Straty alianckiej floty handlowej w czasie II Wojny Światowej

Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji

Autor:
Marek Kędzia

Numer Grupy:
K01-21e

Spis treści

1 Słowo od autora

Podczas II Wojny Światowej, niemieckie okręty podwodne, znane jako U-Booty, stały się przerażającym symbolem potęgi militarnej Trzeciej Rzeszy. Te niesłychanie skuteczne maszyny bojowe, przez lata konfliktu, zdołały zatopić ponad trzy tysiące alianckich statków handlowych, powodując znaczne straty zarówno w ludziach, ale przede wszystkim w sprzęcie, który zamiast wziąć udział w walkach w Europie, spoczął na dnie oceanu.

Przy użyciu zachowanych danych z tamtego okresu, mamy możliwość pogłębienia naszego zrozumienia strategii stosowanej przez U-Booty, jak również próby odkrycia ukrytych wzorców i charakterystyk dotyczących zatopionych jednostek. W erze technologicznej, kiedy narzędzia analizy danych stają się coraz bardziej zaawansowane, możliwe staje się dokładne badanie i interpretacja historycznych danych w nowym świetle.

Celem niniejszej pracy jest zbadanie konkretnego pytania - "Czy na podstawie zgromadzonych danych jesteśmy w stanie określić narodowość zatopionego statku?". Odpowiedź na to pytanie nie tylko dostarczy cennej wiedzy historycznej, ale także umożliwi zastosowanie i ocenę nowoczesnych technik analizy danych i uczenia maszynowego w praktyce.

2 Wprowadzenie

Zebrane dane pochodzą ze strony uboat.net i zawierają dane 3439 statków z czego prawie wszystkie nadawały się do więszkości opracowań, jednak w przypadku wykresów dotyczących lokalizacji zatopień, do analizy nadawało się 2770 rekordów.

W mojej pracy aby znaleźć odpowiedź na postawione na początku pytanie, skupiłem się na dwóch kluczowych aspektach: wizualizacji danych oraz modelowaniu predykcyjnym.

Na samym początku przeprowadzałem tzw. eksploracyjną analizę danych, przy użyciu takich pakietów do języka Python jak Pandas do manipulacji danymi oraz Matplotlib i Seaborn do generowania wykresów.

Następnie, przystąpiłem do budowy modeli predykcyjnych. Do tego celu wykorzystałem dwie popularne metody uczenia maszynowego: Support Vector Machine (SVM) oraz bardziej zaawansowany XGBoost.

SVM to algorytm, który można stosować zarówno do klasyfikacji, jak i regresji. Zasada działania SVM polega na wyznaczeniu hiperpłaszczyzny, która najlepiej oddziela klasy w danych.

Drugim wykorzystanym modelem był XGBoost - zaawansowany algorytm, który stosuje technikę tzw. gradient boosting.

Wybór tych dwóch modeli pozwolił mi na porównanie ich efektywności i wybór tego, który najlepiej radzi sobie z zadaniem przewidywania narodowości zatopionego statku na podstawie dostępnych danych.

3 Dane

Wszystkie dane wykorzystane w eksploracji znajdują się w załączniku o nazwie "data.zip".

Battles.csv zawiera dane o 83 największych bitwach II Wojny Światowej. Nałożenie dat tych potyczek pozwoliło odkryć w rozkładzie częstotliwości charakterystyki oraz wzorce. Ciężko jest wyróżnić jedno kryterium "wielkości"danej bitwy, jednak w zestawieniu zebrano informacje o tych potyczkach, w których brało udział najwięcej żołnierzy. Taki wybór pozwola na łatwiejszą obserwację wpływu potyczek na liczbę zatopień statków z zaopatrzeniem.

W pliku invalid.merchants.data.csv zapisane zostały uszkodzone dane statków, które ze względu na brakujące wartości nie mogły zostać wykorzystane w badaniu. W przypadku większości statków brakowało lokalizacji zatopienia. Nie jest to informacja niemożliwa do znalezienia w dzisiejszych czasach jednak bardzo ciężka w automatyzacji, a powtórzenie tej czynności dla prawie siedmiuset statków mogłoby okazać się bardziej czasochłonne niż przygotowanie całej reszty raportu.

Ostatni plik z danymi, o nazwie merchants.data.csv zawiera wszystkie statki, które dzięki kompletowi danych mogły wziąć udział w badaniu. Łączna liczba wierszy z danymi w tym pliku wynosi 2770, co stanowi wystarczającą bazę do wytrenowania wybranych na początku projektu modeli.

Każdy wiersz składa się z ośmiu kolumn które zawierają informacje o tonażu, narodowości, konwoju do którego należał, lokalizacji, dacie zatopienia oraz podstawowe dane o U-boocie który go zatopił- jego nazwę oraz imię i nazwisko dowódcy.

Przykładowo, dane dla statku "Bosnia"wyglądają nas†ępująco:

Uboat	U-47
Commander	Günther Prien
Ship name	Bosnia
Tonnage	2,407
Nationality	Great Britain
Convoy	-
Coordinates	('45.48', '-9.75')
Date	05.09.1939

W folderze "data"znajduje się także plik html "sinking.map", który po

uruchomieniu w dowolnej przeglądarce pokaże interaktywną mapę z nałożonymi miejscami wszystkich zatopień.

Programy, napisane w języku Python, odpowiedzialne za wygenerowanie wykresów dla odpowiednich kolumn umieszczone zostały w folderze "code/analisis"a przyjęta konwencja zakłada nazwy zgodnie ze schematem:

$$< nazwa.kolumny > .analisis.py \tag{1}$$

z ewentualnym opisem także w przy użyciu notacji kropkowej.

3.1 Commander

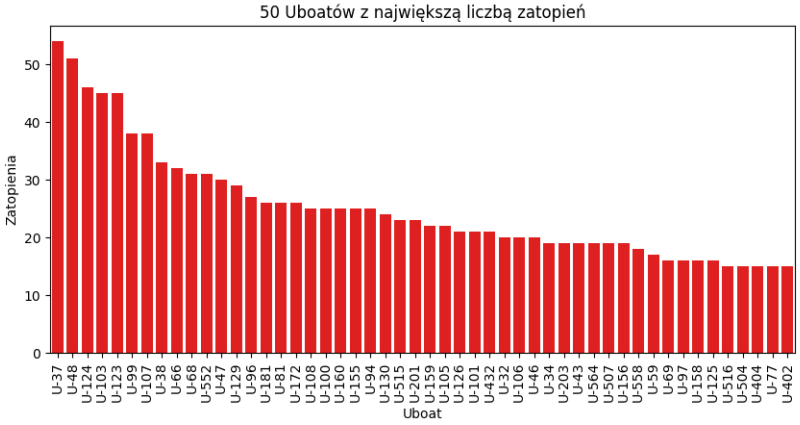
W kolumnie "Commander"zapisane zostały dane oficera, który dowodził okrętem podwodnym w momencie zaliczenia trafienia. Spośród 501 wszystkich podwodniaków, którzy posiadali jakiekolwiek zatopienie tylko 164 osiągnęło ich 5 lub więcej. Do najskuteczniejszych należą:

1. Otto Kretschmer i Wolfgang Lüth, obaj 44 potwierdzone zatopienia
2. Joachim Schepke, 35 potwierdzonych zatopień
3. Erich Topp, 34 potwierdzone zatopienia

3.2 Uboat

Kolumna "Uboat"nawiązuje do okrętu podwodnego, który zaliczył zatopienie. Niemcy przyjęli bardzo pragmatyczną regułę nazywania tych maszyn, ponieważ zamiast przypisywać maszynom imiona, były one po prostu numerowane z prefiksem "U-". Przydzielanie numerów nie zawsze było ściśle sekwencyjne, ponieważ niektóre okręty były zamówione i budowane jednocześnie w różnych stoczniach, a numer były przypisywany na etapie zamówienia, a nie ukończenia budowy.

Spośród wszystkich Uboatów szczególnie należy wyróżnić trzy jednostki: U-37 (53 zatopienia), U-48 (51) oraz U-124 (46). Poniższe zestawienie prezentuje 50 najskuteczniejszych niemieckich okrętów podwodnych.



3.3 Ship name

W tej kolumnie zapisane są nazwy statków, które zostały zatopione. Do najsłynniejszych możemy zaliczyć:

1. Athenia była brytyjskim statkiem pasażerskim, który został zatopiony przez niemiecki okręt podwodny U-30 3 września 1939 roku. Było to pierwsze zatopienie statku przez Niemców podczas II wojny światowej.

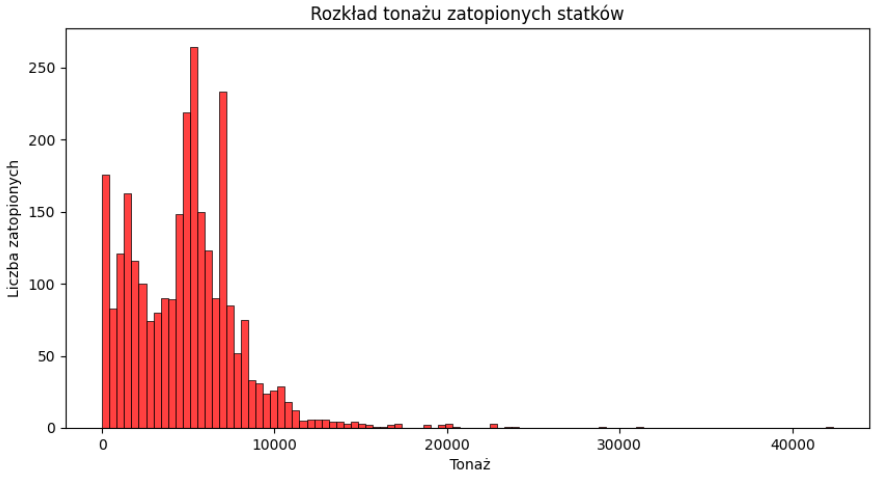
2. City of Benares to brytyjski statek pasażerski, który został zatopiony przez niemiecki okręt podwodny U-48 18 września 1940 roku. Na pokładzie znajdowało się wielu dzieci, które były ewakuowane z Wielkiej Brytanii do Kanady.

3. Robin Moor był amerykańskim statkiem handlowym, który został zatopiony przez niemiecki okręt podwodny U-69 21 maja 1941 roku. Zatopienie tego statku przed formalnym wejściem USA do wojny wywołało kontrowersje.

3.4 Tonnage

Sekcja "Tonnage"to liczba całkowita reprezentująca całkowity tonaż zatopionego statku.

Jak można z łatwością zauważyć zdecydowana większość ofiar uboatów to statki z tonażem większym niż 5000, czyli średniej wielkości jednostki.



Na tej podstawie, można by wysnuć fałszywą tezę, zakładającą, że większe statki były były narażone na ataki, jednak nie jest to do końca prawda. Oczywiście można zakładać, że większe statki były lepiej chronione, jednak różnica w liczności zatopionych okrętów wynika w największym stopniu ze składu konwojów. Jako przykład można przedstawić słynny konwój PQ-16, który składał się z 35 statków handlowych: 21 amerykańskich, 8 brytyjskich, 4 radzieckich, holenderskiego i panamskiego. Spośród tych jednostek aż 28 posiadało tonaż większy niż 5000 i mniejszy od 8000, co w dużej mierze tłumaczy powyższy rozkład.
Największą zatopioną jednostką jest "Empress of Britain". Brytyjski parowiec pasażerski o tonażu 42348 został trafiony 28 października 1940 roku przez U-32.

3.5 Convoy

W kolumnie o nazwie "Convoy"widnieją nazwy konwoju w skład którego wchodził statek w momencie zatonięcia. Ewentualny brak przynależności oznaczony jest znakiem ". Mapowanie nazw zostało załączone w folderze "data/mappings"pod nazwą "convoy.mapping.json".

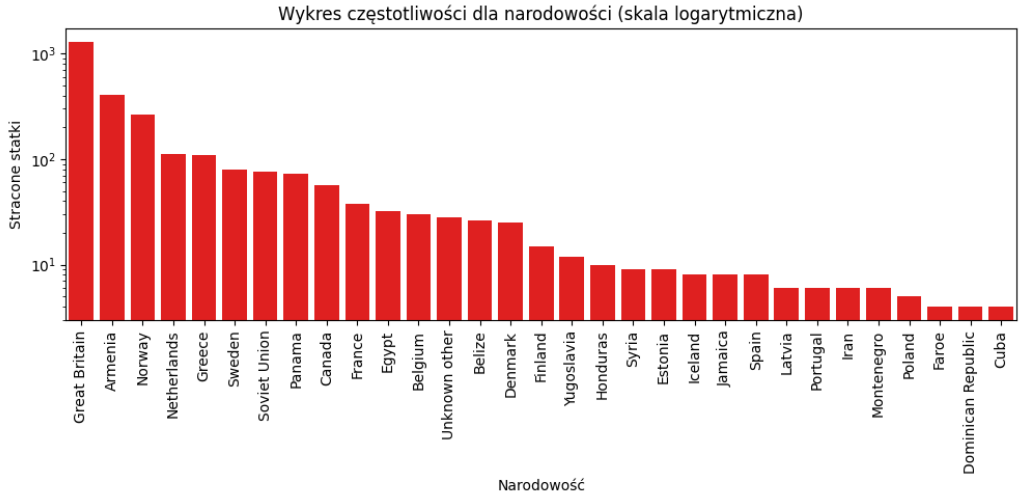
Taktyka konwojów wprowadzona w czasie II Wojny Światowej pozwalała na efektywną ochronę statków handlowych, które transportowały zaopatrzenie do walczącej Europy. Statki podróżujące samodzielnie były łatwym celem dla U-Boatów oraz samolotów jednak zgrupowanie ich razem pod eskortą

okrętów wojennych niwelowało przewagę przeciwnika, co z kolei pomagało w utrzymaniu niezbędnych linii zaopatrzenia.

W nazwie konwoju, pierwsze dwie litery kodu zazwyczaj wskazywały na jego trasę. Na przykład, konwoje "HX"płynęły z Halifaxu w Kanadzie do Liverpoolu w Anglii. Kolejne numery były numerami porządkowymi. Na przykład, konwój "HX-231"był 231. konwojem na tej trasie.

3.6 Nationality

Kolumna "Nationality"informuje o przynależności do jednego z 47 alianckich krajów z listy dostępnej w folderze "data/mappings"pod nazwą "nationality.mapping.json". Poniższy wykres prezentuje statystyki tych państw, które w wyniku ataków Uboatów utraciły więcej niż jeden statek handlowy.



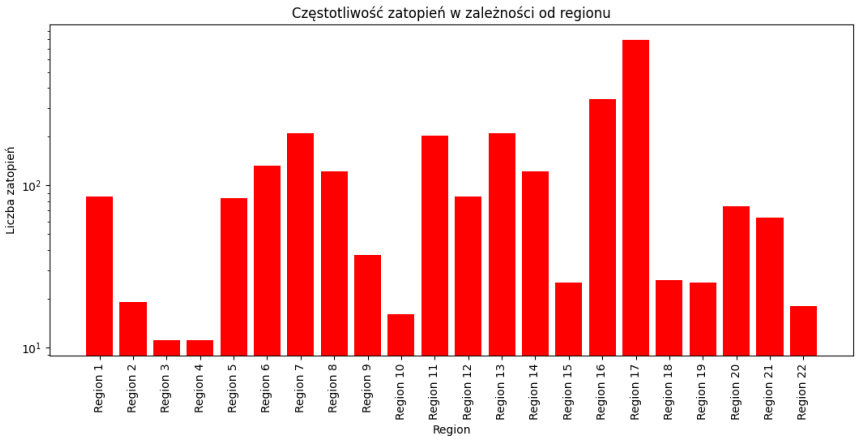
Łatwo zauważyć, że największe straty poniosła flota Wielkiej Brytani oraz Stanów Zjednoczonych, co nie powinno dziwić, ponieważ były to dwa państwa najbardziej zaangażowane w konflikt ze wszystkich aliantów zachodnich. Zaskoczeniem mogą być stosunkowo wysokie straty Norwegii, która przecież znajdowała się pod niemiecką okupacją przez większość wojny. Wskazuje to na bardzo znaczący i niedoceniony udział jednostek pływających pod norweską banderą w dostawach zaopatrzenia dla walczących żołnierzy.

3.7 Coordinates

Kolumna "Coordinates"zawiera informacje dotyczące współrzędnych geograficznych na których został zatopiony okręt.

Analiza tych danych może dostarczyć informacji na temat obszarów, w których konwoje były najbardziej narażone na ataki, czyli miejsca największej aktywności U-boatów.

Z wykresu na następnej stronie możemy odczytać jak rozkładały zatopienia pod względem lokalizacji. Łatwo zauważyć, że do największej liczby zatopień doszło w regionie 17, czyli u wybrzeży Wielkiej Brytanii co wskazuje na dużą aktywność Uboatów w tym kwadracie, co mogło być spowodowane szczególnie przyjaznymi warunkami do podejmowania ataku ale także bliskością portów Kriegsmarine w Niemczech oraz Francji, które zapewniały schronienie oraz zaopatrzenie.



Region			
1	(-40.58,8.88)-(-23.68,44.34)	12	(27.04,-62.04)-(43.94,-26.58)
2	(-23.68,-62.04)-(-6.77,-26.58)	13	(27.04,-26.58)-(43.94,8.88)
3	(-23.68,-26.58)-(-6.77,8.88)	14	(27.04,8.88)-(43.94,44.34)
4	(-23.68,8.88)-(-6.77,44.34)	15	(43.94,-97.50)-(60.85,-62.04)
5	(-6.77,-62.04)-(10.13,-26.58)	16	(43.94,-62.04)-(60.85,-26.58)
6	(-6.77,-26.58)-(10.13,8.88)	17	(43.94,-26.58)-(60.85,8.88)
7	(10.13,-97.50)-(27.04,-62.04)	18	(43.94,8.88)-(60.85,44.34)
8	(10.13,-62.04)-(27.04,-26.58)	19	(60.85,-62.04)-(77.75,-26.58)
9	(10.13,-26.58)-(27.04,8.88)	20	(60.85,-26.58)-(77.75,8.88)
10	(10.13,44.34)-(27.04,79.80)	21	(60.85,8.88)-(77.75,44.34)
11	(27.04,-97.50)-(43.94,-62.04)	22	(60.85,44.34)-(77.75,79.80)

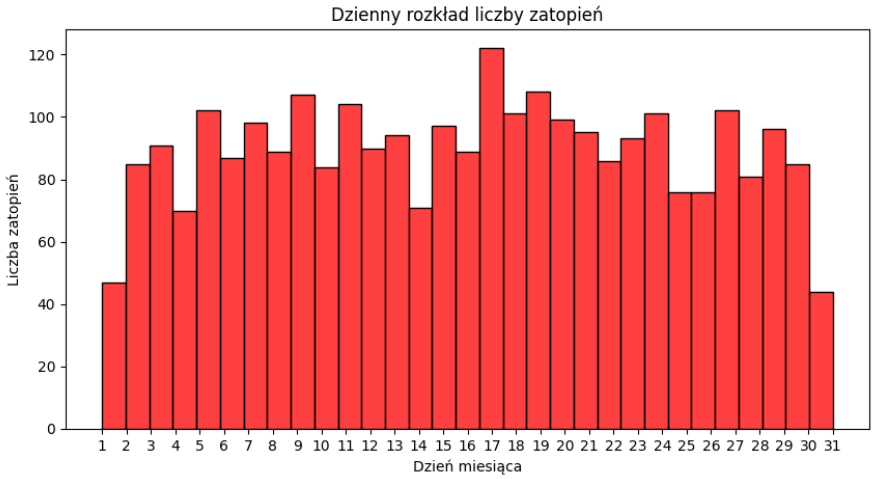
4 Eksploracyjna Analiza Danych

Podczas eksploracji danych zatopionych statków, jednym z najważniejszych i najciekawszych z czynników może być data zatopienia okrętu.

Jak wiadomo, na tą wartość wpływa bardzo wiele czynników, w tym przede wszystkim sytuacja na arenie międzynarodowej, wielkie bitwy oraz wprowadzanie nowych technologii i taktyk.

4.1 Dni

Poniższy wykres przedstawia przenalizowaną liczbę zatopień dla każdego dnia miesiąca.



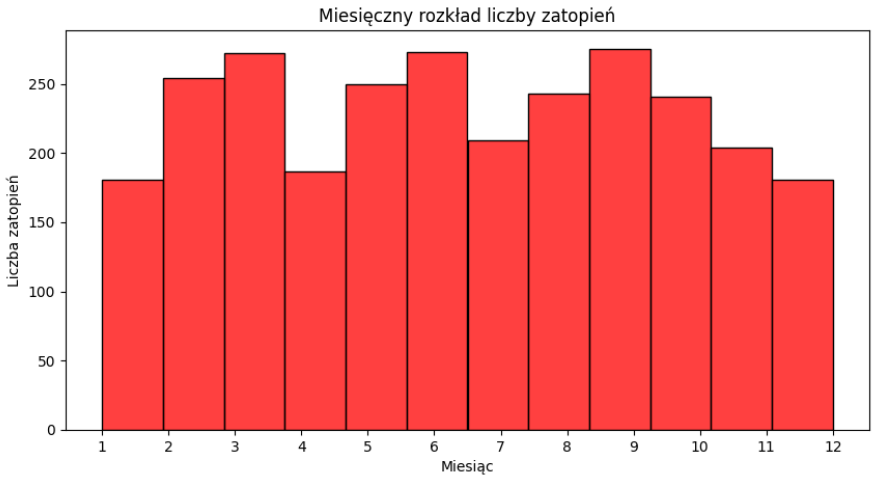
Z przyczyny liczności każdego z miesięcy, najmniejesza liczba zatopień dla 31 dnia miesiąca nie powinnna nikogo dziwić. Uwagę przykuwa także wynik dnia 17, jednak ciężko doszukiwać się w tym przypadku jakiekolwiek wzorca.

W analizie danych statystycznych, szczególnie gdy mamy do czynienia z dużą liczbą danych istnieje możliwość przypadkowych fluktuacji co w nomenklaturze jest często nazywane problemem "wielokrotnych porównań"lub "wielokrotnego testowania".

W przypadku danych o zatopieniach statków, jest 31 potencjalnych dni w miesiącu, które mogą zostać porównane. Przy tak wielu porównaniach, prawdopodobieństwo, że przynajmniej jedno z nich wykaże istotne różnice tylko przez przypadek, jest stosunkowo wysokie.

Ponadto nie ma żadnego logicznego powodu, dla którego 17. dzień miesiąca miałby być czarnym dniem dla alianckiej floty, dlatego najprawdopodobniej obserwowane różnice są po prostu wynikiem losowości.

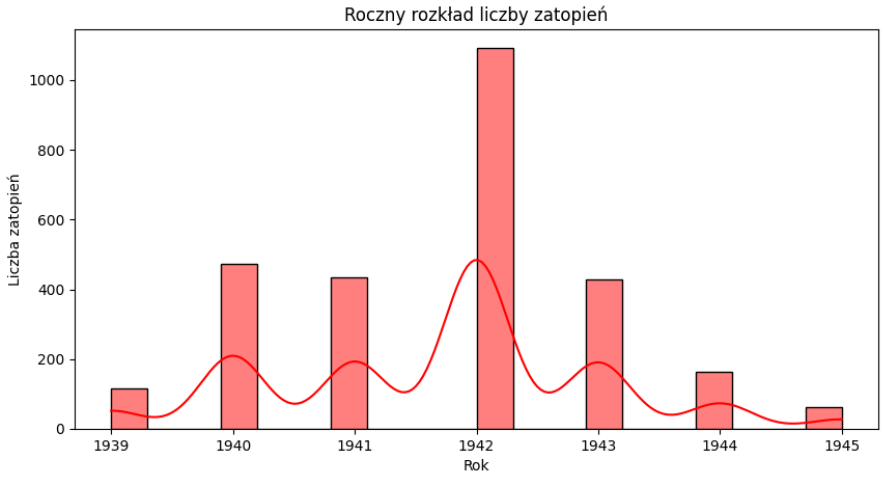
4.2 Miesiące



W przypadku analizy miesięcznej liczby zatopień statków natrafiamy na ciekawą zależność, którą można postarać się uzasadnić historycznymi faktami. Mianowicie, patrząc na wykres łatwo zauważyć, że liczba zatopień w każdym kwartale stopniowo rosła aż do ostatniego miesiąca kwartału, aby ponownie spaść do poziomu początkowego. Wyjątkiem jest ostatni kwartał roku, w którym liczba zatopień stopniowo spadała, jednak łatwo jest to wytłumaczyć ze względu na stopniowo pogarszające się warunki walki dla uboatów oraz zmniejszoną liczbę konwojów płynących przez Atlantyk.

Przyczyny stopniowo rosnącej liczby zatopień w każdym kwartale można doszukiwać się w strategii polowania na alianckie konwoje przez Kriegsmarine. Uboaty w ramach taktyki "Wilczych stad"wspólnie atakowały alianckie statki w jednym momencie w taki sposób aby zadać konwojowi jak największe straty. Okręty podwodne dysponują ograniczonymi zasobami, dlatego po udanym ataku musiały powracać do baz aby uzupełnić zaopatrzenie oraz dokonać napraw. Potwierdza to wzorzec, który może zakładać powrót Uboatów do baz po udanym ataku, przec co zmniejszała się liczba "podwodnych myśliwych"co bezpośrednio wpływa na liczbę zatopień.

4.3 Lata



Analiza roczna zatopień statków handlowych rzuca światło na zmianę dynamiki działań wojennych na morzach i oceanach. Rozkład liczby zatopień wyraźnie wskazuje zwrotne momenty II Wojny Światowej, dzięki czemu z łatwością możemy przeprowadzić analizę tego wykresu.

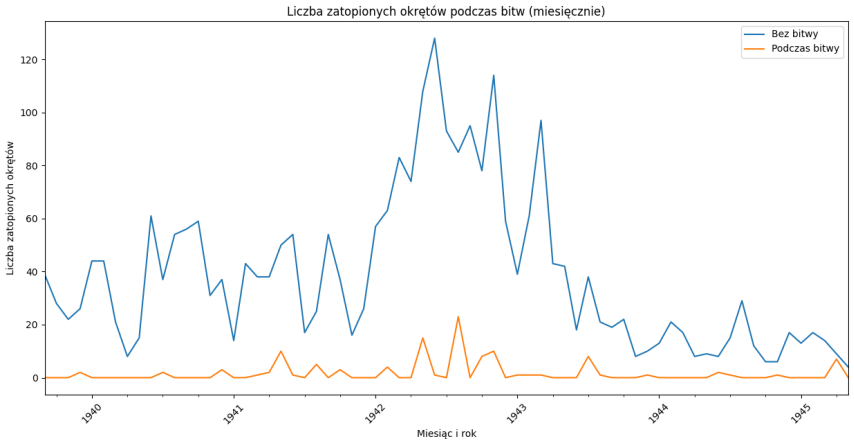
Najbardziej intensywny okres wojny na morzach przypada na rok 1942, co zbiega się z "bitwą o Atlantyk". Niemcy w tym czasie intensyfikowały swoje operacje podwodne, stawiając w obliczu zagrożenia znaczną część alianckich konwojów. W tym okresie zanotowano także największą liczbę zatopień. Był to okres, w którym III Rzesza wchodziła w przełomowy okres wojny. Szala zwycięstwa nie była przechylona w żadną stronę a rozwiązanie dynamicznej sytuacji w Afryce oraz przerwanie stagnacji na froncie wschodnim mogło przynieść jednej ze stron konfliktu zwycięstwo.

Od 1944 roku obserwujemy gwałtowny spadek liczby zatopień. Można to przypisać poprawie taktyk konwojowych Aliantów, lepszemu wyposażeniu i technologii zwalczania okrętów podwodnych, a także rosnącej przewadze alianckiej na morzach. W efekcie liczba zatopień zaczęła drastycznie spadać, aż do końca wojny w 1945 roku.

4.4 Liczba zatopień a największe bitwy II Wojny Światowej

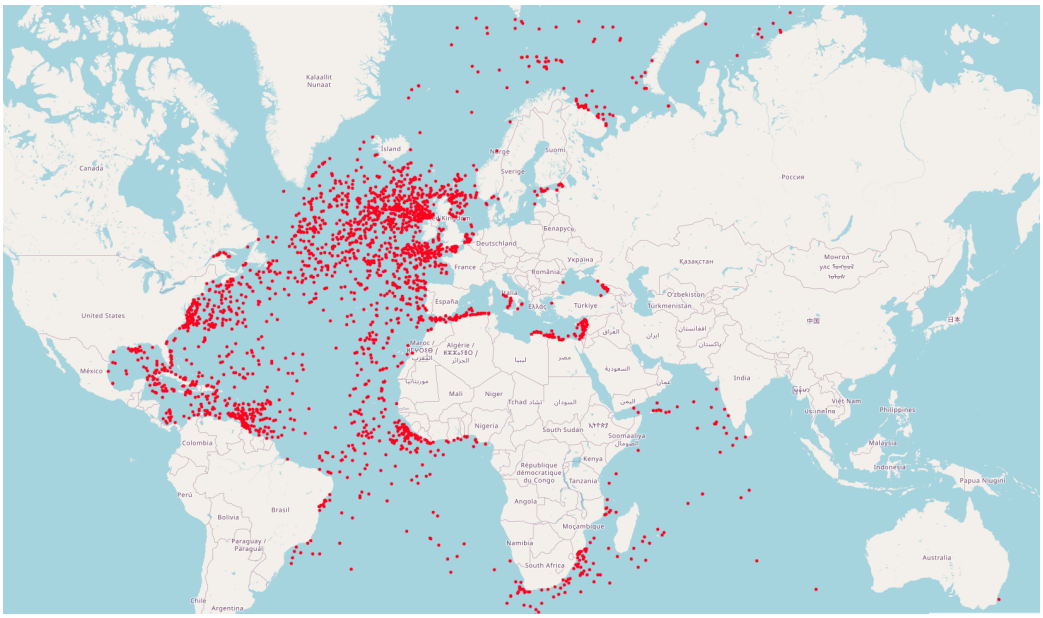
Po przeanalizowaniu dat największych bitew drugiej wojny światowej, można dojść do wniosku, że nie wpływały one znacząco na liczbę zatopień statków handlowych. Nieznaczne wzrosty liczby zatopień z końcówki roku 1942 nakładają się czasowe ze zintesyfikowaniem działań na Atlantyku, co w znacznym stopniu wyjaśnia chwilowe oderwania od średniej.

Wpływ na ten fakt może mieć specyfika działań wojennych. Największe ilości sprzętu oraz zaopatrzenia muszą zostać zgromadzone przez rozpoczęciem kampanii, w przeciwnym wypadku znacząco wpływałoby to na skuteczność walczących oddziałów. Zgromadzone zapasy paliwa oraz żywności pomagają w szybkim przerzucaniu jednostek wgłąb linii nieprzyjaciela, a zapasy amunicji i sprzętu pozwalają uzupełnić braki spowodowanymi ciężkimi walkami.



Analizując ten wykres można jednak z łatwością zauważyć momenty w którym znacząco zwiększał się wysiłek wojenny. Gwałtowny wzrost liczby zatopionych okrętów rozpoczyna się mniej więcej w grudniu 1941 roku, czyli w momencie dołączenia Stanów Zjednoczonych do wojny. Z kolei wyraźne spadki widać na koniec roku 42 oraz w połowie roku 43, czyli w momentach w których oczy III Rzeszy a także Kriegsmarine musiały być zwrócone w innym kierunku niż Atlantyk, ze względu na klęski na froncie wschodnim oraz afrykańskim.

4.5 Lokalizacje zatopień



Mapa z nałożonymi miejscami zatopień (pełna wersja znajduje się w folderze data pod nazwą sinking.map.html) ukazuje ogrom strat alianckiej floty handlowej. Praktycznie całe wybrzeża Wielkiej Brytanii i Francji pokryte są czerwonymi punktami. Główne szlaki handlowe, biegnące przez Atlantyk, będące najkrótszą drogą ze Stanów Zjednoczonych do Wielkiej Brytanii oraz przemysłowej części Związku Radzieckiego są łatwo zauważalne, ze względu na dużą liczbę straconych okrętów w tych rejonach. Stosunkowo bezpieczny był Ocean Indyjski, jednak najdalsze trafienie zdarzyło się nawet u wschodnich wybrzeży Australii. Pokazuje to zasięg ataków okrętów podwodnych Kriegsmarine, która mogła siać postrach praktycznie w każdej części globu.

5 Metodologia

5.1 Przygotowanie danych

Analizowane dane zostały najpierw poddane procesowi filtrowania i przygotowania. Format danych został dostosowany do wymagań biblioteki pandas. Lokalizacja zatopienia znaleziona została na podstawie załączonych na stronie uboat.net map, a narodowość kraju odczytana na podstawie zdjęcia bandery załączonego do każdego z rekordów.

Statki dla których brakowało danych w kolumnie innej niż "Coordinates"stanowiły ułamek procenta wszystkich rekordów (0,14%), dlatego zostały usunięte ze zbioru testowego. Rekordy z brakującą lokalizacją, w liczbie 669 zostały odzielone i zapisane w osobnym pliku oraz użyte jako dodatkowe dane to sprawdzenia poprawności podejmowania decyzji przez gotowy model, ponieważ koordynaty zatopienia nie wpływały znacząco na wyniki przewidywać. Następnie rekordy zostały podzielone na zbiór treningowy i testowy w standardowym stosunku 80:20, aby odpowiednio wytrenować model i umożliwić ocenę jego skuteczności.

5.2 Metodyka Modelowania

Celem projektu jest znalezienie odpowiedzi na pytanie czy dysponując takimi danymi jesteśmy w stanie rozwiązać problem klasyfikacji narodowości zatopionego statku. W tym celu wykorzystane zostały dwie popularne techniki klasyfikacyjne: Support Vector Machines (SVM) oraz Extreme Gradient Boosting (XGBoost).

Możemy sobie wyobrazić mało realną sytuację, w której superkomputer w podziemiach tajnego bunkra w Berlinie, próbuje przewidzieć narodowość zatopionego przez podwodniaków statku handlowego. Taka analiza mogłaby być szczególnie przydatna w przypadku chęci oszacowania strat zadanych konkretnej nacji oraz w celu oceny zaangażowania państwa w konflikt zbrojny.

5.2.1 Support Vector Machines (SVM)

Maszyny wektorów nośnych, znane jako SVM, to jedne z najpotężniejszych dostępnych algorytmów klasyfikacji. SVM zasadniczo tworzy hiperpłaszczyznę w przestrzeni wielowymiarowej, która najefektywniej oddziela różne klasy

danych. SVM wykorzystuje koncepcję 'marginesu', aby maksymalizować odległość między najbliższymi punktami (wektorami nośnymi) z różnych klas. W ten sposób algorytm stara się zminimalizować błąd generalizacji i uniknąć overfittingu.

5.2.2 Extreme Gradient Boosting (XGBoost)

XGBoost to zaawansowany algorytm uczenia maszynowego, który korzysta z techniki Gradient Boosting. Boosting polega na łączeniu wielu słabych predyktorów, tworząc model, który lepiej radzi sobie z predykcją. XGBoost wykorzystuje drzewa decyzyjne jako podstawowe predyktory, które są połączone w sposób, który minimalizuje funkcję straty. Algorytm ten jest niezwykle skuteczny w różnych zadaniach klasyfikacji i regresji, dając często lepsze wyniki od innych algorytmów uczenia maszynowego.

5.3 Ewaluacja modelu

Oba modele były oceniane na podstawie trzech kluczowych metryk: 'precision', 'recall' oraz 'f1-score'. 'Precision' odnosi się do ilości prawidłowo sklasyfikowanych pozytywnych instancji spośród wszystkich instancji sklasyfikowanych jako pozytywne. 'Recall' natomiast to ilość prawidłowo sklasyfikowanych pozytywnych instancji spośród wszystkich rzeczywistych instancji pozytywnych. 'F1-score' to średnia harmoniczna 'precision' i 'recall'. Ostatnim parametrem jest 'support', który wskazuje liczbę rzeczywistych przypadków w każdej klasie.

Dla modelu XGBoost, obliczono również obszar pod krzywą ROC (AUC). AUC jest metryką używaną w klasyfikacji binarnej, która mierzy zdolność modelu do odróżniania między klasami. Im wyższa wartość AUC, tym lepsze są zdolności klasyfikacyjne modelu.

Dla każdego z modeli wyliczona została także macierz pomyłek (confusion matrix), która pozwoliła na zrozumienie, jakie błędy popełniają modele i jakie klasy są dla nich najtrudniejsze do rozróżnienia.

6 Wyniki

6.1 Przedstawienie wyników modelu SVM

Generalnie, wyniki klasyfikacji są dość słabe. Opracowany model działa niezadowalająco, z dokładnością ogólną wynoszącą jedynie 41%. Wartości precyzji, recall i F1-score dla poszczególnych klas wskazują na niejednolite wyniki klasyfikacji, gdzie niektóre klasy są rozpoznawane zdecydowanie lepiej niż inne.

Poniższe tabele prezentują wyniki poprawności dla każdej klasy oraz skróconą macierz pomyłek, której pełna wersja znajduje się w załączonym folderze data, pod nazwą "confusion.matrix.svm.txt".

Class	Precision	Recall	F1-score
0	0.35	0.14	0.20
1	0.33	0.51	0.40
2	0.36	0.16	0.22
3	0.80	0.43	0.56
4	0.39	0.73	0.51
5	0.00	0.00	0.00
6	0.00	0.00	0.00
7	0.33	0.88	0.48
8	0.14	0.00	0.01
9	0.00	0.00	0.00
10	0.00	0.00	0.00
11	0.86	0.72	0.78
12	0.25	0.12	0.16
13	0.52	0.18	0.27
14	0.45	0.43	0.44
15	0.49	0.88	0.63
16	0.00	0.00	0.00
17	0.71	0.78	0.74
18	0.18	1.00	0.30
19	0.78	0.45	0.57
20	0.49	1.00	0.66
21	0.80	0.50	0.62

	0	1	2	3	4	...
0	35	37	2	3	22	...
1	4	138	22	0	15	...
2	1	0	41	0	57	...
3	8	8	2	110	13	...
4	0	0	0	0	183	...
...	...	...	...	...	...	...

Analizując dane, możemy wyciągnąć kilka wniosków na temat poszczególnych klas:

Klasy 5, 6, 9, 10 i 16 wydają się mieć najgorsze wyniki, z punktu widzenia precyzji, czułości i wyniku F1, które wszystkie wynoszą 0.00. Może to sugerować, że model nie był w stanie poprawnie przewidzieć żadnego z przypadków z tych klas, co jest bardzo niepokojące. Prawdopodobnie jest to spowodowane brakiem wystarczającej ilości danych dla tych klas.

Macierz pomyłek jest używana do wizualizacji wyników klasyfikacji. Na przekątnej macierzy (od lewej górnej do prawej dolnej) mamy liczby prawidłowo przewidzianych klas - im wyższa liczba, tym lepiej.

Na przykład, klasa 1 została prawidłowo sklasyfikowana 138 razy, co jest dość wysoką liczbą w porównaniu do innych klas..

Im więcej wartości znajduje się na przekątnej, a im mniej poza nią, tym lepiej działa model. W idealnym przypadku wszystkie wartości poza przekątną powinny wynosić zero, ale w praktyce jest to rzadkością, zwłaszcza w skomplikowanych problemach klasyfikacji.

Jak możemy zauważyć na przykładzie załączonego fragmentu macierzy oraz jej pełnej wersji, w przypadku modelu SVM bardzo często dokonywał on mylnej klasyfikacji, przez co przedstawione przez niego wyniki są raczej niskiej jakości.

Model z pewnością mógł zostać dopracowany aby wyniki były bardziej dokładne, jednak początkowy rezultat był przynajmniej zniechęcający do dalszych prac przy użyciu tej metody, dlatego już w początkowej fazie projektowania zdecydowałem się porzucić tą metodę na rzecz Extreme Gradient Boostingu.

6.2 Przedstawienie wyników modelu Extreme Gradient Boosting (XGBoost)

Poniższa tabela prezentuje rozkład precyzji, pełności, f1 oraz wsparcia dla poszczególnych klas. Warto zauważyć, że w celu zmaksymalizowania dokładności wyniku klasy o liczności mniejszej niż 3 (nacje które straciły mniej niż trzy statki w czasie całej II Wojny Światowej) zostały sklasyfikowane razem jako klasa 0. Upraszcza do działanie algorytmu, a prawdopodobieństwo, że jakieś małe państewko utraciło kolejny statek jest racze niskie.

Class	Precision	Recall	F1-score	Support
0	1.00	0.99	0.99	262
1	0.69	0.80	0.74	279
2	1.00	0.99	1.00	250
3	1.00	1.00	1.00	274
4	1.00	0.99	0.99	263
5	1.00	1.00	1.00	274
6	0.99	0.97	0.98	231
7	0.92	0.79	0.85	259
8	1.00	1.00	1.00	260
9	0.97	0.97	0.97	259
10	0.98	0.96	0.97	261
11	1.00	0.99	1.00	240
12	1.00	1.00	1.00	235
13	1.00	1.00	1.00	246
14	0.99	0.95	0.97	255
15	1.00	1.00	1.00	260
16	0.99	0.97	0.98	238
17	0.99	0.99	0.99	248
18	1.00	1.00	1.00	269
19	1.00	1.00	1.00	258
20	1.00	1.00	1.00	250
21	0.77	0.87	0.82	227
22	1.00	1.00	1.00	257
23	1.00	1.00	1.00	273
24	0.99	0.99	0.99	269
25	0.99	1.00	1.00	256
26	1.00	1.00	1.00	259
27	1.00	0.99	0.99	282
28	1.00	1.00	1.00	267
29	1.00	1.00	1.00	246
30	1.00	1.00	1.00	266
Średnia	0.98	0.97	0.97	265.6

6.3 Końcowa analiza wyników

Wyniki modelu wskazują na wysoki poziom dokładności, wynoszący 98%. Wyniki dla każdej klasy są ogólnie bardzo dobre, ale niektóre klasy (0, 5, 6, 9, 12, 16) mają nieco niższe parametry oceny.

Na przykład, dla klasy 1 'precision' wynosi 0,69, 'recall' 0,80, a 'f1-score' 0,74. 'Precision' dla tej klasy jest relatywnie niskie, co wskazuje, że model często mylnie przypisywał innym klasom instancje, które rzeczywiście należały do klasy 1. Natomiast niski 'recall' wskazuje, że model nie był w stanie wykryć wszystkich instancji tej klasy. Jest to uwarunkowane historycznie. Pod mapowaniem "1"kryje się bowiem Wielka Brytania, która swoje terytoria posiadała rozsiane po całym świecie, przez co modelowi jest ciężej używać parametru "Coordinates"w przewidywaniach. Wojna toczyła się u wybrzeży Brytanii od samego września '39, dlatego nie tonęły tylko wielkie jednotki pływające pod banderą zjednoczonego królestwa ale także średnie i małe statki, co utrudnia rozpoznanie po kolumnie "Tonnage". Z tego samego powodu dane zawarte w sekcji "Convoy"są mało przydatne, ponieważ statki brytyjskie bardzo często wchodziły w skład konwojów, jednak równie często tonęły samotnie.

Macierz pomyłek dodatkowo pokazuje, które klasy są najczęściej mylone. Na przykład, model często myli klasy 1 i 6, co widać w odpowiednich komórkach macierzy pomyłek. Nie powinno dziwić, że także i w tym przypadku z łatwością możemy wskazać fakty historyczne oraz geograficzne, tłumaczące to zjawisko. Klasa 6 odpowieda Norwegii, której flota po przegranej kampanii norweskiej w 1940 roku przeszła pod kontrolę brytyjczyków, którzy chętnie wykorzystywali jej statki do przerzutu zaopatrzenia zza oceanu.

Bardzo wysokie parametry odpowiadają klasie 2 czyli Stanom Zjednoczonym. Nie ma w tym nic dziwnego, ponieważ to na ten kraj przypadł główny wysiłek II Wojny Światowej, a co za tym idzie znaczący udział w konwojach alianckich. Co odróżnia Stany od reszty narodowości to powtarzający się wzorzec wśród statków, które zostały zatopione. Wszystkie okręty były średnie lub duże, brały udział w konwojach oraz tonęły na stałych i określonych szlakach handlowych. Łatwo jest więc dla tak zaawansowanego algorytmu jakim jest XGBoost odnaleźć te zależności i z dużą skutecznośćią znajdować takie instancje tej klasy.

7 Wnioski

Podsumowując, pomimo pewnych problemów z niektórymi klasami, ogólna skuteczność modelu jest bardzo wysoka. Przy dalszych badaniach warto by było przyjrzeć się bliżej klasom, które sprawiają modelowi najwięcej problemów, i zastanowić się, jak można by poprawić ich klasyfikację.

Analizując dane historyczne oraz geopolityczne bez większego wysiłku możemy odnaleźć związki i charakterystyki, które później przekładają się na wyniki algorytmu w przewidywaniu narodowości zatopionego statku.

Nie powinno dziwić, że najlepsze parametry oceny otrzymały te narodowości, których statki wpisywały się w pewne określone przez sytuację geograficzną i historyczną ramy.

Najtrudniejsze do rozpoznania klasy to oczywiście te z najmniej charakterystycznymi cechami. W czasie eksploracji okazało się, że są to przykładowo takie narodowości jak Wielka Brytania czy Norwegia, co jest jak najbardziej uzasadnione.

Jak można się spodziewać, w najmniejszym stopniu na poprawną predykcję wpływa nazwa statku (chociaż i tu można doszukiwać się pewnych zależności) oraz dane na temat Uboata, który go zatopił. Najcenniejsze dane to te z kolumn Coordinates, Convoy oraz Tonnage, ponieważ to one dostarczają nam najlepszych parametrów dzięki którym możemy sklasyfikować dany statek.

Badania pokazały, że pomimo początkowych trudności z algorytmem SVM, odpowiednia manipulacja danymi oraz dobór innego modelu pozwoliły osiągnąć znakomite wyniki, kosztem czasu wykonywania predykcji (dla bardzo wysoko ocenianego w benchmarkach procesora Apple M2 w konfiguracji z 16GB pamięci RAM wynosił on między 18 a 20 minut, a dla nieco bardziej wiekowego Intel Core i7 9 generacji z taką samą ilością pamięci operacyjnej prawie 40 minut(sic!)). Długośc obliczeń była pewnym ograniczeniem, ponieważ każda drobna poprawka do algorytmu kosztowała bardzo dużo czasu czekania na wyniki.

Warto przypomnieć, że obliczenia zostały w pewien sposób "uproszczone", ponieważ narodowości z liczbą wystąpień mniejszą od trzech zostały zgrupowane razem jako klasa "Other"co pozwoliło podnieść wydajność algorytmu z 0.97 do 0.98 dokładności.

Niniejsza analiza udowadnia tezę zakładającą, że na podstawie informacji o nazwie statku, uboacie, który go zatopił, lokalizacji zatopienia, konwoju

oraz tonażu jestemśmy w stanie kosztem skomplikowanych i długich obliczeń przewidzieć narodowość pechowego okrętu. Gdyby Kriegsmarine w czasie II Wojny Światowej udało się opracować tak zaawansowany model jakim jest XGBoost, dostarczyłoby to cennych danych wywiadowczych niemieckiemu dowódctwu, co mogłoby przyczynić się do innych wyników poszczególnych kampanii i potyczek, jednak z pewnością nawet taki (niemożliwy do osiągnięcia) przełom, nie był w stanie odwrócić losów wojny.

8 Ograniczenia i możliwe ulepszenia

Spośród wszystkich ograniczeń podczas przygotowywania analizy najbardziej mogły dać się we znaki ograniczenia sprzętowe a co za tym idzie czasowe. Nawet postrzegana jako wydajna konfiguracja (procesor Apple M2 + 16GB pamięci RAM) potrzebowała do 20 minut na ukończenie zadania klasyfikacji. Tak długi czas wykonania programu wprowadzał istotne ograniczenia, ponieważ każda, nawet najdrobniejsza poprawka wymagała bardzo długiego czasu oczekiwania na wyniki.

Scrapowanie danych odbybyło się bez większych problemów, jednak napisanie odpowiedniego programu do zczytywania pukntów z mapy okazało się dość czasochłonne i dla niektórych statków kończyło się błędem. Początkowo, problemem było także odczytanie narodowości statku, jednak dopasowując "alt"z obiektu html "img"można było przypisać okręt do odpowiedniej nacji.

Z pewnością, gdyby modelowi SVM poświęcić więcej uwagi oraz czasu, możnaby osiągnąć wyniki wyższe niż 40%, jednak dysponując tak potężną technologią jaką jest XGBoost najlepszym rozwiązaniem okazało się zgłebienie jego tematu i dostosowanie właśnie tego modelu do problemu analizy.

Jako możliwe ulepsze można wskazać rozszerzenie danych o dodatkowe kolumny. Pomimo tego, że sam wynik działania jest bardzo wysoki, kolejne parametry mogłyby przynieść jeszcze lepsze rezultaty.