Raymond A. Mason
School of Business
WILLIAM & MARY

Predicting Future Sales

Team 20 - You Wu, Vanessa Guzman, Frank Wan, Elie Baaklini

Agenda

1. Overview
2. Critique of Other Competition Submissions
3. Our Solution
4. Final Prediction

Raymond A. Mason
School of Business
WILLIAM & MARY

Overview

Objective: Predict total sales (quantity) for every product and store in the next month (November 2015)

Dataset: includes 6 csv files from 1C company (Russian software company)

- items:

	A	B	C
1	item_name	item_id	item_category_id
2	ÐÐ¡Ð¢Ð˜ ÐÐÐ’ÐÐ–Ð”Ð•ÐÐ˜Ð¯ (ÐŸÐ›ÐÐ	0	40
3	Professional Edition Full [PC, Ð¦Ð¸Ñ„Ñ€Ð	1	76
4	Ð’ Ð›Ð£Ð§ÐÐ¥ Ð¡Ð›ÐÐ’Ð« (UNV)	2	40
5	“ÐžÐ›Ð£Ð‘ÐÐ¯ Ð’ÐžÐ›ÐÐ (Univ)	3	40
6	ÐšÐžÐ ÐžÐ‘ÐšÐ (Ð¡Ð¢Ð•ÐšÐ›Ðž)	4	40
7	œÐ•Ð Ð˜ÐšÐÐÐ¡ÐšÐ˜Ð• Ð“Ð ÐÐ¤ÐФÐ˜Ð	5	40

- item_categories:

	A	B
1	item_category_name	item_category_id
2	PC - Ð“Ð°Ñ€Ð½Ð¸Ñ‚ÑƒÑ€Ñ‹/ÐÐ°ÑƒÑˆÐ½Ð¸ÐºÐ¸	0
3	ÐÐºÑÐµÑÑÑƒÐ°Ñ€Ñ‹ - PS2	1
4	ÐÐºÑÐµÑÑÑƒÐ°Ñ€Ñ‹ - PS3	2
5	ÐÐºÑÐµÑÑÑƒÐ°Ñ€Ñ‹ - PS4	3
6	ÐÐºÑÐµÑÑÑƒÐ°Ñ€Ñ‹ - PSP	4
7	ÐÐºÑÐµÑÑÑƒÐ°Ñ€Ñ‹ - PSVita	5

- shops:

	A	B
1	shop_name	shop_id
2	ÑƒÑ‚Ñ‚Ðº ÐžÑ€Ð´Ð¶Ð¾Ð½Ð¸ÐºÐ¸Ð´Ð·Ðµ, 56 Ñ„Ñ€	0
3	ƒÑ‚Ñ‚Ðº Ð¢Ð¦ "Ð¦ÐµÐ½Ñ‚Ñ€Ð°Ð»ÑŒÐ½Ñ‹Ð¹" Ñ„	1
4	ÐÐ´Ñ‹Ð³ÐµÑ Ð¢Ð¦ "ÐœÐµÐ³Ð°"	2
5	Ñˆ Ð¸Ñ…Ð° Ð¢Ð Ðš "ÐžÐºÑ‚Ñ Ð±Ñ€ÑŒ-ÐšÐ¸Ð½Ð¾	3
6	Ð¾Ð»Ð¶ÑÐºÐ¸Ð¹ Ð¢Ð¦ "Ð’Ð¾Ð»Ð³Ð° ÐœÐ¾Ð»Ð	4
7	Ð’Ð¾Ð»Ð¾Ð³Ð´Ð° Ð¢Ð Ð¦ "ÐœÐ°Ñ€Ð¼ÐµÐ»Ð°Ð´	5

Overview

- sales-train: training set

	A	B	C	D	E	F
1	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
2	02.01.2013	0	59	22154	999	1
3	03.01.2013	0	25	2552	899	1
4	05.01.2013	0	25	2552	899	-1
5	06.01.2013	0	25	2554	1709.05	1
6	15.01.2013	0	25	2555	1099	1
7	10.01.2013	0	25	2564	349	1

- test: test set

	A	B	C
1	ID	shop_id	item_id
2	0	5	5037
3	1	5	5320
4	2	5	5233
5	3	5	5232
6	4	5	5268
7	5	5	5039

- sample: sample submission file in the correct format

	A	B
1	ID	item_cnt_month
2	0	0.5
3	1	0.5
4	2	0.5
5	3	0.5
6	4	0.5
7	5	0.5
8	6	0.5
9	7	0.5
10	8	0.5
11	9	0.5
12	10	0.5
13	11	0.5
14	12	0.5

Critique of Other Competition Submissions

```
#Random forest regressor model building
from sklearn.ensemble import RandomForestRegressor

RF_model = RandomForestRegressor()
RF_model.fit(X_prepared, y)
```

Random Forest

- **Pros**:
  - Suitable for large dataset
  - Easy data preparation comparing to other algorithms
- **Cons**:
  - Did not tune the hyperparameter
  - Computationally intense when number of trees is big
  - Will cause overfitting when noise is large
  - Lower output accuracy because it cannot guarantee the best tree
- **RMSE: 2.0182**

Critique of Other Competition Submissions

LightGBM (Light Gradient Boosting Machine)

- **Pros:**
  - Faster training speed, Light GBM uses a histogram-based algorithm i.e it buckets continuous feature values into discrete bins which fasten the training procedure
  - Replaces continuous values to discrete bins which results in lower memory usage
- **Cons:**
  - Light GBM split the tree leaf-wise which can lead to overfitting as it produces much complex trees
  - Did not do cross validation for hyperparameters
- **RMSE: 0.9610**

```
evals_result = {}
gbm = lgb.train(
        params,
        lgb_train,
        num_boost_round=3000,
        valid_sets=(lgb_train, lgb_eval),
        feature_name = feature_name,
        #categorical_feature = categorical_features,
        verbose_eval=5,
        evals_result = evals_result,
        early_stopping_rounds = 10)
```

Raymond A. Mason School of Business
WILLIAM & MARY

Our Solution

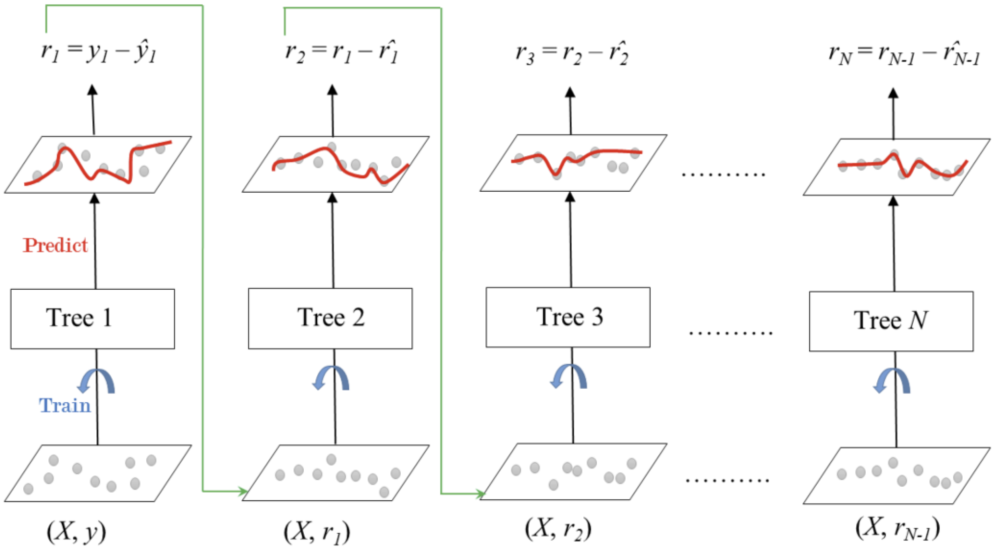
XGBoost: eXtreme Gradient Boosting--"ALL in One" algorithm

As a popular supervised-learning algorithm, XGBoost uses decision trees as base learners; combining many weak learners to make a strong learner to speed up and increase the performance of gradient boosted decision trees

- Regularization
- Parallel Processing
- Handling Missing Values
- Effective Tree Pruning

Our Solution

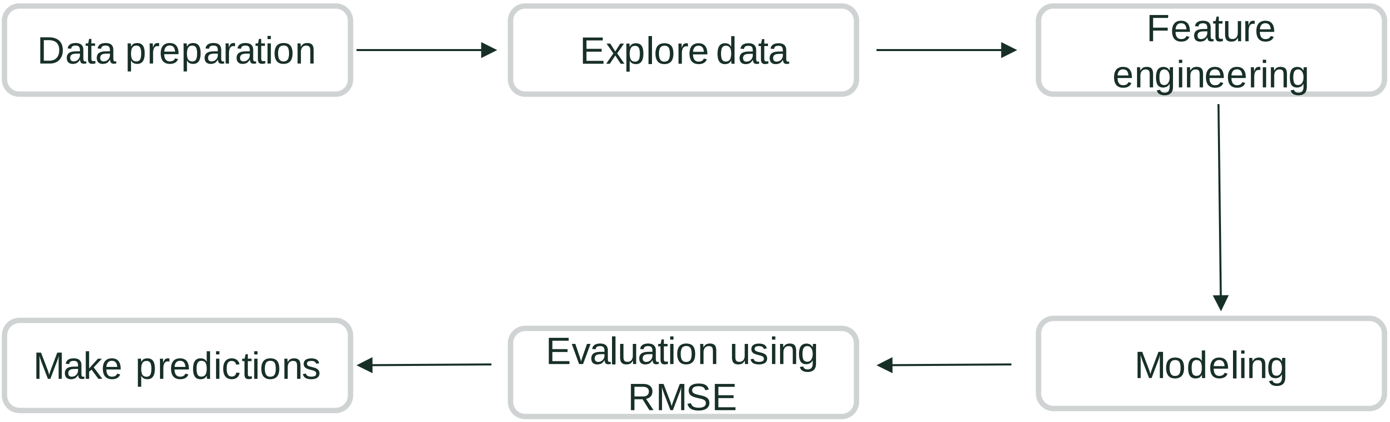
How does XGBoost work?



Raymond A. Mason School of Business
WILLIAM & MARY

Our Solution

Flow chart



Final Prediction

Performance: RMSE=0.877

Final prediction results:

ID	item_cnt_month
33050	0.01800462
32455	0.003886187
35412	0.019796828
32228	0.020295562
35385	0
32229	0.073716491
33048	0.007104043
32454	0
32446	0.036807135
33047	0.215048745
31748	0.042575739
33046	0.073391959
32370	0.073716491
34153	0.073716491
33045	0.027066618
30784	0.031092696
34150	0.073716491
31880	0.036855627
32286	0.121237382
30785	0.008510824
33044	0.015567578
32371	0.004501749

Raymond A. Mason School of Business
WILLIAM & MARY