

# **ResellPredictor – Przewidywanie cen obuwia kolekcjonerskiego na rynku wtórnym przy użyciu uczenia maszynowego.**

Przedmiot: Podstawy Sztucznej Inteligencji  
Autorzy: Antoni Wojcieszek, Igor Schaffer



# Autorzy projektu

**Antoni Wojcieszek** - EDA, ocena jakości i analiza wyników

**Igor Schaffer** - przygotowanie danych, implementacja modeli i eksperymenty

Decyzje projektowe oraz raport z realizacji były tworzone wspólnie



# Opis Problemu

## Cel Projektu:

Zbudowanie modelu regresyjnego, który na podstawie cech obuwia (marka, model, rozmiar, data premiery, cena sklepową) przewidzi cenę odsprzedaży (Resell Price) na platformie aukcyjnej.

## Oczekiwany wynik:

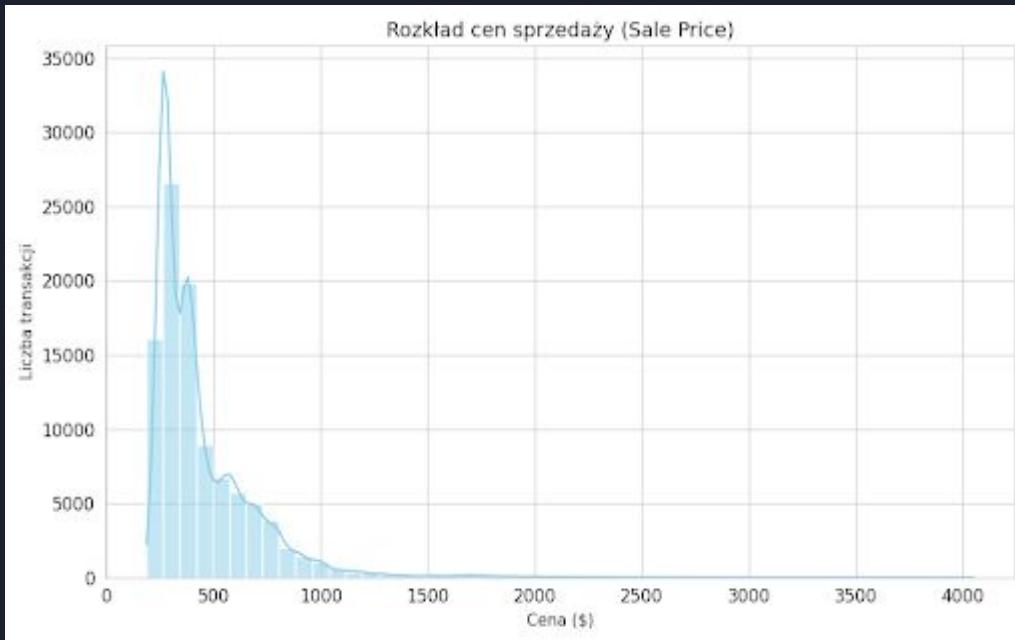
Wytrenowany model uczenia maszynowego realistycznie przewidujący cenę odsprzedaży pary butów oraz analiza czynników wpływających na wzrost wartości obuwia (tzw. "hype").

## Co ma wnieść wynik i jakie jest jego znaczenie:

Pomocne narzędzie dla inwestorów, które pozwala oszacować potencjalny zysk (ROI) nowych par mających premierę w przyszłości. Model pozwala również zidentyfikować czynniki najbardziej wpływające na cenę (rozmiar, marka, model) a także zautomatyzować proces wyceny dla sklepów zajmujących się odsprzedażą butów

# Exploratory Data Analysis (cz. 1)

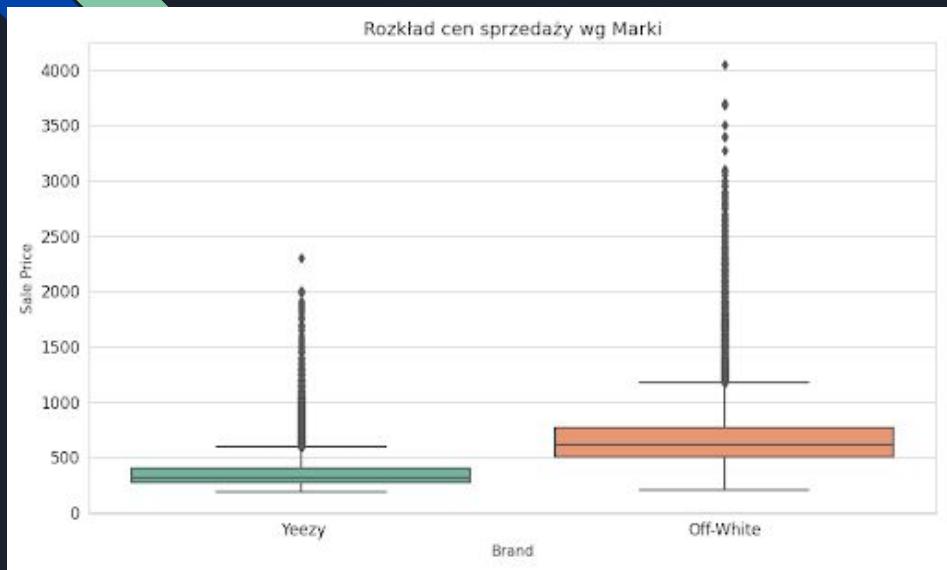
Wykorzystano zbiór StockX Data Contest 2019 zawierający ok. 100 000 transakcji sprzedaży sneakersów w latach 2017-2019 ze strony StockX, zajmującej się odsprzedażą butów a także innych pożądanych przedmiotów przez użytkowników platformy



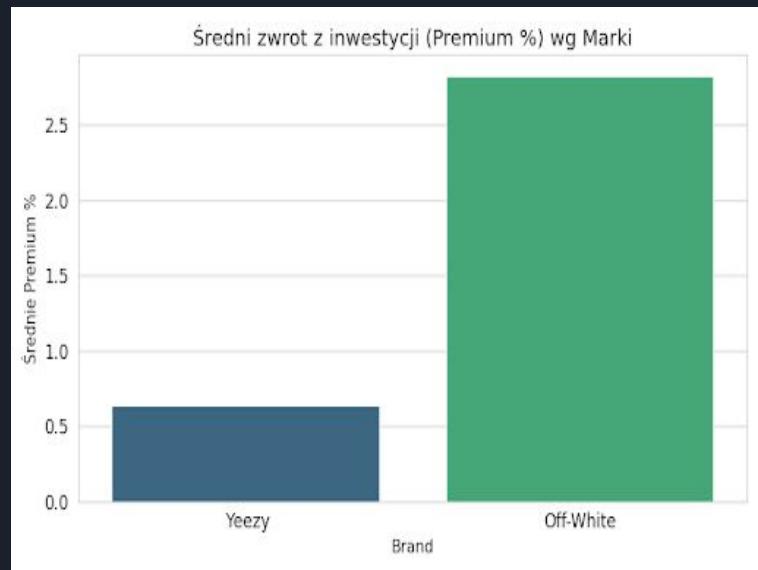
Zbiór zawiera dane dotyczące sprzedaży butów marki Adidas Yeezy, oraz Nike w kolaboracji z firmą Off-White. Posiada on następujące cechy: Order Date, Brand, Sneaker Name, Sale Price, Retail Price, Release Date, Shoe Size i Buyer Region

Ad 1. Na wykresie możemy zauważyć, że rozkład cen jest silnie prawostronnie skośny - większość par butów kosztuje pomiędzy 200 a 400 dolarów, ale istnieją również modele sprzedawane za kwoty powyżej 1000 USD.

# Exploratory Data Analysis (cz. 2)



Ad 2. Na tym wykresie można zauważyc, że marka Off-White ma znacznie wyższą medianę ceny niż Yeezy. Dodatkowo wykres dla marki Off-White jest bardziej "rozciągnięty" - co oznacza, że ceny są mniej stabilne. Niektóre z nich sprzedają się za ogromne kwoty, inne za dużo mniejsze. Buty Yeezy natomiast są bardziej "zbite" - przewidywalne, choć na niższym poziomie.



Ad 3. Do zbioru dodano nową zmienną - "Premium%" - będącą różnicą ceny odsprzedaży i ceny detalicznej pary, podzieloną przez cenę detaliczną. Na wykresie możemy zauważyc, że buty marki Off-White generują średnio wyższy zwrot z inwestycji (ROI) niż Yeezy.



# Ocena wyników

Do oceny jakości modeli zostały wykorzystane standardowe miary:

1. **MAE** - Mean Absolute Error - Średni Błąd Bezwzględny

Ta miara mówi, o ile dolarów średnio myli się model. Jest łatwa w interpretacji, ponieważ zwraca bezpośrednią odpowiedź: "Model myli się o X dolarów".

2. **RMSE** - Root Mean Squared Error - Pierwiastek Błędu Średniokwadratowego

Mierzy przeciętną wielkość błędu predykcji, przy szczególnej wrażliwości na duże błędy (outliery).

3. **R kwadrat** - współczynnik determinacji

Określa, jaki procent zmienności ceny jest wyjaśniany przez model



# Opis zastosowanych metod

## **Metoda 1: Random Forest Regressor**

Tworzy wiele drzew decyzyjnych na losowych podzbiorach danych i uśrednia ich wyniki. Metoda ta jest dosyć prosta, dobrze radzi sobie z danymi o nieliniowej zależności i jest odporna na przeuczenie.

## **Metoda 2: Gradient Boosting Regressor**

Buduje drzewa sekwencyjnie, gdzie każde kolejne koryguje błędy poprzednich. Zazwyczaj osiąga wysoką dokładność i dobrze nadaje się do modelowania złożonych zależności.



# Eksperymenty (cz. 1)

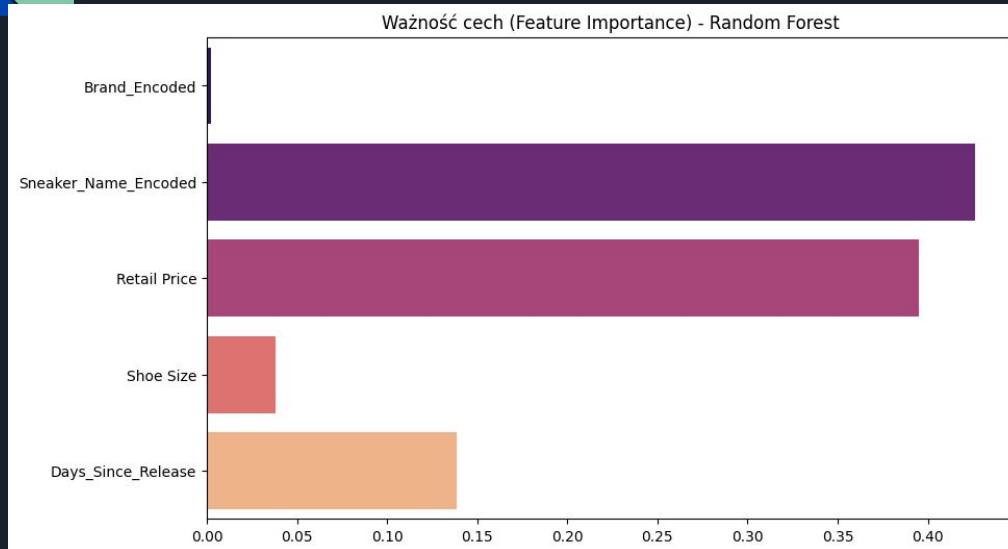
## Inżynieria cech

- Przetestowano wpływ dodatkowej zmiennej **Days\_Since\_Release** i okazało się, że czas, który minął od premiery jest istotnym czynnikiem wpływającym na cenę
- Nazwy marek i modeli zostały zakodowane w celu lepszego przetworzenia przez modele
- Wyczyszczono dane zawierające błędy ("śmieciowe dane")

## Porównanie modeli

- Trening obu modeli został przeprowadzony na tym samym podziale danych - 80% trening, 20% test
- Dla obu metod ustawiono liczbę drzew na 100, w celu zapewnienia porównywalnych warunków

# Eksperymenty (cz. 2)



Ad 4. Wykres przedstawiający stopień w jakim dana cecha wpływa na wynik modelu

## Analiza ważności cech

Według modelu, zdecydowanie najważniejszymi cechami wpływającymi na cenę sprzedaży jest konkretny model buta, jego cena detaliczna oraz liczba dni od daty wydania na rynek. Marka buta staje się zbędną informacją, ponieważ jest już zawarta w nazwie modelu. Rozmiar buta ma dużo mniejszy wpływ na cenę niż pozostałe cechy, chociaż nadal ma on pewien wpływ (popularne lub trudno dostępne rozmiary mają wyższą cenę ze względu na wysoki popyt/niską podaż)

# Analiza uzyskanych wyników

Tabela przedstawia porównanie metod na podstawie wybranych miar oceny wyników.

Metoda	MAE (\$)	RMSE (\$)	R2
<b>Random Forest</b>	<b>16.90</b>	<b>35.99</b>	<b>0.9835</b>
<b>Gradient Boosting</b>	<b>45.48</b>	<b>84.63</b>	<b>0.9086</b>

Model **Random Forest** okazał się skuteczniejszy w tym przypadku.

- Średni Błąd Bezwzględny wynosi niecałe 17 USD, co oznacza, że średnia różnica między rzeczywistą ceną sneakersów a predykcją modelu wynosi 17 dolarów. W kontekście cen sięgających często kilkuset dolarów lub więcej, jest to satysfakcjonujący wynik.
- R “kwadrat” na poziomie 0,98 oznacza bardzo dobre dopasowanie do danych.



# Rekomendacja

Przy tworzeniu projektu, udało się:

- Skutecznie wyczyścić dane i stworzyć zmienne czasowe
- Zidentyfikować kluczowe czynniki wpływające na cenę
- Uzyskać wysoką skuteczność predykcji

Kierunki dalszych prac:

- Wzbogacenie danych o analizę wpisów z mediów społecznościowych, aby móc wykrywać nastawienie społeczności do różnych modeli “hype” i ocenić jak przekłada się on na wynik końcowy
- Dodanie analizy wizualnej zdjęć butów, by sprawdzić wpływ kolorystyki na cenę



# Bibliografia

- Zbiór danych StockX Data Contest 2019  
[https://www.kaggle.com/datasets/hudsonstuck/stockx-data-contest\](https://www.kaggle.com/datasets/hudsonstuck/stockx-data-contest)

Dziękujemy!

