CS229 Project Milestone
Uplift Modeling : Predicting incremental gains

Akshay Kumar
akshayk@stanford.edu

Rishabh Kumar
rkumar92@stanford.edu

October 2018

1 Introduction

Uplift modelling is a predictive response modelling technique which models the "incremental" effect of a treatment on a target group. This measures the incremental gains of a particular treatment on a population. More precisely, to measure uplift effect, we divide the population into two groups : control and exposed. Exposed group is exposed to the treatment whereas control group is suppressed from the treatment. The difference in their responses is used to gauge the "uplift" effect. Uplift modelling has been applied to political election and ad products to study the effective of a political campaign and ad effectiveness.

Traditional response modelling techniques [KIJ15, CMT87] build a predictive model to predict the response of an individual to a treatment (for e.g., seeing an ad campaign) based on prior response of treated individuals. In contrast, uplift model [JJ12] predicts the response of a treatment based both on treated and control population.

Uplift modelling is unique in the sense that it is composed of two different factors: let's assume we are interested in the effect ad on the purchase of a particular product. Uplift modelling is interested in the modelling what "additional" purchases an ad brings in i.e. $P(\text{purchase}|\text{treatment}) - P(\text{purchase}|\text{no treatment})$. Here, treatment implies watching the ad and no treatment implies not watching the ad.

2 Problem

In this project, we will apply uplift modelling on Hillstrom email dataset [hil]. Hillstrom email is an email campaign for an online retailer where the responses of target audience was captures for two weeks after the campaign was over.

3 Dataset

Hillstrom email dataset contains email campaign related data for 64,000 customers with some purchase in past twelve month. The overall population is divided into three different groups: one-third received a mail featuring men's merchandize, other one-third featuring women's merchandize and the remaining one third received no email. Each record in the dataset contains 8 features.

The 8 features are: 1. Recency: months since last purchase 2. History segment: bucketed spend in past 12 months 3. History: actual spend in past 12 months 4. Mens: Indicator variable indicating men's merchandize purchase in past 12 months 5. Womens: Indicator variable indicating women's merchandize purchase in past 12 months 6. Zip Code: Zip Code classification 7. Newbie: Indicator variable indicating if a purchaser in newbie 8. Channel: Channel of the last purchase

Corresponding to each input example, we have three outputs: indicator variables indicating a visit, a conversion and a spend respectively.

4 Algorithms Used

We will study tackle this problem from two different perspective: predictive response modelling and uplift modelling. Response modelling tries to predict the probability of purchase (or a visit

or conversion in our example) based on the input features. Here, input also includes the email campaigns. Uplift modelling, on the other hand, models the "incremental" probability of purchase (visit or conversion, respectively) based on exposing to the email campaign.

For project milestone, we will mainly focus on predictive response modelling. We will tackle uplift modelling in final project report.

Before discussing the algorithms used, we will briefly talk about data sanitization step:

4.1 Data Preprocessing

All the feature were either real values or enums. For enums, we decided two different approaches:
- Directly encode it as an ordinal corresponding to each of the enum.
- Encode it as a one hot vector.

When a feature was represented as a one hot vector, the final feature vector was the concatenation of all the one hot feature vectors and other real values features. When using the one hot representation, each training data was encoded as a 20 dimensional vector.

4.2 Prediction Model Details

We experimented with two different models: logistic regression and two layer neural net.
- Logistic regression model had a fully connected layer followed by sigmoid activation layer.
- 2 layer neural network had a fully connected layer followed by relu activation followed by fully connected layer followed by sigmoid activation layer.

Both the models were trained independently once for all three output variables: visit, conversion and spend. We decided to use adam optimizer instead of gradient descent since it gave better accuracy. Loss function used was cross entropy. We did a batch size optimizer and used a batch size of 32. We ran the algorithm for 6 epochs – the increment in accuracy after 6 epochs was neglible and the model was beginning to overfit.

5 Preliminary Results

We did the analysis using tensorflow library on a Google Compute Engine powered backend using a NDIDIA Tesla K80 GPU.

We split the whole data into a 60:20:20 split. 60% for training, 20% for validation and the remaining 20% for testing. Overall, two layer neural network achieved better results than logistic regression model. However, the different in quality was not large.

We first experimented with encoding enum based feature as a single dimensional discrete feature. However, the accuracy on validation set in this approach was only 60%. Table 1 presents the results in a tabular form.

Architecture	Visit Accuracy	Conversion Accuracy	Spend Accuracy
Logistic Regression	57.85 %	59.67 %	59.44 %
2 layer NN	59.34 %	60.61 %	61.26 %

Table 1: Model accuracy on test set without using one hot vector representation for data representation

We believe this is the case because each of different enums impacts the purchase probability differently and have no interconnection with other values of the same enum. As such, the weights learnt for each enum should be different.

Based on this, we decided to adopt a one hot vector approach for representing input data. The accuracy shot up to 85 % using one hot vector representation – the detailed results are in Table 2.

Architecture	Visit Accuracy	Conversion Accuracy	Spend Accuracy
Logistic Regression	84.31 %	84.78 %	83.64 %
2 layer NN	85.82 %	86.47 %	85.14 %

Table 2: Model accuracy on test set with one hot vector representation for data representation

Additionally, Figure 1 shows the loss function during training for visit model.

Two layer neural network with one hot vector input data representation achieved a precision of 85.39 % and a recall of 87.71 %.

5.1 Ablative Analysis

We tried to evaluate the importance of various feature by using logistic regression by looking at drop in accuracy by selectively removing each feature. After selectively removing each of the features, the model accuracy we got was:

- Recency: 82.40 %
- History segment: 81.29 %
- History: 80.28 %
- Mens: 78.05 %
- Womens: 84.56 %
- Zip Code: 84.71 %
- Newbie: 84.62 %
- Channel: 85.14 %

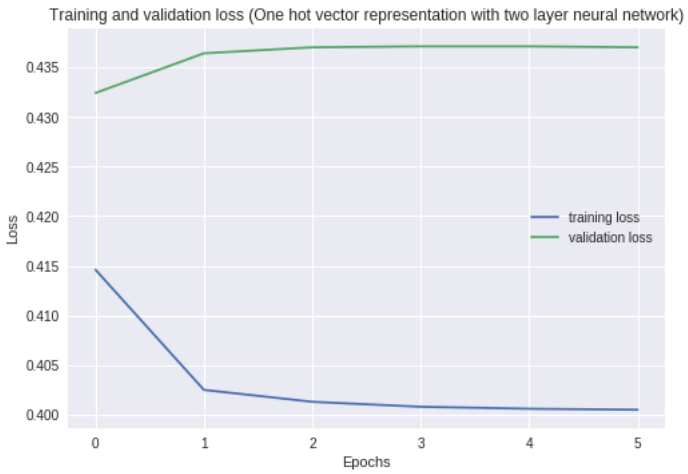

Figure 1: Loss function during model training for the case when we we used one hot vector representation for training data and a 2 layer NN for training

This shows that the most powerful signal was purchasing a men's merchanize in the past 12 months.

6 Next Steps

Most of what we have done till now posits the problem as a classification problem. We still haven't entered tackled it as an uplift modelling problem. Towards that end, we want to develop models that model the incremental effect of showing an ad.

The first approach we want to try is to simply build two separate models: one using training examples which were exposed to the email campaign and the other model using training examples which were not exposed to the email campaign. The uplift achieved is simply the difference of two probabilities.

One problem we encounter with uplift modelling is prediction accuracy : the output labels correspond to either the case when a validation data was either exposed to the email campaign or the validation data was not exposed to an example. Hence, we can't directly compare the model output (since it predicts an increment gain).

For model evaluation, we will bucketize the test data and compare the observed uplift for each of the buckets. Two training examples will be in same bucket if all their enums valued dimensions are same and real valued dimension are close to each other. For each bucket, we can get an uplift score based on difference in visit rates for exposed to email campaign group and not exposed to email campaign groups. We will compare this against model's output. In a way, this changes the problem from a classification problem to a regression problem.

References

[CMT87] David W Clarke, Coorous Mohtadi, and PS Tuffs. Generalized predictive control—part i. the basic algorithm. *Automatica*, 23(2):137–148, 1987.

[hil] Kevin hillstrom: Minethatdata. https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html. Accessed: 2008-03-20.

[JJ12] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, 2012.

[KIJ15] DKM Kufoalor, L Imsland, and TA Johansen. High-performance embedded model predictive control using step response models. *IFAC CAO*, 15, 2015.