IUM - Koncepcja

Oczekiwanie klienta

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

Definicja problemu biznesowego

Przewidywanie ilość wydanych pieniędzy przez danego klienta w nadchodzącym miesiącu, na podstawie aktywności z poprzedniego miesiąca.

Obecna sytuacja

ESzoppping korzysta obecnie z pewnego modelu predykcyjnego. Ilość generowanych, trafnych przewidywań dla danego miesiąca plasuje się na poziomie minimum 70%. Za poprawne przewidywanie uznawana jest wartość, które nie różni się od wartości rzeczywistej o więcej niż 35%.

Firma chciałaby znacząco zwiększyć odsetek prawidłowych przewidywań dla danego miesiąca, podnosząc obecny minimalny poziom o 20 punktów procentowych.

Biznesowe kryterium sukcesu

Ilość poprawnych przewidywań dla danego miesiąca jest minimum na poziomie 90%. Za poprawne przewidywanie uznajemy wartości, które nie odbiegają od rzeczywistych wartości o więcej niż 35%.

Analityczne kryterium sukcesu

Poprawność przewidywania zdefiniowana jako błąd względny δ:

$$\delta = \frac{|pw - wr|}{wr} * 100\%$$

gdzie:

- *pw* - przewidywane wydatki danego klienta w następnym miesiącu,
- *wr* - rzeczywiste wydatki danego klienta w następnym miesiącu,
- $\delta \leq 35\%$ zgodnie z kryterium biznesowym.

Skuteczność modelu odzwierciedlona jako procent poprawnych przewidywań α:

$$\alpha = \frac{tp}{wp} * 100\%$$

gdzie:

- *tp* - ilość trafnych przewidywań,
- *wp* - ilość wszystkich przewidywań,
- $\alpha \geq 90\%$ zgodnie z kryterium biznesowym.

Definicja zadania modelowania

Model będzie realizować **zadanie regresji**. Na podstawie aktywności danego użytkownika z poprzedniego miesiąca (np.: ilość wydanych pieniędzy, przeglądanych ofert, zakupionych produktów) i innych danych model prognozuje ilość wydanych pieniędzy przez klienta w nadchodzącym miesiącu. Wyższa wartość prognozy oznacza większą skłonność do wydania więcej pieniędzy przez klienta sklepu.

Dane wykorzystane do analizy

- Baza użytkowników:
  - Id użytkownika
  - Miasto zamieszkania

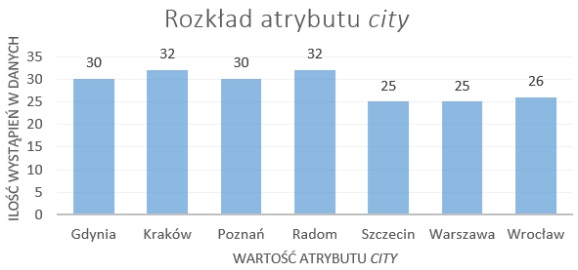
- Historia sesji
  - Id użytkownika
  - Timestamp
  - Typ aktywności
  - Oferowana zniżka
  - Id produktu
  - Id sesji

- Baza produktów
  - Id produktu
  - cena produktu

Wstępna analiza danych - iteracja 1

- *Baza użytkowników*

Nie wykryto błędów ani braków w wartościach atrybutów. Rozkład osób w zależności od miejsca zamieszkania jest równomierny:



Wątpliwość wzbudza jedynie kwestia reprezentatywności danych - wszyscy klienci z bazy pochodzą z większych miast. Miejsce zamieszkania może wpływać na skłonność do wydawania pieniędzy, w związku z czym przewidywania modelu dla osób z mniejszych miejscowości mogą być nietrafione.

- *Baza produktów*

  Znaleziono błędne i odbiegające od średniej wartości atrybutu *price* dla poszczególnych produktów. Dla niektórych z nich, cena była ujemna (24) lub alarmująco niska (co najmniej 10); dla innych z kolei cena była w milionach czy też w miliardach (12). Nie ma możliwość odtworzenia właściwej ceny w sposób inny niż ponowne zebranie danych.

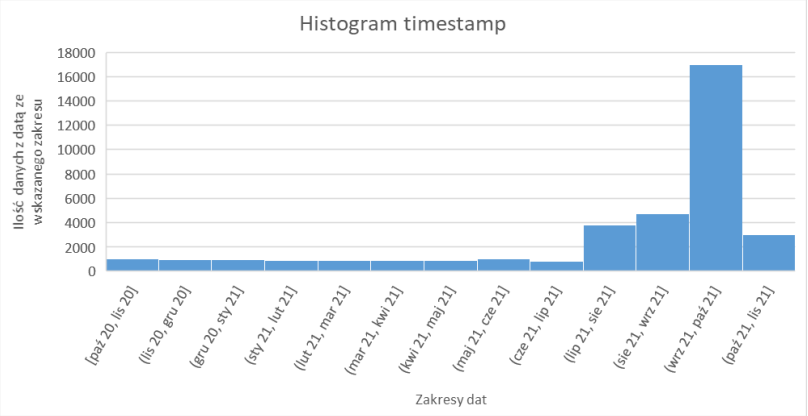
  Produkty z błędnymi wartościami atrybutu *price* stanowią co najmniej 13% wszystkich rekordów (283). Powoduje to, że całkiem sporo danych będzie odrzuconych - usunięcie produktu pociąga za sobą również odrzucenie sesji, w których był on kupowany. Biorąc pod uwagę jak bardzo istotną rolę odgrywa atrybut *price* w wyznaczaniu miesięcznych wydatków klienta, może to poważnie wpłynąć na jakość przewidywań modelu oraz wielkość zbioru dostępnych danych.

- *Historia sesji*

  Zauważono braki wartości w atrybutach *user_id* oraz *product_id*. Około 5% wszystkich sesji nie posiada id użytkownika. Na takim samym poziomie 5% plasuje się część sesji nie posiadającej id produktu. O ile *user_id* można w większości przypadków efektywnie odtworzyć, posługując się identyfikatorem sesji, o tyle w przypadku *product_id* rekonstrukcja się nie powiedzie, w związku z czym wiersze z pustym id produktu zostaną usunięte. W kontekście rozmiaru zbioru sesji (36 862), ten ubytek danych nie ma większego znaczenia.

  Na zamieszczonym poniżej histogramie można zauważyć, że dysponujemy większą ilością danych o aktywności użytkowników sklepu z okresu lipiec-październik 2021 w porównaniu do miesięcy wcześniejszych. Ponadto danych z października jest znacząco więcej niż w pozostałych miesiącach. Nie wiemy czy takie

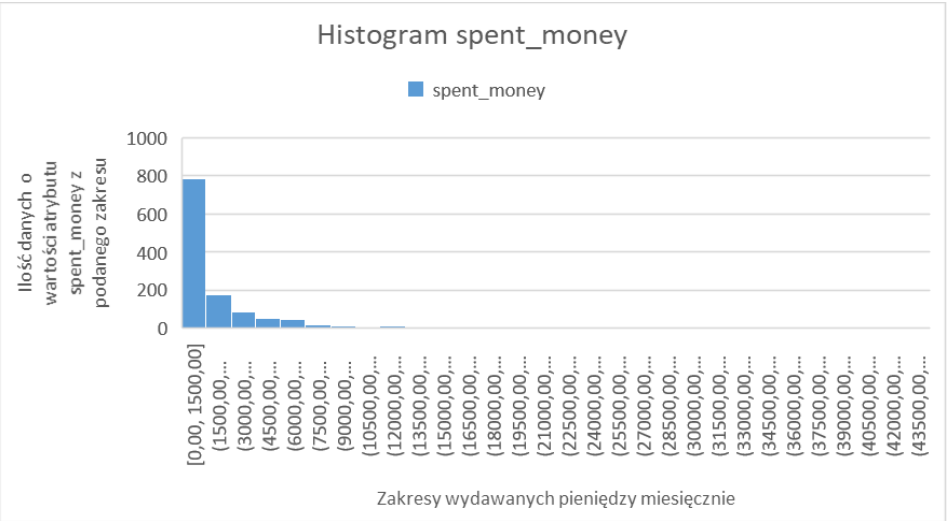
niezbalansowanie wynika z przypuszczalnego wzrostu popularności sklepu w ostatnich miesiącach czy z braku wszystkich danych. Jeżeli powodem jest brak danych to prawdopodobnie będziemy potrzebować więcej danych z wcześniejszych miesięcy.



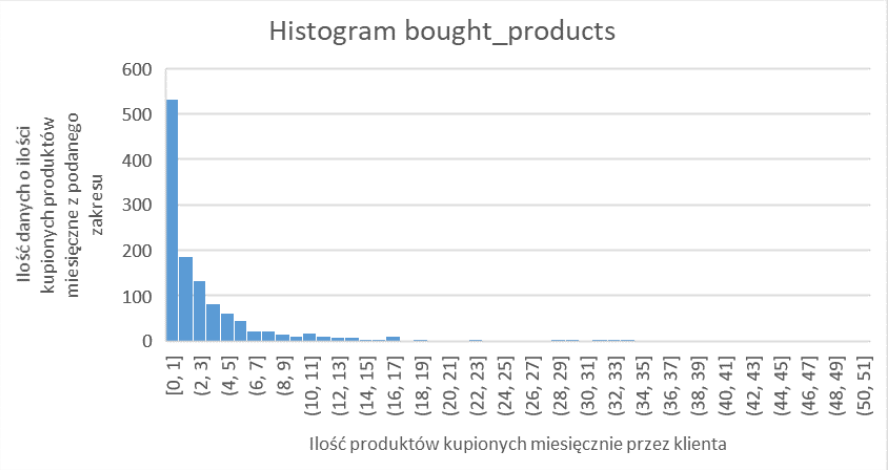
Przekształcenie danych

W celu przeprowadzenia dokładniejszej analizy danych, mając na uwadzę nasze zadanie, dokonaliśmy przekształcenia danych. Korzystając z informacji zawartych w tabelach użytkowników, produktów i sesji wygenerowaliśmy tabelę, która przedstawia miesięczną aktywność poszczególnych użytkowników sklepu. Nowa tabela posiada kolumny: **miasto**, **wydatek pieniędzy w danym miesiącu**, **rok**, **miesiąc**, **liczba zakupionych produktów**, **liczba wyświetleń produktów**, **ilość wydanych pieniędzy** w sklepie przez klienta **w miesiącu następującym**. Taki układ danych pomoże nam w realizacji oraz ocenie predykcji modelu rozwiązującego nasze zadanie.

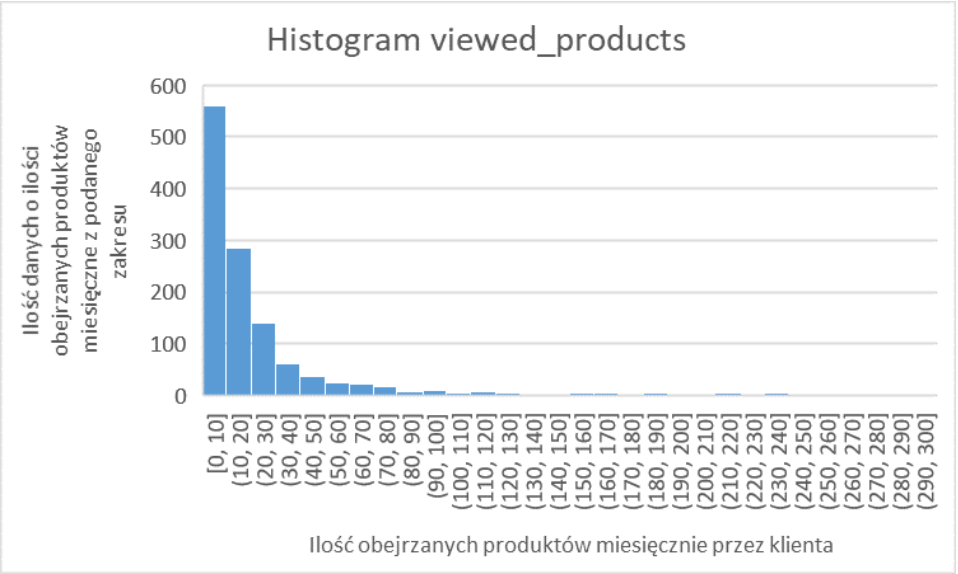
Analiza danych przekształconych



Powyższy histogram pokazuje, z jaką częstotliwością użytkownicy wydają ilości pieniędzy z danego przedziału w ciągu miesiąca. Można wyciągnąć wniosek, że zazwyczaj klienci nie wydają w sklepie więcej niż 1500 zł miesięcznie.



Histogram *bought_products* przedstawia ile produktów najczęściej kupują klienci w ciągu miesiąca - jak często określona ilość produktów zostaje zakupiona przez klientów w ciągu miesiąca. Klienci sklepu najczęściej kupują nie więcej niż 5 produktów miesięcznie.



Z histogramu powyżej można odczytać jak często klienci wyświetlają strony produktów w ciągu miesiąca. Zazwyczaj klient nie wyświetla produktów częściej niż 30 razy w ciągu miesiąca.

W celu zbalansowania liczebności wartości kolumny *spent_money* wartości zostały przeskalowane funkcją logarytmiczną o podstawie 10.

Podsumowanie wstępnej analizy danych

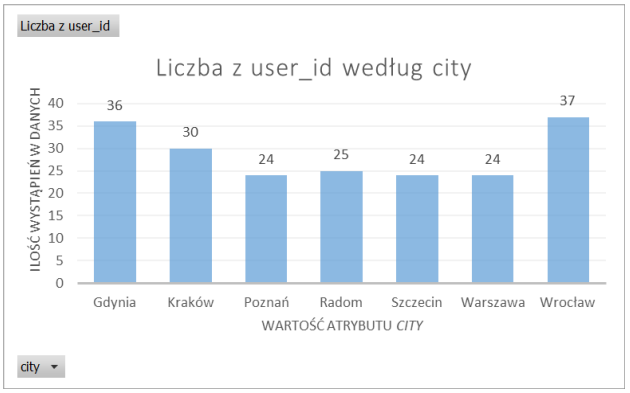
Jeżeli wzrost ilości danych w historii sesji w ostatnich miesiącach nie jest spowodowany wzrostem popularności sklepu, a nie kompletnością danych, to potrzebowalibyśmy więcej danych historii sesji sprzed września/października, żeby lepiej przewidywać wydatki miesięczne klientów.

Część danych wymaga usunięcia. W przypadku niemożność odtworzenia identyfikatora produktu w logach historii sesji, lepiej by było jakbyśmy dostali uzupełnione dane. Natomiast poważnymi błędami są te w cenach produktów - powinny zostać naprawione.

Wstępna analiza danych - iteracja 2

- *Baza użytkowników*

Nie wykryto błędów ani braków w wartościach atrybutów. Rozkład osób w zależności od miejsca zamieszkania jest mniej równomierny niż w poprzedniej iteracji:



Pozostała nierozwiązana kwestia nie reprezentatywności danych - wszyscy klienci z bazy pochodzą z większych miast. Przewidywania modelu mogę się nie sprawdzać w przypadku klientów z mniejszych miejscowości.

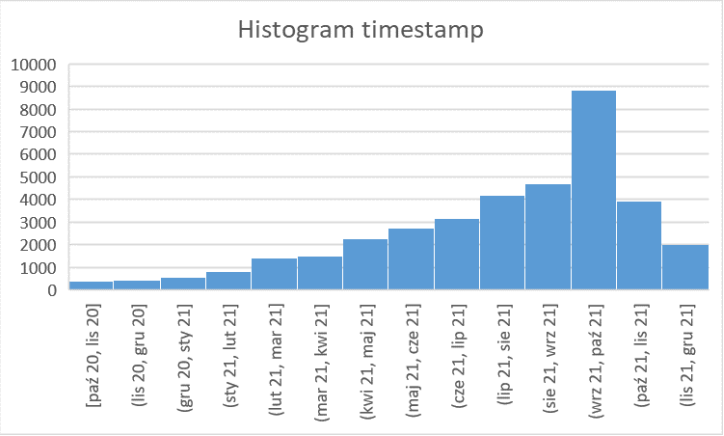
- *Baza produktów*

W nowej iteracji danych zostały naprawione błędy związane z atrybutem *price* - nie zdarza się już, aby przyjmował wartości ujemne lub znacznie odbiegające od średniej. Dla kilku produktów cena wydaje się być nadzwyczaj niska, jednak można przyjąć te wartości za poprawne - być może wynikają one z wyjątkowo dobrej promocji. Błędów ani braków nie znaleziono również w pozostałych analizowanych atrybutach.

- *Historia sesji*

W tej iteracji, braki wartości w atrybutach *user_id* oraz *product_id* występują odpowiednio w ok. 1,6% oraz ok. 1,8% wszystkich sesji. Odsetek wybrakowanych rekordów jest więc na tyle mały, że wiersze z brakującymi wartościami zostaną po prostu odrzucone.

Na podstawie poniższego histogramu można zauważyć, że w tej iteracji posiadamy znacząco więcej danych z miesięcy poprzedzających szczyt popularności sklepu, który przypada na październik 2021 roku. Przewidujemy, że większa ilość tych danych pozytywnie wpłynie na proces uczenia modelu i jakość jego przewidywań.

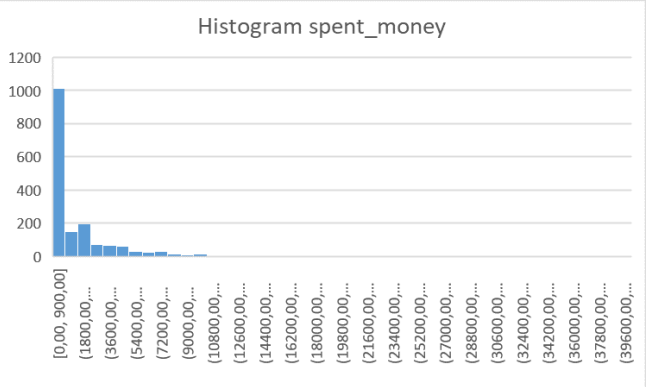


Przekształcenie danych

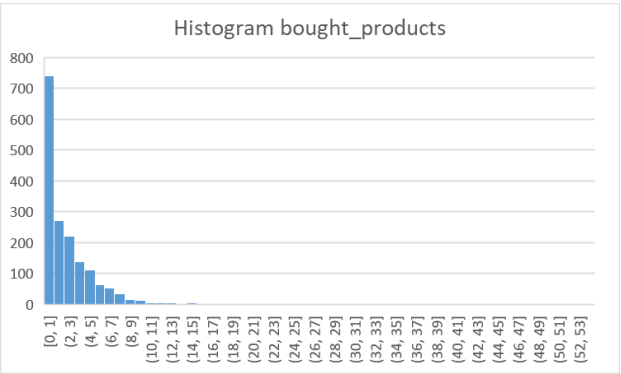
Na potrzeby dalszej analizy danych dokonaliśmy ich przekształcenia w sposób analogiczny do tego z iteracji pierwszej. Dla przypomnienia - tabela poddawana analizie posiada kolumny: **miasto**, **wydatek pieniędzy w danym miesiącu**, **rok**, **miesiąc**, **liczba**

zakupionych produktów, **liczba wyświetleń produktów**, **ilość wydanych pieniędzy** w sklepie przez klienta **w miesiącu następującym**.

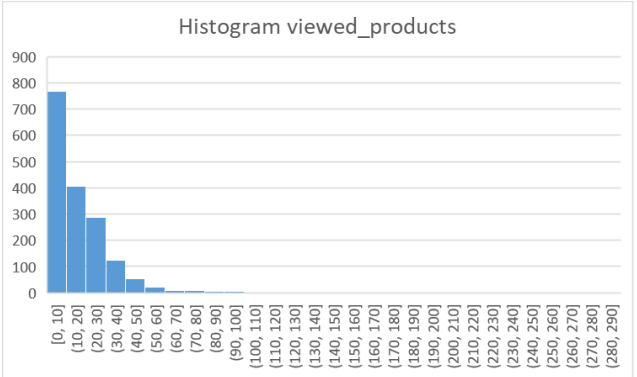
Analiza danych przekształconych



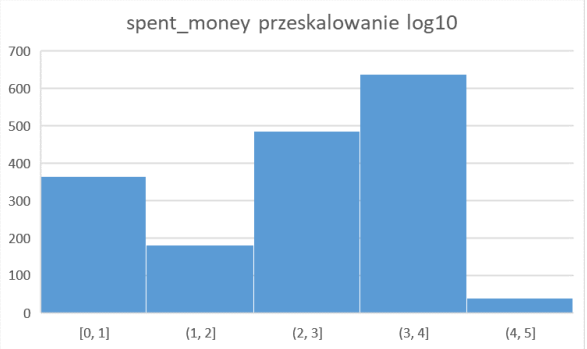
Powyższy histogram pokazuje, z jaką częstotliwością użytkownicy wydają ilości pieniędzy z danego przedziału w ciągu miesiąca. W tej iteracji można wyciągnąć wniosek, że zazwyczaj klienci nie wydają w sklepie więcej niż 900 zł miesięcznie.



Histogram *bought_products* przedstawia ile produktów najczęściej kupują klienci w ciągu miesiąca - jak często określona ilość produktów zostaje zakupiona przez klientów w ciągu miesiąca. Ponownie okazało się, że klienci sklepu najczęściej kupują nie więcej niż 5 produktów miesięcznie.



Z histogramu powyżej można odczytać jak często klienci wyświetlają strony produktów w ciągu miesiąca. Zazwyczaj klient nie wyświetla produktów częściej niż 40 razy w ciągu miesiąca.



W celu zbalansowania liczebności wartości kolumny *spent_money* wartości zostały przeskalowane funkcją logarytmiczną o podstawie 10.

Badanie współczynnika informacji wzajemnej pomiędzy zmiennymi

Poniżej zamieszczona jest tabela przedstawiająca wartości informacji wzajemnej pomiędzy poszczególnymi atrybutami, dla różnych zestawów danych. W celu możliwości porównania tych wartości dla “naszych” danych, wygenerowaliśmy zestawy danych wylosowanych poprzez wymieszanie kolejności wartości w kolumnach atrybutów. Na podstawie tabeli można stwierdzić, że wszystkie atrybuty wejściowe w większym lub mniejszym stopniu wnoszą informację o atrybucie wyjściowym *spent_money_next_month* - w danych nielosowych wartości informacji wzajemnej między atrybutem wyjściowym a poszczególnymi innymi atrybutami są większe od odpowiednich wartości informacji wzajemnej dla danych losowych.

Ze względu na to, że dane pochodzą z przestrzeni jednego roku i ze względu na to co możemy zobaczyć w tabeli można uznać, że atrybut *year* oraz *year*month* można pominąć i zostawić tylko atrybut *month*. Atrybut *year* mało wnosi do atrybutu wyjściowego; *year*month* z kolei wnosi tyle samo, co sam atrybut *month*.

Na podstawie tabeli wydawać się może, że znaczący wpływ na atrybut wyjściowy *spent_money_next_month* mają atrybuty *bought_products* i *viewed_products*.

Iw(x1, x2)							
x1	x2	Nasze dane	Losowe 1	Losowe 2	Losowe 3	Losowe 4	Losowe 5
city	spent_money	0,0784	0,0637	0,0639	0,0622	0,0590	0,0637
city	year	0,0073	0,0028	0,0013	0,0028	0,0032	0,0028
city	month	0,0065	0,0214	0,0223	0,0167	0,0211	0,0214
city	bought_products	0,0518	0,0269	0,0237	0,0226	0,0235	0,0269
city	viewed_products	0,1255	0,1064	0,1009	0,1202	0,1134	0,1064
city	spent_money_next_month	0,0777	0,0556	0,0679	0,0688	0,0698	0,0556
city	year*month	0,0116	0,0216	0,0259	0,0319	0,0247	0,0216
spent_money	year	0,0184	0,0077	0,0070	0,0111	0,0134	0,0077
spent_money	month	0,1568	0,1145	0,1276	0,1074	0,1020	0,1145
spent_money	bought_products	0,8984	0,1247	0,1077	0,1209	0,1194	0,1247
spent_money	viewed_products	0,7899	0,4689	0,4702	0,4584	0,4604	0,4689
spent_money	spent_money_next_month	0,3182	0,2835	0,2899	0,2870	0,2840	0,2835
spent_money	year*month	0,1899	0,1376	0,1293	0,1270	0,1244	0,1376
year	month	0,1075	0,0027	0,0032	0,0022	0,0053	0,0027
year	bought_products	0,0099	0,0024	0,0049	0,0068	0,0056	0,0024
year	viewed_products	0,0215	0,0160	0,0143	0,0189	0,0191	0,0160
year	spent_money_next_month	0,0161	0,0096	0,0121	0,0083	0,0097	0,0096
year	year*month	0,1963	0,0027	0,0038	0,0033	0,0020	0,0027
month	bought_products	0,1019	0,0434	0,0534	0,0368	0,0480	0,0434
month	viewed_products	0,2251	0,1953	0,2063	0,2075	0,2106	0,1953
month	spent_money_next_month	0,1747	0,1055	0,1070	0,1075	0,1197	0,1055
month	year*month	2,3830	0,0415	0,0442	0,0440	0,0403	0,0415
bought_products	viewed_products	0,7335	0,2047	0,2069	0,2025	0,1805	0,2047
bought_products	spent_money_next_month	0,1792	0,1157	0,1097	0,1232	0,1288	0,1157
bought_products	year*month	0,1203	0,0477	0,0508	0,0510	0,0679	0,0477
viewed_products	spent_money_next_month	0,5481	0,4769	0,4683	0,4522	0,4787	0,4769
viewed_products	year*month	0,2619	0,2326	0,2301	0,2240	0,2151	0,2326
spent_money_next_month	year*month	0,2020	0,1339	0,1255	0,1289	0,1388	0,1339

Podsumowanie drugiej iteracji wstępnej analizy danych

W stosunku do poprzedniej iteracji, widoczna jest zdecydowana poprawa jakości w analizowanych danych. Dane nie zawierają już błędów, jest też mniej brakujących wartości. Mamy też teraz więcej informacji z miesięcy przed szczytem popularności sklepu. Wydaje się, że korzystających z tych danych można już próbować uczyć model.