Future Sales Prediction

ATAKAN FILGOZ ENES KARANFIL BURAK BARAN OMER SEKER

Abstract

Nowadays, one of the largest goals of craftsmen is to predict the sales for the future that they get from the products they sell. There is a considerable number of factors which affect the revenue. These factors can be guessed by reviewing the characteristics of daily historical data. Prediction of sales of the craftsmen can be seen as a machine learning problem. Therefore, in this study we try to predict the future sales of products based on their properties.

1. Introduction

Owing to the advancing technology, methods such as machine learning and deep learning are used in solving big problems in our present time. The most important criteria in efficiency of these methods and obtaining results is having enough data on the problem. Recently, data collection and processing the collected data in all fields used for large problems. One of these fields is the commerce industry. This article consists of the problem in commerce data. In order for a shop to generate high revenue, a shop has to meet a number of criterion. The data on the subject and grip of a shops content, the current price of items,the number of products sold in the past are some of the criterion determinative of future sales. The data provided in this project will help to estimate the future sales of the shops. Carrying out works and studies on commerce data and carrying out analysis to estimate future sales and other content are particularly important information for craftsmen. The most important reasoning is that such analysis sets forth financial risks in advance and provides the opportunity to operate dependently. With a credible estimation, necessary measures may be taken by the craftsmen in the early stages of receiving items and financial risks may be reduced with improvements. Especially, it is important for craftsmen to solve this problem to make provision against fluctuating incomes. Thus, tradesmen can determine their living standards by having this knowledge.

Studies on this subject using different data sets are carried out. CNN algorithms are used in future sales predictions.

2. Methods

2.1. Dataset

The data set used in the project, is taken from a competition from Kaggle. In this section, details about the data set will be shared. The training data set consists of 2.935.849 samples and 10 columns which consist of data item id which represents unique identifier of a product, shop id for unique identifier of a shop, ID that related with shop-item tuple within the test set, item category id,item count day is number of product thay they sold, item price current price of an item, date, date block num is used for a consecutive month number that used for convenience, item name, shop name, item category name. The test data has 214.201 samples. After examining the data set, we have noticed that data set consists of time series. Sales over time of each store-item is a time-series itself. Thus, it is foreseen that it may cause difficulties in the future.

2.2. Data Preprocessing

We need to preprocess the data before we start developing models to predict future sales. At first, observed the number of missing values in the data. And there was not any missing values in the data. All of our data held in a single CSV. Data has attributes which can be seen in Figure-1 and given without any order.

Data fields

- **ID** - an Id that represents a (Shop, Item) tuple within the test set
- **shop_id** - unique identifier of a shop
- **item_id** - unique identifier of a product
- **item_category_id** - unique identifier of item category
- **item_cnt_day** - number of products sold. You are predicting a monthly amount of this measure
- **item_price** - current price of an item
- **date** - date in format dd/mm/yyyy
- **date_block_num** - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
- **item_name** - name of item
- **shop_name** - name of shop
- **item_category_name** - name of item category

Figure 1

For an easier understanding we divided data in two different ways and sorted by date. First way we divide our data is according to shop id. We get 59 different data set for 59 different shop. And the second one is according to item categories. We split data in item categories since there were lots of unique items, more than 20000, so using item

categories for creating new data sets was more sensible. A part of sorted data set for first shop can be seen in Figure-2.

Index	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
165355	01.02.2…	1	1	15660	76	1
166023	01.02.2…	1	1	12134	189	1
167439	02.02.2…	1	1	19811	221	3
179913	02.02.2…	1	1	7893	1473	3
165877	02.02.2…	1	1	18539	119	2
180457	02.02.2…	1	1	4907	1136	1

Figure 2