# Inżynieria Uczenia Maszynowego

Etap 1

*Michał Matak, Jakub Robaczewski*

### Polecenie od klienta:

*“Mamy co prawda dodatkowe benefity dla naszych najlepszych klientów, ale może dałoby się ustalić kto potencjalnie jest skłonny wydawać u nas więcej?”*

### Zadanie biznesowe

Sugerowanie klientów, którzy mogą wrócić do serwisu.

### Zadanie modelowania:

Model regresyjny, szacujący prawdopodobieństwo powrotu klienta do serwisu.

### Założenia:

- Mamy dostęp do danych dotyczących:
    - przebiegu sesji użytkowników na serwerze
    - dostaw
    - produktów
    - użytkowników
- Klient poruszając się po stronie internetowej generuje pewne zdarzenia zapisane m.in. w logach dotyczących sesji
- Użytkownicy powracający do naszego sklepu posiadają pewne charakterystyczne cechy, które wskazują, że dla nich prawdopodobieństwo powrotu do sklepu jest większe

### Testowanie:

Model będziemy testować poprzez wycięcie pewnego odcinka czasu z końca szeregu czasowego i potraktowanie go jako zbiór testowy

### Kryteria sukcesu

A) Efektywność na zbiorze testowym lepsza niż stałe zwracanie jednej wartości
B) Powierzchnia pola pod krzywą ROC powyżej 0.5

### Wymaganie techniczne

1. projekt realizowany w pythonie
2. wstępnie przetwarzanie batchowe

## Dane:

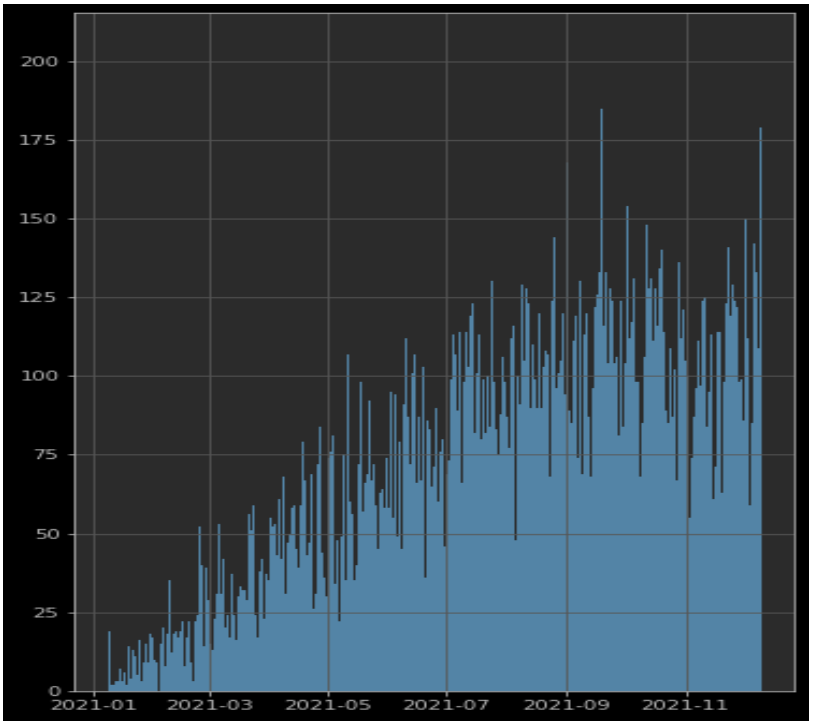
Ewidentne błędy w danych (w samej tylko liście dotyczącej produktów):

- za duże ceny
- ceny ujemne
- ceny z wieloma miejscami po przecinku

## Informacje o danych (pochodzące z 3 iteracji):

**Dane sesji:**

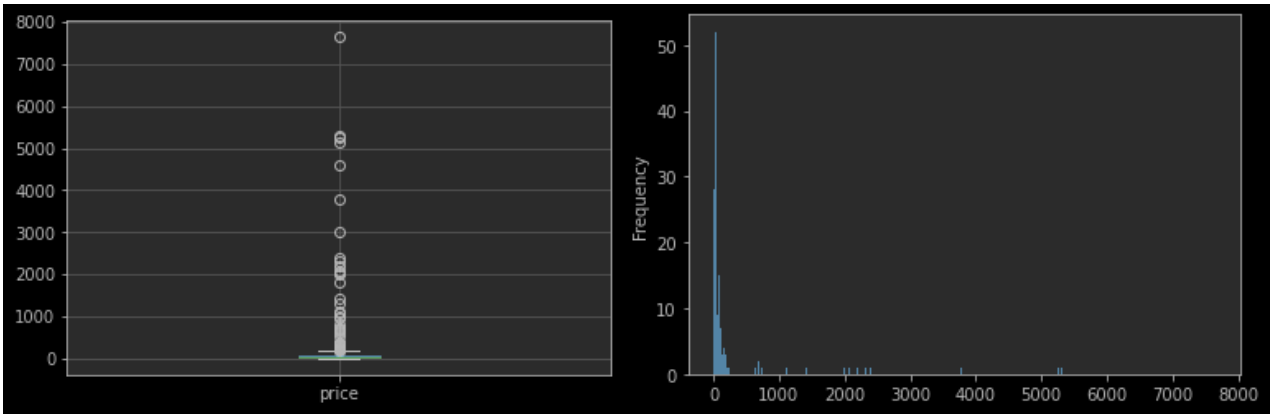
- 24574 rekordy
- informacje o 6758 sesjach
- Typ danych:
    - session_id - int64 - kolejne liczby całkowite zaczynające się od 124 do 6881
    - user_id - int64
    - product_id - int64
    - offered_discount - int64
    - purchase_id - float64
    - event_type - string przyjmujący wartość VIEW_PRODUCT lub BUY_PRODUCT
    - timestamp - string odpowiadający za przechowywanie informacji o dacie
- W 20961 rekordach product_id jest równy NULL - odpowiada to 20961 eventom typu VIEW_PRODUCT
- dane nieposortowane po czasie
- podczas 3613 z nich dokonano przynajmniej jednego zakupu
- dane z pochodzą 335 dni (pierwsze z 8 stycznia, ostatnie z 10 grudnia)
- rozkład danych (na lewej osi liczba eventów wyemitowanych w danym dniu, na dolnej osi - liczba dni):



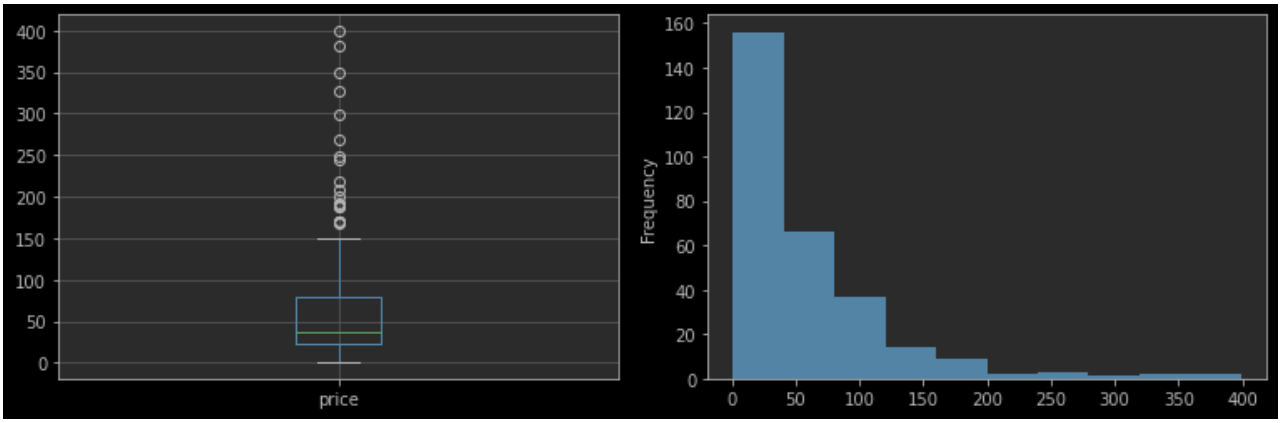

- Liczba eventów emitowanych w ciągu dnia rośnie w przybliżeniu liniowo do przełomu sierpnia i września, następnie utrzymuje się na podobnym poziomie

**Dane o produktach:**

- 319 rekordów
- Typ danych
    - product_id - int64 - zaczynający się od 1001 rosnący o 1
    - product_name - string opisujący nazwę produktu
    - category_path - string opisujący miejsce produktu w drzewie kategorii
    - price - float64
- Brak brakujących wartości
- Rozkład ceny w produktach:

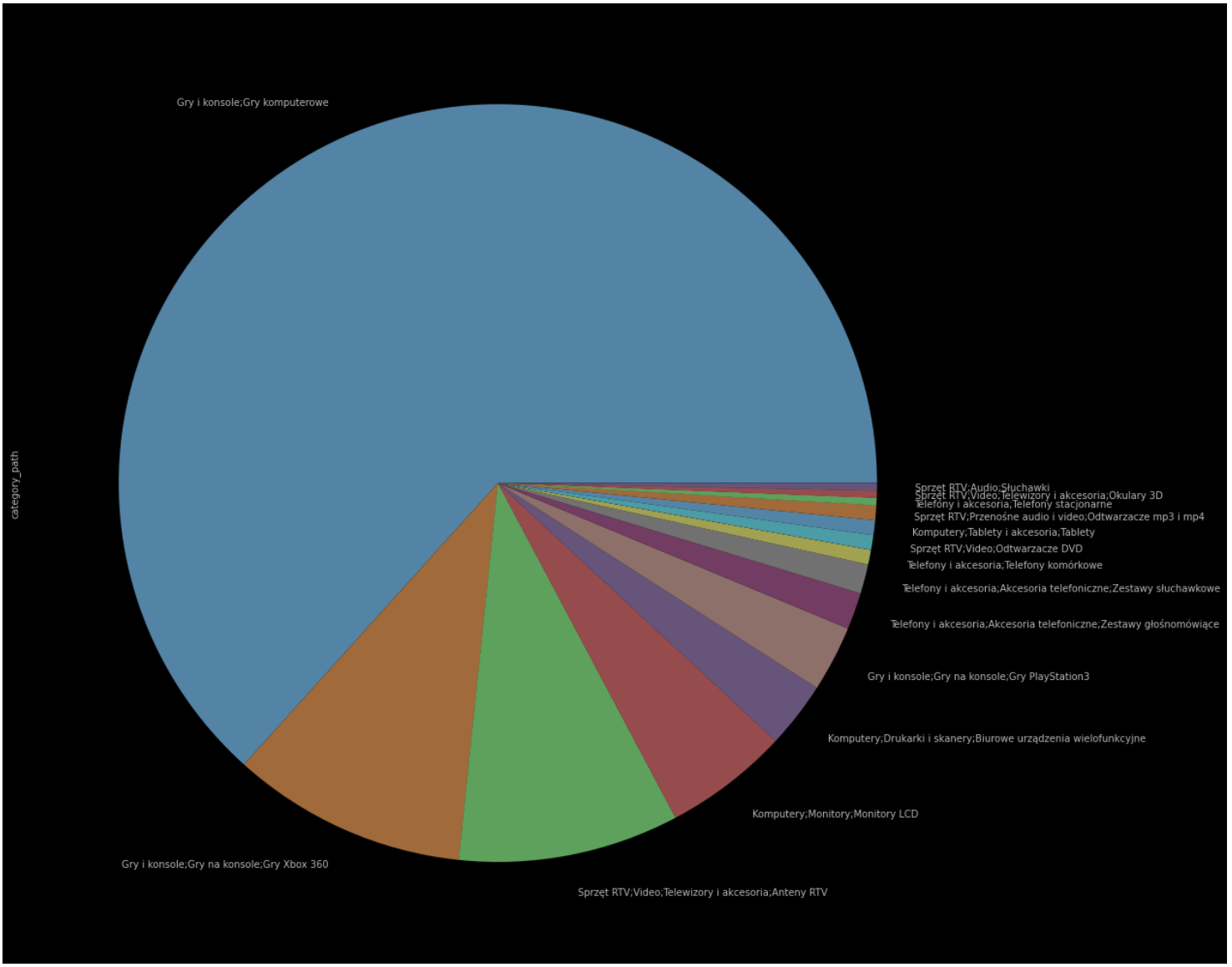



- Jak widać na powyższych wykresach, występuje tam kilka produktów o bardzo wysokiej cenie (27 na 319 wszystkich produktów), które zaburzają wizualizację rozkładu. Poddaliśmy zatem analizie rozkład cen wśród produktów o wartości poniżej 500 zł:



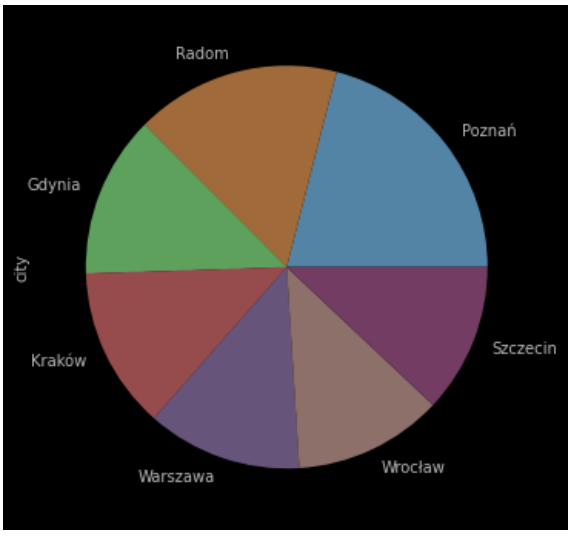

Na podstawie tych danych jesteśmy w stanie stwierdzić, że większość produktów w sklepie jest warta około 50 złotych, a ich liczba zmniejsza się wraz ze wzrostem ceny

- Podział na kategorie produktów



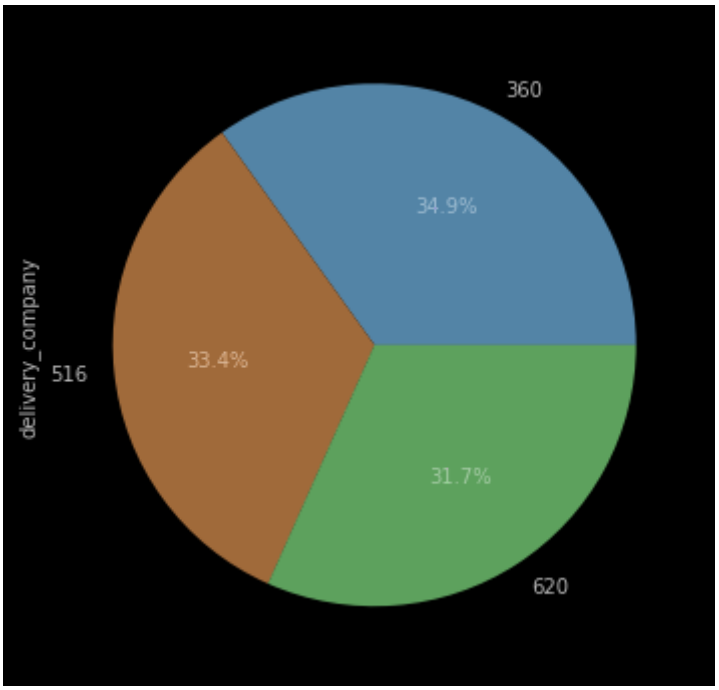

**Dane użytkowników:**

- 200 rekordów
- Typ danych:
  - user_id - int64 - zaczynający się od 102 rosnący o 1
  - name - string opisujący imię i nazwisko
  - city - string opsiujący miasto
  - street - string opisujący ulicę
- Brak brakujących wartości
- Brak powtarzających się adresów
- Miasta z których pochodzą użytkownicy:



**Dane dostaw:**

- 3613 rekordów
- Typ danych**:**
    - purchase_id - int 64 zaczyna się od 20000 rośnie o 1
    - purchase_timestamp - string opisujący moment zakupu
    - delivery_timestamp - string opisujący moment dostawy
    - delivery_company - int_64 - opisuje prawdopodobnie identyfikator dostawcy
- dane nieposegregowane po czasie (zarówno dostawy jak i zakupu)
- realizacja dostaw przez dostawców




jak widać na powyższym wykresie udział dostawców w dostawach był sobie równy

**Dane w kontekście predykcji**:
Aby określić wartościowość danych w kontekście predykcji należy określić horyzont czasowy w jakim będzie określane prawdopodobieństwo powrotu klienta do serwisu. Na przestrzeni całego roku (wszystkich danych) tylko 4 klientów z 200 odwiedziło sklep tylko jeden raz (innymi słowy nie powrócili do serwisu), co stanowi około 2% wszystkich klientów. Tak mały odsetek może utrudnić zadanie. Jednakże na tym tle gdy weźmiemy pod uwagę tylko jeden miesiąc (w tym przypadku to był wrzesień) już 31 użytkowników z 181 dokonało zakupu tylko jeden raz, zatem powołując się na pierwotne polecenie klienta istnieje szansa na wskazanie tych klientów, którzy do nas wrócą, ale moglibyśmy przyspieszyć ten proces co byłoby równoważne temu, że kupowaliby więcej.