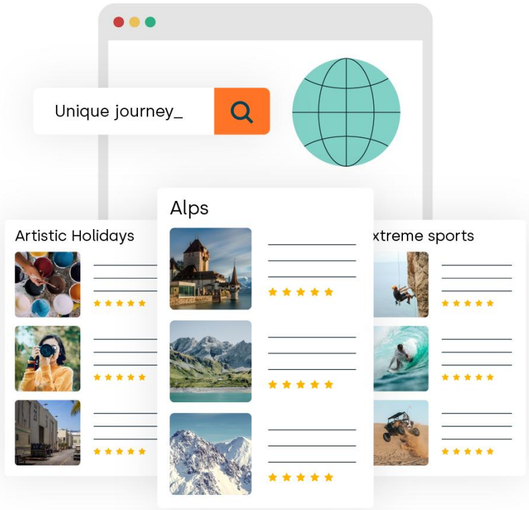
Qtravel.ai

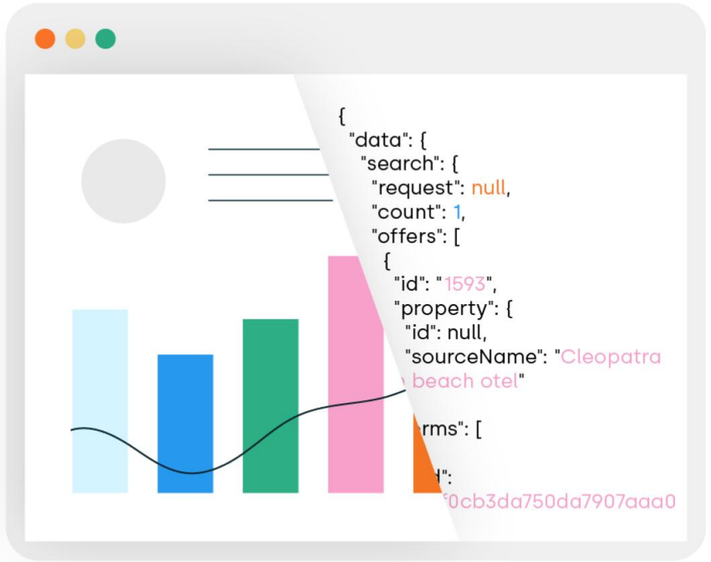
Nowoczesne technologie w analizie danych turystycznych.

Jak wykorzystać narzędzia Big Data do lepszego zrozumienia potrzeb uczestników rynku turystycznego?

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Narodowe Centrum Badań i Rozwoju
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Kim jesteśmy?





Qtravel.pl:

- **14 lat doświadczenia** jako agent turystyczny w Polsce (TUI, Itaka, Rainbow, Coral Travel i inni)
- **2010** – wyszukiwarka wycieczek analogiczna do wyszukiwarki Google

Qtravel.ai:

- **2020** – 2 granty z NCBiR na prace badawczo-rozwojowe w zakresie budowy narzędzi dla turystyki opartych na AI
- **2024** – narzędzia SaaS dla turystyki do wyszukiwania w jęz. naturalnym

Projekty badawczo-rozwojowe

Inteligentny system wyszukiwania podróży oparty na algorytmach rozumienia języka naturalnego

- **wartość projektu:** 4,4 mln zł
- **cel:** budowa systemu do wyszukiwania w danych imprez turystycznych wykorzystująca narzedzia Sztucznej Inteligencji (uczenie maszynowe, uczenie głębokie, przetwarzanie języka naturalnego)
- budowa architektury systemu do przetwarzania i analizy danych tekstowych

Opracowanie inteligentnego systemu predykcyjnego dla branży turystycznej wykorzystującego zaawansowane metody wielowymiarowej fuzji danych i uczenia maszynowego

- **wartość projektu:** 3,65 mln zł
- **cel:**
  (1) budowa systemu do analizy danych historycznych dla rynku turystyki wyjazdowej
  (2) Budowa systemu do przewidywania trendów i cen ofert turystycznych
- budowa architektury systemu do przetwarzania dużych zbiorów danych (Big Data)

Potrzeby uczestników rynku turystycznego

Turyści

Zakup oferty tur. dostosowanej do potrzeb
i w optymalnej cenie nie wymagający analizy setki stron internetowych

- dostęp do narzędzi podpowiadających najlepsze dni/terminy zakupu oferty
- dostęp do narzędzi prognozujących ceny i trendy cenowe
- predykcja cen na poziomie oferty tur. (np.: jak wzrośnie/spadnie cena do danego hotelu)

Dostawcy

Kształtowanie polityki cenowej i ofertowej na podstawie aktualnej wiedzy na temat rynku tur. :

- dostęp do danych historycznych z jak najmniejszym opóźnieniem
- analiza danych tur. w ujęciu historycznym
- analiza cen i trendów cenowych w ujęciu historycznym
- wykrywanie trendów w danych (np.: wzrost zainteresowania daną destynacją)
- predykcja cen (trendy krótko, długookresowe)

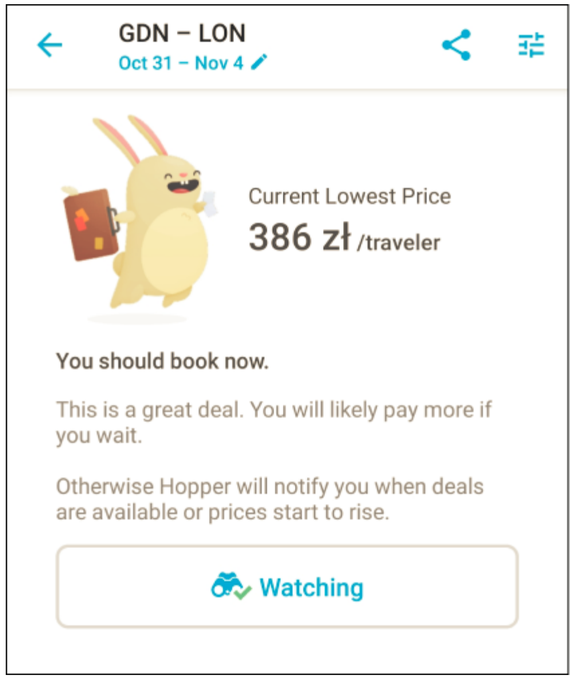
Inni uczestnicy

Analiza rynku turystycznego (organizacje, turystyczne, media branżowe, agenci i pośrednicy tur.)

- dostęp do danych historycznych z jak najmniejszym opóźnieniem
- zaawansowana analiza trendów na rynku tur.
- możliwość porównywania danych tur. z danymi z innymi danymi (np.: ekonomicznymi, gospodarczymi, itp.)

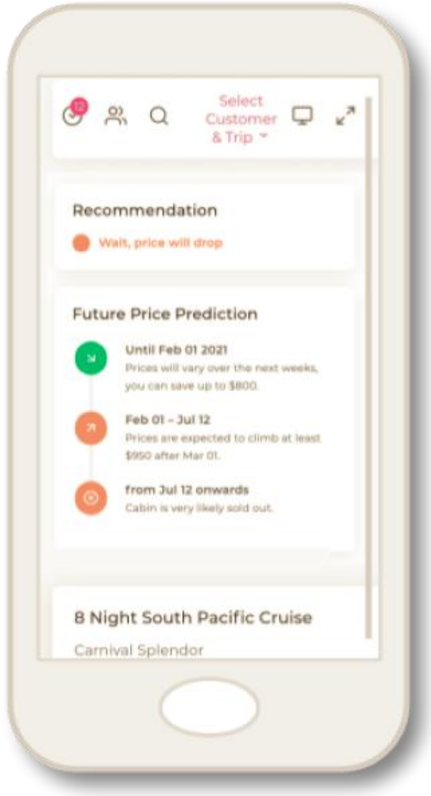
Analiza danych i systemy predykcji cen (rozwiązania komercyjne)

hopper



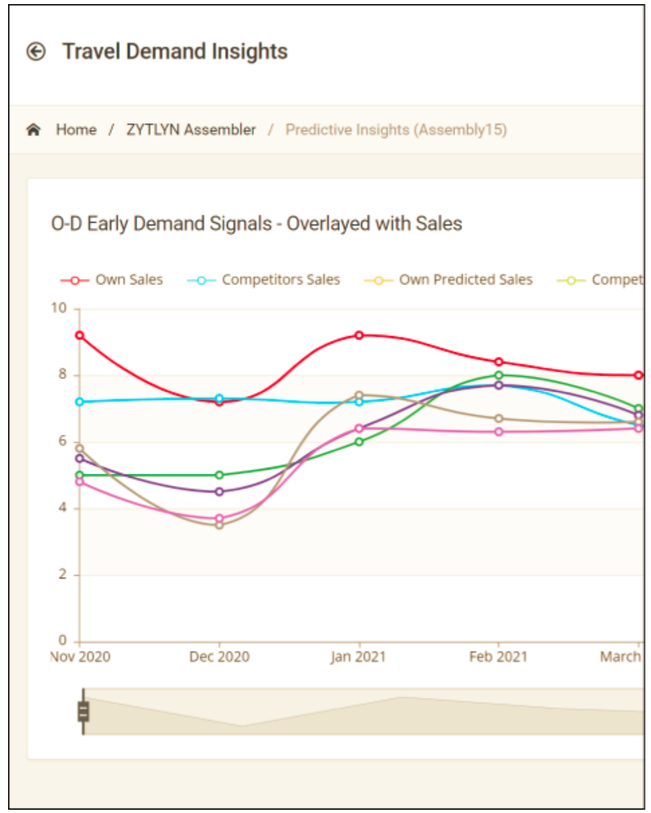
Aplikacja mobilna do predykcji cen biletów lotniczych

cruisewatch



Narzędzie do przewidywania cen rejsów

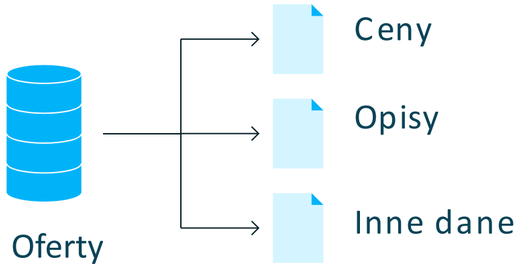
ZYTLYN



Platforma do analizy danych turystycznych i predykcji trendów

Platforma do analizy danych turystycznych

Dostawcy usług



Inne źródła danych

- Dane geolokalizacyjne
- Dane baz geograficznych
- Dane pogodowe
- Dane ekonomiczne
- Dane turystyczne
- Dane tekstowe (np.: Wikipedia)

Platforma AI & Big Data

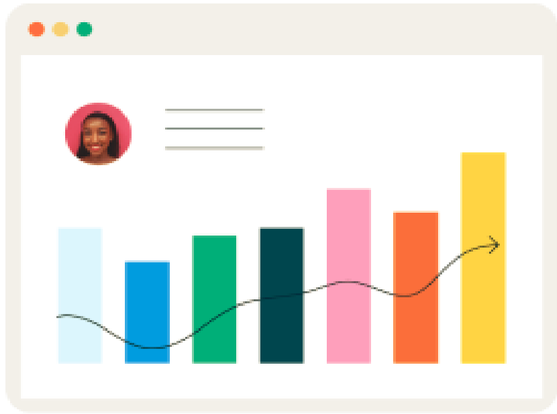
ezioro danyc
Data Lake

XML,
CSV,
JSON,
TXT,
PDF,
inne

- Analityka danych
- Uczenie maszynowe
- Uczenie głębokie (Sieci neuronowe)
- Przetwarzanie i rozumienie języka naturalnego (NLP/NLU)

Narzędzia Big Data
Apache Spark, Apache Spark Streaming, Hadoop, HDFS, Airflow, Apache Kafka, Clickhouse, ElasticSearch, ….

Dashboard do analizy i predykcji danych tur.



- Statystyki wyszukiwań klientów wyszukiwarki w języku naturalnym
- Trendy w wyszukiwaniach
- Statystyki ofert turystycznych na podstawie danych historycznych (np.: minimalne, średnie, maksymalne ceny ofert)
- Predykcje cen i trendów cenowych

z opóźnieniem maksymalnie 1 dnia

Qtravel.ai

Budowa platformy Big Data

Jaki jest cel budowy platformy?

2 cele: **analiza danych historycznych** i **predykcja cen/trendów cen** w horyzoncie około 1 miesięca

Jakie dane chcemy analizować i skąd je pozyskiwać?

Rynek wycieczek zorganizowanych
Dane pozyskiwane od dostawców tur. (touroperatorów)
Dane zewnętrzne istotne do budowania modeli predykcyjnych

Jakie raporty/wizualizacje/dane nas interesują?

Specyfikacja około 20 typów raportów z prezentacją danych turystycznych, np.:
„Minimalna/Maksymalna/średnia cena za wycieczkę z Gdańska w obserwowana w okresie ostatnich 2 tygodni (01.12.2024-13.02.2024) z datami wyjazdu w kolejnych tygodniach /miesiącach”

Analiza danych historycznych

Dostawca (touroperator)	Wolumen Cen	Odsetek zmian cen w ciągu dnia	Wolumen cen po 365 dniach	Wolumen po 2 latach
Duży	85 000 000	1%	395 250 000	790 500 000
Duży	85 000 000	10%	3 187 500 000	6 375 000 000 (6,3 mld cen)
Mały	3 000 000	1%	13 950 000	27 900 000
Mały	3 000 000	10%	112 500 000	225 000 000
50 małych + 1 duży	235 000 000	1%	945 750 000	1 891 500 000
50 małych + 1 duży	235 000 000	10%	8 666 500 000	17 331 000 000 (17 mld cen)

- Statystyki cen z okresu realizacji projektu (2021r. ,covid)
- Analiza danych w tabeli zakłada przechowywanie tylko zmienionych cen
- Rozmiar danych w 2024r per tourooperator wzrósł kilkakrotnie (duży TO – 280 mln cen)
- Architektura systemu budowana w latach 2020-2023 musiała uwzględniać ogromne przyrosty danych r.d.r (100% - 300%)
- Konieczne uwzględnienie ograniczenia danych

Analiza danych historycznych

Przechowywanie danych

- zastosowanie tzw. architektury medalionu (dane przechowywanie w 3 warstwach):
  - **warstwa brązowa :**
    - dane surowe, oryginalne
    - formaty XML, CSV, JSON i inne
    - przechowywane w chmurze, skompresowane
  - **warstwa srebrna:**
    - dane oczyszczone, przetworzone, zwalidowane, znormalizowane, gotowe do dalszych operacji (np.: jako dane wejściowe do modeli predykcyjnych)
    - specjalne formaty plików – AVRO, Parquet
    - przechowywane w klastrze Big Data (Hadoop, HDFS)
    - do przetwarzania danych między warstwami wykorzystywany jest Apache Spark
  - **warstwa złota:**
    - kolejna warstwa przetworzonych danych, przechowywane w sposób optymalny do wykorzystania w celach biznesowych (na przykład do prezentacji w raportach/wizualizacjach)
    - przechowywane w bazie danych typu OLAP (*online Analitical processing*) zoptymalizowanej pod kątek generowania raportów analitycznych (baza danych ClickHouse)
    - istotny czas dostępny do danych (zapytanie SQL powinno być szybkie, liczone w milisekundach)

Predykcja cen turystycznych

Analiza czynników wpływających na cenę oferty tur. (pakietów turystycznych)

- Założenie do predykcji cen turystycznych:
  - **opracowanie modelu, który uwzględnia czynniki zewnętrzne – społeczno-ekonomiczne, polityczno-prawne, geopolityczne, przyrodnicze, kulturowe, technologiczne**
- Identyfikacja 15 kluczowych kategorii pakietów turystycznych
  - podział na podstawie typu transportu (samolot, autokar, dojazd własny) i rodzaju wycieczki (wypoczynek, zwiedzanie, city break, zwiedzanie + wypoczynek, narty, rejsy)
- opracowanie matrycy czynników cenotwórcznych i pakietów:
  - lista czynników cenotwórczych
  - czy dany czynnik występuje dla danego pakietu turystycznego
  - wagi cząstkowei końcowe poszczególnych czynników
  - kierunek oddziaływania na cenę (wzrost/spadek)
  - źródła danych
- Podział czynników na 3 poziomy:
  - mikro – dotyczące wyłącznie cech usług/usługodawców włączonych do danego pakietu
  - mezo – dotyczące otoczenia regionalnego
  - makro – dotyczące zjawisk globalnych (np.: ceny paliw, wskaźniki koniunktury)

Predykcja cen turystycznych

l.p.		01.1	01.2	01.3	02.1	02.2	02.3	02.4	03.1	03.2	04.1	04.2	04.3	05.1	05.2	05.3	Źródło danych
1	ceny paliw	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	OPEC
2	kursy walut w okresie t	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	E/MFW OECD
3	kursy walut w okresie t-1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	j.w.
4	koszty transportu (ceny biletów lotniczych)	✓	✓		✓		✓		✓				✓			✓	j.w.
5	PKB/1 mieszk. w kraju wysyłającym	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	j.w.
6	stopa bezrobocia w kraju wysyłającym	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	j.w.
7	stopa oszczędności w kraju wysyłającym	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	j.w.
8	poziom inflacji w kraju wysyłającym	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	j.w.
9	poziom inflacji w kraju docelowym	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	j.w.
10	zmiany poziomu cen (CPI) w kraju docelowym	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	j.w.

Poziom mezo – dotyczący bezpośredniego otoczenia miejsca realizacji usług zawartych w pakiecie turystycznym

1	wskaźnik atrakcyjności kulturalnej uwzględniający: (1) szczególne wydarzenia przyciągające turystów (sportowe, kulturalne, polityczne); (2) imprezy kulturalne o charakterze cyklicznym; (3) zabytki i muzea na liście 100 najczęściej odwiedzanych; (4) obecność obiektów z listy światowego dziedzictwa UNESCO.	Wskaźnik atrakcyjności kulturalnej przyjmuje postać gradacji w skali 0-4: jeżeli dla oferty nie wskazano żadnych atrakcji wskaźnik=0; jeżeli 1 atrakcja wskaźnik=1 itd., dla 4 i więcej atrakcji wskaźnik=4	Lista A Lista B Lista C Lista D
2	wielkość ruchu turystycznego w stosunku do 1. mieszkańców danego państwa		number of arrivals | Data
3	wskaźnik atrakcyjności środowiskowej uwzględniający: (1) dostępność morza; (2) dostępność akwenów do uprawiania sportów wodnych; (3) występowanie górskich szlaków turystycznych; (4) występowanie obszarów chronionych (parków narodowych i krajobrazowych, rezerwatów, obszarów Natura 2000); (5) atrakcyjność klimatu (temperatura powietrza, wód morskich, liczba dni słonecznych).	wskaźnik atrakcyjności środowiskowej przyjmuje postać gradacji w skali 0-5: jeżeli dla oferty nie wskazano żadnych atrakcji wskaźnik=0; jeżeli 1 atrakcja wskaźnik=1 itd., dla 5 i więcej atrakcji wskaźnik=5.	pogoda
4	średnia temperatura wody	skala 0-2 (0-niekorzystna, 1- korzystna, 2-b. korzystna)	pogoda
5	średnia liczba dni słonecznych	skala 0-2 (0-niekorzystna, 1- korzystna, 2-b. korzystna)	pogoda
6	średnia temperatura powietrza	skala 0-2 (0-niekorzystna, 1- korzystna, 2-b. korzystna)	pogoda
7	dbałość o zrównoważony rozwój środowiska	0/1	raport

•Zdefiniowaliśmy:

- 61 determinant na poziomie mikro
- 13 determinant na poziomie mezzo
- 16 determinant na poziomie makro

• Niektóre determinanty były złożone i składały się z wielu wskaźników cząstkowych, np.: wskaźnik atrakcyjności kulturalnej czy wskaźnik atrakcyjności środowiskowej

• Do dalszych prac wybraliśmy:

- 31 determinant mikro
- 9 determinant mezzo
- 9 determinant makro

• Ze względu na różnorodność determinant dla różnych pakietów skupiono się tylko na **pakietach wypoczynkowych** z różnymi formami transportu

travel.ai

Predykcja cen turystycznych

Problemy

- **Źródła danych**
  - proces znalezienia i weryfikacji źródeł danych dla około 50 determinant cen był **czasochłonny** i **skomplikowany**
  - niektóre źródła danych zapewniały tylko **dane dla części państw** (np.: Eurostat udostępnia dane tylko dla krajów UE, dane dla krajów innych niż UE wymagały dostępu do innych baz danych)
  - niektóre determinanty były zdefiniowane zbyt ogólnie:
    - zagrożenia wystąpieniem katastrof naturalnych
    - zagrożenia zdrowotne
    - zagrożenia atakami terrorystycznymi lub konfliktami zbrojnymi
- **Formaty danych:**
  - często dostęp tylko do danych w formacie XLS, PDF
  - brak API
  - ręczne uzupełnianie danych dla niektórych determinant
  - analiza danych tekstowych (np.: Wikipedia na potrzeby danych o wydarzeniach)
- **Aktualność danych**
  - dane często dostępne z opóźnieniem (miesięcznym, kwartalnym lub rocznym)
  - dla modeli predykcyjnych istotne okazały się cechy (czynniki) zmienne w czasie, które wpływają na zmianę ceny wycieczki w przyszłości
- **Koszt pozyskania danych**
  - dane open source vs dane udostępniane przez firmy komercyjne
  - czas opracowywania danych open source vs dostęp do zweryfikowanych baz danych komercyjnych (np.: bazy danych ekonomicznych, dane pogodowe)
  - dane komercyjne zawierają predykcje (np.: prognozy pogody)
  - koszt niektórych danych komercyjnych przekraczał możliwości projektu:
    - **$15K USD / rok** za dostęp do danych (live) o nieprzewidywalnych wydarzeniach dla 1 państwa

Predykcja cen turystycznych

Statystyki z ULC dotyczące lotnisk i liczby pasażerów dostępne tylko w PDF

Liczba pasażerów obsłużonych w polskich portach lotniczych według miast w międzynarodowym ruchu czarterowym w drugim kwartale 2023 i 2024 roku						
Miasto	2024 l.p.	2024 liczba pasażerów	2024 udział	2023 l.p.	2023 liczba pasażerów	2023 udział
Antalya	1	688 603	28,32%	1	525 570	28,50%
Hurghada	2	181 134	7,45%	2	116 659	6,33%
Marsa Alam	3	124 484	5,12%	4	89 593	4,86%
Heraklion	4	122 119	5,02%	5	85 168	4,62%
Rodos	5	101 660	4,18%	3	93 082	5,05%
Bodrum	6	97 485	4,01%	6	73 901	4,01%
Kos	7	76 399	3,14%	10	53 208	2,88%
Zakintos	8	74 037	3,04%	7	67 614	3,67%
Palma de Mallorca	9	72 528	2,98%	8	57 774	3,13%
Szarm el-Szejk	10	67 420	2,77%	21	21 983	1,19%
Djerba	11	60 493	2,49%	11	41 137	2,23%
Tirana	12	52 112	2,14%	13	35 640	1,93%
Izmir	13	49 337	2,03%	12	39 722	2,15%
Chania	14	48 944	2,01%	15	31 418	1,70%
Enfidha	15	48 003	1,97%	20	22 066	1,20%
Burgas	16	42 612	1,75%	9	55 305	3,00%
Korfu	17	40 523	1,67%	14	33 210	1,80%
Monastir	18	38 085	1,57%	16	29 702	1,61%
Dalaman	19	31 318	1,29%	22	19 777	1,07%
Larnaka	20	27 947	1,15%	17	25 973	1,41%

Wskaźniki Infor RISK (np.: zagrożenia naturalne) – format XLS, publikacja raz na rok

INFORM RISK

COUNTRY (a-z)	ISO3 (a-z)	INFORM RISK (0-10)	RISK CLASS (Very Low	Rank (1-191)	Lack of Reliability (*) (0-10)	HAZARD & EXPOSURE (0-10)	Natural (0-10)	Earthquake (0-10)	River Flood (0-10)	Tsunami (0-10)
Panama	PAN	3,6	Medium	86	2,9	3,0	5,1	7,8	0,2	8,5
Papua New Guinea	PNG	5,7	High	27	3,4	4,0	6,1	9,2	4,6	7,1
Paraguay	PRY	2,5	Low	136	2,9	1,4	2,5	0,1	5,6	0,0
Peru	PER	4,9	Medium	42	1,7	5,1	6,4	9,6	6,5	9,1
Philippines	PHL	5,4	High	35	2,1	8,4	8,3	9,7	6,7	9,4
Poland	POL	2,5	Low	136	4,1	1,5	2,8	0,8	5,9	0,0
Portugal	PRT	1,9	Very Low	169	3,2	1,8	3,2	3,4	3,8	4,3
Qatar	QAT	1,4	Very Low	186	4,6	1,0	1,9	0,1	0,0	0,0
Romania	ROU	2,5	Low	136	3,3	2,0	3,6	6,1	6,1	0,0
Russian Federation	RUS	5,1	High	39	5,0	7,5	5,2	4,2	8,4	4,2
Rwanda	RWA	3,5	Medium	89	0,8	1,6	2,8	4,0	2,7	0,0
Saint Kitts and Nevis	KNA	1,8	Very Low	173	6,7	1,6	2,9	3,6	0,0	0,0
Saint Lucia	LCA	2,7	Low	126	7,2	1,9	3,5	4,4	0,0	0,0
Saint Vincent and the Grenadines	VCT	2,4	Low	144	7,1	1,8	3,3	4,8	0,0	0,0
Samoa	WSM	3,0	Low	109	6,3	1,7	3,1	4,4	0,0	4,2
Sao Tome and Principe	STP	2,4	Low	144	2,7	0,7	1,4	0,1	0,0	0,0
Saudi Arabia	SAU	2,8	Low	123	2,6	3,4	3,5	1,8	4,8	0,0
Senegal	SEN	4,1	Medium	65	0,8	2,7	4,6	0,1	6,3	3,9
Serbia	SRB	2,9	Low	115	2,5	1,9	3,4	5,3	7,7	0,0
Seychelles	SYC	1,4	Very Low	186	5,3	1,2	2,2	0,1	0,0	7,8
Sierra Leone	SLE	4,2	Medium	61	1,2	2,1	3,8	0,1	5,8	2,9

Predykcja cen turystycznych

	Amazon Forecast API
Maksymalny rozmiar danych	30 GB (5 GB w 2020 r.)
Maksymalna liczba kolumn w zbiorze danych (time series)	13 (3 + 10)
Maksymalna liczba kolumn w zbiorze metadanych	10
Maksymalna liczba kolumn w powiązanym zbiorze danych (time series)	25 (2 + 10 + 13)
Maksymalna liczba wierszy w zbiorze danych	3 000 000 000 (100 mln w 2020 r.)
Maksymalna liczba plików	10 000
Maksymalny horyzont czasowy	14 dni
Modele predykcyjne	ARIMA, ETS, Prophet, DeepAR+, NPTS

- Rozwiązania komercyjne nie są dostosowane do budowy modeli predykcyjnych dla turystyki (predykcja cen turystycznych)

- Limity nie są dostosowane do danych turystycznych (np.: liczba kolumn metadanych nie wystarczy do opisania parametrów wycieczki)

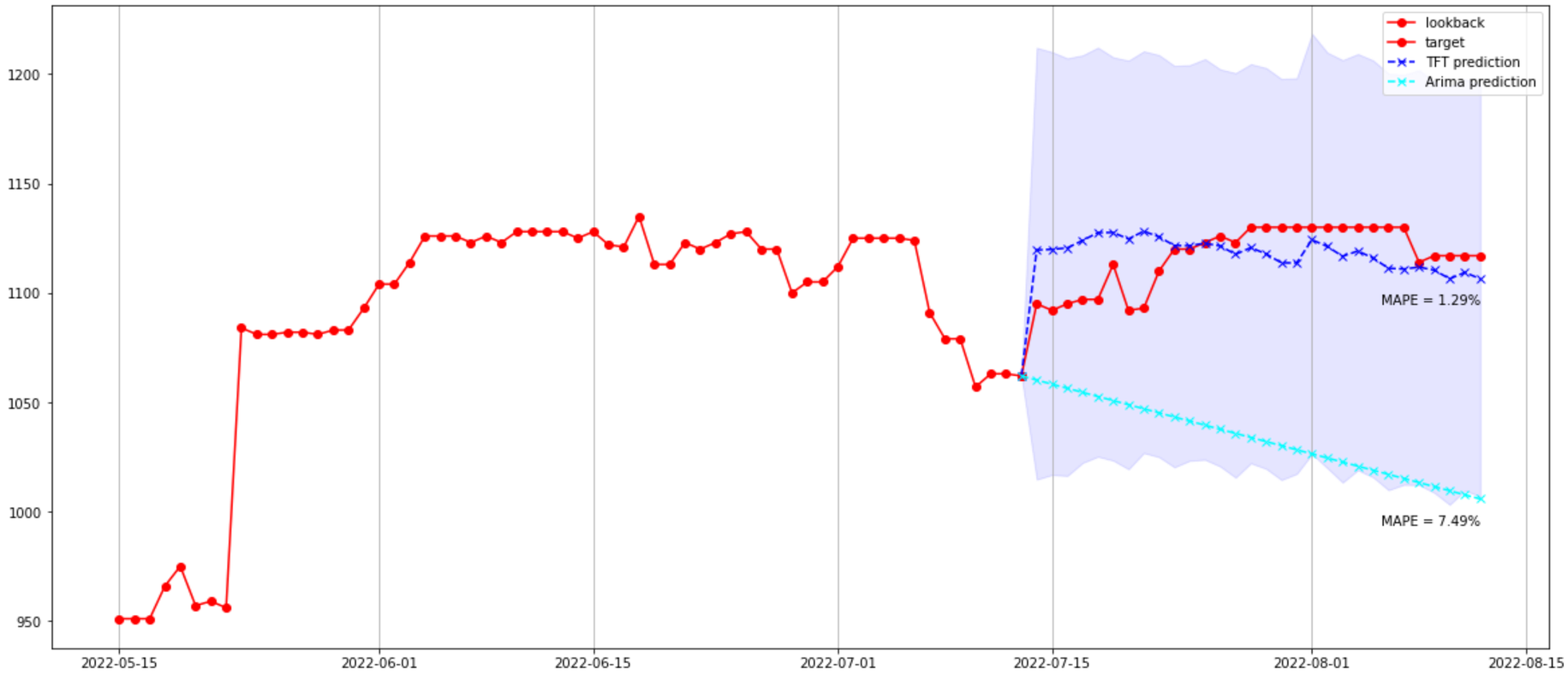
- Modele klasyczne (ARIMA, ETS) – modele lokalne:
  - nie są dostosowane do wielomilionowych szeregów czasowych
  - konieczne tworzenie wielu modeli predykcyjnych (per szereg czasowy)

- Modele, które analizowaliśmy:
  - DeepGLO (sieci neuronowe)
  - GBRT (*Gradient Boosted Regression Tree*)
  - TFT (Temporal Fusion Transfomer)

travel.ai

Predykcja cen turystycznych

Przykładowe wizualizacja predykcji modelu TFT i porównanie z predykcją modelem klasycznym (ARIMA)



travel.ai

Wnioski z budowy systemów Big Data dla turystyki

1. Poprawnie zdefiniowana funkcjonalność systemu Big Data:
   - ✓ Do czego ma służyć system?
   - ✓ Jakie dane chcemy analizować?
   - ✓ Do jakich raportów chcemy mieć dostęp?
   - ✓ Skąd pozyskiwać dane?
   - ✓ Z jaką częstotliwością aktualizować dane?
   - ✓ Jakie opóźnienie w raportach (np.: dane z poprzedniego dnia, dane z ostatniego miesiąca, dane z ostatniego kwartału)

2. Poprawnie zaprojektowana architektura systemu Big Data:
   - ✓ 3 warstwy danych i jakie systemy do ich przechowywania potrzebujemy?
   - ✓ Czy potrzebujemy danych aktualizowanych na bieżąco (minimalne opóźnienia)?
   - ✓ Jaki jest przewidywany wolumen danych?
   - ✓ Jaki przewidywany wzrost danych w ciągu roku, 2 lat?

3. Czy przewidujemy funkcjonalności predykcji (np.: predykcji cen, prognozowanie trendów)?
   - ✓ Jak bardzo zawansowane modele predykcyjne nas interesują?
   - ✓ Czy wystarczą nam modele klasyczne?
   - ✓ Czy potrzebujemy uwzględniać w predykcji czynniki zewnętrzne (problem się komplikuje)?

Dziękuję !

Agnieszka Kukałowicz
CEO & founder Qtravel.ai
Next-generation Travel Search

mail: agnieszka@qtravel.ai
linkedin: https://www.linkedin.com/in/agnieszka-kukalowicz/