



Master Thesis

Development of R Shiny Web App for Incremental
Response Modeling of Data

Master of Engineering in Information Technology

Submitted by

Akshay Laddha (1098383)

in

FRANKFURT UNIVERSITY OF APPLIED SCIENCE
FACULTY OF COMPUTER SCIENCE

Supervisor at Frankfurt University of Applied Science

Prof. Dr. Christina Andersson

Prof. Dr. Andreas Orth

DECLARATION OF AUTHORSHIP

I assure that this Thesis is single handed composition of my original Research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

Signed :

Date :

Acknowledgements

I would first like to thank my thesis adviser Prof. Dr. Christina Andersson of Frankfurt University of Applied Sciences, for guiding me and for her support and assistance throughout the master thesis. Without her passionate participation and input, the development of the application could not have been successful.

I would also like to acknowledge Prof. Dr. Andreas Orth of Frankfurt University of Applied Sciences as the second reader of this thesis, and I am gratefully indebted for his very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and my elder brother and sister in law for their support through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Akshay Laddha

Abstract

John Wanamaker known as father of modern marketing has once said *“I know half of the money I spend on Advertising is wasted, but the trouble is I don’t know which half.”*

Incremental Response Modelling also known as Uplift Modelling or Net lift Modelling is a technique which is used to identify those customers who are likely to respond positively when they are targeted by any marketing action and not likely to respond if not approached. Incremental Response modelling has applications in different areas of customer retention and demand generation. The other big successes have come in the area of cross-sell and up-sell, particularly of high-value financial products. Here, purchase rates are often low, and the overall incremental impact of campaigns is often small. Uplift modelling often allows dramatic reduction in the volumes targeted while losing virtually no sales. In some case, where negative effects are present, incremental sales actually increase despite a lower targeting volume.

The main purpose of this web application is to develop a user interface where user can select input data, perform data pre processing tasks, perform incremental response modeling on the data and download the results for further analysis. These findings could be used in future campaigns to find out the customers whom should the company target for next marketing action.

Contents

Acknowledgements	ii
Abstract	iii
Contents	iv
List of Figures	vi
List of Tables	vii
Abbreviations	viii
1 Introduction	1
1.1 Thesis Structure	2
2 Background and Motivation	3
2.1 Obama's 2012 Presidential Election Campaign	4
2.2 Introduction to Predictive modelling	4
2.3 How Incremental Response model is different from Traditional Predictive model?	6
3 Incremental Response Modelling	8
3.1 Literature of Incremental Response Modelling	10
3.2 Difference score from two separate models	11
3.3 Difference score from combined model	13
4 Data Pre-Processing	15
4.1 Missing Values	16
4.2 Special Values	17
4.3 Outliers	17
5 Variable Pre Screening	19
5.1 Problems which can be solved using Variable Pre Screening	20
5.2 Net Information Value Method (NIV)	20
5.2.1 Weight of evidence (WOE) and Information Value (IV)	20
5.2.2 Weight of Evidence (WOE)	21
5.2.3 Information Value(IV)	23
5.2.4 Extension of WOE/IV Analysis for Incremental Response Modeling	24

5.2.5	Penalized Net Information Value	25
6	Incremental Response Score Modeling	26
6.1	Difference Score from Two Separate Models	27
6.2	Difference Score from One Combined Model	27
6.3	Model Diagnostics	28
7	R Shiny Web App for Incremental Response Modeling	31
7.1	What is R Shiny?	31
7.2	Web App for ICM	33
7.2.1	How to use this Web App	33
7.2.2	Data Selection	34
7.2.3	Data Pre-Processing	34
7.2.4	Variable pre screening and Data Preparation	35
7.2.5	Variable Selection using NIV Analysis	35
7.2.6	Incremental Response Model Building	36
7.3	Results	36
7.3.1	Two Model Approach	37
7.3.2	Combined or Single Model Approach	38
8	Conclusion and Future Work	39
8.1	Conclusion	39
8.2	Future Scope and Improvements	41
9	Appendix	42
9.1	Web Application Screen-shots	42
9.1.1	Tab 1: Select and View Data	42
9.1.2	Tab 2: Data Pre-Processing	43
9.1.3	Tab 3: Variable Pre Screening	43
9.1.4	Tab 4: Data Preparation	44
9.1.5	Tab 5: Variable Selection for Modeling	45
9.1.6	Tab 6: Incremental Response Modeling	45
	Bibliography	48

List of Figures

2.1	Categories of Customers	6
2.2	Incremental Response Group in the Data	7
3.1	Traditional Response Modeling Process	8
3.2	Basic Incremental Response Model	10
3.3	Two Separate model method	12
3.4	Combined Model Method	13
5.1	Basic Example fo WOE IV Calculations	23
6.1	Sample of Model Diagnosis for IRM	29
7.1	How to Start with Shiny	32
7.2	How to use Shiny App	33
7.3	Model Evaluation for Two Model Approach	37
7.4	Model Evaluation for Two Model Approach	37
7.5	Model Evaluation for Combined Model Approach	38
7.6	Model Evaluation for Two Model Approach	38
9.1	Data Selection Tab for Shiny App	42
9.2	Data Processing tab of Shiny App	43
9.3	Variable Pre Screening tab of Shiny App	43
9.4	Data Preparation Tab of Shiny App-1	44
9.5	Data Preparation Tab of Shiny App-2	44
9.6	NIV Analysis for Variable Selection	45
9.7	Two model Approach	45
9.8	Two model Approach-2	46
9.9	Combined model Approach	46
9.10	Combined model Approach-2	47

List of Tables

Abbreviations

IV	Information Value
WOE	Weight of Evidence
NIV	Net Information Value
NWOE	Net Weight of Evidence
WOE_T	Weight of Evidence Treatment Group
WOE_C	Weight of Evidence Control Group
IRM	Incremental Response Modeling
AIC	Akaike information criterion
IR	Incremental Response
UI	User Interface
DT	Data Table
CRAN	Comprehensive R Archive Network
HTML	Hyper Text Markup Language
CSS	Cascading Style Sheet

Chapter 1

Introduction

When a customer is not completely unknown, company/banks can directly approach them to advertise their product. For example a telecomm operator company can advertise their product by SMS or Voice calls. The online shopping giant like amazon can send a promotional email to the customers promoting their service.

However these marketing efforts are only valuable if the customers respond positively to the promotional offer or service. It is always good to target the customers after proper analysis rather than targeting any random customers. Properly targeted customers provides greater returns than randomly targeted customers. Also, it should also be noted that the customers can get annoyed by the marketing action and this could result in loss of these customers.

Usually response models targets those customers who are likely to purchase if they are included in the marketing campaigns. However out of these customers, there are some who were likely to take the offer irrespective of the marketing action. Hence spending amount on such customers is a waste of money and man power. However uplift models or Incremental response models target those customers who are likely to respond only when they are subjected to the marketing action.

1.1 Thesis Structure

This thesis is divided into the following chapters:

- Chapter 2 provides introduction about Incremental Response Modelling and its motivation
- Chapter 3 deals with the detailed overview of Incremental Response Modelling
- Chapter 4 explains the test data used in this Thesis Work.
- Chapter 5 provides various data cleaning tasks performed prior to incremental response modelling
- Chapter 6 provides details about the variable pre-screening step using Net information value method (NIV)
- Chapter 7 gives information about the two different methods used to calculate the incremental score
- Chapter 8 deals with the R Shiny Web App implementation for creating a user interface for Incremental Response Modelling
- Chapter 9, finally conclusion for this thesis work. It also highlights some approaches which can be used to improve the developed application

Chapter 2

Background and Motivation

Let us consider a brand/company/bank running a campaign of calling their customers to promote their newly launched product. Let's say that the company approached 1000 customers and out of these 1000 customers, 700 actually bought the promoted product. But out of these 700 customers, 400 customers would have bought the product even if the product was not marketed to them explicitly. Which means there are actually 300 customers whom should the company/bank should have targeted instead of targeting all 1000 customers. Also there are some customers who usually don't like the company calling/mailling them with offers. The marketing action applied on such customers usually backfires. Because this could cause the customers not to buy the product even though they were planning to buy it. Incremental Response Modelling is all about finding out such customers who are likely to respond positive if they are subjected to the marketing action but are not likely to respond if not targeted.

One of the major task in the field of marketing is how wisely one make use of the money spent on marketing to inflict a positive change in customer's behaviour. Although all these efforts of marketing are done just to change the uncertain future behaviour of the customer. But building a proper model to predict the course of marketing action and the people to target with that marketing action helps a lot in saving money spent on marketing. A very popular example of uplift modelling is Obama's 2012 Presidential Election Campaign. It is explained in brief in below section.

2.1 Obama's 2012 Presidential Election Campaign

Back in 2010, many of the analysts looking ahead to 2012 Presidential Elections started predicting loss for President Obama. However, Obama's campaign managers knew the fact that Obama will not be able to win depending upon only Obama's loyal voter base. This encouraged them to hire team of 50 analytics experts including Data Scientist to predict which voters will respond positively to political contact such as call, flyer, door knock or TV advertisement. This team used Uplift modelling aka Incremental Response modelling to predict or find such base of voters. Daniel Porter, who is now a partner at analytics services provider BlueLabs in Washington, D.C., led a team of eight data analysts who were charged with determining which voters the campaign should focus on. They targeted those voters who were planning to vote for the Republican Nominee Mitt Romney but might decide to vote for Obama if they were contacted.

This resulted in significant rise in Obama's voter base and thus helped President Obama to successfully achieve win in the 2012 Presidential Elections. [1]

2.2 Introduction to Predictive modelling

Predictive modelling has been playing an important role in customer management. As the time passed on, there has been a huge growth in the use of predictive techniques by companies to find their important customers. Predictive modelling can be defined in different types in brief as:

- **Penetration models or lookalike models:** This type of modelling looks for those customers who have already bought the product. In the next step to find out their target customers, this model looks for the customers who has similar characteristics to those who have already bought. Thus the name Lookalike models.
- **Purchase Models:** These type of modelling techniques involve the historical data. They follow the similar approach like penetration models but they look only in the historical data. And depending on the history, they predict the customers who are likely to buy the product.

- **Response models:** These type of modelling techniques try to characterize the customers who have purchased the product based on their response to the marketing actions such as a promotional email or a phone call. Generally the identification of the responders is based on the time interval between promotional offer and the actual purchase. Such response models are considered as more powerful and effective than Penetration and Purchase models because it actually try to connect the customer's behaviour with the marketing action. [2]

But Response Model also have a significant drawback. It did not account those customers who did not receive that marketing action. This may lead to targeting those customers who would have bought the product irrespective of marketing action. Even worse that this may lead to target those customers who might not want to spend more money because they received the marketing action. This is the common problem in Telecommunication industry where retention offer at the end of customer's contract might turn negative for the company. Because at this point the customers get the reminder that they can switch their carrier. Incremental response modelling helps in overcoming these drawbacks to an extent. How it is different from Traditional Predictive modelling is explained in brief in next section.

2.3 How Incremental Response model is different from Traditional Predictive model?

Depending on above discussion, the customers can be categorized in following four categories.





		Will buy if not received an offer	
		No	Yes
Buy if received an offer	Yes	Persuadable 	Sure Things 
	No	Lost Cause 	Do not disturb 

FIGURE 2.1: Categories of Customers

- Buy only if targeted with marketing action: Persuadable
- Buy regardless of marketing action: Sure Things
- Never buy regardless of marketing action: Lost Cause
- Will not buy if targeted with marketing action: Do not disturb or sleeping dogs

Clearly, as a company one will never want to disturb the sleeping dogs category. Also, company don't want to invest money on those customers who will buy regardless of marketing actions i.e. Sure Things. This leaves us with only one category to look after. The Persuadable!!!

Traditional predictive models (Response Models) usually divide customers into following two categories:

- Respondents (1)
- Non Respondents (0)

Thus the data analyst often asked with the two questions: a. Who are the Persuadable?
b. Which marketing activity will influence more?

To answer these questions, one surely need a different approach than traditional response models. The net effect, a marketing action has done on the customer with those customers who were not targeted with it need to be modeled. Such models which uses the approach of net lift are called Incremental response models or uplift models. They directly models the increment in the response probability due to marketing action.

Traditional response models have long been used to predict who is likely to respond to an action, such as a marketing incentive. In such models, all customers in a group receive the promotion, their responses are recorded, and a predictive model is built to separate likely responders from those unlikely to respond. This is done through a number of predictive modelling methods such as decision trees, neural networks, or regression models.

An incremental response model uses two randomly selected data sets, which are called control and treatment groups. The treatment group is targeted with the marketing action, but the control group does not. Figure shows where the incremental response can be identified in the process of modelling.

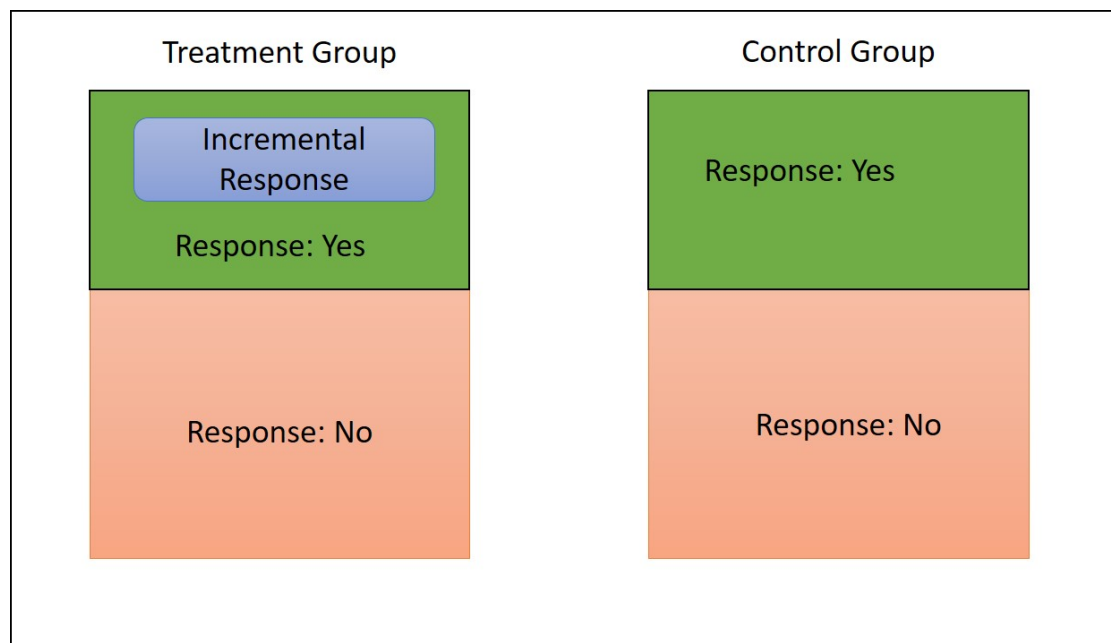


FIGURE 2.2: Incremental Response Group in the Data

Chapter 3

Incremental Response Modelling

This section provides a more detailed overview of Incremental Response Modelling.

Traditional Response models are built on the sample of the customer's data set where every row in the dataset represents customer's information. This information generally describes the characteristics of customer's behavior towards the marketing action. In penetration models, generally historical information about the purchase (or any other response variable) is used. Whereas in response models, information about all the customers who were subjected to marketing action is used or all the customers who are selected as a sample data are targeted with the marketing action. After finishing the marketing campaign, customer's response is recorded into a variable. This dataset is then used to build a prediction model. Once the model is built, it is applied to the whole customer database to find out the customers with highest probability of purchasing the product. This traditional response modelling process is illustrated in below figure

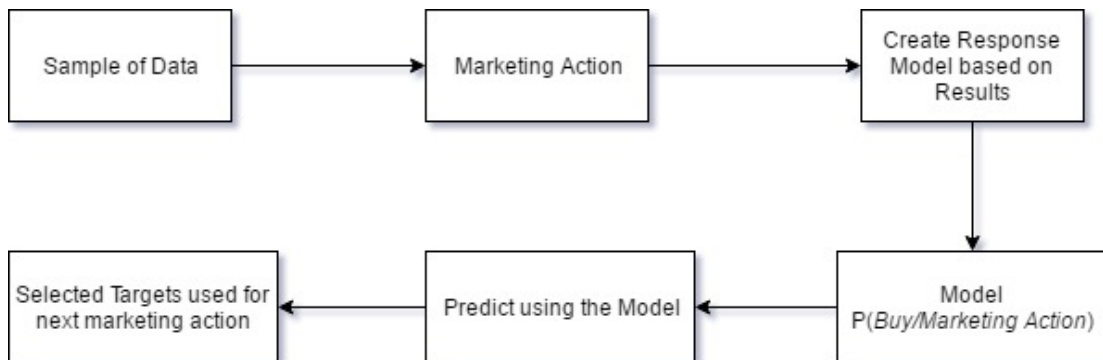


FIGURE 3.1: Traditional Response Modeling Process

However in general the customers are divided into 4 categories:

- Responded *because of* the action
- Responded *regardless of* the action
- Did not respond because action had *no impact* (Unnecessary costs
- Did not respond because action had *negative impact* (Customer got annoyed with the marketing or customer changed his mind after getting the offer)

Traditional response models are not designed to distinguish between the above four categories of customers. They treat customers in only two possible categories. One who responded positively and other who responded negatively. Thus the response models predicts by using conditional class probability.

$$P(\text{response}/\text{Treatment})$$

Traditional response models score customers based on their likelihood to purchase so that marketing campaigns can be targeted towards those customers who will maximize the response rate. Although correlation may exist between the response rate and marketing incentive, traditional predictive models fail to address if the response rate was caused by the marketing campaign, because they cannot distinguish between customers who would have responded positively regardless of the marketing offer and those who responded positively only because of the offer. Clearly company only want to target the later group to measure the extra revenue they generate.

Whereas Incremental Response Models predicts the customer by using the change in customer's behaviour because of the marketing action.

$$P(\text{response}/\text{Treatment}) - P(\text{response}/\text{No Treatment})$$

The basic Incremental Response Model is explained in below block diagram:

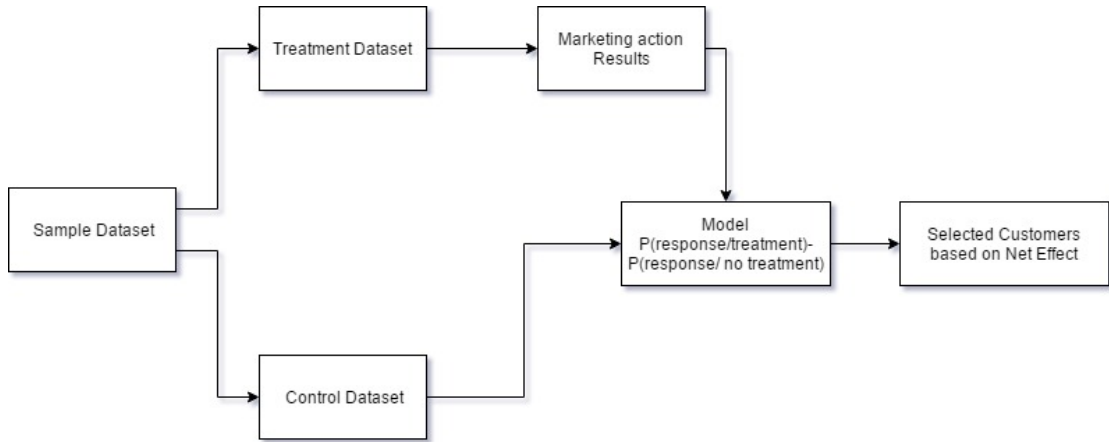


FIGURE 3.2: Basic Incremental Response Model

3.1 Literature of Incremental Response Modelling

Although uplift models aka incremental response models have a very significant results to improve the way companies look at their customers, the problem of incremental response modelling has received a very less research.

The very first paper which explicitly discussed the incremental models was *Differential Response Analysis: modeling true response by isolating the effect of a single action* by N. J. Radcliffe and P. D. Surry in 1999.[3] After this a detailed description of their decision tree algorithm was published in a white paper *Real-World Uplift Modelling with Significance-Based Uplift Trees*[2]

In 2002, Victor S.Y. Lo who is Vice President of Modeling at Fidelity Investments also wrote a research paper *The True Lift Model - A Novel Data Mining Approach to Response Modeling in Database Marketing*. In this paper he explains the new approach to improve current response modelling techniques with the use of True net lift models. In the same year, Behram Hansotia and Brad Rukstales published a research paper on the topic *Incremental value modeling*.[4]

In incremental response modeling, one of the method to calculate the net lift is to find out the differences. Difference models brings the difference of predicted outcome from Treatment group and from the control group. These difference values are known as difference scores. These score are then ranked and divided into number of bins in descending order of ranked difference scores and the customers in top bin are considered

as the probable responders to the marketing action. This predictive model can be built in two different ways:

- Difference score from two separate models
- Difference score from a combined model

The sample Dataset is first divided into two separate datasets as Treatment Data D_T and Control Data D_C . Treatment group contains those customers who were targeted with the marketing action and Control group with those who were not targeted with any. Let us denote the *RESPONSE* variable as Y and all the explanatory variables as X_i . Number of observation in Treatment group as n_T and in Control group as n_C .

$$D_T = \left\{ X_i, Y_i \right\}_{i=1}^{n_T} \quad (3.1)$$

$$D_C = \left\{ X_i, Y_i \right\}_{i=1}^{n_C} \quad (3.2)$$

$$D = D_T \cup D_C = \left\{ X_i, Y_i \right\}_{i=1}^n, n = n_T + n_C \quad (3.3)$$

Thus a simple linear model can be formed as:

$$Y = X\beta + \epsilon \quad (3.4)$$

Now lets discuss these two methods in brief:

3.2 Difference score from two separate models

The steps involved in this approach is explained in below block diagram:

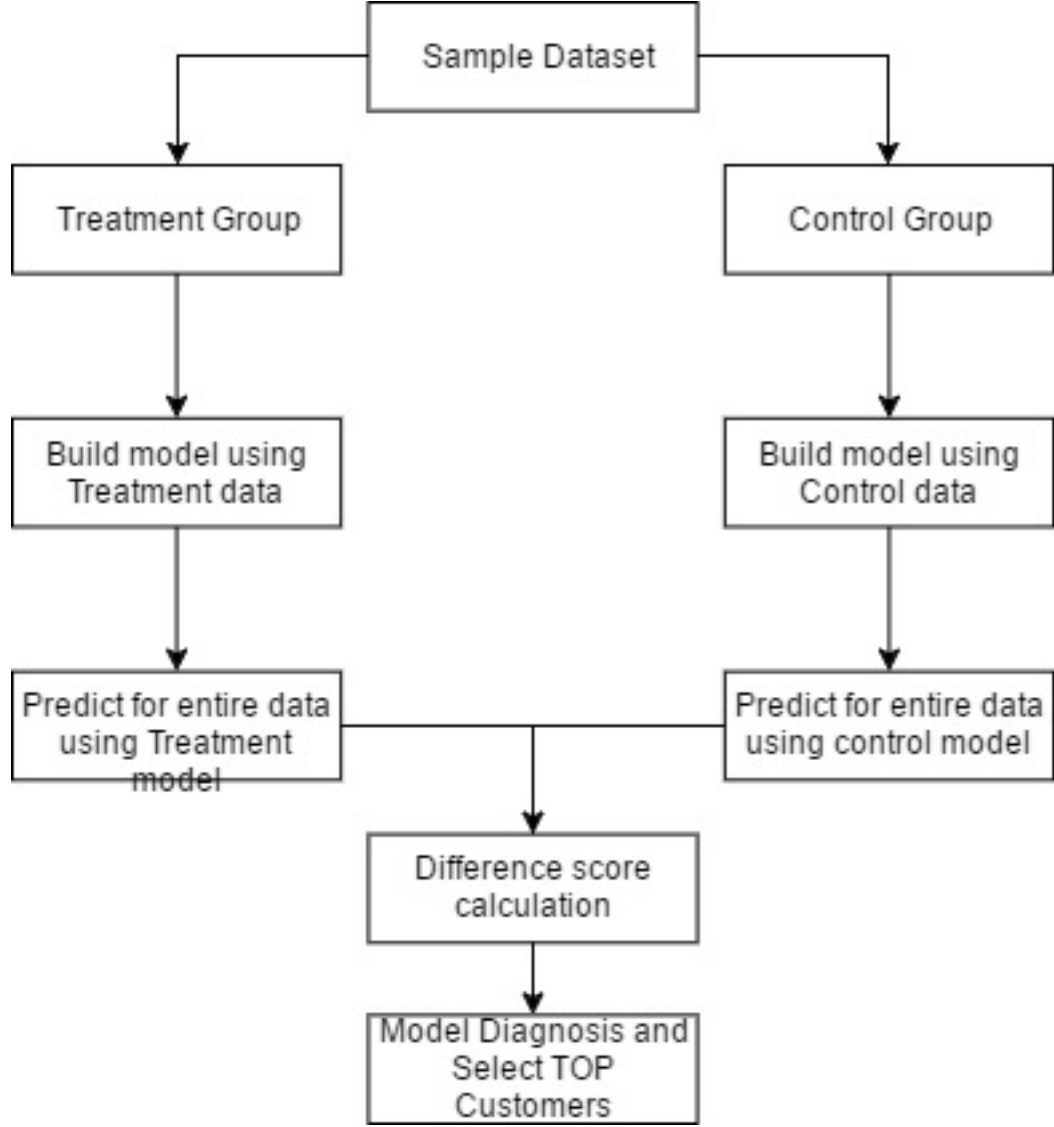


FIGURE 3.3: Two Separate model method

Once the data is divided into Treatment and Control group, two separate models are built on D_T and D_C respectively.

$$\hat{Y}_T = X_T \hat{\beta}_T \quad (3.5)$$

$$\hat{Y}_C = X_C \hat{\beta}_C \quad (3.6)$$

These two models are used to predict response value for the entire data ($D = D_T + D_C$) and also the probabilities are calculated for entire data. Thus the probability value for all n records with both models is calculated. Then, the difference score can be calculated

as follows:[5]

$$\hat{D}S_i = (\hat{Y}_T - \hat{Y}_C)_i \text{ for } i=1,2,\dots,n \quad (3.7)$$

- Pros
 - Uses standard logistic regression modelling techniques.
 - Easy to implement and maintain.
- Cons
 - Does not fit the target correctly.
 - Introduces modeling error twice.
 - Sensitive to predictive variable selection and parameter estimation

3.3 Difference score from combined model

Steps involved in this approach is explained in below block diagram.

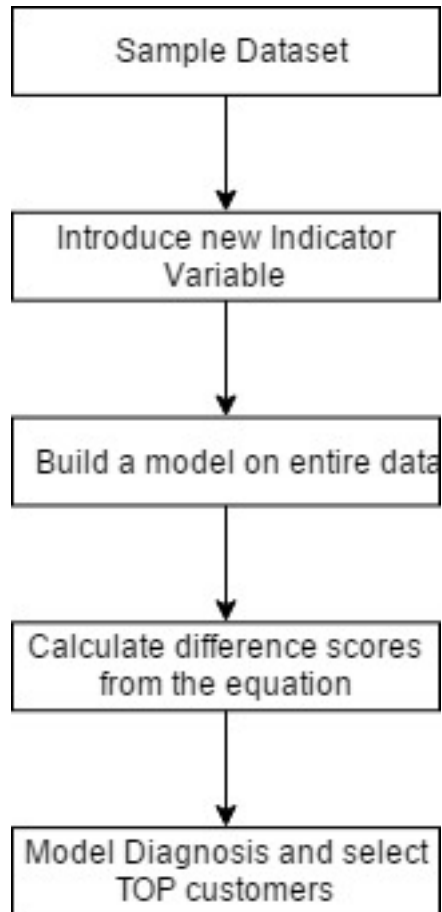


FIGURE 3.4: Combined Model Method

This model was introduced by Victor S.Y. Lo (2002). He introduced a new variable, T_i , where $T_i = 1$ for D_T and $T_i = 0$ for D_C . The model was built on entire data i.e. ($D=D_T \cup D_C$):[5]

$$Y = X\beta + T\gamma + (XT)\psi + \epsilon \quad (3.8)$$

Then the difference scores are calculated from the equation 3.8, as follows:

$$\hat{Y}_T = X\hat{\beta} + \hat{\gamma} + (X)\hat{\psi} \quad (3.9)$$

$$\hat{Y}_C = X\hat{\beta} \quad (3.10)$$

$$\hat{DS}_i = \hat{Y}_T - \hat{Y}_C \quad (3.11)$$

In simpler words, difference scores can be calculated as:

$$Score = P(response = X, treatment=1) - P(response = X, treatment=0)$$

- Pros
 - Uses standard logistic regression modelling techniques.
 - Better robustness compared to two model approach.
- Cons
 - Does not fit the target correctly.
 - Introduces model complexity due to assumptions of non linearity.

Customers who have a positive value of \hat{DS}_i are initially considered as the incremental responders as a result of the promotion campaign. However, to determine the final set of incremental responders, further analyses must be performed with the ranked difference scores in decreasing order. This is explained in chapter 7.

Chapter 4

Data Pre-Processing

Data pre processing is very important step in Data mining. But it is often neglected. It is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. A low quality data leads to a low quality of mining results.

In real world, data can be,

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- Noisy: containing errors or outliers
- Inconsistent: containing discrepancies in codes or names

The quality of data affects the Data Mining results. To improve the quality of data and consequently the quality of mining results, data is pre processed so as to improve the efficiency and the ease of mining process. Data pre processing methods are divided into following categories:

- **Data Cleaning:** Fill in missing values, Smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data Integration:** Using multiple databases, data cubes, or files.

- **Data Transformation:** Normalization and Aggregation.
- **Data Reduction:** Reducing the volume but producing the same or similar analytic results.

Consistent data is technically correct data which is fit for statistical analysis. But in practice data can contain missing values, special values, (obvious) errors and outliers, which need to be either removed, corrected or imputed. Following sections describe the data cleaning tasks performed in this thesis work:

4.1 Missing Values

A missing value, represented by NA in R, is a placeholder for a data of which the type is known but its value isn't. Therefore, it is impossible to perform statistical analysis on data where one or more values in the data are missing. One may choose to either omit elements from a dataset that contain missing values or to impute a value, but missing value is something to be dealt with prior to any analysis.

In this thesis work, hot deck imputation is used to treat missing values. In hot deck imputation, missing values are imputed by copying values from similar records in the same dataset.

$$\hat{x}_i = x_j$$

where x_j values are taken from observed values. Hot-deck imputation can be applied to numerical as well as categorical data but is only effective when there are enough records to look for similarities.

In the user interface (R Shiny Web App) where user can upload any dataset, user can process the missing values if its needed. The full working functionality of R Shiny web app is explained in chapter 8.

4.2 Special Values

Generally, in a raw data, numeric values are often endowed with several formalized special values such as $\pm Inf$ and NaN where Inf stands for Infinity value and NaN stands for Not a Number value.

Thus any calculations which involves such special values results in a special value. However in practical scenario, such special values does not weight much significance. Hence it is recommended to treat such special values prior to modeling process. The number of records with NaN and infinity values can be checked using *is.nan()* and *is.infinite()* functions.

4.3 Outliers

There are numerous definitions available for outliers. A general definition by Barnett and Lewis defines an outlier in data as an observation or set of observations which appear to be inconsistent with that set of data. Detection of outliers is the most important step in data cleaning because of following reasons:

- An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).
- In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, user typically do not want to simply delete the outlying observation. However, if the data contains significant outliers, one must consider the use of robust statistical techniques.

In this master thesis work, outliers are detected using Cook's distance method.

Cook's Distance is commonly used estimate of influence of a data point.

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE} \left[\frac{h_i}{(1 - h_i)^2} \right] \quad (4.1)$$

The significant observation about Cook's Distance is that it depends on both the residual, $e_i = (y_i - \hat{y}_i)^2$ (in the first term), and the leverage, h_{ii} (in the second term). That is, both the x value and the y value of the data set play a role in the calculation of Cook's distance. The measurement is a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance.

- A general rule of thumb is that observations with a Cook's Distance of more than 3 times the mean, μ , is a possible outlier.
- An alternative interpretation is to investigate any point over $\frac{4}{n}$, where n is the number of observations.
- An alternative (but slightly more technical) way to interpret cook's distance is to find the potential outlier's percentile value using the F-distribution. A percentile of over 50 indicates a highly influential point.

Below is a code snippet from the thesis work, used to detect the outliers in the data:

Chapter 5

Variable Pre Screening

Once the data is prepared for modeling, next step is to do the variable pre screening. This step is described in brief in this chapter.

Variable selection also known as feature selection is one of the very important step in data mining. Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, where as feature selection methods include and exclude attributes present in the data without changing them.

Feature selection step carries a meaningful magnitude, but it mostly acts as a filter, where one remove out those variables which are not useful for the modeling process[6]

Variable selection or variable pre screening is used for following reasons:

- It makes the models more simpler and easier to interpret.
- It reduces the time to train the model.
- It reduces the possibility of probable overfitting.
- It helps in avoiding *curse of dimensionality*. [7]

Feature selection techniques are different from feature extraction. Feature extraction is the process of creating new features from functions of the original features, whereas feature selection is a method to select subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples

and selecting these many features/variables will increase the training time for the model and also increases the complexity of the model. Thus selecting only the features which are important for the model improves the model quality.

5.1 Problems which can be solved using Variable Pre Screening

Selecting variables helps to build more accurate predictive models. It helps in selecting those variables which carry more importance for the model.

Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model.

Fewer attributes is desirable because it reduces the complexity of the model, and a simpler model is simpler to understand and explain.

The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data.[5]

5.2 Net Information Value Method (NIV)

In incremental response modeling, the difference of score between Treatment and Control sets i.e. the net lift is calculated. Such calculation increases the possibility of overfitting of the model and thus can hamper the performance of the predictive model. In practice, the incremental effect is relatively small, hence variable pre screening gets more importance when dealing with incremental response modeling. Net Information Value method was first used with Incremental Response Models by Larsen in 2010.

5.2.1 Weight of evidence (WOE) and Information Value (IV)

Weight of evidence and Information value provides an explanatory analysis on variable pre screening for binary classifiers. WOE and IV have been used extensively in credit

risk modeling for many years. However this method has not got much research under data modeling. Some of the features of WOE and IV analysis are as follows:[8]

- It provides each variable's individual contribution for the prediction.
- It provides the user with linear and non linear relationship of the variables.
- It provides the rank to the variable depending upon their predictive strength.
- It compares the strength of continuous and categorical variables without creating dummy variables.
- It evaluates the predictive power of missing values and Seamlessly handle missing values without imputation.

In this thesis, Information Package is used which provides easy way to calculate WOE and IV for Uplift models.[9]

Let's assume to have a binary dependent variable Y and a set of predictive variables X_1, \dots, X_p . WOE and IV play two distinct roles when analyzing data:

- WOE describes the relationship between a predictive variable and a binary target variable.
- IV measures the strength of that relationship.

5.2.2 Weight of Evidence (WOE)

WOE/IV is based on following relationship:

$$\log \frac{P(Y = 1|X_j)}{P(Y = 0|X_j)} = \underbrace{\log \frac{P(Y = 1)}{P(Y = 0)}}_{\text{sample log-odds}} + \underbrace{\log \frac{f(X_j|Y = 1)}{f(X_j|Y = 0)}}_{\text{WOE}}, \quad (5.1)$$

where $f(x_j|Y)$ denotes the conditional probability density function (or a discrete probability distribution if X_j is categorical).

This relationship says that the conditional logit of $P(Y=1, \text{ given } X_j)$, can be written as the overall log-odds (i.e., the “intercept”) plus the log-density ratio – also known as the weight of evidence.

Weight of Evidence reveals the predictive power of the independent variable in relation to the dependent variable. As described in above section, concept of WOE is been derived from Credit Risk Scoring world. Hence it is generally described in terms of good and bad customers. In our case, good customers are those customers who have responded to the marketing action and bad are those who haven't.

$$WOE = \ln \left[\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}} \right] \quad (5.2)$$

where,

Distribution of Goods- % of good customers in particular group

Distribution of Bads- % of bad customers in particular group

The value of WoE will be 0 if the odds of *Distribution Goods* / *Distribution Bads* is equal to 1. If the *Distribution Bads* in a group is greater than the *Distribution Goods*, the odds ratio will be less than 1 and the WoE will be a negative number; if the *Distribution of Goods* is greater than the *Distribution Bads* in a group, the WoE value will be a positive number.

Estimating WOE General steps which are followed in calculating WOE are:

- The continues variables are split into number of bins depending on the distribution of data.
- Calculate the number of Goods and Bads in all the bins
- Calculate % of Goods and % of Bads in each bin.
- Calculate *WOE* using above formula for all bins.

Note: For a categorical variables, data is not required to be split. Rest all the steps can be followed as it is. Simple example of these steps is shown in below screen-shot:

Range	Bins	Bads	Goods	% of Bads	% of Goods	WOE	IV
0-50	1	150	54	5%	6%	-0.2364	0.0029
51-100	2	365	75	11%	10%	0.1222	0.0016
101-150	3	524	85	16%	12%	0.3466	0.0166
151-200	4	425	45	13%	15%	-0.1310	0.0024
201-250	5	225	54	7%	16%	-0.8242	0.0739
251-300	6	257	65	8%	16%	-0.7095	0.0587
301-350	7	855	95	27%	12%	0.7862	0.1140
351-400	8	256	23	8%	7%	0.1584	0.0018
>401	9	154	21	5%	6%	-0.2589	0.0037
	Total	3211	517				0.2756

FIGURE 5.1: Basic Example fo WOE IV Calculations

If B_1, \dots, B_k denote the bins for X_j , the WOE for X_j for bin i can be written as:

$$WOE_{ij} = \log \frac{P(X_j \in B_i | Y = 1)}{P(X_j \in B_i | Y = 0)} \quad (5.3)$$

5.2.3 Information Value(IV)

Estimating Information Value is the most important step in WOE/IV analysis. WOE describes the relationship between predictors and target, whereas IV gives the strength of that relationship. Thus the predictor variables which are more important for model building process are derived as a result. In the book "Credit Risk Scorecards" by Naeem Siddiqi,[10] author has provided with some guidelines to follow while selecting variables using IV.

IV Rules of thumb for evaluating the strength of predictor:

- Less than 0.02 \rightarrow Unpredictive.
- 0.02 - 0.1 \rightarrow Weak
- 0.1 - 0.3 \rightarrow Medium
- $>0.3 \rightarrow$ Strong

5.2.4 Extension of WOE/IV Analysis for Incremental Response Modeling

As discussed in previous chapter, IRM is used in direct marketing programs where a treatment group receives an offer or is targeted with marketing action and a control group which is not subjected to such actions. Then find out the net lift between behaviour of Treatment set customers and Control group customers. And use this model to predict which customers can be our persuadables. Thus while building such model, using just WOE/IV analysis to select variables is not effective. Since considering only log odds of Response=1 will not be as effective. One must consider the log odds of Response=1 for the Treatment group versus the control group. This leads to an extension to WOE/IV Analysis. It is known as Net Weight of Evidence and Net Information value.(NWOE and NIV).

Net information value can be calculated as the difference of WOE's for the Treatment group(WOE_t) and Control group(WOE_c). In other words, it can be formulated as,

$$NWOE_j = \log \frac{f(x_j|Y=1)_t f(x_j|Y=0)_c}{f(x_j|Y=1)_c f(x_j|Y=0)_t} \quad (5.4)$$

Or in simple terms it is calculated as,

$$NWOE = \log \frac{P(X = x_i|Y=1)_T / (P(X = x_i|Y=0)_T)}{P(X = x_i|Y=1)_C / (P(X = x_i|Y=0)_C)} \quad (5.5)$$

where, $P(X|Y)_T$ and $P(X|Y)_C$ are the conditional probabilities for Treatment and Control group, respectively. NWOE is estimated using the log odd ratio of response odd for treatment group and response off for control group. Then the Net Information Value (NIV) can be defined as,

$$\begin{aligned} & \int (f(X_j|Y=1)_t f(X_j|Y=0)_c - f(X_j|Y=1)_c f(X_j|Y=0)_t) \\ & \times \log \frac{f(X_j|Y=1)_t f(X_j|Y=0)_c}{f(X_j|Y=1)_c f(X_j|Y=0)_t} dx_j \end{aligned} \quad (5.6)$$

Or in simple terms it is calculated as,

$$\begin{aligned} NIV = \sum_i & (P(X = x_i|Y=1)_T (P(X = x_i|Y=0)_C - \\ & (P(X = x_i|Y=0)_T (P(X = x_i|Y=1)_C) \times NWOE_i \end{aligned} \quad (5.7)$$

These NIV values are arranged in descending order of their ranks. Finally variables are chosen either by selecting those above some threshold or by selecting some specific number of variables which tops the list.

5.2.5 Penalized Net Information Value

In general, any predictive model is built on the Training Data and then the model is tested against some new data or test data for checking the predictiveness of the model. Since NIV is calculated using the Training data, there is a possibility that the predictive power of the selected variable drops in test or validation data than training data. Thus the variables selected after NIV analysis lacks in the predictive abilities in validation data. Thus the concept of Penalized Information Values is introduced. In such scenarios, the NIV is adjusted with a penalty term. This penalty term describes the difference of WOE calculated from training data and WOE calculated from validation datasets.

For each bin of predictor variable, NWOE is calculated from Training data as $NWOE_{train}$ and NWOE calculated from Validation Data as $NWOE_{valid}$. Then the penalty is calculated using the difference of these two WOE's denoted as ω .

$$\omega = |NWOE_{train} - NWOE_{valid}| \quad (5.8)$$

$$Penalty = \sum_i (P(X = x_i|Y = 1)_T(P(X = x_i|Y = 0)_C - P(X = x_i|Y = 0)_T(P(X = x_i|Y = 1)_C) \times \omega_i \quad (5.9)$$

$$PenalizedNIV = NIV - Penalty \quad (5.10)$$

This method of penalizing the Net Information Values was introduced by Larsen in 2010. [\[11\]](#)

Chapter 6

Incremental Response Score Modeling

Incremental response modeling, one of the methods to find out persuadables customers is by calculating difference score models. Such models basically calculate the difference between the prediction probability for Treatment group and for Control group. These difference values are called difference scores. The ranked difference scores are then grouped in descending order for model assessment. And the top group is selected as true responders or persuadables customers for the marketing action.

As explained in chapter 3, the incremental response model can be built in two ways. One with two different models on Treatment and Control data and other with combined model on both Treatment and Control data.

Before building the models, an important and required step is to create the required subsets of the data.

The whole data is first divided into Training and Validation subsets. 70% is kept as the Training subset and rest 30% as Validation subset. However, in the WebApp, user is provided with an option to choose the percentage of Training and Validation data from entire data. Depending on the user input, the data is divided into either 80% or 70% or 60% of Training and respective Validation subset

The data is also divided into Treatment and Control groups based on the PROMOTION Value. Treatment group are the ones who received the Promotion i.e. Data with PROMOTION=1 and Control group are the ones who did not received the Promotion i.e. Data with PROMOTION=0

These subsets are then used in the variable pre screening process. NIV Analysis is used to discover the important variables. This is described in chapter 6. Once the variables are selected, they are used to build the logistic models to calculate the difference scores. The two methods for calculating the difference score are described in next sections in brief:

6.1 Difference Score from Two Separate Models

In this method of score modeling, two different logistic models are built separately, one on the Treatment data subset and other on the Control data subset.

$$\hat{Y}_T = X_T \hat{\beta}_T \quad (6.1)$$

$$\hat{Y}_C = X_C \hat{\beta}_C \quad (6.2)$$

These prediction probabilities are the used to calculate difference scores.

$$DS_i = (\hat{Y}_T - \hat{Y}_C)_i \text{ for } i=1,2,\dots,n \quad (6.3)$$

```
%Score_Model <- PredictTreatmentGroup - PredictControlGroup
```

The ranked difference scores are then grouped in descending order for model assessment and the top group is selected as true responders or persuadables customers for the marketing action.

6.2 Difference Score from One Combined Model

This method was introduced by Victor S.Y. Lo in 2002. He introduced the new variable (named as Treatment Indicator) where $T_i = 1$ for Treatment Data and $T_i = 0$ for Control

Data. Once this new variable created, it is used to built the model on the entire set of data ($D = D_T \cup D_C$)

$$Y = X\beta + T\gamma + (XT)\psi + \epsilon \quad (6.4)$$

To build this equation into a glm model, the term (XT) is required to be evaluated before adding it in the equation(formula). Thus the required new columns are calculated by simply multiplying the variables with the new Treatment Indicator variable.

Once these new varaibles are created, the combined model is built on entire set of data. Then the difference scores are calculated using the coefficients estimated by the log model.

$$\hat{Y}_T = X\hat{\beta} + \hat{\gamma} + X\hat{\psi} \quad (6.5)$$

$$\hat{Y}_C = X\hat{\beta} \quad (6.6)$$

$$\hat{DS}_i = \hat{Y}_T - \hat{Y}_C = \hat{\gamma} + X\hat{\psi} \quad (6.7)$$

In simpler terms,

$$Score = P(response|X, treatment = 1) - P(response|X, treatment = 0) \quad (6.8)$$

The ranked difference scores are then grouped in descending order for model assessment and the top group is selected as true responders or persuadables customers for the marketing action.

6.3 Model Diagnostics

In both the methods, at the end, difference scores (\hat{DS}_i) are calculated. These difference scores are ordered in descending order i.e. the customer (record) with highest difference score value at the top and lowest at the bottom. These ranked observation with the decreasing order of scores are then grouped in number of bins. For each bin, the average predicted values are calculated. For two model method, average predicted values are calculated for both Treatment and Control models and the difference of these values is considered as average predicted increment. Following table shows the sample of such model diagnosis. These sample values are taken from a paper published under Data Mining and Text Analytics in SAS Global forum in 2013.[5]

Bin Nr	Predicted Treatment	Predicted Control	Predicted Increment
1	0.731977	0.354541	0.377436
2	0.711739	0.36625	0.345489
3	0.703177	0.375665	0.327512
4	0.693125	0.380453	0.312672
5	0.669399	0.371129	0.29827
6	0.650663	0.367496	0.283167
7	0.625271	0.358865	0.266406
8	0.575347	0.330758	0.244589
9	0.486066	0.27287	0.213196
10	0.311545	0.165989	0.145556

FIGURE 6.1: Sample of Model Diagnosis for IRM

To constitute a good incremental response model, the incremental response model should satisfy below two criteria:

- The top most bin of record should have higher (highest) incremental rate and the bottom most bin of record should have the lower (lowest) incremental rate.
- The incremental rates should decrease monotonically from top bin to bottom bin.

As seen in the above sample model diagnosis table 6.1, it fulfills the criteria and can be considered that the estimated model is reasonably good. None of the bin violates any of the criteria.

*** Please note: The table 6.1 is a sample of Model Diagnosis table. These are not actual values. The actual results of the thesis are provided in chapter 8.*

Chapter 7

R Shiny Web App for Incremental Response Modeling

7.1 What is R Shiny?

Shiny is an open source R package that provides a powerful web framework for building web applications using R. Shiny helps in presenting the data analysis work done in R into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

A web application with R shiny can be summarised as an easy way to make an interactive web page, and that web page can seamlessly interact with R and display R objects (plots, tables, of anything else a user can do in R). It is a new package from RStudio that makes it incredibly easy to build interactive web applications with R. Some of the main features of R Shiny web App development are listed below:

- Build useful web applications with only a few lines of code—no JavaScript required.
- Shiny applications are automatically "live" in the same way that spreadsheets are live. Outputs change instantly as users modify inputs, without requiring a reload of the browser.
- Shiny user interfaces can be built entirely using R, or can be written directly in HTML, CSS, and JavaScript for more flexibility. Works in any R environment (Console R, RGUI for Windows or Mac, ESS, StatET, RStudio, etc.)

- Attractive default UI theme based on Twitter Bootstrap.
- A highly customizable slider widget with built-in support for animation.
- Pre-built output widgets for displaying plots, tables, and printed output of R objects.
- Fast bidirectional communication between the web browser and R using the websockets package.
- Uses a reactive programming model that eliminates messy event handling code, so you can focus on the code that really matters.

Shiny is available on CRAN, so you can install it in the usual way from your R console:

```
install.packages("shiny")
```

Or if you have R studio installed on your system, you can just click on new project and can select Shiny Web Application.

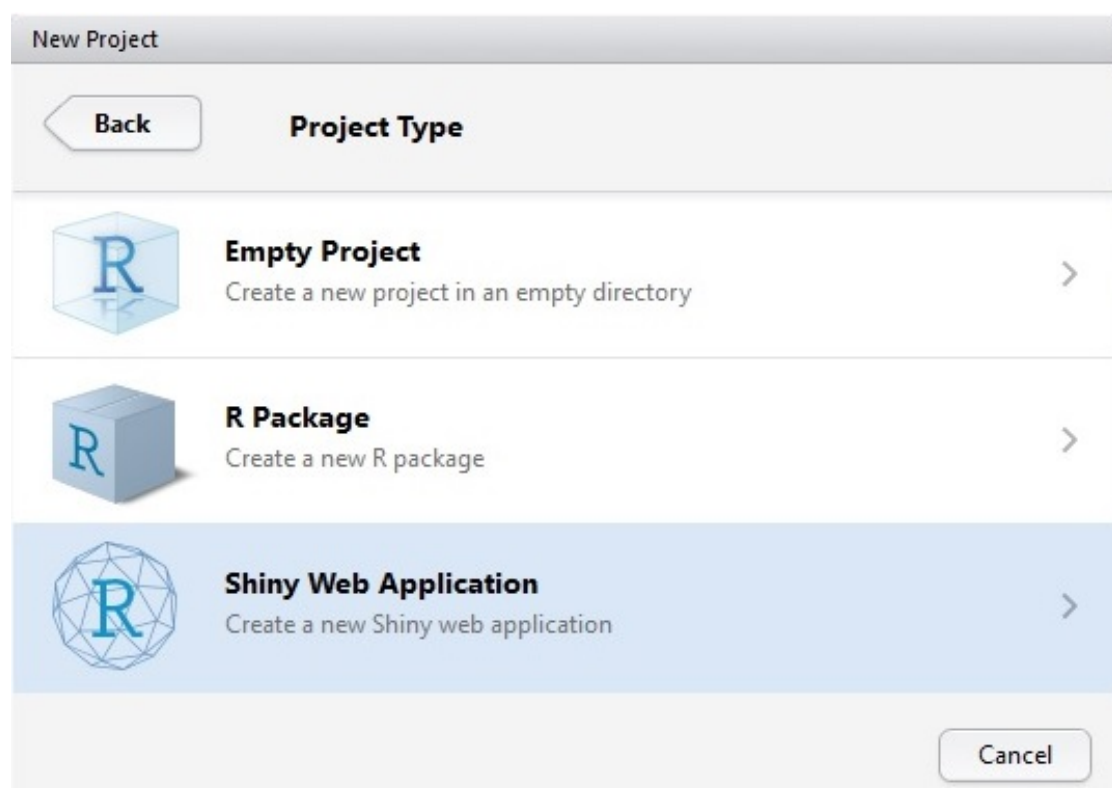


FIGURE 7.1: How to Start with Shiny

7.2 Web App for ICM

The user interface for Incremental Response Modeling consists of following:

- Horizontal Navigation Bar
- File upload option to select data.
- View Data in tabular form with search, order and filter options.
- A independent tab for Analysing loaded data (View summary, NA Check, etc).
- Along with analysing data, a option to preform preprocessing tasks such as missing records prcessing, Outliers detection.
- A separate tab for Data preparation (Treatment-Control Group, Training-Validation Group).
- A separate tab for Variable Pre screening using NIV Analysis.
- A dedicated tab for performing Incremental Response modeling.
- A option to extract top customers as the result of uplift models in an excel file.

7.2.1 How to use this Web App

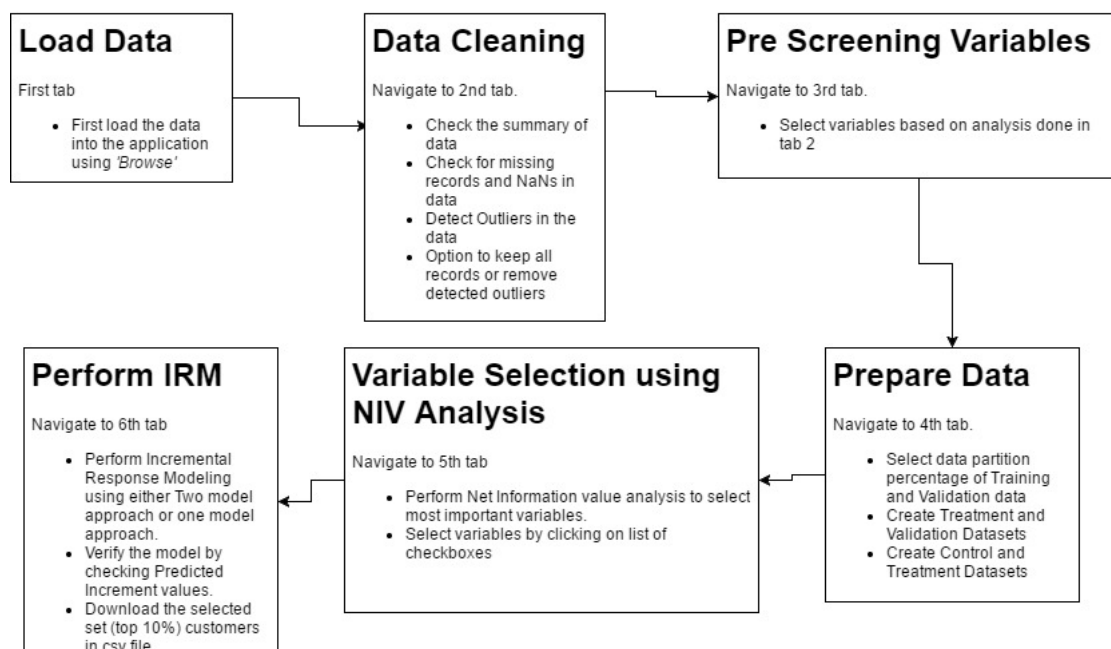


FIGURE 7.2: How to use Shiny App

7.2.2 Data Selection

When user navigates to the web application, Data selection tab will be the first tab. In this tab the user can select the raw data for IRM. The data can be either in SAS7BDAT file or CSV file. However, further options can be added to the selection (such as excel file or any other database connection).

This tab page is divided into sidebar panel and main content panel. In the sidebar panel, the UI for data selection is implemented. Whereas, in main panel, UI for viewing the data is implemented. The data is displayed in main panel using a *DataTableOutput* of DT package. The screen-shot below displays the first tab of the web application.

User can check the uploaded data file in the Data Table in the main panel. This data table widget provides user with many options. User can restrict the number of records displayed by selecting *Show Entries* option to show 10, 25, 50, 100 entries per page. User can also sort the data depending upon specific column in both ascending and descending order by clicking the up-down arrows present next to the column names. Refer to appendix [9.1](#) for screenshot.

Once the data load is complete, the next tab "Data Pre-Processing" will be enabled and user can navigate to next tab.

7.2.3 Data Pre-Processing

After the data is loaded into working environment, first step of data analysis is data pre-processing. In the User interface, I have a dedicated tab for Data Pre processing step. Here user can check the summary of the uploaded data. Also the basic checks like NA and NaNs. Also in the summary output, user can do some more checks like, number of missing records, number of distinct values, mean, median of the values of each individual column. Also user can perform data cleaning tasks on the records with missing values and NAs. A separate button to perform such operations is provided in the UI.

User can also detect the outliers in the dataset. As described in section 5, cook's distance method is used to detect the outliers. When user clicks on the button to detect outliers, the detected number of records is shown with the graphical representation. Along with

this, user is further provided with an option to keep these records and proceed or remove these records and proceed. Also, these records can be downloaded into an excel file for further analysis purpose. Code snippet of Outlier processing is shown below:

Once the outlier records are detected, they are either removed from original data or been kept as it is depending upon user's choice. User also get an option to download these detected outlier records in csv file. The complete code for this is provided in CD along with this thesis document. Please refer to appendix [9.2](#) for screenshot.

7.2.4 Variable pre screening and Data Preparation

After looking at the data summary, user can decide some variables which will not be very useful while building the logistic models. The Variable pre screening tab provides the user with the option of selecting some specific set of variables from the whole dataset.

For eg., ORDER_JAN column in the sample data provided in this thesis contains only two values as 0 and 1. Since this data is not useful since the order values for the month of January for particular customer can not be a a logical value (0 or 1). It should be a significant value when compared to order values in other months.

This tab is also divided into sidebar panel and a main panel. The variables to select are shown in the sidebar panel in the form of list. User can select multiple values using *Ctrl*. Once the variables are selected then it is displayed in the main panel using the *DT DataTableOutput*.

After this user can navigate to next tab "Data Preparation" where user can select the percentage of Training data and Validation data. In this tab user can also divide the data into treatment and control set. Please refer to appendix [9.3](#) , [9.4](#) and [9.5](#) for screenshots.

Once data is divided, the next tab of Variable selection using NIV analysis gets activated.

7.2.5 Variable Selection using NIV Analysis

In this tab user can perform NIV analysis to figure out most important features for the model building process. The detailed explanation about the NIV method is explained in chapter 6.

The group of check-boxes at the top of page shows the top variables from NIV Analysis. User can select the variables with which the Incremental response models will be built. The next tab "Incremental Response Modeling" will be activated and user can navigate to next tab.

7.2.6 Incremental Response Model Building

This is the last step in the process. Once user selects the variables in the previous tab, next step is to build a model.

In this tab, user is provided with the two options to get incremental response results. One option is to follow two model approach and other option is to follow single combined model approach.

On the click of Two model approach button, the user gets the summary of the Treatment and Control model in the dedicated summary tab. To get the detailed result of analysis, user has to navigate to the next tab "Model diagnosis for two model approach". The results of the analysis is explained in next subsection. Similar details can be obtained for combined model approach. User also get the option to download the results into csv file. The option to select all records, top 10%, 20% or 30% of results is also provided. Refer to appendix [9.7](#) , [9.8](#) , [9.9](#) and [9.10](#) for screenshots. The complete *UI.R* and *Server.R* files for the web app are provided in CD along with this thesis document.

7.3 Results

The criteria for the good model is explained in section 7. It can be summarized as follows:

- The top most bin of record should have higher (highest) incremental rate and the bottom most bin of record should have the lower (lowest) incremental rate.
- The incremental rates should decrease monotonically from top bin to bottom bin.

Below are the results of this thesis work:

7.3.1 Two Model Approach

As shown in below picture, results of the two model approach are satisfactory. The top most bin has the highest incremental response rate where as the bottom most bin has lowest. Also the the incremental rates are decreasing monotonically from top bin to bottom bin as shown in the graph.

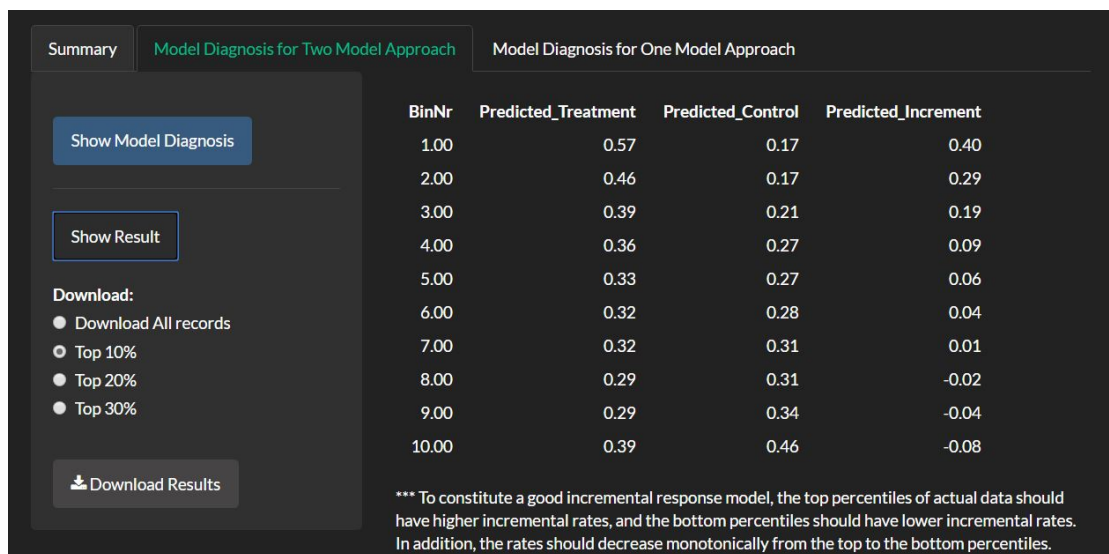


FIGURE 7.3: Model Evaluation for Two Model Approach

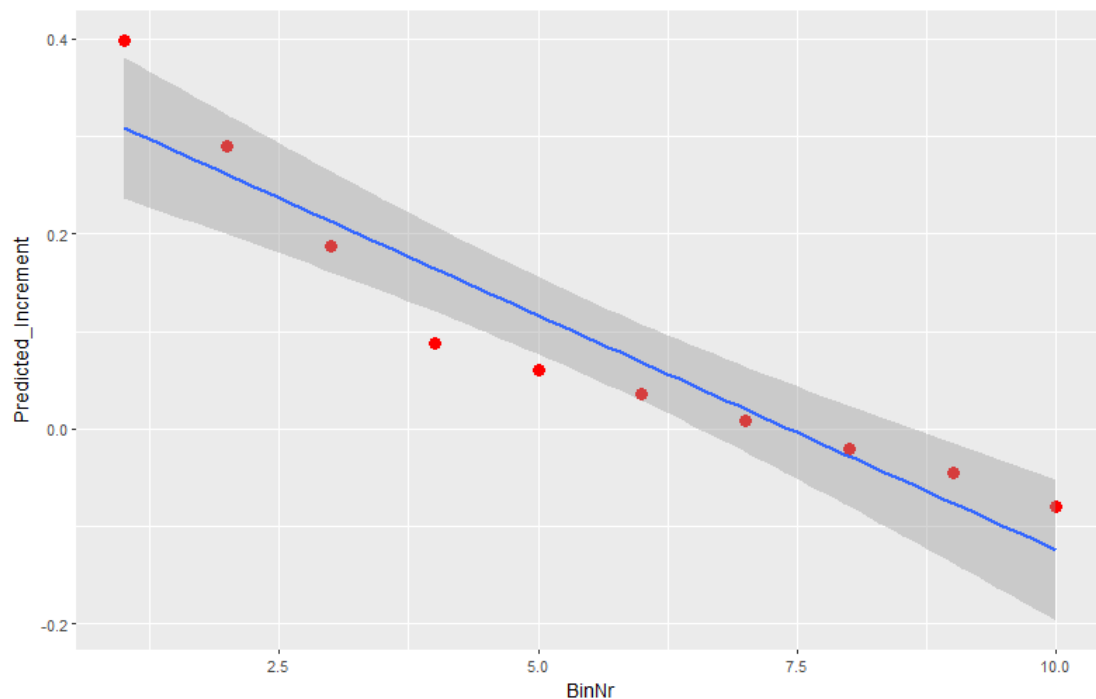


FIGURE 7.4: Model Evaluation for Two Model Approach

7.3.2 Combined or Single Model Approach

As shown in below picture, results of the single model approach are satisfactory but are less accurate compared to two model approach. However it still follows the satisfactory result's criteria, the top most bin has the highest incremental response rate where as the bottom most bin has lowest. Also the the incremental rates are decreasing monotonically from top bin to bottom bin as shown in the graph.

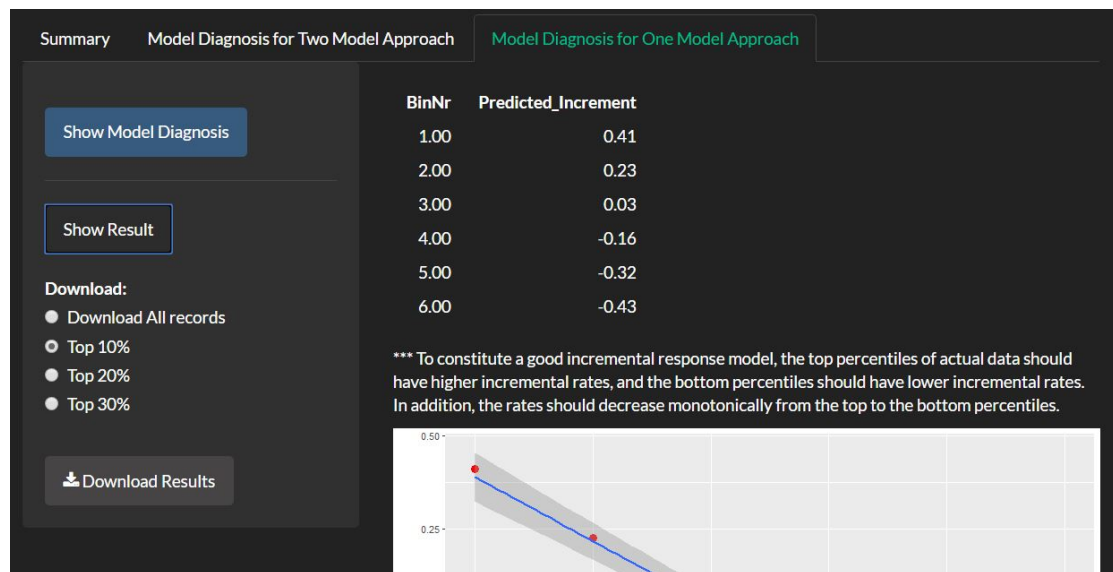


FIGURE 7.5: Model Evaluation for Combined Model Approach

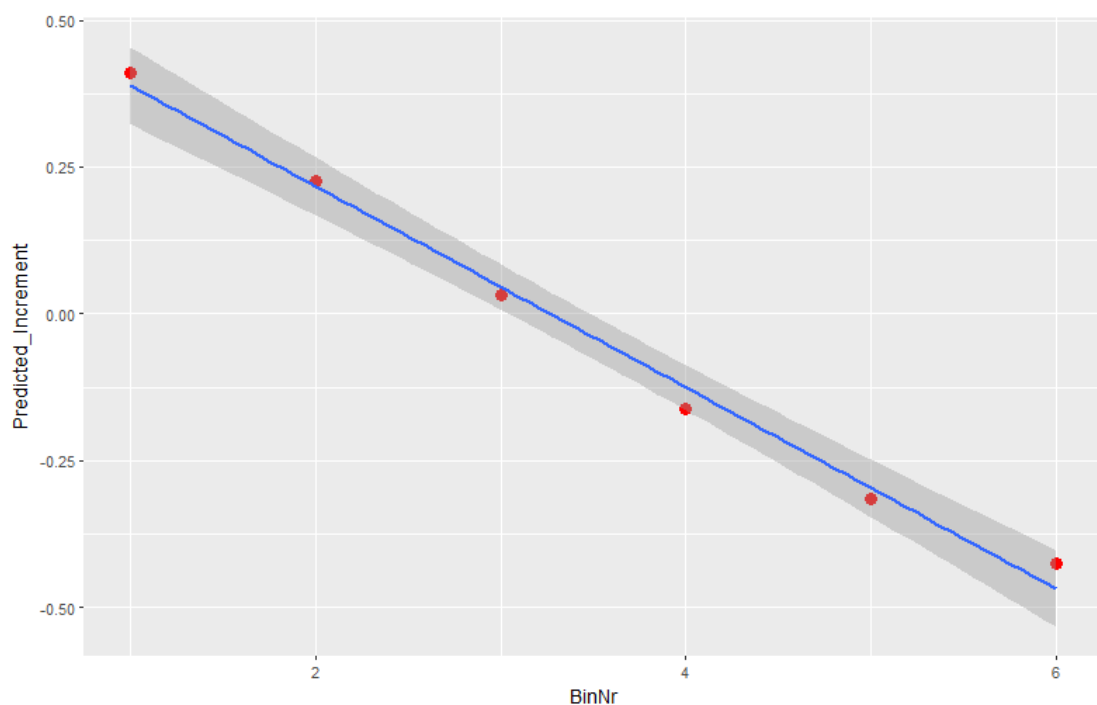


FIGURE 7.6: Model Evaluation for Two Model Approach

Chapter 8

Conclusion and Future Work

8.1 Conclusion

The main aim of this thesis work is to develop a web application in R shiny for building Incremental response model of the data. Theoretically to implement an incremental response model which can be an improvement to the traditional response modeling techniques. IRM has seen a rapid growth in the field of predictive analytics. Although such models are hard to implement but are worth implementing.

The thesis started with defining the goal of thesis work, explaining the background and motivation behind using incremental response models over traditional response models. Later on, it defines the data used and performed data pre processing tasks such as Data cleaning, Variable pre screening, Data partitioning, tasks, etc. Data cleaning tasks such hot deck imputation for missing records, outlier detection using cook's distance are implemented in this thesis work. Variable pre screening is very important step in any model building process. In this thesis work, Penalised Net information Value method for extracting set of important features is also implemented. It further highlights the theoretical concepts of incremental response modeling and their respective implementation in R. It also covers the two different approaches to build an incremental model.

However the ultimate goal of this thesis is to develop an web application which can be used by any user to build IRM on his selected dataset. To achieve this an web application with a User interface styled with Bootstrap Css style is implemented using

R shiny web development. The web app provides user to select the dataset, perform data pre processing tasks, perform IRM and download the results for further analysis.

Based on the thesis work done, the results obtained are quite satisfactory and can be improved with the better quality and quantity of the treatment and control group. The results does selects the top set of customers as the true responders or persuadable customers. These customers can be used to determine the characteristics of the customer to target during the next marketing campaign. Companies saw a tremendous cost saving and revenue increases by using incremental response models. However, this does not come without cost. To build such models we need to have a good data quality and significant control and treatment group to be able to estimate good results. These results are further used in next marketing campaign to save extra marketing costs. Following points could help in deciding when is uplift modeling techniques are worth useful: [2]

- **Existence of valid control group:** To build IR model, the most important thing to have is the valid control group data i.e. the data where the customer is not subjected to marketing offer. Only having the treatment data i.e. the customer data who were subjected to marketing action is not useful alone.
- **Volume of data:** Not only the control data should exist but also it should be large enough to be used for IR model building.
- **Do not disturb type of customers:** IR models are should be applied to those datasets where there is possibility of having customers with negative effect of marketing. This could help the organisation to identify the customers whom they do not want to disturb.
- **Anti-correlated Outcomes:** In some areas, a direct marketing campaign can have better effect on the high spending customers. Thus in such scenarios the sales and the uplift are directly co-related, Thus the traditional response model and uplift models gives the similar type of result. Thus it is not advised to build an uplift model on such customer base.

To conclude, Incremental response modeling is pretty hard nut to crack in the field of predictive analytics. However, the reality though, it should not be a surprise if Incremental Response Modeling or Uplift Modeling is used as a replacement to the

traditional response models. Next section describes the future scope and improvements which can be done in this thesis topic.

8.2 Future Scope and Improvements

The future scope and improvements are listed below:

- *Incremental Sales model:* This thesis covers building the incremental response models. However, this can be further extended to build the Incremental Sales model. The goal of the incremental sales model is to find customers who are likely to spend incrementally when they receive a promotion.
- *Variable Reduction:* Many variable reduction or feature extraction techniques have been proposed in data mining. In this thesis penalised net information value method is implemented. However, one can consider using any other technique such as step-wise regression or interaction detection algorithms such as decision trees
- *Determine volume of Control Data:* The performance of incremental response modeling is mainly dependent on the volume of control data available. In this thesis, a small dataset of 4500 records of control data is used. One can further analyze the most acceptable volume of control data set to use.
- *Multiple treatments:* When more than one treatment (i.e. multiple offers, channels, messages, etc.) are available, the proposed methodology can be easily extended. In this thesis implementation, treatment data contains only yes or no values. However, data could have an separate variable to define the type of treatment given to the customer. But, this could result in more interaction variables generated which will complicate the model building procedure.

Chapter 9

Appendix

9.1 Web Application Screen-shots

9.1.1 Tab 1: Select and View Data

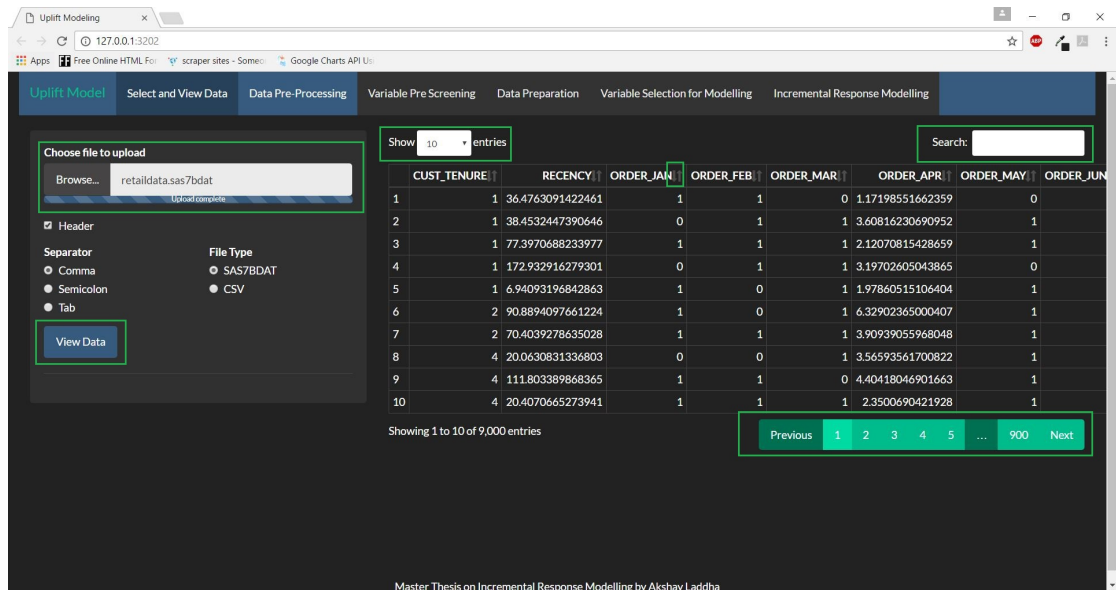


FIGURE 9.1: Data Selection Tab for Shiny App

9.1.2 Tab 2: Data Pre-Processing

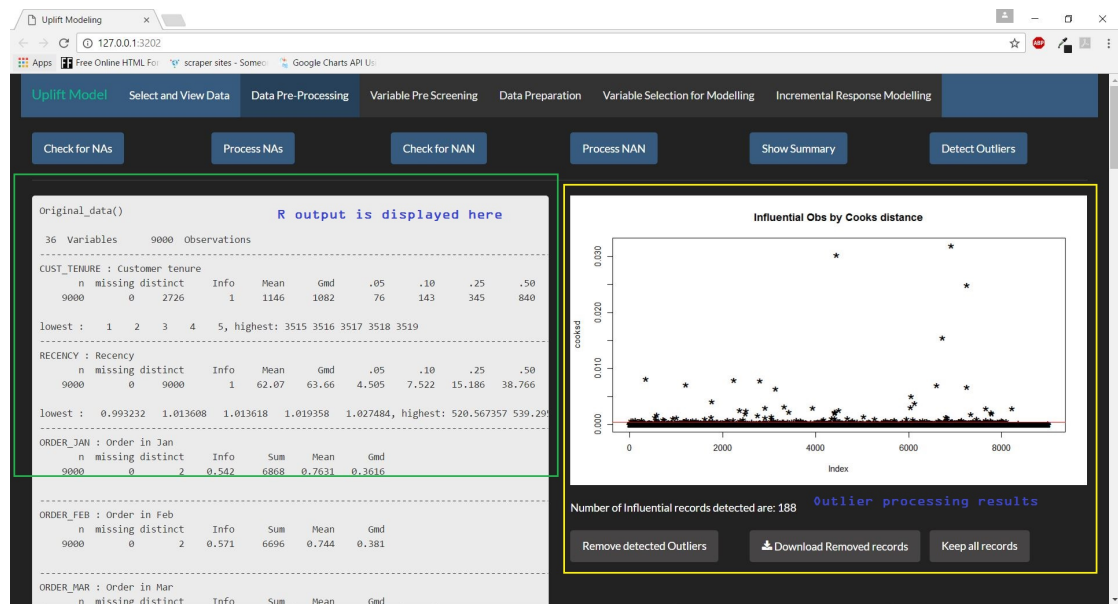


FIGURE 9.2: Data Processing tab of Shiny App

9.1.3 Tab 3: Variable Pre Screening

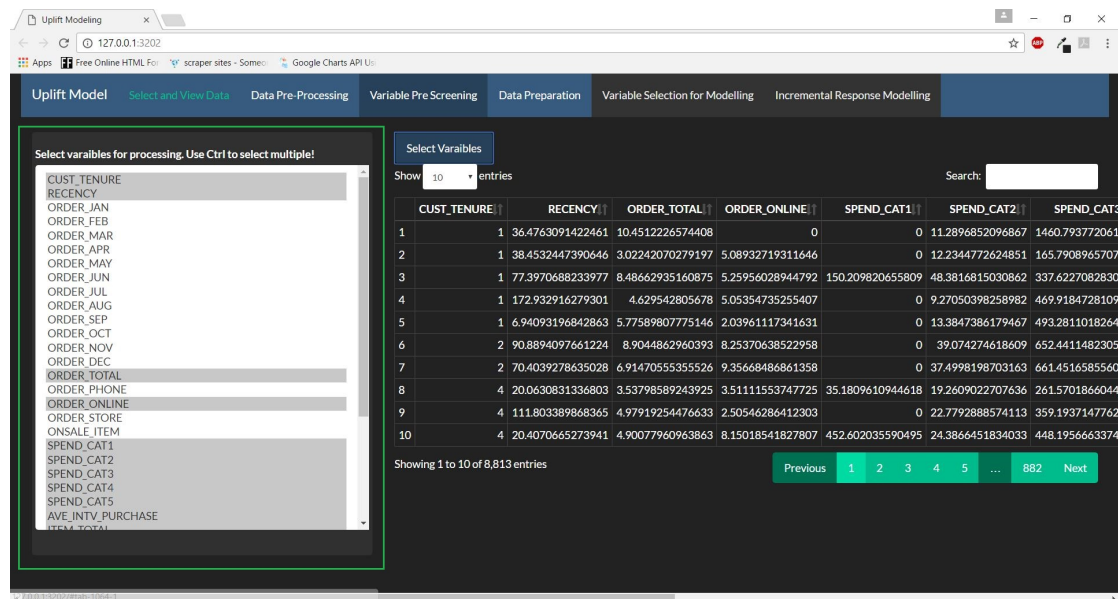


FIGURE 9.3: Variable Pre Screening tab of Shiny App

9.1.4 Tab 4: Data Preparation

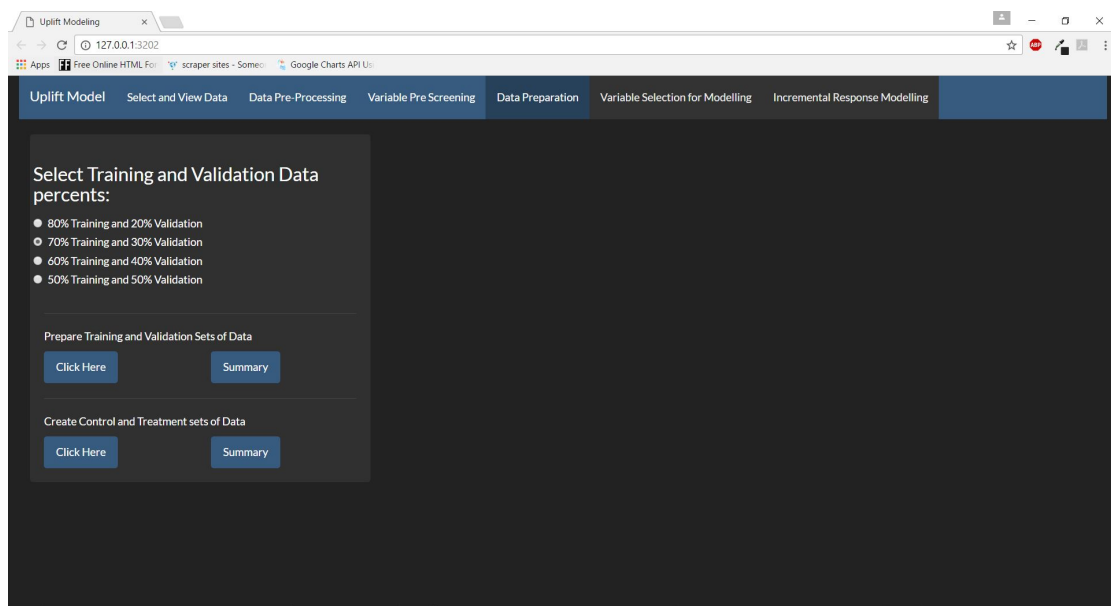


FIGURE 9.4: Data Preparation Tab of Shiny App-1

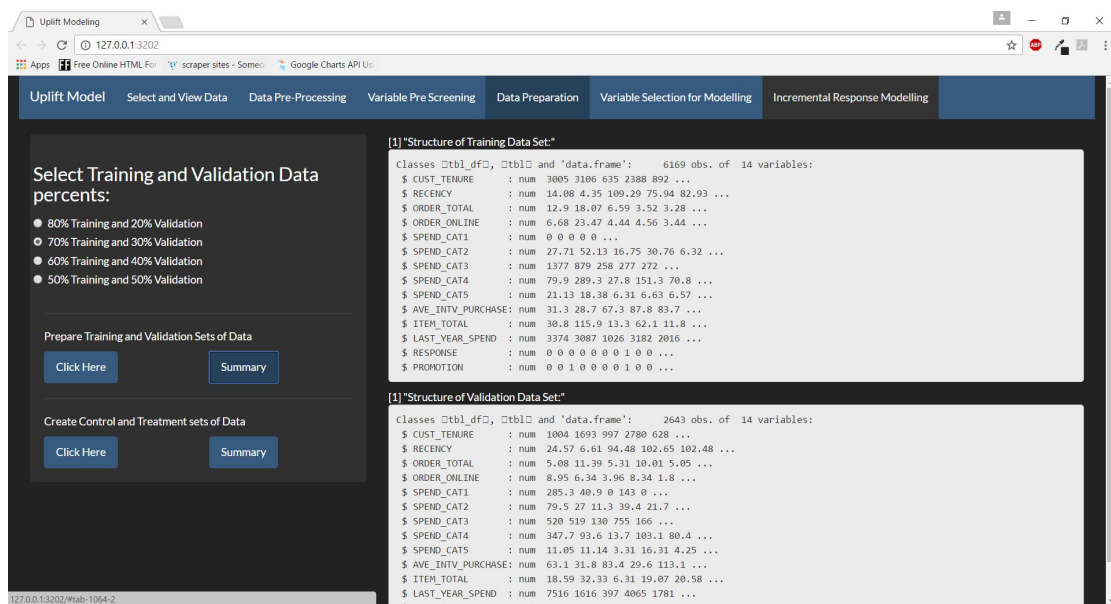


FIGURE 9.5: Data Preparation Tab of Shiny App-2

9.1.5 Tab 5: Variable Selection for Modeling

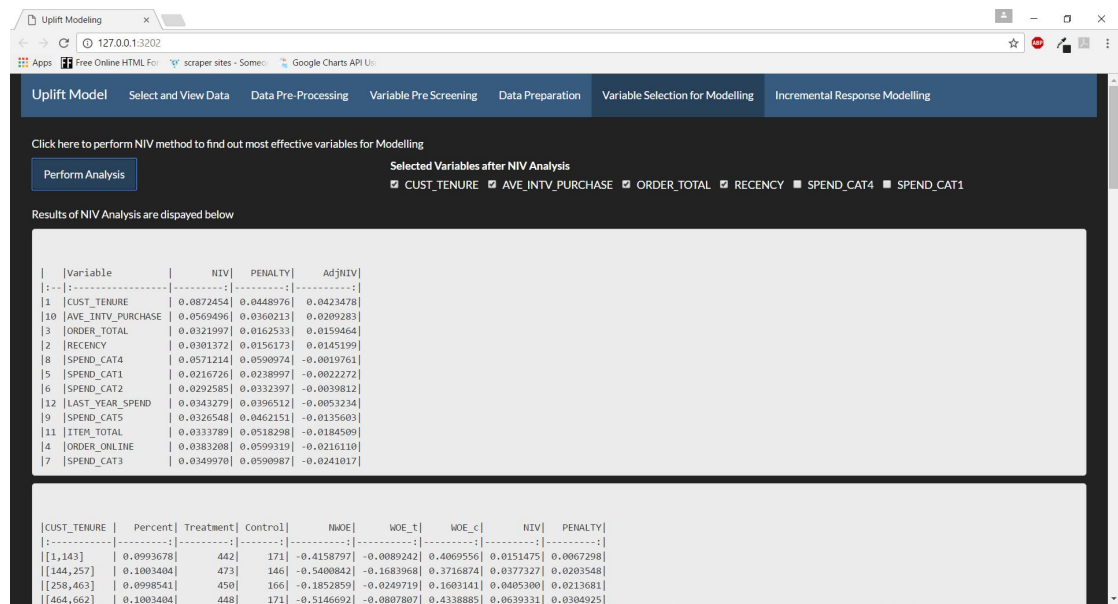


FIGURE 9.6: NIV Analysis for Variable Selection

9.1.6 Tab 6: Incremental Response Modeling

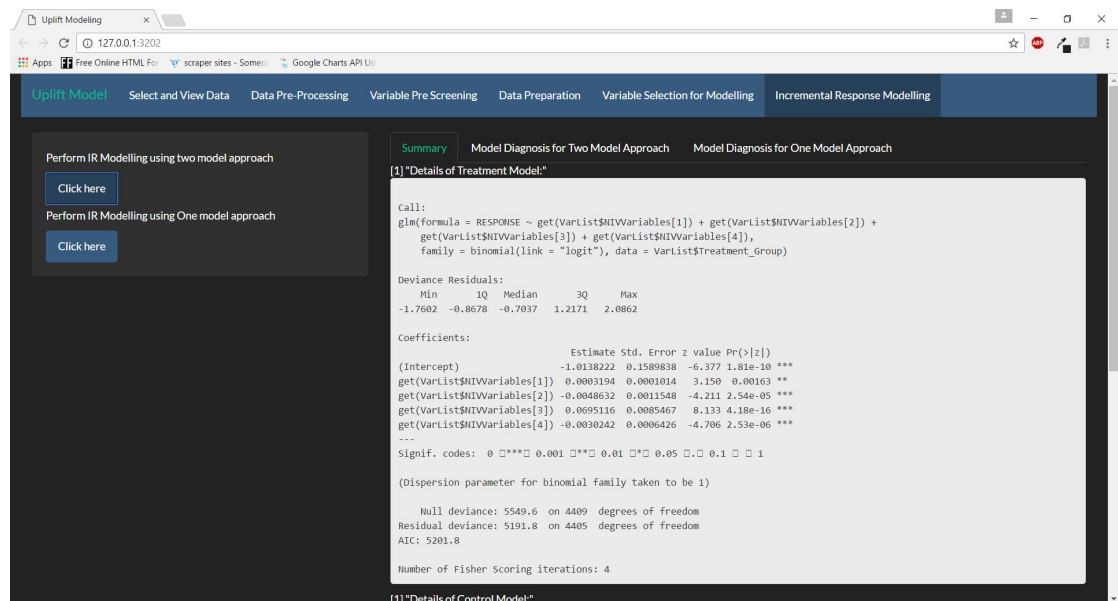


FIGURE 9.7: Two model Approach

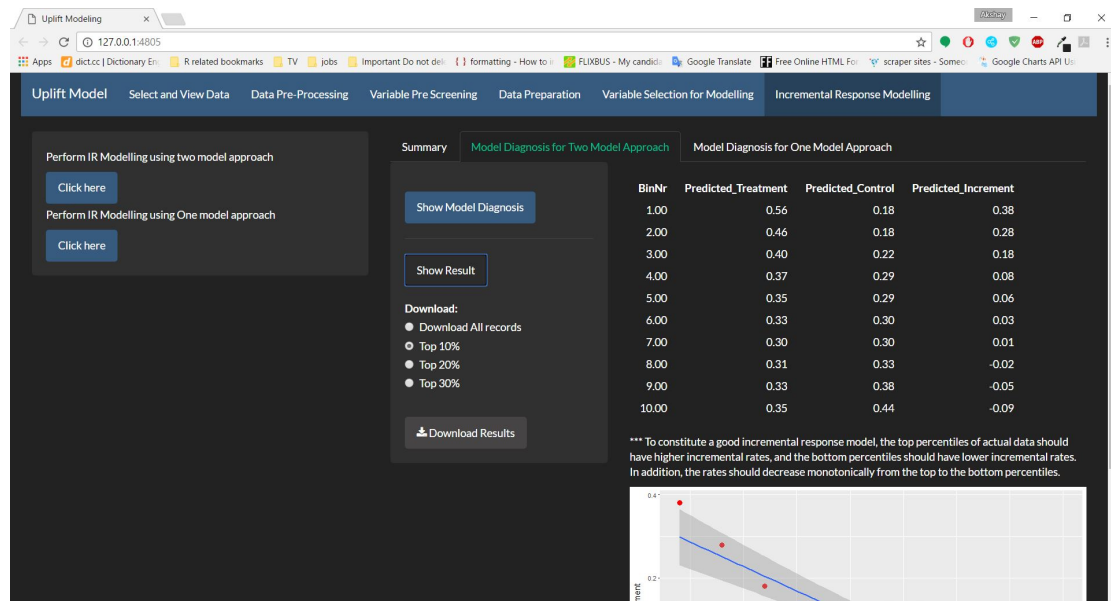


FIGURE 9.8: Two model Approach-2

Following screenshot shows the Combined Model Approach :

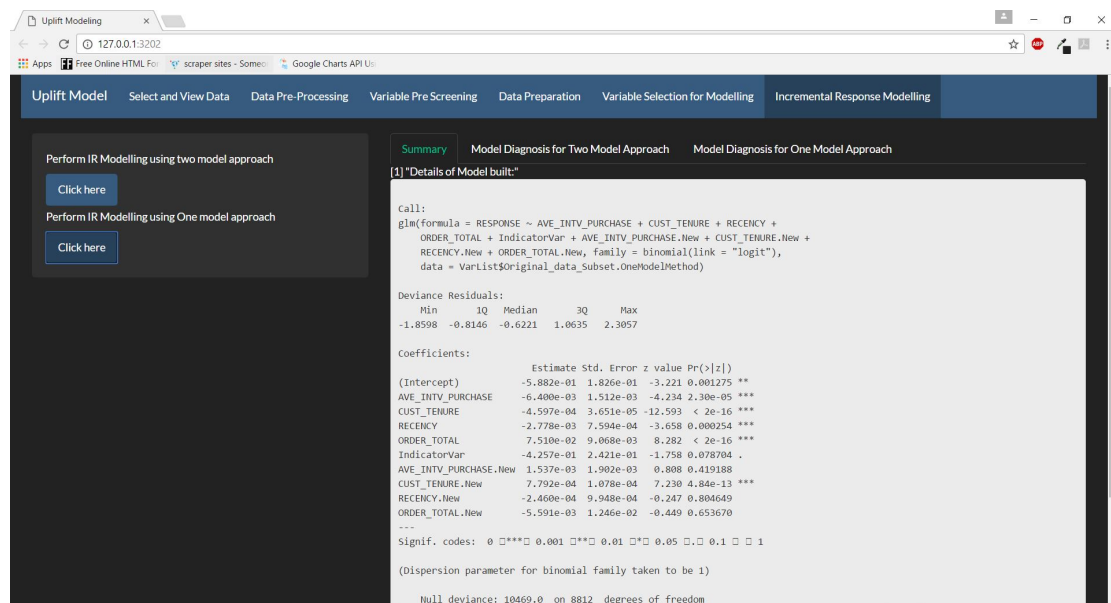


FIGURE 9.9: Combined model Approach

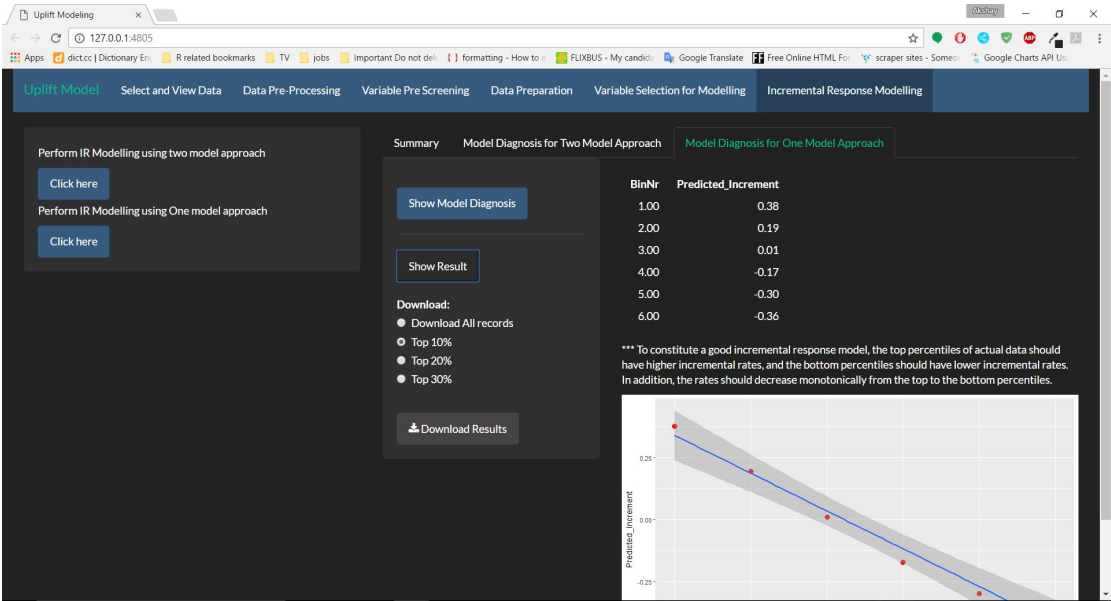


FIGURE 9.10: Combined model Approach-2

Bibliography

- [1] Eric Siegel. Article at the fiscal times. January 21, 2013. URL <http://www.thefiscaltimes.com/Articles/2013/01/21/The-Real-Story-Behind-Obamas-Election-Victory>.
- [2] N. J. Radcliffe and P. D. Surry. Real-world uplift modelling with significance-based uplift trees. 2011. URL <http://stochasticolutions.com/pdf/sig-based-up-trees.pdf>.
- [3] N. J. Radcliffe and P. D. Surry. Differential response analysisanalysis: modeling true response by isolating the effect of a single action. 1999. URL <http://www.maths.ed.ac.uk/~mthdat25/uplift/csc99-1>.
- [4] Victor S.Y. Lo. The true lift model - a novel data mining approach to response modeling in database marketing. URL <https://pdfs.semanticscholar.org/d087/44957d3f7a54c346df79b3a50f72d72f72f2.pdf>.
- [5] Xiangxiang Meng Taiyeong Lee, Ruiwen Zhang and Laura Ryan. Incremental response modelingusing sas enterprise miner. Oct 2014. URL <https://support.sas.com/resources/papers/proceedings13/096-2013.pdf>.
- [6] Wikipedia. Feature selection. . URL https://en.wikipedia.org/wiki/Feature_selection.
- [7] Wikipedia. Curse of dimensionality. . URL [:https://en.wikipedia.org/wiki/Curse_of_dimensionality](https://en.wikipedia.org/wiki/Curse_of_dimensionality).
- [8] Kim Larsen. Blog on data exploration with weight of evidence and information value in r. 2015. URL <http://multithreaded.stitchfix.com/blog/2015/08/13/weight-of-evidence/>.

-
- [9] Kim Larsen. Data exploration with information theory (weight-of-evidence and-information value). 2016. URL <https://cran.r-project.org/web/packages/Information/Information.pdf>.
 - [10] Naeem Siddiqi. Credit risk scorecards: Developing and implementing intelligent credit scoring. 2016.
 - [11] Kim Larsen. Net lift models: Optimizing the impact of your marketing efforts. 2010.