**Datathon 2019 -- Predicting Rare Events**

Github link: https://github.com/Yuhan-Liu-Heidi/Datathon-2019

**Background**

In digital advertising, a “conversion” refers to the event when the shopper clicks on the ad and performs a valuable action such as signup, registration, or makes a purchase. Since “conversion” is a measurable event, it represents a reasonable proxy for the number of customers acquired during the ad campaign. Increasingly, brands and agencies looking to put a value on the Return on Advertising Spend (ROAS) require marketers such as us to optimize the ad spend such that customer acquisition is maximized.

In order to wisely spend the limited marketing dollars, we need to identify the shoppers who are more likely to respond to our ad and convert. While the number of devices to target is nearly one billion, the number of conversion events range from just a few hundreds to few thousands during the period of the ad campaign. In other words, these conversion events are extremely rare.

Data: Provided by Valassis, a leader in marketing technology and consumer engagement.

**Methods**

**1. Convert with minimal false alarm**

Introduction

**Question**: Will this shopper convert with minimal false alarm?

**Importance**: With a given shopper and their interest profile, this machine learning algorithm will be able to tell whether they are likely to convert, thus advise the marketers on whether to send this customer more digital advertisements.

Process: How did you clean and prepare the data, and what data did you use?

Data used: training.csv, validation.csv

Clean and prep:

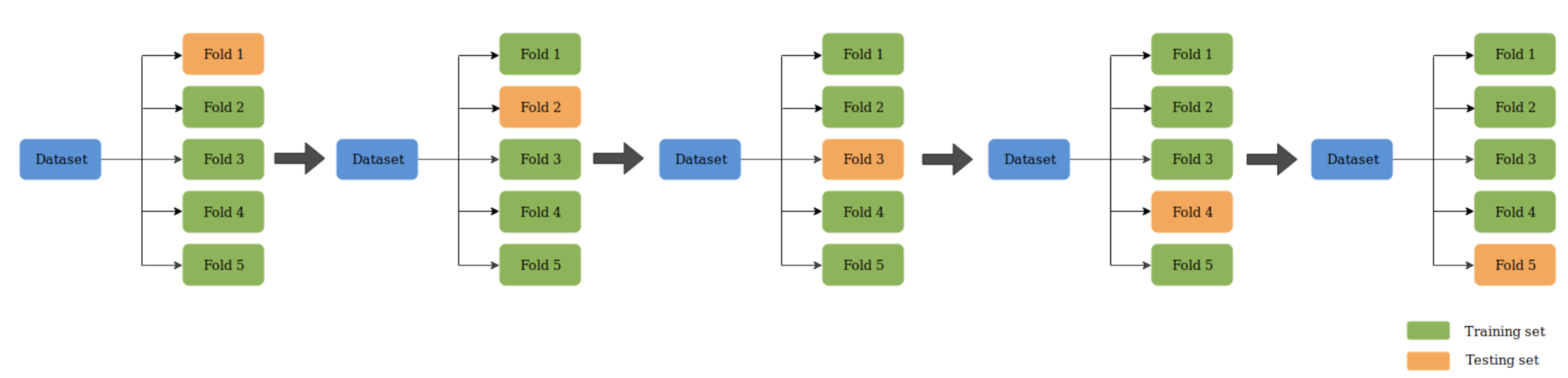
If the ltiFeature/stiFeature value is empty, assume value is zero. Normalize each user’s interest feature values to have a sum of one. Find the maximum interest feature index to create a 3D matrix of zeros (maxindex by 2 by num.ofshoppers), where 2 is ltiFeature and stiFeatures. The normalized feature values are then filled into the matrix as pixel value and converted to image.

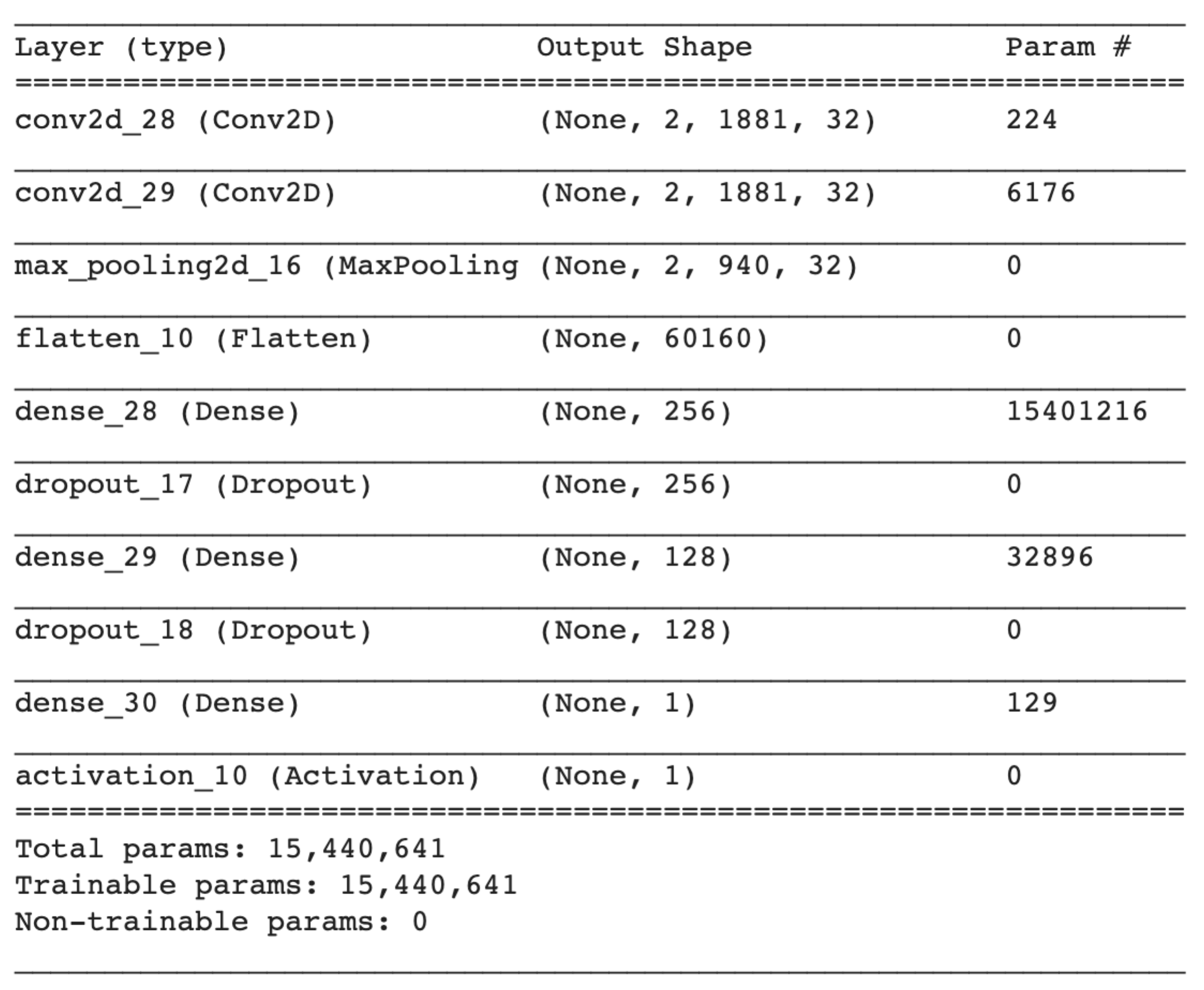
Analysis: What analytical techniques did you use, and why?

Machine learning:

The image matrix is used in a machine-learning algorithm described below.

K-Fold CV is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. Lets take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

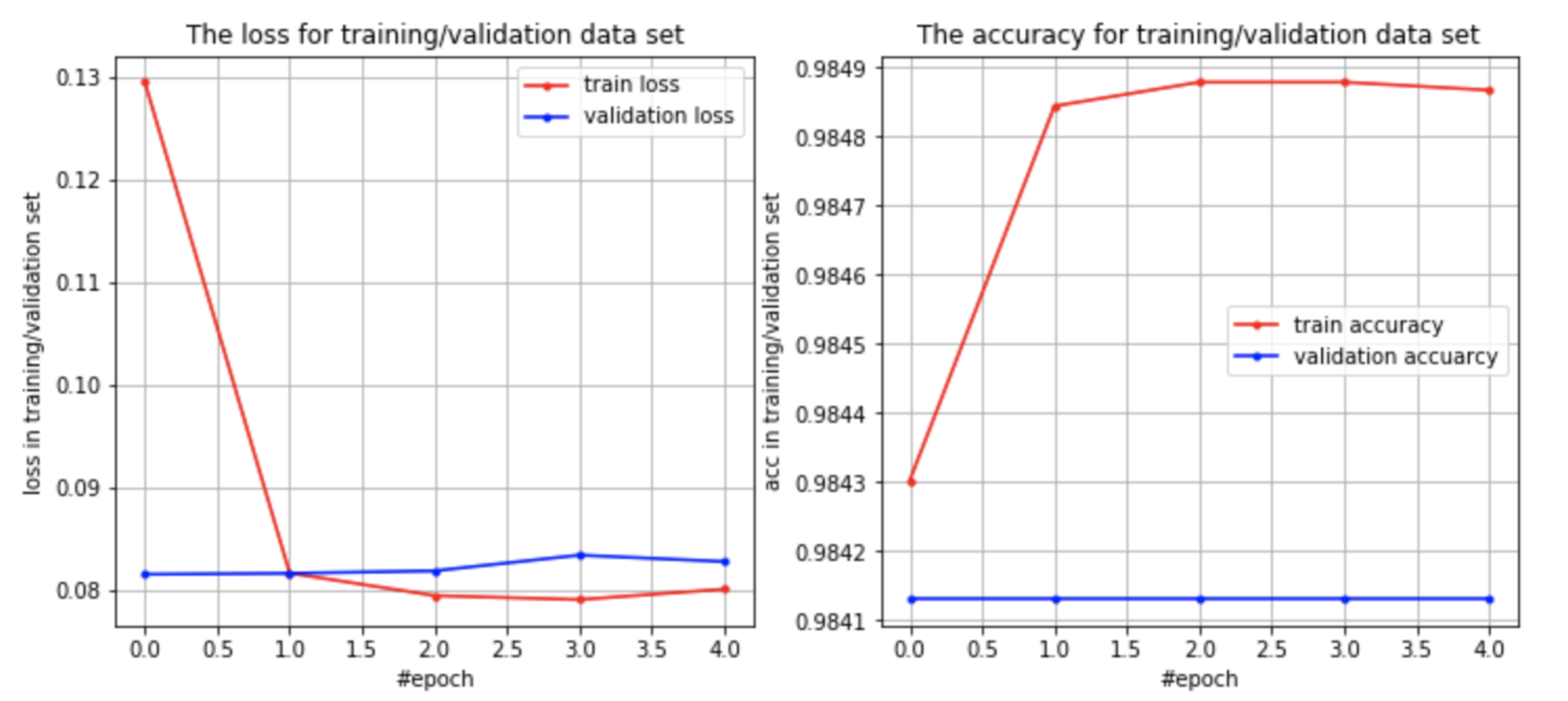


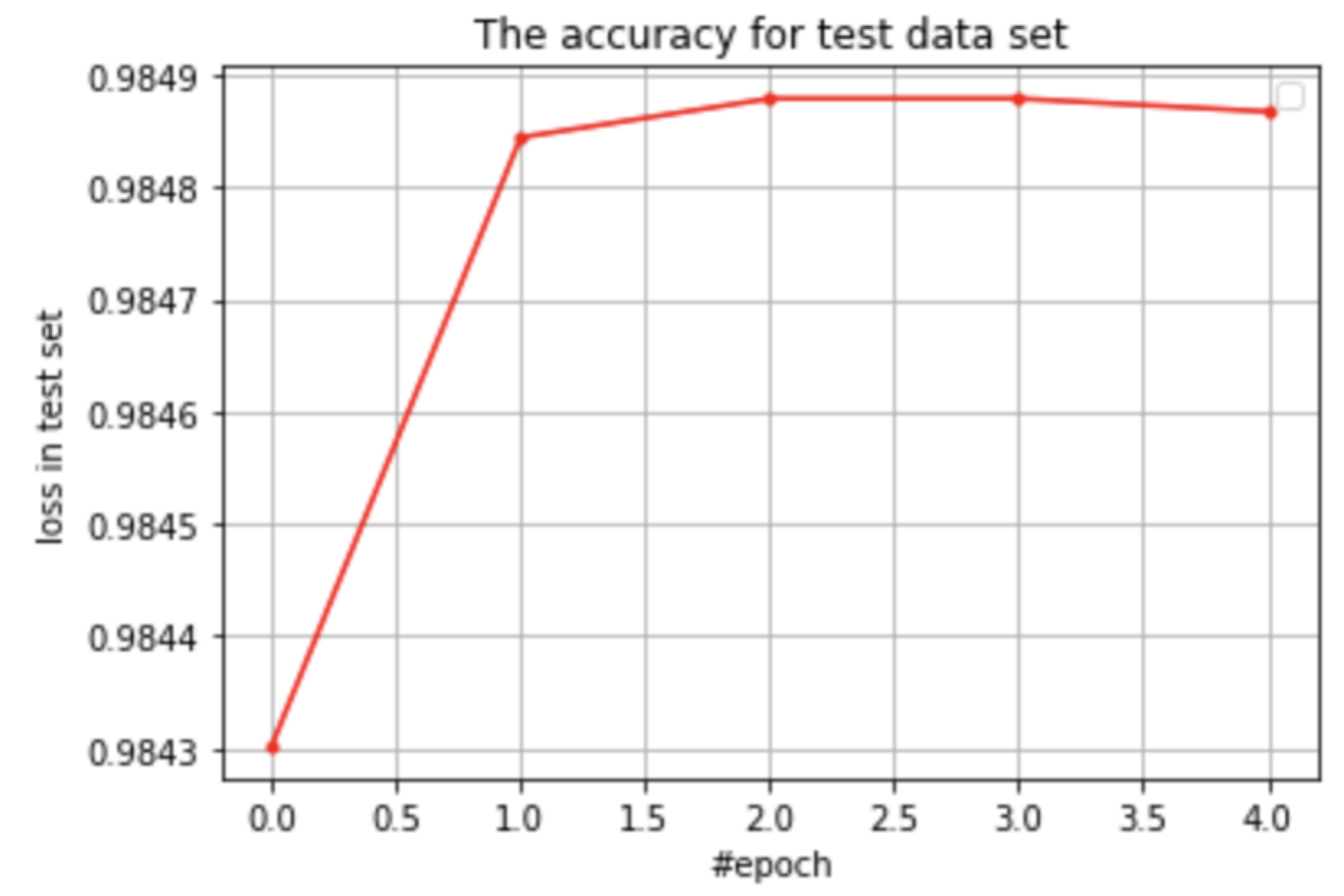


Model architecture is shown above.

Findings:

Using the network created, within 4 layers, the loss on the training/validation data set is down to 0.08, and the predicting accuracy around 0.984.





Conclusion:

Using this algorithm, the marketing person can predict, with around 98% confidence, the conversion possibility of each customer they can obtain interest features for.

**2. Convert rate within category**

Introduction

**File**: customer\_analysis.py

**Question**: Which categories of shoppers are more likely to convert?

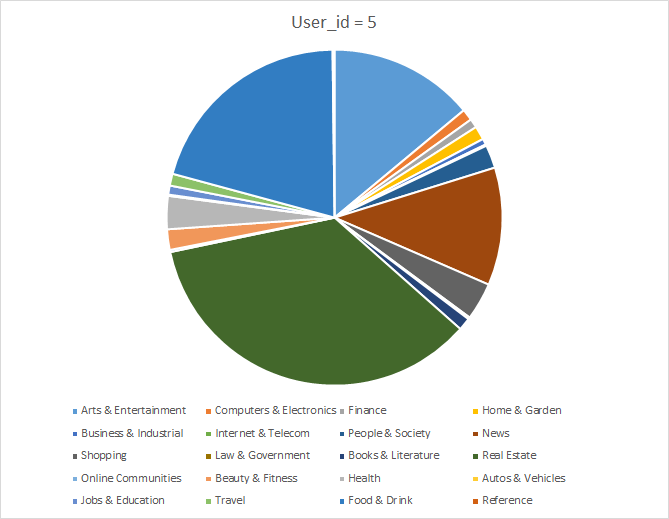
**Importance**: With the given data set of shoppers and their interest profiles, this program finds the interest category with the highest shopper conversion rate. This will help the marketer decide which category of customers to gear their advertisements towards.

Process: How did you clean and prepare the data, and what data did you use?

Data used: training.csv, interest\_topics.csv

Clean and prep:

Read interest\_topics.csv and training.csv, and categorize input topics by first level category (e.g. /Arts & Entertainment/Performing Arts 🡪 category: Arts & Entertainment). For each shopper, sum their interest for each category, and the category with the highest sum will be this shopper’s assigned category.



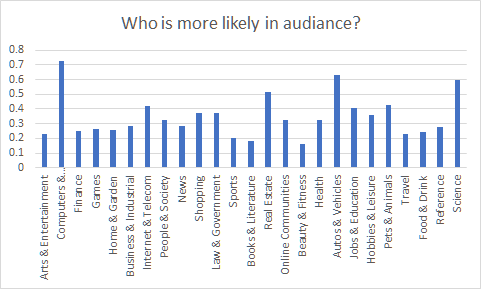
This is an example of interest weight for one customer. The customer is assigned to Real Estate.

Analysis: What analytical techniques did you use, and why?

Within each category of customers, find the percentage of customers that were converted using data from inAudience in training.csv. Plot the result as a bar graph. This enables us to compare the percentage of converted customers across categories.

Findings:

From the training data set provided, more than 70% of customers most interested in computers & electronics were converted, and more than 50% of customers with most interest in autos & vehicles, science, and real estate were converted.



Conclusion:

It can be concluded with some uncertainty (due to difference in sample size and limited total samples) that people with a lot of interest in these categories are most likely to be converted, and digital advertisements targeting these customer groups may yield higher conversion rates and be more effective.