# Projekt IUM

Raport końcowy


- Michał Pałasz
- Michał Sadlej


## Model bazowy

Model bazowy został zaprojektowany jako punkt odniesienia do oceny skuteczności predykcji w procesie automatycznego wypełniania formularza ofert. Wykorzystuje on wyłącznie proste reguły statystyczne oparte na danych historycznych.

- Dla zmiennych liczbowych (price, beds, accommodates, bathrooms, bedrooms) wyznaczana jest mediana z wcześniejszych ofert danego gospodarza (host_id).
- Dla zmiennych kategorycznych (room_type, property_type, bathrooms_text) wybierana jest najczęściej występująca wartość.
- W przypadku braku danych dla konkretnego gospodarza, wykorzystywane są dane globalne.

Model został przetestowany na zbiorze testowym. Predykcje oceniano za pomocą:

- błędu bezwzględnego (MAE) dla zmiennych liczbowych,
- trafności (accuracy) dla zmiennych kategorycznych.

Model został zapisany jako base_model.pkl i wykorzystany jako wariant referencyjny w mikroserwisie A/B.

## Model zaawansowany

Model zaawansowany AdvancedModel wykorzystuje przetwarzanie języka naturalnego (NLP) oraz algorytmy uczenia maszynowego w celu zwiększenia trafności predykcji.

- Wejście stanowią tekstowe pola formularza: name, description, neighbourhood, łączone w jedną sekwencję i wektoryzowane metodą TF-IDF.
- Predykcja zmiennych liczbowych odbywa się za pomocą RandomForestRegressor w układzie MultiOutputRegressor.
- Predykcja zmiennych kategorycznych odbywa się za pomocą RandomForestClassifier z MultiOutputClassifier.
- Kategorie kodowane i dekodowane są za pomocą LabelEncoder.

Model został wytrenowany na oczyszczonych danych i zapisany jako advanced_model.pkl. Został również zintegrowany z mikroserwisem i użyty w eksperymencie A/B.

## Mikroserwis

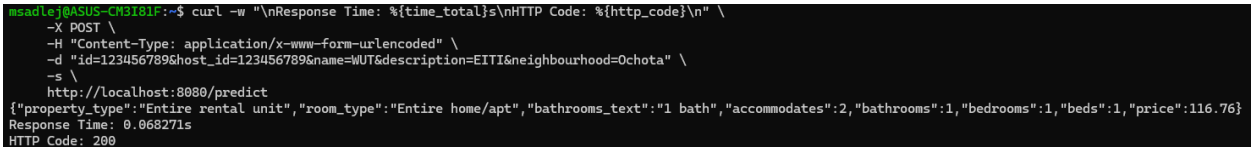
Mikroserwis serwujący predykcje został zaimplemetowany przy użyciu framework'u FastAPI. Udostępnia on 3 endpoint'y:

- /predict/base - zwraca predykcje wygenerowane przez model bazowy
- /predict/advanced - zwraca predykcje wygenerowane przez model zaawansowany
- /predict - służący do przeprowadzania testów A/B

Mikroserwis pozwala na przeprowadzanie testów A/B. Użytkownicy są dzieleni na dwie grupy przy użyciu funkcji hashującej z "host_id". Dzięki temu rozwiązaniu podział na grupy jest stabilny, losowy oraz równomierny. Użytkownicy przydzieleni do grupy A dostają predykcje wygenerowane przez model bazowy, natomiast z grupy B – model zaawansowany. Z perspektywy użytkownika nie ma różnicy jaki model został użyty - odpowiedź przyjmuje tą samą postać niezależnie od grupy.

## Wnioski

1. Model bazowy, oparty na medianach i modach z danych historycznych danego gospodarza, działa dobrze w przypadku użytkowników z wcześniejszymi ofertami, ale traci skuteczność przy nowych użytkownikach bez historii.
2. Model zaawansowany, wykorzystujący analizę tekstu (TF-IDF) oraz algorytmy Random Forest, poprawia trafność predykcji.
3. Eksperyment A/B wykazał:
    a. Accuracy dla obu modeli przekroczyło 50%, co oznacza spełnienie pierwszego kryterium analitycznego.
    b. Spośród 552 zapytań, pole z najmniejszą liczbą predykcji zostało przewidziane w 404 przypadkach (73%), spełniając drugie kryterium analityczne (pokrycie powyżej 60%).
    c. Czas potrzebny na wygenerowanie sugestii był znacznie krótszy niż 1 sekunda, co potwierdza spełnienie trzeciego kryterium dotyczącego wydajności.



4. Architektura mikroserwisu umożliwia łatwą wymianę modeli i dalsze eksperymenty.

## Podsumowanie

Celem projektu było stworzenie systemu predykcyjnego wspomagającego automatyczne wypełnianie formularza dodawania ofert w serwisie noclegowym. Stworzono dwa modele:

- BaseModel – oparty na danych historycznych gospodarzy,
- AdvancedModel – oparty na wektoryzacji treści tekstowych i Random Forest.

Choć wyniki obu modeli były zbliżone, model zaawansowany wykazał nieco lepszą skuteczność, zwłaszcza w predykcji zmiennych kategorycznych. Możliwym kierunkiem dalszego rozwoju byłoby rozdzielenie problemu na dwa osobne modele – jeden wyspecjalizowany w przewidywaniu zmiennych tekstowych, drugi w wartościach liczbowych. Takie podejście mogłoby pozwolić na lepsze dopasowanie algorytmów do charakterystyki danych i potencjalnie poprawić ogólną jakość predykcji.