Mariusz Dramski, Marcin Mąka

Analiza opóźnień samolotów pasażerskich z wykorzystaniem reguł asocjacyjnych

JEL: O18 DOI: 10.24136/atest.2018.492
Data zgłoszenia: 19.11.2018 Data akceptacji: 15.12.2018

Wydajność transportu pasażerskiego w tym lotnictwa cywilnego, jest kluczowa dla światowej gospodarki. Jednym z głównych czynników oceny linii lotniczych przez pasażerów jest punktualność. Należy tu uwzględnić również fakt, że sieć połączeń między lotniskami na całym świecie jest niezwykle skomplikowana. Powyższe fakty prowadzą do wniosku, że można stworzyć narzędzie, które pomoże pasażerom planować ich podróż w sposób optymalny. W niniejszym artykule do analizy ponad 7 milionów lotów na terenie Stanów Zjednoczonych, zastosowano reguły asocjacyjne. Dane pozyskano z Departamentu Transportu USA i obejmują one loty, które odbyły się w 2008 roku.

Słowa kluczowe: associacion rules, filght's delays, air transport, data mining

Wstęp

Celem każdego pasażera jest dostanie się z punktu początkowego do punktu docelowego w założonym czasie. Naturalnie kluczowym elementem staje się punktualność. O ile w przypadku lotów bezpośrednich, pasażer jest w stanie sam określić opóźnienie lotu (np. na podstawie informacji udzielanych pasażerom na lotnisku), o tyle w przypadku lotów łączonych sytuacja staje się nieco bardziej skomplikowana. Załóżmy, że pasażer podróżuje z Los Angeles do Nowego Jorku z przesiadką w Dalllas. Tam ma zaledwie jedną godzinę na zmianę samolotu. Niekoniecznie musi on sobie zdawać sprawę, że np. samolot startujący z Warszawy będzie miał niewielkie opóźnienie. Wówczas istnieje niewielkie ryzyko, że nie zdąży on na następny lot. Oczywiście renomowane linie lotnicze udzielą wszelkich wskazówek i ewentualnie zapewnią nocleg hotelu, tym niemniej pasażer nie jest w stanie sam określić z jakim wymiarem opóźnienia ma do czynienia.

1 Czym jest opóźnienie lotu ?

Opóźnienie lotu jest terminem zdefiniowanym przez odpowiednie regulacje prawne. Jest to bardzo ważne z uwagi na to, że pasażerowie mogą wnosić swoje roszczenia w stosunku do linii lotniczych. Dlatego zatem należało wyraźnie określić kiedy lot jest opóźniony, a kiedy nie.

Na podstawie regulacji Parlamentu Europejskiego z dn. 11 lutego 2004 [1], można zdefiniować następujące rodzaje opóźnień:

- early arrivals – wczesny przylot, samolot ląduje przed oczekiwanym czasem;
- on time – samolot przylatuje o czasie lub jest opóźniony nie więcej niż 15 minut;
- delayed – opóźnienie w przedziale od 15 minut do 2 godzin;
- long delayed – opóźnienie większe niż 2 godziny.

Powyższa regulacja uwzględnia również odległość pomiędzy konkretnymi lotniskami, ale nie jest to przedmiotem badań w niniejszym artykule. Jest ona również bardzo istotna również z uwagi na fakt iż na jej podstawie przyznawane są stosowne odszkodowania, jeśli wystąpią roszczenia pasażerów. Opóźnienie może zostać zdefiniowane równaniem:

$$D = C + W + NAS + S + LA \quad (1)$$

gdzie:

$D -$ opóźnienie całkowite
$C -$ opóźnienie z winy przewoźnika
$W -$ opóźnienie z powodu warunków pogodowych
$NAS -$ opóźnienie z powodu kontroli ruchu lotniczego
$S -$ opóźnienie z powodu kontroli bezpieczeństwa
$LA -$ późne przybycie samolotu na lotnisko tzw. late arrival

Nawiązując do regulacji prawnych, na potrzeby niniejszego artykułu, proponuje się podział wielkości opóźnień na 5 klas:

a) klasa 1 – wczesny przylot;
b) klasa 2 – przylot o czasie;
c) klasa 3 – opóźnienie lotu, ale pasażerom nie przysługuje odszkodowanie (15 minut – 2 godziny) – niewielkie opóźnienie;
d) klasa 4 – znaczne opóźnienie (powyżej 2 godzin);
e) klasa 5 – rekordy, których nie można zaklasyfikować do klas 1-4.

2 Opis danych

Jak wspomniano wcześniej, dane dotyczące lotów pozyskano z Departamentu Transportu USA [2]. Zbiór zawiera informacje o wszystkich komercyjnych lotach pasażerskich nad terytorium USA w roku 2008. Struktura danych zawarta jest w postaci jednej dużej tabeli. Każdy lot może być zidentyfikowany poprzez zmienną FlightNum, które reprezentuje numer lotu. Dane nie zawierają informacji o trasie, ale ta może być odtworzona. Tabela zawiera 29 kolumn i znajdują się w niej również informacje o opóźnieniach. Z punktu widzenia niniejszego artykułu szczególnie ważne są atrybuty: ArrDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay oraz LateAircraftDelay. Zmienna ArrDelay jest wyrażona poprzez sumę pozostałych zmiennych (równanie (1)).

Tab. 1. Opis danych

Lp	Nazwa	Opis
1	Year	Rok (1987-2008)
2	Month	Miesiąc (1-12)
3	DayOfMonth	Dzień miesiąca (1-31)
4	DayOfWeek	Dzień tygodnia (1- poniedziałek, 7 – niedziela)
5	DepTime	Aktualny czas odlotu
6	CRSDepTime	Planowany czas odlotu
7	ArrTime	Aktualny czas przylotu
8	CRSArrTime	Planowany czas przylotu
9	UniqueCarrier	Unikalny kod przewoźnika
10	FlightNum	Numer lotu
11	TailNum	Numer samolotu (na ogonie)
12	ActualElapsedTime	Faktyczny upływ czasu
13	CRSElapsedTime	Planowany upływ czasu
14	ArrTime	Czas przylotu w minutach
15	ArrDelay	Całkowity czas opóźnienia w minutach
16	DepDelay	Czas opóźnienia odlotu w minutach
17	Origin	Lotnisko odlotu (wg IATA)
18	Dest	Lotnisko przeznaczenia (wg IATA)
19	Distance	Odległość w milach
20	TaxiIn	Kołowanie (przylot) w minutach
21	TaxiOut	Kołowanie (odlot) w minutach
22	Cancelled	Czy lot był odwołany ?

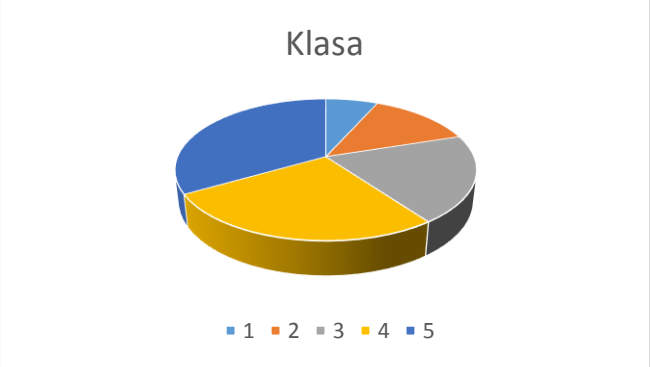
23	CancellationCode	Przyczyna odwołania lotu
24	Diverted	Czy lot był przekierowany ?
25	CarrierDelay	Opóźnienie z winy przewoźnika w minutach
26	WeatherDelay	Opóźnienie z winy warunków pogodowych w minutach
27	NASDelay	Opóźnienie z uwagi na kontrolę lotów nad danym terytorium (NAS w przypadku USA) w minutach
28	SecurityDelay	Opóźnienie z przyczyn bezpieczeństwa w minutach
29	LateAircraftDelay	Opóźnienie powstałe w wyniku wcześniejszego późnego przybycia na lotnisko w minutach

Łączna liczba rekordów: 7009729
Wielkość pliku: 673256 KB

Uwzględniając proponowany podział na klasy:

Tab. 2. Podział danych na klasy

Klasa	Opis	Liczba instancji	Liczba instancji w [%]
1	Wczesny przylot	3690606	52.6%
2	Przylot o czasie	1639688	23.4%
3	Niewielkie opóźnienie	1371198	19.6%
4	Znaczne opóźnienie	153537	2.2%
5	Niesklasyfikowane w 1-4	154699	2.2%



Rys. 1. Ilustracja klas

Już na podstawie wstępnej analizy danych zawartych w Tab. 2 widać wyraźnie, że odsetek lotów opóźnionych znacznie jest niewielki. Dominują loty punktualne lub tzw. early arrivals. Odsetek lotów o niewielkim opóźnieniu jest na poziomie ok. 20%. Rekordy, których nie dało się sklasyfikować w klasach 1-4 stanowią nieco ponad 2% ogólnej liczby danych. Można zatem przyjąć, że są to dane błędne lub niekompletne z powodu różnych przyczyn.

3 Analiza

Głównym celem analizy jest oszacowanie czy tzw. późne przybycie na lotnisko (late arrival) powoduje finalne opóźnienie podróży. Jest to informacja bardzo istotna dla pasażera. Wzięto również pod uwagę kolejny czynnik: dzień tygodnia.

Można więc zatem przyjąć, że finalna postać reguły asocjacyjnej może przyjąć postać:

$$Jeżeli\ (klasa = x)\ i\ (dzeń\ tygodnia = y)\ to\ (klasa\ wyjściowa = z) \quad (2)$$

3.1 Parametry reguł asocjacyjnych

Aby ocenić jakość reguł asocjacyjnych należy określić trzy parametry:

- wsparcie (ang. support);
- ufność (ang. confidence);
- korelacja (ang. lift).

Wartość wsparcia powinna być jak najwyższa, ponieważ mówi ona ile instancji spełnia daną regułę. Wsparcie wyrażamy wzorem:

$$support(X \Rightarrow Y) = \frac{N_{XY}}{N} \quad (3)$$

gdzie:
N_{XY} – łączna liczba instancji pokrywających X (warunek) oraz Y (wynik);
N – łączna liczba instancji

Ufność z kolei mówi o tym jaki jest wpływ warunku X na wynik Y. Pożądana wartość musi być bliska 1. Wyrażana jest wzorem:

$$confidence(X \Rightarrow Y) = \frac{N_{XY}}{N_X} \quad (4)$$

gdzie:
N_{XY} – łączna liczba instancji pokrywających X (warunek) oraz Y (wynik);
N_X – łączna liczba instancji pokrywających warunek X

Ostatnim parametrem jest korelacja i wyrażana jest wzorem:

$$lift(X \Rightarrow Y) = \frac{N_{XY} \cdot N}{N_X \cdot N_Y} \quad (5)$$

gdzie:
N_{XY} – łączna liczba instancji pokrywających X (warunek) oraz Y (wynik);
N_X – łączna liczba instancji pokrywających X (warunek);
N_Y – łączna liczba instancji pokrywających Y (wynik);
N – łączna liczba instancji

Interpretacja korelacji jest następująca:

$lift(X \Rightarrow Y) > 1$ - X i Y są pozytywnie skorelowane
$lift(X \Rightarrow Y) = 1$ - X i Y są niezależne
$lift(X \Rightarrow Y) < 1$ - X i Y są negatywnie skorelowane

3.2 Wyniki

Tab. 3. ilustruje wyniki przeprowadzonych obliczeń. Poszczególne kolumny opisują: klasę, dzień, klasę wyjściową, wsparcie 1, wsparcie 2, ufność, korelację oraz liczbę instancji. Jak łatwo zauważyć wsparcie wyrażone jest w dwóch kolumnach. W pierwszym przypadku rozważone są wszystkie występujące instancje, w drugim natomiast wzięto pod uwagę jedynie loty opóźnione. Można przyjąć takie założenie jeśli przyjmiemy, że w przypadku braku opóźnienia, pasażer nie będzie oczekiwał żadnych dodatkowych informacji. W ostatniej kolumnie zawarto liczbę instancji, które są zgodne z proponowaną regułą.

Pierwszą obserwacją jest fakt, że odbywanie lotu w konkretny dzień tygodnia nie może być powiązane z faktem wystąpienia ewentualnego opóźnienia. Nie ma znaczenia czy w podróż wybierzemy się w weekend czy dzień roboczy. Mimo wszystko można zaobserwować trend, że odsetek opóźnień wzrasta nieznacznie w piątki, po czym w soboty wraca z powrotem do normy. Można jednak przyjąć, że jest to spowodowane po prostu większą liczbą lotów w te dni. Dla ułatwienia interpretacji obliczeń można przyjąć, że następujące reguły będą dobrze opisywać zależność pomiędzy opóźnieniami na trasach łączonych:

- jeśli samolot będzie miał niewielkie opóźnienie, to na lotnisku docelowym również będzie miał niewielkie opóźnienie;

– jeśli samolot będzie miał znaczne opóźnienie, to na lotnisku docelowym będzie miał niewielkie opóźnienie;
– jeśli samolot będzie miał niewielkie opóźnienie, to na lotnisku docelowym będzie miał znaczne opóźnienie;
– jeśli samolot będzie miał znaczne opóźnienie, to na lotnisku docelowym również będzie miał znaczne opóźnienie;
– jeśli samolot przybędzie o czasie, to na lotnisku docelowym będzie miał niewielkie opóźnienie;
– jeśli samolot przybędzie o czasie, to na lotnisku docelowym będzie miał znaczne opóźnienie.

Pierwsza grupa reguł mówi, że gdy samolot będzie miał niewielkie opóźnienie, to również wystąpi ono na lotnisku docelowym. Wszystkie parametry potwierdzają ten fakt. Niski poziom wsparcia spowodowany jest jednak przez niską liczbę instancji zgodnych z tą regułą. Jednakże ufność i korelacja wskazują na prawdziwość reguły.

Następnie skupiono się na tym czy samolot z dużym opóźnieniem może nadrobić nieco czasu i zmniejszyć opóźnienie na lotnisku docelowym. Wszystkie parametry przyjęły wartość 0 – wskazuje to więc na to, że taka reguła nie jest prawdziwa.

Trzecia grupa to sytuacja odwrotna. Sprawdza się tu czy samolot może zwiększyć swoje opóźnienie. Na podstawie obliczeń można stwierdzić, że istnieje korelacja, natomiast takie sytuacje w praktyce nie występują.

Czwarta grupa reguł mówi, że jeśli mamy znaczne opóźnienie, to na lotnisku docelowym również będzie ono znaczne. Ta reguła jest jak najbardziej prawdziwa, choć jej wsparcie jest nieznaczne. Spowodowane jest to faktem, że odsetek lotów opóźnionych ponad 2 godziny jest niewielki. Linie lotnicze starają się uniknąć opóźnień, aby zapobiec ewentualnym roszczeniom pasażerów.

Piąta grupa reguł mówi nam o tym, że pomimo punktualnego odlotu, samolot może mieć niewielkie opóźnienie na lotnisku docelowym. Wszystkie parametry potwierdzają ten fakt.

Grupa szósta jest podobna do piątej, uwzględnia jednak znaczne opóźnienie samolotu na lotnisku docelowym. Ta regułą może być w praktyce wyeliminowana pomimo wystąpienia korelacji.

Tab. 3. Wartości parametrów

Klasa	Dzień	Klasa w.	Wsp. 1	Wsp. 2	Ufność	Kor.	Instancje
3	PON	3	0.01	0.05	0.94	4.83	71375
3	WTO	3	0.01	0.05	0.94	4.79	63268
3	ŚRO	3	0.01	0.05	0.95	4.85	63402
3	CZW	3	0.01	0.05	0.95	4.83	72266
3	PIĄ	3	0.01	0.05	0.94	4.80	85099
3	SOB	3	0.01	0.05	0.94	4.83	48431
3	NIE	3	0.01	0.05	0.93	4.77	67629
4	PON	3	0.00	0.00	0.00	0.00	0
4	WTO	3	0.00	0.00	0.00	0.00	0
4	ŚRO	3	0.00	0.00	0.00	0.00	0
4	CZW	3	0.00	0.00	0.00	0.00	0
4	PIĄ	3	0.00	0.00	0.00	0.00	0
4	SOB	3	0.00	0.00	0.00	0.00	0
4	NIE	3	0.00	0.00	0.00	0.00	0
3	PON	4	0.00	0.00	0.06	2.54	4210
3	WTO	4	0.00	0.00	0.06	2.92	4318
3	ŚRO	4	0.00	0.00	0.05	2.30	3365
3	CZW	4	0.00	0.00	0.05	2.50	4186
3	PIĄ	4	0.00	0.00	0.06	2.80	5567
3	SOB	4	0.00	0.00	0.06	2.53	2847
3	NIE	4	0.00	0.00	0.07	3.01	4776
4	PON	4	0.00	0.00	1.00	45.65	7521
4	WTO	4	0.00	0.00	1.00	45.65	7032
4	ŚRO	4	0.00	0.00	1.00	45.65	5541
4	CZW	4	0.00	0.00	1.00	45.65	6874
4	PIĄ	4	0.00	0.00	1.00	45.65	8884
4	SOB	4	0.00	0.00	1.00	45.65	5117
4	NIE	4	0.00	0.00	1.00	45.65	9109
2	PON	3	0.02	0.09	0.92	4.73	131438
2	WTO	3	0.02	0.09	0.92	4.70	127345
2	ŚRO	3	0.02	0.09	0.93	4.76	127264
2	CZW	3	0.02	0.09	0.93	4.75	136635
2	PIĄ	3	0.03	0.14	0.92	4.73	146432
2	SOB	3	0.01	0.05	0.93	4.74	103505
2	NIE	3	0.02	0.09	0.91	4.66	127009
2	PON	4	0.00	0.00	0.08	3.43	10692
2	WTO	4	0.00	0.00	0.08	3.72	11291
2	ŚRO	4	0.00	0.00	0.07	3.19	9555
2	CZW	4	0.00	0.00	0.07	3.23	10402
2	PIĄ	4	0.00	0.00	0.08	3.44	11928
2	SOB	4	0.00	0.00	0.07	3.28	8022
2	NIE	4	0.00	0.00	0.09	4.03	12300

Podsumowanie

Celem niniejszego artykułu była ekstrakcja reguł, które mają na celu sprawdzenie czy założenia dotyczące opóźnień lotów znajdują potwierdzenie w rzeczywistości. W zasadzie prawdziwe będą tu trzy reguły (wysoka wartość parametru ufności):

– jeśli opóźnienie samolotu jest niewielkie, to nie zmieni się ono na lotnisku docelowym;
– jeśli opóźnienie samolotu jest znaczne, to nie zmieni się ono na lotnisku docelowym;
– jeśli samolot jest o czasie, to może mieć on niewielkie opóźnienie na lotnisku docelowym.

Wartości korelacji mówią o tym, że jest ona pozytywna. Jedynym wyjątkiem jest wartość 0, która spowodowana jest tym, że nie ma reguł spełniających założenia (wsparcie równe 0).

Ponadto nie ma większego znaczenia czy lot odbywa się w konkretny dzień tygodnia. Wszystkie te obserwacje znajdują swoje potwierdzenie w rzeczywistości. Ponadto dostępne dane mogą posłużyć również innym badaniom (np. czy częstotliwość lotów zmieniła się po 11 września 2001 r.).

Bibliografia:

1. Regulation EC No 261/2004 of the European Parliament and of the Council of 11 Feb 2004.
2. http://stat-computing.org/dataexpo/2009/the-data.html (dostęp dn. 15.09.2018 r
3. Agrawal R., Srikant R., Proceedings of the 20th VLDB Conference, p.p. 487-499, Santiago de Chile, 1994

Analysis of flights' delays using association rules

The efficiency of air passenger transport in world's economy is crucial. For this kind of flights, one of the most important features is punctuality. The network of connections between the airports, very often is significantly complicated. It leads to the conclusion that there is a need to do some research in this field which will help the passengers to plan their optimal journeys. In this paper one of the data mining techniques (association rules) was applied to the analysis of flights' delays. The data consists of over 7 millions records was taken from the US Department of Transportation (year 2008) [2]. Then the research was carried out and conclusions were described.

Keywords: association rules, data mining, flights' delays, air transport

Autorzy:

dr inż. **Mariusz Dramski** – Akademia Morska w Szczecinie, Wydział Nawigacyjny, ul. Wały Chrobrego 1-2, 70-500 Szczecin, e-mail:m.dramski@am.szczecin.pl

dr inż. **Marcin Mąka** – Akademia Morska w Szczecinie, Wydział Nawigacyjny, ul. Wały Chrobrego 1-2, 70-500 Szczecin, e-mail: m.maka@am.szczecin.pl