Weight Loss Prediction Based on Network Data

Wenying Gu [†]
Data Science Institute
Vanderbilt University
TN, USA
wenying.gu@vanderbilt.edu

Yilin Yang
Data Science Institute
Vanderbilt University
TN, USA
yilin.yang@vanderbilt.edu

Tyler Derr
Data Science Institute
Vanderbilt University
TN, USA
tyler.derr@vanderbilt.edu

ABSTRACT

Using a mobile application to lose weight has becomes more and more popular worldwide. Many weight-loss mobile applications are being developed in various countries. Users could not only edit their profile on the APP to store personal information but also record their weight changes and purchase weight loss products. Some apps even provide more social attributes for users to form a weight-loss community. In the community, users may add friends, post articles, and comment on others' posts. These features allow us to use social network analysis to study the likelihood of weight loss and adopt social factors to predict weight loss [3]. This study crawls a weight-loss mobile application's social network data set of 10 million users. Results indicate that users lose weight successfully by using this app. Besides, among users who clocked in for about 400 days, males lose more weight than females. However, females on weight loss apps are more engaged. They keep recording their weight changes and more likely to become long-term users. Also, this study finds older people, especially those over 40 is harder to lose weight since their metabolism is no longer flourishing. In addition, the Graph Convolutional Networks (GCN) model is introduced in this study to predict the success of weight loss. The accuracy of the model was 65%.

KEYWORDS

Social Network Analysis, Weight Loss, GCN Model, Mobile Applications

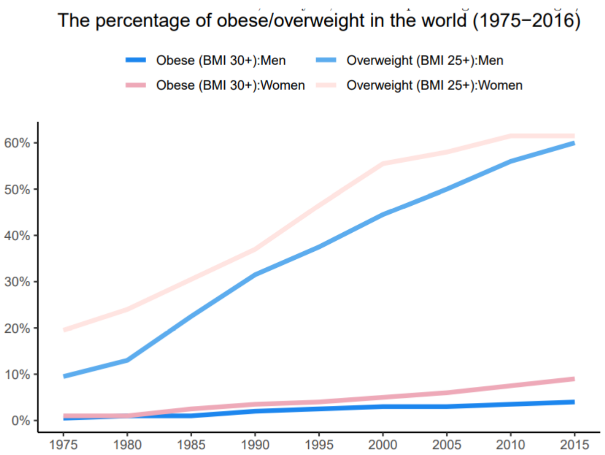
ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Obesity is a complex chronic disease characterized by excess body fat. The trend of being overweight is becoming more and more pronounced, with obese adults and children on the rise worldwide. How to control weight has become a global problem, and obesity has become the biggest public health concern in some countries. According to the obesity data released by NCD-RisC Lancet in 2019, we plot a time series graph and find that both women and men are overweight in 60% of countries [1]. It's worth noting that some recent studies have suggest that if we can keep track of our weight and be part of the weight-loss community [2], it can help us lose weight. With the rise of weight loss mobile applications, it is possible for people to practice these methods to lose weight more effectively. In this study, we will predict the user's future weight loss based on his personal information and social network information with other users. The assumption of our study is that whether a person lose weight or not is not only relevant to the user himself, but also to his local neighborhood community. Because users could get inspired by the people around them. The data set comes from an online weight loss app where users can connect, comment on each other's posts and more. The weight-loss app helps users track their weight and build connections within the community. We suggest using these associations better predict the weight loss of a user based on a Graph Convolutional Network model.

Figure 1: The Percentage of Obese/Overweight in the World



[*] Article Title Footnote needs to be captured as Title Note
[†] Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

2 Data Description

We use 11 data sets in this study[1]. What variables are contained in each data set is shown in the table below. Data sets present the behavior of 10 million users on this weight loss mobile application in various dimensions. Some of these data structures are one-to-one relationship and some are one-to-many. This makes it difficult for us to process the data later. For example, profile data set is one to one, it includes user's personal information like age, gender, BMI, etc. The weight record data set illustrate chronic changes of weights as time passing so it is one to many. In addition to the basic information related to weight loss, our data also includes other users' daily behaviors. We have diet data for calculating calories intake and disclose their purchasing history on this app. Last, besides those feature variables data, the most important information that cover in our data set is the relationship information between users. There are three data set (comment_rela, friend_consolidated, mention_rela) to represent users' social network links which make our network analysis possible.

Table 1: Data Variables Information

File Name	# Distinct Users	Column Names
userprofile	9,967,290	user_id, avatar, gendar, height, target_weight, lastdays, AvgPost, AvgComment, AvgMention, age, hasPost, hasComment, hasMention, hasHeight, hasWeightTarget
users_consolidated	9,967,290	id, avatar, gendar, birthday, location, height, weight, target_weight, target_date, latest_weight, bmi, created_date
can_record_consolidated	2,823,657	record_id, user_id, eating_calory, activity_calory, record_on, created_date
comment_consolidated	498,718	comment_id, post_id, user_id, date
order_detail_record_consolidated	257,164	order_id, user_id, good_id, quantity, created_date, price_each_goods
post_consolidated	9,360,045	post_id, user_id, date
mention_consolidated	264,969	mention_id, user_id, type, post_id, comment_id, date
weight_record_consolidated	3,952,879	user_id, weight, record_on, date

Relation file name	# of links
comment_rela	25,960,524
friend_consolidated	127,425,354
mention_rela	26,520,741

3 Data Processing

In exploring this study, we faced three significant data processing challenges. First, our data contains 10 million users, and even 25GB of memory may not handle this data. Also, when we merge 11 data sets, we get a lot of NAs. The reason is that not all users complete every behavioral variable. Filling in the missing values is a big challenge. In addition, since each user may have many weight records, how we split our data for modeling is so tricky and how to define the loss is essential.

3.1 Subsampling

We solve the memory problem by subsampling the data and analyzing it using PySpark. Although, there are about 10 million users but only 4 million of them have their weight recorded on this weight loss mobile application. The primary subsample strategy is to filter the data using the weight record duration of the user. And the concept duration is defined as the number of days that elapsed between the last time a particular user records his weight and the first time he records the weight. This way, we can remove some of the extreme values and make our sample more representative. To better understand the distribution of the data, we plot two graphs of users with different durations. As you can see from the distribution in Figure 2 and Figure 3, most users only record their weight changes for a few days and then give up. 70.7% lasted less than ten days. On the other hand, 3.5% of users keep track of their weight for more than 400 days. We screened users with a duration of between 300 and 400 days based on distribution graphs and finally subsampled the data to get about 298,497(7.6%) users.

Figure 2: The Distribution Graph of Weight Record for All Users

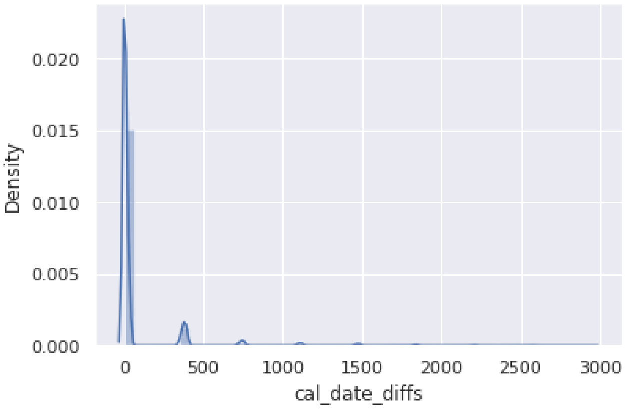
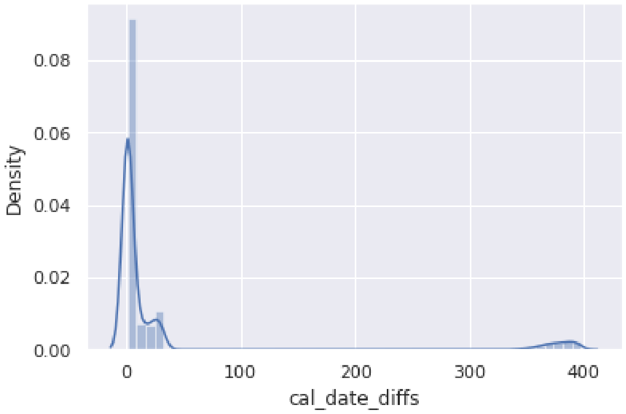


Figure 3: The Distribution Graph of Weight Record for Users Whose Duration is Between 0 – 400 Days

[1] Weight loss application dataset download link: https://vanderbilt.box.com/s/fshvoq2u7j4ybs54lrs9a4ya0k81a0ug



3.2 Missing Value

Another data processing must be done is filling in NAs. As Table 1 encapsulates, a considerable number of registered users will not record their daily calorie intake and consumption, nor post and comment on articles, nor do they shop on this mobile application. Therefore, there are many missing values. According to Z Wang et al. study in 2017[3], people's weight loss is highly correlated with age and gender. Therefore, we intend to classify users with missing values by age group and gender group. The defined age groups are: 0-10, 10-20, 20-30, 30-40, 40-55, 55-70, 70+. Then filling in NA with means of a specific variable that contains NAs.

3.3 Feature Engineering

After subsampling and impute the missing, we start doing the feature engineering. We have completed the feature engineering part to combine all our user features into one dataset. For example, one user may have many orders, it is one to many data. But we calculate the total goods for each user so that we can have one to one data for combining. Also, we labeled our weight loss into different groups and for later classification purposes. We use all the provided files, clean them, aggregate them and then add more features to get a combined data, one user one row for modelling. The additional features that we add are calories intake, daily comments, daily post, spend per order and order frequency. To get feature calories intake, for each user, we sum their difference between eating calories and activity calories and divided by the total number of records for that user; then we count daily unique comments for a user and also the daily post. Some users will purchase goods on the app, so it is interesting to know how much money each user spend per order and how many times they buy things on this app.

Table 2: Feature Engineering

Features	Methods
Calorie intake	sum (eating calories - activity calories)/number of records
Daily Comments	unique comments/ duration
Daily Post	unique post/duration
Spend per order	sum(money)/unique order frequency
Order frequency	Unique orders

3.3 Data Split

The final step of data processing is splitting data into the training dataset, the validation dataset, and the testing dataset with weight loss labels. However, how to choose our loss and split data for modeling is complex since each user may have multiple weight records. We first broke 70% of the dataset to be training and 30% as the testing. But our validation dataset is using the same user as training but counts their weight loss in different ways. For training, the weight loss is mid-weight loss. We are using the difference between the weight record that near 250 days and the weight record on the first day. Instead, the weight loss for validation is the final weight loss. To Use the difference between the weight record on the last day and the weight record on the first day. The reason why we do this is that people's weight changes are not linear but fluctuate. Girls in summer may lose some weight but gain weights in winter then start keeping fit again. But we find that if we do the classification and label our weight loss as 0 or 1, there is a significant overlap between mid-weight loss for training and final weight loss for validation. So, we finally decide to split our data based on different users, and all of them use final weight loss instead.

4 Modeling

4.1 Graph Convolutional Networks

Graph Convolutional Networks (GCN) is a type of convolutional neural network that can work directly on graphs and take advantage of their structural information [4]. Our initial idea is, the weight loss is not only associated with the user himself but also his local neighborhood since the user will get inspired by people around him. This project seeks to predict if a user will successfully lose weight at this time frame with GCN. The user is labeled as 1 if they have successfully lost weight within 400 days and labeled as 0 if they failed. Our definition of weight change is the weight difference between their last weight record and first weight record.

4.2 Feature Selection

After feature engineering, a total of 12 features were extracted to represent each user. There is individual information such as gender, age, BMI and initial weight and there are features to measure user's social activity such as daily post and daily comment as well as their purchase history in the platform.

4.3 Build Network

In this app, users can mainly perform three types of social activities: they can follow users they are interested in; they can mention others in a post or comment; and they can comment on others' posts. There are three types of network in the dataset: following, commenting, and mentioning network. The following network data is used to build the network. Since the network would change over time, one graph is created based on all the edges in the date range of our subset of data. We assume the user would keep the following relationship since the date they built the connection. To work with smaller dataset, all users with degree less than 5 is removed from the graph.

4.4 Prepare Data Object for Modeling

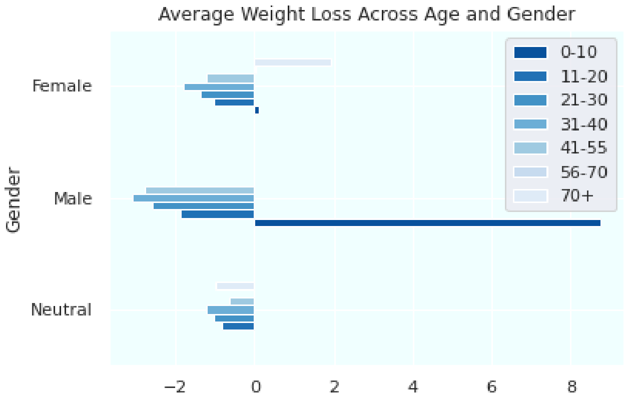
To feed data into the Graph Convolutional Network, we need to prepare data in a PyTorch Geometric Data format, which holds the attributes of node feature, edge, and classification label. The data is split into training set, validation set and testing set with a ratio of 70%, 15% and 15%. The Data object we created does not inherently have a train_mask, val_mask, or test_mask. Thus, we define them as tensors and add them as attributes to the Data object. The train_mask, val_mask and test_mask contains a list of boolean values to help filter the dataset. After hyperparameter tuning with different combinations of number of layers, hidden size, learning rate, the best accuracy we get is 65% with 4 layers, hidden size 16, learning rate 0.0001.

5 Results

5.1 Brief Data Description After Feature Engineering

After all the data processing steps, we have a summary graph to show average weight loss across age and gender among our 298,497 users. The ratio of Female to Male to natural gender is 27 to 2 to 1. It is quite unbalanced. Thus we conclude girls on this app are more engaged. They keep recording their weight changes and more likely to become a long-term user. Okay. Another finding that I want to share is overall, people lose weight than gain weight, but those who clocked in for about 400 days mam lose more weight than women. Also, we find age group 30 to 40 is more likely to lose weight regardless of gender. Older people, especially those over 40 is harder to lose weight since their metabolism is decreasing rapidly.

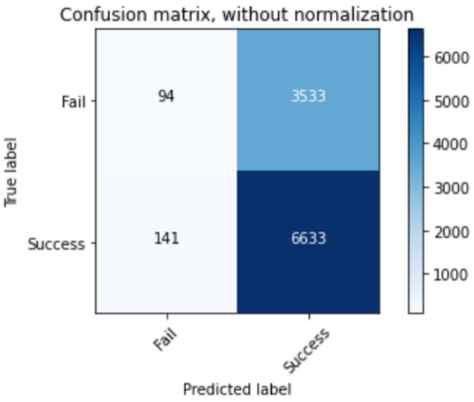
Figure 4: Average Weight Loss Across Age and Gender



5.2 Modeling Results

The classification accuracy is lower than we expected. After checking the confusion matrix, we found that there is a high recall value but low precision because our model is biased towards one class. The model tends to predict that most users will lose weight successfully. There might be several reasons contributing to this. First, in our subset of data, 95% is female. We lost most of information in gender difference, but the gender difference could be an important social factor to analyze since the impact of friends with different genders may not be the same. Besides, after testing the model performance with different feature sets, we found that the feature we add such as calory intake, order counts does not seem to make much difference in the model performance.

Figure 5: Confusion Matrix without Normalizations



6 Discussion

In conclusion, the GCN we built have a relatively lower accuracy that we expected and it is strongly biased towards one class. We need to think about new strategies to subset a balanced dataset, impute missing values in a reasonable way and add features to measure users' social activity with their neighbors of different gender and age groups. Also, we plan to train the model with different graph models and compare the model performance. Since there are induvial factors and social factors in the features, the next step would be conducting feature analysis to understand the importance of features. The feature analysis could help us understand if the result of weight loss is largely dependent the user itself or their social activities in this platform. Besides, since the app records user's weight change over time, it would be interesting to conduct time series analysis to track the weight loss on a monthly, quarterly, and yearly basis.

The most likely reason for the low accuracy may be that there may be something wrong with the date cutoff of our mid-weight loss. Because people who have adhered to 250 days have a high probability of having developed good fitness and eating habits, the results of losing weight within 250 days and the results of losing weight within 400 days will not change. Therefore there is a significant overlap between mid-weight loss for training and final weight loss for validation. And we finally decide to split our data based on different users, and all of them use final weight loss instead. In the future, we will try monthly cutoff instead of long-term users whose duration lasting 400 days.

In addition, when selecting data, we should consider the number of times he records his weight to determine whether he is an active user, not just the duration of the recording.

Finally, we will try other social network models in the future, such as Graph Attention Networks Model (GAT), and use all the data of Network Relation. This time, we just use the following relation for each user.

ACKNOWLEDGMENTS

Thank you very much to our instructor Tyler. Over the past semester, Tyler has been working to help us learn more about the concepts and models of social network analysis. He is a conscientious and responsible professor, and we have learned a lot in the past four months. Tyler often responds to our messages on Slack outside of work hours and spends a lot of time meeting with us to guide this study.

REFERENCES

[1] “NATIONAL ADULT BODY-MASS INDEX.” NCDRISC, 2017, ncdrisc.org/data-downloads.html.

[2] Sasan Adibi. 2015. Mobile health: a technology road map. Vol. 5. Springer

[3] Wang, Zhiwei, et al. "Understanding and predicting weight loss with mobile social networking data." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.

[4] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).