

Data science project proposal

By Cesar Miranda



ESIEE Paris

M.Sc. Management of Technology and Information Systems

**Predicting the future sales**

1. **The company**

The 1C company is a Russian firm that has presence in all over Russia having several stores in different parts of this country. Each store works with different stock and products.

1. **The goal**

Since the company has some stores distributed in Russia, the goal of this project is to predict the number of products that each store is going to sell the next month. For the purpose and the data of this project, I consider the term “next month” as November 2015.

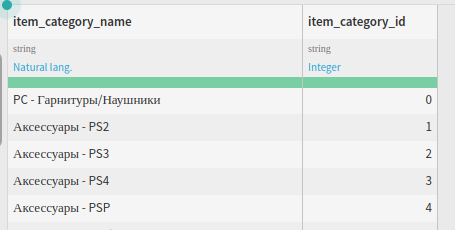
1. **The benefits for the customer**

1C Company can have different benefits that would help it to take better decisions and know a little bit more about the sales of the company as a conglomerate.

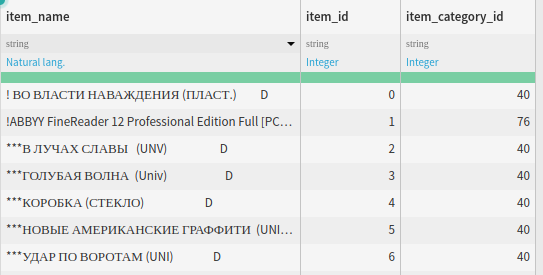
* Have the sufficient volume of stock to cover all the sales.
* Take decisions based on accurate predictions such as taking out products from a store where the minimum volume of sales is not reached.
* Improve the sales plan the company has for the store or the entire company.

1. **The datasets**

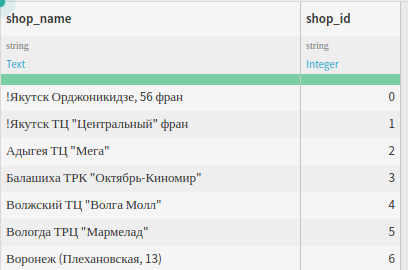
The datasets provided by the company contains the data of the store, the list of available products, the categories of these products, the sales information per store and per product and a test dataset that will help me to test the model I am going to build, for this scenario, I consider that the test dataset belongs to November 2015 (next month). The following pictures represent the data sets mentioned lines above.



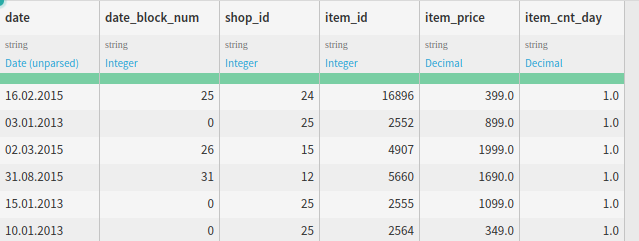
Item Category dataset



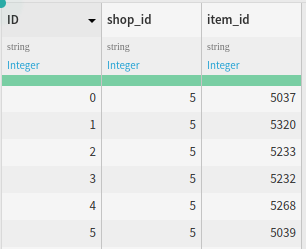
Item dataset



Shops dataset



Sales dataset



Test Dataset

1. **The deliverable**

Since the company wants to predict the number of sales per product and store for the next month, the deliverable of this project is to build a prediction model that can support this idea having at least 60% of accuracy.

1. **Proposed methodology**

Based on the project and after examining the available datasets provided by the company, I suggest three steps to build the model.

1. Analyze and generate the features for each dataset

In this step, I firstly analyze the datasets shops, items and item categories and I observe that there is one common characteristic. This characteristic is that all of these datasets are basically lists of objects (store, product and category) and I believe that I need to learn from the description of each line, so I generate more features based on the text column to get some insights.

1. Consolidate the daily information to monthly

Because of the prediction has to be on monthly basis and all the sales information is on daily basis, I need to group the sales per product and month to obtain the total number of items per product and store. After that, I need to get the mean of the sales and the number of items sold the previous month.

1. Analyze the algorithms and choose the best one

After adding the necessary features to each dataset and consolidate the sales information, all these datasets need to be joined and I have to start creating our possible models. For this case, I suggest to use Neural Network, Extra trees, random forest, or decision trees and try to do some ensemble algorithms.

1. **Ethical Implications**

The unique ethical implication I could find for this project is that I, as a consultant, do not have to show or share the data of the company to any other person who does not belong to the project because of the several consequences it could bring to the customer company. For example, this data can be useful for other competitors to have the knowledge of how well or bad 1C company is going and therefore, take decisions to affect it.

1. **Future considerations**

The approach that is going to be taken to achieve the goal of this project does not take into consideration any external data. Therefore, if I wanted to use any other factor, I could get some information about the average temperature in those months, the season of each month and any special event that happened in inside a month. Furthermore, the model and the data linked to the model can be redefined in order to predict the volume of sales not only for a specific month, but also for any month in advance. Unfortunately, all these considerations will not be taken as scope of this project.