Rozwiązanie zadania rekrutacyjnego

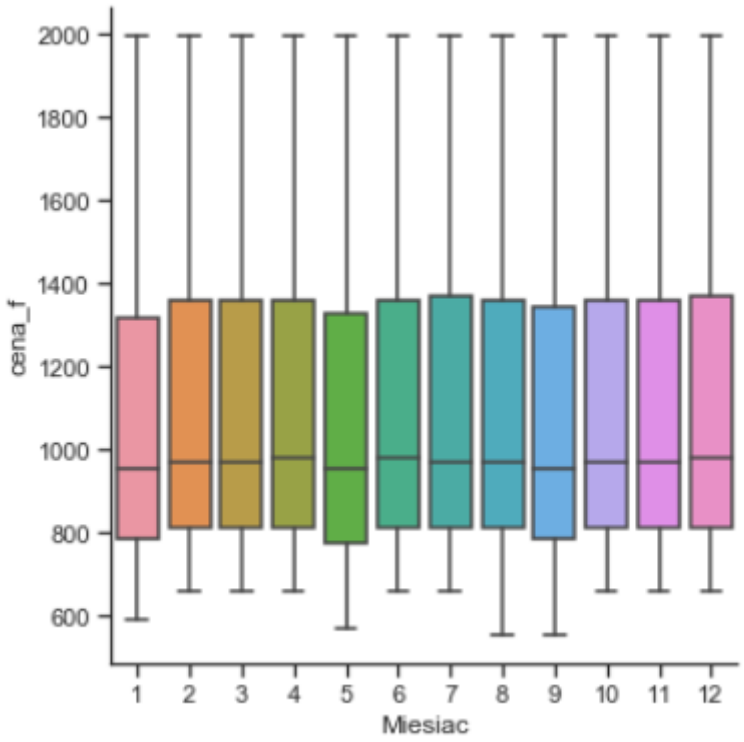
Jakub Majcher

1. **Opis problemu.**
   Celem zadania było przeprowadzenie analizy wartości sprzedaży sprzętu RTV/AGD. Dane do przetworzenia zostały zebrane w pliku csv. Plik zawierał 475966 rekordów danych i 28 kolumn. Kolumny zawierały informacje o dacie i miejscu zakupu, informacji o produkcie (ID produktu, producent, grupa produktowa), kosztach i marży, oraz oczywiście o cenie sprzedaży. Na podstawie tych danych należało sporządzić tabelę podsumowań oraz stworzyć model regresji liniowej.

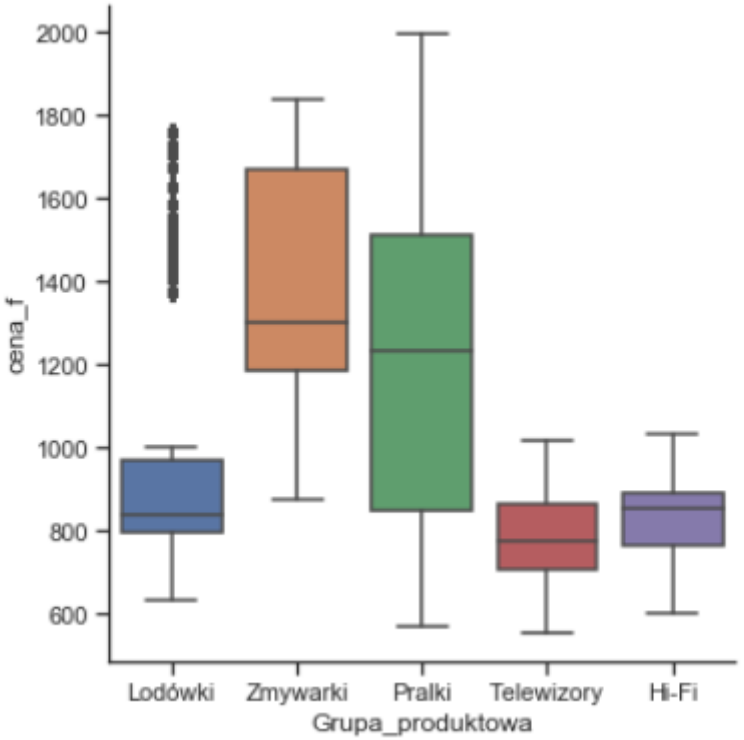
2. **Rozwiązanie problemu.**
   Pierwszym etapem było zaznajomienie się ze strukturą danych i z typami zmiennych. Następnie należało sprawdzić czy występują brakujące wartości. W celu prawidłowego rozwiązania problemu brakujących wartości przystąpiłem do analizy poszczególnych kolumn potrzebnych do stworzenia tabeli podsumowań. Z analizy zmiennej 'Platnosc' dowiedziałem się, że rekordom z brakującymi wartościami tej zmiennej brakuje również wartości w innych polach. Wywnioskowałem z tego, iż takie rekordy oznaczają niesprzedane produkty. Postanowiłem usunąć je ze zbioru danych. Okazało się, że po usunięciu tych rekordów brakujące dane nie występują w zbiorze danych.
   Następnym krokiem było wydobycie informacji liczbowej o cenie sprzedaży. W tym celu usunąłem z kolumny 'cena_sprzedazy' wszystkie odstępy i skrót 'zł', po czym przekonwertowałem kolumnę na liczbę zmiennoprzecinkową. Po otrzymaniu numerycznych wartości ceny przystąpiłem do tworzenia wyjściowego zbioru danych. Po ustaleniu nazw kolumn przepuściłem dane przez pętlę tworzącą tabelę podsumowań.
   Pętla składała się z pięciu małych pętli tworząc każdą możliwą i istniejącą kombinację wartości grup produktowych, rodzajów płatności, województw, miesięcy i lat. Dla każdej kombinacji obliczone zostały:
   a. średnie wartości,
   b. mediany,
   c. minima,
   d. maksima,
   e. odchylenie standardowe.

   Kombinacje te wraz z wyliczonymi miarami zostały zapisane w tabeli podsumowań.

3. **Stworzony model.**
   Tworzenie modelu rozpocząłem od graficznej analizy zmiennych. Do przedstawienia zależności zmiennej wyjściowej od zmiennych kategorycznych użyłem wykresów pudełkowych. Największy wpływ na cenę wykazywały zmienne 'Grupa_produktowa' i 'Producent'. Mniejszy wpływ wykazywała zmienna 'Miesiac'. Postanowiłem użyć ich do tworzenia modelu mając na uwadze, iż te zmienne 'Grupa_produktowa' i 'Producent' są najpewniej skorelowane (producenci specjalizujący się w produkowaniu jednej grupy). Następnie postanowiłem zbadać zależność kosztów stałych i zmiennych od ceny sprzedaży. Zgodnie z moimi przypuszczeniami zmienne te miały istotny wpływ na cenę sprzedaży. W celu uniknięcia problemów z korelacją miedzy tymi dwoma zmiennymi postanowiłem je zamienić na jedną zmienną 'koszty' będącą ich sumą.



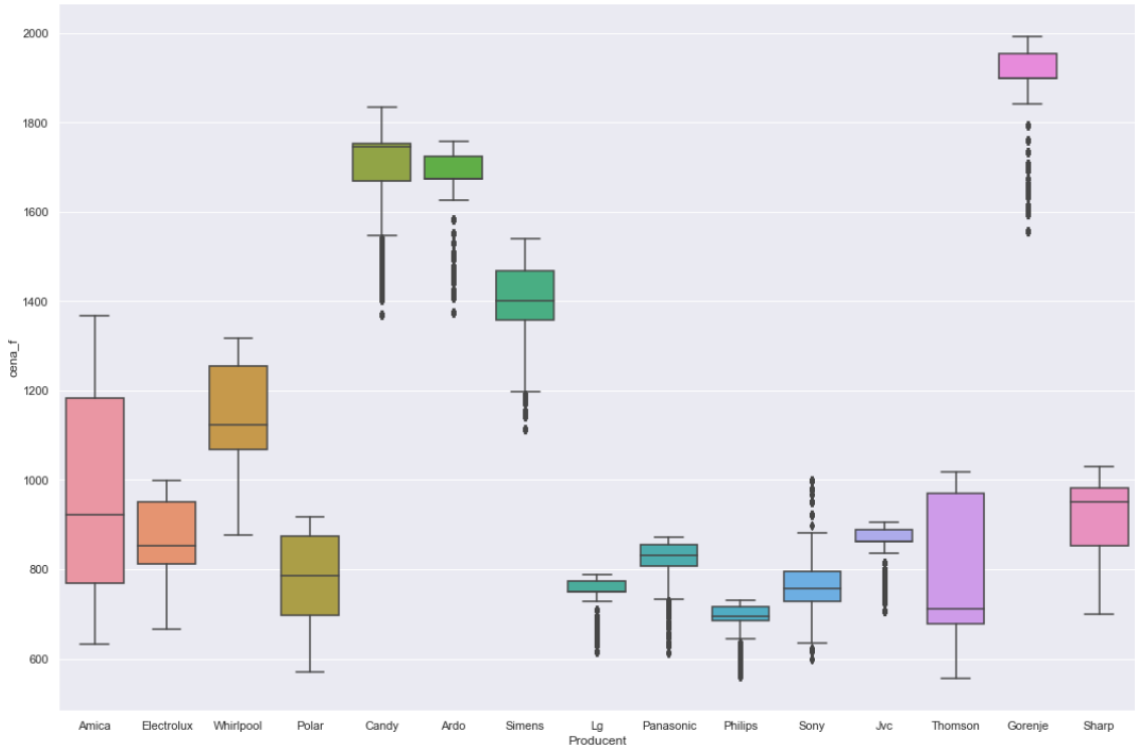
Rys. 1: Wykres pudełkowy dla ceny sprzedaży w zależności od miesiąca.



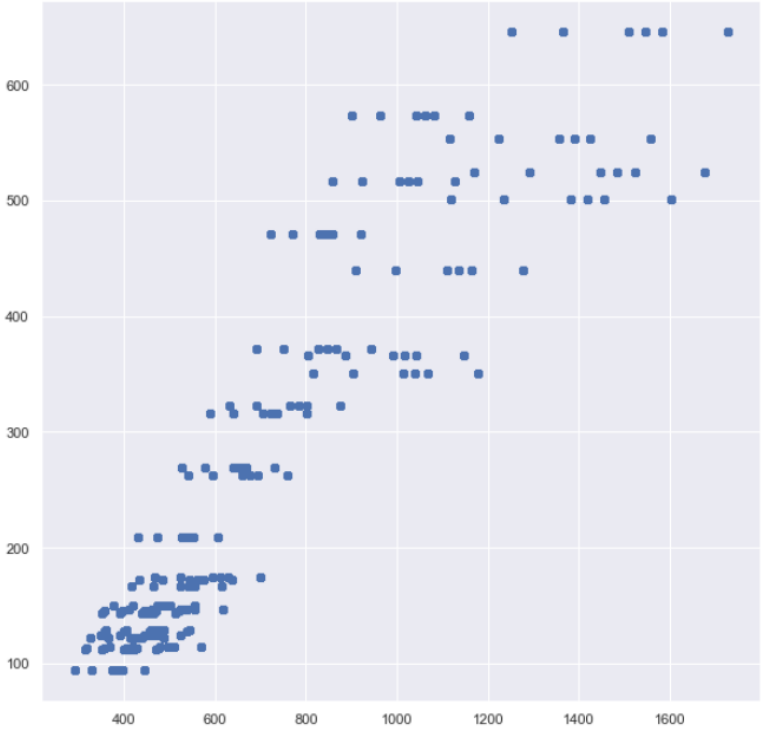
Rys. 2: Wykres pudełkowy dla ceny sprzedaży w zależności od grupy produktowej.

Do użycia danych kategorycznych w regresji liniowej potrzebne jest stworzenie tzw. ‘dummy variables’: zmiennych oznaczającej pojedyncze kategorie, przyjmujących wartość 0 lub 1. Po stworzeniu tych zmiennych sprawdziłem współzależność między nimi poprzez obliczenie współczynnika VIF. Wartości współczynnika VIF potwierdziły moje przypuszczenia co do

współzależności zmiennych ‘Grupa_produktowa’ i ‘Producent’. Postanowiłem wyeliminować jedną z nich. Do podjęcia decyzji wykorzystałem modele regresyjne stworzone z biblioteką statsmodels. Model używający zmiennej ‘Producent’ wykazał większą dokładność. Zmiennymi wykorzystanymi przy tworzeniu modelu były koszty, marka produktu i miesiąc zakupu.



Rys. 3: Wykres pudełkowy dla ceny sprzedaży w zależności od producenta.

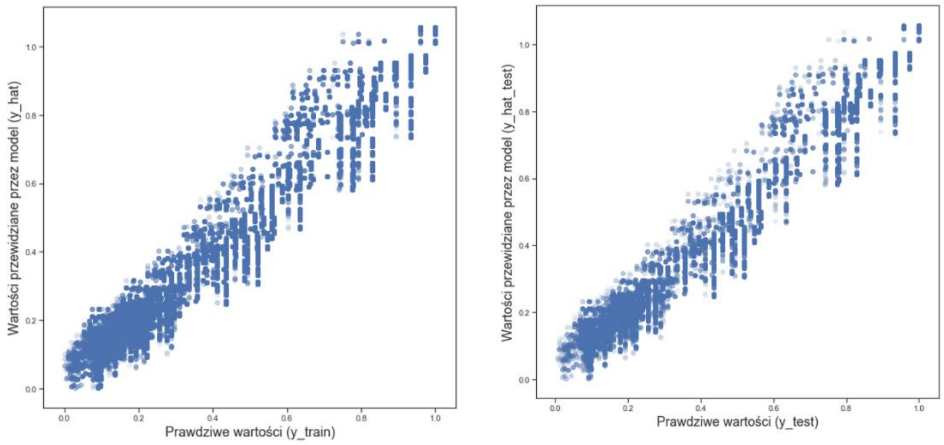


Rys. 4: Wykres zależności ceny sprzedaży od kosztów.

Analizując wartości p zmiennych użytych do tworzenia modelu mogę stwierdzić, iż wszystkie zmienne są istotne. Patrząc na wartość F-statictic mogę stwierdzić, że model również jest istotny. Do testowania modelu użyłem biblioteki sklearn. Po podzieleniu danych na wejściowe i wyjściowe oraz na dane trenujące i testowe przystąpiłem do uczenia i testowania modelu.

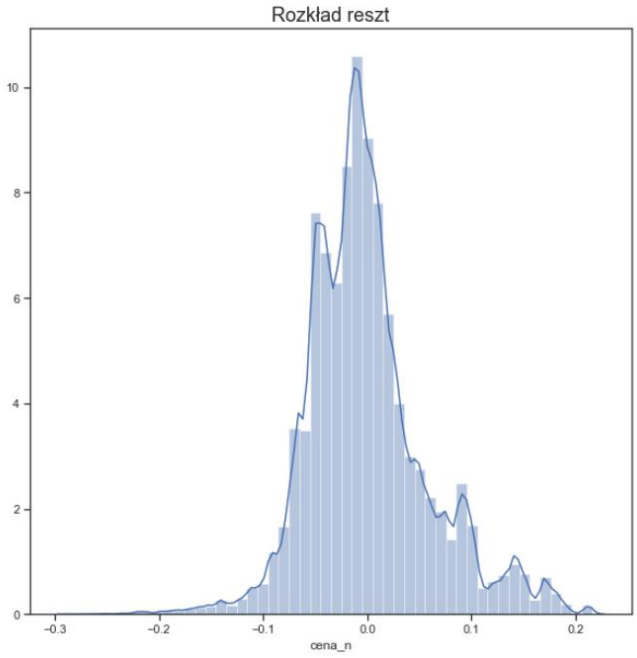
4. Analiza wyników.

Obserwując wykresy zależności wartości prognozowanych przez model od wartości rzeczywistych dla danych trenujących i testowych można stwierdzić, że model trafnie przewiduje wartości ceny produktu.



Rys. 5: Zależność wartości prognozowanych przez model od wartości rzeczywistych dla danych trenujących i testowych.

Dokładność prognozy jest większa dla mniejszych cen, jednakże nie na tyle, żeby się martwić o niespełnienie założenia o niezależności błędu od zmiennej.



Rys: 6: Rozkład błędu dla danych trenujących.

W celu dalszego zbadania postanowiłem stworzyć wykres rozkładu reszt. Wykres ten przypomina rozkład normalny ze średnią w okolicach zera. Założenia regresyjne nie zostały więc naruszone.
Dokładność modelu dla zmiennych testowych wynosi 94,86% co jest wynikiem akceptowalnym. Po analizie współczynników zmiennych można dojść do wniosku, że największy wpływ na cenę sprzedaży mają koszty produktu. W przypadku zmiennych określających producenta odniesieniem była Amica: ujemne wartości oznaczają mniejszą cenę dla danego producenta przy porównaniu z Amicą, a dodatnie- wyższe ceny.
Odniesieniem miesięcy był styczeń, który (jak można zauważyć na wykresie pudełkowym dla miesięcy) odnotowywał najmniejsze sprzedaże. Błąd modelu wahał się od 0,002% do 34,162%.