# Przewidywanie wartości zamówień

*kompleksowa analiza i uzasadnienie*

*Tomasz Boguszewski (s25237)*
*Mikołaj Polecki (s23034)*
*Oskar Paciorkowski (s25488)*
*Kamil Kłodawski (s24777)*
*Mikhai Kramushchanka (s24124)*

# Spis treści

**1. Streszczenie..... 3**
**2. Definicja problemu.....3**
**3. Opis zbiorów danych.....3**
3.1. Zestaw danych "customers_orders"..... 3
3.2. Zestaw danych "pages"..... 3
**4. Wstępna obróbka danych..... 4**
4.1. Łączenie zbiorów danych.....4
4.1.1. Uzasadnienie..... 4
4.2. Usunięcie skorelowanych cech.....4
4.2.1. Uzasadnienie..... 5
4.3. Zbalansowanie danych..... 5
4.3.1. Uzasadnienie..... 6
4.4. Podział zbioru na zbiory treningowy i testowy..... 6
4.4.1. Uzasadnienie..... 7
4.5. Wyłączenie wybranych cech.....8
4.5.1. Uzasadnienie..... 8
4.6. Zmniejszenie złożoności danych..... 9
4.6.1 Uzasadnienie..... 9
4.7. Wygenerowanie nowych cech..... 9
4.7.1. Pairwise Linear Combinations..... 10
4.7.1.1. Uzasadnienie..... 10
4.7.2. Pairwise Polynomial Combinations.....10
4.7.2.1. Uzasadnienie..... 10
**5. Wybór modelu uczenia maszynowego.....11**
5.1. Opis wybranego modelu..... 11
5.2. Wyniki wybranego modelu.....12
5.2.1. Wyniki innych modeli..... 13
5.2.2. Zastosowane parametry wybranego modelu.....14
5.2.3. Wpływ poszczególnych cech..... 14

# 1. Streszczenie

Celem tego projektu było opracowanie modelu predykcyjnego, zdolnego do oszacowania wartości zamówień na podstawie danych o użytkowniku. Wykorzystanie mocy uczenia maszynowego miało na celu zapewnienie sklepowi e-commerce praktycznego narzędzia do optymalizacji strategii marketingowych i ostatecznie zwiększenia rentowności i zysku. Niniejszy raport przedstawia kompleksową analizę projektu, skrupulatnie wyszczególniając przesłanki stojące za każdą decyzją projektową i metody zastosowane do osiągnięcia pożądanych celów.

# 2. Definicja problemu

Projekt dotyczył problemu regresji, w którym głównym celem było opracowanie modelu predykcyjnego zdolnego do dokładnego oszacowania docelowej cechy "order_value" (wartość zamówienia) na podstawie obszernego zbioru danych. To podejście do modelowania predykcyjnego wykorzystywało zaawansowane techniki uczenia maszynowego do identyfikowania skomplikowanych wzorców i relacji w danych, umożliwiając szacowanie wartości zamówień dla nowych interakcji z zarówno obecnymi, jak i nowymi klientami.
Problemy regresji dobrze nadają się do przewidywania wyników, w których celem jest przewidywanie ciągłego wyniku liczbowego, takiego jak wartość zamówienia, na podstawie zestawu cech wejściowych. Określając wyzwanie jako zadanie regresji, projekt mógł wykorzystać moc algorytmów uczenia maszynowego zaprojektowanych specjalnie do tego celu, zapewniając dokładne i wiarygodne prognozy.

# 3. Opis zbiorów danych

W projekcie wykorzystano dwa odrębne zbiory danych jako podstawę do analizy:

## 3.1. Zestaw danych "customers_orders"

Ten zbiór danych zawierał wiele informacji, w tym datę zamówienia, płeć, pierwszą odwiedzoną przez klienta stronę, wykorzystany kod kuponu, liczba wcześniejszych zamówień dokonanych przez klienta, wiek, źródło ruchu, identyfikator źródła ruchu, user agent (agent użytkownika) i docelową, przewidywaną wartość: kwotę zamówienia.

## 3.2. Zestaw danych "pages"

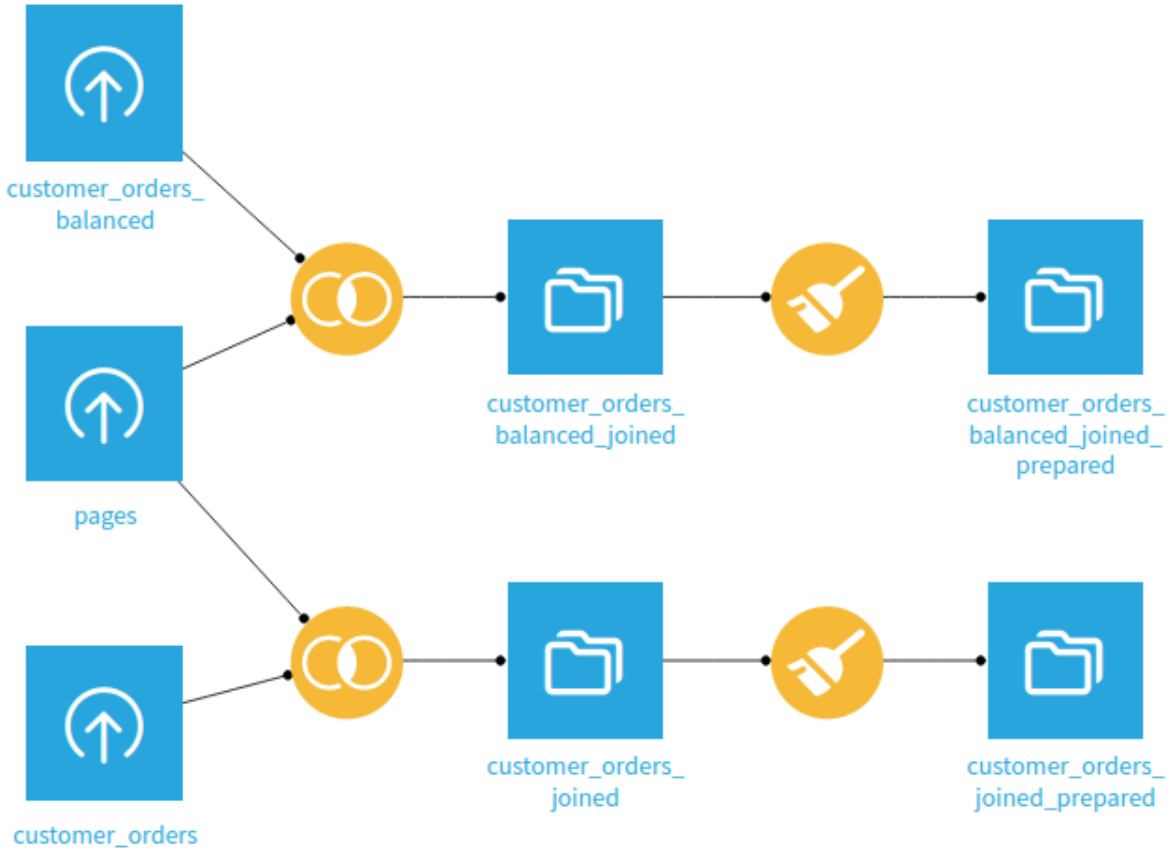
Ten dodatkowy zbiór danych zawierał informacje opisowe na temat różnych stron odwiedzanych przez klientów, składające się z kolumn page_id i page_name.

# 4. Wstępna obróbka danych

Aby zapewnić optymalny stan danych do rozpoczęcia fazy uczenia maszynowego, przeprowadzono fazę przetwarzania wstępnego:

## 4.1. Łączenie zbiorów danych

Zbiory danych "customers_orders" i "pages" zostały złączone w celu włączenia nazw stron odpowiadających identyfikatorom stron odwiedzanych przez klientów, wzbogacając dostępne informacje o przyjazne dla użytkownika nazwy stron zastępujące identyfikatory liczbowe.

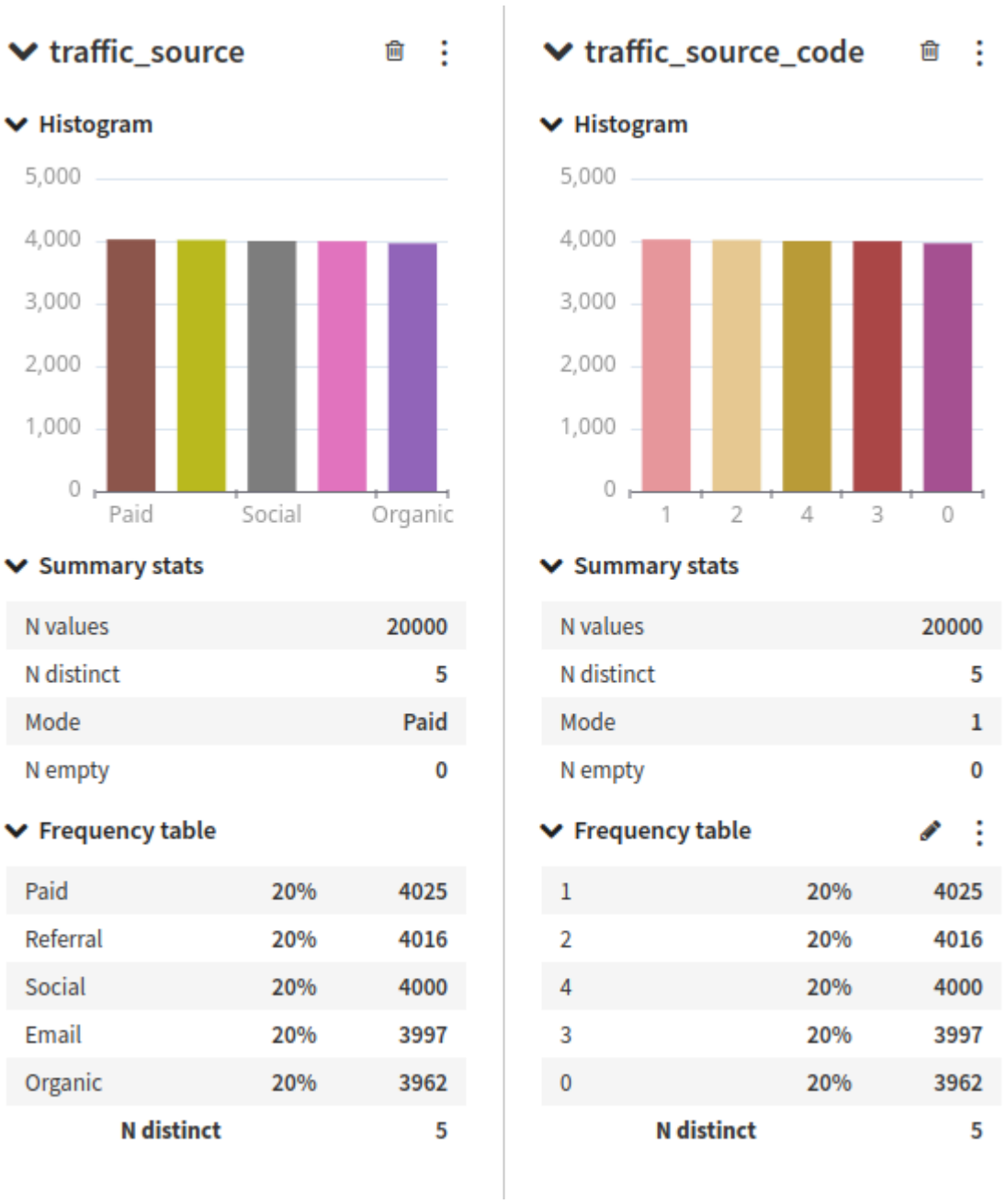



### 4.1.1. Uzasadnienie

Łącząc te dwa zbiory danych, pozwoliło wykorzystać opisowe nazwy stron, w celu uzyskania głębszego wglądu w zachowania i preferencje klientów. Ten dodatkowy kontekst pozwala lepiej zrozumieć dane, co może przyczynić się do stworzenia dokładniejszego przewidywania wartości zamówienia.

## 4.2. Usunięcie skorelowanych cech

Aby złagodzić zwiększyć wydajność modelu, usunięto skorelowaną kolumnę "traffic_source_code", ponieważ jej informacje zostały już przechwycone przez kolumnę "traffic_source". Korelację danych widać na poniższym obrazku:



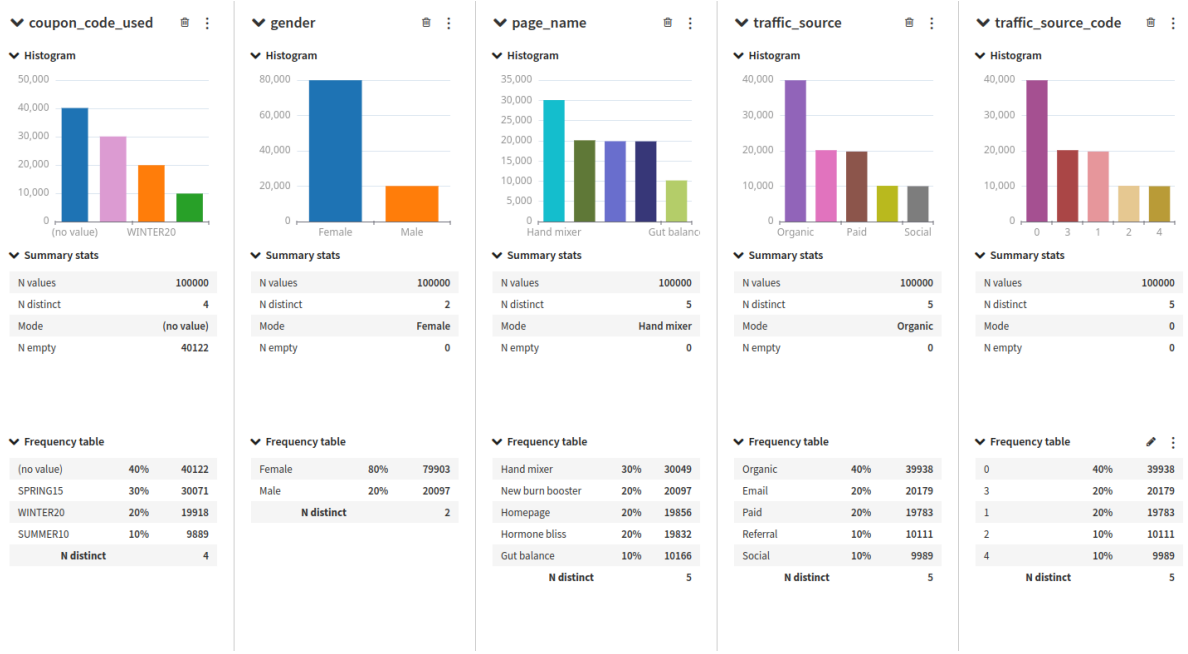

### 4.2.1. Uzasadnienie

Wysoce skorelowane cechy mogą wprowadzać wieloliniowość, co może niekorzystnie wpływać na wydajność i interpretowalność modelu. Usunięcie nadmiarowej kolumny "traffic_source_code" miało na celu usprawnienie zbioru danych i wyeliminowanie potencjalnych źródeł szumu lub stronniczości w modelu.

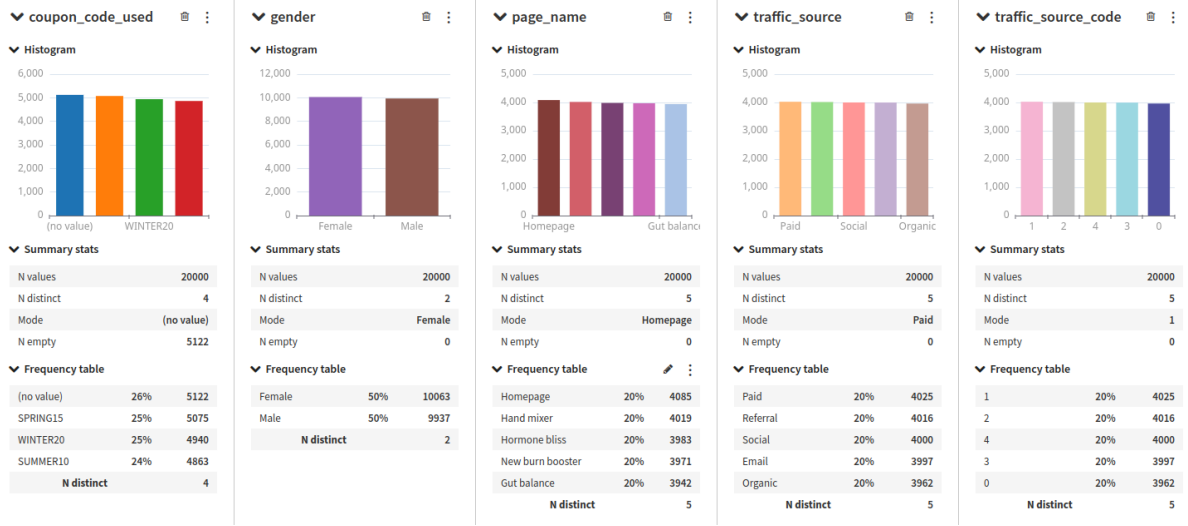
## 4.3. Zbalansowanie danych

Biorąc pod uwagę potencjalny wpływ niezrównoważonych danych na wydajność modelu, opracowano oddzielny zestaw treningowy przy użyciu skryptów w Pythonie. Ten krok zapewnił, że model został wytrenowany na reprezentatywnym rozkładzie danych, zmniejszając tendencyjność i poprawiając jego możliwości pracy na uogólnionych zbiorach danych.

Poniższa grafika przedstawia rozkład danych przed balansacją:




A poniższa grafika rozkład danych w zbalansowanym zbiorze:



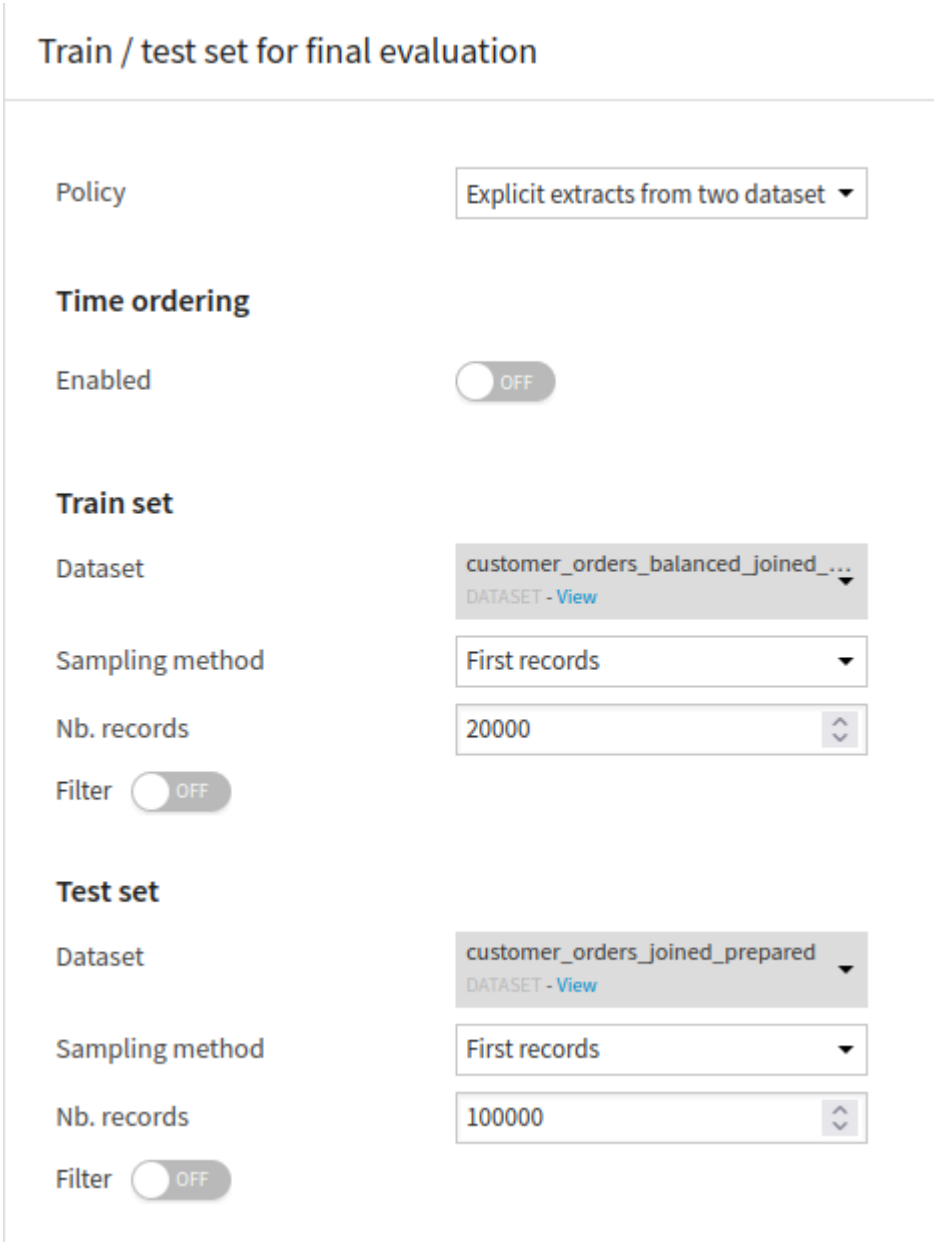

### 4.3.1. Uzasadnienie

Niezbalansowane zestawy danych mogą prowadzić do uzyskania tendencyjnych modeli, które mają trudności z dokładnym przewidywaniem klas mniejszościowych lub wartości ekstremalnych. Stworzenie zrównoważonego zestawu danych treningowy miało na celu złagodzenie tej kwestii, zapewniając, że model był narażony na zróżnicowany zakres wartości zamówienia podczas procesu szkolenia, co ostatecznie poprawiło jego zdolność do uogólniania i dokonywania dokładnych prognoz w całym spektrum wartości zamówienia.

## 4.4. Podział zbioru na zbiory treningowy i testowy

Aby dokładnie ocenić wydajność modelu, oryginalny zestaw danych został podzielony na dwa podzbiory: zrównoważony zestaw danych służył jako zestaw treningowy, ułatwiając uczenie się modelu, podczas gdy oryginalny niezrównoważony zestaw danych został wyznaczony jako zestaw testowy, umożliwiając realistyczną ocenę zdolności predykcyjnych

modelu na niewidocznych danych. Wadą zbalansowania zbioru treningowego było zmniejszenie liczby danych z 100 tysięcy rekordów w zbiorze testowym do 20 tysięcy rekordów w zbiorze treningowym, co może negatywnie wpłynąć na wyniki przewidywań.

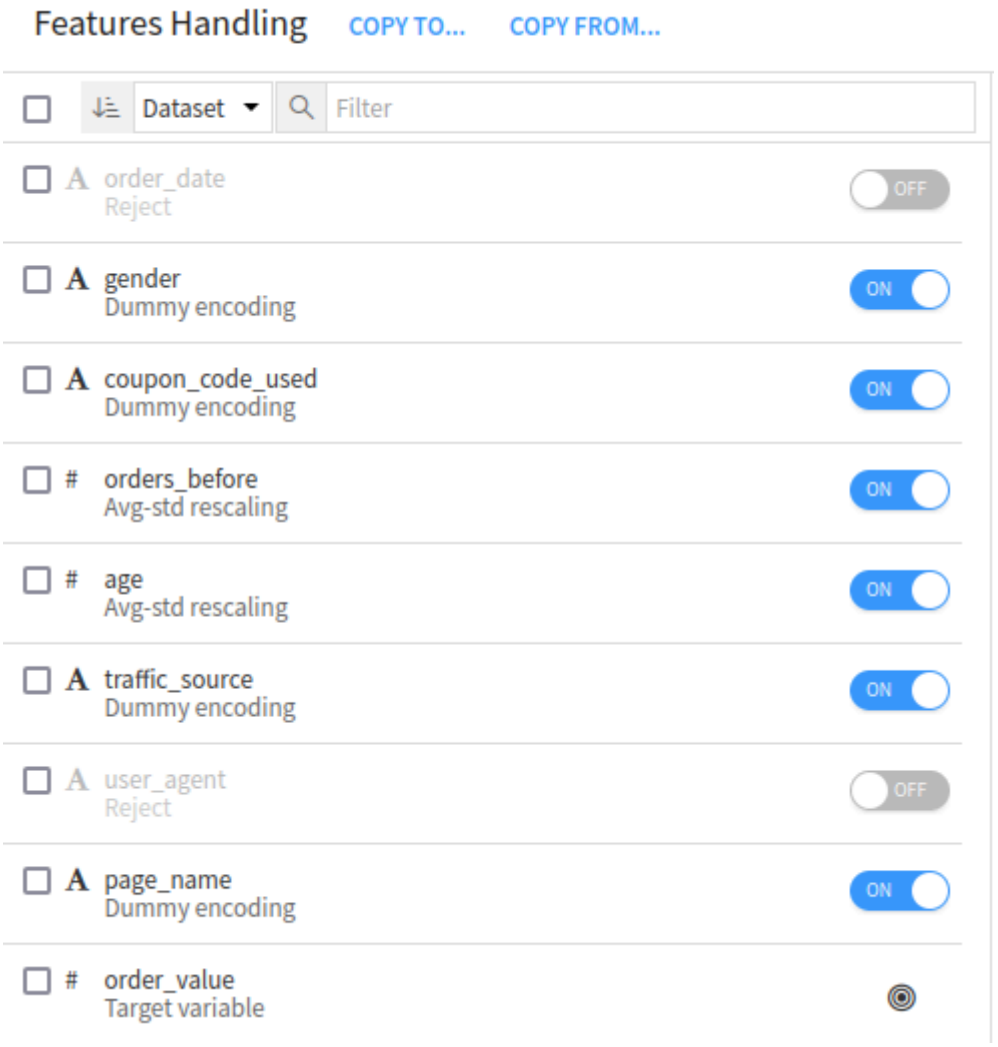



### 4.4.1. Uzasadnienie

Rozdzielenie zbioru danych na odrębne zestawy treningowe i testowe jest kluczowym krokiem w uczeniu maszynowym, aby zapewnić, że model jest oceniany na danych, na które nie był narażony podczas procesu uczenia. Takie podejście pomaga ocenić zdolność modelu do uogólniania na nowe dane, zapewniając realistyczne oszacowanie jego wydajności w rzeczywistych scenariuszach.

## 4.5. Wyłączenie wybranych cech

Po analizie danych ustalono, że cechy "order_date" i "user_agent" nie miały wpływu na przewidywaną wartość zamówień. W związku z tym te cechy zostały wykluczone z procesu uczenia maszynowego.

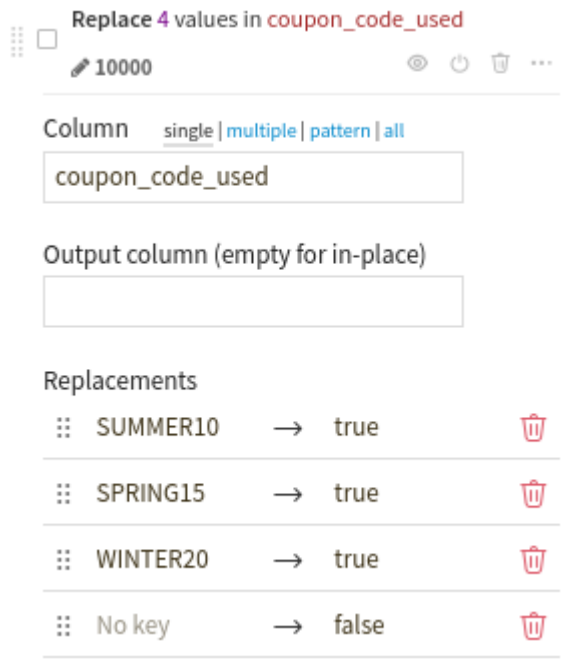



### 4.5.1. Uzasadnienie

Włączenie nieistotnych lub zbędnych cech do modelu uczenia maszynowego może niepotrzebnie zwiększyć złożoność modelu, potencjalnie prowadząc do nadmiernego dopasowania lub słabej wydajności na ogólnych zbiorach danych. Wyłączając cechy "order_date" i "user_agent", które uznano za mało znaczące dla wartości zamówień, proces ten miał na celu uproszczenie modelu i skupienie się na najbardziej istotnych cechach, poprawiając jego wydajność, ułatwiając analizę wyników i zapobiegając przetrenowaniu.

## 4.6. Zmniejszenie złożoności danych

Zespół projektowy podjął decyzję o przekształceniu cechy "coupon_code_used" z jej oryginalnej reprezentacji kategorialnej (kod kuponu) na binarny format logiczny (prawda/fałsz). Ta transformacja była motywowana spostrzeżeniem, że konkretny kod kuponu nie miał znaczącego wpływu na wartość zamówienia, lecz na samą moc predykcyjną gdy kod kuponu był obecny lub nie.

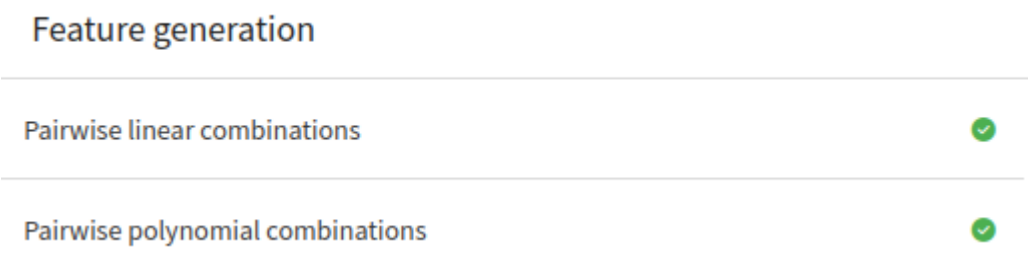



### 4.6.1 Uzasadnienie

Zastosowanie tej transformacji miało na celu uproszczenie reprezentacji cech, zmniejszenie wymiarowości, zachowanie istotnych informacji o wykorzystaniu kuponów przy jednoczesnym wyeliminowaniu niepotrzebnej ziarnistości, umożliwienie wydajnego kodowania i obsługi cech binarnych przez modele uczenia maszynowego oraz potencjalną poprawę wydajności modelu poprzez redukcję cech kategorycznych o wysokiej kardynalności.

## 4.7. Wygenerowanie nowych cech

Aby jeszcze bardziej zwiększyć skuteczność predykcyjną modelu, zastosowano następujące techniki generacji nowych cech:




Zastosowanie tych technik inżynierii cech było kluczowym krokiem w maksymalizacji możliwości predykcyjnych modelu, wykorzystując w pełni bogate informacje zawarte w zbiorze danych. Włączenie tych cech inżynieryjnych miało na celu zapewnienie modelowi

kompleksowej reprezentacji podstawowych wzorców i relacji, ostatecznie zwiększając jego zdolność do dokładnego i wiarygodnego przewidywania wartości zamówień.

### 4.7.1. Pairwise Linear Combinations

Podejście to obejmowało tworzenie nowych cech poprzez łączenie istniejących cech za pomocą operacji liniowych. Poprzez uchwycenie potencjalnych interakcji i relacji między zmiennymi, te zaprojektowane cechy miały na celu dostarczenie modelowi dodatkowych spostrzeżeń i wzorców w celu poprawy dokładności przewidywania.

#### 4.7.1.1. Uzasadnienie

Wiele zjawisk w świecie rzeczywistym wykazuje złożone relacje i interakcje między zmiennymi, które mogą nie zostać uchwycone przez poszczególne cechy. Tworzenie parami liniowych kombinacji istniejących cech miało na celu uchwycenie tych skomplikowanych relacji i zapewnienie modelowi bardziej kompleksowej reprezentacji podstawowych wzorców w danych, potencjalnie prowadząc do poprawy dokładności przewidywania.

### 4.7.2. Pairwise Polynomial Combinations

Opierając się na technice kombinacji liniowych, która wygenerowała nowe cechy, łącząc istniejące już cechy za pomocą operacji wielomianowych. Umożliwiło to modelowi uchwycenie nieliniowych relacji i złożonych interakcji w danych, potencjalnie odkrywając skomplikowane wzorce, które mogłyby dalej udoskonalić przewidywania wartości zamówienia.

#### 4.7.2.1. Uzasadnienie

Podczas gdy relacje liniowe są powszechne w wielu zbiorach danych, zjawiska w świecie rzeczywistym często wykazują nieliniowe wzorce i interakcje. Włączenie kombinacji wielomianów parami miało na celu uchwycenie tych nieliniowych relacji, umożliwiając modelowi lepsze reprezentowanie i uczenie się na podstawie złożonej dynamiki obecnej w danych, potencjalnie prowadząc do dokładniejszych przewidywań wartości zamówienia.

# 5. Wybór modelu uczenia maszynowego

W projekcie zastosowano Gradient Boosted Trees, potężną technikę uczenia zespołowego, jako wybrany model uczenia maszynowego do przewidywania wartości zamówień. Decyzja ta była podyktowana zdolnością algorytmu do radzenia sobie z nieliniowymi zależnościami, złożonymi interakcjami oraz jego odpornością na wartości odstające i szum - cechy powszechnie obecne w rzeczywistych zbiorach danych, takich jak dane dotyczące wartości zamówień w handlu elektronicznym.

## 5.1. Opis wybranego modelu

Gradient Boosted Trees wykazuje się doskonałą wydajnością w zadaniach regresji, dając zwykle wyższe wyniki niż popularne modele jak Random Forest, szczególnie w przypadku danych wielowymiarowych i złożonych. Dodatkowo, interpretowalność algorytmu poprzez wyniki ważności cech zapewnia cenny wgląd w najbardziej wpływowe czynniki przyczyniające się do przewidywania wartości zamówień.
Co więcej, algorytm Gradient Boosted Trees płynnie radzi sobie z brakującymi danymi i cechami kategorycznymi, eliminując potrzebę obszernego wstępnego przetwarzania danych i pozwalając zespołowi projektowemu skupić się na głównym zadaniu modelowania.

## 5.2. Wyniki wybranego modelu

Wybór Gradient Boosted Trees w projekcie został potwierdzony przez wysokie wskaźniki wydajności modelu, w tym imponujący wynik $R^2$ wynoszący 0,8367, wskazujący na silną korelację między przewidywanymi a rzeczywistymi wartościami zamówień. Inne godne uwagi wskaźniki obejmowały średni błąd bezwzględny (MAE) wynoszący 16,16, średni bezwzględny błąd procentowy wynoszący 3,15%, średni błąd kwadratowy (MSE) wynoszący 410, pierwiastek ze średniej kwadratowej błędu (RMSE) wynoszący 20,26, pierwiastek ze średniej kwadratowej błędu logarytmicznego (RMSLE) wynoszący 0,03953 oraz współczynnik korelacji Pearsona wynoszący 0,9149.

Metrics and assertions

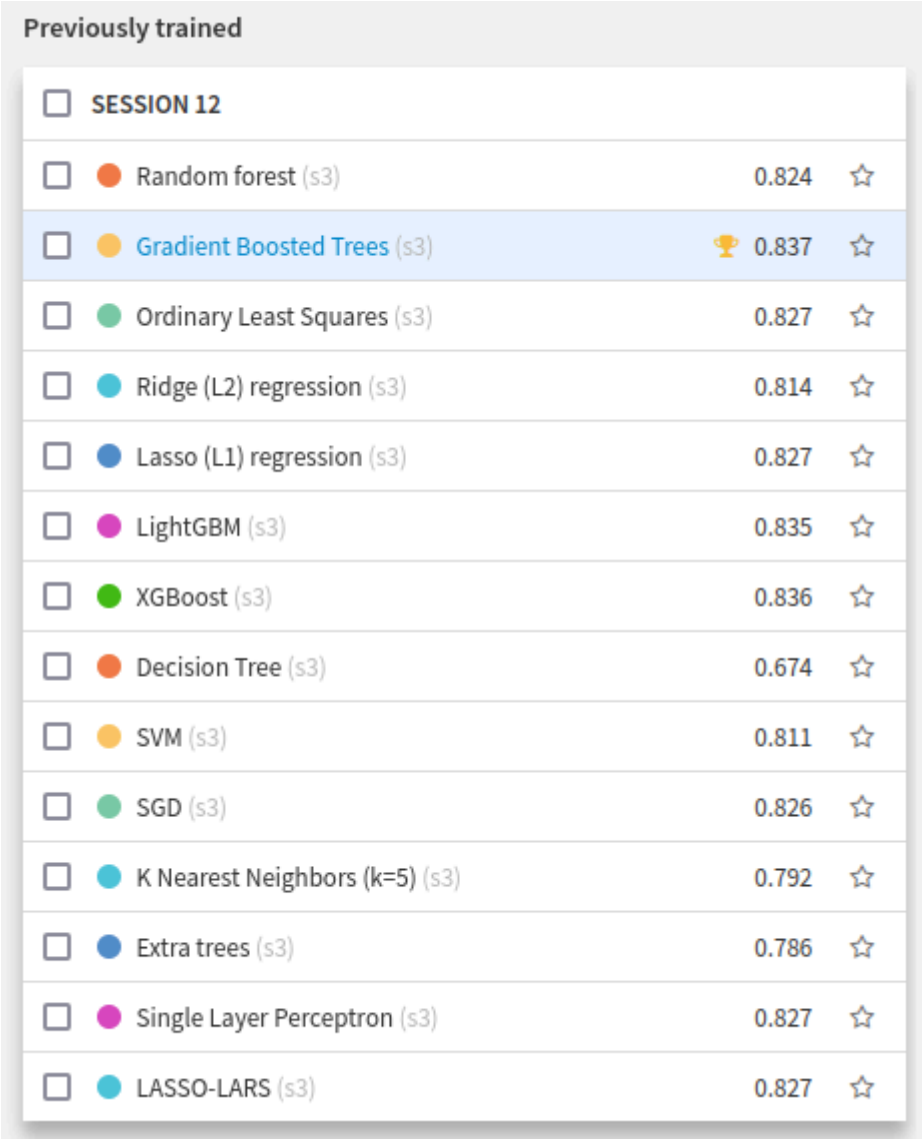
**Detailed metrics**

| | |
|---|---|
| Explained Variance Score | **0.8367** |
| Mean Absolute Error (MAE) | **16.16** |
| Mean Absolute Percentage Error | **3.15%** |
| Mean Squared Error (MSE) | **410** |
| Root Mean Squared Error (RMSE) | **20.26** |
| Root Mean Squared Logarithmic Error (RMSLE) | **0.03953** |
| Pearson coefficient | **0.9149** |
| R2 Score | **0.8367** |

Te znakomite wyniki wykazały zdolność modelu do dokładnego uchwycenia podstawowych wzorców i relacji w danych, prowadząc do precyzyjnych i wiarygodnych prognoz wartości zamówienia. Wykorzystanie mocnych stron Gradient Boosted Trees zapewni sklepowi e-commerce potężne narzędzie do podejmowania decyzji opartych na danych i optymalizacji strategicznej, co ostatecznie zwiększy rentowność i sukces w konkurencyjnym środowisku e-commerce.

### 5.2.1. Wyniki innych modeli

Najważniejszą metryką przy ocenie modelu była metryka $R^2$. Wyniki innych modeli uczenia maszynowego przedstawiono na obrazku poniżej:



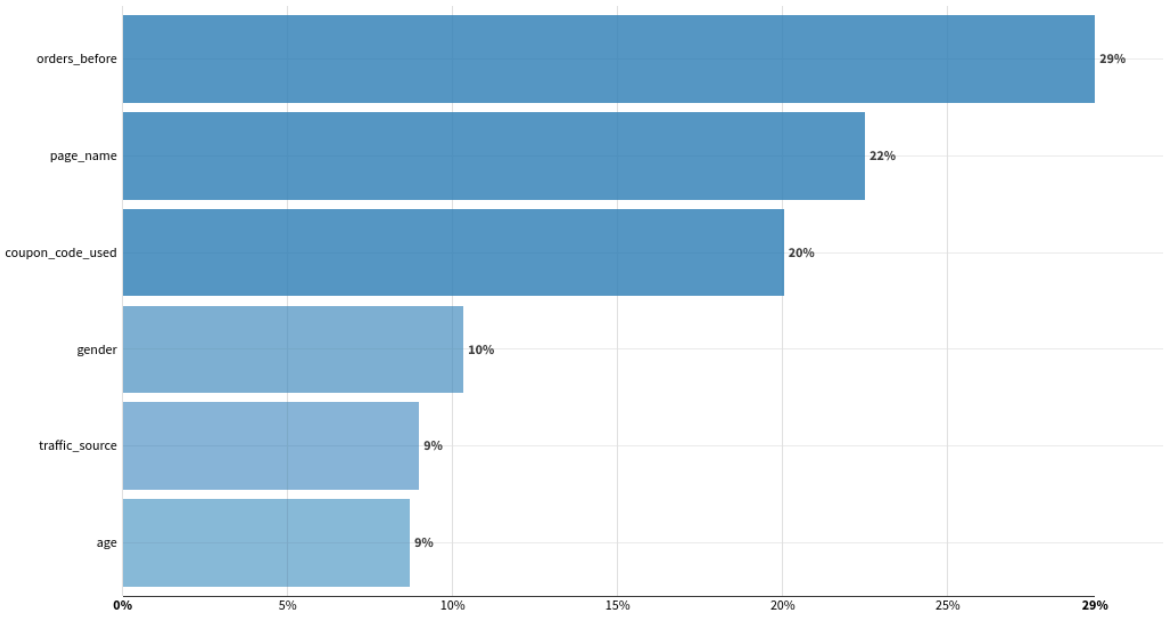

Wyniki o podobnej skuteczności uzyskały również algorytmy XGBoost i LightGBM. Jednak mimo iż oba modele są znane ze swojej wysokiej wydajności, czasami mogą być postrzegane jako modele "czarnej skrzynki", co utrudnia zrozumienie leżącego u ich podstaw procesu decyzyjnego. Decydując się na bardziej przejrzyste podejście Gradient Boosted Trees, zespół projektowy mógł lepiej interpretować zachowanie modelu, przeanalizować znaczenie cech i uzyskać wgląd w relacje między cechami a wartościami zamówień.

### 5.2.2. Zastosowane parametry wybranego modelu

| | |
|---|---|
| **Algorithm** | Gradient Boosted Trees (Regression) |
| **Loss** | Least Square |
| **Feature sampling strategy** | Default |
| **Number of boosting stages** | 100 |
| **Eta (learning rate)** | 0.1 |
| **Max trees depth** | 3 |
| **Minimum samples at leaf** | 1 |

### 5.2.3. Wpływ poszczególnych cech

Algorytm uczenia maszynowego za najważniejszą cechę wybrał liczbę zamówień złożonych do tej pory przez klienta. Inne ważne wskaźniki to nazwa pierwszej odwiedzanej strony i czy został użyty kupon. Wpływ poszczególnych cech przedstawiono na poniższych grafikach:



Feature effects