Aleksandra Mitura
320969
Kierunek studiów: Lotnictwo i Kosmonautyka
Specjalność: Statki Powietrzne

Politechnika Warszawska
Wydział Mechaniczny Energetyki i Lotnictwa

Projektowanie inżynierskie w chmurze

"Analiza i predykcja opóźnień przylotów lotniczych dla lotniska EPWA"

Warszawa 2025/2026

Spis treści

1 Wprowadzenie

Współczesny transport lotniczy stanowi jeden z najbardziej złożonych systemów operacyjnych, w którym punktualność wykonywania operacji lotniczych zależy od wielu wzajemnie powiązanych czynników. Ograniczona przepustowość lotnisk, zmienne warunki meteorologiczne, przeciążenie przestrzeni powietrznej, a także uwarunkowania organizacyjne powodują powstawanie opóźnień, które wpływają zarówno na efektywność przewoźników, jak i komfort pasażerów.
Jednym z kluczowych mechanizmów zarządzania ruchem lotniczym w Europie jest system Air Traffic Flow Management (ATFM), którego zadaniem jest regulowanie przepływu ruchu w sytuacjach, gdy przewidywane zapotrzebowanie przekracza dostępne możliwości operacyjne. W ramach tego systemu opóźnienia przydzielane są już na etapie planowania lotu, aby zapobiegać przeciążeniom sektorów przestrzeni powietrznej oraz portów lotniczych. Informacje o takich opóźnieniach stanowią cenne źródło danych analitycznych, pozwalające na ocenę funkcjonowania systemu zarządzania ruchem lotniczym.
W ostatnich latach dynamiczny rozwój technologii chmurowych oraz narzędzi do przetwarzania danych wielkoskalowych stworzył nowe możliwości w zakresie analizy i prognozowania zjawisk operacyjnych w lotnictwie. Platformy takie jak Amazon Web Services umożliwiają elastyczne przetwarzanie dużych wolumenów danych bez konieczności utrzymywania własnej infrastruktury sprzętowej, natomiast silniki obliczeniowe typu Apache Spark pozwalają na równoległą analizę danych historycznych oraz budowę modeli predykcyjnych.
Realizacja projektu pozwala na połączenie wiedzy z zakresu lotnictwa, analizy danych oraz technologii chmurowych, a także stanowi przykład praktycznego wykorzystania narzędzi Big Data w analizie problemów operacyjnych transportu lotniczego.

2 Cel i zakres projektu

2.1 Cel projektu

Celem projektu było zaprojektowanie i wykonanie kompletnego procesu analizy oraz predykcji opóźnień przylotowych w lotnictwie z wykorzystaniem technologii chmurowych Amazon Web Services oraz silnika przetwarzania Apache Spark. Projekt koncentruje się na analizie danych typu Airport Arrival ATFM Delay, udostępnianych przez EUROCONTROL, które opisują wielkość opóźnień operacji lotniczych wynikających z ograniczeń przepustowości w przestrzeni powietrznej i na lotniskach. Analiza została zawężona do jednego portu lotniczego — Lotniska Chopina w Warszawie (EPWA), co umożliwiło szczegółowe zbadanie charakterystyki opóźnień w ujęciu dziennym, miesięcznym oraz wieloletnim.
Głównym celem analitycznym projektu było:

- zrozumienie struktury i sezonowości opóźnień przylotowych,
- identyfikacja czynników operacyjnych mających największy wpływ na poziom opóźnień,
- budowa modelu regresyjnego umożliwiającego predykcję średniego dziennego opóźnienia,
- przeprowadzenie walidacji jakości predykcji,
- wykonanie scenariuszowej prognozy opóźnień dla roku 2026 na podstawie danych historycznych.

Dodatkowym celem projektu było zdobycie praktycznych kompetencji w zakresie projektowania architektury danych w chmurze, obejmującej pełny cykl życia danych: od warstwy surowej, poprzez przetwarzanie i analizę eksploracyjną, aż po modelowanie predykcyjne i wizualizację wyników.

2.2 Zakres pracy

Zakres pracy obejmuje zaprojektowanie oraz realizację kompletnego procesu analizy danych lotniczych z wykorzystaniem środowiska AWS oraz silnika Apache Spark. W ramach projektu zrealizowano następujące etapy:

1. Pozyskanie i przygotowanie danych wejściowych, wczytanie surowych plików csv oraz ich organizcja w chmurze Amazon S3
2. Projekt architektury warstwowej danych, zgodnej z podejściem Bronze–Silver–Gold
3. Implementacja procesu ETL w środowisku Apache Spark
4. Przeprowadzenie eksploracyjnej analizy danych (EDA)
5. Budowa modelu predykcyjnego o charakterze regresyjnym (Random Forest Regressor)

6. Walidacja modelu predykcyjnego
7. Interpretacja wyników modelu
8. Prognoza opóźnień dla roku 2026

2.2.1 Wykorzystane technologie

W projekcie wykorzystano następujące technologie:

- **Amazon Web Services** - podstawowe środowisko wykonawcze projektu. Platforma ta umożliwia uruchamianie aplikacji analitycznych bez konieczności utrzymywania własnej infrastruktury sprzętowej oraz zapewnia dostęp do szerokiego zestawu usług wspierających przetwarzanie danych.
  - **S3** - centralne repozytorium danych. W S3 przechowywano zarówno dane surowe, jak i dane przetworzone oraz wyniki analiz. Zastosowanie S3 umożliwiło łatwe zarządzanie strukturą danych
  - **EC2** - instancja EC2 została wykorzystana jako środowisko obliczeniowe do uruchamiania Apache Spark. Pozwoliło to na pełną kontrolę nad konfiguracją środowiska, parametrami pamięci oraz sposobem uruchamiania skryptów analitycznych.
  - **Athena** - Athena została wykorzystana jako narzędzie do wykonywania zapytań SQL bezpośrednio na danych zapisanych w Amazon S3. Dzięki temu możliwe było szybkie tworzenie zestawień tabelarycznych oraz weryfikacja wyników uzyskanych w procesach ETL i modelowania.
- **Apache Spark** - stanowił główny silnik przetwarzania danych w projekcie. Jest to platforma umożliwiająca rozproszone przetwarzanie dużych zbiorów danych w pamięci operacyjnej, co znacząco zwiększa wydajność w porównaniu do tradycyjnych narzędzi przetwarzania sekwencyjnego. Do implementacji wykorzystano interfejs PySpark, który umożliwia tworzenie aplikacji Spark w języku Python, zachowując jednocześnie wydajność obliczeń rozproszonych.
- **Spark MLib** - biblioteka została wykorzystana do budowy modelu obliczeniowego.Biblioteka ta udostępnia zestaw algorytmów uczenia maszynowego przystosowanych do pracy na dużych zbiorach danych. W projekcie zastosowano algorytm Random Forest Regressor

Dobór technologii podyktowany był ich powszechnym zastosowaniem w środowiskach przemysłowych oraz wysoką skalowalnością

2.2.2 Architektura rozwiązania

Architektura rozwiązania została zaprojektowana w oparciu o warstwowy model przetwarzania danych typu Bronze–Silver–Gold. Zastosowanie takiego podejścia umożliwiło czytelny podział odpowiedzialności poszczególnych etapów przetwarzania danych oraz ułatwiło kontrolę jakości danych.

- **Warstwa Bronze** – surowe dane wejściowe w formacie CSV,
- **Warstwa Silver** – dane oczyszczone i przetworzone, na tym etapie realizowany był proces ETL. W warstwie Silver wykonano: parsowanie dat i typów liczbowych, usunięcie niepoprawnych rekordów, agregację danych do poziomu dziennego, obliczenie średnich dziennych opóźnień, wyznaczenie liczby operacji i opóźnionych lotów, ograniczenie danych do lotniska EPWA.
- **Warstwa Gold** – dane końcowe, przeznaczone do analizy, modelowania oraz raportowania. W jej skład wchodziły: wyniki eksploracyjnej analizy danych, wykryte anomalie, dane walidacyjne modelu, predykcje dzienne i miesięczne, prognoza opóźnień na rok 2026, wizualizacje wykorzystywane w raporcie.

2.3 Opis danych

Dane wejściowe pozyskane z EUROCONTROL zawierają zbiór opóźnień operacji lotniczych w ramach systemu zarządzania przepływem ruchu lotniczego (ATFM) w sytuacjach, gdy przewidywane zapotrzebowanie na ruch lotniczy przekracza dostępne możliwości operacyjne lotnisk lub sektorów przestrzeni powietrznej. Dane te zostały zebrane dla wszystkich europejskich portów lotniczych realizujących operacje o charakterze rejsowym lub charterowym i raportujących do EUROCONTROL.

Spośród wszystkich dostępnych danych, dane wejściowe obejmowały dzienne dane operacyjne lotnisk europejskich z lat 2014-2025, ograniczone w procesie ETL do lotniska EPWA. Zakres danych został przedstawiony poniżej:

Cecha	Wartość
Lotnisko	EPWA
Typ danych	Arrival ATFM delay
Zakres lat	2014-2025
Liczba dni	4291
Źródło	Eurocontrol

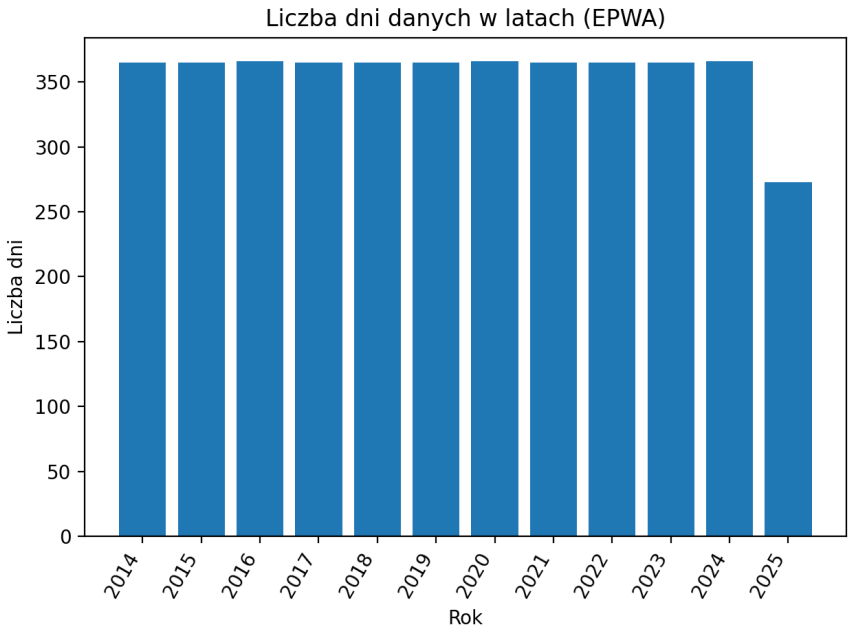
Rys. 1: Zakres danych

Wykorzystane dane zawierają datę operacji, kod oraz nazwę lotniska, liczbę przylotów, liczbę opóźnień powyżej 15 minut oraz minuty opóźnień podzielone według następujących przyczyn:

Column name	Label
DLY_APT_ARR_A_1	A - Accident/Incident
DLY_APT_ARR_C_1	C - ATC Capacity
DLY_APT_ARR_D_1	D - De-icing
DLY_APT_ARR_E_1	E - Equipment (non-ATC)
DLY_APT_ARR_G_1	G - Aerodrome Capacity
DLY_APT_ARR_I_1	I - Industrial Action (ATC)
DLY_APT_ARR_M_1	M - Airspace Management
DLY_APT_ARR_N_1	N - Industrial Action (non-ATC)
DLY_APT_ARR_O_1	O - Other
DLY_APT_ARR_P_1	P - Special Event
DLY_APT_ARR_R_1	R - ATC Routeing
DLY_APT_ARR_S_1	S - ATC Staffing
DLY_APT_ARR_T_1	T - Equipment (ATC)
DLY_APT_ARR_V_1	V - Environmental Issues
DLY_APT_ARR_W_1	W - Weather
DLY_APT_ARR_NA_1	NA - Not specified

Rys. 2: Przyczyny opóźnień

W ramach wykorzystanych dostępne były jedynie częściowe wyniki dotyczące roku 2025. Dla jasności, liczba dostępnych dni danych w latach 2014-2025 została przedstawiona poniżej:



Rys. 3: Liczba danych w latach

3 Proces ETL

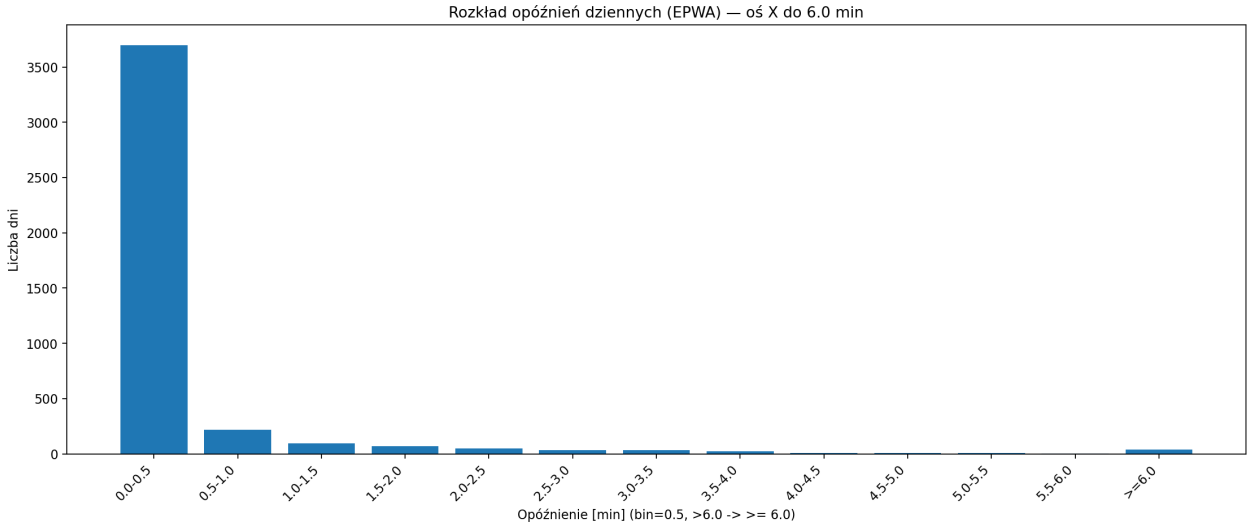
Proces ETL (Extract–Transform–Load) stanowił kluczowy element realizacji projektu, odpowiadając za przygotowanie danych wejściowych do dalszej analizy oraz modelowania predykcyjnego. W projekcie zastosowano architekturę warstwową, w której kolejne etapy przetwarzania danych realizowane były w sposób sekwencyjny i logicznie rozdzielony. Celem procesu ETL było przekształcenie za pomocą skryptu etl.py surowych danych wejściowych do postaci spójnego, ujednoliconego zbioru analitycznego umożliwiającego efektywne przetwarzanie w środowisku Apache Spark.
Etap Extract polegał na wczytaniu surowych plików wejściowych z magazynu danych Amazon S3. Dane źródłowe zapisane były w formacie CSV i podzielone na pliki miesięczne.
W etapie Transform dane zostały oczyszczone, przekształcone oraz wzbogacone o nowe cechy analityczne. Pierwszym krokiem było ograniczenie danych wyłącznie do Lotniska Chopina w Warszawie (EPWA). W kolejnym etapie dokonano konwersji typów danych: daty zostały przekształcone do formatu typu date, wartości liczbowe opisujące liczbę operacji i opóźnień zostały przekonwertowane do typów numerycznych, usunięto rekordy zawierające niekompletne lub błędne dane.Pomimo że dane źródłowe posiadały charakter dzienny, wykonano dodatkową agregację w celu zapewnienia spójności struktury danych. Na tym etapie wyznaczono m.in. łączną liczbę przylotów danego dnia, liczbę lotów objętych opóźnieniem ATFM, całkowity czas opóźnień, średnie dzienne opóźnienie przylotowe. W szczególności obliczona została zmienna:
AVG_ARR_DELAY_MIN, która stanowiła główną zmienną objaśnianą w dalszym modelowaniu predykcyjnym. W celu umożliwienia analizy sezonowości oraz cykliczności ruchu lotniczego dane zostały wzbogacone o dodatkowe cechy czasowe, takie jak: rok, miesiąc, dzień tygodnia, dzień miesiąca, numer tygodnia w roku. Po zakończeniu transformacji dane zostały zapisane do warstwy Silver w formacie Parquet.

4 Analiza opisowa EDA

Analiza eksploracyjna danych (Exploratory Data Analysis – EDA) stanowiła istotny etap projektu, którego celem było wstępne poznanie charakterystyki danych, identyfikacja zależności czasowych oraz wykrycie potencjalnych anomalii. EDA umożliwia zrozumienie struktury danych przed zastosowaniem metod modelowania predykcyjnego oraz pozwala na ocenę ich jakości i kompletności. Analiza została przeprowadzona na danych przetworzonych do warstwy Silver, obejmujących dzienne wartości opóźnień przylotowych dla lotniska EPWA.

4.1 Średnie opóźnienia

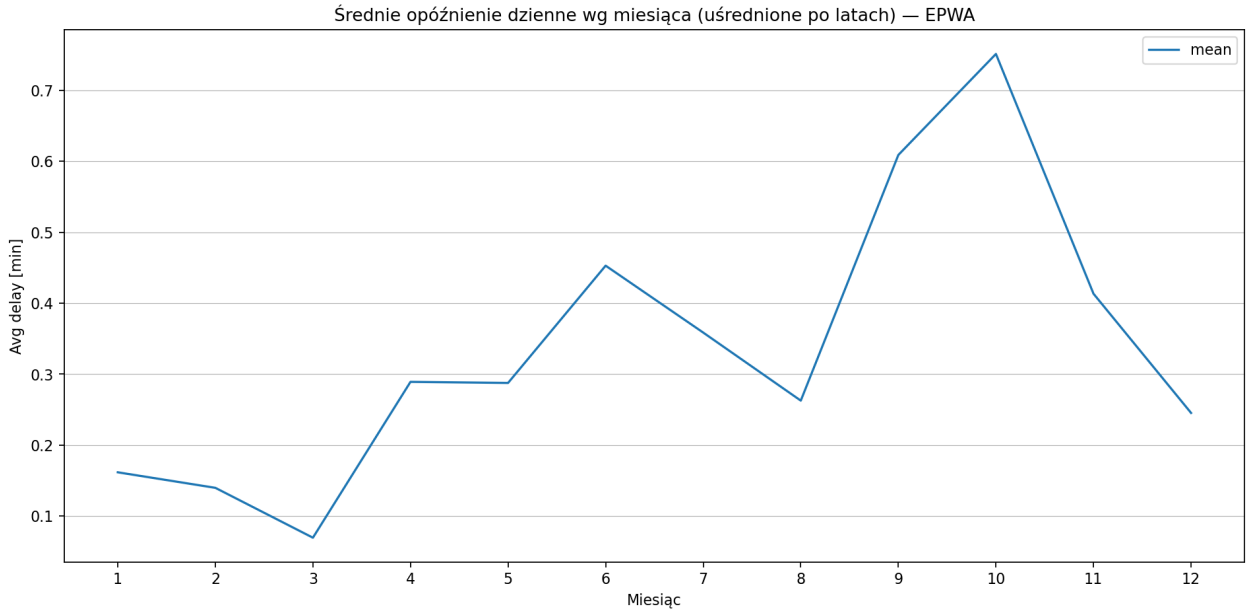
W pierwszym etapie EDA sprawdzono liczbę dostępnych dni operacyjnych w poszczególnych latach. Analiza wykazała, że w zbiorze danych brakuje miesięcy październik - grudzień 2025, ze względu na brak opublikowanych danych dla tych dat.
Kolejnym krokiem była analiza rozkładu dziennych wartości średniego opóźnienia przylotowego.Wyniki przedstawiono na poniższym histogramie:



Rys. 4: Rozkład opóźnień dziennych

Przeprowadzona analiza histogramów wykazała, że zdecydowana większość dni charakteryzuje się niewielkim poziomem opóźnień, a sporadycznie występują dni o znacznie podwyższonym poziomie opóźnień.Taki charakter rozkładu jest typowy dla danych operacyjnych w lotnictwie, gdzie większość dni przebiega w sposób stabilny, natomiast pojedyncze zdarzenia systemowe powodują znaczące odchylenia od wartości przeciętnych.

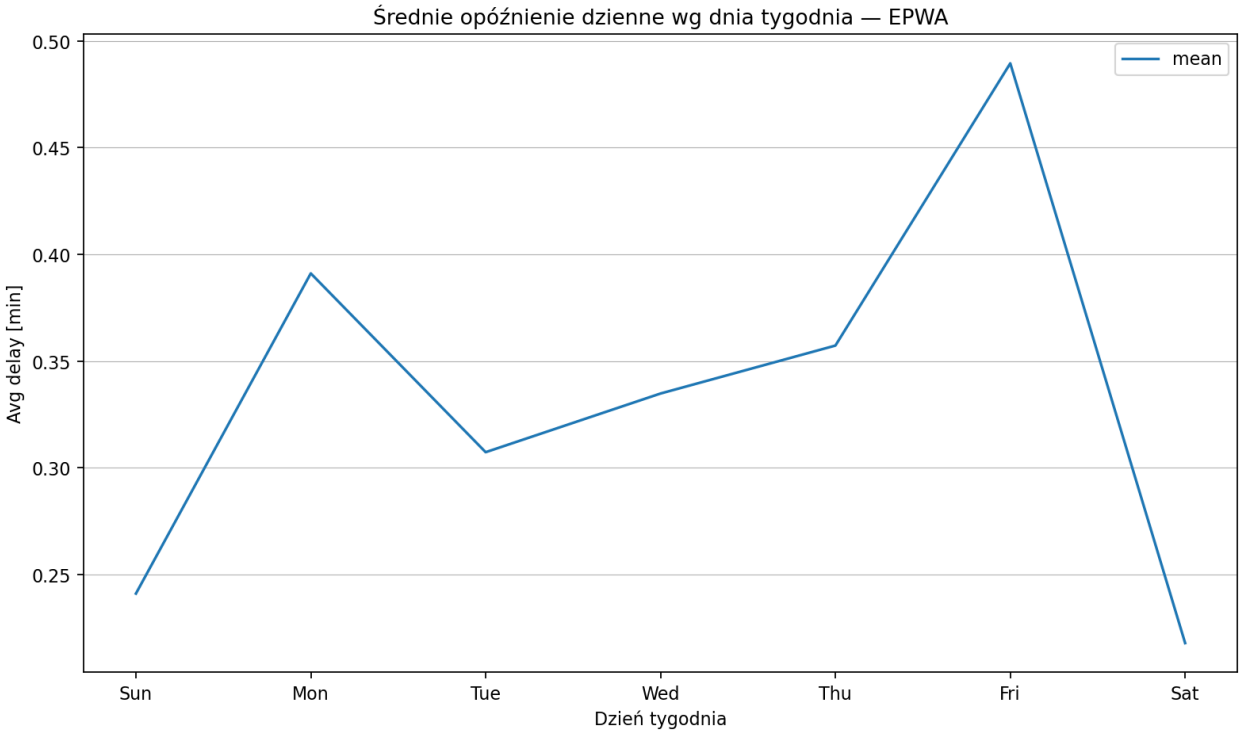
W ramach EDA przeprowadzono analizę średnich opóźnień w podziale na miesiące. Wyniki przedstawiono na poniższym wykresie:



Rys. 5: Średnie opóźnienie według miesięcy

Większe średnie wartości opóźnień odnotowano w sezonie letnim niż zimowym, co może wiązać się ze zwiększonym natężeniem ruchu lotniczego w sezonie wakacyjnym, natomiast peak opóźnień występuje w

października. Może wiązać się to ze zmianą sezonu z letniego na zimowy, gdzie część linii lotniczych wykonuje operacje jeszcze w sezonowym grafiku, a zmniejsza się przepustowość ATC. W tym okresie następuje również zmiana pogody, zatem zwiększają się separacje. Należy pamiętać, że opóźnienia ATFM są przydzielane profilaktycznie, aby zapobiec przeciążeniu systemu.



Rys. 6: Średnie opóźnienie dzienne według dnia tygodnia

Analiza przeprowadzona w podziale na dni tygodnia wykazała istnienie umiarkowanych różnic pomiędzy poszczególnymi dniami.Zaobserwowano tendencję do nieco wyższych opóźnień w dniach roboczych,a stabilniejszych i niższych wartości w weekendy, przy czym największe opóźnienia obserwowano w piątki i poniedziałki. Różnice te nie były jednak na tyle znaczące, aby stanowiły jedyny czynnik decydujący o poziomie opóźnień, lecz potwierdziły zasadność uwzględnienia cech kalendarzowych w modelu predykcyjnym.

4.2 Anomalie

W kolejnym etapie przeprowadzono analizę anomalii, polegającą na identyfikacji dni charakteryzujących się nietypowo wysokimi wartościami średnich opóźnień. Za anomalie uznano dni, w których wartość opóźnienia znacząco odbiegała od rozkładu typowych obserwacji. Zidentyfikowane przypadki mogą odpowiadać: zdarzeniom o charakterze systemowym, istotnym ograniczeniom przepustowości, czy skumulowanym zakłóceniom operacyjnym. Zestawienie dni anormalnych przedstawiono w poniższej tabeli:

date	anomaly_type	severity	avg_arr_delay_min
25.11.2016	TOP_AVG_DELAY	EXTREME	49.13
25.11.2016	TOP_TOTAL_DELAY	EXTREME	49.13
19.12.2016	TOP_AVG_DELAY	HIGH	23.04
19.12.2016	TOP_TOTAL_DELAY	HIGH	23.04
09.11.2018	TOP_AVG_DELAY	HIGH	19.03
09.11.2018	TOP_TOTAL_DELAY	HIGH	19.03
22.09.2023	TOP_TOTAL_DELAY	HIGH	15.82
22.09.2023	TOP_AVG_DELAY	HIGH	15.82
21.09.2023	TOP_AVG_DELAY	HIGH	14.66
21.09.2023	TOP_TOTAL_DELAY	HIGH	14.66
02.10.2024	TOP_AVG_DELAY	HIGH	13.04
02.10.2024	TOP_TOTAL_DELAY	HIGH	13.04
26.09.2024	TOP_AVG_DELAY	HIGH	12.88
26.09.2024	TOP_TOTAL_DELAY	HIGH	12.88
10.10.2024	TOP_AVG_DELAY	HIGH	12.35
10.10.2024	TOP_TOTAL_DELAY	HIGH	12.35
27.12.2024	TOP_AVG_DELAY	HIGH	11.93
27.12.2024	TOP_TOTAL_DELAY	HIGH	11.93
03.02.2018	TOP_AVG_DELAY	HIGH	11.40
03.02.2018	TOP_TOTAL_DELAY	HIGH	11.40

Rys. 7: Dni anormalne

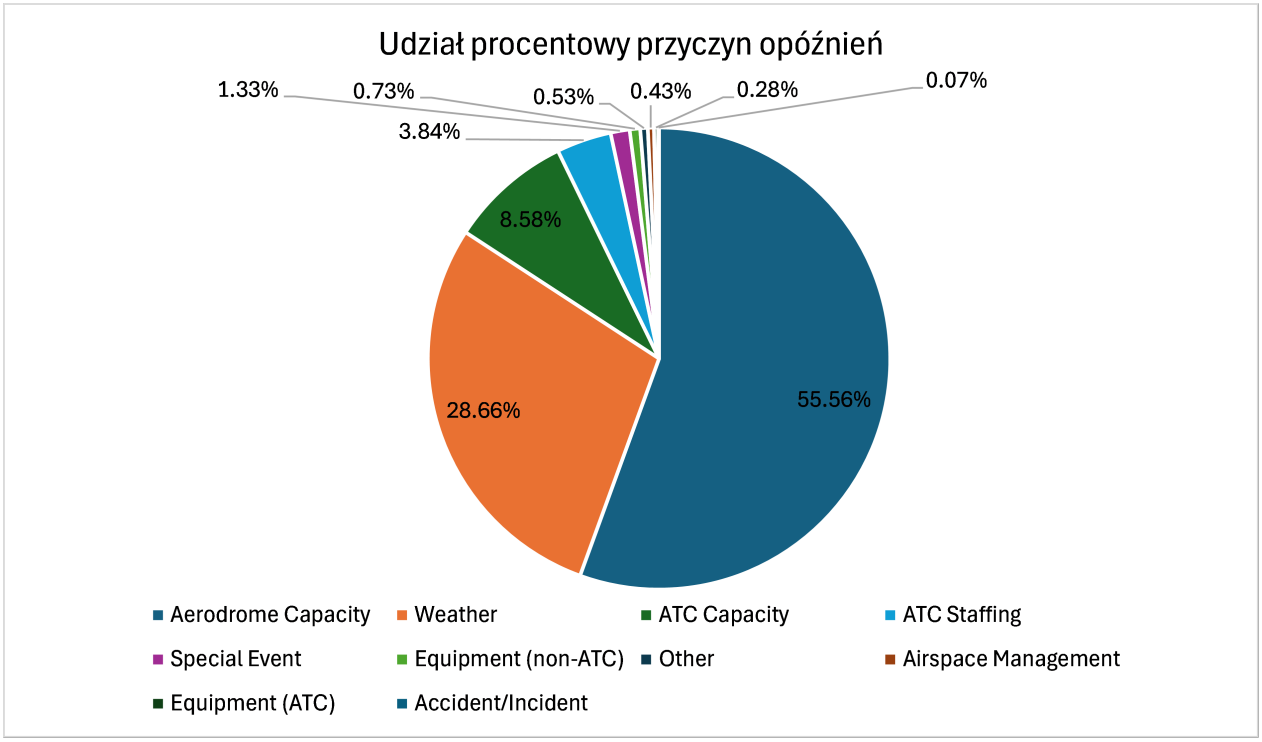
Wyniki sugerują, że w ciągu analizowanych lat nastąpił jeden dzień, gdzie zaobserwowano opóźnienia ponad dwukrotnie wyższe od kolejnych anomalii.

4.3 Przyczyny opóźnień

Istotnym elementem analizy eksploracyjnej była ocena struktury przyczyn opóźnień przylotowych przypisywanych w systemie ATFM. Dane wejściowe zawierały szczegółowy podział opóźnień według kategorii operacyjnych, co umożliwiło identyfikację dominujących źródeł zakłóceń w ruchu lotniczym. Każda kategoria przyczyn reprezentowała określony obszar systemu zarządzania ruchem lotniczym, m.in.:

- ograniczenia przepustowości lotnisk i sektorów przestrzeni powietrznej,
- czynniki operacyjne,
- warunki meteorologiczne,
- zarządzanie ruchem i regulacje sieciowe,
- pozostałe oraz niejednoznacznie sklasyfikowane przyczyny.

W ramach analizy obliczono udział poszczególnych kategorii w całkowitej liczbie minut opóźnień:



Rys. 8: Procentowy udział przyczyn opóźnień

Największy udział w strukturze opóźnień stanowiły przyczyny związane z ograniczeniami przepustowości lotnisk oraz pogoda.
Analiza struktury przyczyn opóźnień pozwoliła również na zaobserwowanie zmian w ich udziale w zależności od pory roku. W okresach zwiększonego natężenia ruchu udział przyczyn operacyjnych był wyraźnie wyższy, natomiast w miesiącach zimowych relatywnie wzrastał udział czynników pogodowych.
Wyniki tej analizy potwierdziły zasadność wykorzystania zmiennych opisujących przyczyny opóźnień jako cech wejściowych w modelu predykcyjnym. Uwzględnienie struktury przyczyn pozwala bowiem na lepsze odwzorowanie złożonych zależności operacyjnych, które nie są widoczne jedynie na podstawie liczby operacji lub wartości opóźnień.

5 Model analityczno–predykcyjny

Celem modelu analityczno-predykcyjnego było przewidywanie średniego dziennego opóźnienia przylotowego ATFM dla lotniska EPWA na podstawie danych historycznych. Model opracowano w środowisku Apache Spark z wykorzystaniem biblioteki Spark MLlib, co umożliwiło skalowalne przetwarzanie danych oraz integrację z wcześniej przygotowanym procesem ETL.

5.1 Zmienna objaśniana

Zmienną objaśnianą (target) modelu było AVG_ARR_DELAY_MIN — średnia wartość opóźnienia przylotowego w danym dniu, wyrażona w minutach.Zmienna ta została wyznaczona w warstwie Silver jako iloraz całkowitego czasu opóźnień oraz liczby opóźnionych operacji danego dnia.

5.2 Zmienne objaśniające

Jako zmienne objaśniające wykorzystano zestaw cech opisujących zarówno intensywność ruchu lotniczego, jak i charakterystykę operacyjną oraz czasową - liczbę przylotów danego dnia, liczbę lotów objętych opóźnieniem ATFM,liczbę lotów opóźnionych powyżej 15 minut, zmienne opisujące strukturę przyczyn opóźnień (kategorie operacyjne), cechy kalendarzowe: (rok, miesiąc, dzień tygodnia, dzień miesiąca, numer tygodnia w roku)

5.3 Wybór algorytmu

Do budowy modelu zastosowano algorytm Random Forest Regressor. Wybór tego algorytmu uzasadniono następującymi cechami:

- zdolnością do modelowania nieliniowych zależności,
- odpornością na obecność wartości odstających,
- stabilnością działania przy danych o zróżnicowanej skali,
- możliwością oceny istotności zmiennych wejściowych,
- dobrą skutecznością w zadaniach regresyjnych o charakterze operacyjnym

Random Forest stanowi zespół drzew decyzyjnych, których wyniki są uśredniane, co pozwala ograniczyć nadmierne dopasowanie modelu do danych uczących.

Element	Opis
Typ modelu	Random Forest Regressor
Dane wejściowe	Dzienne
Zmienna docelowa	AVG_ARR_DELAY_MIN

Rys. 9: Konfiguracja modelu

5.4 Proces uczenia modelu

Preferowanym podejściem było wykorzystanie ostatniego dostępnego roku jako zbioru testowego, natomiast wcześniejsze lata posłużyły do uczenia modelu. W sytuacjach, w których liczba obserwacji była niewystarczająca, stosowano losowy podział danych w proporcji 80/20.
Uczenie modelu zostało przeprowadzone w ramach pipeline Spark ML, obejmującego transformację danych wejściowych do wektora cech, trenowanie algorytmu Random Forest Regressor i generowanie predykcji dla zbioru testowego.

5.5 Wyniki predykcji

Po zakończeniu uczenia model wygenerował predykcje średniego dziennego opóźnienia przylotowego. Wyniki zostały zapisane w warstwie Gold i stanowiły podstawę do dalszej walidacji oraz analizy jakości predykcji.

6 Analiza wyników modelu

6.1 Ocena jakości modelu

Do oceny skuteczności modelu zastosowano następujące metryki:

- RMSE – miara średniego błędu predykcji,
- R^2 miara stopnia wyjaśnienia zmienności danych przez model

Uzyskane wyniki przedstawiono w tabeli poniżej:

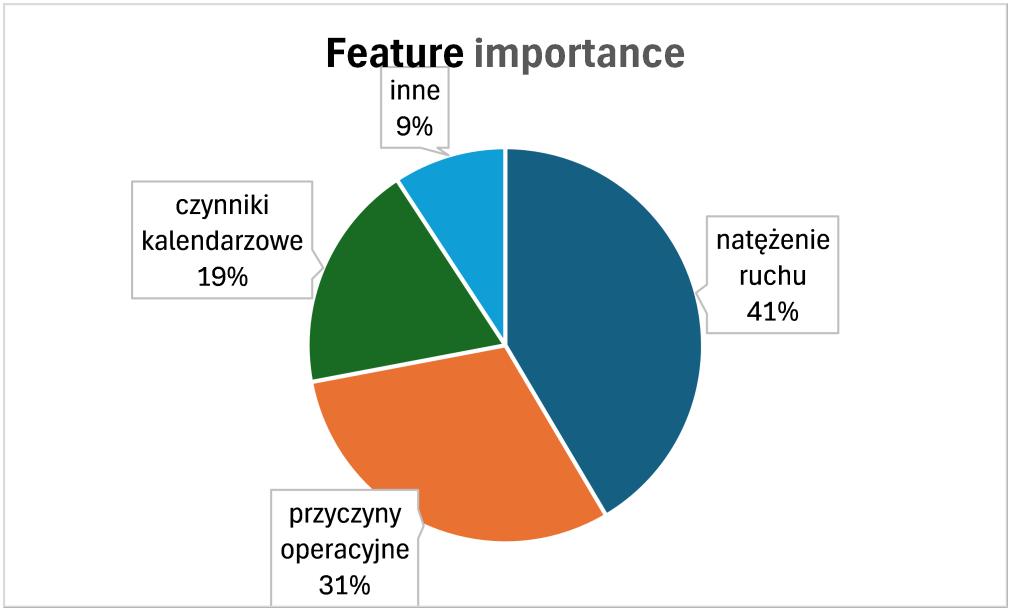
Wskazują one na dobrą zdolność modelu do odwzorowania zależności pomiędzy danymi wejściowymi a poziomem opóźnień. W szczególności wysoka wartość współczynnika R^2 potwierdza, że model wyjaśnia znaczną część zmienności danych. Warto zauważyć, że błędy predykcji są nieuniknione w przypadku danych operacyjnych o charakterze niestacjonarnym i podatnych na zdarzenia losowe. Pomimo tego uzyskane wyniki potwierdzają użyteczność modelu do analizy trendów i sezonowości.

RMSE	0.348
R^2	0.918

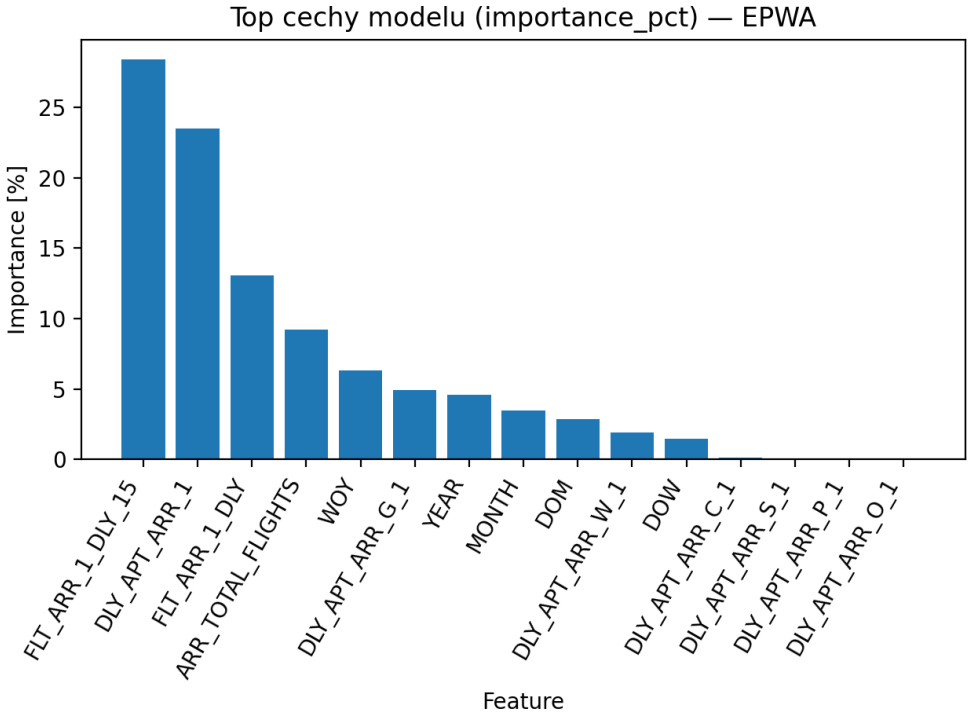
Rys. 10: Metryki modelu

6.2 Interpretacja modelu

W celu zwiększenia przejrzystości działania modelu przeprowadzono analizę istotności cech (feature importance), udostępnianą przez algorytm Random Forest. Miara ta pozwala określić względny wkład poszczególnych zmiennych wejściowych w proces podejmowania decyzji przez model. Poniżej przedstawiono rozkład czynników mających największy wpływ na prognozowanie opóźnień oraz cechy modelu, które w największym stopniu te prognozy generowały:



Rys. 11: Feature importance

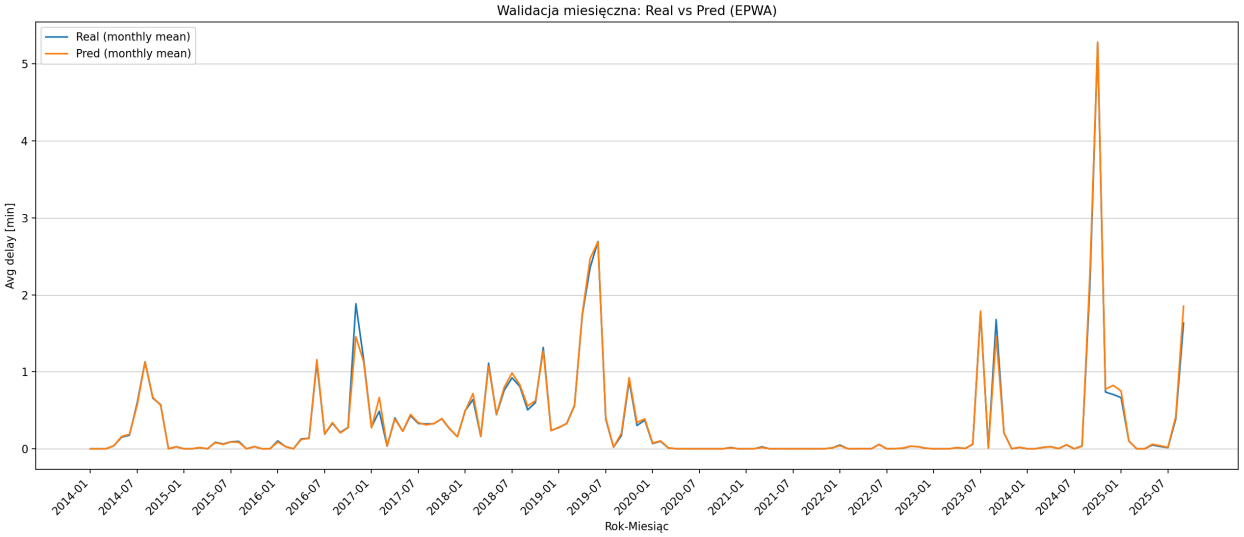


Rys. 12: Top cechy modelu

Według analizy, największy wpływ na prognozy opóźnień miało natężenie ruchu. Ważnymi zmiennymi są również przyczyny operacyjne oraz czynniki kalendarzowe.
Analiza top 10 cech wykazała, że największy wpływ na poziom prognozowanych opóźnień miały zmienne opisujące liczbę lotów objętych opóźnieniem ATFM, zmienne związane z intensywnością ruchu lotniczego, wybrane kategorie przyczyn opóźnień operacyjnych, cechy kalendarzowe, w szczególności miesiąc oraz dzień tygodnia.
Wyniki te są spójne z wnioskami uzyskanymi na etapie analizy eksploracyjnej danych, gdzie wykazano wyraźną sezonowość oraz istotny wpływ struktury operacyjnej na poziom opóźnień. Interpretacja modelu potwierdziła, że algorytm nie opiera się wyłącznie na zmiennych czasowych, lecz uwzględnia także rzeczywistą charakterystykę operacyjną ruchu lotniczego.

6.3 Walidacja modelu

Walidacja modelu została przeprowadzona w celu oceny jego stabilności oraz zdolności do generalizacji na dane nienależące do zbioru uczącego. Poniżej przedstawiono wyniki walidacji w ujęciu miesięcznym, polegającą na agregacji predykcji i wartości rzeczywistych:



Rys. 13: Walidacja miesięczna modelu

Porównanie średnich miesięcznych wykazało bardzo dobrą zgodność trendów pomiędzy wartościami rzeczywistymi i prognozowanymi. Model poprawnie odtwarzał sezonowość roczną, relacje pomiędzy miesiącami o wysokim i niskim natężeniu opóźnień oraz ogólny poziom opóźnień w poszczególnych latach.

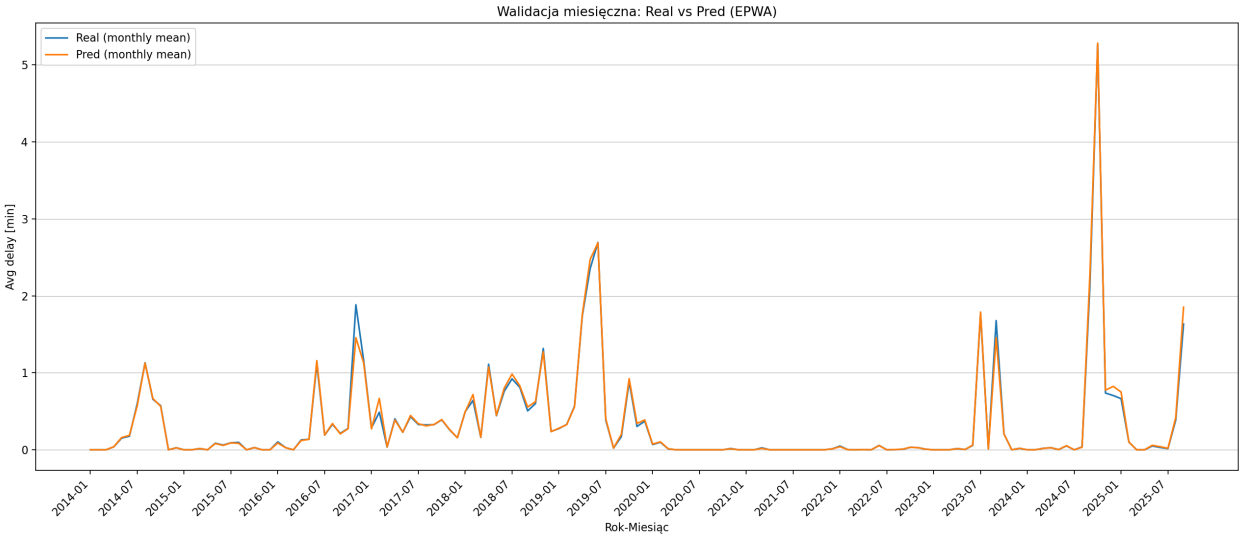
7 Prognoza opóźnień

Ostatnim etapem projektu była wykonana scenariuszowa prognoza opóźnień przylotowych typu ATFM dla lotniska Chopina w Warszawie (EPWA). Celem tego etapu było oszacowanie możliwego poziomu opóźnień w przyszłym okresie na podstawie historycznych danych operacyjnych z lat 2014-2025 oraz zależności zidentyfikowanych przez model predykcyjny.
Prognoza stanowiła analizę długookresową, umożliwiającą ocenę potencjalnych trendów w funkcjonowaniu systemu zarządzania ruchem lotniczym.
Prognoza opóźnień została wykonana przy następujących założeniach:

- zachowanie struktury operacyjnej systemu ATFM z lat historycznych,
- brak istotnych zmian infrastrukturalnych na lotnisku,
- brak uwzględnienia zdarzeń losowych o charakterze nadzwyczajnym,
- stabilność relacji pomiędzy natężeniem ruchu a poziomem opóźnień,
- wykorzystanie danych z lat 2022-2025 (względna odbudowa systemu po okresie pandemii COVID-19)

Proces prognozowania oparto na wcześniej wytrenowanym modelu regresyjnym Random Forest. W pierwszym kroku wyznaczono średnie wartości cech wejściowych obliczone na podstawie danych z lat 2022–2025. Następnie wygenerowano kalendarz dni dla roku 2026 i połączono go z wyznaczonymi średnimi wartościami. Tak przygotowany zbiór danych został wykorzystany jako wejście do modelu, który wygenerował dzienne prognozy średniego opóźnienia przylotowego. Wyniki zostały następnie zagregowane do poziomu miesięcznego w celu ułatwienia interpretacji oraz porównań historycznych.

Agregacja wyników do poziomu miesięcznego pozwoliła na porównanie prognozowanego roku 2026 z wybranymi latami historycznymi, w szczególności z rokiem 2014 oraz 2024:



Rys. 14: Porównanie predykcji średnich wartości opóźnień

Porównanie to wykazało, że ogólny poziom opóźnień w 2026 roku mieści się w zakresie obserwowanym w danych historycznych, natomiast sezonowość miesięczna pozostaje wyraźna. Miesiące letnie nadal charakteryzują się największym ryzykiem opóźnień. Prognoza nie wskazuje na gwałtowny wzrost średnich opóźnień, lecz raczej na utrzymanie trendu obserwowanego w ostatnich latach.

8 Wnioski

Przeprowadzona analiza eksploracyjna wykazała wyraźną sezonowość opóźnień przylotowych oraz istotny wpływ natężenia ruchu lotniczego na poziom opóźnień. Najwyższe wartości średnich opóźnień występowały w okresie letnim oraz w miesiącach przejściowych pomiędzy sezonami rozkładowymi. Jednocześnie większość dni charakteryzowała się niskim poziomem opóźnień, a znaczną część całkowitych zakłóceń generowały pojedyncze dni anormalne. Analiza struktury przyczyn opóźnień potwierdziła dominujący wpływ ograniczeń przepustowości lotnisk, czynników pogodowych oraz regulacji operacyjnych. Wyniki te uzasadniły wykorzystanie zmiennych operacyjnych i kalendarzowych jako cech wejściowych w modelu predykcyjnym. Zastosowany model regresyjny Random Forest wykazał dobrą zdolność odwzorowania trendów i sezonowości danych historycznych. Walidacja miesięczna potwierdziła wysoką zgodność wartości rzeczywistych i prognozowanych, co wskazuje na poprawną generalizację modelu. Wykonana prognoza dla roku 2026 wskazuje na utrzymanie poziomu opóźnień w zakresie obserwowanym w danych historycznych, przy zachowaniu sezonowości rocznej. Opracowana architektura chmurowa oraz metodologia analizy mogą stanowić podstawę do dalszego rozwoju systemów analitycznych wspomagających zarządzanie ruchem lotniczym.

9 Źródła

- https://ansperformance.eu/data/
- https://docs.aws.amazon.com/
- https://www.datacamp.com/tutorial/aws-s3-efs-tutorial
- https://medium.com/@otengcode/building-modern-data-lakes-on-aws-s3-with-the-medallion-architecture-baa6cb9d028a
- https://www.kerno.io/aws/ec2
- https://matplotlib.org/stable/tutorials/pyplot.html
- https://docs.github.com/en/get-started/start-your-journey/hello-world
- https://www.geeksforgeeks.org/machine-learning/random-forest-regression-in-python/