CECS 550 Pattern Recognition

Spring 2023

Deliverables

1. **Source code** on GitHub repo
    a. Each team member of the group should collaborate on GitHub and use the "issue" feature to mark their task as complete. This will be reviewed for actual contribution of each team member.
    b. All code should be working as expected
2. **Presentation slides** (upload on Canvas)
3. **Project report** in Microsoft word file (upload on Canvas)

Submission instruction and deadline: Please zip "presentation slide" and "project report" and upload them on Canvas by 05/01/2023.

Note: Only one submission is required from each group and Group leader should upload the files on Canvas.

Repeat Buyers prediction for E-Commerce

1. Problem statement

Merchants often gain many new customers through promotions, but a significant portion of these customers are only interested in one-time deals. Therefore, the impact of promotions on long-term sales may be limited. To maximize return on investment (ROI) and reduce promotion costs, it is crucial for merchants to distinguish between one-time buyers and potential loyal customers and focus their efforts on converting the latter group.

In this project, you are provided a dataset with information on promotional shopping event from e-commerce platform. Your task is to design a system which will increase the ROI (in other words, you need to predict the probability that these new buyers would purchase items from the same merchants again within 6 months), reduce promotional cost, and identify one-time buyers.

2. Data description

The data set contains anonymized users' shopping logs in the past 6 months before and on the "Double 11" day (*Promotional Event*), and the label information indicating whether they are repeated buyers. But it will not affect the applicability of the solution. The files for the training and testing data sets can be found in "data_format2.zip".

Details of the data format can be found in the table below.

Note: It is up to you to use any of the dataset format 1 or 2 as they both have similar information.

Data Fields	**Definition**
user_id	A unique id for the shopper.
age_range	User' s age range: 1 for <18; 2 for [18,24]; 3 for [25,29]; 4 for [30,34]; 5 for [35,39]; 6 for [40,49]; 7 and 8 for >= 50; 0 and NULL for unknown.
gender	User' s gender: 0 for female, 1 for male, 2 and NULL for unknown.
merchant_id	A unique id for the merchant.
label	Value from {0, 1, -1, NULL}. ' 1' denotes ' user_id' is a repeat buyer for ' merchant_id' , while ' 0' is the opposite. ' -1' represents that ' user_id' is not a new customer of the given merchant, thus out of

	our prediction. However, such records may provide additional information. ' NULL' occurs only in the testing data, indicating it is a pair to predict.
activity_log	Set of interaction records between {user_id, merchant_id}, where each record is an action represented as ' item_id:category_id:brand_id:time_stamp:action_type' . ' #' is used to separate two neighbouring elements. Records are not sorted in any particular order.

3. Data in another format

We also provide the same data set in another format, which contains 4 files and may be more user-friendly for feature engineering (files can be found in "data_format1.zip"). The details of the data formats can be found below:

User Behavior Logs

Data Fields	**Definition**
user_id	A unique id for the shopper.
item_id	A unique id for the item.
cat_id	A unique id for the category that the item belongs to.
merchant_id	A unique id for the merchant.
brand_id	A unique id for the brand of the item.
time_tamp	Date the action took place (format: mmdd)
action_type	It is an enumerated type {0, 1, 2, 3}, where 0 is for click, 1 is for add-to-cart, 2 is for purchase and 3 is for add-to-favourite.

User Profile

Data Fields	**Definition**
user_id	A unique id for the shopper.
age_range	User' s age range: 1 for <18; 2 for [18,24]; 3 for [25,29]; 4 for [30,34]; 5 for [35,39]; 6 for [40,49]; 7 and 8 for >= 50;0 and NULL for unknown.
gender	User' s gender: 0 for female, 1 for male, 2 and NULL for unknown.

Training and Testing Data

Data Fields	Definition
user_id	A unique id for the shopper.
merchant_id	A unique id for the merchant.
label	It is an enumerated type {0, 1}, where 1 means repeat buyer, 0 is for non-repeat buyer. This field is empty for test data.

6. **Dataset access**: CECS550 (Download from OneDrive)

7. Dataset allocation for individual group – each group will be working on unique subset of the data which is explained below.

Note: The below table explains how each group is going to use the dataset. For example, from the above dataset, Group 1 only needs to work on item_id from 1 – 160. (please refer the tables in the dataset to extract your dataset).

Group	item_id
1	1 - 160
2	161 - 320
3	321 - 480
4	481 - 640
5	641 - 800
6	801 - 960
7	961 - 1120
8	1121 - 1280
9	1281 - 1440
10	1441 - 1600

Tasks

1. **Data visualization:** Visually analyze the given dataset and use the right visualization technique to present the information. For example, you only use Box plot to show distributions of numeric data values, especially when you want to compare them between multiple groups.

   a) Use your best understanding about the dataset and come up with visualizations which can help provide insight from the data.
   Hints:
   https://www.tableau.com/learn/whitepapers/which-chart-or-graph-is-right-for-you
   https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization

2. Feature engineering

Please note that `user_log` (user interaction log) and `user_info` (information about users) does not provide any structured features that can be directly embedded in some model. These datasets need to be analyzed to create valuable features that can correlate users and merchants.

Task: Create new features from the given information.

3. Dataset statistics and feature ranking

Please provide data set statistical description and the rank of the feature, and possible explanations for the same. Use the right feature importance/ranking method.

- Provide statistical summary of the dataset
- Perform feature ranking
- Demonstrate your understanding of PCA for feature reduction and later in section 4 perform the comparative analysis of the model before and after performing PCA (identify optimal number of features).

4. Prediction model

1. Iterate through different combinations of features to identify the optimal features and remove potential correlated features (if any) for your predictions (you can use the feature ranking results from section 3). Add the derived features from section 2 to your

features used to build the model. The motive is to create a consolidated dataset for building a model.

2. Start with a Bayes Classifier as discussed in Lecture 5 to get identify the customers who will be a repeat buyer or not.
    a. Use 80% of data for training and 20% of data for testing. Compare the model accuracy for training and test data sets. (5 points)
    b. Extend the use case of Bayer classier to design a recommendation system for the customers.
3. Use a non-parametric technique, for example nearest neighbors and Parzen windows for classification (as discussed in Lecture 7).
    a. For nearest neighbor demonstrate your understanding of how to choose the value of ***K*** and use various distance measure techniques.
    b. Perform a comparative study of performance analysis for Parzen window and nearest neighbor.
4. Implement a neural network model for classification and perform the model evaluation as mentioned in section 5.

5. Use the best practices as discussed in lecture – “Advice for applying machine learning”.

5. Model evaluation

1. Performance Evaluation
    - Evaluate the performance of each classification technique using metrics such as accuracy, precision, recall, F1 score, ROC curve, and confusion matrix.
    - Compare and contrast the results of each technique and provide insights on which method works best for the given classification problem.
2. Discussion
    - Discuss the strengths and weaknesses of each technique and provide insights on how to improve the classification task using the best-performing technique.

Grading

The project will be graded based on the following criteria:

- Source code
- Project report
- Presentation

Note: All the three points above also evaluates the project tasks.