Decrypt

BLACK FRIDAY

Data Analytics Project Presentation

Shangyou Wu, Haoran Tang, Chi Zhang, Jing Qian

Contents

JHSC Group

1. Data Description: 550k transaction records

```
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 12 columns):
User_ID                     537577 non-null int64
Product_ID                  537577 non-null object
Gender                      537577 non-null object
Age                         537577 non-null object
Occupation                  537577 non-null int64
City_Category               537577 non-null object
Stay_In_Current_City_Years  537577 non-null object
Marital_Status              537577 non-null int64
Product_Category_1          537577 non-null int64
Product_Category_2          370591 non-null float64
Product_Category_3          164278 non-null float64
Purchase                    537577 non-null int64
```

Features： 12

Records： 537577

User_ID: 5891 Unique Numbers

Product: 3623 Unique Numbers

Age: Age in bins(6)

City_Category: Category of the City (A,B,C)

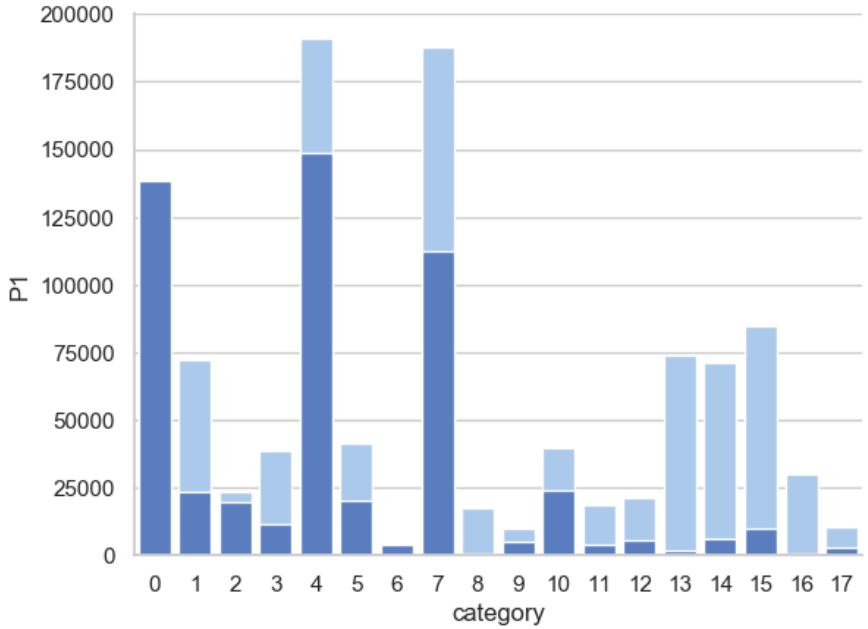
Stay_In_Current_City_Years: Number of years stay in current city

Purchase: Purchase amount in dollars

Dataset of 550 000 observations about the black Friday in a retail store
it contains different kinds of variables either numerical or categorical.
It contains missing values.

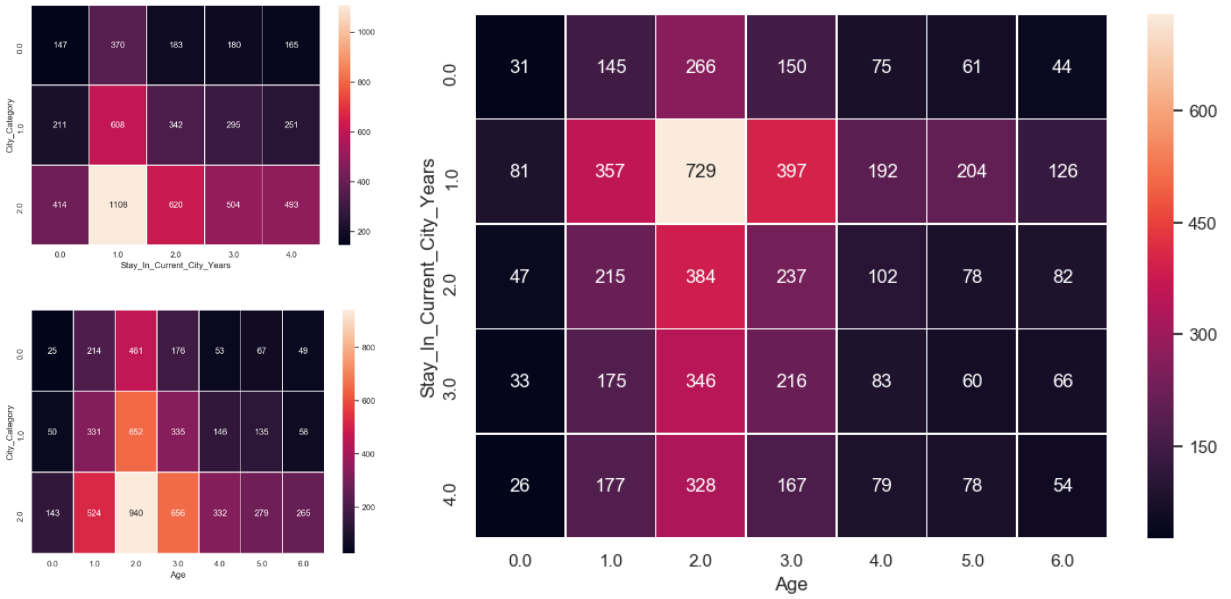
JHSC Group

1. Data Description: Basic Profile



Category Sales

Category 4,7,0
dominated the sales in Black Friday



User Profile

People in 26-35 ages living in City B&C for 1 year are our dominated customers

JHSC Group

2. Purchase Prediction: Linear Regression

Gender	17588.242528
Occupation	-391.366199
Marital_Status	-17759.951708
Stay_In_Current_City_Years	-48891.810739
Age_0	-17283.825241
Age_1	-40276.842929
Age_2	44130.282035
Age_3	16255.056812
Age_4	50337.512858
Age_5	-7672.635788
Age_6	-45489.547747
City_Category_0	1565.290560
City_Category_1	-31352.441224
City_Category_2	29787.150664

Objective

We wanted to use linear regression to construct a model to study the relationship between sales and different characteristics of consumers, and then to predict sales using the model.

Explanations

Gender: 1 means male and 0 means female.

Marital status: 1 means is married and 0 means not.

Age: Age_0 – Age_6 respectively means age of 0-17, 18-25, 26-35, 36-45, 46-50, 51-55, 55+.

JHSC Group

2. Purchase Prediction: Outcome

Gender	17588.242528
Occupation	-391.366199
Marital_Status	-17759.951708
Stay_In_Current_City_Years	-48891.810739
Age_0	-17283.825241
Age_1	-40276.842929
Age_2	44130.282035
Age_3	16255.056812
Age_4	50337.512858
Age_5	-7672.635788
Age_6	-45489.547747
City_Category_0	1565.290560
City_Category_1	-31352.441224
City_Category_2	29787.150664

Outcome

Gender: Since the coefficient of gender is positive, it means that for these three categories of products, males tend to purchase more than females.
Marital status: The negative shows that people tend to constrain their purchases after they are married. It's reasonable to guess that the product categories listed here are not necessity goods but "luxury" goods for families as married groups need to save some money for other spending for their families.
Age: Age group 2, 3 and 4 have positive coefficients, and they represent the group of people aging from 26 to 50. It's reasonable because people within this age range usually have higher purchase power and more time and strength to devote into shopping.
Stay_In_Current_City_Years: The relationship is negative. The goods may be durable goods like furniture (people usually purchase them when they newly come to a city and use them for a relatively long period of time).

JHSC Group

2. Purchase Prediction: Label the Users

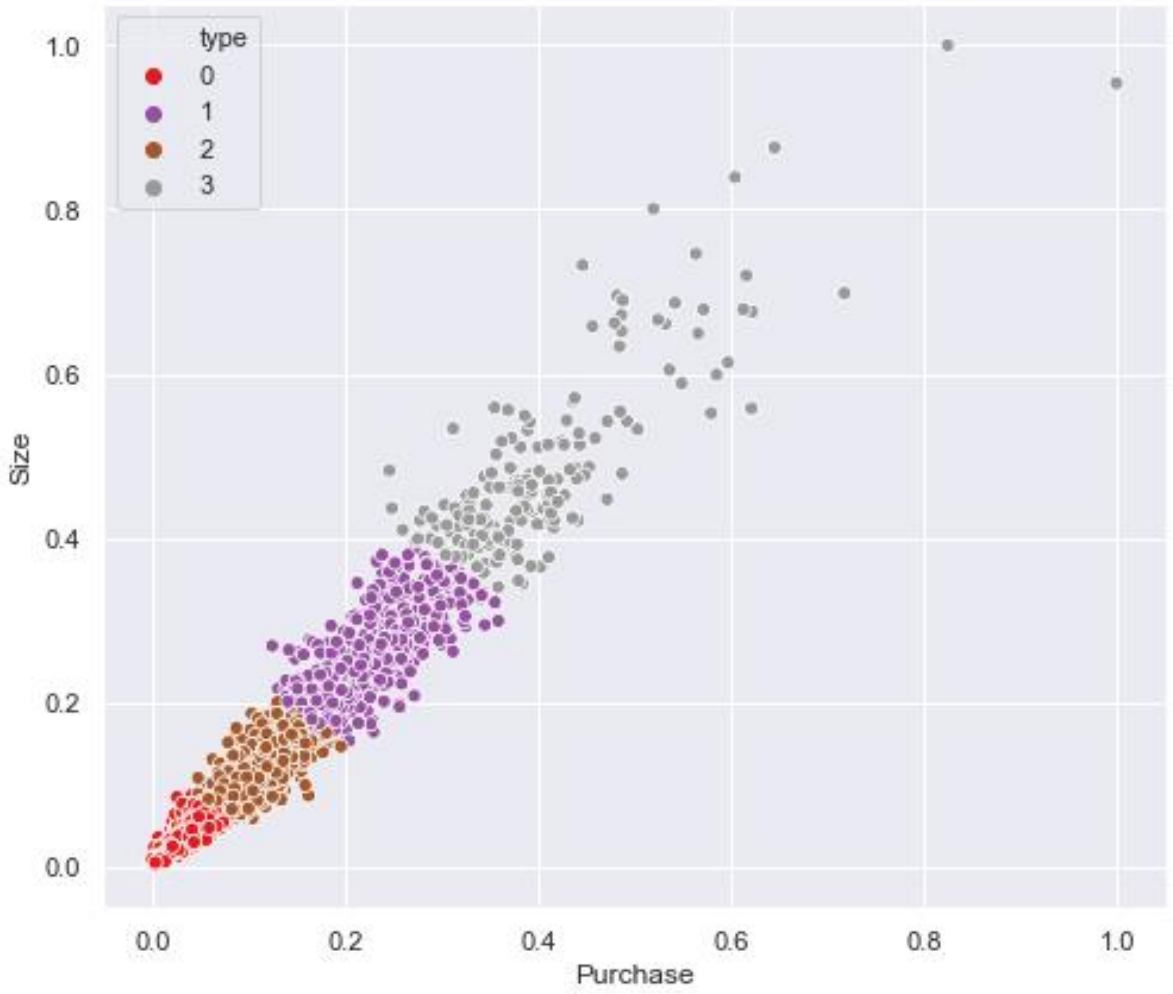
K-MEANS Algo.

k-means clustering aims to partition n observations into k clusters in which each record belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

4 Clusters

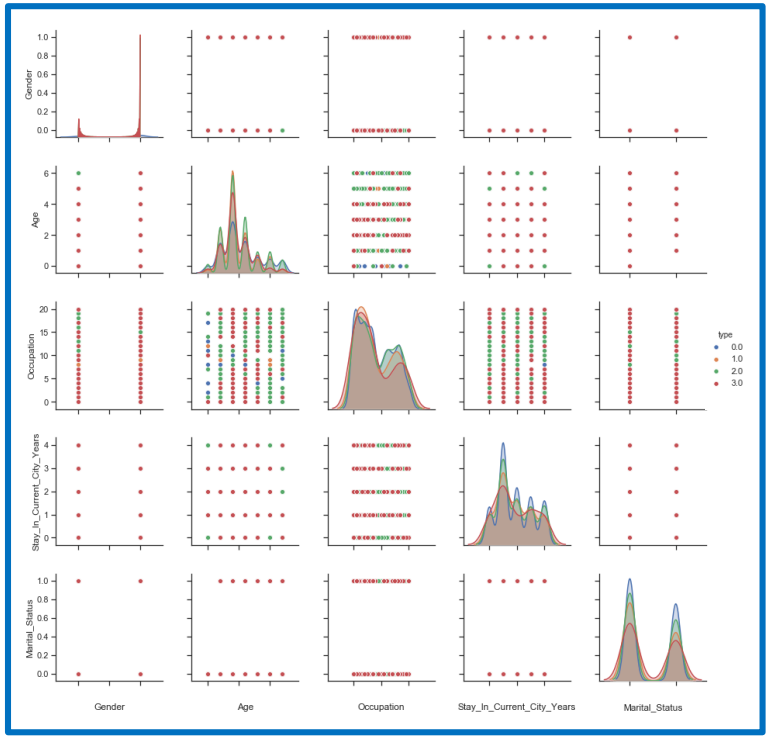
K-means algorithm enable us to automatically label the customers into 4 groups in terms of basket size and purchase amount

Type 4 (type=3) represents the customer with high purchase amount and big basket size, **which is Very Important for our business (VIP)**



JHSC Group

2. Purchase Prediction: Feature Issues



Plot graph shows no distinctions in 4 types among each pair

Customer behavior preference can help us have a more efficient prediction!

We add top 3 frequently bought categories as features

Decision Tree

Accuracy score
Training set: 0.89
Test set: 0.49
Overfitting

Random Forest

Accuracy score
Training set: 0.86
Test set: 0.53
Overfitting

SVM

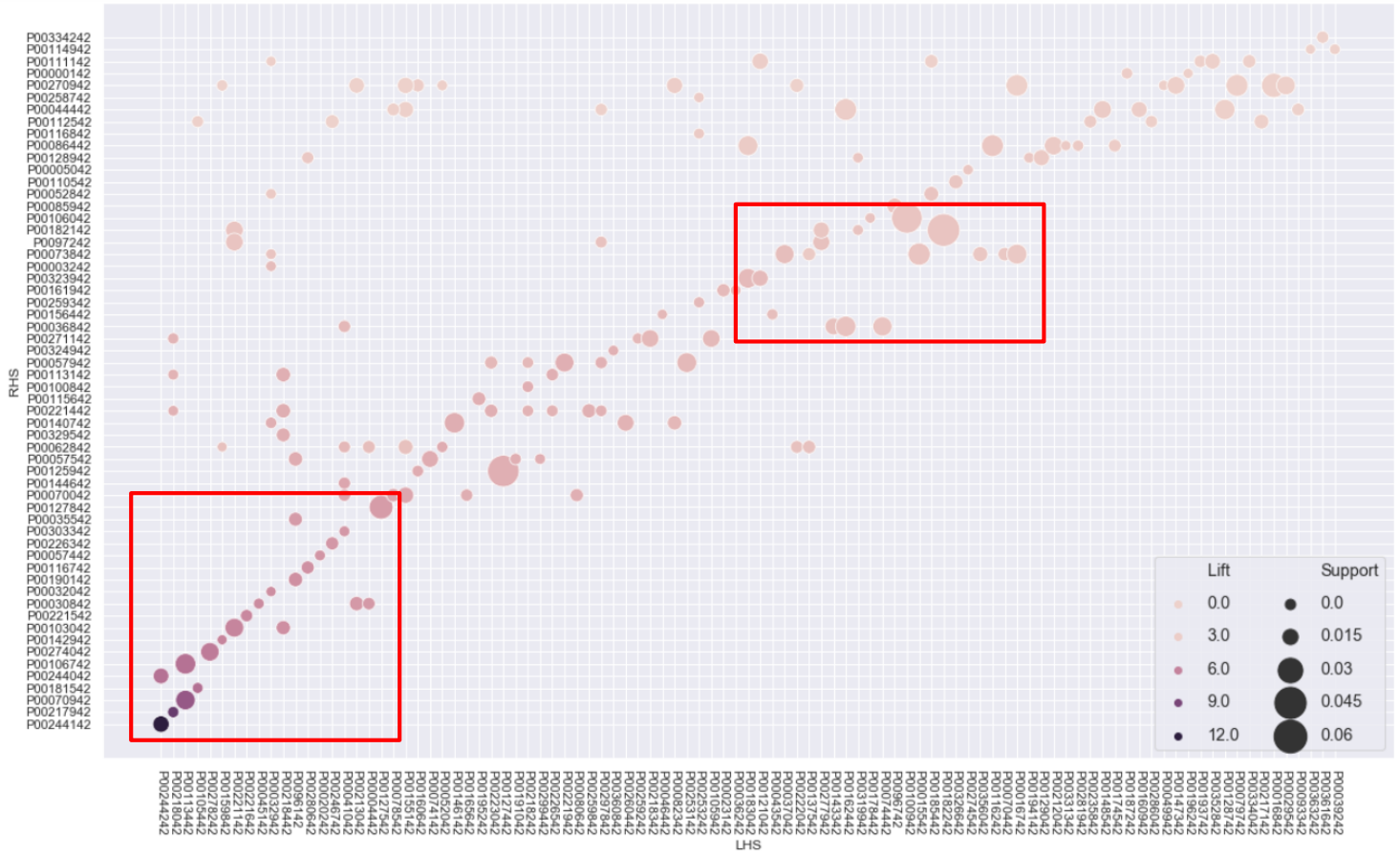
Gaussian Kernel

Accuracy score
Training set: 0.65
Test set: 0.61

JHSC Group

3. Recommendation System: Association Rule Mining

Product Rules



Apriori algorithm

For association rule X → Y

Support

$$support = \frac{(X \cup Y).count}{n}$$

Confidence

$$confidence = \frac{(X \cup Y).count}{X.count}$$

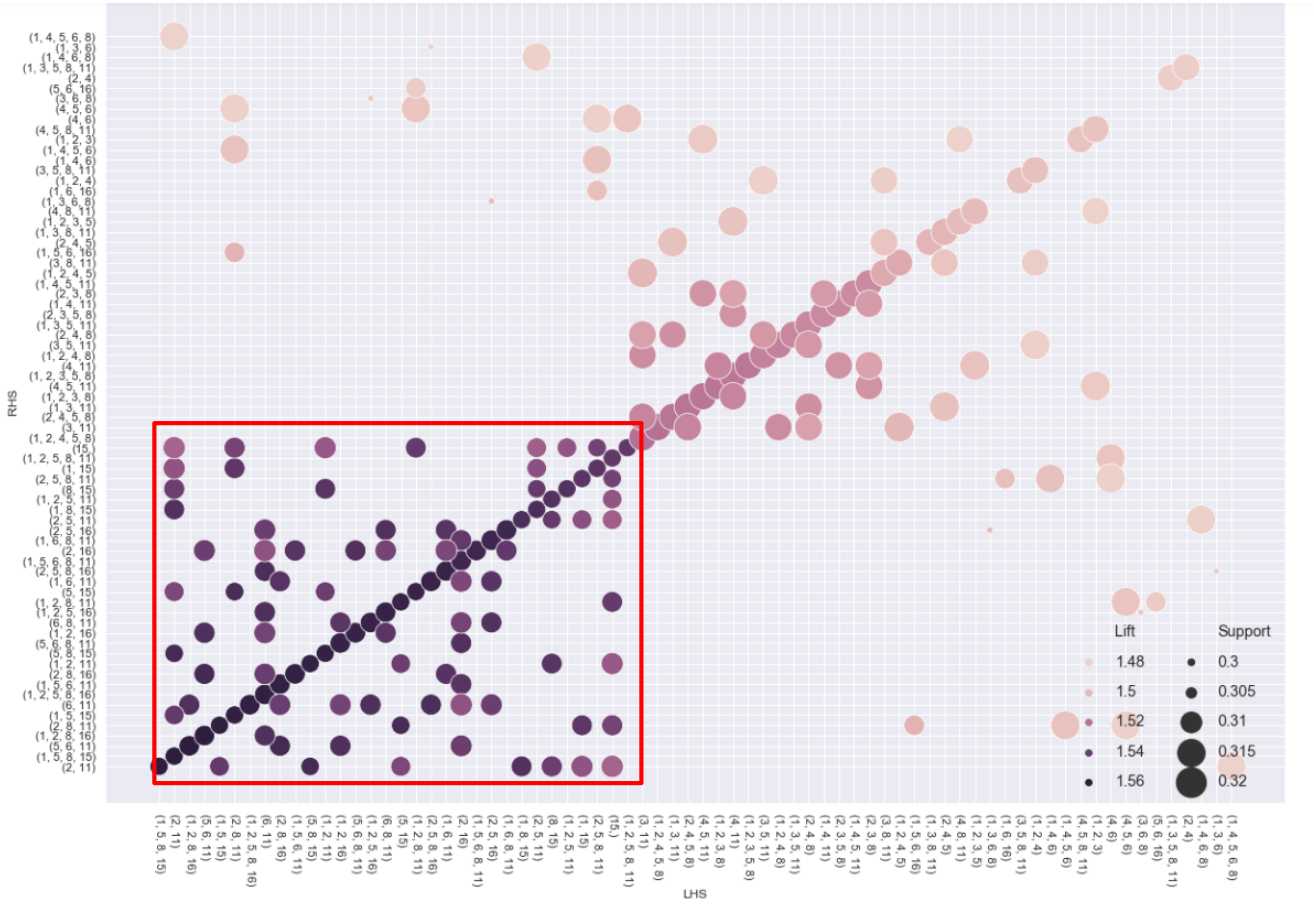
Lift

It tells us how much more likely it is to buy Y GIVEN THAT X are bought as well

JHSC Group

3. Recommendation System: Association Rule Mining

Category- multi-class rules



Product Rules

P00244242 → P00244142
P00133442 → P00070942
P00133442 → P00106742
P00278242 → P00274042
……

Category Rules

1,5,8,15 → 2,11
1,2,8,16 → 6,11
1,5,15 → 2,11
2,8,11 → 5,15
……

JHSC Group

3. Recommendation System: Personalization

This prediction model help recommend products to specific customers and boost sales.

Input

The customers purchase amount for each product

Model

Matrix decomposition model

$$\hat{r}_{i,u} = \mu + b_u + b_i + p_u^T q_i$$

Coefficients

Mean rating μ is the average rating of all users over all movies in our training set $\mathcal{T}$

$$\mu = \frac{1}{N_{\mathcal{T}}} \sum_{(i,u) \in \mathcal{T}} r_{u,i} \quad (4)$$

User Bias b_u is a *per user* bias that will be higher for users that give high average ratings to all movies.

Item Bias b_i is a *per item* bias that will be higher for the more popular (higher ranked) movies.

User Embedding $p_u \in \mathbb{R}^F$ is a *per user* F dimensional vector that maps user u into some kind of abstract *taste space*.

Item Embedding $q_i \in \mathbb{R}^F$ is a *per item* F dimensional vector that maps item i into the taste space.

Output

ranking matrix $r_{i,u}$

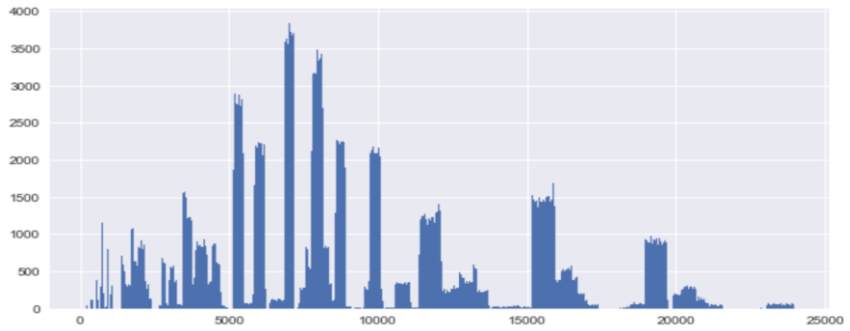
The amount that a specific customer would like to buy for a certain product given the historical records

JHSC Group

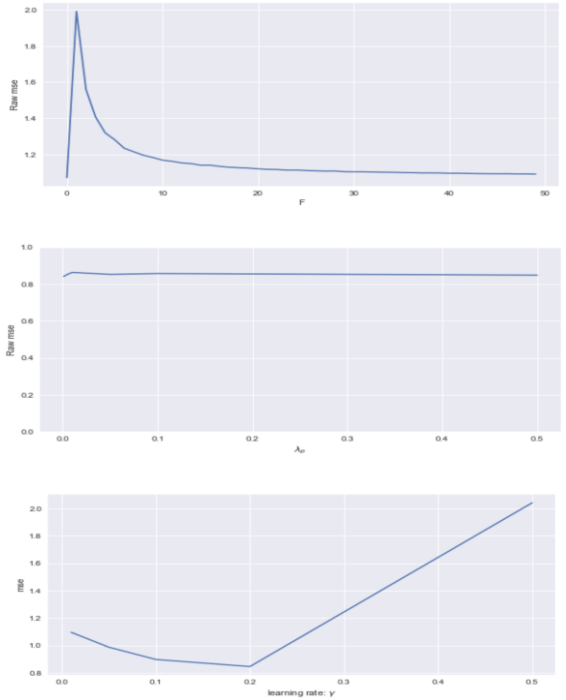
3. Recommendation System: Personalization

This prediction model help recommend products to specific customers and boost sales.

Preprocessing & Parameters Setting



- ✓ Split data into 3 sets: train, test, validation
- ✓ Group 'Purchase' label into 'Rating'



Set F =20 to tune other parameters

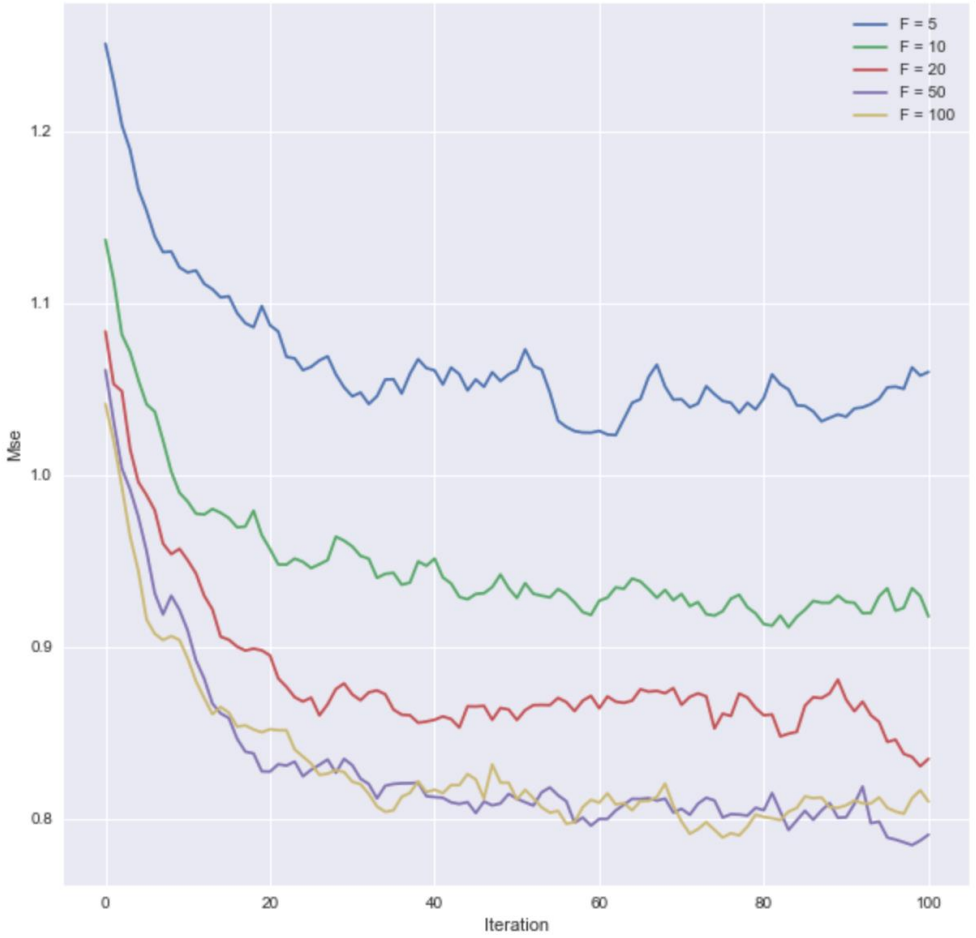
Penalty parameter has no significant influence

Set learning rate = 0.2

JHSC Group

3. Recommendation System: Personalization

This prediction model help recommend products to specific customers and boost sales.



- ✓ When F = 50, our model has the best performance
- ✓ Best MSE = 0.79

Output: The amount that a specific customer would like to buy for a certain product given the historical records.

Model Benefit:

- 5891 customers and 3623 products
- predicts every customer's willing for every product
- $5000 /person*score*product
- Capture 21% of the whole market

Around 20 billion dollars!

JHSC Group

Thanks!

Shangyou Wu, Haoran Tang, Chi Zhang, Jing Qian

JHSC Group