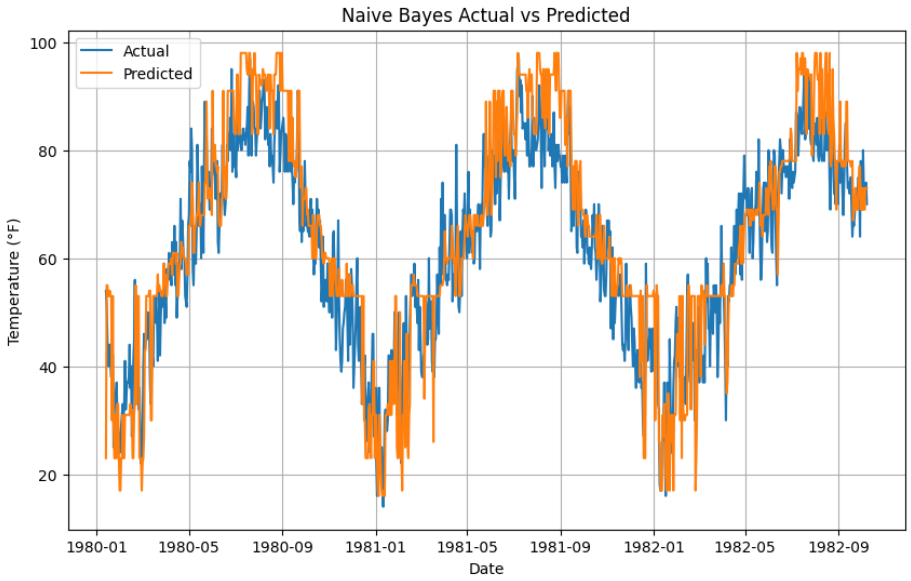
Algorytm Naive Bayes to popularna metoda klasyfikacji oparta na teoretycznych podstawach statystycznych, zwłaszcza na twierdzeniu Bayesa.

Naive Bayes zakłada niezależność między cechami (stąd "naive" w nazwie), co oznacza, że prawdopodobieństwo wystąpienia danej kombinacji cech jest iloczynem prawdopodobieństw wystąpienia każdej z cech indywidualnie. Pomimo tego uproszczenia, algorytm ten może być bardzo skuteczny, szczególnie w przypadku dużej liczby cech.

Stosujemy nasz model do danych i oceniamy jego skuteczność. Poniżej przedstawiamy wyniki dla naszego zbioru danych:

- Średni błąd bezwzględny (MAE): 7.810791781348012

DATE	actual	predicted	diff
1980-01-13	54.0	23.0	31.0
1980-01-14	51.0	55.0	4.0
1980-01-15	45.0	55.0	10.0
⋮	⋮	⋮	⋮
2022-10-19	61.0	64.0	3.0
2022-10-20	64.0	63.0	1.0
2022-10-21	64.0	63.0	1.0



Wyniki Naive Bayess a faktyczne dane (1000 dni)

Jak widać, ten model okazał się być niezbyt skuteczny.

4.5 Algorytm K-Nearest Neighbors (KNN)

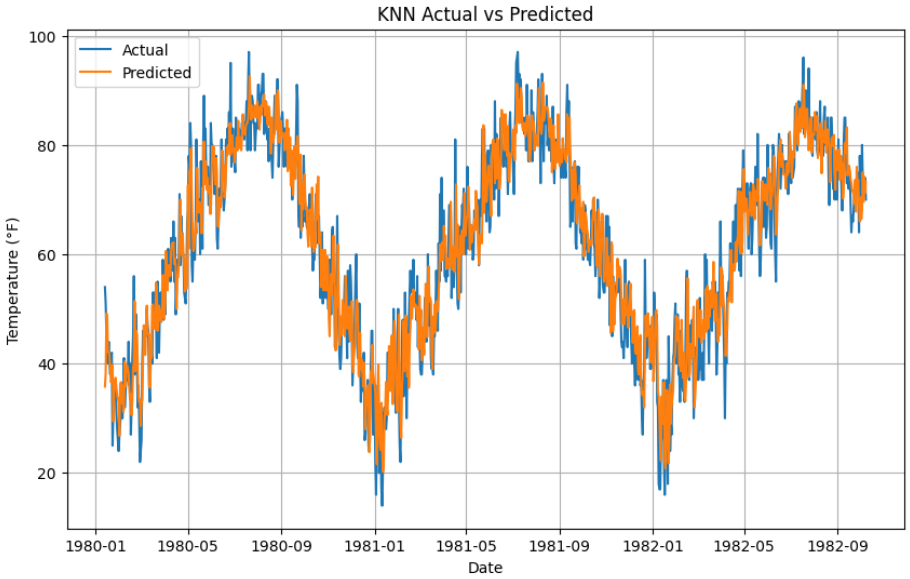
Algorytm K-Nearest Neighbors (KNN) to metoda uczenia się maszynowego wykorzystywana zarówno w zadaniach klasyfikacji, jak i regresji. Kluczową cechą KNN jest to, że nie wymaga uczenia się modelu na podstawie danych treningowych. Zamiast tego, dla nowej obserwacji, algorytm znajduje K najbliższych obserwacji w zbiorze treningowym i przewiduje wartość na podstawie ich średniej (dla regresji) lub najczęstszej klasy (dla klasyfikacji).

W naszym modelu wybraliśmy wartość parametru `n_neighbors=5`, co oznacza, że algorytm będzie brał pod uwagę 5 najbliższych sąsiadów.

Aby ocenić skuteczność modelu, przeprowadziliśmy testowanie wsteczne na naszym zbiorze danych pogodowych. Wynik mean absolute error (MAE) dla tego modelu wynosi 5.198783844332075.

Poniżej znajduje się tabela przedstawiająca rzeczywiste wartości, wartości przewidziane przez model KNN oraz różnicę między nimi:

DATE	actual	predicted	diff
$1980-01-13$	54.0	35.8	18.2
$1980-01-14$	51.0	38.2	12.8
$1980-01-15$	45.0	49.2	4.2
⋮	⋮	⋮	⋮
$2022-10-19$	61.0	58.2	2.8
$2022-10-20$	64.0	59.4	4.6
$2022-10-21$	64.0	60.8	3.2



Wyniki KNN a faktyczne dane (1000 dni)

4.6 Lasso Regression

Regresja Lasso (Least Absolute Shrinkage and Selection Operator) to technika regresji liniowej wykorzystująca regularyzację L1. Regularyzacja L1 pomaga w redukcji nadmiernych współczynników modelu poprzez dodanie kary do funkcji

kosztu proporcjonalnej do wartości bezwzględnej współczynników. Wartość kary jest kontrolowana przez parametr α, który jest określany przez użytkownika.

W naszym przypadku testowaliśmy różne wartości parametru α dla modelu Lasso.

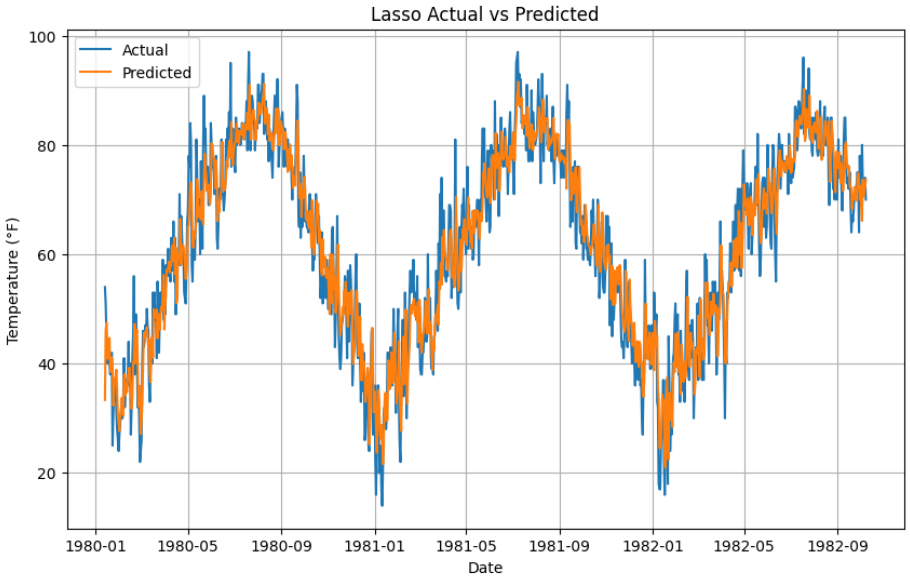
Parametr α jest istotny dla regularyzacji modelu. Im większa wartość α, tym mocniejsza regularyzacja, co prowadzi do bardziej ograniczonych współczynników i prostszych modeli. Natomiast mniejsza wartość α pozwala modelowi dopasować się bardziej swobodnie do danych treningowych, ale może prowadzić do nadmiernego dopasowania.

Poniżej znajdują się wyniki mean absolute error (MAE) dla różnych wartości parametru α:

Wartość α	**MAE**
0.1	4.824728030340125
0.01	4.817524660398157
0.2	4.830461765027311
0.002	4.813694464424817

Poniżej znajduje się tabela przedstawiająca rzeczywiste wartości, wartości przewidziane przez model Lasso oraz różnicę między nimi dla standardowego modelu Lasso:

DATE	actual	predicted	diff
1980-01-13	54.0	33.334492	20.665508
1980-01-14	51.0	46.011756	4.988244
1980-01-15	45.0	47.491400	2.491400
⋮	⋮	⋮	⋮
2022-10-20	64.0	62.246059	1.753941
2022-10-21	64.0	63.093437	0.906563



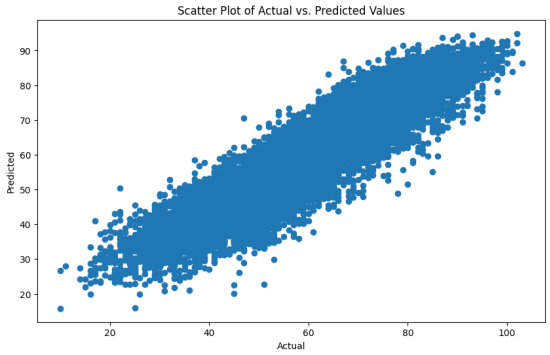
Wyniki Lasso a faktyczne dane (1000 dni)

5 Podsumowanie

Po przeprowadzeniu analizy skuteczności różnych modeli regresji w przewidywaniu pogody, najbardziej skutecznym okazał się model XGBoost, osiągając mean absolute error (MAE) na poziomie 4.788. Jednakże, pomimo wysokiej skuteczności modelu XGBoost, warto zauważyć, że prognozowanie pogody jest zadaniem trudnym ze względu na jej zmienność i liczne czynniki wpływające na warunki atmosferyczne.

Każdy z testowanych modeli miał swoje zalety i ograniczenia, a wybór odpowiedniego modelu zależy od specyfiki problemu oraz dostępnych danych. Zbadajmy trochę dokładniej model XGBoost.

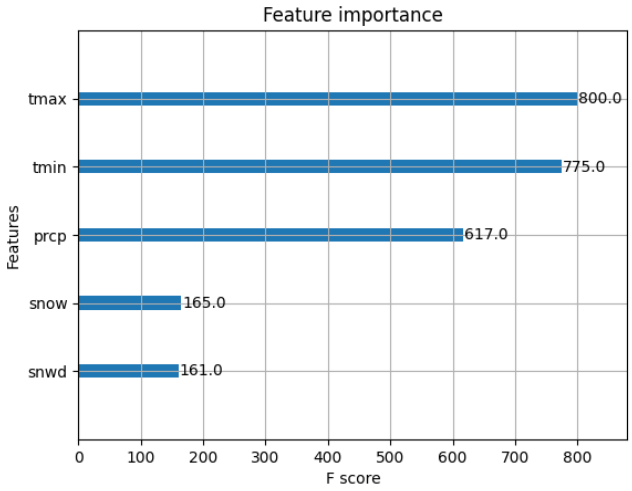
5.1 Rozrzut danych dla XGBoost



Wyniki rozrzutu

Jak widać dane skupiają się wokół środka. Jest to wynik oczekiwany i dowodzi skuteczności naszego modelu.

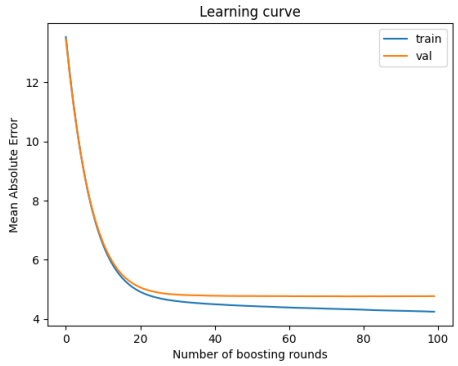
5.2 Najbardziej istotne współczynniki



Istotność współczynników

Dla modelu XGBoost najbardziej istotne okazały się kolumny **tmax** oraz **tmin**.

5.3 Learning curve



Learning curve

Krzywe uczenia w przypadku modeli opartych na boosting, takich jak XGBoost, są wykresami, które pokazują zmianę błędu (np. błąd średniokwadratowy lub błąd średniowy bezwzględny) na zbiorze treningowym i zbiorze walidacyjnym w zależności od liczby rund boostingowych.

References

[1] URL: https://www.youtube.com/watch?v=aLOQD66Sj0g

[2] URL: https://xgboost.readthedocs.io/en/stable/install.html

[3] URL: https://medium.com/@varun.tyagi83/a-deep-dive-into-building-a-weather-prediction-model-using-neural-networks-5d0fe7fd149f

[4] URL: https://www.researchgate.net/publication/352110820$_{W}eather_{p}rediction_{u}sing_{r}andom_{f}orest_{m}achine$

[5] URL: https://medium.com/@varun.tyagi83/a-deep-dive-into-building-a-weather-prediction-model-using-neural-networks-5d0fe7fd149f

[6] URL: https://towardsdatascience.com/weather-forecasting-with-machine-learning-using-python-55e90c346647