# Catch The Fashion Trend!
# A Fashion Bundle Analysis

Heather Liu

McGill University

# Agenda

- Data processing and limitations
- Modeling approaches
- Results
- Future plans

# Roadmap

### Data processing

☑ Oct. 2018

- ☑ Extract the item from title
- ☑ Created an item list with a 78% match rate with titles
- ☑ Extract the features based on the item list

### Approach from machine learning aspect

□ Nov. 2018 - Jan. 2019

- ☑ Split dataset into training and testing sets
- ☑ 1ed regression model is built and trained
- ☑ Built different regression architecture to evaluate the correlation between the bundle information and the 'likes' predicted for the bundle
- □ Increase the accuracy by using larger training set

### Approach from business aspect

□ Dec. 2018

- ☑ Check the existence of star product
- ☑ Market lift analysis between every pair of product
- □ Market lift analysis between every 3 products

**Note:** In the following slides, we will use "bundle" to indicate a set of wear and "likes" as a measure of the quality of the bundle.

# Data Processing

- Independent variables
  - Features (e.g. item, color)
- Dependent variables:
  - Bundle likes
- Data processing
  - Cleaned the dataset (Nah)
  - Selected the data falling into the interval of 0.05 to 0.95 quantile
  - Extract the nouns from title
  - Created an feature set with a 78% match rate with the dataset

# Modeling

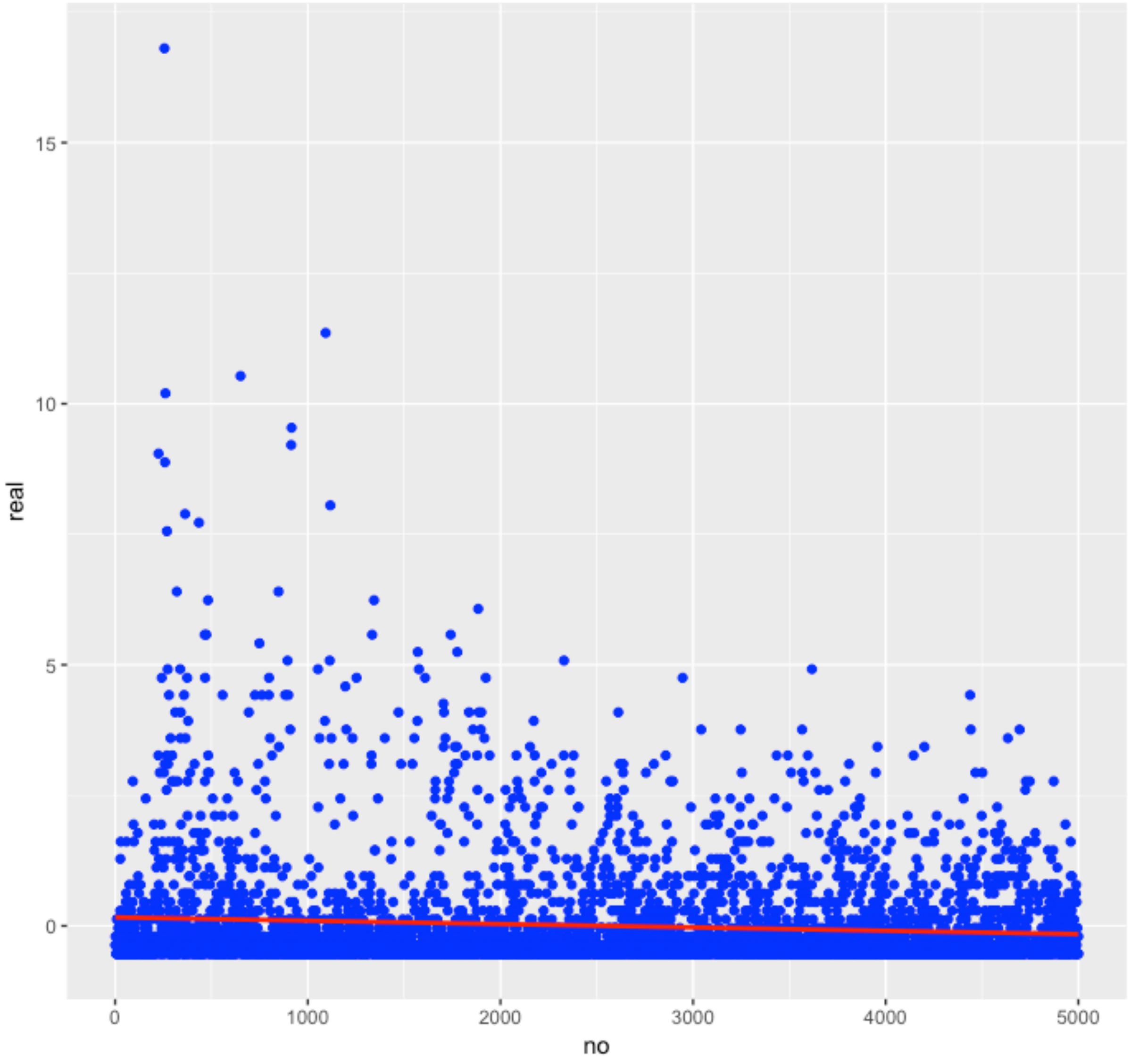
- Model 1: Regression Model (data mining aspect)
- Model 2: Xgboost Model (data mining aspect)
- Model 3: Market Lift Model (business aspect)

# Model 1: Regression Model

- What we did:
  - Separate the testing and training data
    - For every 5 units, one as testing, four as training
  - Feature Set :
    - An array at shape(1, 253) to store the type of the items in the bundle
  - Choosing the proper model/machine learning methods to handle the data: Regression
  - Set up proper standards to evaluate the model
    - Real likes VS Predicted likes
  - Improvements:
    - Optimized the model
    - Optimized the feature set
    - Adjusted the dependent variable

# Model 1: Regression Model

**Result:**



**The red line is the regression line and the blue points are the real “likes” in the bundle.**

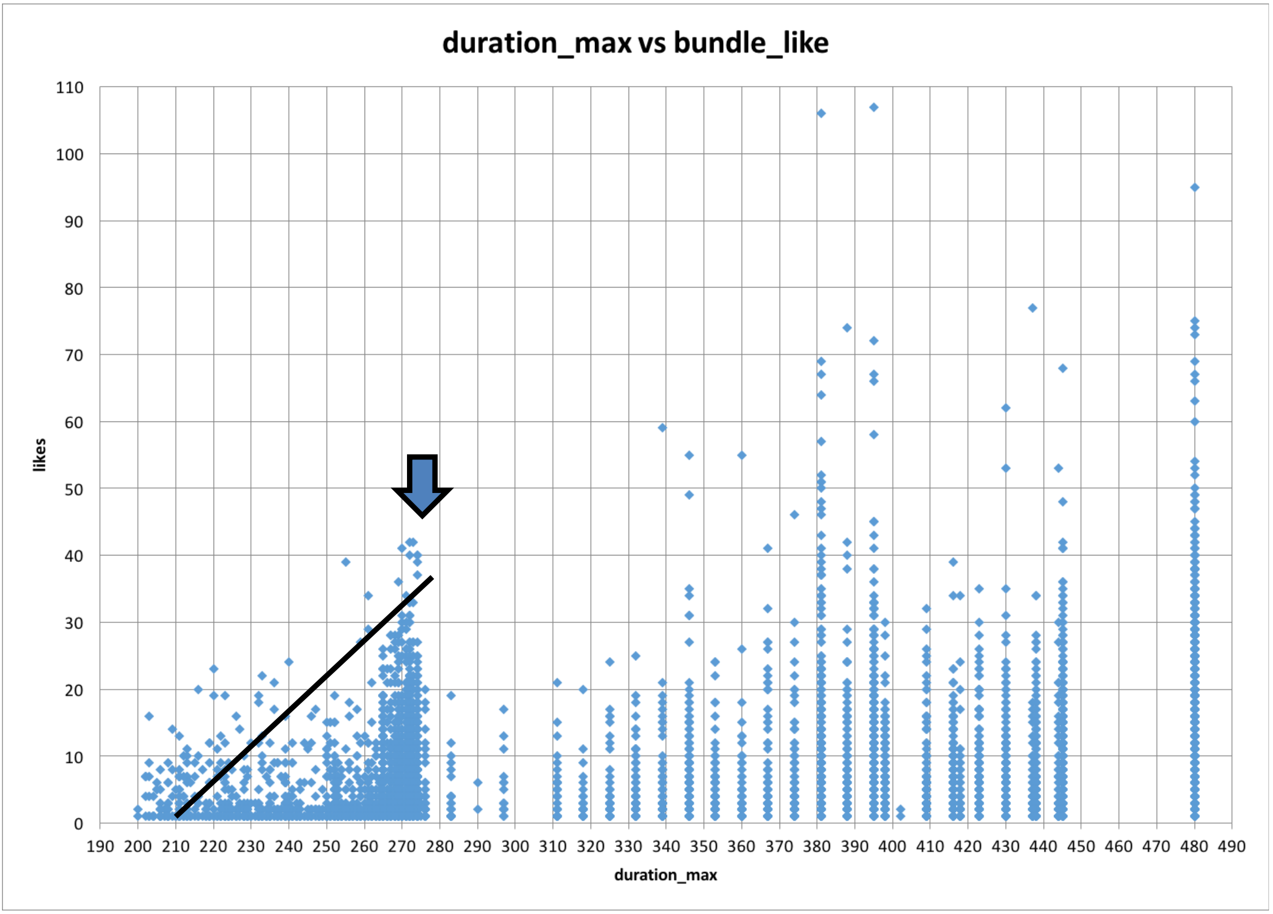
# Data Limitation

- Data Limitation:

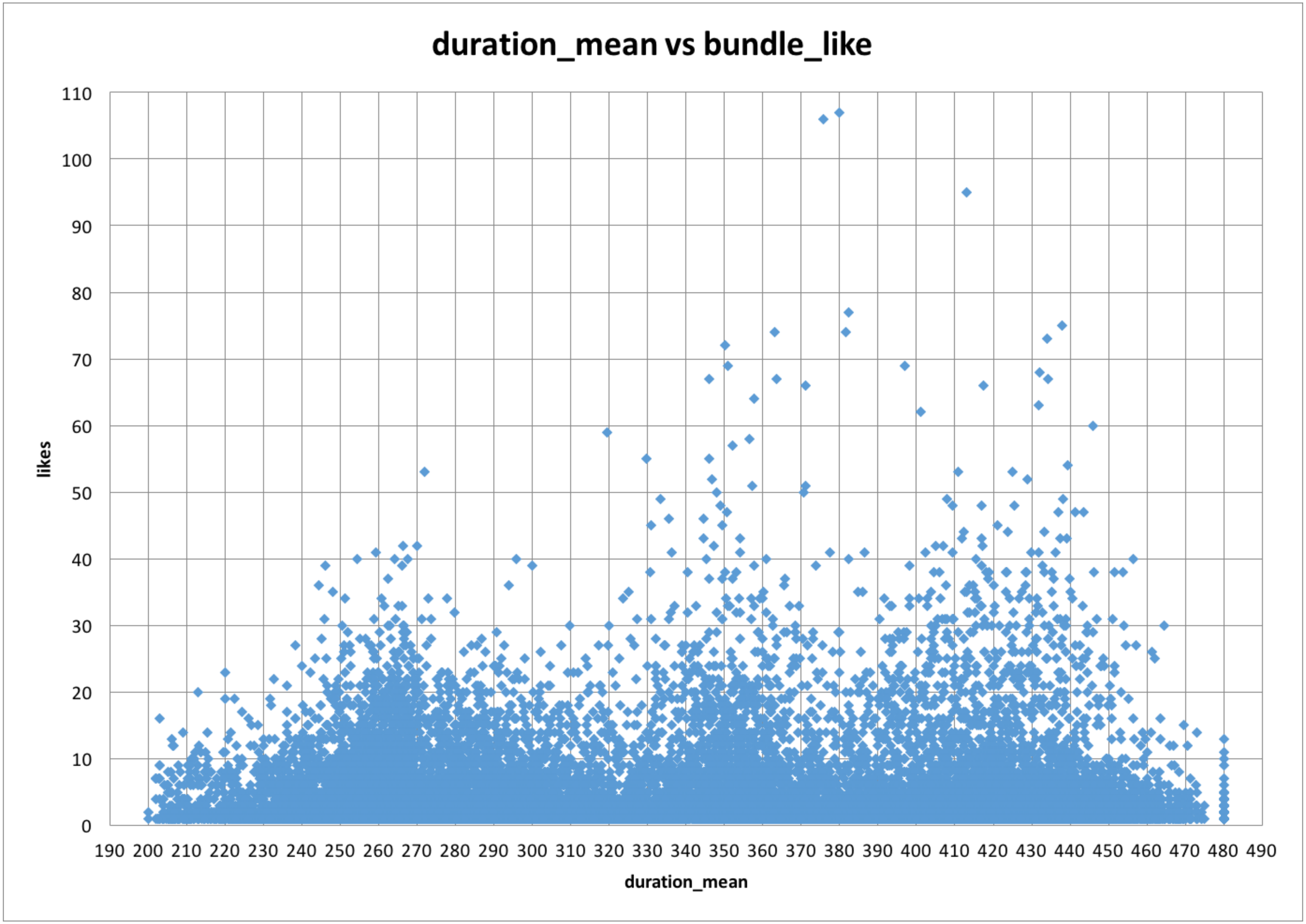
  When using different index as the dependent variable, we found the existing time of bundles has undesirable relation with the bundle like:

  – Intermittent collective appearance after a time point

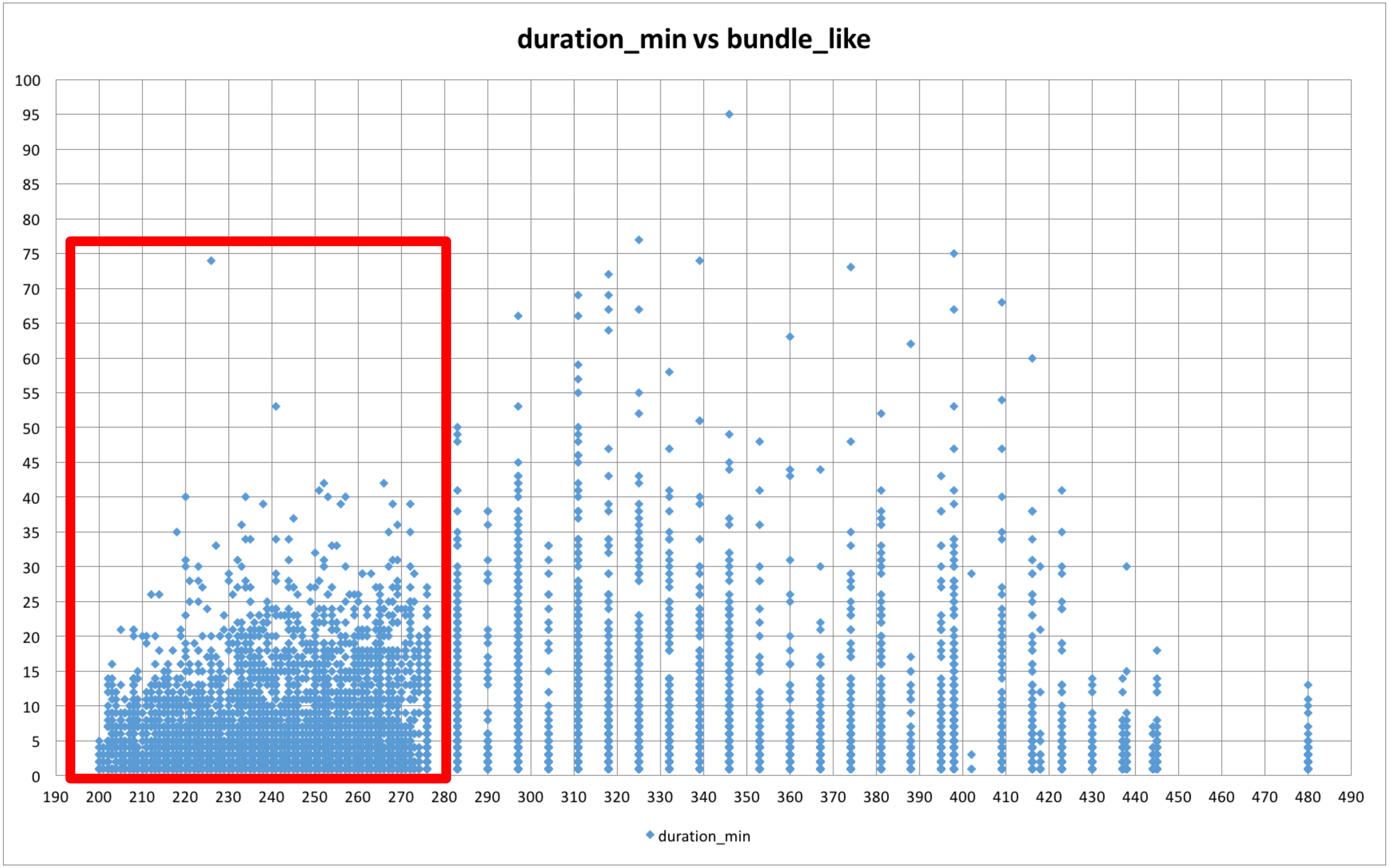
  – Positive correlation between likes and existing time



# Data Limitation



# Data Limitation



# Model 2: Xgboost Model

- Input data:
  - To avoid the affect of uncertain data, we finally choose 11008 bundles (80% training+20% testing)
- Training label:
  - The likes for this bundle normalized by the duration of existing, at the shape (1, 1)
- Model:
  - Random forest: is suitable for large size of the feature(272 numbers)
  - Support Vector Machine(SVM): is a traditional machine learning model widely used in regression
  - FNN: feedforward neural network is a basic structure widely used in regression
  - Xgboost: can do feature selection, it is more suitable for this task

# Model 2: Xgboost Model

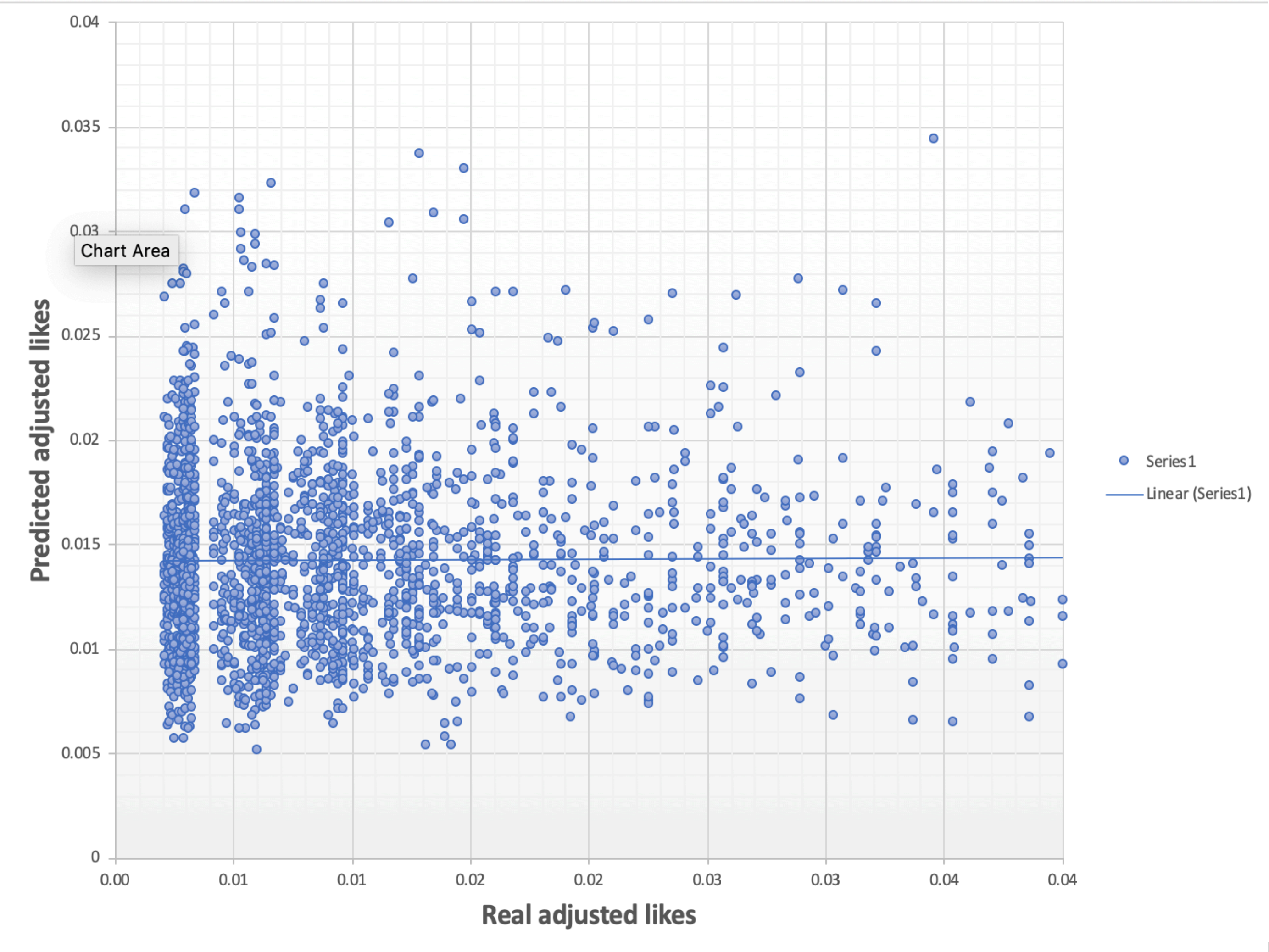
- Feature set:
  - An array at shape(1, 253) to store the type of the items in the bundle
  - An array at shape(1,10) to store the color information in the bundle
    - We selected 10 basic color: brown, black, blue, red, grey, white, gold, rose, pink, silver
  - An array at shape(1, 3) to store the sum, mean, standard deviation for the likes of every product in the bundle
  - An array at shape(1, 3) to store the sum, mean, standard deviation for the outfit count of every product in the bundle
  - An array at shape(1, 3) to store the sum, mean, standard deviation for the price of every product in the bundle
  - As a result, for every bundle, the training feature shape is (1, 272)
- How is works?
  - If one of the item/color appears in the bundle, the index of this color is 1, otherwise is 0.

# Model 2: Xgboost Model

- Optimizing:
  - For the models except FNN, we used the Gridsearch to select the optimized parameter in a wide range. For FNN, training failed( the loss and acc shows NaN. The reason would be the feature needs to be normalized due to the wide range of price, likes, etc. appeared in the dataset.
- Result:
  - Xgboost has the best performance in the testing data and obtained a pearson's correlation at about 0.28.
  - A correlation between 0 - 1 represents a positive correlation
    - It's a measure of the performance of the model(prediction vs real).

# Model 2: Xgboost Model

- Limitation: Since our training dataset is relatively small. This correlation is not enough to accurately predict likes for a bundle.
- We could expect a higher correlation with larger dataset.



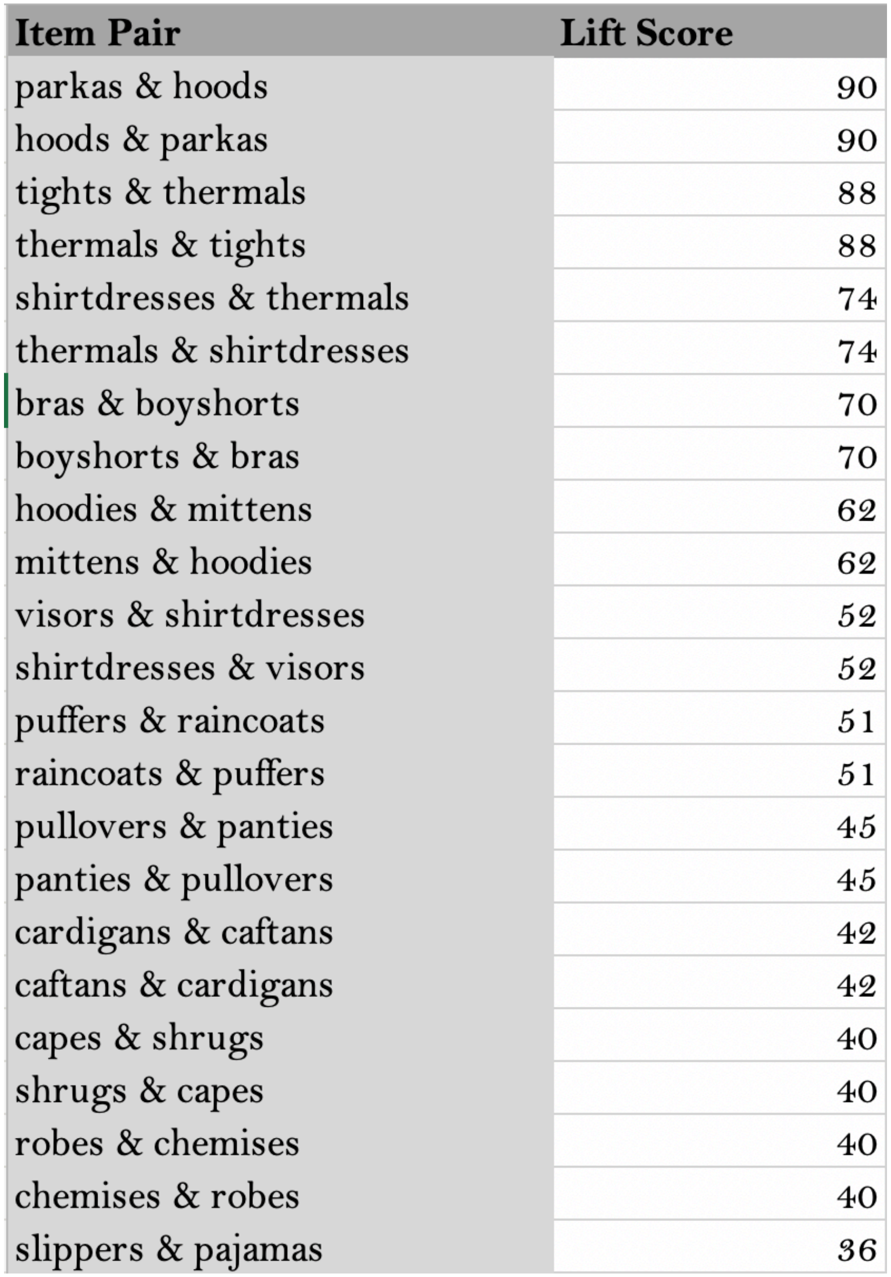
# Model 3: Market Lift Analysis

- What is a lift?
  - **In marketing, “lift”** represents an increase in sales in response to some form of advertising or promotion.
  - **In our experiment, “lift”** represents an increase in the existence of product A in response to the existence of product B when A and B are in the same bundle.
- Why this analysis?
  - Want to know whether there is a coexistence between items.
- What we did?
  - Tested the correlation between every two items
  - Computed the lift score between every pair of items

# Model 3: Market Lift Analysis

**Results:**

There is a higher coexistence probability between two more correlated products.



| Item Pair | Lift Score |
|---|---|
| parkas & hoods | 90 |
| hoods & parkas | 90 |
| tights & thermals | 88 |
| thermals & tights | 88 |
| shirtdresses & thermals | 74 |
| thermals & shirtdresses | 74 |
| bras & boyshorts | 70 |
| boyshorts & bras | 70 |
| hoodies & mittens | 62 |
| mittens & hoodies | 62 |
| visors & shirtdresses | 52 |
| shirtdresses & visors | 52 |
| puffers & raincoats | 51 |
| raincoats & puffers | 51 |
| pullovers & panties | 45 |
| panties & pullovers | 45 |
| cardigans & caftans | 42 |
| caftans & cardigans | 42 |
| capes & shrugs | 40 |
| shrugs & capes | 40 |
| robes & chemises | 40 |
| chemises & robes | 40 |
| slippers & pajamas | 36 |

# Model 3: Market Lift Analysis

- Star product
  - **Definition:** A product will increase the level of likes of bundles when it exists.
    - Ex:
      - Must-have items
      - New release
      - Star style
  - **Idea:** Trying to detect a “star product” effect
    - If we could extract  all the star product, we could have a better understanding of the relation between bundle and likes, thereby to build bundles with high likes by adding or subtracting star products.
  - **Test:** Brands VS Likes, Seasonal trends VS Likes, Prices VS Likes
  - **Limitations:** We need a larger sample set to build criterion to discover these products.

# Future Plan

- Data:
  - Expand the feature set with more key words such as photos, brands, size, etc.
  - Collect cleaner data used for training, or acquire larger dataset
- Idea and method:
  - Apply deep learning model on images
  - Explore the sequential relation between items added to the bundle
- Technique:
  - Use more computational resources such as high performance computing to work on larger dataset

Thank you!