Politechnika Wrocławska

Wydział Informatyki i Telekomunikacji

Jakub Andrzejewski

Nr indeksu: 266514

Analiza szans przeżycia katastrofy lotniczej na podstawie historycznych danych

Czerwiec 12, 2023

Spis treści

1. Problemu badawczy

Celem badania jest analiza szans przeżycia katastrofy lotniczej na podstawie danych historycznych. Pozyskane dane obejmują informacje na temat praktycznie wszystkich, zarejestrowanych wypadkach lotniczych. Dane zawierają informację dotyczące samolotów zarówno, cywilnych jak i wojskowych. Obejmują one szczegóły takie jak: data, czas, lokalizacja, operator lotu, numer lotu, trasa, typ samolotu, rejestracja, ilość osób na pokładzie, liczba ofiar śmiertelnych, liczba ofiar na ziemi, a także podsumowanie.

Problemem badawczym, jest próba odpowiedzenia na pytanie: "Jakie są szanse przeżycia katastrofy lotniczej, biorąc pod uwagę różne zmienne, takie jak region, miesiąc, dzień tygodnia, typ operatora(cywilny czy wojskowy), producent samolotu, oraz porę dnia?"

Model, który zostanie stworzony w ramach tego projektu, będzie próbował przewidzieć szansę przeżycia katastrofy lotniczej. Model ten zostanie nauczony na danych historycznych, a następnie przetestowany, aby sprawdzić, jak jest jego skuteczność.

2. Pozyskanie danych

Dane wykorzystane do analizy zostały pozyskane z witryny `https://www.planecrashinfo.com/database.htm` za pomocą techniki skrapowania danych(web scraping). Technika ta pozwala na ekstrakcję informacji z witryn internetowych.

Wykorzystując bibliotekę BeautifulSoup, stworzyłem skrypt w języku Python, który automatycznie przechodzi przez strony z powyższej witryny, zawierające szczegółowe dane, na temat każdej z katastrof, zaczynając od roku 1920 aż do 2023. Każda strona ma unikalny adres URL składający się z roku i numeru wypadku w danym roku. Zatem skrypt iteracyjnie generuje adresy URL i przechodzi do każdej strony, dopóki nie napotka strony, która nie istnieje (co znaczy, że nie ma więcej wypadków z danego roku).

Na każdej stronie skrypt lokalizuje odpowiednią tabelę z danymi, a następnie przechodzi przez każdy jej wiersz, zapisując dane do pliku CSV. Przy każdym zapisie, skrypt oczekuje przez 5 sekund, aby nie doprowadzić do przeciążenia serwera strony(co działo się, bez ustawionego opóźnienia).

3. Przygotowanie i wyczyszczenie danych

Zestaw danych zawiera następujące kolumny:

- Date - data katastrofy.
- Time - godzina katastrofy.
- Location - miejsce katastrofu.
- Operator - operator lotu.
- Flight - numer lotu.
- Route - trasa lotu.
- Type - model samolotu.
- Registration - numer rejestracyjny samolotu.
- cn/In - numer konstrukcyjny lub seryjny / Numer liniowy lub numer kadłuba
- Aboard - liczba osób na pokładzie.
- Fatalities - liczba ofiar śmiertelnych.
- Ground - liczba ofiar na ziemi.
- Summary - podsumowanie zdarzenia.

Do modelowania szans na przeżycie katastrofy lotniczej, wybrałem następujące cechy: Region, Miesiąc, Dzień tygodnia, Typ Operatora, Typ Samolotu oraz Porę dnia. Region jest pobierany z kolumny Location, Miesiąc i Dzień tygodnia z kolumny Date. Typ Operatora jest określany na podstawie kolumny Operator, wartość tej kolumny jest mapowana na "Military"lub "Civilian". Typ Samolotu określany jest na podstawie kolumny Type, pobieram z niej tylko producenta samolotu. Natomiast Pora dnia jest określna na podstawie godziny z kolumny Time. Używam etykietowania, aby przekształcić wartości kategorialne na liczby.

W zbiorze danych, występują puste wartości w kolumnach Time, Typ Operatora, Typ Samolotu i Pora dnia. Postępuje z nimi w następujący sposób:

- Time - puste wartości są zastępowane średnią wartością czasu.
- Typ Operatora, Typ Samolotu, Pora dnia - puste wartości są zastępowane wartością najczęściej występującą (modą).

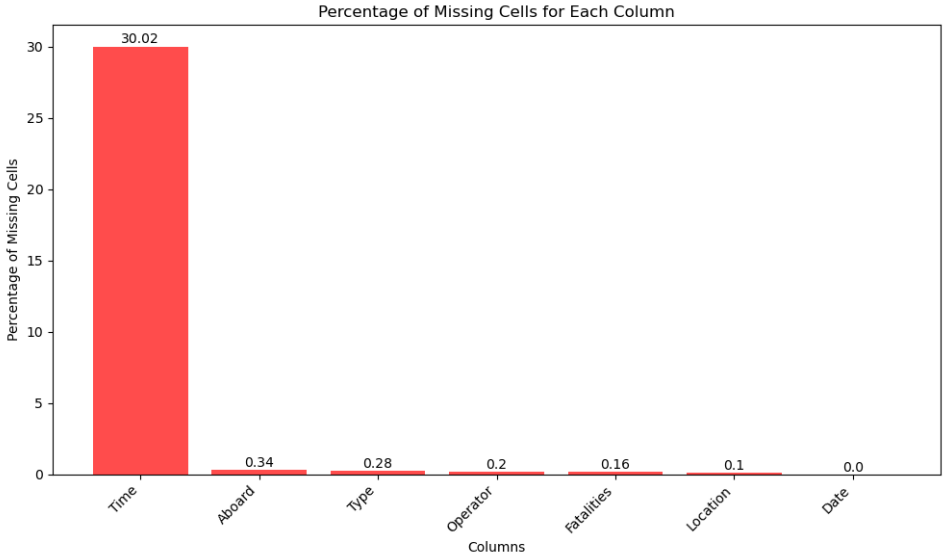
Wykorzystanie średniej i mody do zastępowania pustych wartości pozwala na minimalizację wpływu brakujących danych na jakość naszego modelu. Średnia jest dobrą miarą dla danych ciągłych, takich jak czas, podczas gdy moda jest najbardziej odpowiednia dla danych kategorialnych, takich jak Typ Operatora, Typ Samolotu i Część dnia.

W końcowym etapie przygotowania danych, tworzymy kolumnę Wskaźnik Przetrwania, obliczając różnicę między liczbą osób na pokładzie (Aboard) i liczbą ofiar śmiertelnych (Fatalities), a następnie dzieląc ją przez liczbę osób na pokładzie(Aboard). Puste wartości w Wskaźniku Przetrwania są zastępowane średnią wartością Wskaźnika Przetrwania.

4. Wstępna analiza danych

Przed przystąpieniem do modelowania danych, ważne jest przeprowadzenie wstępnej analizy danych, która pozwoli na lepsze zrozumienie naszego zestawu danych i jego cech.

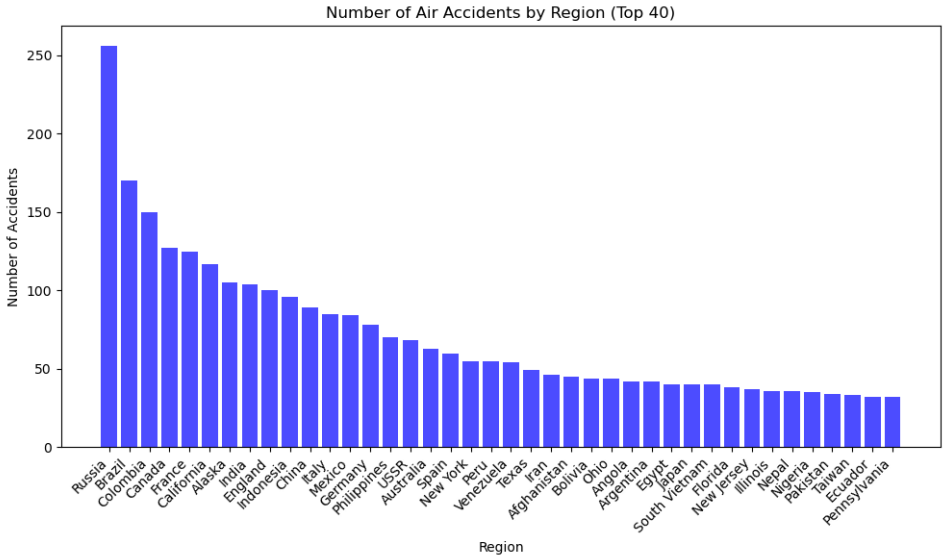
Na początek przeanalizujmy, jaki procent poszczególnych danych jest pusty, aby móc ocenić jak dużo danych zostanie zastąpionych średnią lub medianą.



Wykres 1. Procent brakując danych w wykorzystywanych kolumnach

Z powyższego wykresu wynika, że wszystkie kolumny, poza kolumną Time, praktycznie nie mają pustych wartości. Jednak w kolumnie Time prawie 1/3 komórek jest pusta, co może świadczyć o tym, że podczas przewidywania szans przeżycia, pora dnia(która jest wyznaczana na podstawie kolumny Time) będzie miała mały wpływ na wynik.

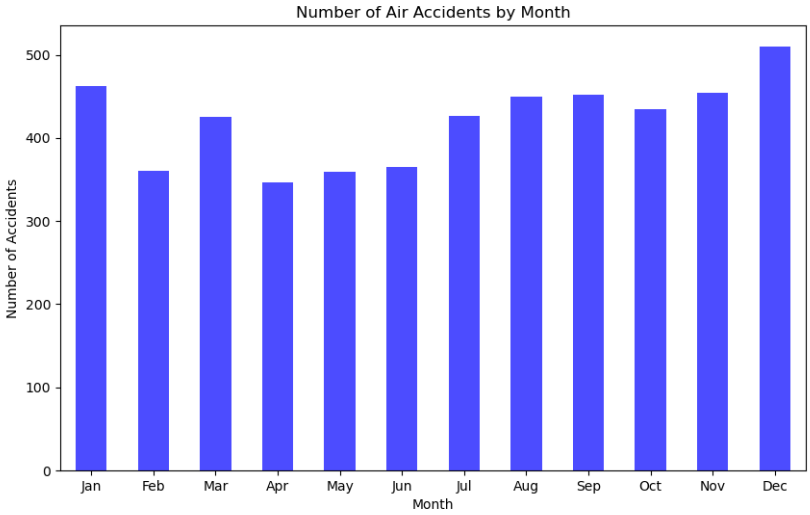
Teraz przeanalizujmy, jak wygląda ilość katastrof w zależności od poszczególnych cech.



Wykres 2. Ilość katastrof w zależności od regionu

Na powyższym wykresie widzimy, ilość wypadków lotnicznych w 40 regionach w których było ich najwięcej. Ewidetnie najwięcej wypadków miało miejsce w Rosji, następnie prawie o 100 mniej w Brazyli, kilka mniej w Kolumbii itd. Z powyższego wykresu, jasno wynika, że istnieje zależność pomiędzy liczbą katastrof a regionem w którym miały one miejsce.

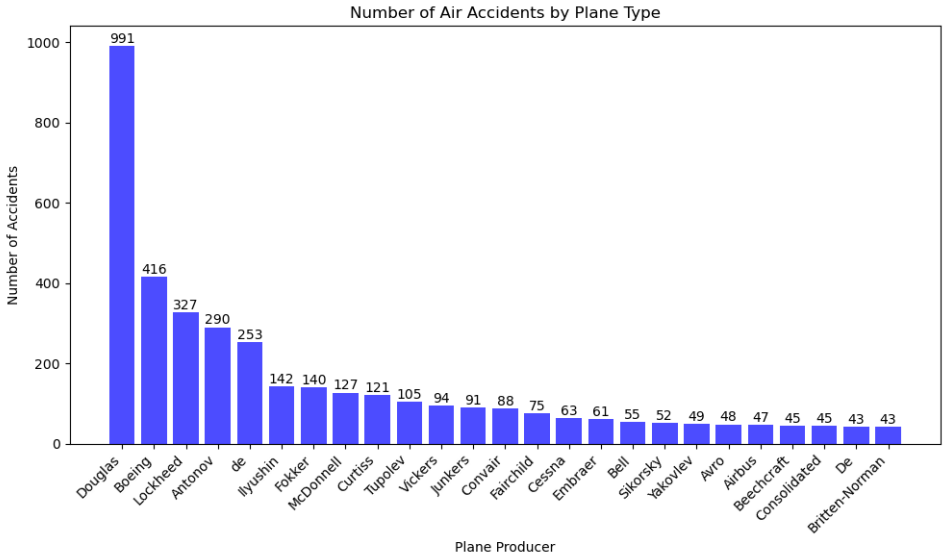
Następnie zobaczmy, jak rozkłada się ilość katastrof w zależności od miesiąca w roku.



Wykres 3. Ilość katastrof w zależności od miesiąca

Analizując powyższy wykres widzimy, że istnieje zależnośc między miesiącem a ilością katastrof. W miesiącach wiosenno-letnich(kwiecień, maj, czerwiec) było najmniej wypadków. Natomiast w miesiąch jesienno-zimowych(szczególnie w grudniu) było ich zdecydowanie najwięcej.

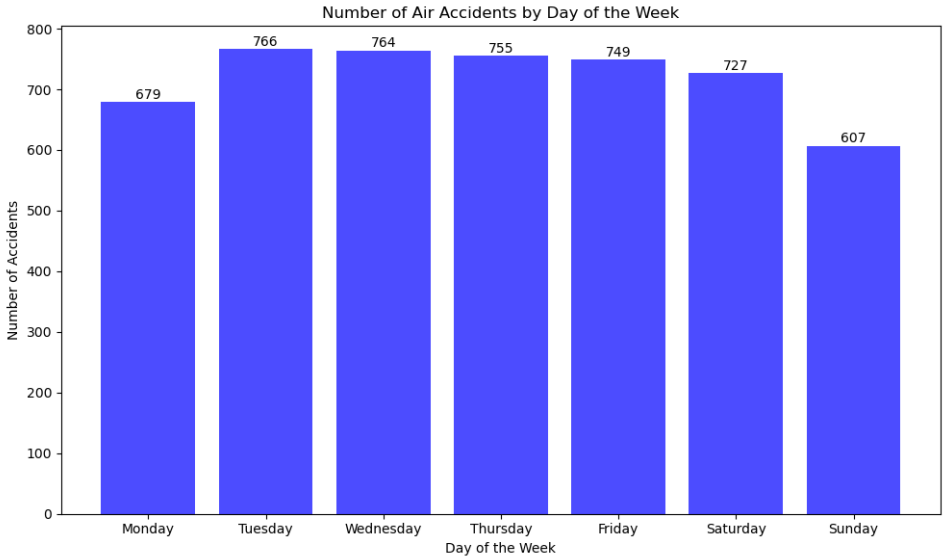
Teraz zobaczmy, jaka zależność występuję między producentem samolotu, a ilością wypadków.



Wykres 4. Ilość katastrof w zależności od producenta samolotu

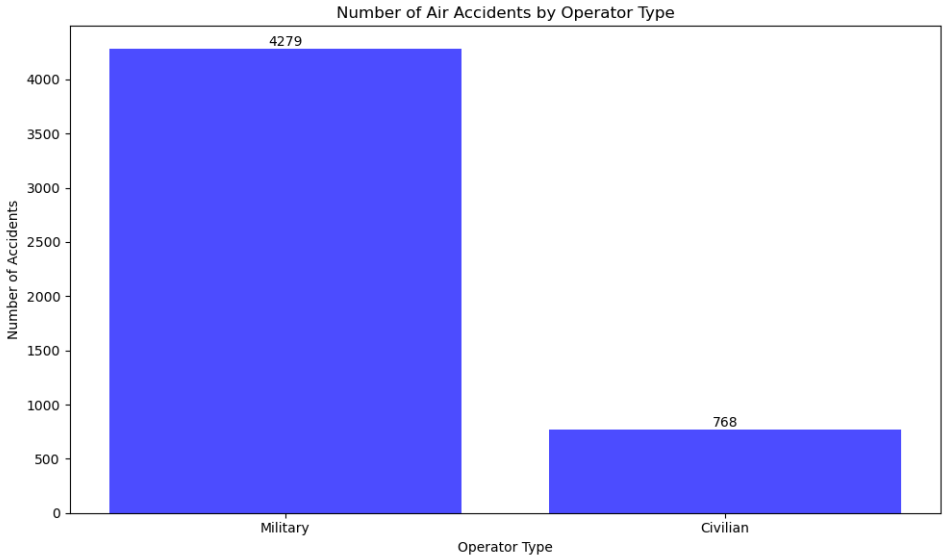
Na powyższym wykresie możemy zauważyć, że na ponad dwukrotne prowadzenie przed drugim w kolejności producentem wyłania się firma Douglas. Również, ciekawą zależnością jest

różnica w ilości katastrof między dwoma największymi producentami samolotów na świecie - Boeing i Airbus, która jest prawie 10-krotna. Jednak może mieć na to wpływ fakt, że firma Airbus została założona około 50 lat później. Po raz kolejny widzimy zależność między ilością katastrof, a w tym przypadku producentem samolotu.



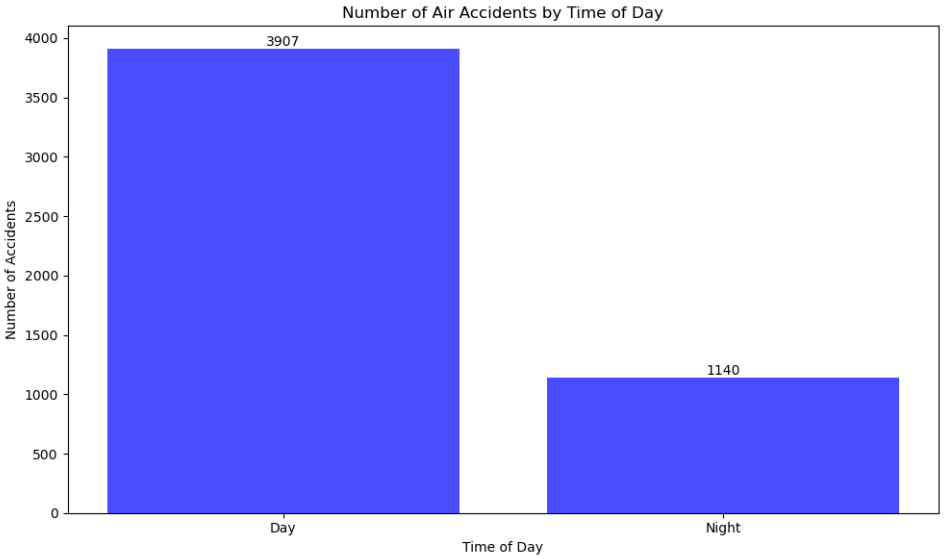
Wykres 5. Ilość katastrof w zależności od dnia tygodnia

Z powyższego wykresu wynika, że nie występują większe różnice między ilością katastrof dla poszczególnych dni tygodnia. Jednak widać, że najmniej katastrof miało miesjsce w Niedziele, może to być spowodane tym, że w niedziele odbywa się najmniej lotów w tygodniu. Prawdopodbnie, bedzię to skutkowało tym, że dzień tygodnia wprowadzany do modelu nie będzie miał dużego wpływu na szanse przeżycia.



Wykres 6. Ilość katastrof w zależności operatora

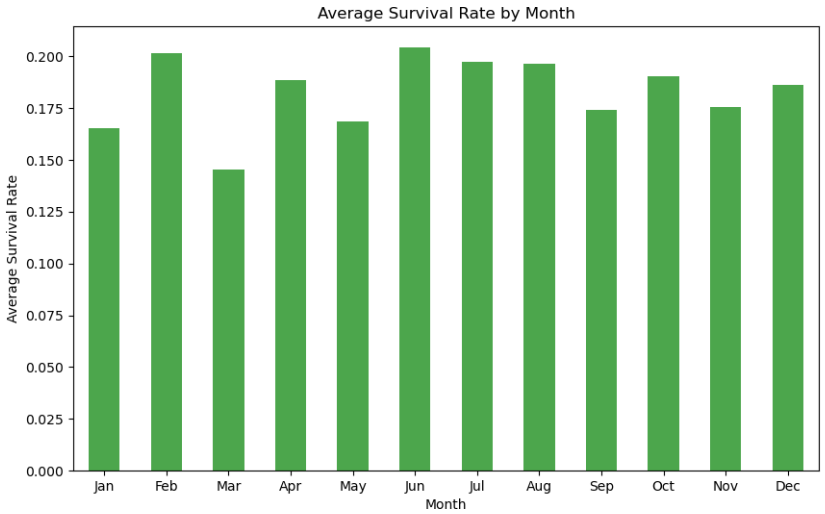
Na powyższym wykresie widać kolosalną różnicę pomiedzy ilością katastrof wojkowych statków powietrznych a cywilnych. Biorąc pod uwagę, jakie ryzyko niosą za sobą loty i operację wojskowe, wydaje się to uzasadnione. Prawdopodobnie ta cecha, będzie miała bardzo duży wpływ na wysokość szans przeżycia, generowaną przez model.



Wykres 7. Ilość katastrof w zależności od pory dnia

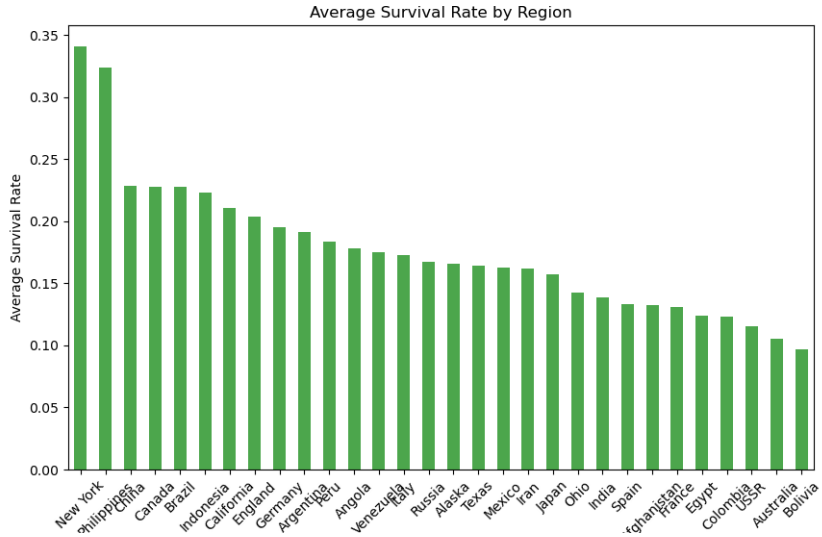
Z wykresu zależności ilościa katastrof od pory dnia, wynika, że prawie 4 razy więcej katastrof miało miejsce w dzień. Jednak należy wziąć pod uwagę, że blisko 30% katastrof nie miało podanej informacji o jej godzinie i te dane zostały zastąpione średnią, dlatego powyższy wykres prawdopodbnie nie jest do końca prawdziwy. Ponadto, znacznie więcej lotów odbywa się w dzień, co też może mieć przełożenie na wiekszą ilość wypadków.

Teraz przeanalizujmy jak wygląda odsetek ludzi którzy przeżyli katastrofęy w zależności od wybranych cech.



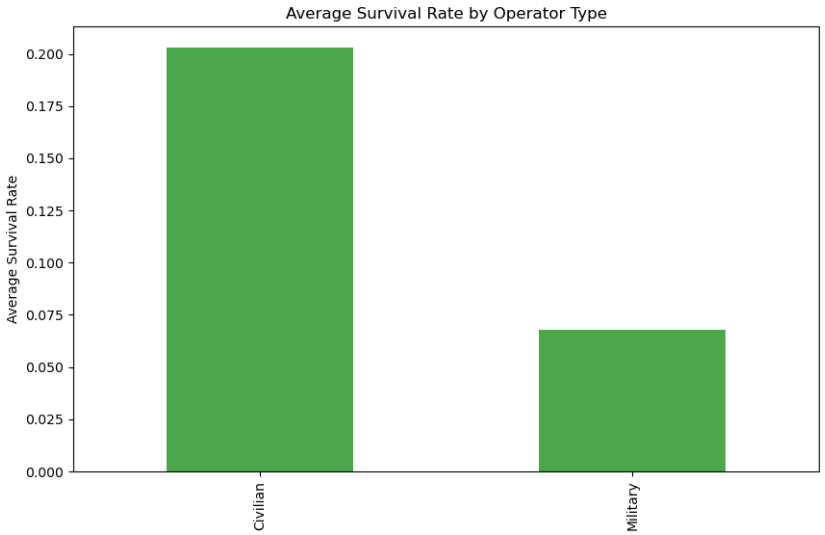
Wykres 8. Wskaźnik przetrwania w zależności od miesiąca

Z powyższego wykresu wynika, że miesiąc w którym miała miejsce katastrofa, nie ma większego wpływu na procent ludzi którzy przeżyli. Możemy jednak zauważyć, że najmniej ludzi przeżyło w wypadkach które miały miejsce w Marcu, natomiast najwięcej w Lutym i Czerwcu.



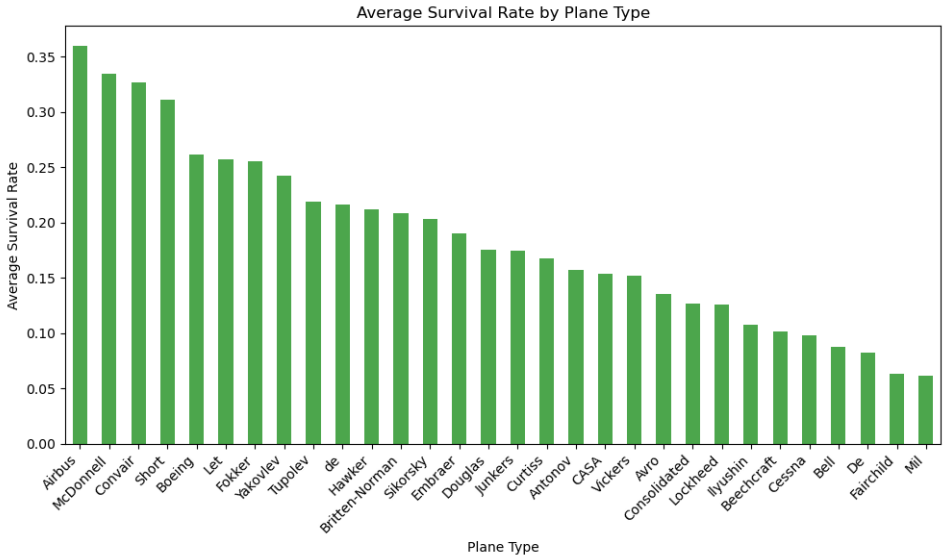
Wykres 9. Wskaźnik przetrwania w zależności od regionu

Na wykresie zależności wskaźnika przetrwania od regionu, ewidetnie wyróżniają się 2 rejony w których szanse przeżycia są największe, jest to Nowy York i Filipiny. Natomiast kolejne regiony mają szanse przeżycia mniejsze aż o blisko 15%. Z powyższego wykresu wynika, znacząca zależność pomiedzy szansami na przeżycie a Regionem katastrofy.



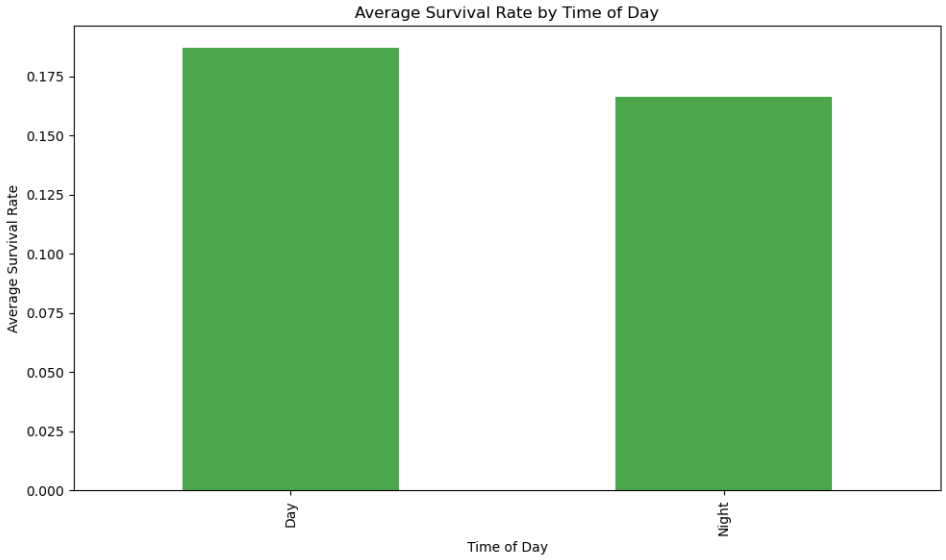
Wykres 10. Wskaźnik przetrwania w zależności od typu operatora

Z powyższego wykresu wynika, że prawie 4 razy więcej ludzi przeżywało katastrofy samolotów cywilnych niż wojskowych. Tak jak w przypadku ilości katastrof, biorąc pod uwagę ryzyko lotów i operacji wojskowych wydaje się to być uzasadnione.



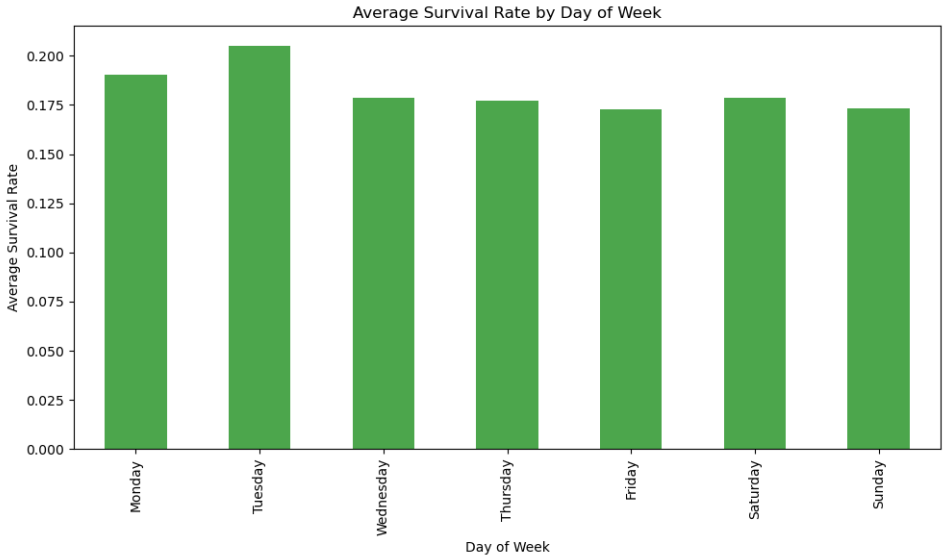
Wykres 11. Wskaźnik przetrwania w zależności od rodzaju samolotu

Podobnie jak w przypadku wykresu zależności ilości katastrof od rodzaju samolotu, widzimy, że występuje znacząca zależność między procentem osób które przeżyły, a typem samolotu jakim leciały. Co ciekawe, największy odsetek ludzi przeżyło lecąc samolotem firmy Airbus, która miała 10 raz mniej katastrof od firmy Boeing. Natomiast Boeing plasuje się na 5 miejscu na tym wykresie, wypadki na pokładzie tych samolotów przeżyło o 10% ludzi mniej.



Wykres 12. Wskaźnik przetrwania w zależności od pory dnia

Z tego wykresu wynika, że nie ma większej zależności szans przeżycia, od pory dnia, mimo, że w dzień było około 4 razy więcej katastrof. Różnica pomiędzy szansami przeżycia w dzień i w nocy wynosi zaledwie około 2%.



Wykres 13. Wskaźnik przetrwania w zależności od dnia tygodnia

Podobnie wygląda rozkład szans przeżycia w zależności od dnia tygodnia. Również nie ma większej różnicy, między wartością średniej szansy przeżycia dla poszczególnych dni, jedynie wtorek ma delikatnie wyższe szanse, ale jest to nieznacząca przewaga.

5. Przygotowanie modelu

Na podstawie zebranych i przygotowanych danych, zdecydowałem się na zastosowanie modelu `RandomForestRegressor` z biblioteki Scikit-learn. Model lasów losowych jest jednym z popularniejszych modeli uczenia maszynowego. Wspomniany model to implementacja algorytmu Random Forest dla problemów regresji. Algorytm Random Forest jest typem zbiorowego algorytmu uczenia maszynowego, który integruje wiele drzew decyzyjnych do generowania prognozy. Model `RandomForestRegressor` działa poprzez tworzenie wielu drzew decyzyjnych podczas treningu i generowania prognozy poprzez uśrednianie prognoz z każdego drzewa. Wszystkie drzewa w modelu `RandomForestRegressor` są trenowane równolegle, co sprawia, że algorytm Random Forest jest niezwykle wydajny.

Wybrałem go do przewidywania szans przeżycia katastrofy lotniczej z kilku powodów:

- **Praca z danymi historycznymi**: RandomForest jest znany ze swojej zdolności do radzenia sobie z danymi historycznymi. Jego mechanizm oparty na drzewach decyzyjnych pozwala na analizę sekwencji zdarzeń w przeszłości, co jest istotne w tym przypadku, gdyż szanse przeżycia katastrofy lotniczej mogą być ściśle powiązane z historycznymi trendami i wzorcami.

- **Zdolność do modelowania złożonych zależności**: W danych istnieje wiele zmiennych, które mogą wpływać na szanse przeżycia katastrofy lotniczej. Model RandomForest jest zdolny do modelowania nieliniowych i interakcyjnych zależności między zmiennymi, co sprawia, że jest doskonałym wyborem dla tego problemu.

- **Obsługa różnych typów danych**: Nasze dane zawierają zarówno zmienne kategoryczne, jak i numeryczne. Model RandomForest dobrze radzi sobie z obydwoma typami danych.

- **Oszacowanie ważności cech**: Las losowy pozwala na ocenę ważności cech, co jest szczególnie przydatne w tym przypadku, gdzie mamy wiele zmiennych, które mogą wpływać na szansę przeżycia. Pozwala to na lepsze zrozumienie, które zmienne mają największy wpływ na przewidywaną szansę przeżycia.

Następnie, przy pomocy funkcji train_test_split podzieliłem dane na zestawy treningowe i testowe w proporcji 80/20. Wykorzystanie zbioru testowego pozwoli na sprawdzenie skuteczności modelu na nowych, niewidzianych wcześniej danych.

Przy wyborze optymalnych parametrów dla modelu skorzystałem z metody GridSearchCV. Pozwala ona na przeszukanie przestrzeni hiperparametrów i wybranie takiej kombinacji, która daje najlepsze wyniki. Metoda ta jest szczególnie użyteczna, gdy nie ma pewności co do optymalnych wartości parametrów, czyli dokładnie tak jak w tym przypadku.

Po znalezieniu optymalnych parametrów, utworzyłem model RandomForestRegressor, który został następnie wytrenowany na zestawie treningowym.

Wykorzystałem walidację krzyżową (cross_val_score) z 5 podziałami, aby sprawdzić stabilność modelu. Po wytrenowaniu modelu, przewidziałem wartości dla zestawu testowego i obliczyłem błąd średniokwadratowy (mean_squared_error). Jest to podstawowa miara oceny jakości modelu regresji, która mierzy różnicę pomiędzy przewidywanymi a rzeczywistymi wartościami.

Oprócz tego, stworzyłem funkcję predict_survival_rate, która pozwala przewidzieć szansę przeżycia dla określonego przypadku. Funkcja ta pobiera dane wejściowe, konwertuje je do odpowiedniego formatu, a następnie dokonuje przewidywania za pomocą modelu.

6. Ocena jakości modelu

Cross-Validation jest techniką używaną do oceny jakości modelu na różnych podzbiorach danych uczących. W tym przypadku, uzyskane wyniki MSE (Mean Squared Error - średni błąd kwadratowy) na 5 różnych podzbiorach są równe odpowiednio:

- 0.09576163
- 0.08908814
- 0.09260809
- 0.10031849
- 0.10300803

Wartość MSE jest miarą, która pokazuje, jak blisko nasze przewidywania są do rzeczywistych wartości - im niższa wartość, tym lepiej.

Średnia tych pięciu wyników, która jest wyznaczona jako miara globalnej jakości modelu, wynosi 0.09615687612099895. Ten wynik pokazuje, że model generalizuje się dość dobrze na różnych podzbiorach danych uczących.

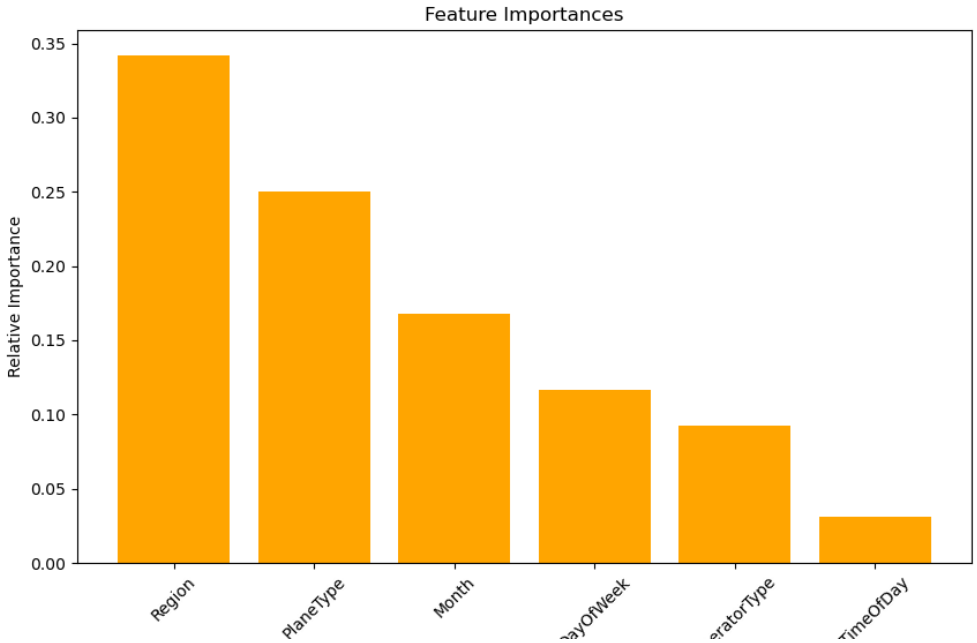
Dodatkowo, na zbiorze testowym, który nie był używany do treningu modelu, uzyskałem MSE równe 0.09192227847995471. Ta wartość jest nawet niższa niż średnia wartość z cross-validation, co sugeruje, że nasz model działa dość dobrze na nowych, nieznanych mu wcześniej danych.

Podsumowując, wyniki te wskazują, że model przewiduje szanse przeżycia katastrofy lotniczej z akceptowalnym błędem.

7. Działanie modelu

Przeprowadzona analiza oraz zbudowany model pozwoliły na dokładne zrozumienie i przewidzenie szans na przeżycie katastrofy lotniczej. Za pomocą modelu `RandomForestRegressor`, udało się również zidentyfikować najważniejsze cechy, które wpływają na szanse przeżycia.

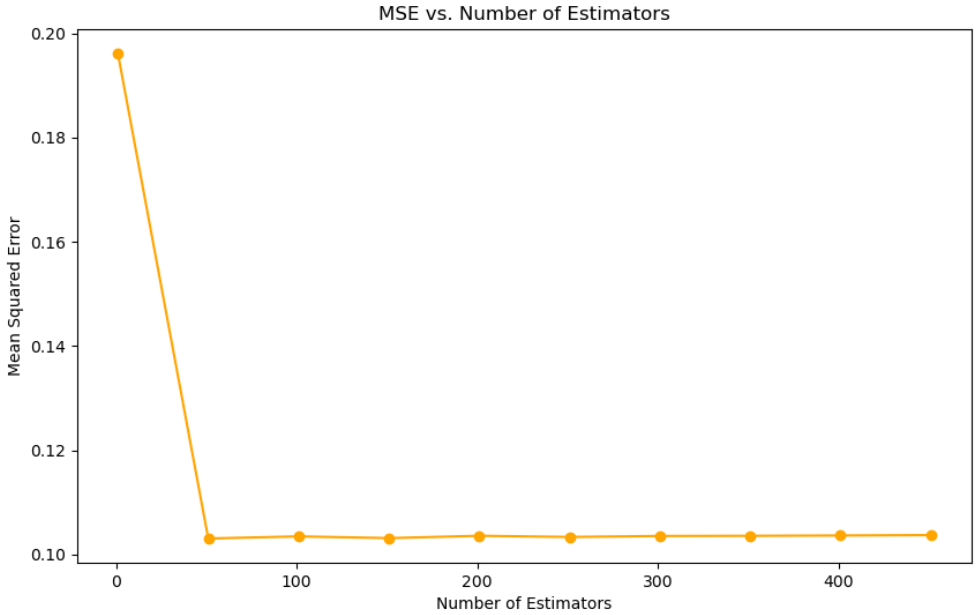
Poniższy wykres prezentuje istotność poszczególnych cech w naszym modelu. Najważniejsze cechy mają największy wpływ na końcowy wynik.



Wykres 14. Wpływ cech na oszacowane szanse przeżycia

Analiza istotności cech pokazała, że niektóre czynniki mają większy wpływ na szanse przeżycia katastrofy lotniczej niż inne. Możemy zauważyć, że Region wraz z typem samolotu mają łącznie wynik bliski 0.6, co oznacza, że pozostałe 4 cechy, są znacznie mniej ważne. W takim razie, najprawdopodbniej, podczas przewidywań z wykorzystaniem tego modelu, największy wpływ na wynik, będzie miała zmiana Regionu lub typu samolotu. Warto jednak zauważyć, że nawet mniej istotne cechy, takie jak np. czas dnia czy typ operatora, mogą mieć znaczenie w konkretnych sytuacjach.

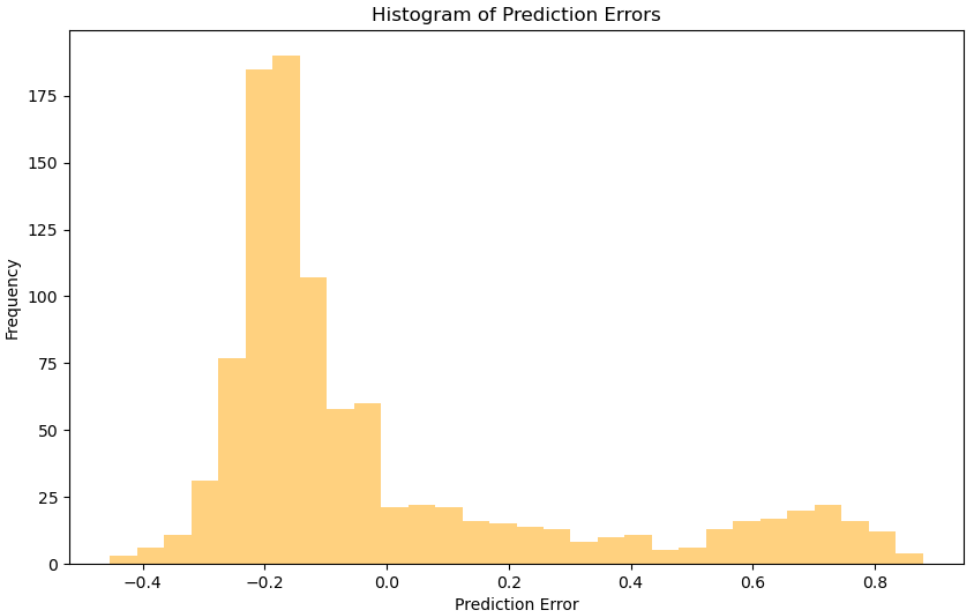
Poniższy wykres, prezentuje zależność MSE (Mean Squared Error - średni błąd kwadratowy), od ilości estymatrów, czyli pokazuje jak skuteczność modelu zmienia się wraz ze zmianą liczby estymatorów. Taki wykres może pokazać, czy model jest przetrenowany (overfitting - kiedy model jest zbyt skomplikowany i "uczy się"danych treningowych na pamięć, co może prowadzić do gorszej skuteczności na nowych, nieznanych danych) lub niedotrenowany (underfitting - kiedy model jest zbyt prosty i nie jest w stanie uchwycić zależności w danych).



Wykres 15. Wykres skuteczność modelu od liczby estymatorów

Wykres, sugeruje, że model zaczyna stabilizować się i poprawiać swoją wydajność po przekroczeniu pewnej liczby estymatorów, około 50. MSE (Mean Squared Error) jest miarą błędu, więc niższa wartość MSE oznacza lepszą wydajność modelu. Po przekroczeniu liczby estymatorów około 50, błąd średniokwadratowy modelu osiąga wartość około 0.1, co wskazuje na poprawę dokładności predykcji. Z tego wynika, że model osiąga lepszą wydajność, gdy liczba estymatorów przekracza 50. Nie wskazuje to jednak na to, że model jest przetrenowany ani niedotrenowany, ponieważ MSE stabilizuje się i utrzymuje na niskim poziomie dla większej liczby estymatorów. Wskazuje to, że model jest w stanie dobrze generalizować i robić dokładne predykcje na nowych, nieznanych danych, a nie tylko na danych treningowych.

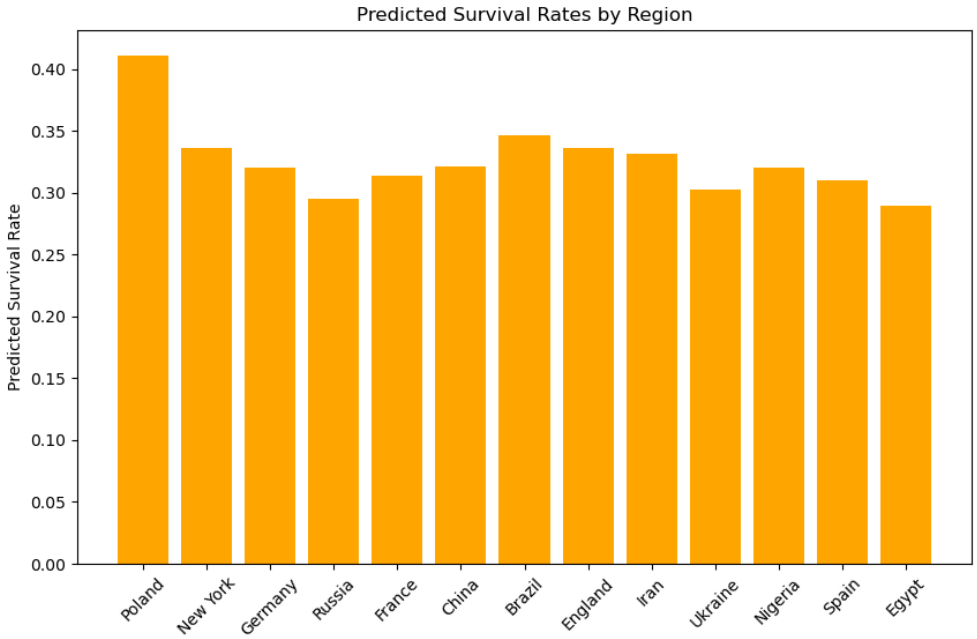
Następnie, przedstawiam histogram błędów predykcji, który ilustruje rozkład różnic między wartościami rzeczywistymi a przewidywanymi przez model. Analiza tego wykresu pozwoli zrozumieć, jak dobrze model radzi sobie z przewidywaniem danych, identyfikować potencjalne obszary, w których model może być niedokładny, oraz rozpoznać, czy występują systematyczne błędy.



Wykres 16. Historgram błędów predykcji

Na podstawie powyższego histogramu, widzimy, że większość błędów koncentruje się wokół zera. Jest to dobra wiadomość, ponieważ oznacza to, że model jest w stanie dokładnie przewidzieć wiele przypadków. Jednak, obserwujemy również kilka innych interesujących rzeczy. Po pierwsze, zauważamy, że rozkład błędów jest niesymetryczny - widzimy większe wartości błędów po lewej stronie zera. Może to sugerować, że model ma tendencję do przeszacowywania prawdopodobieństwa przeżycia. Po drugie, zauważamy długie ogony na obu końcach histogramu. Oznacza to, że mamy kilka przypadków, w których błąd predykcji jest stosunkowo duży. Może to świadczyć o tym, że model może mieć trudności z przewidywaniem skrajnych wartości. Warto podkreślić, że mimo pewnych niedoskonałości nasz model jest dość skuteczny w przewidywaniu szans na przeżycie.

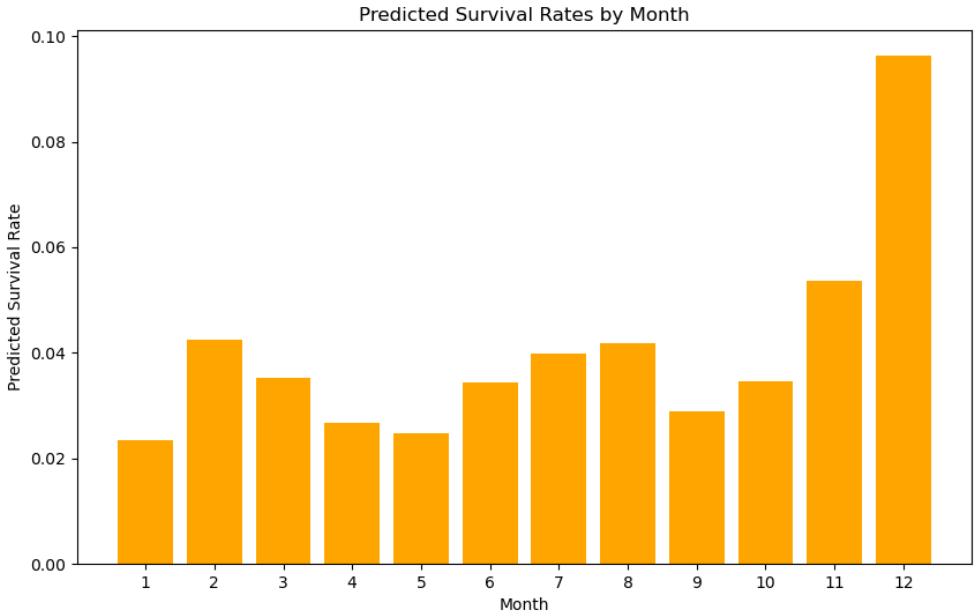
Aby zilustrować działanie naszego modelu i jego zdolność do przewidywania szans przeżycia katastrofy lotniczej w różnych warunkach, stworzyłem przykładowe wykresy przedstawiające przewidywane szanse przeżycia dla konkretnych cech. Poniższe przewidywanie zostało wykonane dla różnych regionów oraz identycznych pozostałych cech - loty cywilne, w lipcu, w poniedziałek, samolotem marki Boeing, podczas dnia.



Wykres 17. Przewidywane szanse przeżycia dla wybranych regionów

Analizując wyniki przedstawione na wykresie, można zauważyć, że przewidywane szanse przeżycia są nieco wyższe dla Polski, Nowego Jorku i Brazylii. Natomiast najniższe szanse przeżycia przewiduje model dla Rosji i Egiptu. Pozostałe regiony, takie jak Niemcy, Francja, Chiny, Anglia, Iran, Ukraina, Nigeria i Hiszpania, osiągają wyniki pośrednie. Niższy wynik dla Rosji pokrywa się z największą ilością katastrof w historii w tym regionie. Co ciekawe, dla krajów tak zwanego "trzeciego świata"jak Egipt, Nigeria czy Iran, szanse przeżycia są zbliżone do wyników osiąganych przez kraje bardziej rozwinięte.

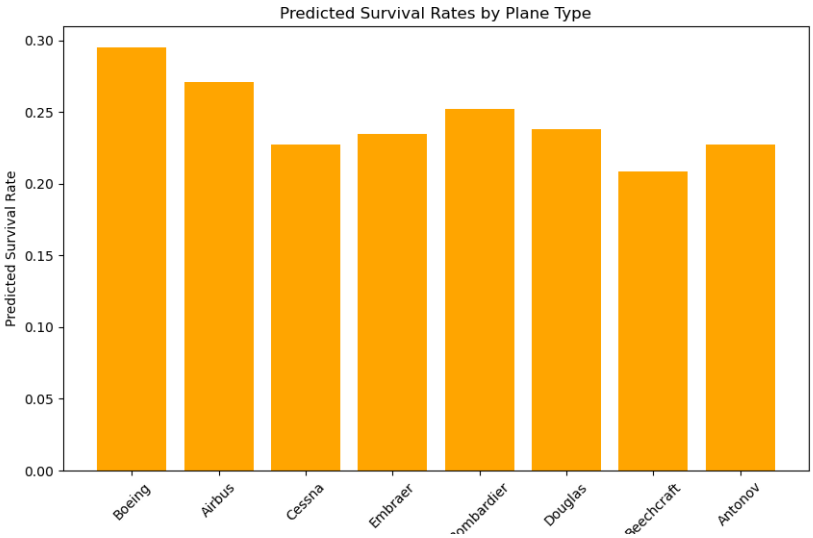
Kolejny wykres obrazuje jak zmieniają się szanse przeżycia, przewidywane przez model w zależności od miesiąca katastrofy. Pozostałe cechy były identyczne dla wszystich przypadków - wypadek w Niemczech, w piatęk, przelot militarny, samolot typu Boeing i przelot odbywa się w dzień.



Wykres 18. Przewidywane szanse przeżycia dla kolejnych miesięcy

Na tym wykresie, przede wszystkim należy zauważyć wartość szans przeżycia. Dla większości miesięcy oscylują one na poziomie 5%. Najprawdopodobniej ma to związek z tym, że tym razem przewidujemy szanse przeżycia dla lotów militarnych. Przy analizie danych zauważyliśmy, że katasrof militarnych było praktycznie 4 razy więcej niż cywilnych, a współczynnik ludzi który je przeżył był znacząco mniejszy niż w lotach cywilnych. Najpradopodobniej, właśnie ta cecha miała wpływ na tak niskie wyniki. Poza tym, możemy zauważyć, że w grudniu szansę przeżycia są prawie 2-3 krotnie wyższe niż w pozostałych miesiącach, jednak w tym przypadku ciężko zauważyć jakaś zależność z której mogłoby to wynikać. W pozostałych miesiącach szansę przeżycia utrzymują się na podobnym poziomie zaczynając się od 2% i nie przekraczając 6%.

Na poniższym wykresie, zilustrowałem szansę przeżycia, jakie oszacował model w zależności od producenta samolotu. Pozostałe cechy pozostawały identyczne dla każdego przewidywania - lot w Rosji, w lipcu, w poniedziałek, cywilny i w nocy.



Wykres 19. Przewidywane szanse przeżycia dla wybranych typów samolotów

Analizując powyższy wykres, widzimy, że dla tych konretnych cech największe szanse przeżycia są w samolocie marki Boeing, mimo, że Boeing miał znacznie więcej wypadków niż Airbus. Jednak firma Airbus uplasowała się na drugim miejscu z bardzo niewielką stratą. Poza tym widzimy, że inne firmy mają mniejszę szanse przeżycia, jednak wyniki bardzo nie odbiegają od siebie. Należy zwrócić uwagę, że jako region w tym przewidywaniu została ustawiona Rosja, w której było najwięcej wypadków i mały odsetek ludzi przeżywał katastrofy, stąd prawdopodobnie wyniki nieprzekraczające 30%.

8. Wnioski

Na podstawie przeprowadzonej analizy, stworzonego modelu i uzyskanych wyników, możemy wyciągnąć kilka interesujących wniosków dotyczących szans na przeżycie katastrofy lotniczej.

Istotność cech: Najważniejsze cechy wpływające na szanse przeżycia katastrofy lotniczej, według naszego modelu, to region oraz typ samolotu.

Różnice regionalne: Model wskazał na różnice w przewidywanych szansach przeżycia w zależności od regionu. Przewidywane szanse na przeżycie katastrofy lotniczej były najwyższe dla Polski, Nowego Jorku i Brazylii, podczas gdy najniższe dla regionów takich jak Indie, Rosja i Chiny. Może to odzwierciedlać różnice w normach bezpieczeństwa, jakości infrastruktury lotniczej i warunkach pogodowych w tych regionach.

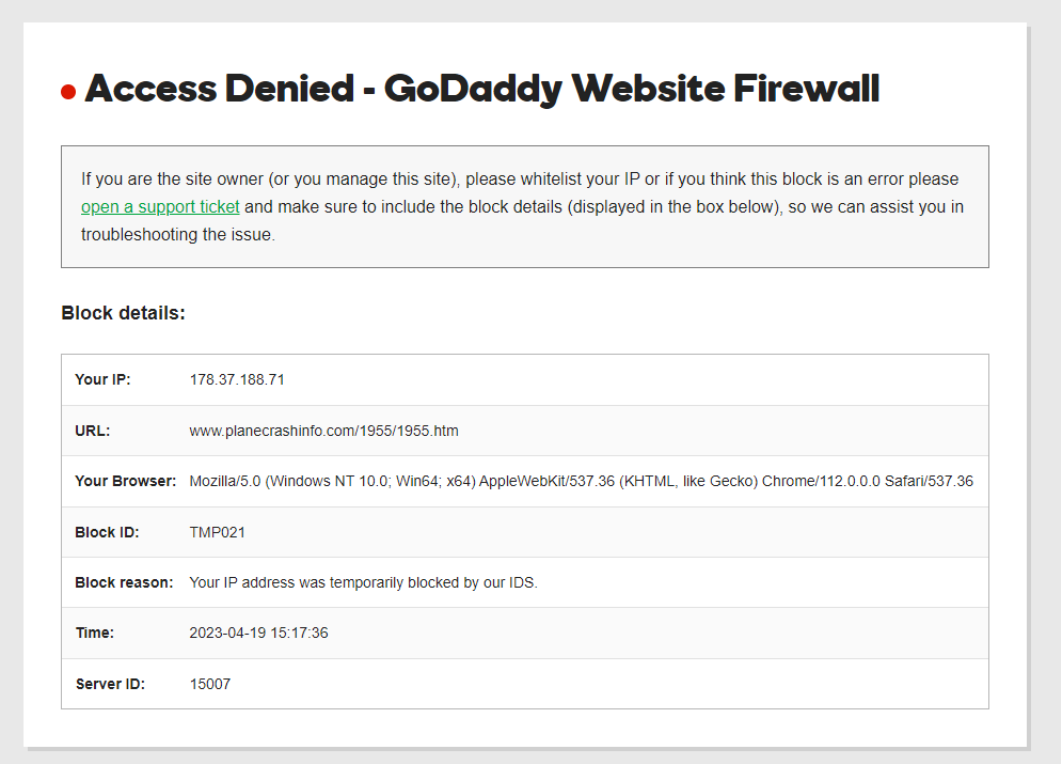
Typ samolotu: Znaczenie typu samolotu w przewidywaniu szans na przeżycie katastrofy lotniczej może sugerować, że niektóre modele samolotów są bezpieczniejsze niż inne w razie katastrofy.

Jakość modelu: Model, który stworzyłem, osiąga dobre wyniki zarówno na danych treningowych, jak i testowych, co wskazuje na jego solidność i zdolność do generalizacji. Błąd średniokwadratowy (MSE) jest dość niski, co oznacza, że różnica między przewidywanymi a rzeczywistymi szansami na przeżycie jest z reguły niewielka.

9. Komentarze, ulepszenia

Podczas analizy zbioru danych i tworzenie modelu zauważyłem kilka kwestii, które mogłyby wymagań poprawy w przyszłości:

Zbieranie Danych: Zauważyłem, że niektóre rekordy w naszym zbiorze danych były niekompletne. Poza tym, dane nie zawierały wszystkich katastrof jakie miały miejsce (np. nie było w nich katastrofy samolotu Skyvan który rozbił się, po zrzucie skoczków spadochronowych w Piotrkowie Trybunalskim w 2022 roku czy innego wypadku związanego ze skokami spadochronowymi z pod Częstochowy z 2014 roku). Dodatkowo baza danych z której pobierałem dane miała znaczące ograniczenia, przez co przy próbie scrapowania bez opóźnienia po kilkunastu rekordach otrzymywałem taki komunikat o zablokowaniu mi dostępu do strony:



Komunikat o czasowym zablokowaniu dostępu do bazy danych katastrof

W związku z tym, w przyszłości, można byłoby spróbować znaleźć lepszą bazę danych, która umożliwiałaby scrapowanie danych w szybszy sposób(nie całą noc jak w tym przypadku) i zawierałaby większą ilość danych.

Inne Techniki Modelowania: W analizie skupiłem się głównie na regresji liniowej. W przyszłości można by rozważyć zastosowanie innych technik modelowania, takich jak maszyny wektorów nośnych czy sieci neuronowe, które mogą lepiej radzić sobie z nieliniowymi zależnościami.

Mimo tych potencjalnych możliwości ulepszeń, badanie dostarczyło istotnych wniosków dotyczących czynników wpływających na szanse przeżycia w katastrofach lotniczych.