**Projekt zaliczeniowy „Titanic” przygotowany na przedmiot Podstawy Uczenia Maszynowego (PUM)**

Autor: Dominik Suszek s23396

Spis treści

[Opis problemu 1](#_Toc155533788)

[Dane 1](#_Toc155533789)

[Sposób rozwiązania problemu 4](#_Toc155533790)

[Dyskusja wyników i ewaluacja modelu 5](#_Toc155533791)

[Podsumowanie 5](#_Toc155533792)

[Załączniki 6](#_Toc155533793)

# Opis problemu

Poniższy raport został opracowany na podstawie zbioru danych dotyczących katastrofy brytyjskiego statku pasażerskiego Titanic, kursującego na trasie pomiędzy USA i Europą. Ówcześnie był uważany za największy, najbardziej luksusowy i najbezpieczniejszy statek pasażerski na świecie. Niestety, w nocy z 14 na 15 kwietnia 1912 roku, podczas pierwszego rejsu na trasie Southampton – Cherbourg – Queenstown – Nowy Jork, statek zderzył się z górą lodową i zatonął.[[1]](#footnote-1) Ta tragedia zwróciła uwagę na niedostateczny poziom zabezpieczeń i potrzebę odświeżenia zasad bezpieczeństwa morskiego.

# Dane

Zbiór danych dotyczących katastrofy statku Titanic to bardzo dobrze znany zbiór wykorzystywany podczas nauki podstaw uczenia maszynowego. Zgodnie z rekomendacją, dane zostały pobrane ze strony: <http://analityk.edu.pl/wpcontent/uploads/2020/02/titanic.csv>

Zbiór zawiera następujące informacje (nie zawsze kompletne) o wszystkich pasażerach Titanica:

* **survival** – czy pasażer przeżył (1 – tak, 0 – nie)
* **pclass** – w której klasie dany pasażer wykupił bilet na podróż (1 – najlepsza, 2, 3)
* **name** – imię i nazwisko pasażera (razem z tytułem, np. Mr.)
* **sex** – płeć
* **age** – wiek
* **sibsp** – małżonek/małżonka i łączna liczba rodzeństwa, które znajdowało się razem z pasażerem/pasażerką na pokładzie statku
* **parch** – łączną liczba dzieci i/lub rodziców, którzy znajdowali się razem z pasażerem/pasażerką na pokładzie statku
* **ticket** – numer identyfikacyjny biletu
* **fare** – cena biletu
* **cabin** – numer kajuty
* **embarked** – port, w którym dana osoba wsiadła na pokład Titanica (C = Cherbourg, Q = Queenstown, S = Southampton)
* **boat** – numer szalupy, na której dana osoba się ewakuowała
* **body** – numer identyfikacyjny znalezionego ciała

Przed rozpoczęciem prac nad czyszczeniem zbioru danych, został on podzielony na dwie części – treningową i testową. W celu sprawdzenia czy rozkłady zmiennych dla różnych atrybutów są zbilansowane, przygotowano histogramy ([Załącznik 1](#_Załącznik_1._Histogramy)).

W związku z tym, że zmienne nie są zbilansowane, wykorzystano próbkowanie warstwowe na podstawie zmiennych *survived, sex, pclass.* Dzięki temu zarówno w zbiorze treningowym, jak i testowym powinny znaleźć się:

* takie same odsetki osób, które przeżyły i zginęły,
* taki sam udział kobiet i mężczyzn,
* taki sam udział osób podróżujących w poszczególnych klasach.

W poniższej tabeli przedstawiono statystyki opisowe zmiennych numerycznych:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **pclass** | **survived** | **age** | **sibsp** | **parch** | **fare** | **body** |
| **count** | 1309 | 1309 | 1046 | 1309 | 1309 | 1308 | 121 |
| **mean** | 2,29 | 0,38 | 29,88 | 0,50 | 0,39 | 33,30 | 160,81 |
| **std** | 0,84 | 0,49 | 14,41 | 1,04 | 0,87 | 51,76 | 97,70 |
| **min** | 1,00 | 0,00 | 0,17 | 0,00 | 0,00 | 0,00 | 1,00 |
| **25%** | 2,00 | 0,00 | 21,00 | 0,00 | 0,00 | 7,90 | 72,00 |
| **50%** | 3,00 | 0,00 | 28,00 | 0,00 | 0,00 | 14,45 | 155,00 |
| **75%** | 3,00 | 1,00 | 39,00 | 1,00 | 0,00 | 31,28 | 256,00 |
| **max** | 3,00 | 1,00 | 80,00 | 8,00 | 9,00 | 512,33 | 328,00 |

W celu zbadania zależności występujących pomiędzy zmiennymi, przygotowano macierz korelacji ([Załącznik 2](#_Załącznik_2._Macierz)). Jest to macierz kwadratowa, której elementy są wartościami współczynników korelacji dla wszystkich możliwych par zmiennych losowych.

Zauważono, że pomiędzy zmienną *survived*, a zmienną *pclass* występuje istotna ujemna korelacja (-0.56). Jest to intuicyjne, ponieważ wraz ze wzrostem wartości zmiennej *pclass* (klasa, w której podróżowała dana osoba) prawdopodobnie malał jej priorytet podczas akcji ewakuacyjnej. W związku z tym, dla większych wartości *pclass* maleje średnia wartość zmiennej *survived* (maleje szansa danej osoby na uratowanie się).

Kolejną istotną korelację (0.24) zaobserwowano pomiędzy zmiennymi *survived*oraz *fare*. Oznacza to, że wraz ze wzrostem ceny biletu, który dana osoba kupiła, rosła jej szansa na przeżycie. Można to wytłumaczyć w następujący sposób: droższe bilety były zapewne sprzedawane na miejsca w 1 klasie, które były położone w najbezpieczniejszej części statku. Dodatkowo, te osoby miały najwyższy priorytet podczas akcji ratunkowej.

Mniej istotną korelację (0.08) znaleziono między zmiennymi *survived* oraz *parch****.*** Oznacza to, że im więcej osób z najbliższej rodziny (rodzice i dzieci) dana osoba miała przy sobie w momencie katastrofy, tym większe prawdopodobieństwo przeżycia. Prawdopodobnie ten wynik jest podyktowany faktem, że krewni pomagali sobie w tym tragicznym momencie, wspierali się wzajemnie i walczyli o siebie.

Na wykresie słupkowym ([Załącznik 3](#_Załącznik_2._Wykres)) przedstawiano zależność pomiędzy dwiema zmiennymi – *survived* oraz *pclass*, dla których znaleziono najsilniejszą ujemną korelację.

Na kolejnym wykresie ([Załącznik 4](#_Załącznik_3._Wykres)) przedstawiono liczbę brakujących rekordów dla poszczególnych atrybutów. Najwięcej braków jest w odpowiednio w kolumnach *body, cabin, boat, home\_dest*. Zdecydowano o usunięciu tych zmiennych ze zbioru danych, ponieważ nie są to dane numeryczne i prawdopodobnie nie miałyby wysokiej istotności podczas trenowania modelu. Dla zmiennej *age*również zaobserwowano 263 brakujące wartości – jednakże jest to istotna zmienna. W związku z tym, najpierw pogrupowano dane na podstawie zmiennej *title*, a następnie do brakujących rekordów w kolumnie *age*przypisano średnią wieku dla danej grupy. Jeden wiersz, dla którego nie było przypisanej wartości dla zmiennej *fare,* został całkowicie usunięty ze zbioru danych.

W kolejnym kroku dodano trzy dodatkowe zmienne:

* ***family\_size*** – powstała poprzez dodanie do siebie wartości zmiennych *parch,* oraz *sibsp* oraz liczby 1, odpowiadającej za daną osobę,
* ***age\_range*** – zmienna kategoryczna, dla której przypisano następujące wartości:
  + 1 dla małych dzieci w wieku od 0 do 6 lat,
  + 2 dla większych dzieci w wieku od 6 do 12 lat,
  + 3 dla nastolatków w wieku od 12 do 18 lat,
  + 4 dla osób dorosłych powyżej 18 roku życia,
* ***MPC*** – zmienna liczbowa, której zadaniem było uwypuklenie szans na przeżycie dzieci i osób z pierwszej klasy. Powstała jako wynik mnożenia wieku danej osoby i klasy, w której podróżowała.

Następnie usunięto wartości odstające. Ustalono, że takimi wartościami są rekordy powyżej 200 dla zmiennej *fare* ([Załącznik 5](#_Załącznik_4._Wykres)),oraz wiek powyżej 67 lat ([Załącznik 6](#_Załącznik_6._Wykres)). Zastąpiono je średnimi wartości, wyliczonymi bez uwzględniania tychże wartości odstających.

Ostatnią modyfikacją zbioru danych było przekształcenie etykiet tekstowych zmiennej *sex* na liczby. Przypisano wartość 0 dla wszystkich rekordów ‘female’, a wartość 1 dla ‘male’.

Przed rozpoczęciem trenowania modelu, wszystkie dane numeryczne zostały znormalizowane za pomocą funkcji MinMaxScaler.

# Sposób rozwiązania problemu

Do wytrenowania modelu wybrano algorytm losowych lasów, który polega na uczeniu wielu drzew decyzyjnych skonstruowanych za pomocą różnych podzbiorów cech. Następnie otrzymane prognozy są uśredniane. Dzięki temu można uzyskać jeszcze lepszą wydajność. Innymi zaletami lasu losowego są:

* odporność na przeuczenie,
* możliwość dokładniejszego odtworzenia zależności pomiędzy zmiennymi, niż byłyby to w stanie zrobić drzewa decyzyjne,
* odporność na różnorodne problemy związane z danymi.

Pierwszym etapem realizacji projektu było zapoznanie się z danymi i próba zrozumienia zależności występujących pomiędzy zmiennymi. Do tego celu wykorzystano wykres przedstawiający macierz korelacji pomiędzy zmiennymi, statystyki opisowe zmiennych, a także pozostałe wykresy przedstawiające związki pomiędzy. Ostatecznie wybrano następujące zmienne: *pclass, sex, age, sibsp, parch, fare, family\_size, age\_range, MPC.*

Po przygotowaniu danych, rozpoczęto proces trenowania modelu. Do wyboru najlepszych parametrów wykorzystano metodę przeszukiwania siatki (*grid search*). Znalezione najlepsze parametry to: {'max\_depth': 5, 'min\_samples\_split': 3, 'n\_estimators': 100}.

Jako główny punkt odniesienia do oceny jakości modelu wykorzystano poziom trafności (accuracy) – założono, że docelowo powinien on przekroczyć próg 80%.

# Dyskusja wyników i ewaluacja modelu

Poniżej przedstawiono względną istotność każdego atrybutu w generowaniu prognoz.

|  |  |
| --- | --- |
| pclass | 0.0985557 |
| sex | 0.4570433 |
| age | 0.05881217 |
| sibsp | 0.02966779 |
| parch | 0.01910072 |
| fare | 0.11064422 |
| family\_size | 0.0573775 |
| age\_range | 0.0132404 |
| MPC | 0.15555819 |

Dla powyższego zbioru atrybutów uzyskano następujący poziom trafności: 0.8192. Poniżej przedstawiono macierz pomyłek:

|  |  |
| --- | --- |
| **TP**  148 | **FP**  12 |
| **FN**  35 | **TN**  65 |

Pozostałe miary pozwalające na ocenę klasyfikatora wyglądają następująco:

* precyzja (precision): 0.8442 – odsetek przykładów zaprognozowanych pozytywnie, które rzeczywiście są pozytywne,
* czułość (recall): 0.65 – prawdopodobieństwo, że klasyfikacja będzie poprawna pod warunkiem, że przypadek jest pozytywny,
* F1-score: 0.7345 – średnia harmoniczna pomiędzy precyzją i czułością.

# Podsumowanie

Ostatecznie udało się osiągnąć wyższy poziom trafności (ponad 81.9%) względem zakładanego (80%).

Najbardziej czasochłonnym zadaniem była inżynieria cech. Koniecznym okazało się dodanie dodatkowych atrybutów, uzupełnienie brakujących wartości, przycięcie odstających rekordów oraz normalizacja danych numerycznych.

Inspiracją do prac nad modelem oraz niniejszym raportem był podręcznik „Uczenie maszynowe z użyciem Scikit-Learn, Keras i TensorFlow. Wydanie II” autorstwa Aurélien Géron.

# Załączniki

A group of blue bars

Description automatically generated

## Załącznik 1. Histogramy zmiennych *sex*, *survived*,oraz *pclass*

A screenshot of a graph

Description automatically generated

## Załącznik 2. Macierz korelacji

A graph of blue and orange bars

Description automatically generated

## Załącznik 3. Wykres słupkowy przedstawiający zależność pomiędzy klasą, w której podróżowała dana osoba, a tym czy przeżyła katastrofę

A graph with numbers and a number of people

Description automatically generated with medium confidence

## Załącznik 4. Wykres słupkowy przedstawiający brakujące wartości w zbiorze danych

## A graph with red and blue dots Description automatically generated

## Załącznik 5. Wykres przedstawiający wartości odstające dla zmiennej *fare*

A screen shot of a graph

Description automatically generated

## Załącznik 6. Wykres przedstawiający wartości odstające dla zmiennej *age*

1. <https://www.britannica.com/topic/Titanic> [dostęp: 04.01.2024 r.] [↑](#footnote-ref-1)