Business Understanding/Data Understanding:

/\*

What problems you specifically addressed, including details in technical as well as business terms.

What is the domain and what are the potential benefits to be derived from data mining.

How will your data client benefit from your analysis?  
- attribute and interpretation

\*/

**1. Business Understanding**

**1.1 Background**

H&M is a fast fashion brand that was first established in Sweden. The brand offers the latest style and the best prices to customers. H&M consists of inexpensive clothing necessities, perfect appearance accessories, and sportswear for women, men, teenagers, and children. It also includes living items from H&M Home. H&M operates 4420 stores, 72 store markets, and 49 online markets worldwide. H&M was considered the second most valuable clothing brand in the world after Nike in 2018.

**1.2 Assess Situation**

Every day, H&M handles large amounts of orders from its multiple sales channels. And the volume is tremendous.

It is essential for it to predict the items that the customers want so that it can have good control of its supply chain management. You don’t want to produce the inventories that keep staying in stores or warehouses and these inventories become outdated or deadstock. That will be a huge cost and lose the customers.

For that huge number of customers and items, to improve the profit, H&M needs to develop a plan to control its operation cost and maintain its customers.

**1.3 Models Objectives**

Here, we want to help H&M to develop the models to identify its valuable customers and predict the most likely items that these customers will buy in the near future.

**1.4 Produce Project Plan**

We will develop two types of models to help H&M to improve its operation and customers management.

**1.4.1 Classification Models**

First, a classification model will be built to help H&M to identify the different classes of its customers.

We will use the previous year’s transactions history, to get these attributes such as age, buying channels, and different amounts of items a person bought to see the customers’ features and behaviors. After that, we will build a classification model to predict the class of this customer in the following year. Generally speaking, we are going to use the attributes of a customer’s characteristics and buying behaviors to predict the class of this customer in the future.

Then we will get a list to see the customers with different priorities for H&M.

As we know the long-tailed theory. 20% of the customers contribute to 80% of its total sales. By using the classification model, H&M can get a list of its valuable customers.

H&M can use this list to do customers analysis. It can analyze some attributes of these customers, such as age, buying channel, and so on. By analyzing these data, H&M can have a good knowledge of its current business and find some useful information for its future strategy. For example, they can measure the buying channels to see the favorable channels of the customers so it could have the right marketing budget for different channels. They can also find its main age group and design specific clothes to maintain and attract more customers. Moreover, it can also know the performance of different categories of the items. Some items, which are popular, should be ordered or produced more. Some items, whose sales are pale, should be considered removed from the product catalog.

**1.4.2 Association Models**

Secondly, after knowing the list of these valuable customers, H&M can know the behaviors and make a prediction of the possible items these customers will buy in the near future. Association Model can help H&M discover the bundle sales of items for its valuable group. Thus, H&M can have a more efficient and light approach of inventory management because, if the prediction is good, the items will be sold out in a short while and H&M’s holding cost will be reduced. (Oliver跟Shengyi的部分，可能可以再补充一点？)

（Shengyi 的补充）:

By applying association rules to the model, we can analyze customers’ past transaction data and give 10 recommendations for each customer by sorting pairs of items that are bought together from all the customers in HM.

Data Understanding and Preparation:

/\*

Present your understanding of the data including quality and other issues.

– missing values

* Outliers
* Weird values

What is in the data, and what preprocessing was done to make it amenable for the visualization and analysis.

What kinds of data analysis or data mining techniques are suitable for this data? Where choices were made (e.g., parameter settings for discretization, or decisions to ignore an attribute), describe your reasoning behind the choices and the process you followed. \*/

**2. Data Understanding**

H&M provides us with 3 datasets, the articles.csv, the customers.csv, and the transaction.csv.

(<https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data>)

**The Articles Dataset**

The articles Dataset includes all the items that H&M has currently. It consists of 105, 542 articles with 25 attributes.

图片包含 应用程序

描述已自动生成

表格

描述已自动生成 手机屏幕截图

描述已自动生成

Besides some common attributes such as **article\_id**, **prod\_name** and **product\_code**. The dataset also provides us with some useful attributes such as the color, the perceived color, and the description of the article.

Among these attributes, most of them are types or groups. These groups are divided based on H&M’s different categorizing methodology. They can be categorized based on types, departments, sections, garment groups, and so on.

For example, **product\_type\_no** and **product\_type\_name** are two attributes that H&M uses to categorize the articles. H&M put the articles into different types and assign them a type series number.

图表, 条形图

描述已自动生成

For another attribute, **index\_name**, is also similar.

图表, 条形图

描述已自动生成

**The Customers Dataset**

The Customers Dataset includes all the customer information. It consists of 1,371,980 customers with 7 attributes.

图片包含 图形用户界面

描述已自动生成

文本

描述已自动生成 文本

描述已自动生成

Also, there are some missing values for some attributes.

文本

描述已自动生成

**FN** and **Active**, due to their bad data quality with many missing values, we decide not to choose and use in our models.

For **club\_member\_status**, this is an attribute to show the status of the customers. It has 3 unique values, ACTIVE, LEFT CLUB, and PRE-CREATED. We can see most of these customers are active members and a few of them are at other statuses. We believe this attribute has little effect on the model, so we decide to drop it.

图表, 瀑布图

描述已自动生成

The **fashion\_news\_frequency** describe the frequency of the customers receiving the fashion news. We can see the distribution as the following:

图表, 瀑布图

描述已自动生成

Moreover, the dataset also provides the age of the customers. Most of these customers are between 20 and 30 years old. The maximum age could reach over 80. The median age is 32 years old.

图表, 直方图

描述已自动生成

**The Transaction Dataset**

The Transaction Dataset includes the transaction history from 2018-09-20 to 2020-09-20. It consists of 31,788,324 transactions with 5 attributes. Each instance represents a transaction of a customer for a single article.

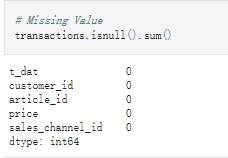
文本

描述已自动生成

图形用户界面, 文本

描述已自动生成

There is no missing value for this dataset, which is good.



图表, 直方图

描述已自动生成

We can see some patterns in the line chart. We have a huge volume of transactions happening in October 2019.

图表, 条形图

描述已自动生成

If we just look at the monthly distribution, most transactions occurred in June and July and few in winter and spring.

图表, 条形图

描述已自动生成

We have the highest volume of transactions on Saturday. For Monday, Tuesday and Sunday, the sales are lower. We can arrange our employee schedule based on this. Also, for Sunday, the management team should go to discover the reason why Sunday, a non-working day, has this few numbers of transactions.

**3. Data Preparation**

**3.1 Divide the Transaction into 2 period**

We are going to build a model to predict the future importance of a customer. If we want to use many X to predict Y. We need to collect our X, the customers’ buying behavior, and feature attributes from last year. We get the Y, our target variable, from the data for next year. Thus, first, we need to divide the transaction dataset into two time periods.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

We split the transaction in the point of 2019-9-20.

**3.2 Merge the Transactions and Articles Dataset**

We are going to combine the transaction data with articles data. And we only want to keep the attributes that are useful for our models.

There are many attributes describing the category of the article, eventually, we pick the most distinct and meaningful attributes as shown below.

图片包含 表格

描述已自动生成

**3.3 Modify the fashinon\_news\_frequency NONE value**

To improve the classification models’ performance, we decide to turn the string description into numbers. 0 represents NONE, 1 represents Monthly, and 2 represents the other circumstances.

图片包含 文本

描述已自动生成

**3.4 Drop useless attributes**

Some attributes, because they have many null values and some of them are hard to analyze and interpret, such as the encrypted postal code. We decided to drop them.

图形用户界面, 文本, 应用程序

描述已自动生成

After dropping, the Customers Dataset is shown below:

表格

中度可信度描述已自动生成

**3.5 Fill null value in Age with the median**

Age has many null values, we decided to fill it with the median.

图片包含 图示

描述已自动生成

**3.6 Merge Customers Dataset with the merged Transactions\_Artibles Dataset**

图形用户界面

低可信度描述已自动生成

**3.7 Turn nominal attributes into dummy variables.**

As mentioned above, we have some attributes about the type and the color and some other nominal attributes in the current dataset. We decide to make them into dummy variables for our next step.

图形用户界面, 文本, 应用程序

描述已自动生成

文本

描述已自动生成

**3.8 Select the customers with more than 5 transactions**

In order to get meaningful insight, we decided to concentrate on the customers who had more than 5 transactions. We don’t want some outlier or noise in our dataset.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

**3.9 Group the dataset**

In order to calculate the difference and preference of each customer, we can sum up the 1 in a column. If a customer often buys a certain type or a certain color of a product, we will know.

图形用户界面, 文本

中度可信度描述已自动生成

Group some attributes that do not need to sum up.

文本

描述已自动生成

Group some attributes that need to sum up.

图形用户界面, 文本

中度可信度描述已自动生成

表格

描述已自动生成

Combine the sum and the non-sum-up data frame.

表格

中度可信度描述已自动生成

**3.10 Create the target variable based on total consumption.**

We are going to create our target variable, the importance of a customer based on his/her next year’s total consumption at H&M.

First, we sum the price of all the transactions for one customers.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Then, we could see the total consumption (the sum of price in 2020) distribution.

图形用户界面, 文本, 应用程序

描述已自动生成



We decide to divide them into 3 groups by using the 25% and the 75% percentile as our split points. Therefore, the top 25% of customers will be labeled as 2 - the valuable customer. The last 25% will be labeled as 0 – less valuable, and the rest will be 1 – normal customer,

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Finally, we merge this column into our previous dataset.

表格

描述已自动生成

**4. Data mining and Evaluation**

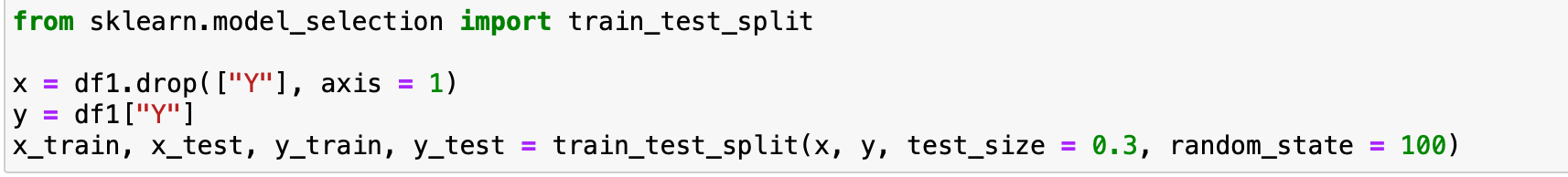
**4.1 Classification Modeling**

After the data cleaning and exploration, we have finalized the dataset and firstly we decide to use the classification modeling to find the most and least valuable customers. In this dataset, we have used the data of 2020 to classify our customers into 3 groups in which the value 2 corresponds to the most valuable customers and the value 2 corresponds to the least valuable ones. And we use the data of 2019 to do the classification modeling.

Considering that we have a large number of independent variables and a huge dataset, we decide not to use the support vector machine and KNN algorithms. Instead, we decide to try 3 classification algorithms: naïve bayes, decision tree and random forest, and select the best result one for our client. And before coding the algorithms in Python, we decide to drop the customer\_id column out of the independent variables and preprocess both the dependent and independent variables. Then we split the data into 70% train and 30% test sets.

表格

描述已自动生成



* + 1. **Gaussian Naive Bayes**

We first try the Gaussian Naïve Bayes model using scikit-learn package. After fitting the model and generating the prediction, we then calculate its accuracy and generate the confusion matrix for the model.

文本

描述已自动生成

文本

描述已自动生成

图表, 树状图

描述已自动生成

As the result showing above, we only get around 31% accuracy for the gaussion naïve bayes models.

* + 1. **Decision Tree**

Next, we try the decision tree model. After fitting the model and generating the prediction, we then calculate its accuracy and generate the confusion matrix for the model.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

As the result showing above, we get a better accuracy than naïve bayes model which is around 51%.

图表, 树状图

描述已自动生成

* + 1. **Random Forest**

Finally, we try the random forest model. We find the max\_depth = 5 yields to the great result. Like always, after fitting the model and generating the prediction, we then calculate its accuracy and generate the confusion matrix for the model.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图表, 条形图

描述已自动生成

It turns out that the random forest model performs the best among the 3 models. It has around 63% overall accuracy.

Because the main goal of the classification modeling is to help the client find the most valuable customers, compared to the rest 2 groups, the value 2 group is the “important” group that we care about most.

To focus on the value 2 group, we deeply analyze the confusion matrices of the 3 models.

**For Gaussian Naïve Bayes:**

Sensitivity = 22978 / (26876 + 12553 +22978) = 37%

Specificity = (8732+15533) / (15533+1447+1202+56536+8732+7284) = 27%

**For decision tree model:**

Sensitivity = 35326 / (4276 + 22805 + 35326) = 57%

Specificity = (39577+3279) / (3279+10592+4311+10359+39577+22616) = 47%

**For random forest model:**

Sensitivity = 36127 / (36127+26280) = 58%

Specificity = 60296 / (16762 + 60296 + 1420 + 12256) = 67%

In this perspective, we find it that decision tree model and the random forest has the similar sensitivity to detect the important group members. But the random forest model is more likely to rule out the other 2 group members.

Another aspect for model evaluation is to check how fast the model could capture the essential information, in our case is to find the most valuable customers. We decide to use the lift chart to compare among the 3 models.

图表, 折线图

描述已自动生成**For Gaussian Naïve Bayes:**

**For decision tree model:**

图表, 折线图

描述已自动生成

**For random forest model:**

图片包含 图表

描述已自动生成

Based on the results of lift charts, we find it that the random forest model is the one capture the most valuable customers quickest.

To conclude, after the model evaluation, we decide to choose the random forest model to classify the customers to find the most valuable ones.

**4.2 Olivia’s part**

**4.3 Association Rules**

We applied association rules to the dataset and give recommendations for customers.

**4.3.1 Reason for applying association rules**

By exploring the data, we found that many customers bought more than 1 articles at a time and we believe that there may be some relations between articles.

The simplest example is the association between a holiday dress and a straw hat. We may find that when a customer buys a holiday dress, he/she is likely to buy a straw hat through the purchase records of a large number of customers in the past. Therefore, it would be a meaningful recommendation to recommend a straw hat to a customer who just bought a white holiday dress recently.

By applying association rules, we can figure out the relation between articles and recommend articles to customers that they are most likely to buy.

**4.3.2 Choice of parameters**

Customer id and article id are the main parameters to analyze the problem.

One traditional case for the association rule is in 1990s, Walmart found that beer and diapers, two completely insignificant items, have a high probability of being purchased together. Therefore, to analyze the problem, we cannot presuppose possible outcomes from experience, sometimes unexpected results occur.

In this case, we choose customer id to help filter out some low-quality users and only keep the customers who have high potentials to buy articles in HM. Articles ids are grouped by customer id and date parameters to find out what articles are bought together by a customer at a time.

**4.3.3 Accuracy of the models**

Analyze the problem from the total recommended articles side: we give 542155 recommendations in total. Among them, 9669 are correct, 532486 are incorrect. The accuracy rate is 1.8%.

Analyze the problem from the customer side: we give recommendations for 72515 people, among them, 7228 have bought at least one articles from the recommendations we gave them, the accuracy rate is 10.0%.

For a recommender system, we have a considerable prediction accuracy by applying association rule.

**5. Analysis of Models:**

**5.1 Classification Models**

Based on the modeling results, we finally decide to use the decision tree model to classify our customers into 3 groups with different purchasing powers. The results of the model indicate that around 60% customers follow the same pattern on spending money for H&M products.

**图表, 条形图

描述已自动生成**

And around 40% customers are classified as the most valuable customers (labeled with value 2). For the group of the most valuable customers, H&M decision makers may consider make the marketing team to push ads of top gear products, send new products emails in the first places and may design customized products or gifts for certain shopping seasons. For those customers, it’s important to keep their loyalty and generate more profits from them. On the other hand, for those one-time or new customers (labeled with value 0), they should consider attracting them to place at least a new order through sending promotion codes for instance.

And another thought is to expand our modeling by make models to recommend the “right” products for “right” customers.

**5.2 Olivia’s part**

**5.3 Association Rules**

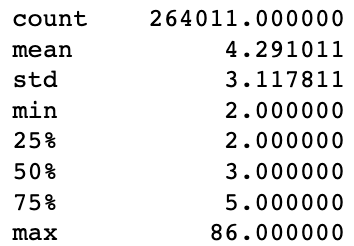
**5.3.1 General description**

We use transactions data in May 2020 as the training data to figure out the association rules between pairs of articles. By sorting the confidence score by pairs of articles, recommendations are given for the customers. The testing dataset are the people who have transaction history in May 2020 and in either the following 3 months (June, July, August 2020). People who have low intends tag in HM are also filtered out.

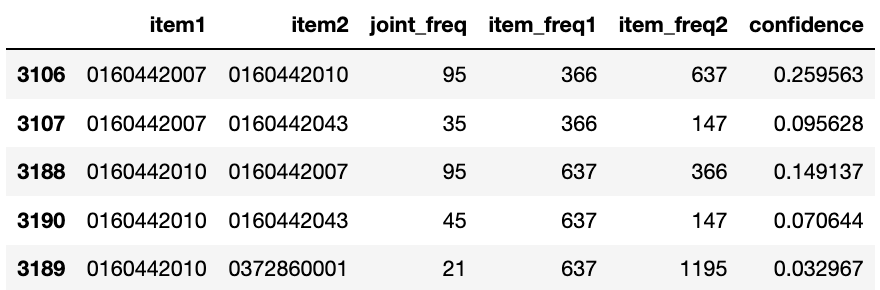
**5.3.2 Results in detail**

Use data from “transactions\_train.csv” where date range from 2020-05-01 to 2020-05-31 as training data for the association rules.

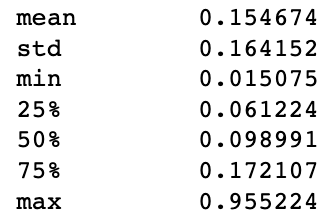
Extract all transactions where people bought at least two items a time as the original table. Descriptive statistics for origin table：



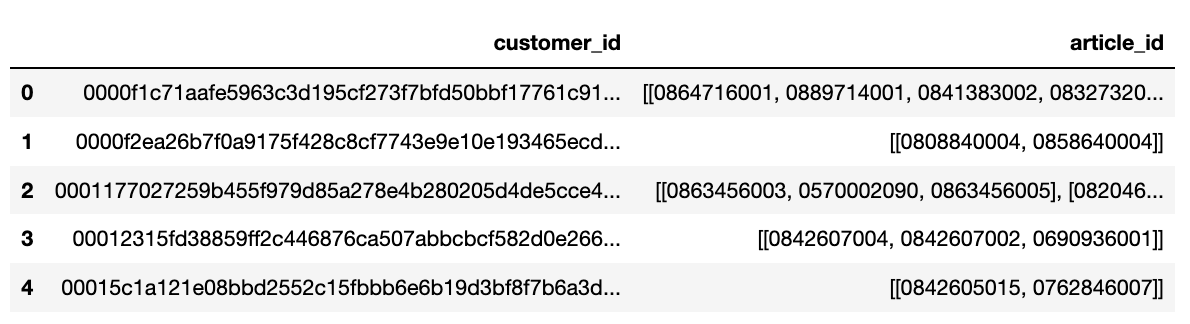
Construct a frequency table for pairs of articles bought in May 2020. Calculate the joint frequency and confidence to evaluate the association between articles. The frequency table is as following:



Descriptive statistics for confidence in origin table：



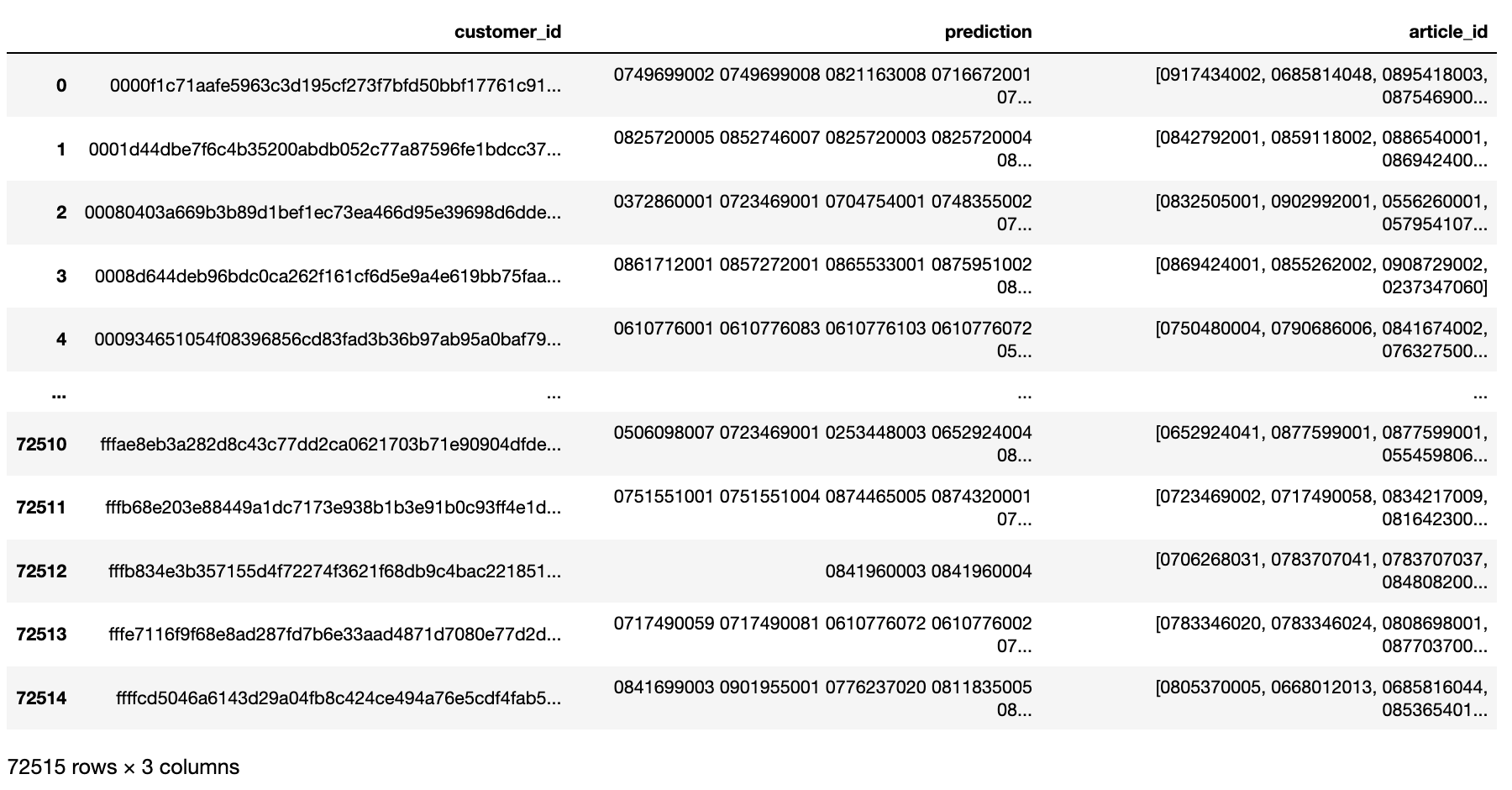
For the prediction part, we first construct a transaction table where date range from 2020-06-01 to 2020-08-31.



Then we inner join the transaction table with original table to get the customers who bought article in May 2020 and also in either the following 3 months of 2020. We also use the tags we get when clustering the people to filter out the low potential customers. Thus, we get a new customer list for prediction.

For these customers, we get their articles bought in May and extract all the association articles from the frequency table. In this step, we limit the association articles by setting the joint frequency > 20 (items1 and items2 are bought more than 20 times together by a same customer in May 2020). Finally, we sort the related articles by sorting confidence value from the highest to the lowest. For each customer, we give at most 10 recommendations.

We then get a prediction table as following:



For each customer, we give at most 10 predictions and extract what they actually buy in June, July, August 2020 for comparison (article\_id in the table). There are 72515 people in total that we give recommendations.

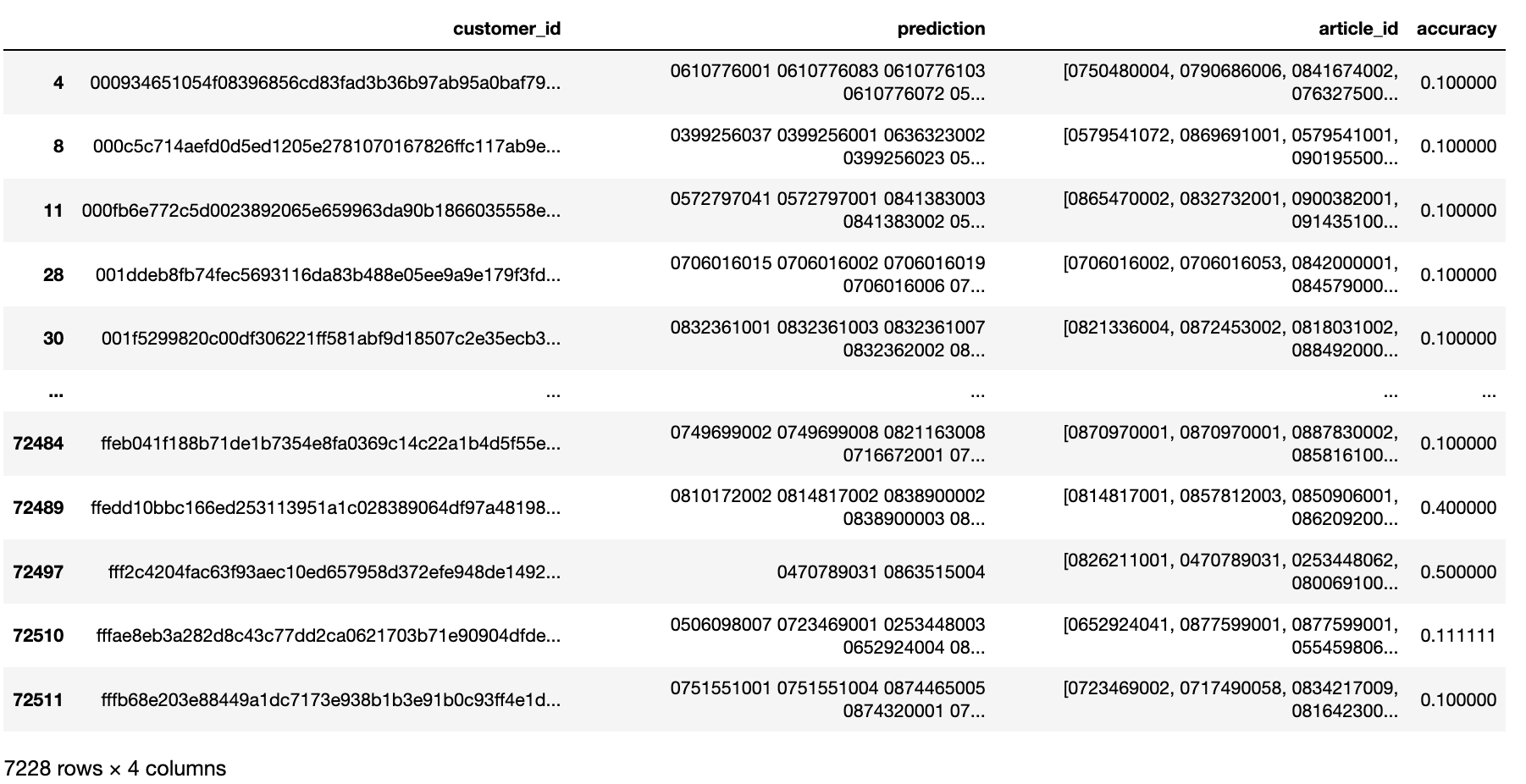
To evaluate the model, we calculate the accuracy of each predictions we give. In total, we give 542155 predictions and 1.8% of them are correct.

We also calculate the accuracy for each customers:

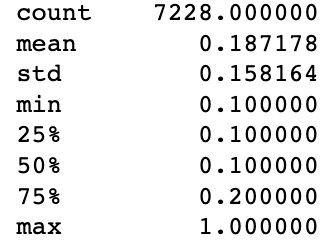


For all the 72515 customers whom we give recommendations, 7228 (10%) of them bought at least 1 article we recommended to them.

Construct a predict correct table for all the customers whom we have at least 1 correct predictions.



Descriptive statistics for accuracy in predict correct table:



**5.3.3 Business meaning of the result**

The business meaning behind the result is if we predict 10 items to each high-potential customers, around 10% of the customers will actually buy at least 1 items from the recommendations, which is a really high prediction accuracy rate. The result can give HM the confidence that reasonable personalized recommendations are useful Therefore, in reality, if we actually push the 10 items’ product information to the customers by front page or advertisements, we may probably get a more than 10% accuracy rate. In this case, HM customers can benefit from reducing searching time to get what they want and have a better customer experience. HM can also make front pages meaningful and earn profits from directly giving good recommendations to customer.

**6. Recommendations:**

For the group of the most valuable customers, H&M decision makers may consider make the marketing team to push ads of top gear products, send new products emails in the first places and may design customized products or gifts for certain shopping seasons. For those customers, it’s important to keep their loyalty and generate more profits from them. On the other hand, for those one-time or new customers (labeled with value 0), they should consider attracting them to place at least a new order through sending promotion codes for instance.

Based on the result of the association rule applied to the data, we found that giving 10 predictions for each high-value customers of HM are meaningful. Therefore, for our client – HM, we recommend them to dynamically scrolling every week or month and recommend products to customers based on their historical transaction records.

The recommended methods may include：1. Pushing the predicted product on the homepage；2. Using the predicted product as an opening advertisement on the website.

At the same time, by analyzing the most related products, HM can also use business strategies such as bundling sales to promote customer purchases (For example, if customers purchase certain types of goods at the same time, they can have discounts).