PREDICT FUTURE SALES

Kaggle Competition

Hubert Jóskowiak

Poznań, 2018

Spis treści

1. Opis projektu

- Dane wejściowe: informacje dotyczące sprzedaży produktów w danym sklepie między styczniem 2013 a październikiem 2015.
- Cel: prognoza sprzedaży produktów znajdujących się w zbiorze testowym w poszczególnych sklepach, również wyróżnionych w zbiorze testowym, w listopadzie 2015.
- Dostępne pliki:
  - *sales_train.csv* - dane sprzedaży między styczniem 2013 a październikiem 2015 (zbiór treningowy),
  - *test.csv* - wyróżnione sklepy i produkty, dla których należy przeprowadzić prognozę sprzedaży (zbiór testowy),
  - *items.csv* - informacje o poszczególnych produktach,
  - *item_categories.csv* - informacje o przynależności danego produktu do poszczególnych kategorii,
  - *shops.csv* - dodatkowe informacje o sklepach.

2. Zaproponowane rozwiązanie

- Wykorzystanie danych dotyczących jedynie sklepów znajdujących się w zbiorze testowym:
  - Cechy: informacje o produkcie, sklepie, kategorii oraz sprzedaży w poprzednich miesiącach
  - Etykieta: informacja o liczbie sprzedanych produktów
- Wykorzystanie algorytmów uczenia maszynowego:
  - XGBoost
  - Wielowarstwowych sieci neuronowych

- Jako punkt odniesienia (baseline) wykorzystano dane sprzedaży z ostatniego dostępnego miesiąca (października 2015). Dane zostały przycięte, tak by zawierały się w przedziale [0; 20]. Wartość RMSE na zbiorze testowym dla wygenerowanych w ten sposób danych wynosi **1,16811**.
- Cel projektu: Osiągnąć wartość RMSE na poziomie 1,00

2.1. Wykorzystane projekty

- Przygotowanie danych (ekstrakcja cech), baseline
  www.kaggle.com/szhou42/predict-future-sales-top-11-solution
- Poprawa błędnych danych
  www.kaggle.com/dlarionov/feature-engineering-xgboost
- Model XGBoost
  www.kaggle.com/jamesguo89/sklearn

2.2. Wykorzystane biblioteki

- numpy
- pandas
- matplotlib.pyplot
- sklearn
- itertools
- tensorflow
- xgboost

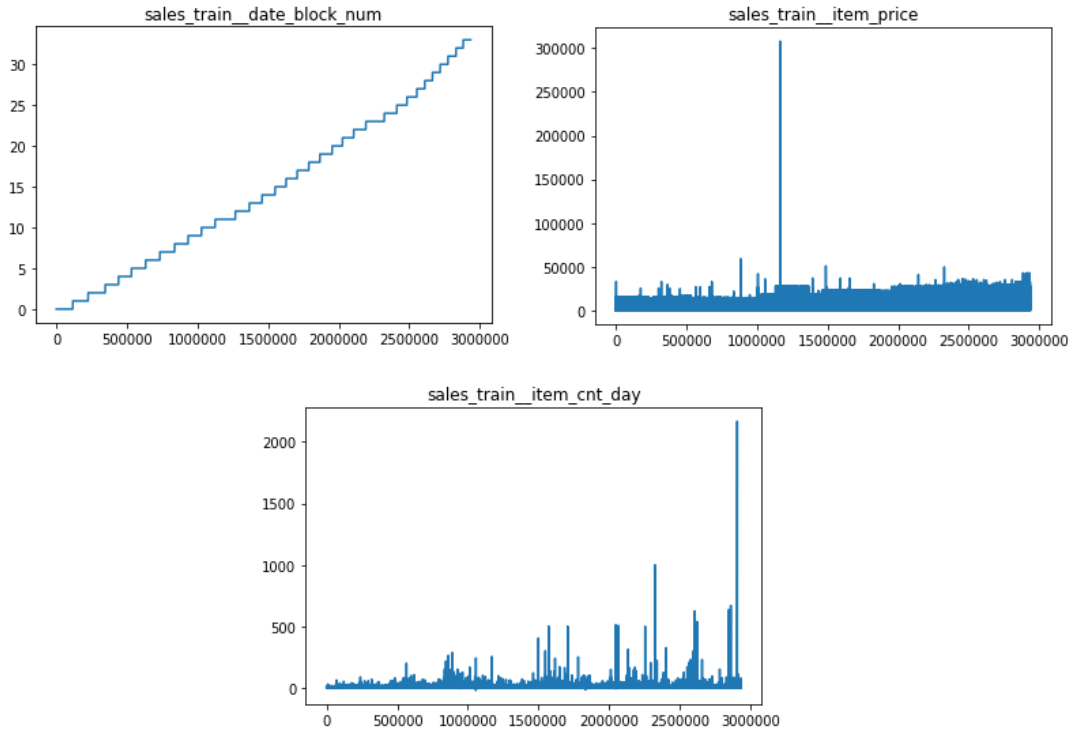
3. Przygotowanie danych

3.1. Wstępne przetwarzanie danych

1. Pobranie danych i przygotowanie tabeli z wszystkimi sklepami/produktami znajdującymi się w zbiorze testowym dla każdego miesiąca.
2. Usunięcie duplikatów
3. Wykrycie błędów w danych (rys 3.1)

4. Poprawa błędów:

   - cena produktu równa 0 - wykorzystanie mediany cen tego samego produktu (*item_id* = 2973) dostępnego w danym sklepie (*shop_id* = 32) w tym samym miesiącu (*date_block_num* = 4)
   - przycięcie cen tak by zawierały się w zbiorze [0; 300000] - likwidacja cen ujemnych
   - przycięcie liczby sprzedanych produktów, tak by zawierały się w zbiorze [0; 1000] - likwidacja wartości ujemnych oraz wartości znacznie odbiegających od średniej.
   - zamiana nazw i indeksów sklepów, w nazwach których pojawiały się literówki - zamiana wartości ID sklepów o ID równym 11 (na 10), 57 (na 0) oraz 58 (na 1) i ponowne numeracja sklepów.



Rys. 3.1: Wykrycie błędów

3.2. Przygotowanie danych do uczenia maszynowego

1. Dodanie informacji o mieście (*city_id*) oraz przegrupowanie kategorii (18 głównych zamiast 84 często zbliżonych do siebie kategorii)

2. Zsumowanie miesięcznej sprzedaży w każdym sklepie i pobranie informacji tylko o sklepach znajdujących się w z zbiorze testowym. Zbudowany zbiór danych ma zatem 214200 danych ze zbioru testowego * 35 miesięcy = **7497000** pozycji.

3. Dodanie dodatkowych wektorów cech:

    - sprzedaż produktu w danym sklepie w poprzednich miesiącach
    - sprzedaż produktu w danym mieście w poprzednich miesiącach
    - sprzedaż produktu we wszystkich sklepach w poprzednich miesiącach
    - sprzedaż produktów przynależących do danej kategorii w danym sklepie w poprzednich miesiącach
    - sprzedaż produktów przynależących do danej kategorii w danym mieście w poprzednich miesiącach
    - sprzedaż produktów przynależących do danej kategorii we wszystkich sklepach w poprzednich miesiącach

4. Podział zbioru danych na zbiór treningowy, walidacyjny oraz testowy:

    - **zbiór treningowy**: dane zebrane między 12 a 32 miesiącem (dane z miesięcy 0-11 nie są brane pod uwagę, ponieważ cechy opisane w poprzednim punkcie są generowane dla maksymalnie 12 miesięcy wstecz - dla miesiąca 11 nie istnieją informacje o sprzedaży 12 miesięcy wstecz co może zakłamywać dane w procesie uczenia) - 7068600 danych
    - **zbiór walidacyjny**: dane pochodzące z 33 miesiąca (październik 2015); zbiór walidacyjny został wykorzystany w celu umożliwienia szybszego zakończenia treningu (jeśli wartość RMSE rośnie przez 10 kolejnych iteracji/epok trening zostaje zakończony, a na wyjściu programu znajduje się model, który uzyskał najniższą wartość błędu) - 214200 danych.
    - **zbiór testowy**: dane pochodzące z 34 miesiąca (listopad 2015) - 214200 danych

4. Wykorzystane algorytmy uczenia maszynowego

4.1. Model XGBoost

Tab. 4.1: Parametry modelu XGBoost

Parametr	Wartość
Maksymalna wysokość drzewa	10
Liczba drzew	1000
Maksymalna liczba iteracji	1000
Minimalna suma wag krawędzi	300
Współczynnik podpróbkowania zbioru treningowego	0.8
Współczynnik podpróbkowania kolumn przy tworzeniu każdego drzewa	0.8
Współczynnik uczenia	0.3

4.2. Sieć neuronowa

Tab. 4.2: Architektura sieci neuronowej

Typ warstwy	Liczba neuronów	Funkcja aktywacji
w pełni połączona	256	ReLU
w pełni połączona	128	ReLU
w pełni połączona	64	ReLU
w pełni połączona	32	ReLU
w pełni połączona	1	-

Tab. 4.3: Parametry sieci neuronowej

Parametr	Wartość
Rozmiar pojedynczego zbioru uczącego (batch)	512
Maksymalna liczba epok	25
Współczynnik uczenia	0.001
Funkcja kosztu	RMSE
Algorytm optymalizacji	Adam
Dropout	1.0

5. Przedstawienie wyników

5.1. Wykorzystane zbiory danych

- Dataset_1 - 10 cech : miesiąc, ID sklepu, ID miasta, ID produktu, ID kategorii, sprzedaż produktu w danym sklepie w poprzednich miesiącach (5 cech).
- Dataset_2 - 20 cech : miesiąc, ID sklepu, ID miasta, ID produktu, ID kategorii, sprzedaż produktu w danym sklepie w poprzednich miesiącach (5 cech), średnia sprzedaż produktu w danym mieście w poprzednich miesiącach (5 cech), średnia sprzedaż produktu w poprzednich miesiącach (5 cech)
- Dataset_3 - 25 cech : miesiąc, ID sklepu, ID miasta, ID produktu, ID kategorii, sprzedaż produktu w danym sklepie w poprzednich miesiącach (5 cech), średnia sprzedaż produktu w danym mieście w poprzednich miesiącach (5 cech), średnia sprzedaż produktu w poprzednich miesiącach (5 cech), średnia sprzedaż produktów należących do danej kategorii w danym sklepie w poprzednich miesiącach (5 cech)

Dane sprzedaży dla danego miesiąca uwzględniają sprzedaż produktów sprzed 1,2,3,6 i 12 miesięcy w stosunku do tego miesiąca.

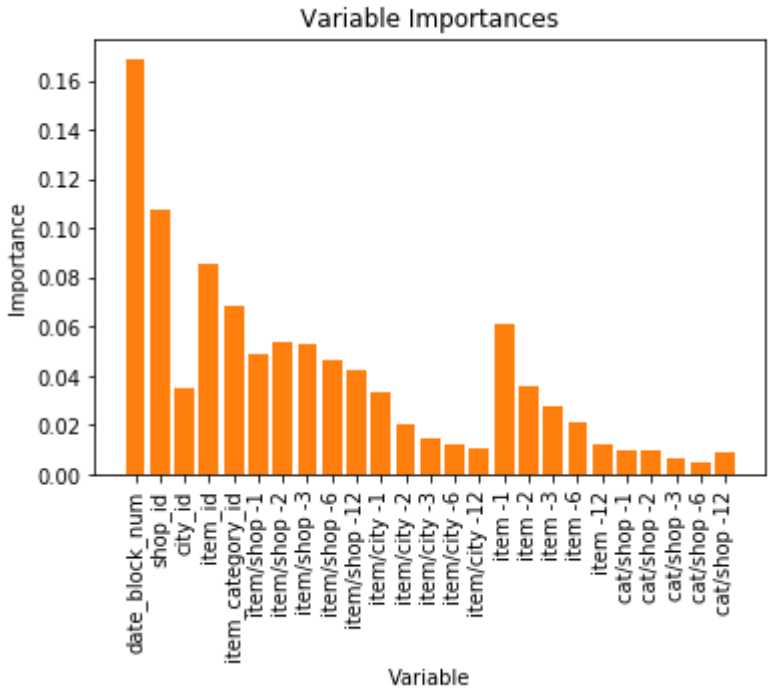
5.2. Uzyskane wartości RMSE

Tab. 5.1: Wyniki przeprowadzonych testów

Zbiór danych	Algorytm	RMSE zb. treningowego	RMSE zb. walidacyjnego	RMSE zb. testowego
Baseline				1,168
Dataset_1	XGBoost	0.785	0.882	1.014
	Sieć neuronowa	0.790	0.892	1.026
Dataset_2	XGBoost	0.776	0.879	1.008
	Sieć neuronowa	0.781	0.888	1.024
Dataset_3	XGBoost	0.783	0.878	**1.006**
	Sieć neuronowa	0.783	0.888	1.022

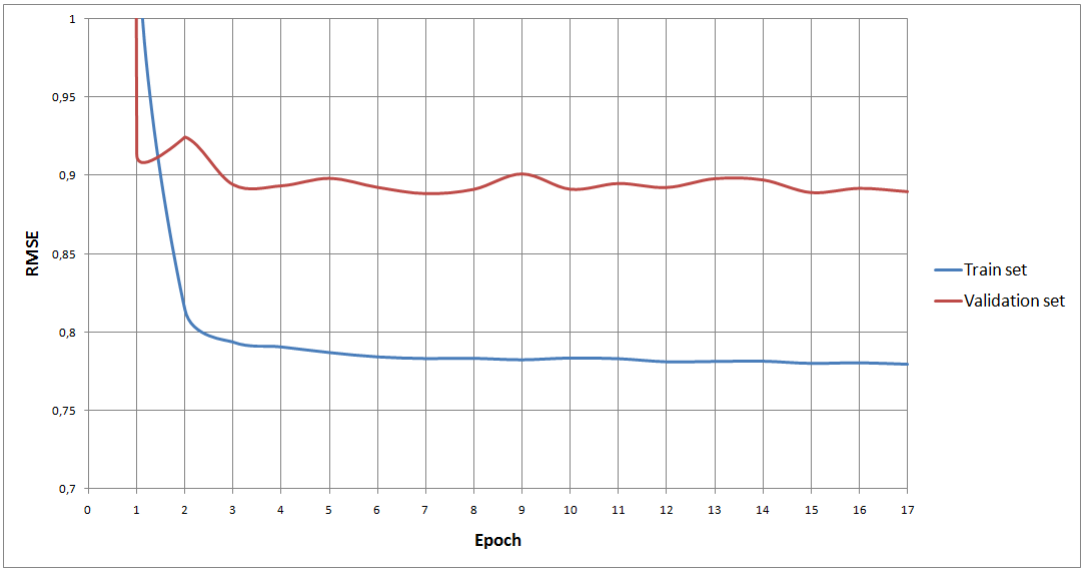
Zbiór testowy stanowi 35% pełnego zbioru testowego (214200 wektorów).

5.3. Wpływ cech na proces uczenia



Rys. 5.1: Znaczenie cech podczas uczenia modelu XGBoost.

5.4. Trening sieci neuronowej



Rys. 5.2: Wartości RMSE dla zbioru treningowego i walidacyjnego w kolejnych epokach

6. Wnioski

- Osiągnięto wartość RMSE na poziomie 1,006, co oznacza poprawę w stosunku do baseline'u o 0,162.
- Uzyskane wyniki pozwoliły na zajęcie 775 miejsca na 2055 projektów biorących udziału w *Kaggle Competition* [stan na 06.01.2019].
- Przetestowano działanie wielowarstwowych sieci neuronowych - osiągnięte wyniki są zbliżone do tych osiągniętych z wykorzystaniem modelu XGBoost (wartość RMSE na poziomie 1,022 pozwoliłaby na zajęcie 842 miejsca).
- W projekcie ograniczono liczbę danych wejściowych do informacji dotyczących par sklep/produkt znajdujących się w zbiorze testowym - pozwoliło to na zwiększenie szybkości treningu i zmniejszenie złożoności obliczeniowej.
- Wykorzystano informację o sprzedaży danego produktu w mieście, w którym znajduje się sklep.

6.1. Możliwe sposoby rozwinięcia projektu

- Zwiększenie liczby danych uwzględnianych w analizie. Możliwe jest uwzględnianie kombinacji wszystkich sklepów i wszystkich produktów występujących w danym miesiącu w zbiorze treningowym (10913850 wektorów danych), lub kombinacji wszystkich sklepów i wszystkich produktów w każdym z 35 miesięcy (22170 produktów * 57 sklepów * 35 miesięcy = 44229150 wektorów danych).
- Dodanie większej ilości dodatkowych cech: informacji dotyczących sprzedaży produktów z danej kategorii, informacji dotyczących cen sprzedawanych produktów, informacji o miesiącu, w którym po raz ostatni zakupiono dany produkt w danym sklepie itp.
- Niestety z uwagi na ograniczone możliwości obliczeniowo niemożliwe było przeprowadzenie dodatkowych testów z uwzględnieniem większej ilości danych.