Predict Future Sales

Reemet Ammer, Indrek Pertman, Raido Everest, Karl Kuusik

Dataset

Overview

Description

Evaluation

This challenge serves as final project for the "How to win a data science competition" Coursera course.

In this competition you will work with a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company.

We are asking you to predict total sales for every product and store in the next month. By solving this competition you will be able to apply and enhance your data science skills.

Dataset

Shops

shop_name	shop_id
Москва ТЦ "Ареал" (Беляево)	26
Химки ТЦ "Мега"	54
Коломна ТЦ "Рио"	16
Якутск ТЦ "Центральный"	58
Омск ТЦ "Мега"	38
Новосибирск ТЦ "Мега"	37
Жуковский ул. Чкалова 39м²	11

Item Categories

item_name	item_id	item_category_id
Фигурка Rocky Series 2 Drago Yellow Trunks 7"	20719	72
РЫЖАЯ СОНЯ (BD)	17997	37
ВАС ОЖИДАЕТ ГРАЖДАНКА НИКАНОРОВА (rem)	9583	40
Castlevania: Lords of Shadow 2 [Xbox 360, русс...	2356	23
Футболка Watch Dogs Skull XL	21283	61
КОМПАС-3D V14 Home [PC, Цифровая версия]	12729	76
Krater Character DLC Mayhem MK13 [PC, Цифровая...	4275	31

Items

item_category_name	item_category_id
Кино - Blu-Ray	37
Подарки - Настольные игры	64
Книги - Бизнес литература	46
Игры PC - Стандартные издания	30
Игры PC - Цифра	31
Программы - Для дома и офиса (Цифра)	76
Игры PC - Дополнительные издания	28

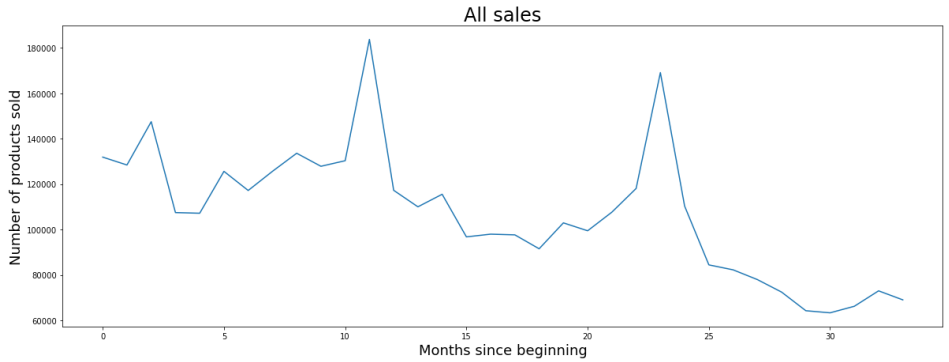
Sales – Training Data

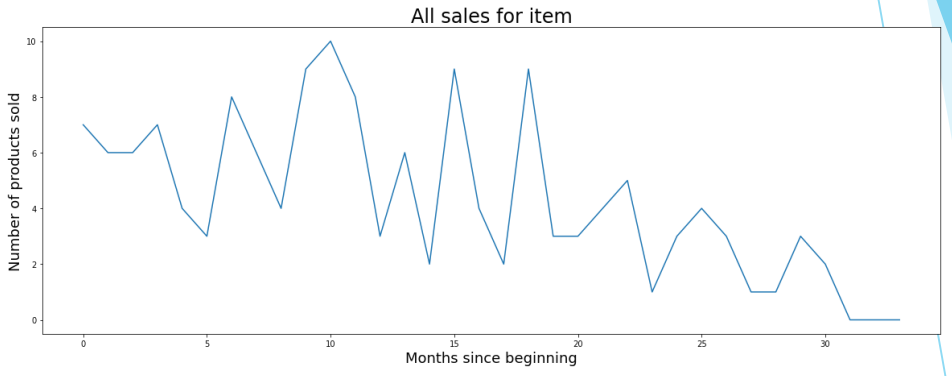
date	date_block_num	shop_id	item_id	item_price	item_cnt_day
09.01.2014	12	22	5380	2491.0	1.0
21.02.2015	25	48	77	149.0	1.0
24.11.2013	10	58	14723	299.0	2.0
23.06.2015	29	45	4901	2399.0	1.0
25.04.2014	15	54	8415	349.0	1.0
14.05.2014	16	18	21385	1199.0	1.0
04.01.2013	0	0	2848	73.0	2.0

Data Cleaning

- 1. Remove rows with negative item_price and item_cnt_day
- 2. Remove duplicate shops, items and categories
- 3. Remove rows with extreme values of item_price and item_cnt_day
- 4. Replace missing values with 0 or mean

Time series


All sales
Number of products sold
Months since beginning


All sales for item
Number of products sold
Months since beginning


Sales for a shop
Number of products sold
Months since beginning


All sales shop and item
Number of products sold
Months since beginning

First Approach
Simple

- Predict the number of sold items directly the same as for previous month.

- RMSE: 1.19724
- Placement: 6470/9800 – among 66%

Second Approach

Pivot Table

	shop_id	item_id	m0	m1	m2	m3	m4	m5	m6	m7	...	m24	m25	m26	m27	m28	m29	m30	m31	m32	m33
0	0	30	0	31	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	31	0	11	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	32	6	10	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	33	3	3	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	35	1	14	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
424092	59	22154	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
424093	59	22155	0	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
424094	59	22162	0	0	0	0	0	0	0	0	...	0	9	4	1	1	0	0	1	0	0
424095	59	22164	0	0	0	0	0	0	0	0	...	0	2	1	2	0	0	1	0	0	0
424096	59	22167	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

LinearRegression, DecisionTreeRegressor, Neural network

RMSE for Neural Network: 1.018, Placement: 4527/9800

Third Approach
Feature Engineering

Lag features:

Lag features are the classical way that time series forecasting problems are transformed into supervised learning problems.

The simplest approach is to predict the value at the next time (t+1) given the value at the previous time (t-1)

New features:

Sold items for 3 previous months (+ average)

Revenue (count*price) for last the months (+ average)

Price change for an item in the last 3 months

Derive month_of_year from the full date

What we learned

The difficult part is not training and figuring out the models, but to figure out the relevant features and structure of the data to be given to the model.

This includes:

Understanding and getting insight of the underlying process generating the data

Derive new features from the given data, which could reveal important structure or relations underlying the dataset

Choose or develop the correct statistical tools for a given problem domain