Airline Passenger Satisfaction

# Opis problemu

Dany dataset powstał w wyniku badań zadowolenia klientów jednej z linij lotniczych.

Za pomocą modelu stworzonego na podstawie tego datasetu da się analizować wyniki ankiet klientów i odnotowywać, czy klient został zadowolony lotem, czy nie.

Ten problem jest ciekawy, bo za pomocą rozwiązania tego problemu można podwyższyć jakość usług w branży, która pomaga dużej ilości ludzi.

# Dane

Źródło datasetu — opis cech na stronie datasetu wygląda na wiarygodny.

Na podstawie analizy danych za pomocą wykresów doszliśmy do wniosku, że dane są sensowne.

Podstawowa analiza danych wykryła, że największą wagę spośród ocen mieli pola „Seat comfort”, „Inflight service”, „On-board service”, „Inflight entertaiment” oraz „Cleanliness”. (zał. 1)

Dane z tego datasetu można użyć do predykcji satysfakcji klient linij lotniczych na podstawie ogólnych ocen oraz danych klienta.😊

# Sposób rozwiązania problemu

Co zostało zrobione dla przygotowania danych:

1. Usuwanie pola numeru linij

2. Dołożone pole średniej oceny klienta

3. Próba usuwania mniej istotnych ocen (to zmniejszyło wynik analizy — próba nieudana)

4. Transformacja atrybutów:

4.1. Usuwanie pola Id

4.2. Zmiana danych w polu Płeć - (Male do M, Female do F)

4.3. Zmiana danych w polu Customer Type - (Loyat do L, Disloyal do D)

4.4. Zmiana danych w polu Type of Travel - (Private do P, Business do B)

4.5. Utworzenie pola IsWifiAvailable na podstawie kolumny Inflight Wifi Service (jeśli zero, to IsWifiAvailable = false)

4.6. Wymazanie danych 0 w kolumnie Inflight Wifi Service (bo to nie jest ocena, tylko informacja, że WiFi na locie nie było)

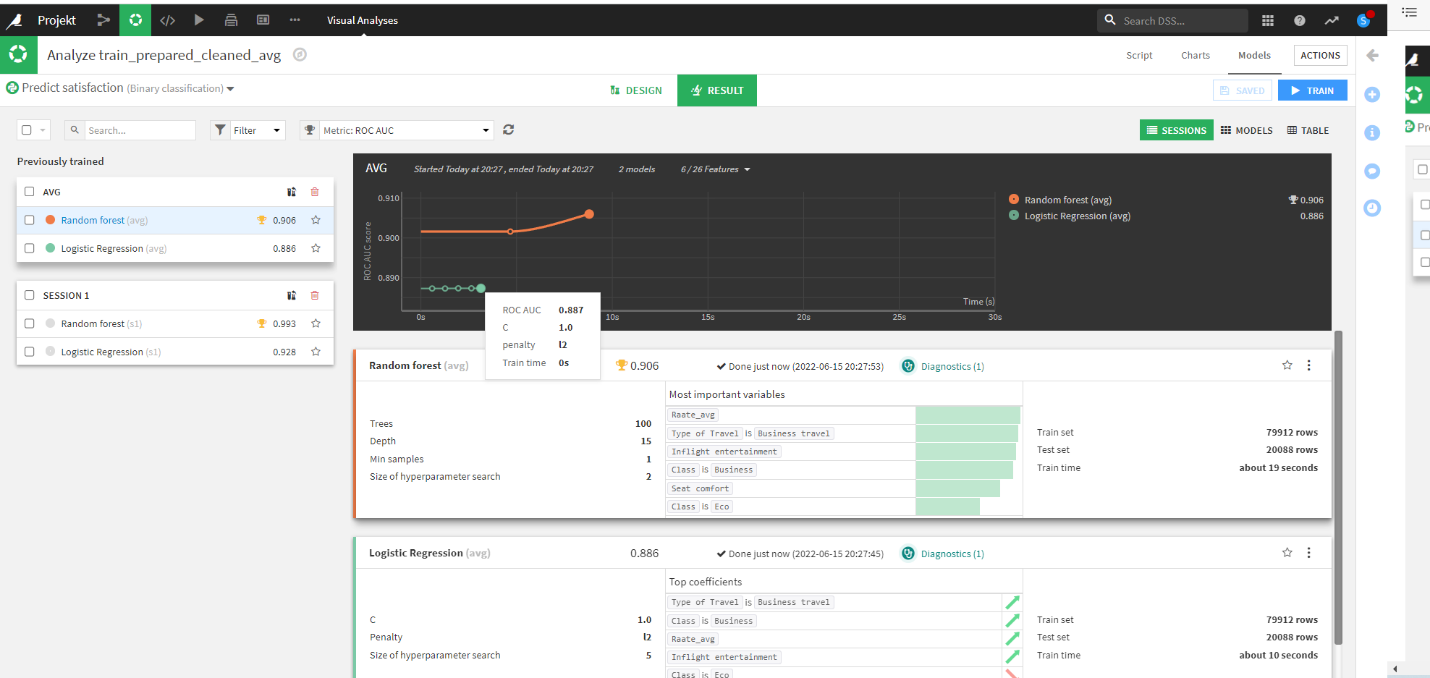
4.7. Zaokrąglenie danych w kolumnie Rate\_avg (utworzoną w kroku 2)

4.8. Zaokrąglenie danych w kolumnie Arrival Delay in Minutes. 4.9. Zmiana kolumny „satisfaction” na IsSatisfied (boolean)

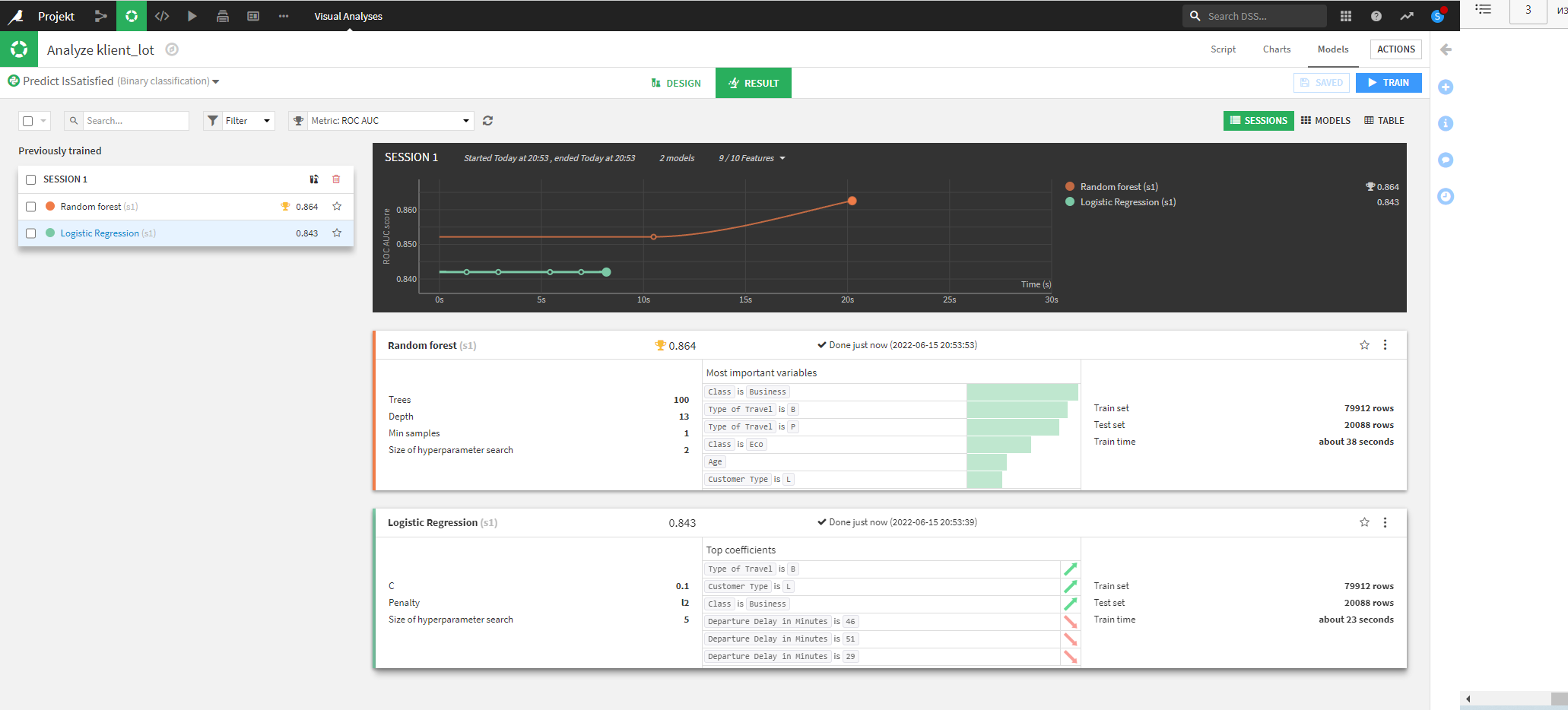
5. Próba rozdzielenia datasetu na dwa osobne datasety Klienty oraz Oceny. To rozdzielenie było spowodowane chęcią pominąć pobieranie ankiet od użytkowników i dostać informacje o satysfakcji klienta na podstawie tylko ogólnych danych o kliencie oraz locie. Niestety, wyniki analizy na podstawie datasetu Klienty były znacznie gorsze od wyników na podstawie Oceny, które z kolei było gorsze od wyników na podstawie połączonego datasetu.

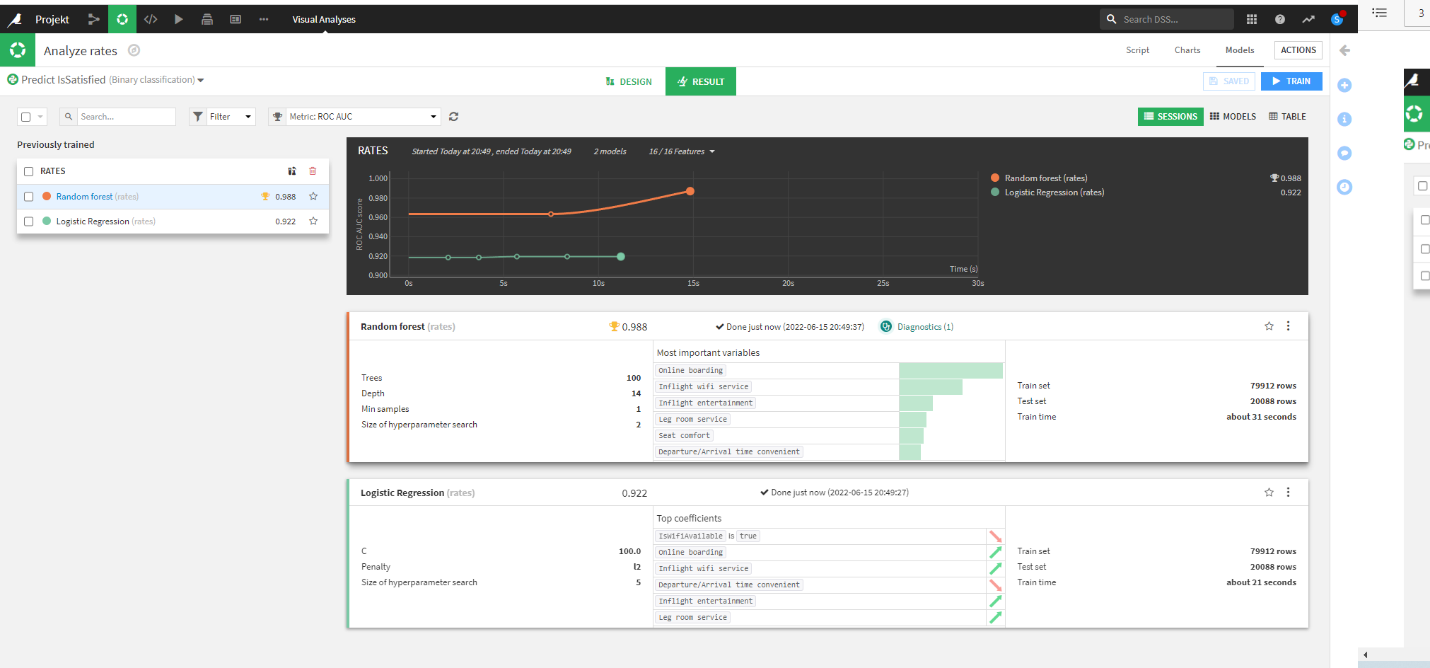
Analizy były prowadzone na modelach algorytmu „Random Forest” oraz „Logical Regression”. Pod uwagę były brane wyniki modelu w metrykach ROC AUC, Accuracy oraz Precision. Także po każdej sesji treningu były sprawdzane wykresy Calibration Curve, ROC Curve oraz Density chart. Po osiągnięciu najlepszych wyników z użyciem „Random Forest” / „Logical Regression” oraz zrozumienia, że eksperymenty z danymi nie pokazują wyników, była podjęta decyzja przetestować inne algorytmy. Okazało się, że pomimo bardzo dobrego wyniku w 0.988 ROC AUC z użyciem Random Forest, użycie algorytmu XGBoost powoduję wynik w 0,992 ROC AUC oraz tak samo na 0,001 lepszy wynik dla Accuracy/Precision.

# Dyskusja wyników i ewaluacja modelu

Model na podstawie danych ogólnych klienta, lota oraz 5 głównych ocen: 

Model na podstawie danych ogólnych klient oraz lota:



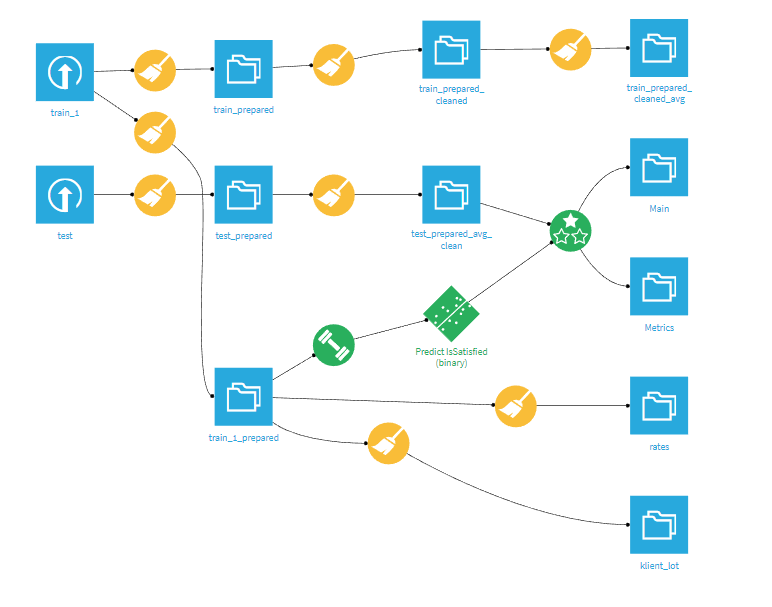
Model na podstawie wyłącznie wszystkich ocen:

Model XGBoost na podstawie danych ogólnych klienta, lota oraz wszystkich ocen (finalny model):

Изображение выглядит как стол

Автоматически созданное описание

# Podsumowanie

Finalny flow: 

Udało się z precyzją 0.945 przewidzieć poprawność zadowolenia klienta z lotu

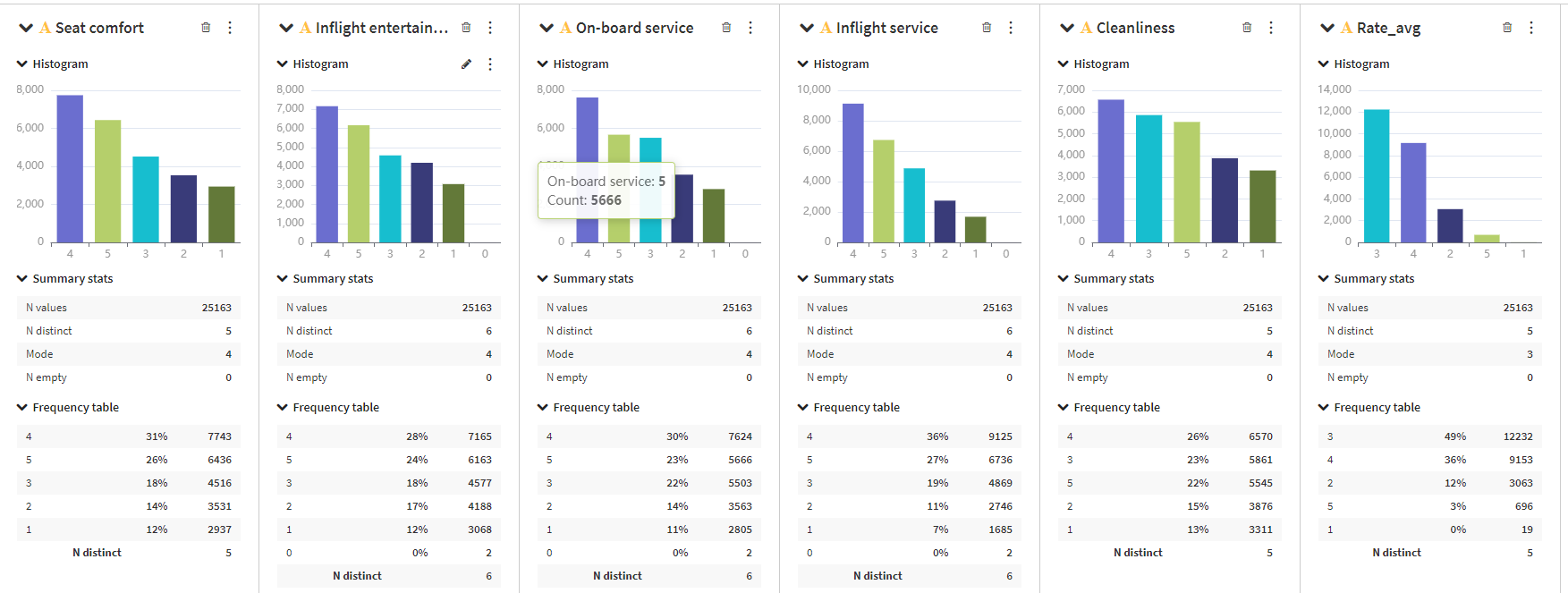
Był problem z eksploracją danych, bo było za dużo różnych ocen, dlatego dołożyliśmy pole średniej oceny.

Dany model w przyszłości można użyć do analizy satysfakcji użytkowników.

Też można rozwinąć ten model, dodając więcej informacji o kliencie/locie dla wycofania z ankiet satysfakcji.

# Załączniki

## Załącznik 1. Eksploracyjna analiza danych



# Załącznik 1. Macierz pomyłek

