Predict Future Sales

Siyu Chen

Xi’an Shiyou University

2020-09-28

Overview

Problem Definition

Analyze Data

Files

predict total sales for every product and store in the next month.

- sales_train.csv: Training set 2013 to October 2015.
- test.csv: Test set forecast sales of these stores and products in November 2015.
- items.csv: Additional information about the merchandise/product.
- item_categories.csv: Additional information on the categories of goods.
- shops.csv: Additional information about the store.

File Field Description

- 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.
- item_category_name : The name of the project category.

Data Statistics

- Statistic the data in the training set

Table 1: Statistic the data in the training set

	datebknum	*shop_id*	*item_id*	*item_price*	*itemcntday*	*item_ctg_id*
count	2935849.00	2935849.00	2935849.00	2935849.00	2935849.00	2935849.00
mean	14.57	33.00	10197.23	890.62	1.24	40.00
std	9.42	16.23	6324.30	1726.44	2.62	17.10
min	0.00	0.00	0.00	-1.00	-22.00	0.00
25%	7.00	22.00	4476.00	249.00	1.00	28.00
50%	14.00	31.00	9343.00	399.00	1.00	40.00
75%	23.00	47.00	15684.00	999.00	1.00	55.00
max	33.00	59.00	22169.00	307980.00	2169.00	83.00

Data Processing

Three ways to process data

- Once we get the data set, we need to process the data.
  - Use only shop_id and item_id that appear in the test set.
    Data set size before: 2935849
    Data set size after: 1224439
  - Use only data whose price is greater than 0.
  - Drop the text features.
  - Delete the unwanted columns, aggregate them monthly, and get a new data set.
    Get new features:Total monthly sales price item_price and Monthly gross sales item_cnt.

Removing Outliers

- Discard outliers

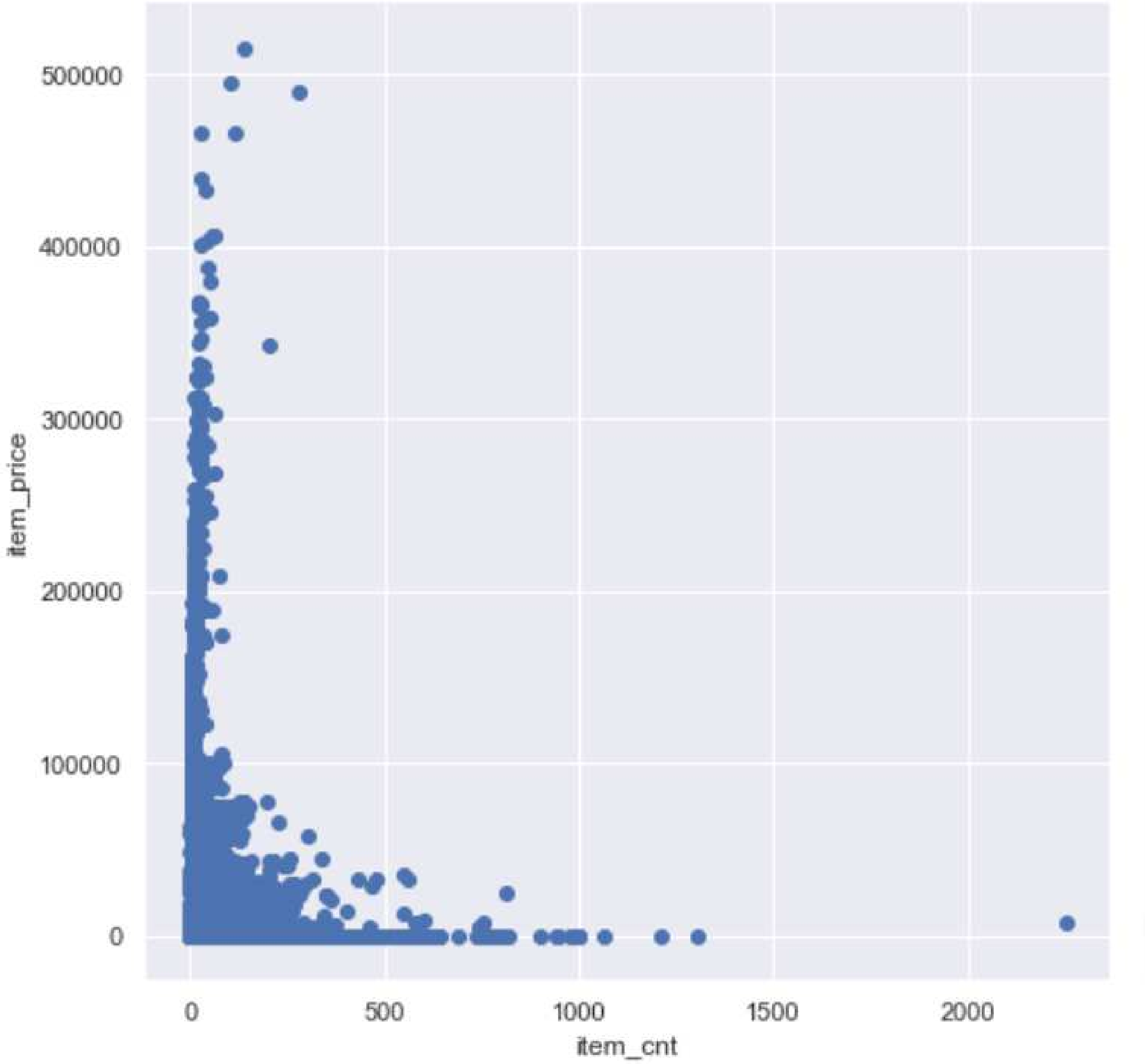

Figure 1: Checking for outliers

Data Visualization

Extraction Time Features

- We need to extract the time characteristics before drawing.
  - Divide date_block_num into months and years.
    Get the title monthly_mean for item_cnt.mean.
    Get the title monthly_sum for item_cnt.sum.

Annual Sales Performance

- Look at the line chart below
  - The average monthly sales are on the rise.
  - Monthly sales reached their peak in December.

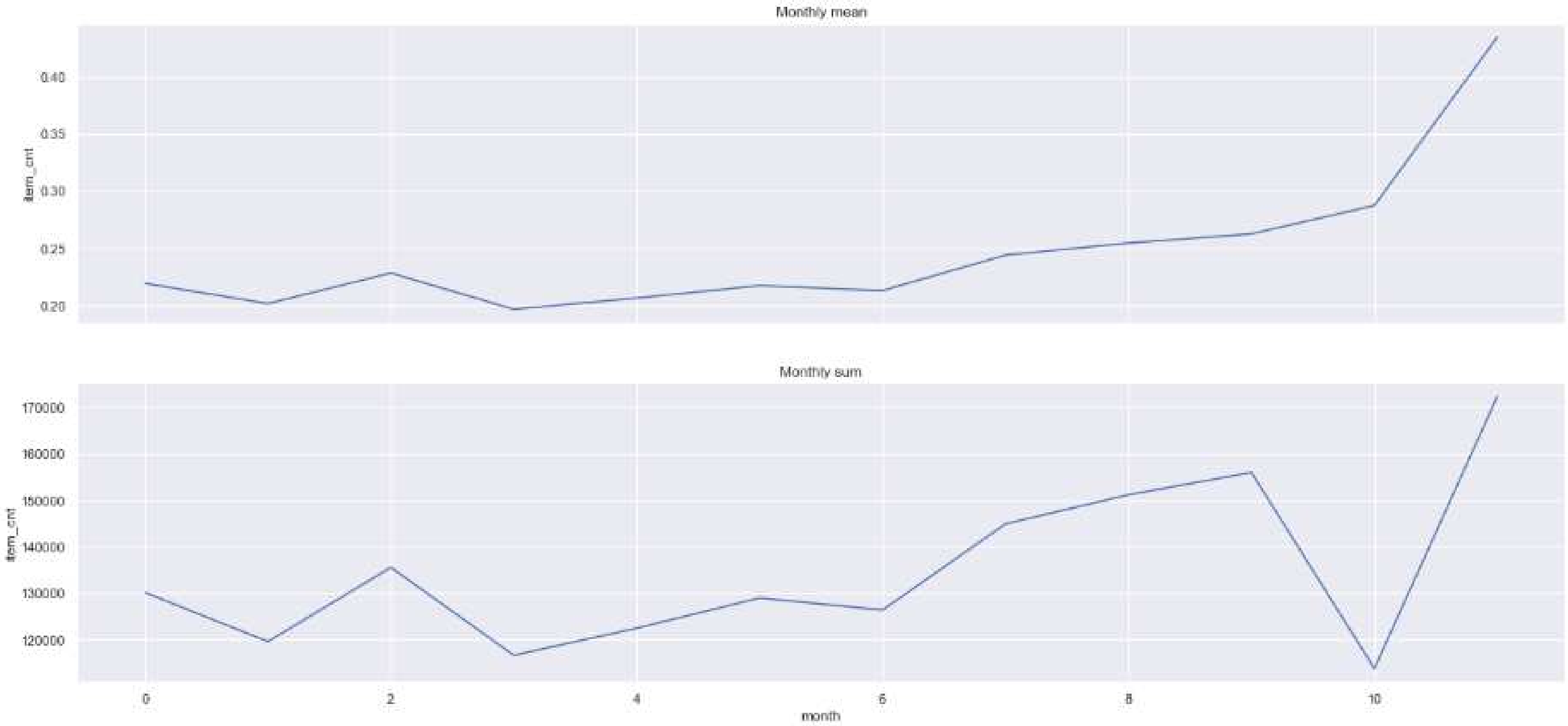

Figure 2: Annual sales performance

What category sells more?

- Category 30 is the best seller.
- Also only few of the categories seems to hold most of the sell count.

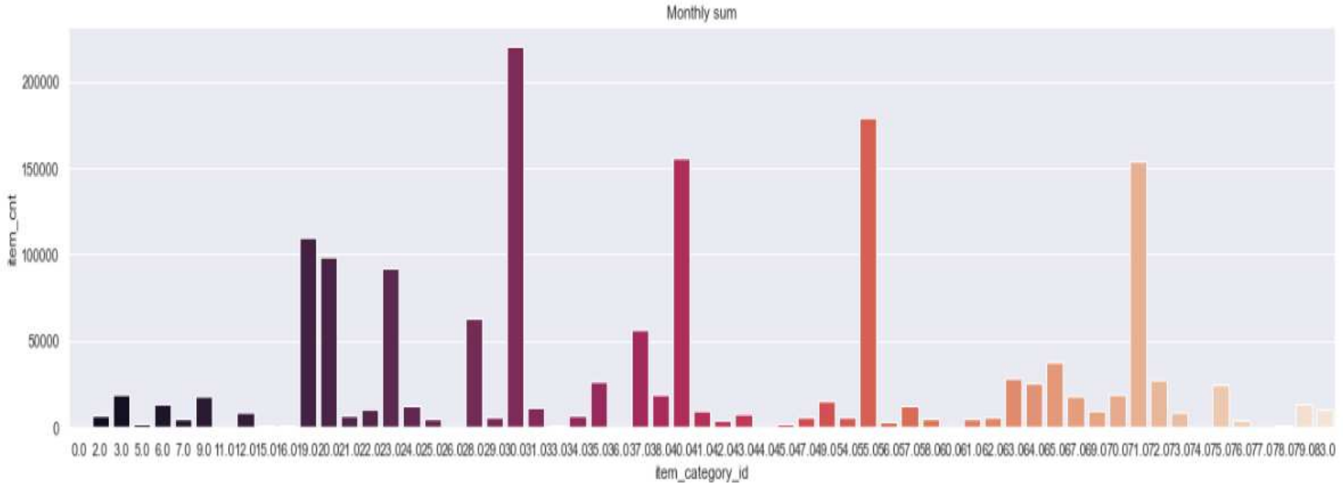

Figure 3: category

What shop sells more?

- Most shops have similar sell rate.
- 3 shops have a much higher rate, this is indicative of the shop size.

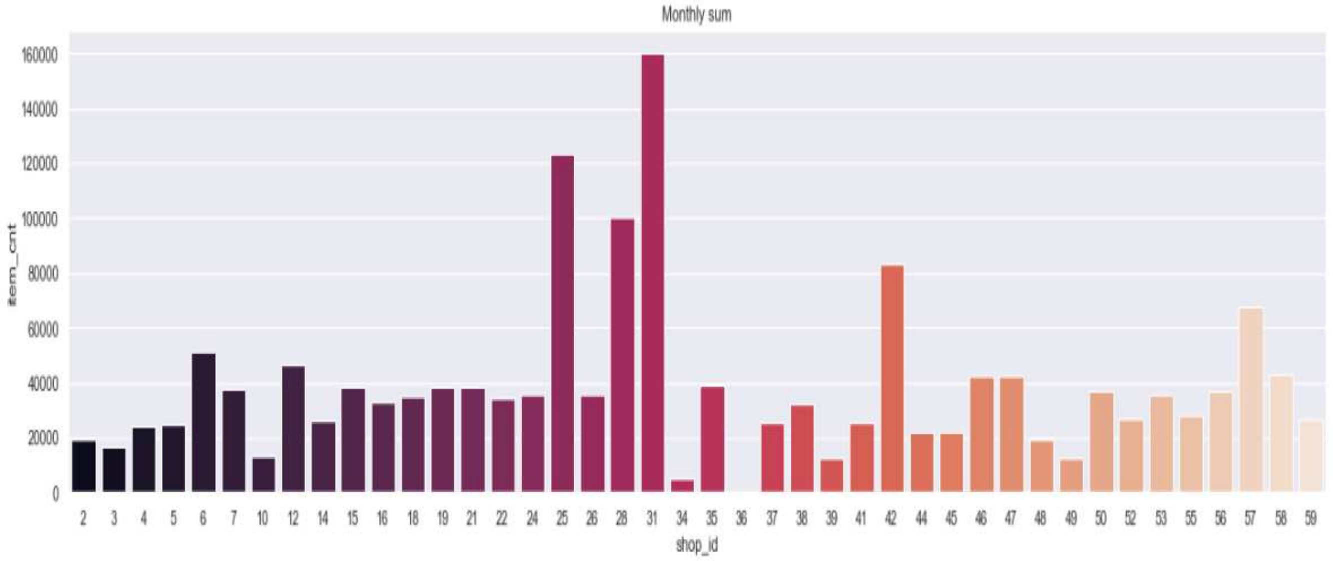

Figure 4: shop

Feature Engineering

Feature Extraction

- We can extract new features from existing data.
  - item_cnt_month:The monthly sales volume of each item in each store, sorted by month.
  - hist_min(max)_item_price:Figure out the maximum and minimum monthly sales of each item by month.
  - price_increace(decreace):How much each item's price changed from its (lowest/highest) historical price.
  - item_cnt_max,item_cnt_mean,item_cnt_std:Maximum, minimum, average, and median monthly sales of each item in each store.

Partition training set

- Training set
  - Use the first three months to generate features and implement functionality. The 3-27 blocks are used for training.
- Validation set
  - the five blocks 28-32 are used for verification.
    It is used to verify the accuracy of the model. The data is divided according to time because of its time characteristics.
- Test set
  - We want to predict for "date_block_num" 34 so our test set will be block 33 and our predictions should reflect block 34 values.
    In other words we use block 33 because we want to forecast values for block 34.

Mean Encoding

- Done after the train/validation split.
  - 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, grouped by store and item number.
  - Group by year, find the average annual sales.
  - Group by month, find the average monthly sales.
- Add meand encoding features to train set. Add meand encoding features to validation set.

Build Model

Linear Regression and Random Forest

- 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.

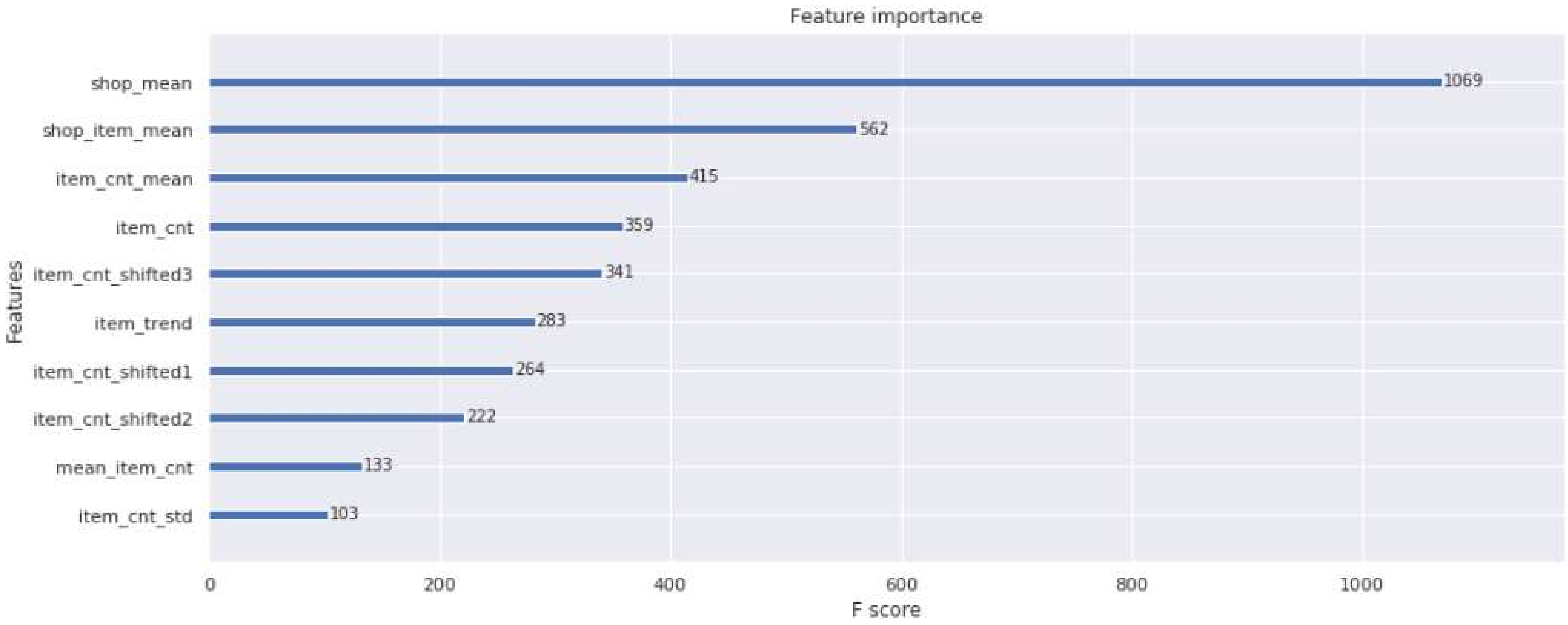

Figure 5: Feature importance

Create New Datasets

Table 2: Predictions from first level models

	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.
  This is the model that will combine the other ones to hopefully make an overall better prediction.
- Make predictions on test set using the 1st level models predictions as features.

Ensemble Diagram

- Look at the line chart below
  - Here is an image to help the understanding

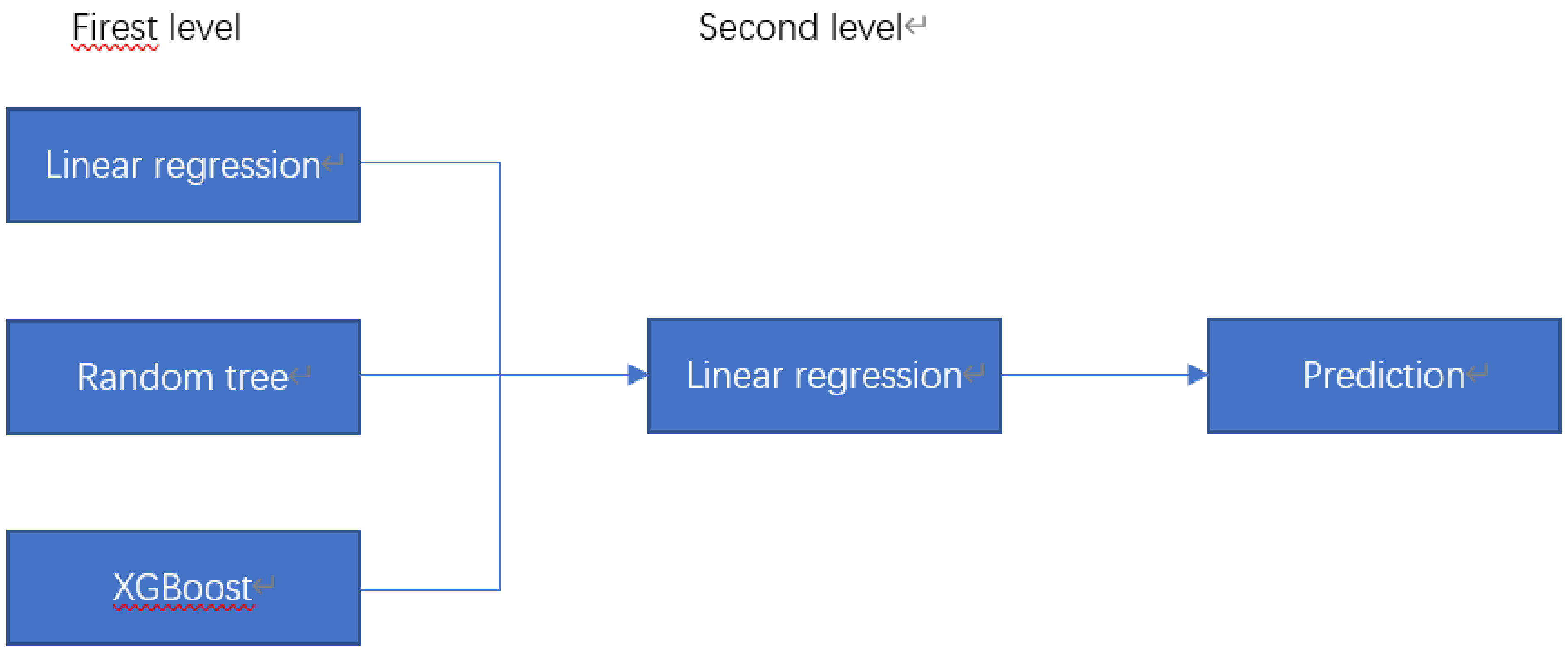

Figure 6: Ensemble diagram

Output Dataframe

Table 3: 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

Contact Information

Thank you for listening!

Siyu Chen

Xi'an Shiyou University

✉ 785987165@QQ.COM