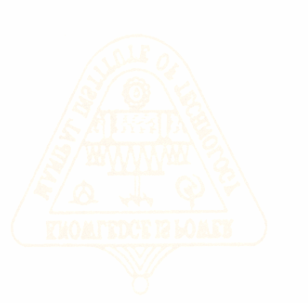


UPLIFT MODEL FOR CUSTOMER PROPENSITY MODELING

SUBMITTED

BY

SAI RAGHU TEJA DAVULURI 170907024

 Under the Guidance of:

Dr. Goutham Simha GD Mr. AAKASH SHARMA

Assistant Professor Sr. Manager

Department of ECE AS Dept

Manipal Institute of Technology Ugam Solutions Pvt. Ltd.

TABLE OF CONTENTS

Contents Page No.

Abstract 2

1. Introduction 3
   1. General 3
   2. Organization 3
   3. Area of Computer Science 4
   4. Software requirements 4
2. Problem Definition 5
3. Objectives 5
4. Background 6
5. Methodology 7
6. Implementation Details 9
7. Project till date & remaining work 11
8. References 12
9. Project Details 13

**ABSTRACT**

The desire to model the true gain from targeting an individual in marketing purposes has lead to the common use of uplift modeling. Efficient methods for estimating the probabilities in uplift models are statistical machine learning methods. The statistical machine learning methods applied are Random Forests along with the standard method Logistic Regression.

The data is collected from a well-established retail company and the purpose of the project is thus to investigate which uplift modeling approach and statistical machine learning method that yields in the best performance given the data used in this project. The variable selection step was shown to be a crucial component in the modeling processes as so was the amount of control data in each data set. For the uplift to be successful, the method of choice should be either model it directly using Random Forest, or ‘class variable transformation’ using Logistic Regression. Furthermore, the ‘subtraction of two models’ did not perform well since each model tended to focus too much on modeling the class in both data sets separately instead of modeling the difference between the class probabilities.

The conclusion is hence to use an approach that models the uplift directly, and also to use a great amount of control data in the data sets.

**CHAPTER-1 INTRODUCTION**

* 1. **General**

Uplift modeling uses a randomized scientific control to not only measure the effectiveness of a marketing action but also to build a predictive model that predicts the incremental response to the marketing action. It is a data mining technique that has been applied predominantly in the financial services, telecommunications, and retail direct marketing industries to up-sell, cross-sell, churn, and retention activities.

This project deals with knowing the metrics on which the customers will be segmented. Based on the segmentations, we will be creating strategies so that customers can be retained. After this comes analyzing the customer’s purchasing pattern (predicting next purchase day) and the sales at a certain period of time. Based on this, we will be creating uplift model along with the market response model.

* 1. **Organization [4]**

Ugam, a Merkle company, is a leading analytics and technology services company. Our customer-centric approach delivers impactful business results for large corporations by leveraging data, technology, and expertise. We consistently deliver superior, impactful results through the right blend of human intelligence and AI. With 2400+ people spread across locations worldwide, we successfully deploy our services to create success stories across industries like Retail & Consumer Brands, High Tech, BFSI, Distribution, and Market Research & Consulting. Over the past 20 years, Ugam has been recognized by several firms including Forrester and Gartner and was recently named the No.1 data science company in India by Analytics Insight.

The name Ugam comes from the ancient Sanskrit language and means “source” or “origin.” Ugam seeks to be the source of solutions for the complex problems of retailers, brands and market research firms worldwide. The name signifies the company’s position as a pioneer in the market for data and analytics, and the source of relevant actionable insights for clients.

Ugam’s **analytics services** help drive growth by answering key business questions across functions:

* Customer & Marketing
* Sales, Channel & eCommerce
* Product & Pricing
* Analytics Center of Excellence

Ugam's **technology services** provide consulting, implementation and managed services in the areas of:

* Cloud & Data Engineering
* Data Visualization
* Experience Management

Ugam’s **digital commerce** **services** help businesses improve customer experience and overall digital performance by addressing business questions across:

* Product
* Channel
* Experience

Ugam’s **market research services** leverage AI to accelerate and automate research in the following areas:

* Research Design & Programming
* Data Collection
* Data Management
* Reporting
  1. **Area of Work**

**Predictive modeling** is the process of using known results to create, process, and validate a model that can be used to forecast future outcomes.  It is a tool used in **predictive** analytics, a data mining technique that attempts to answer the question "what might possibly happen in the future?”[5] **Data mining** is a process used by companies to turn raw data into useful information[6].**A/B testing** is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B and determining which of the two variants is more effective[7]. **Regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features')[8].

* 1. **Software Requirements**

The model is developed using Python 3.6. In this, essential libraries used for classification & regression are numpy, pandas, pylift, matplotlib (for visualization), scikit learn (for splitting and standardizing training & testing set). The IDE on which we will build the model is Jupyter Notebook.

**CHAPTER-2 PROBLEM DEFINITION**

An assumption that Direct Marketing campaign has that it will achieve maximum incremental Response when a group of the highest scored customers is targeted. A Propensity/Response model itself is not going to tell marketers which customers are most likely to contribute to the incremental campaign response. An alternative statistical model is therefore needed to target the customers whose propensities of response are dramatically driven by “touching” customers with a promotion.

So, the question to be answered is how to optimize customer targeting in the marketing domain by using the uplift modeling approach, and at the same time being able to model the true gain from targeting one specific individual. Furthermore, how should the uplift modeling technique be implemented in the best way to obtain the most applicable results given this kind of data?

**CHAPTER-3 OBJECTIVES**

* Know your metrics & segregate the customers according to that.
* Analysing product market fit (PMF) for retaining the customers.
* Predict future sales using predictive analytics.
* Prepare market response models.
* Create a model to identify the individuals who are only likely to respond after receiving a particular ‘treatment’ you decided to give them.

**CHAPTER-4** **BACKGROUND**

In retail and marketing, predictive modeling is a common tool used for targeting and evaluating the response from individuals when an action is taken on. The action is normally refereed to a campaign or offer that is sent out to the customers and the response to model is the likelihood that a specific customer will act on the offer.

Put differently, in traditional response models, the objective is to predict the conditional class probability: [13]

P(Y = 1|X = x)

where the response Y ∈ {0, 1} reflects whether a customer responded positively (i.e. made a purchase) to an action or not (i.e. did not make a purchase). X = (X1, ..., Xp) are the quantitative and qualitative attributes of the customer and x is one observation.

Using traditional response modeling, the resulting classifier can then be used to select what customers to target when sending out campaigns or offers in a marketing purpose. In reality, this is not always the desirable approach to use since the targeted customers are those who are most likely to react positively to the offer after the offer has been sent out. The solution is thus to use a second order approach recognized as uplift modeling.

The original idea behind uplift modeling is to use two separate train sets and test sets, namely one train and test set containing a treatment group and one train and test set containing a control group. The customers in the treatment group are subject to an action whereas the customers in the control group are not. Uplift modeling thus aims at modeling the difference between the conditional class probabilities in the control and treatment group, instead of just modeling one class probability: [13]

P T (Y = 1|X = x) − P C (Y = 1|X = x)

where the superscript T denotes the treatment group, and the superscript C denotes the control group. This method is called Subtraction of Two Model. Each probability in the above equation is estimated using the statistical machine learning methods. If the result is negative, it indicates that the probability that a customer makes a purchase when belonging to the control group is larger than when the customer belongs to the treatment group. This is called a negative effect and is very important to include in the models for being able to investigate how the campaigns are affecting the customers. There also exist other approaches for uplift modeling which directly models the uplift by using one data set instead of two. This data set includes both the treatment data as well as the control data and is split into train and test data sets. The methods for modeling the uplift directly either includes the use of a tree-based method or to use of a class variable transformation.

**CHAPTER-5** **METHODOLOGY**

Uplift models have been called as Differential response analysis, incremental value modeling, and true lift models. Uplift models try to model second-order phenomena as it tries to model the conditional average treatment of two or more different mutually exclusive groups.

The metric depends on company’s product, position, targets etc. We need to segregate the customers based on these metrics because you cannot treat every customer the same way with the same content, same channel, same importance. They will find another option which understands them better. Customers who use your platform have different needs and they have their own different profile. You should adapt your actions depending on that.

Now it’s time to measure one of the most important metric we should closely track: **Customer Lifetime Value.** Calculating Lifetime Value is the easy part. First we need to select a time window. It can be anything like 3, 6, 12, 24 months. By the equation below, we can have Lifetime Value for each customer in that specific time window:

Lifetime Value = Total Gross Revenue – Total Cost

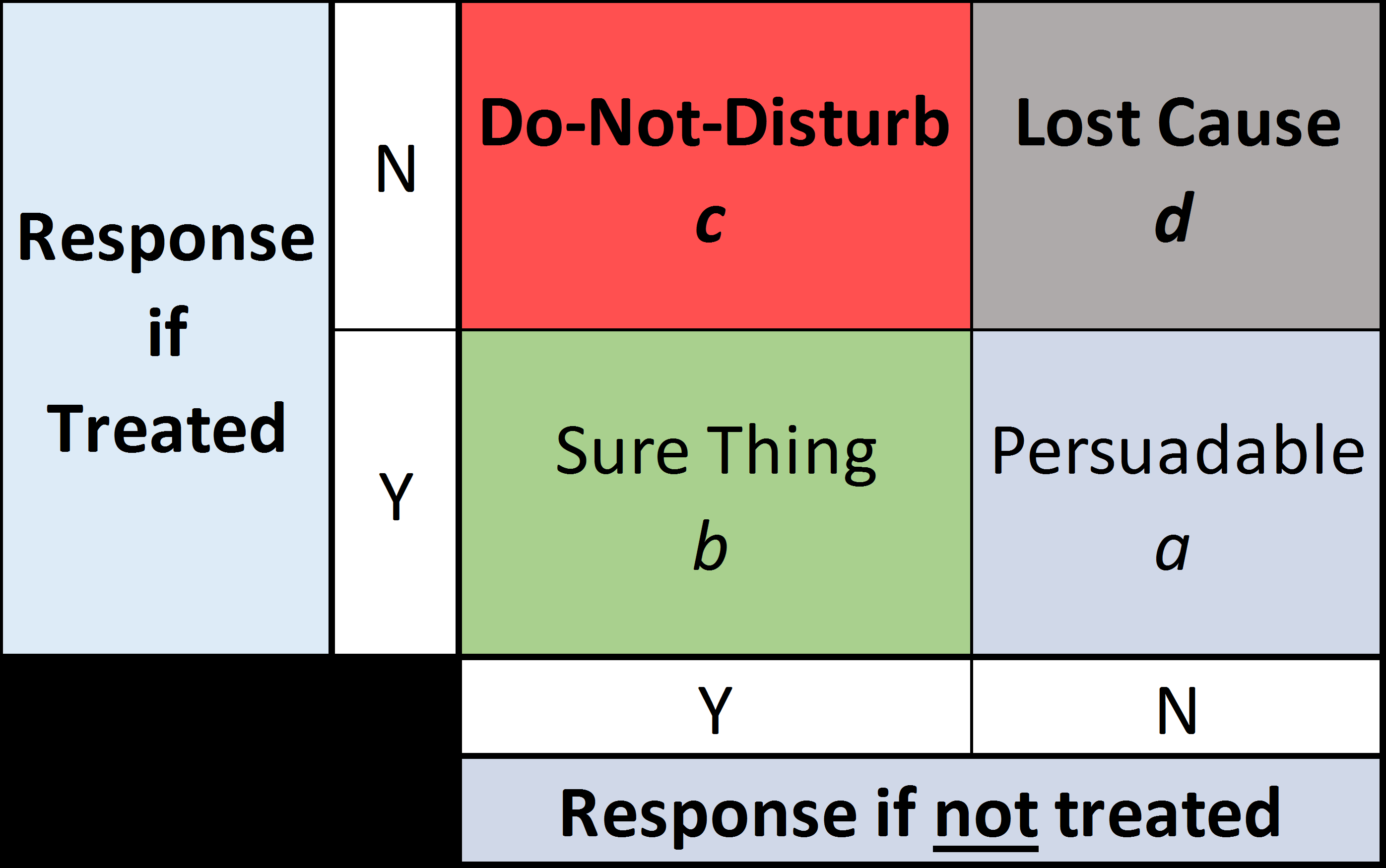
Rate of retaining your customers or Retention Rate is an indication of how good your product market fit (PMF) is. If your PMF is not satisfactory, you should see your customers churning very soon.

Now comes predictive sales, which acts as the benchmark. We can calculate the incremental value of our new actions on top of this benchmark. It can be utilized for planning. We can plan our demand and supply actions by looking at the forecasts. Also, it helps to see where to invest more. It is an excellent guide for planning budgets & targets. It also includes the next purchase day of a customer. We can build strategy on top of that & come up with tactical actions like promotional offer frequency to a customer based on his/her purchasing pattern. For more frequent ones, less/no promotional offers.

The explicit goal in uplift modeling is to model the conditional average treatment effect. The conditional average treatment effect or uplift, estimate the increase of purchase probability given that a customer receives treatment compared to if no treatment is given. By being able to identify which customers who are more likely to purchase before treatment is given, would ideally let a company target a smaller part of the sample and thus reduce marketing costs meanwhile they maintain or even increase their earnings. The uplift model returns a score for each customer, where a higher score means a higher chance of positive outcome. This score should be seen as a priority list of whom to give treatment first. The score is then used to partition the individuals into segments for the treatment group and control group and the uplift is computed per segment.

There are 4 possible outcomes in a binary uplift model [1][2]:

1. The Persuadables: Customers that would be convinced to purchase because they received treatment. They are the optimal customers to target since the response changes from no purchase to a purchase when treatment is given.
2. The Sure Things: Customers that would purchase the product with or without treatment.
3. The Lost Causes: Customers who would not purchase the product with or without treatment.
4. Do-Not-Disturb or the Sleeping Dogs: Customers who are convinced not to purchase when they receive treatment but would have purchased if no treatment were given.



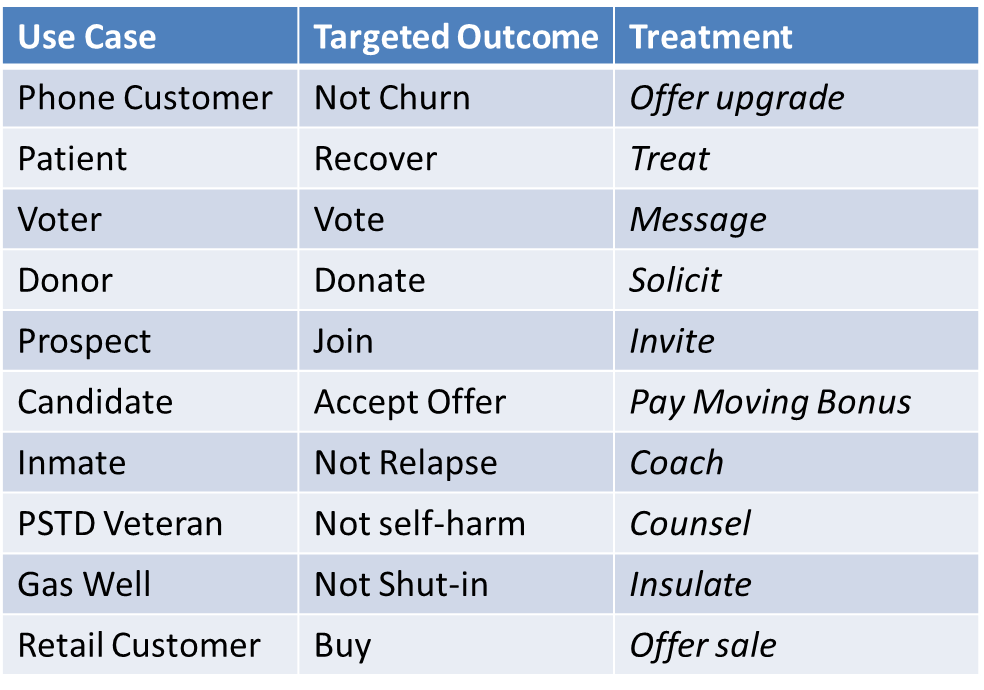
Both the sure things and the lost causes are considered as a waste of money to give treatment to, because the treatment will not affect their response. For sleeping dogs, the treatment has the opposite effect than intended and the customer is lost.

**CHAPTER-6** **IMPLEMENTATION DETAILS**

.

Using **predictive analytics**, information is extracted from existing dataset for determining patterns and predicting the forthcoming trends or outcomes. It uses data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. [11]

There are many **applications of predictive modeling** where the outcome is predicted as advice only to a human decision maker, and no action is directly taken automatically from the model result.  An example is workload prioritization.  For example, in the telecom industry we can predict which customers are most likely to churn (cancel their contracts).  In healthcare we can predict which patients are most likely to recover. For universities or charitable organizations, we can predict which prospective benefactors are most likely to donate. [10]



When the pre-processing step is done, the data is ready to train models for classification. Statistical machine learning methods is used to find relationships between the response and the variables in form of a function, i.e Y = f(X) + ϵ where ϵ is some error term.

The statistical Machine Learning methods: Logistic Regression and Random Forests will be described. These are the classification methods used for uplift modeling in this project. Logistic Regression is the simplest one as it is easy to implement and it is linear. Random Forests is a type of ensemble classifiers and which is tested to see if a more complex classifier yields in a better result. [14]

Logistic Regression and Neural Networks are thus used to make estimations of the probabilities from section 4.2.1 and 4.2.3, i.e. [14]

* For Subtraction of Two Models:

Mˆ U = PˆT (Y = 1|X = x) − PˆC (Y = 1|X = x)

* For the Class Variable Transformation:

Mˆ U = PˆZ (X = x) = 2Pˆ(Z = 1|X = x) – 1

Although it will not tell businesses what will happen in the future, it helps them get to know their individual consumers and understand the trends they follow. This, in turn, helps marketers take necessary, action at the right time, which in turn has a bearing on the future. It focuses on the online behavior of the customer. [11]

Below is the dictionary of the dataset provided by the client:

|  |  |
| --- | --- |
| treatment | HO= Control Group Promo = Test Group |
| zip5 | zip code |
| unique\_key | unique identifier of the record |
| age | age of the customer |
| channel | DM = Direct Mail DM\_EM= Direct Mail & Email |
| state | state |
| inq\_month | months since the last inquiry was made by customer |
| resp | did the customer respond |
| conv | did the customer convert |
| region | region |
| division | division |
| cancel\_reason\_bucket | Policy cancelation reason |
| annual\_premium\_select | annual premium on the policy |
| driver\_cnt\_select | driver count in a household |
| vehicle\_cnt\_select | vehicle count in a household |
| polk\_flag | presence of auto in a household |
| pif\_own\_rent\_cd | O - homeowner R- Renter T - Refused information blank - no information available |
| internet\_sale\_ind | Policy purchased through Internet |
| pif\_risk\_lvl | risk level of customer B - preferred C - Non Preferred D - Reject |
|  | Blank - N/A |

**CHAPTER-7** **PROGRESS TILL DATE AND REMAINING WORK**

The progress on the project till date is as follows:

* **Prerequisite training for the project**

1. Basic practices related to working with SQL, Python & its various libraries, supervised & unsupervised Machine Learning algorithms
2. Excel where the main focus was on preparing Pivot tables, different types of graphs & make conclusions depending on the metrics & corresponding results from the dataset.
3. Statistics which help us to study the distribution of data and significance of the variables by noting p value.

* **Work done on project till date.**

1. Understanding all the metrics in the dataset provided by the client.
2. Made pivot tables for the dataset to get a better idea of what all data in each metric is there & in what frequency. Also checked if any correlations are there or not to prevent redundancy.
3. All the columns have been imputed with their mean
4. All the categorical data are encoded into numeric
5. Filtering out the invalid entries
6. Feature Engineering

The remaining work is as follows:

* Segment & categorize the customers according to their responses.
* Decide treatments for different categories based on the output from the model.
* Prepare market response model & predict future sales.
* Compare the different treatments to different customers using A/B testing

**CHAPTER-8 REFERENCES**

1. <https://medium.datadriveninvestor.com/simple-machine-learning-techniques-to-improve-your-marketing-strategy-demystifying-uplift-models-dc4fb3f927a2>
2. <https://www.researchgate.net/publication/270217235_Identifying_Individuals_Who_Are_Truly_Impacted_by_Treatment_Introduction_to_Recent_Advances_in_Uplift_Modeling>
3. <https://towardsdatascience.com/tagged/data-driven-growth>
4. <https://www.ugamsolutions.com/>
5. <https://www.investopedia.com/terms/p/predictive-modeling.asp#:~:text=Predictive%20modeling%20is%20the%20process,possibly%20happen%20in%20the%20future%3F%22>
6. <https://www.investopedia.com/terms/d/datamining.asp#:~:text=Data%20mining%20is%20a%20process,increase%20sales%20and%20decrease%20costs>.
7. <http://www.pardot.com/blog/abcs-ab-testing/>
8. <https://books.google.com/books?id=fW_9BV5Wpf8C&q=%22regression+analysis%22>
9. <http://www.predictiveanalyticsworld.com/machinelearningtimes/uplift-modeling-making-predictive-models-actionable/8578/>
10. <http://www.predictiveanalyticsworld.com/patimes/wp-content/uploads/2017/03/Mike-Thurber-Graphic-1.png>
11. <https://www.fingent.com/blog/data-mining-and-predictive-analytics-know-the-difference/>
12. <https://www.diva-portal.org/smash/get/diva2:1386196/FULLTEXT01.pdf>
13. <https://doi.org/10.1007/s10618-014-0383-9>
14. Trevor Hastie Robert Tibshirani, Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. First Edition. CRC Press, 2015. ISBN: 9781498712163

CHAPTER-9 PROJECT DETAILS

|  |  |  |  |
| --- | --- | --- | --- |
| *Student Details* | | | |
| **Student Name** | Sai Raghu Teja Davuluri | | |
| Register Number | 170907024 | Section / Roll No | D/21 |
| Email Address | teja12899@gmail.com | Phone No (M) | 8639123779 |
|  | | | |
| *Project Details* | | | |
| **Project Title** | Uplift Model | | |
| Project Duration | 4 months | Date of reporting | 02 March 2021 |
|  |  | | |
| *Organization Details* | | | |
| **Organization Name** | Ugam Solutions Pvt. Ltd. | | |
| Full postal address with pin code | Ground floor, H2 Block, Mountain Ash, Manayata Tech Park, Nagavara, Bengaluru, Karnataka 560045, India | | |
| Website address | https://www.ugamsolutions.com/ | | |
|  |  | | |
| *External Guide Details* | | | |
| **Name of Guide** | Aakash Sharma | | |
| Designation | Senior Manager, AS Dept | | |
| Full contact address with pin code | Ugam Solutions Private Limited | Registered Office: Ground Floor, H2 Block, Mountain Ash, Manyata Tech Park, Nagawara, Bangalore 560045, India | | |
| Email address | aakash.sharma@ugamsolutions.com | Phone No (M) |  |
|  |  | | |
| *Internal Guide Details* | | | |
| **Faculty Name** | Dr. Goutham Simha GD | | |
| Full contact address with pin code | Dept of Electronics and Communication & Engg, Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA | | |
| Email address | goutham.simha@manipal.edu | | |