DATA SCIENCE PROJECT REPORT

By Cesar Miranda

Predict future sales

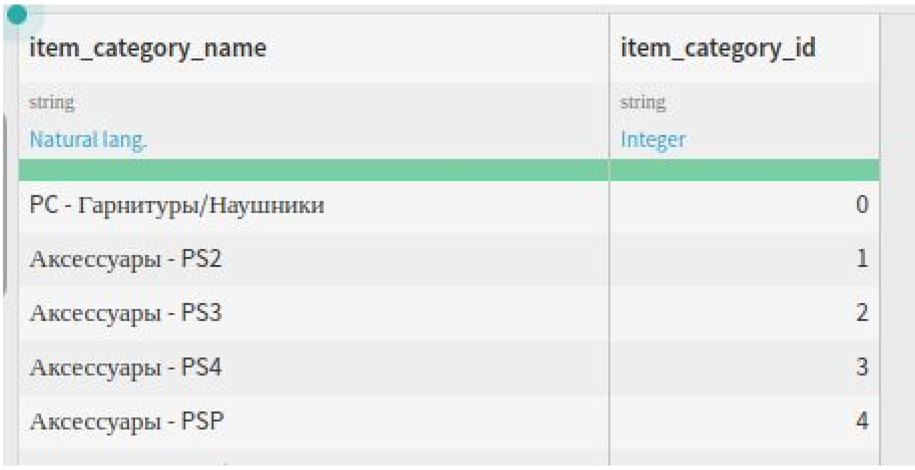
Note: For further explanation, please, see the Jupiter HTML Notebook file.

As was stated in the proposal document, there are 4 data sets that contain the data of the sales, the list of products, the product categories and the list of stores. Furthermore, the execution of this projects was done by following the proposed methodology mentioned in the proposal document which has three steps: analyze and generate the features for each dataset, consolidate the daily information to monthly and analyze the algorithms and choose the best one.

a. Analyze and generate the features for each dataset

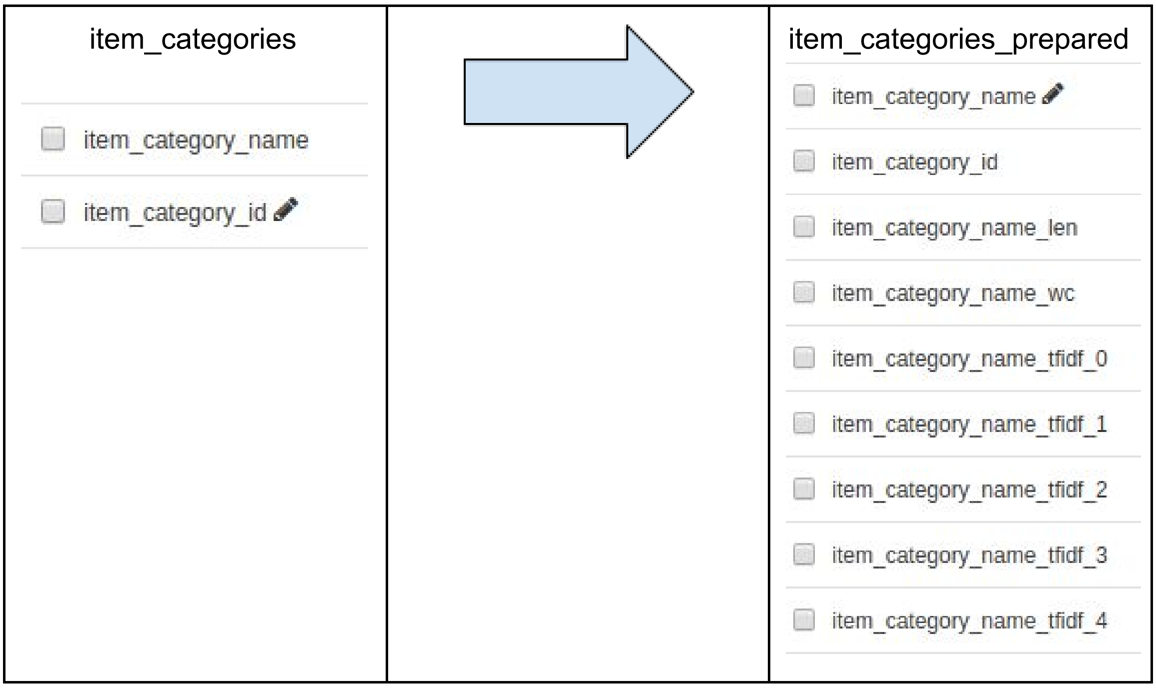
In this step, I firstly analyzed the datasets shops, items and item categories and I observed that there was one common characteristic. This characteristic was that all of these datasets were basically lists of objects (store, product and category) and I believed that I needed to learn from the description of each line, so I generated more features based on the text column to get some insights.

For example, in the product category dataset, at the beginning I only had 2 columns which indicate the name and the ID, as it is seen in the image below.



item_category_name	item_category_id
string Natural lang.	string Integer
PC - Гарнитуры/Наушники	0
Аксессуары - PS2	1
Аксессуары - PS3	2
Аксессуары - PS4	3
Аксессуары - PSP	4

After analyzing this dataset, I added some other features such as the length of the category name, the number of words that the category name has, and some tfid columns that will help the computer to know the importance of a word among all the category names and in this way learn something from them.



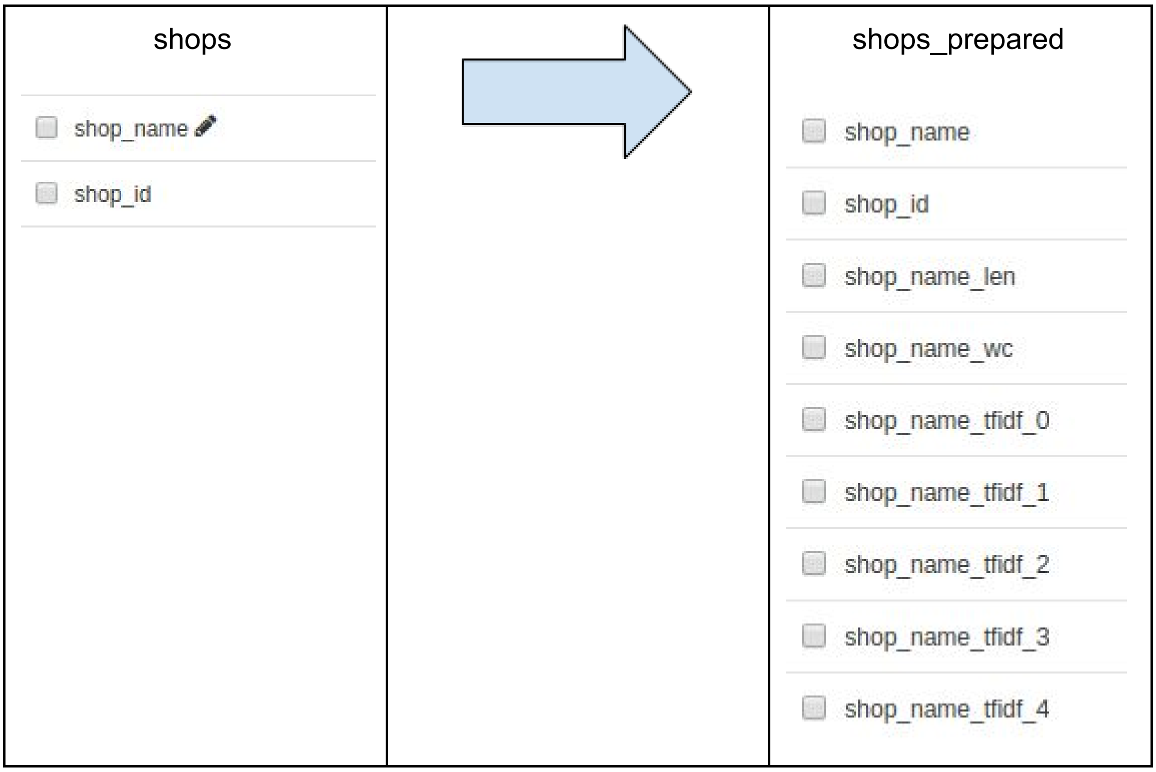
item_categories

- □ item_category_name
- □ item_category_id ✎

item_categories_prepared

- □ item_category_name ✎
- □ item_category_id
- □ item_category_name_len
- □ item_category_name_wc
- □ item_category_name_tfidf_0
- □ item_category_name_tfidf_1
- □ item_category_name_tfidf_2
- □ item_category_name_tfidf_3
- □ item_category_name_tfidf_4

The same principle was applied to the shops dataset and items dataset.

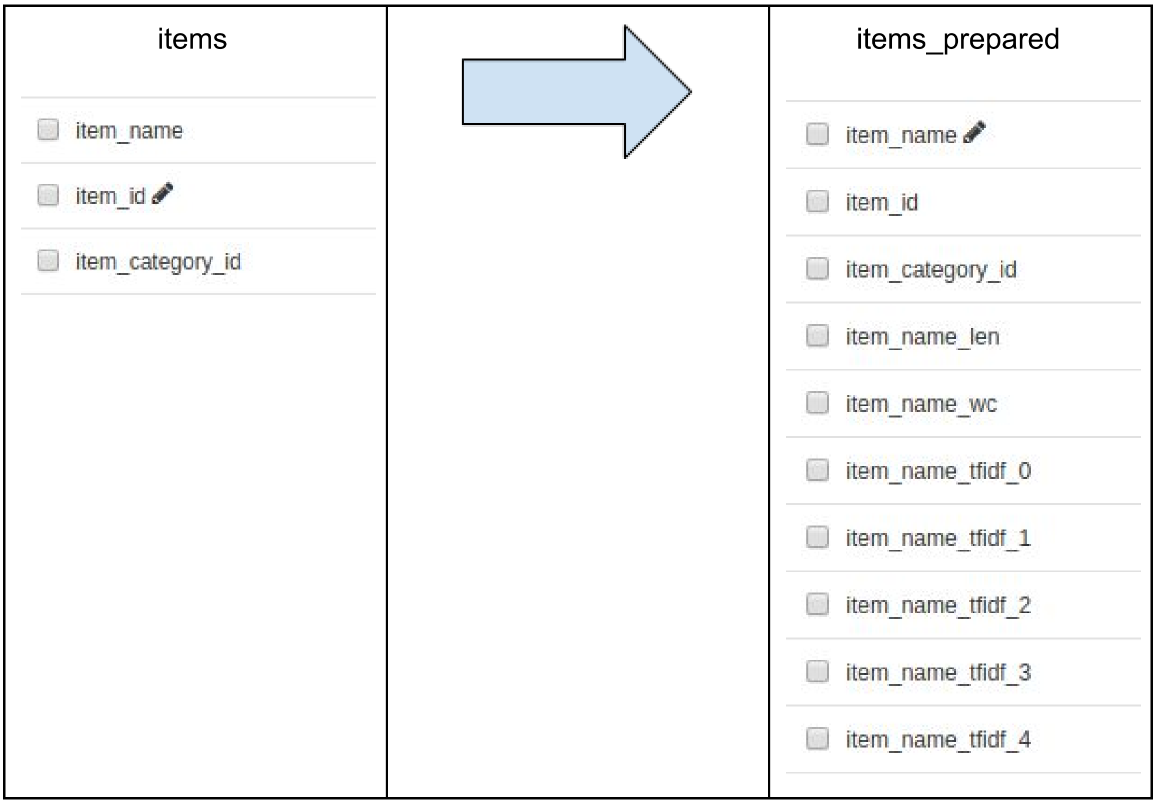


shops

- □ shop_name ✎
- □ shop_id

shops_prepared

- □ shop_name
- □ shop_id
- □ shop_name_len
- □ shop_name_wc
- □ shop_name_tfidf_0
- □ shop_name_tfidf_1
- □ shop_name_tfidf_2
- □ shop_name_tfidf_3
- □ shop_name_tfidf_4



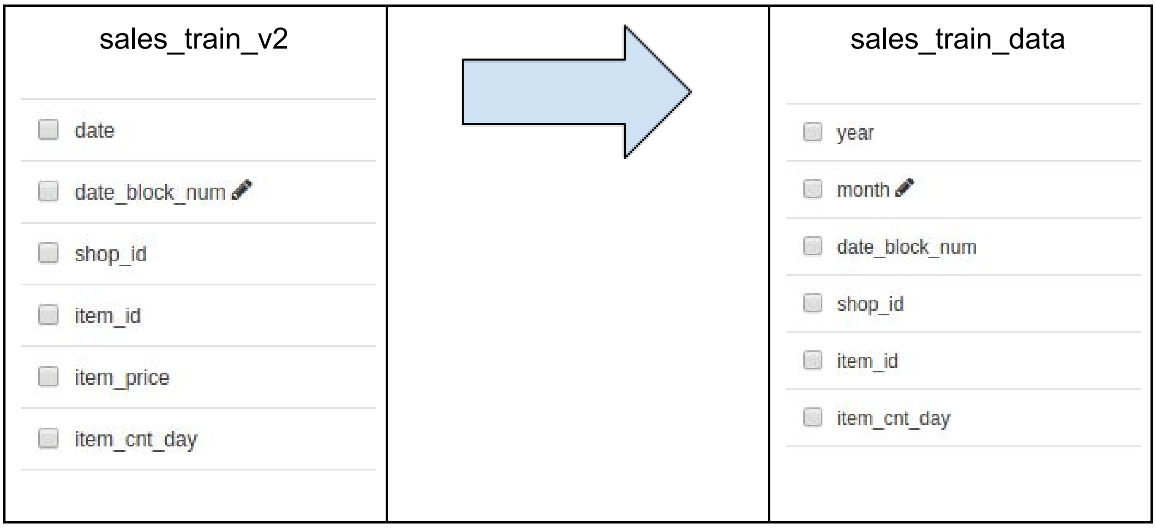
items

- □ item_name
- □ item_id ✎
- □ item_category_id

items_prepared

- □ item_name ✎
- □ item_id
- □ item_category_id
- □ item_name_len
- □ item_name_wc
- □ item_name_tfidf_0
- □ item_name_tfidf_1
- □ item_name_tfidf_2
- □ item_name_tfidf_3
- □ item_name_tfidf_4

Second, I observe that in the sales dataset I can extract the month and year from the date and eliminate the item price because for the purpose of this project, this column is not relevant.



sales_train_v2

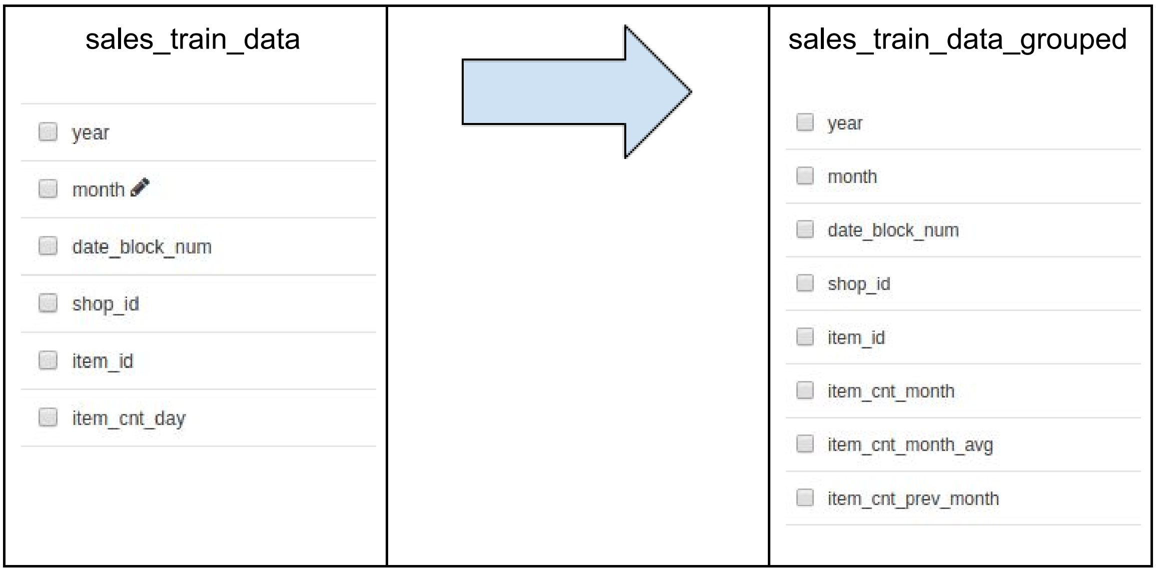
- □ date
- □ date_block_num ✎
- □ shop_id
- □ item_id
- □ item_price
- □ item_cnt_day

sales_train_data

- □ year
- □ month ✎
- □ date_block_num
- □ shop_id
- □ item_id
- □ item_cnt_day

b. Consolidate the daily information to monthly

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



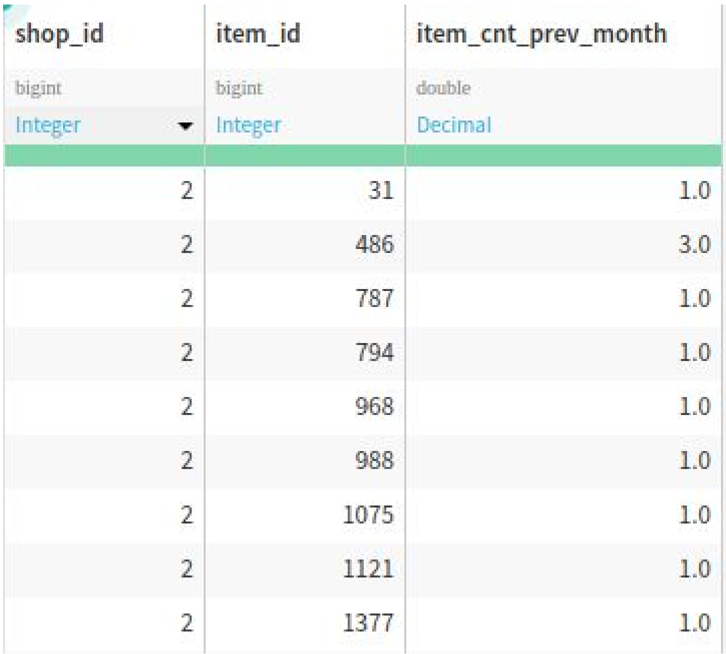
sales_train_data

- ☐ year
- ☐ month ✎
- ☐ date_block_num
- ☐ shop_id
- ☐ item_id
- ☐ item_cnt_day

sales_train_data_grouped

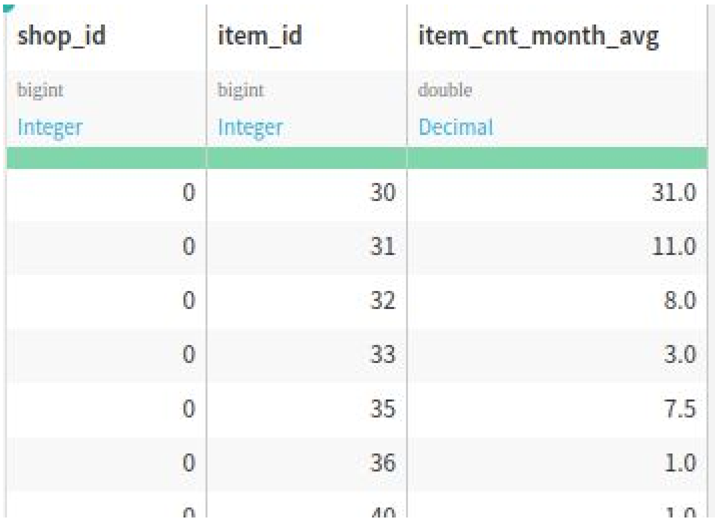
- ☐ year
- ☐ month
- ☐ date_block_num
- ☐ shop_id
- ☐ item_id
- ☐ item_cnt_month
- ☐ item_cnt_month_avg
- ☐ item_cnt_prev_month

I also generated from sales_train_data two other datasets containing the grouped quantity of sold items of the previous month (dataset 1: sales_shop_item_prev_month) and the other dataset the mean of the number of items sold per product and store (dataset 2: sales_shop_item_monthly). These datasets will help me in the next step to make the test dataset in the same structure than the train dataset.



shop_id	item_id	item_cnt_prev_month
bigint Integer	bigint Integer	double Decimal
2	31	1.0
2	486	3.0
2	787	1.0
2	794	1.0
2	968	1.0
2	988	1.0
2	1075	1.0
2	1121	1.0
2	1377	1.0

Dataset 1: sales_shop_item_prev_month

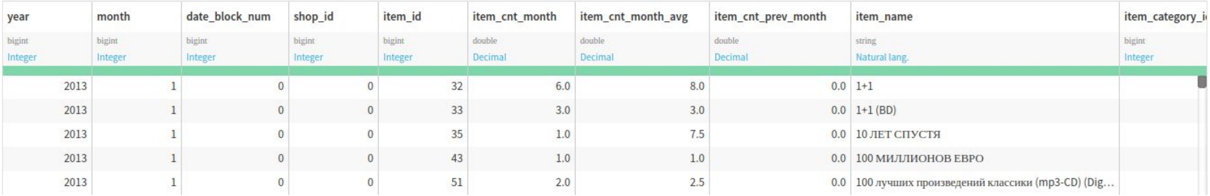


shop_id	item_id	item_cnt_month_avg
bigint Integer	bigint Integer	double Decimal
0	30	31.0
0	31	11.0
0	32	8.0
0	33	3.0
0	35	7.5
0	36	1.0
0	40	1.0

Dataset 2: sales_shop_item_monthly

c. Analyze the algorithms and choose the best one

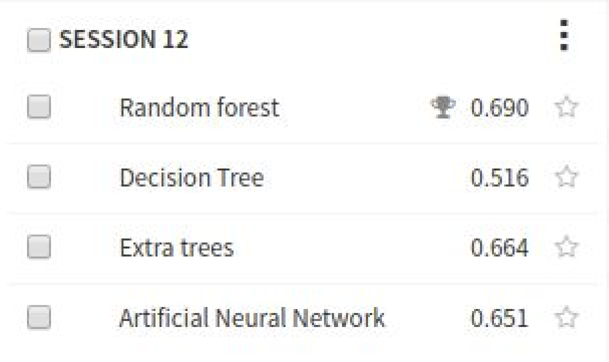
After having generated the new shops, item categories and items datasets with all the new features, I had to join them with the dataset called “sales_train_data_grouped” and start generating the possible models. This new dataset is called “sales_data_prepared”.



year	month	date_block_num	shop_id	item_id	item_cnt_month	item_cnt_month_avg	item_cnt_prev_month	item_name	item_category_i
bigint Integer	bigint Integer	bigint Integer	bigint Integer	bigint Integer	double Decimal	double Decimal	double Decimal	string Natural lang.	bigint Integer
2013	1	0	0	32	6.0	8.0	0.0	1+1	
2013	1	0	0	33	3.0	3.0	0.0	1+1 (BD)	
2013	1	0	0	35	1.0	7.5	0.0	10 ЛЕТ СПУСТЯ	
2013	1	0	0	43	1.0	1.0	0.0	100 МИЛЛИОНОВ ЕВРО	
2013	1	0	0	51	2.0	2.5	0.0	100 лучших произведений классики (mp3-CD) (Dig…	

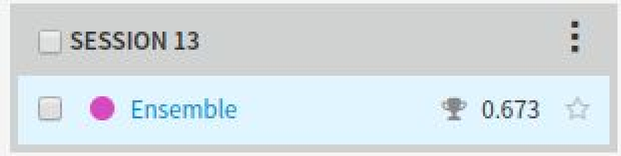
sales_data_prepared

After creating a new analysis and selected the possible algorithms to use, the DSS showed me that the most suitable algorithm is the Random Forest so this is the one with which I built the model.



□ SESSION 12		⋮
□ Random forest	🏆 0.690	☆
□ Decision Tree	0.516	☆
□ Extra trees	0.664	☆
□ Artificial Neural Network	0.651	☆

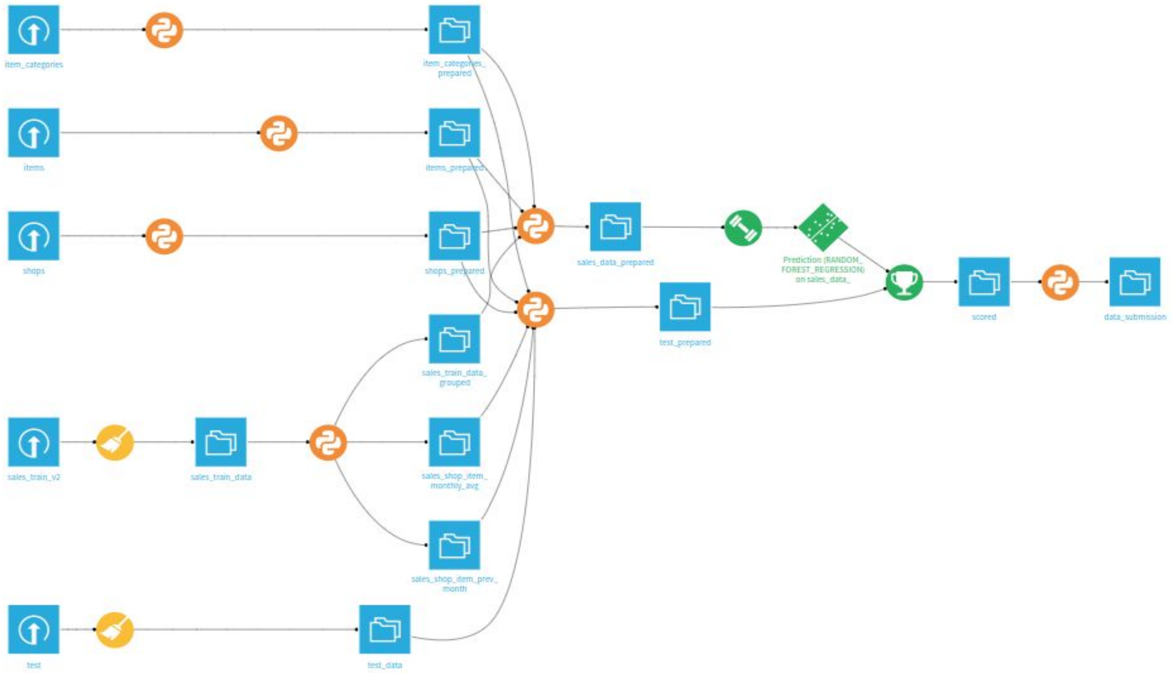
Because I wanted to try other kind of algorithms, I ensembled the Random Forest, Extra trees and ANN, but the result was lower than the others so I kept using Random Forest.



□ SESSION 13

□ ● Ensemble 0.673 ☆

Finally, I joined the new shops, item categories, items datasets, sales_shop_item_prev_month, sales_shop_item_monthly with the test dataset to give it the same structure than the sales_data_prepared dataset and evaluate the generated model.



Ethical Implication

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.

Future Considerations

1. **Use of external data**

Get the average weather conditions for each month of this dataset, the season of each month and if any special event that happened in inside a month to have a more realistic result.

2. **Improve the model to make it for the whole year**

Currently, the model proposed can only predict one month (Nov 2015) so the improvement needed would be to modify the model in order to make it for not only a specific month.