January 3, 2019
Krzysztof Joachimiak
joachimiak.krzysztof@gmail.com
github.com/krzjoa/kaggle-sales

Predict Future Sales - Kaggle competition

Recruitment task for Research Engineer position

Contents

1 Task

The goal of this task is to predict future sales value. This task is a Kaggle competition.

2 Data Analysis

2.1 General information

There are **22170** divided into **84 categories**. In the dataset, there occur **60 shops**. We can find **2'935'849 records** in the training dataset, and **214'200** in the testing one.

Insigths:

- There occur **negative coun values**. As many guys in the competition-related discussion say, it probably expresses the number of returned and refunded items

2.2 Trends in Time Series

At the very beginning, let's check, how many recordings per each month in the measured period we have. As we can see in the figure 2.1, the number of sale records depends on time and we are not sure if it's just a **lack of data** or it really shows us some **meaningful temporal relation**.

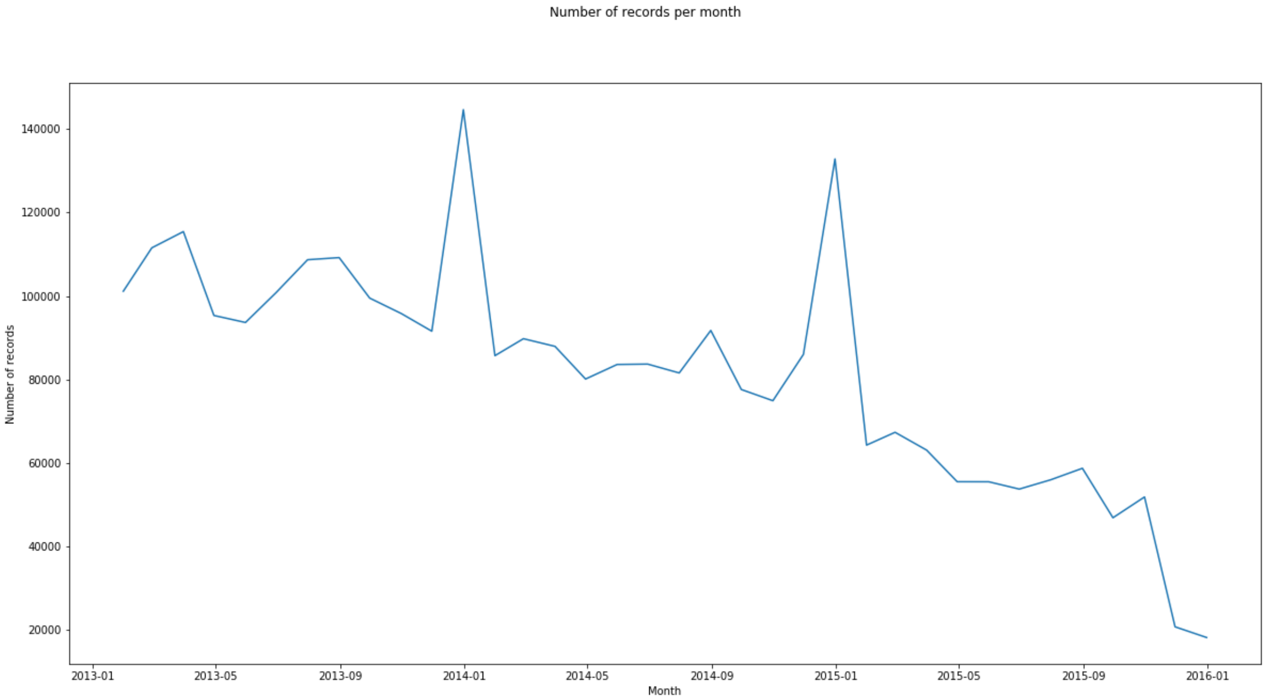

Figure 2.1: Number of records per month in the training dataset

If we sum values of sales in the each month, we can get following plot (figure 2.2).

Seeing only this picture, we cannot state if there exists a clear link between time and summed sales values in each month.

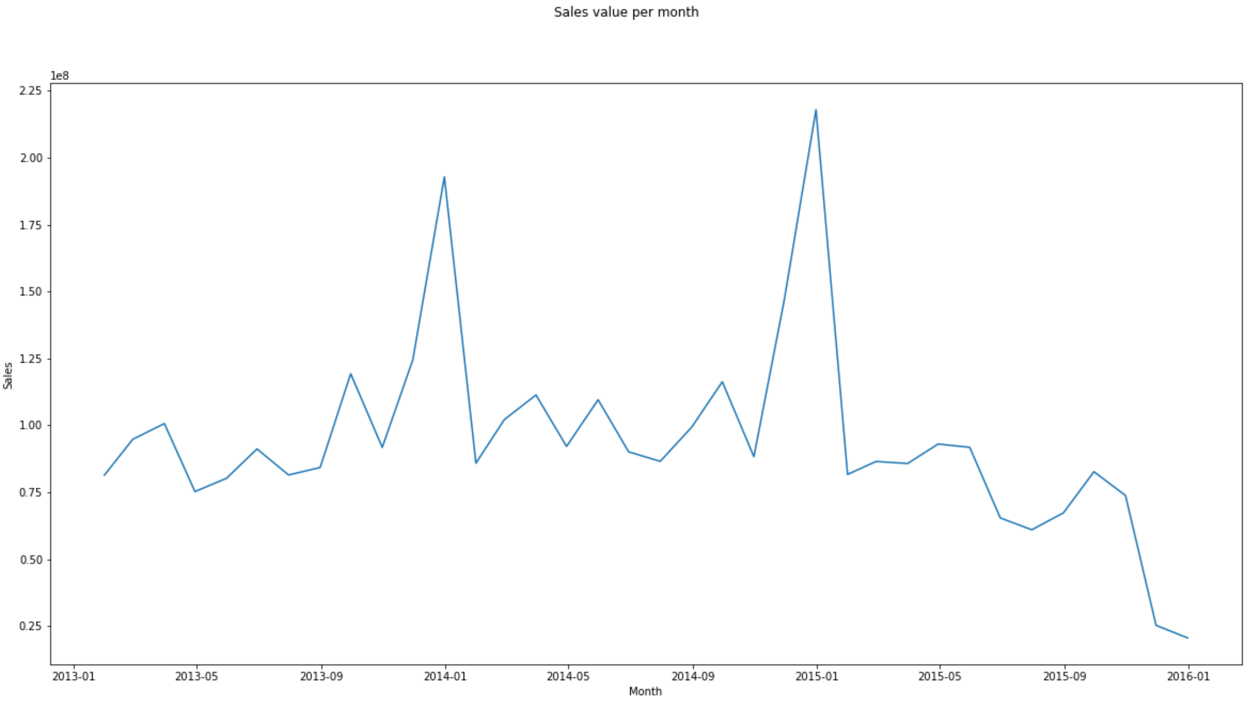

Figure 2.2: Sales values per month in the training dataset