Predykcja opóźnień lotów

Temat projektu

Każdego roku na świecie odbywa się wiele przelotów, zarówno pasażerskich jak i towarowych.
Dodatkowa minuta spędzona w powietrzu oznacza koszty związane z paliwem, czy wynagrodzeniem załogi. Opóźnienia natomiast powodują nieprzyjemności zarówno dla pasażerów, jak i przewoźników.

Zadaniem projektu jest stworzenie odpowiedniego modelu do predykcji czy dany lot zostanie opóźniony czy też nie, przy zachowaniu szczególnej ostrożności podczas przygotowania danych, aby nie doprowadzić do ich wycieku.

Biblioteki użyty w projekcie:

pandas

numpy

seaborn

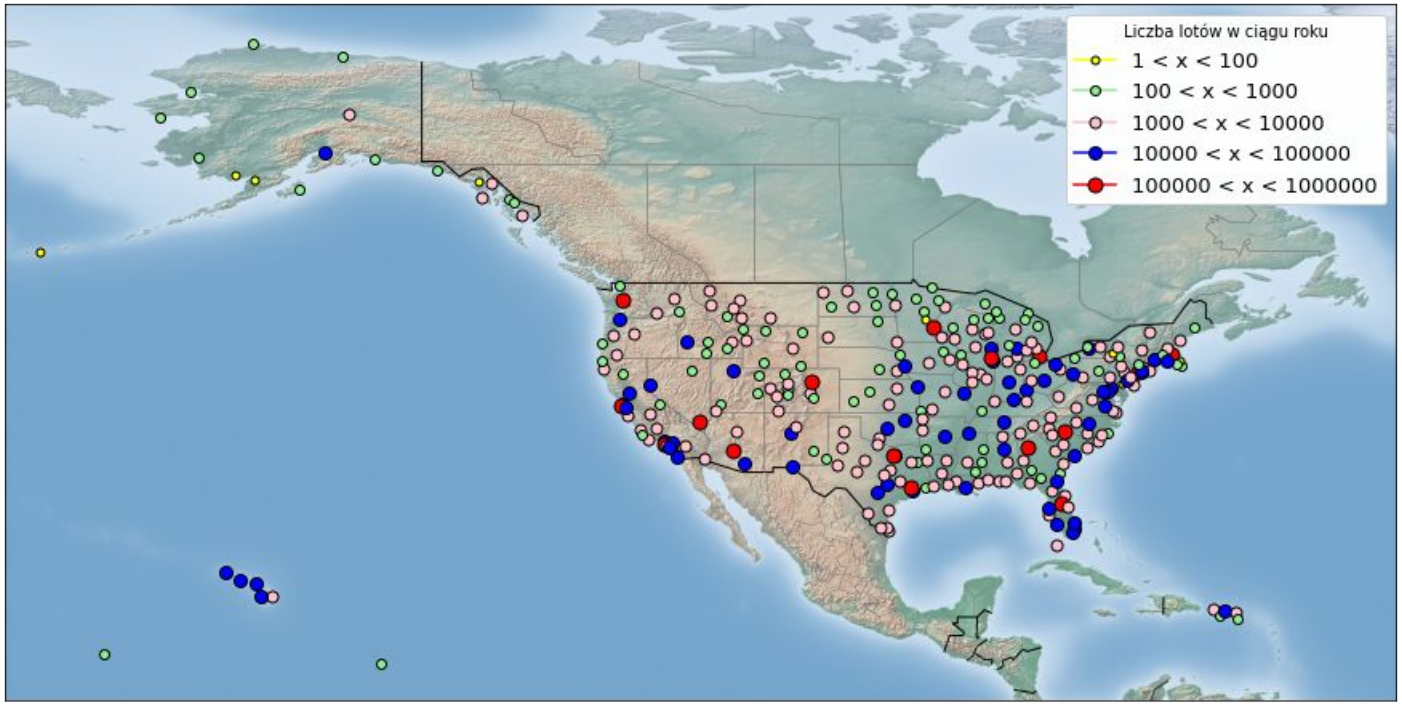
matplotlib

matplotlib.pyplot

datetime, warnings, scipy

Dane zostały pobrane z Kaggle.com

Obszar geograficzny, którego dotyczy zbiór danych oraz lokalizacja lotnisk i liczba lotów zarejestrowanych w każdym z nich w roku 2015 :



Ze względu na bardzo dużą liczbę danych,
bo aż **5 819 079** lotów,
do dalszej analizy danych wybrany został tylko jeden miesiąc:

styczeń 2015

Duża uwaga została poświęcona kolumnom.
Niestety, niektóre z nich musiały zostać usunięte aby uniknąć wycieku danych.

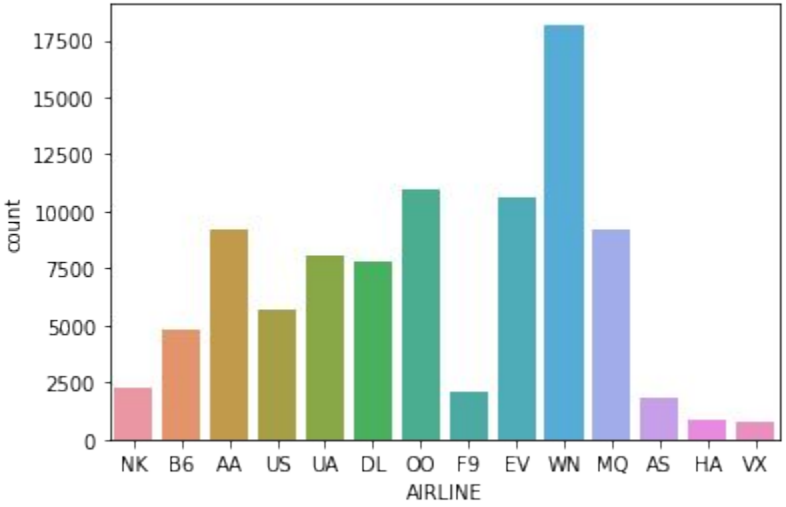
Wyciek danych - w tym przypadku wystąpiłby wtedy, gdy jedna z kolumn zawierałaby informację na temat, np. rzeczywistego lotu, bądź opóźnienia w minutach.

Niektóre dane zostały po prostu pominięte, gdyż nie miały one wpływu na model, np. numer lotu.

Wersy, które zawierały wartości zerowe, zostały także usunięte (około 2,5%).

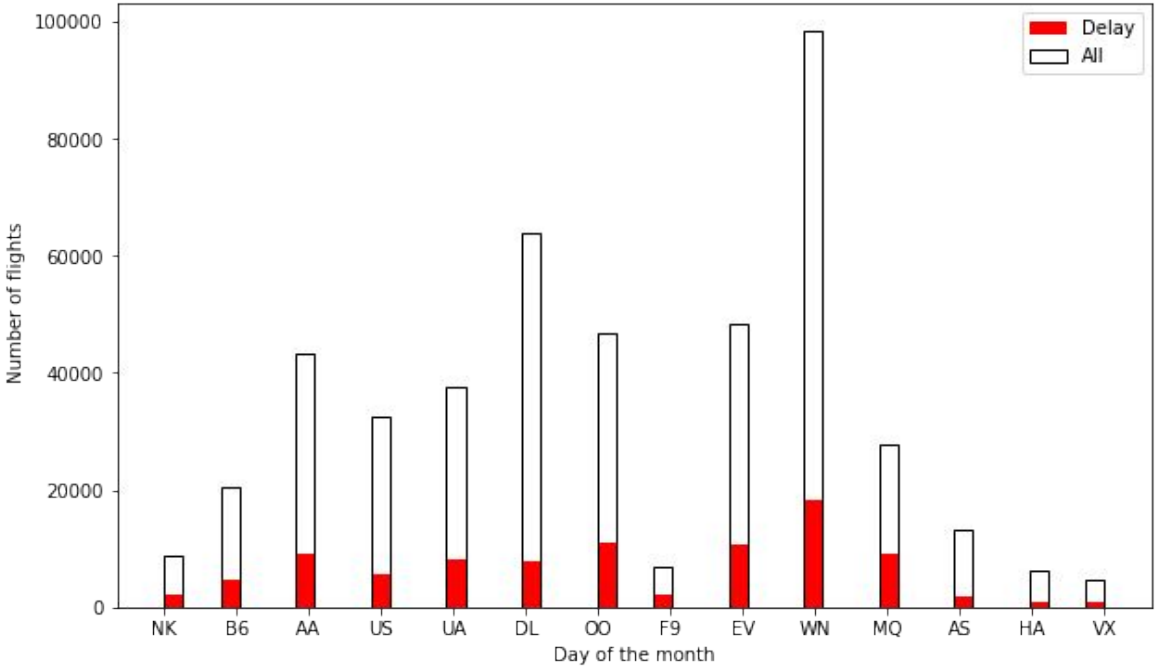
Informacje dotyczące opóźnień lotów

Przed rozpoczęciem podziału na target oraz features, sprawdzone zostało czy linia lotnicza może mieć wpływ na opóźnienie lotu.



Możemy zauważyć, że niektóre linie lotnicze mają dość sporą liczbę opóźnionych lotów.

Porównajmy ilość opóźnionych lotów do wszystkich odbytych lotów w danej linii lotniczej.



Sprawdźmy to na liczbach

AIRLINE	DELAY ▲
DL	0.12215285185417582
HA	0.1314312441534144
AS	0.14058355437665782
VX	0.15867555364437755
US	0.17609811575938278
WN	0.18501861988970514
AA	0.21224688383300125
UA	0.21644582175678564
EV	0.22041171714658872
B6	0.2323881905875475
OO	0.2344889343738007
NK	0.2640832851359167
F9	0.3102373887240356
MQ	0.33426244343891404

Wniosek

Mimo, że może nam się wydawać, że niektóre z linii lotniczych mają dużo opóźnień warto porównać to z wszystkimi ich lotami, aby sprawdzić jak to się ma do całości.

Okazuję się, że niektóre linie lotnicze mają za sobą 20% opóźnionych lotów z wszystkich odbytych w styczniu, a dwie z nich przekraczają nawet 30%.

Wybór atrybutów oraz etykiety

Naszymi atrybutami są:

dni tygodnia

czas opóźnienia w starcie lotu

czas trwania przejścia przez bramki aż do oderwania kół z podłoża samolotu

planowany czas trwania lotu

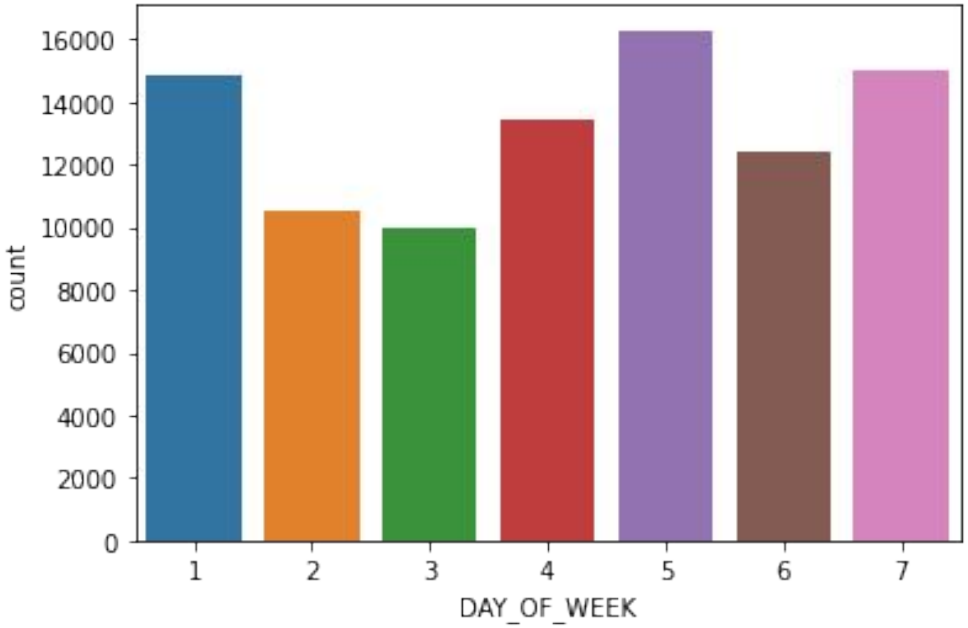
dystans w milach

linie lotnicze

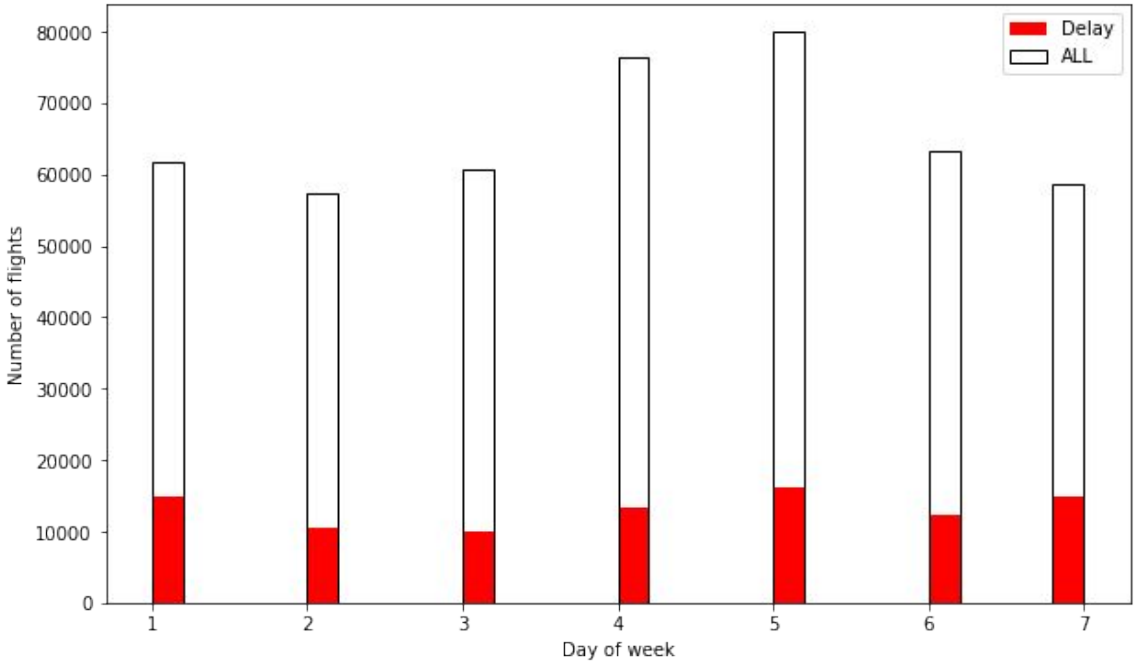
Etykieta (target):

Czy lot jest opóźniony czy nie (opóźnienie lotu powyżej 15 minut)

Loty w poszczególne dni tygodnia.
Na wykresie możemy zobaczyć, że wtorek oraz środa są mniej narażone na opóźnienie lotu.



Porównajmy to do wszystkich odbytych lotów w danym dniu.



Znów zobaczmy to na liczbach

DAY_OF_WEEK	DELAY
1	0.241273
2	0.183229
3	0.163814
4	0.175452
5	0.203669
6	0.195674
7	0.255649

Porównując opóźnienia do wszystkich odbytych lotów w danym dniu, niestety to w niedzielę ponad 25% lotów została opóźniona, w poniedziałek ponad 24%.

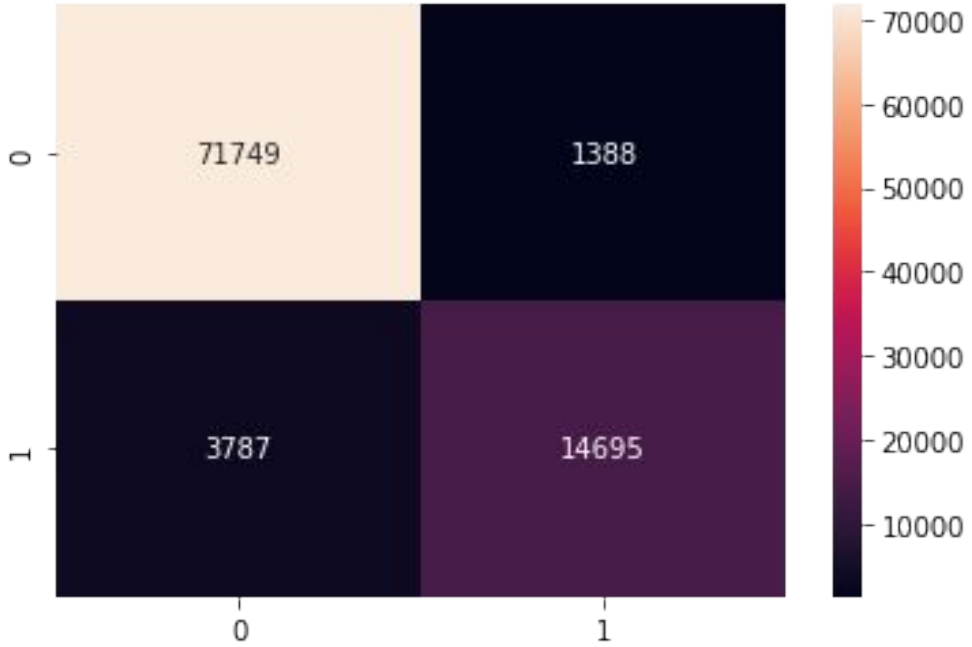
Środa wygrywa w rankingu najmniejszej ilości opóźnionych lotów.

Czas na uczenie modeli.

Ze względu na to, że jest to klasyfikacja, modele wybrane to:

- **Logistic Regression**
- **K-Neighbors Classifier**
- **Decision Tree Classifier**

Regresja Logistyczna
Tablica pomyłek



True Positive:

14695

False Negative:

3787

True Negative:

71749

False Positive:

1388

Metryki

- **Accuracy (dokładność) - procent poprawnie zaklasyfikowanych przykładów**
- **Precision (precyzja) - sprawdza pewność klasyfikatora dla przykładów pozytywnych**
- **Recall (czułość) - określa jaką część dodatnich wyników wykrył klasyfikator**
- **F1-score - łączy precyzję i czułość - średnia harmoniczna**

Dla naszego modelu Recall będzie najważniejszą metryką.

Metryki dla regresji logistycznej

Accuracy: 0.943516082908567

Precision: 0.9136976932164397

Recall: 0.7950979331241208

F1-score: 0.8502820772457688

Metryki dla metody k-najbliższych sąsiadów

Accuracy: 0.9352863489014287

Precision: 0.933856224128091

Recall: 0.9352863489014287

F1-score: 0.9335627207339542

Metryki dla drzewa decyzyjnego

Accuracy: 0.91090276034447

Precision: 0.9116617338607232

Recall: 0.91090276034447

F1-score: 0.9112583920590964

Podsumowanie

Najlepszym modelem okazała się metoda k-najbliższych sąsiadów.

Recall na poziomie 93,5% oznacza, że 93,5% opóźnionych lotów nasz model zaklasyfikował poprawnie jako opóźnione.

Możliwość rozwoju projektu

Dodanie informacji na temat pogody czy może mieć ona wpływ na opóźnienia lotów.

Sprawdzenie czy piloci są w stanie opóźniony start nadrobić na dłuższym dystansie.

Aplikacja do prognozy opóźnień lotu.