Predict Future Sales

Siyu Chen

Xi'an Shiyou University, China

Introduction

To predict total sales for every product and store in the next month. Provided with daily historical sales data. The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations.

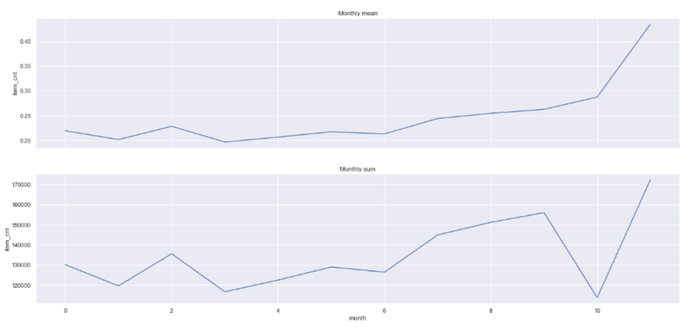
Fields	Field to explain
shop_id	unique identifier for a store.
item_id	a unique identifier for a product.
item_category_id	a unique identifier for a category.
item_cnt_day	quantity of products sold.
item_price	the current price of the goods.
date_block_num	a consecutive month.
item_name	product name.
shop_name	shop name.

Data Processing

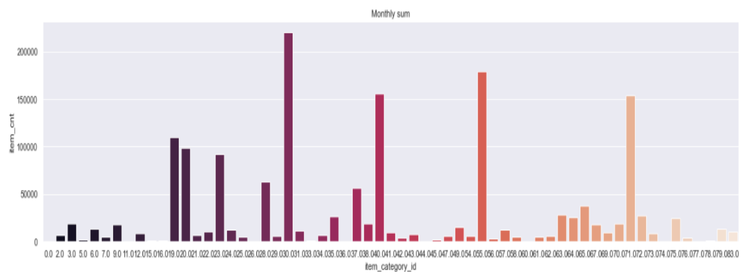
- Use only shop_ID and item_ID that appear in the test set.
- Use only data whose price is greater than 0.
- Delete the unwanted columns, aggregate them monthly, and get a new data set.
- Remove outliers.
- Drop the text features.

Data Visualization

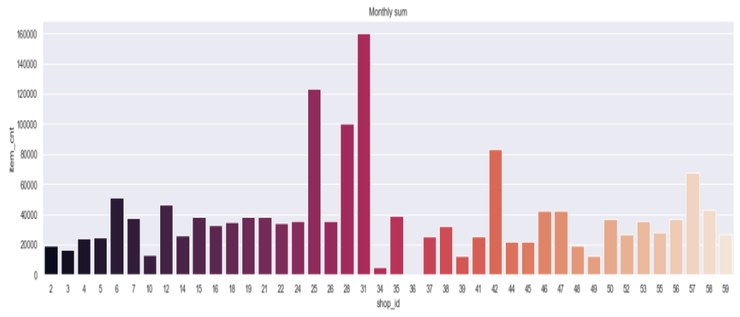
Draw line graph of average monthly sales and total monthly sales, average monthly sales were rising and peaking in December.



By visualizing by category, category 30 sells best, and only a few items sell well. Grouping by store number, most stores sell about the same, with only three doing particularly well.



Grouping by category



Grouping by shop_id

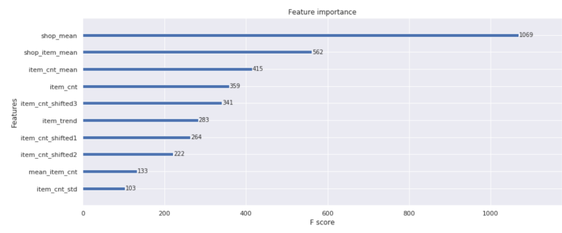
Feature Engineering

- The first 3-27 blocks are used for training.
- the five blocks 28-32 are used for verification.
- test set will be block 33 and our predictions should reflect block 34 values.
- Mean Encoding:
  - Find the average monthly sales volume by store number.
  - Find the average monthly sales volume by commodity number group.
  - Find the average monthly sales volume of each item in each store.
  - Group by year, find the average annual sales.
  - Group by month, find the average monthly sales.

Build Model

- linear Regression
  Train rmse: 0.7347132326333324
  Validation rmse: 0.7755311093532987
- Random Forest
  Train rmse: 0.6985868322226099
  Validation rmse: 0.776123635046122
- XGBoost
  Train rmse: 0.697475453300762
  Validation rmse: 0.798117433161014

XGBoost feature importance.



Feature importance

Output the predictions of the first level model.

	random_forest	linear_regression	xgboost
0	0.98	0.85	0.44
1	0.06	0.06	0.10
2	0.85	1.79	0.50
3	0.00	0.06	0.10
4	0.06	0.06	0.10

Ensembling

- To combine the 1st level model predictions,to use a simple linear regression.
- Trained on validation set using the 1st level models predictions as features.
  Make predictions on test set using the 1st level models predictions as features.
  Train rmse: 0.7654489715389068
- Output Dataframe:

	ID	item_cnt_month
0	0	0.85
1	1	0.08
2	2	1.29
3	3	0.06
4	4	0.08
5	5	0.96
6	6	1.25
7	7	0.21
8	8	1.99
9	9	0.06