Projekt IUM

Dokumentacja wstępna

- *Michał Pałasz*
- *Michał Sadlej*

Definicja problemu biznesowego:

Potrzeba automatyzacji i usprawnienia procesu wypełniania pól podczas dodawania nowych ofert w celu zwiększenia efektywności, zmniejszenia liczby błędów oraz poprawy doświadczenia użytkownika.

Zadania modelowania

1. Automatyczne uzupełnianie danych podstawowych:
    a. przewidywanie wartości pól na podstawie danych użytkownika (np. adres) oraz kontekstu oferty.
2. Klasyfikacja/kategoryzacja ofert:
    a. automatyczne przypisywanie oferty do odpowiedniej kategorii (np. apartament, pokój) na podstawie opisu.
3. Sugerowanie wartości dla pól opisowych:
    a. generowanie rekomendacji dla pól tekstowych na podstawie podobnych, wcześniej utworzonych ofert.
4. Wykrywanie potencjalnych błędów:
    a. identyfikacja niespójności lub brakujących danych podczas wypełniania formularza.

Założenia projektu

1. Wzorce wypełniania formularzy przez użytkowników wykazują pewną powtarzalność i strukturę możliwą do wychwycenia przez algorytmy uczenia maszynowego.
2. Dane historyczne z recenzji zawierają informacje pozwalające określić, które parametry oferty przyczyniają się do jej sukcesu.
3. Użytkownicy będą akceptować sugestie systemu, jeśli będą one trafne i oszczędzające czas.
4. Automatyzacja nie musi być kompletna - częściowe wypełnienie formularza również przyniesie wartość biznesową.
5. System musi być intuicyjny i przyjazny dla oferentów, którzy mogą nie mieć doświadczenia w dodawaniu ofert, a jednocześnie powinien umożliwiać ręczne nadpisywanie automatycznych sugestii, aby zachować pełną kontrolę dla użytkowników nieufnych wobec automatyzacji.
6. System nie może pogorszyć jakości danych.

Proponowane kryteria sukcesu:

Kryteria biznesowe:

1. Łączny czas przeznaczony na dodanie oferty do systemu.
2. Zwiększenie liczby nowych ofert - zauważalny wzrost po wdrożeniu rozwiązania.
3. Redukcja liczby negatywnych opinii spowodowanych nieporozumieniami między klientem a właścicielem obiektu.

Kryteria analityczne:

1. Dokładność predykcji:
    a. określa odsetek pól pozostawionych bez zmian przez użytkownika po ich automatycznym wypełnieniu przez algorytm
    b. ponad 50%
2. Pokrycie formularza:
    a. procent pól, dla których system jest w stanie zaproponować wartości
    b. ponad 60%
3. Czas predykcji:
    a. czas potrzebny na wygenerowanie sugestii
    b. mniej niż 1s

Weryfikacja baseline'u:

1. Porównanie z modelem bazowym, działającym na podstawie danych historycznych (np. najbardziej popularna kategoria, średnia cena).
2. Model zaawansowany (wykorzystujący uczenie maszynowe) powinien wykazywać znaczne polepszenie wyników względem bazowego.

Analiza danych:

Users.csv:

1. Zakładamy, że 'Id' użytkownika powinno być unikalne.
2. Rekordy z pustym polem 'Id' użytkownika mogą być odrzucone.
3. Dane adresowe powinny być poprawne oraz kompletne.

Reviews.csv:

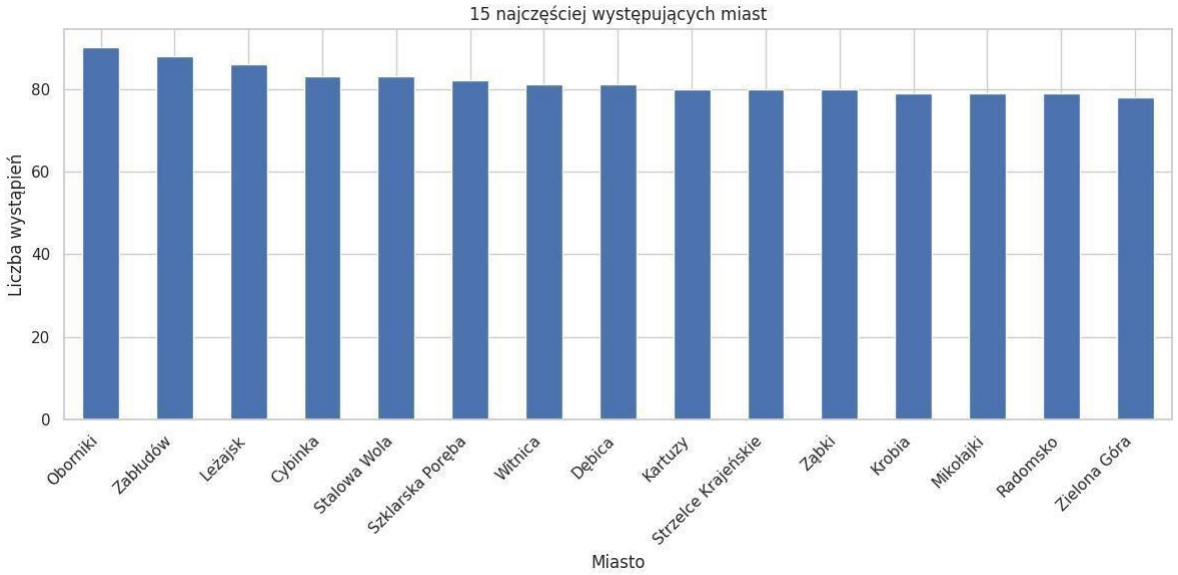
1. 'Listing_id' powinno odpowiadać odpowiedniemu 'Id' oferty w pliku 'listings.csv'.
2. Data i dane autora powinny być poprawne.
3. Pola 'comment' nie powinny być puste. Puste komentarze mogą być odrzucone.

Listings.csv:

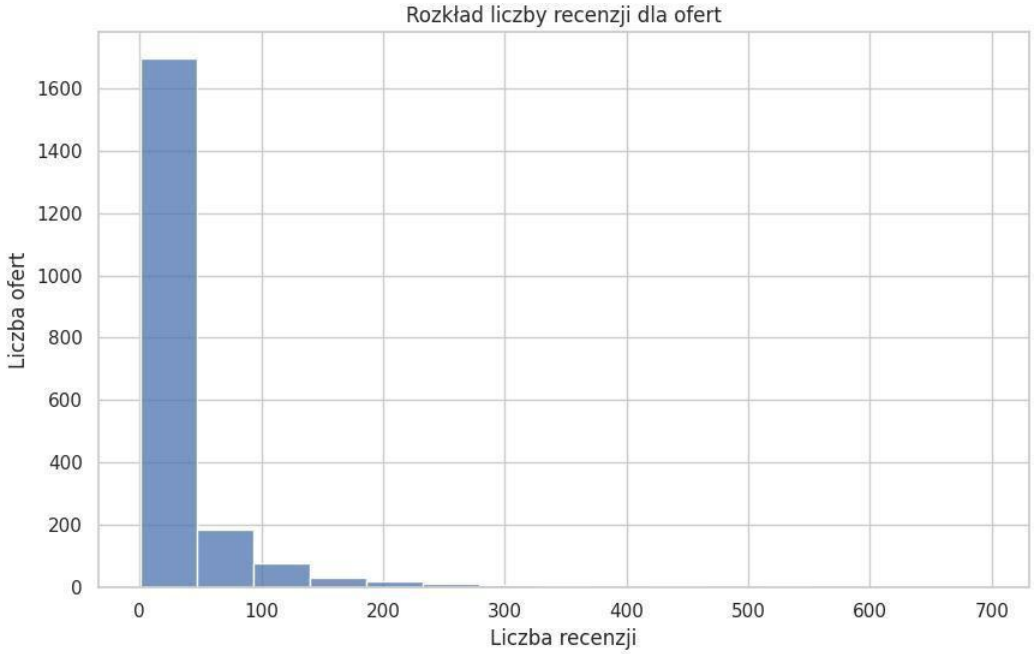
1. W tym pliku powinny znajdować się wartości pól z szczegółami ofert oraz unikalnym ‘Id’ oferty.

Rozkłady kluczowych atrybutów:

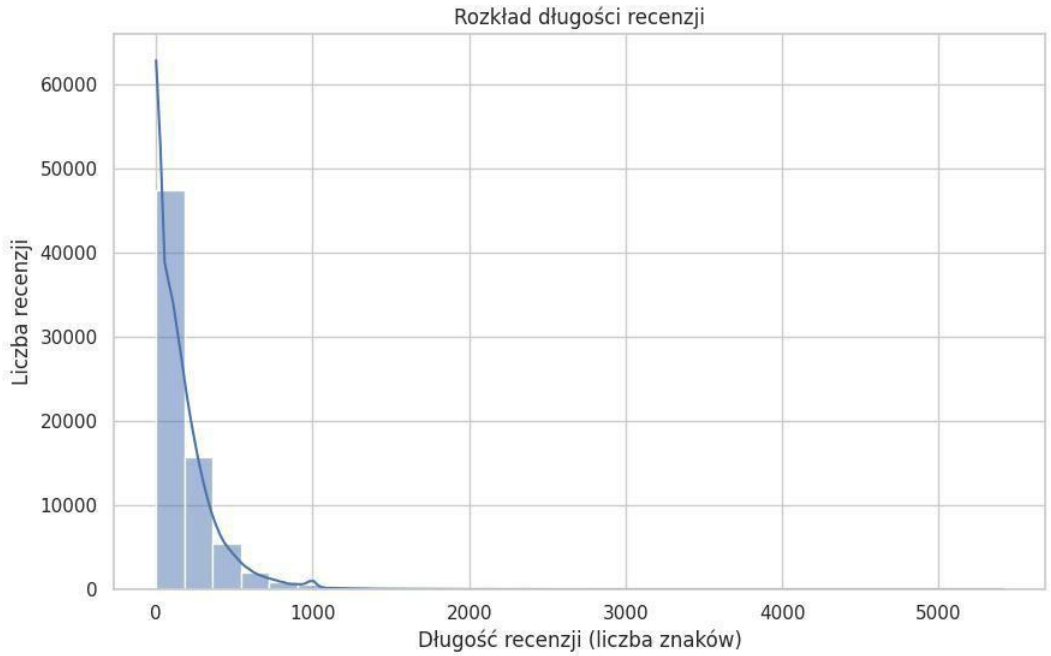
1. Najczęściej występujące miasta



2. Rozkład liczby recenzji dla ofert



3. Rozkład długości recenzji



4. Najbardziej popularne rodzaje nieruchomości