6105 Proposal

Author:

Qiuchi Chen : 001448400
Jing Ren : 001447030

1. TOPIC

Store Management

2. KEYWORDS

TOP 10 Popular Products Prediction
Rush Hour
Busy Day

3. GROUP MEMBER

Hansen Guo	001828028
Xiangyu Liu	001498478
Qiuchi Chen	001448400
Jing Ren	001447030
Yunyi Wu	001490936

4. INTRODUCTION

With the online shopping rising, retailing is severely impacted. Layoffs and reasonable purchase have become the top priority options for river closure. In this program, we aim to find the best schema of layoff and purchase by using machine learning to create the optimal model for predicting future trendy with related history dataset.

5. GOALS

As the trendy of benefit maximization, store management should be optimized with machine learning. We choose history data information about Orders, Products, Hour Of a Day, The Day of a Week and so forth as the dataset, to analyze and predict two things to help reducing the cost of managing a store.

First, we will predict the top 10 most popular products in order to have the store purchasing and selling more targeted.

Second, according to dataset we can predict in one week which day is the most busy and which day order number is the least. Accordingly, store supervisor would like to arrange schedule more reasonable according to this prediction.

6. METHODOLOGY

Data Processing	Algorithms
Multiple table merging	Linear regression
Data cleaning & data munging	Decision tree
Features selection & Dimension reduction	

7. ALGORITHMS

I Linear Regression:

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

II Decision tree:

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. Decision Tree models are created using 2 steps: Induction and Pruning.

8. SPECIFICATION/DESCRIPTION

Our dataset contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.

There are six tables in our dataset. Their names are “aisles”, “departments”, “order_products_prior”, “order_products_train”, “orders”, “products”.

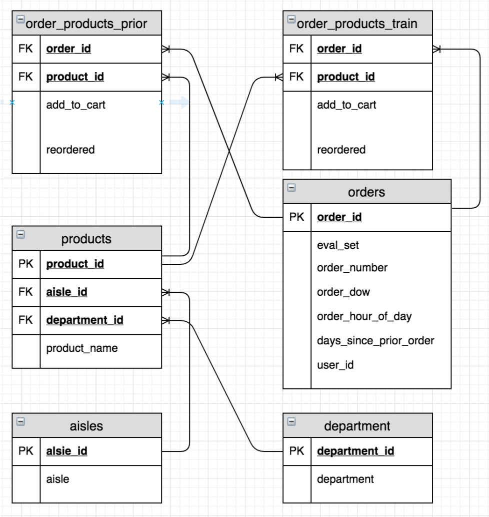
There are two columns in the “department” table, they are department_id and department. This table contains primary categories of products. Every department name uniquely correspond to a department_id number.

Similar to “departments” table, there are two columns in “aisles” table, they are aisle_id and aisle. This table contains the secondary categories and every secondary categories uniquely correspond to an aisle_id number. Primary key of the table is aisle_id.

There are seven columns in the “orders” table. They are order_id, user_id, eval_set, order_number, order_dow, order_hour_of_day, days_since_prior_order. Every single order corresponds to an unique order_id. User_id of an order corresponds to the customer who make the order. Eval_set represents the use of the order data in our analysis. Order_number represents the serial number of the orders which made by the same customer. Order_dow is the day of week. Order_hour_of_day is the time which order made in a single day. Days_since_prior_order is the interval between previous order and this order.

There are four columns in the “products” table. They are product_id, product_name, aisle_id, department_id. Every product corresponds to an unique product_id. Product_name is the name of products. Aisle_id is the id of the secondary category which product belongs to. Department_id is the id number of primary category which the product belongs to.

“Order_products_*” tables are the relationship between order and products. There are two tables which named “order_products_prior” and “order_product_train”.These two tables both have four columns which named order_id, product_id, add_to_cart, reordered. There is many-to-many relationship between order_id and product_id. Values of theses two columns represent the correspondence among orders and products. add_to_cart is the serial number of the products which in the same order. Reordered represents if thee product is reordered by the customer, 1 means reorder, 0 means the first order of the customer. * in the name of tables is the usage of the order, it has the same value with the “eval_set” column in the “orders” table. There are three values for eval_set column — prior, train, test. Train and test are always the last order data recording which the same customer made. Prior means the table(order_products_prior) contains previous order contents for all customers. Assume a situation which we will predicting has causes and results, and we need to use causes to predicting results. If we purpose on predicting new orders of every customers, the new order is the results we should predict. Prior orders with the train data at the end of the customer’s orders are the cause data we use to train the model, train orders are the result data we use to train the model. Prior orders with the test data at the end of the customer’s orders are the cause data we use to test the model, test orders are the result data we user to test the model.
Here is the UML of the dataset.



9. DATA SOURCE

https://www.kaggle.com/c/instacart-market-basket-analysis/data

10. REFERENCE

https://www.kaggle.com/serigne/instacart-simple-data-exploration

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/33128#183176

https://www.kaggle.com/frednavruzov/instacart-exploratory-data-analysis

https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-instacart