Uplift Modeling : Predicting incremental gains

Akshay Kumar
akshayk@stanford.edu

Rishabh Kumar
rkumar92@stanford.edu

December 2018

1 Abstract

Marketing strategies directed to random customers often generate huge costs and a weak response. Sometime, such campaigns tend to unnecessarily annoy customers and make them less likely to react to any communication. Uplift Modelling has turned out to be very successful technique in understanding difference between the behaviour of a treated and a control population and this has been used as the basis for targeted direct marketing activity. Uplift modelling has wide applications in customer relationship management for up-sell, cross-sell and retention modelling. It has also been applied to political election and personalized medicine. We tried to approach this problem with 2 different perspective, predictive response modelling and uplift modelling. For predictive response modelling and uplift modelling, we used 4 models: Logistic Regression (with and without bagging), three layered neural network along with decision tree to apply on a real world example. Three layered neural network demonstrated significant advantages of uplift modeling over traditional, response based targeting.

2 Introduction

In this section we will talk about introduction to uplift modelling. Consider a direct marketing campaign where potential customers receive some advertisement and a typical application of machine learning techniques in this context will involve selecting a small group who received such advertisement and then build classifier for that group. In such techniques, we will figure out which customers are most likely to buy after the campaign and they will be selected as target. Unfortunately this is not optimum strategy as there are some customers who would do the sale regardless of the campaign and there are some who will be annoyed by the campaign. Targeting them results in unnecessary costs. The result is a loss of a sale or even a complete loss of the customer. Uplift modeling techniques provides a solution to this problem implying that we should only target customers who will buy because of the campaign, i.e., those who are likely to buy if targeted, but unlikely to buy otherwise. 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.

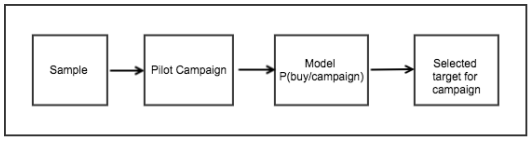

Figure 1: Uplift modelling creation process

Also, as said earlier, uplift modelling is unique in the sense that it's only concerned with incremental effect, *i.*e., Pr[purchase|treatment] − Pr[purchase|no treatment]. Here, treatment implies watching the ad and no treatment implies not watching the ad.

3 Related Work

There has not been many machine learning papers studying similar problems, but learning effectiveness of marketing campaign along with identification of target group has always been a hot topic. 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 on both treated and control population. Talluru [Tal] talks about dynamic-uplift modelling which considers time dependent behavior of the customers. Both the above papers talk about standard classification models such as logistic regression. Motivated by these papers, we performed logistic regression (with and without bagging) on our dataset. Jaroszewicz et. al. [MSR14] talk about ensem-

ble methods for uplift modeling. We were also inspired by Jaskowski et. al. [RJ12] who discuss about causes of low effectiveness in visit prediction and also touch on uplift modelling.

4 Dataset & Feature

We used Hillstrom email dataset [hil] for doing analysis of predictive response and uplift. Main motivation behind choosing this dataset was good number of customers along with uniformity in treatment data for customers. This dataset contains email campaign related data for 64,000 customers with some purchase history in past twelve month. The overall population is divided into three different groups:

- 1/3 were randomly chosen to receive an e-mail campaign featuring Men merchandise.
- 1/3 were randomly chosen to receive an e-mail campaign featuring Women merchandise.
- 1/3 were randomly chosen to not receive an e-mail campaign.

Each record in the dataset can be described as in Figure 2.

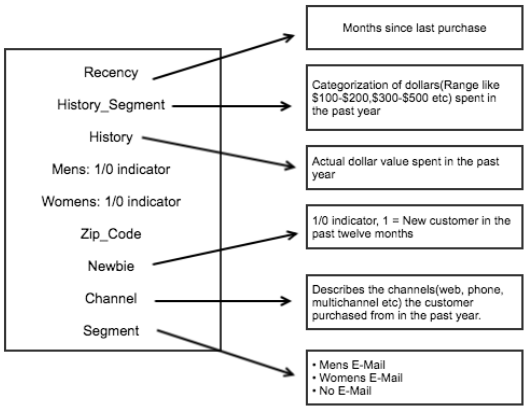

Figure 2: Features in Hillstorm Dataset

Corresponding to each input example, we have three output indicator variables:

- Visit: 1 = Customer visited website in the following two weeks.
- Conversion: 1 = Customer purchased merchandise in the following two weeks.
- Spend: 1 = Customer dollars spent in the following two weeks.

We realized that dataset has categorical features like segment, history segment, channel etc. So we have to preprocess the data before feeding it to the model. Data preprocessing is explained in later section.

5 Algorithms Used

We tackled 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 exposure to the email campaign.

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

5.1 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.

5.2 Prediction Model Details

We experimented with two different models: logistic regression (with and without bagging), decision tree and three layer neural net.

- **Logistic regression (LR) model** : Fully connected (FC) layer followed by sigmoid activation layer.
- **Logistic Regression with bagging (BBLR)** : Same as logistic regression but with bagging.
- **Decision Trees (DT)** : Since many feature were based on enum values, we decided to create decision tree with Gini criteria.
- **3 layer neural network (3NN)** : FC followed by ReLU followed by FC followed by ReLU followed by FC followed by sigmoid activation.

All 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 for logistic and neural network and gini impurity for decision tree. We did a batch size optimizer and used a batch size of 32. We ran the algorithm for 5 epochs – the increment in accuracy after 5 epochs was neglible and the model was beginning to overfit.

5.3 Uplift Modelling

We will now discuss how to model "incremental" ad effectiveness. The major problem with uplift modelling is we do not have training data for "incrementality" : a user has either seen or not seen the ad. It can not both see and not see the ad.

To tackle this, we will create two different models: one for computing probabilities when a user was not exposed to an ad campaign and the other when the user was exposed to the ad campaign. Both of models would output the probability of, *e.g.,* visit. Please refer to 3 below for getting visual representation of uplift modelling with 2 models. To get the incremental effect, we take the difference of the two probabilities.

Pr[visit is driven by ad] = Pr[purchase|ad] − Pr[purchase|no ad].

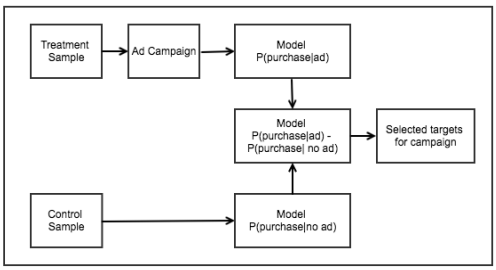

Figure 3: Uplift modelling based on 2 different models

5.3.1 Evaluation

Evaluation also suffers from the same problem : one single test data can not both see and not see the ad. As such, we can not directly compare predicted uplift of a test data to the actual ground truth for a test data (since there's no notion of "ground truth" here).

To solve this we will look at ROC curves for logistic regression and their extension into the realm of uplift modelling. ROC curve is a plot of TPR (True Positive Rate) versus False Positive Rate (FPR). A sample ROC curve is shown in Figure 6. A good model is one which has high TPR and low FPR i.e. it shoots up at the beginning and then stays flat.

We will extend this idea to uplift modelling evaluation as well. Since a single test data can not have ground truth for uplift value, we look at overall uplift value by bucketing test data. We bucket the test data based on similar categorial features : test data whose enums features have the same values go into the same bucket. For a single bucket, we look at overall uplift rate by looking at differences in visit rate for the set of users of that bucket who saw the ad campaign and the set of users in the same bucket who didn't see the ad campaign.

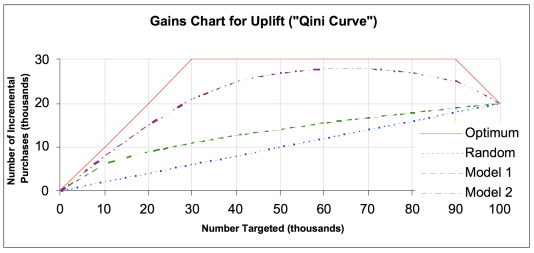

Figure 4: Qini Curve [Rad07]

Note that in this case, we can not compute normal logistic regression metrics like sensitivity, recall, F-score, etc. because the output variables are not 0/1 indicator variables. Instead, they are uplift probabilities. We will extend ROC curve and define a variant of ROC curve(qini like curve) which plots TPR and FPR for both the cases where an incremental visit happened due to ad and existing visit stopped due to ad (perhaps because the ads are annoying).

Qini curve sorts these buckets in descending order based on predicted uplift rate and plots it against number of users targeted. In this case, there is a chance that some users who actually intended to visit might actually get annoyed and end up not visiting after seeing the ad. As such, there can be a dip in the graph as well. See Figure 4. As shown, for the ideal model, there will be a dip at the very end. This segment corresponds to annoying set of users.

6 Results

We did the analysis of both logistic regression and neural networks using tensorflow library on a Google Compute Engine powered backend using a NVIDIA Tesla K80 GPU. In coming sections, we will focus on results of both predictive response and uplift modelling

6.1 Predictive Response Modelling

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

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.

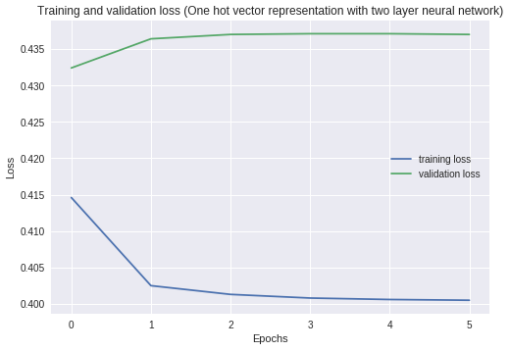

Figure 5: Loss function during model training for the case when we we used one hot vector representation for training data and a 3 layer NN for training

Config	Visit	Conversion	Spend
LR	57.85 %	59.67 %	59.44 %
BBLR	58.12 %	59.83 %	60.02 %
DT	49.81 %	49.34 %	48.89 %
3NN	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.

Config	Visit	Conversion	Spend
LR	85.01 %	85.72 %	84.27 %
BBLR	85.79 %	84.27 %	85.28 %
DT	63.57 %	61.39 %	63.74 %
3NN	87.01 %	87.83 %	86.21 %

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

Additionally, Figure 5 shows the loss function during training for visit model.

3NN performed the best among all the models. In comparison, decision tree didn't perform as good as other models. We believe this happened because decision tree was not complex enough to learn the underlying model.We used f-score as our metric to ensure we do not suffer from class imbalance problem. Decision tree gave us major difference in F-score on training data vs test data which was not really the case with other approaches.

Refer Table 3. This shows F-score only for visit.

Model	Train	Test
LR	0.753	0.7313
BBLR	0.7689	0.749
3NN	0.801	0.79
DT	0.7129	0.6366

Table 3: F-scores for the three models

We also plotted ROC curve for the neural net based models (LR, BBLR and 3NN) for visit. We did this by plotting TPR (True Positive Rate) against FPR (False Positive Rate). As shown in Figure 6, 3NN has the highest AUC metric.

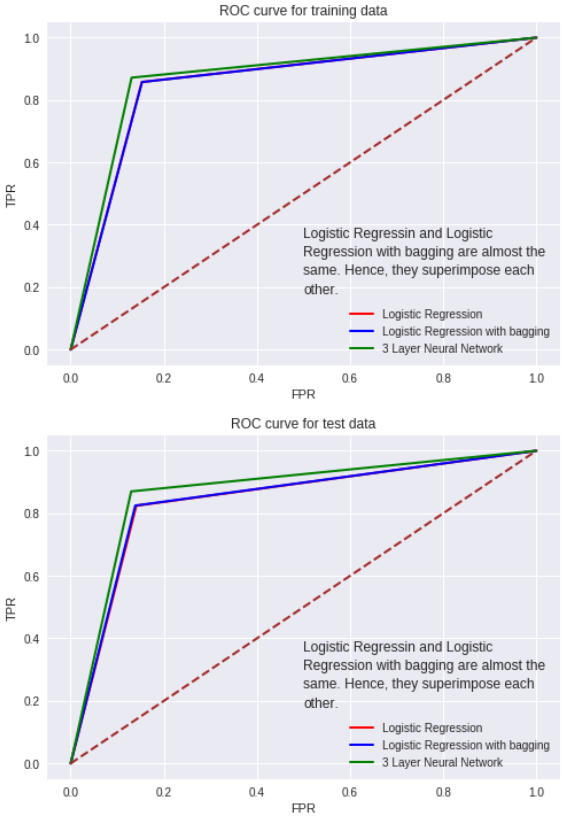

Figure 6: ROC Curves for Predictive Response Modelling

6.2 Ablative Analysis

We had multiple features in our dataset and tried computing the weightage of each of the metric. We tried to evaluate the importance of various feature by using logistic regression by looking at drop in accuracy by selectively removing each feature. Assumption here was mostly results obtained here would be same even if we would have tried neural network or any other techniques. After selectively removing each of the features, the model accuracy we got was:

- Recency: 82.40%
- History segment:81.29 %

- History: 80.28%
- Men: 78.05 %
- Women: 84.56 %
- Zip Code: 84.71 %
- Newbie: 84.62 %
- Channel: 85.14 %

This shows that the most powerful signal was purchasing a men merchandize in the past 12 months as that corresponds to maximum drop in accuracy among all features.

6.3 Uplift Modelling : Results

As described in Section 5.3.1, we look at a variant of Qini curve (by discretizing the predicted uplift rates) for email campaign. Figure 7 shows the two curves for our training and test data. We trained our model using Logistic Regression(With and Without Bagging) along with 3NN. Here again, we observed that 3NN had the best AUUC(Area Under Uplift Curve) amongst LR, BBLR and 3NN.e

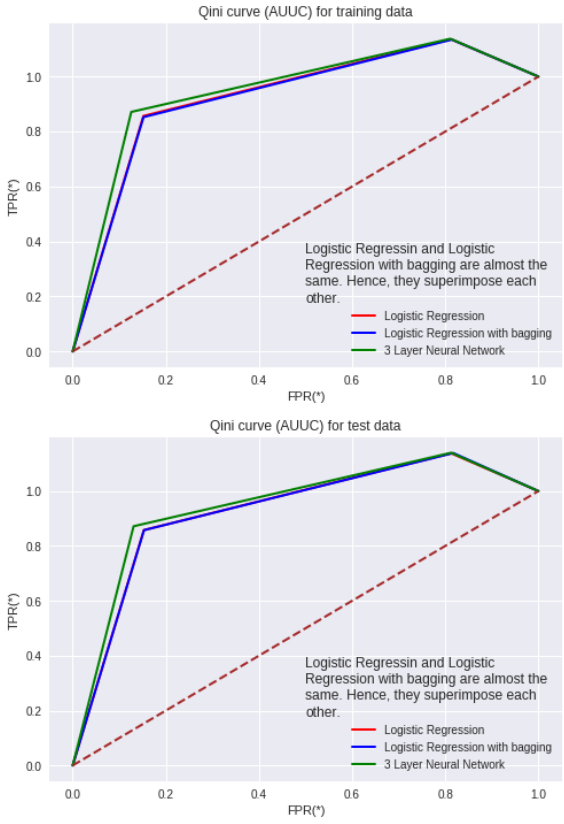

Figure 7: Qini curve for uplift modelling for train and test data

The salient thing to note is after a certain point, uplift actually starts dropping. For our data, this point corresponds to 80% of the whole data. Also, we didn't try training decision tree classifier due to the fact that it performed poorly on predictive model technique and doesn't give a "probability" output.

6.3.1 Ablative Analysis

We also performed ablative analysis for uplift modelling by looking at the drop in AUUC (Area Under Uplift Curve) by selectively removing features from feature set. The feature which corresponded to maximum drop in AUUC was man's merchandize purchase in past 12 months. We found this result to be consistent with our predictive modelling approach.

7 Conclusion & Future Work

Our experiments confirm the usefulness of uplift modeling in ad campaigns. Here are important observations which we noted:

- We experimented with 4 different models for predictive response and neural network gave best F-score out of 4 models. Decision tree overfits the training data and predicts poorly on test set. We recorded appreciable difference between training and test F-score for decision tree.
- Uplift modelling allows us to better target the intended set of users compared to random targeting.
- As seen in Qini curve, beyond a certain point, there is a dip. These user are annoyed by the email campaign and might end of not purchasing the merchandize when they could have otherwise. This illustrates the possibility of achieving more incremental effect by targeting a smaller group. In our case group size comes out to be 80% of total population.
- Ablative analysis for both the models (predictive response and uplift modelling) indicates that men merchandize purchase is the most potent signal for ad targeting.

As future work, we would like to experiment with more complicated and advanced neural net architectures. One approach we would like to experiment with is layered approach: instead of having two separate NN, the neural network for control also feeds into the neural network for exposed – there may be some common weights between control and exposed to learn. We would also like to explore dynamic uplift modelling where data gathered can be from multiple instances of time

We have uploaded all our code here

8 Contributions

Both of us worked together collaboratively on almost all aspects. So there is no clear distinction in the contributions.

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. 2012.

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

[MSR14] Szymon Jaroszewicz Michał Sołtys and Piotr Rzepakowski. Ensemble methods for uplift modeling. 2014.

[Rad07] NJ Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Market J Direct Market Assoc Anal Council*, 1:14–21, 2007.

[RJ12] Piotr Rzepakowski and Szymon Jaroszewicz. Uplift modeling in direct marketing. *Journal of Telecommunication and Information Technology*, 2012.

[Tal] Gowtham Talluru. Dynamic uplift modeling. *GAlLOGLY COLLEGE OF ENGINEERING*, 1.