Midterm Progress Report

KDD Cup 1998 – Direct Marketing for Profit Optimization

Prepared By: Casey Kongpanickul & Janell So

1 Introduction

In this section, we will discuss in detail of the project that we have selected.

1.1 Objective

The project problem definition is cited from KDD Cup 1998 where the goal is to maximize profit from the donors through direct mail. The data set is provided by KDD and it includes 24 months of detailed promotion and giving history, overlay demographics, summary of promotions sent to donors and summary variables of each donor's lifetime giving history. Based on the information, we will need to target the donors who have donated large amounts in the past, have consistently donated more than the cost of asking for the donation to maximize overall profit. We will need to create visualization on the results of the techniques applied.

1.2 Original Proposed Solution/Methods

Given that the goal and data set provided by KDD Cup, there are several data mining techniques that we are planning to implement.

The following data mining techniques will be applied for this project:

- Association – Given the 24 month giving history data, using Association to recapture former donors who donated more than the cost between 13-24 months and to capture donors who have donated more than the cost for the past 12 months. A parallel coordinates can be used to show the results.
- Cluster Analysis – Given the overlay demographics data, using cluster analysis to determine donors. A scatter plot will be used to show the results.
- Classification – Given the overlay demographics data and donor's response to other mail offers data, using classification to determine future response by the donors. A boxplot can be used to show the results.
- Regression – Given the summary variables of donor's, demographics and giving history, using regression technique to predict the donors who will make donations on the next marketing campaign. A heatmap can be used to show the demographics of the possible donors.

In order for the targeted end users to better understand the results, the results will be shown in the form of histogram, scatter plot, boxplot and heatmap. We plan to use the following technologies/tools to implement the data mining techniques mentioned above.

- Programming Language – Python
- Data Cleansing Tool – OpenRefine
- Data Visualization – R Cloud*****, Hadoop, Python

  ***In evaluation mode**

1.3 Original Project Schedule

1.3.1 Resources

The following team members will be involved in working on this project.

- Casey Kongpanickul
- Janell So

1.3.2 Detail Estimated Project Timeline & Resources

The following chart shows an estimated timeline along with the detailed tasks assigned to each team member.

	Task Name	Duration	Start	Finish	Predecessors	Resource Names
1	Project Proposal	3 days	Wed 10/11/17	Fri 10/13/17		Casey,Janell
2	Understanding Data Set	2.5 days	Mon 10/16/17	Wed 10/18/17	1	Casey,Janell
3	**Implementation**	**36 days**	**Wed 10/18/17**	**Wed 12/6/17**		
4	**PreProcessing**	**11 days**	**Wed 10/18/17**	**Wed 11/1/17**	**2**	**Casey,Janell**
5	**Data Integration - Code Development**	**6 days**	**Wed 10/18/17**	**Wed 10/25/17**		**Casey,Janell**
6	Code Development	5 days	Wed 10/18/17	Tue 10/24/17		
7	Unit Testing/Validation	0.5 days	Wed 10/25/17	Wed 10/25/17		Casey,Janell
8	**Data Cleansing**	**6 days**	**Thu 10/26/17**	**Thu 11/2/17**	**5**	**Casey,Janell**
9	Code Development	5 days	Thu 10/26/17	Wed 11/1/17	7	
10	Unit Testing/Validation	0.5 days	Thu 11/2/17	Thu 11/2/17	9	
11	**Data Mining**	**19 days**	**Wed 11/1/17**	**Mon 11/27/17**	**4**	**Casey,Janell**
12	**Data Mining Technique (1) - Association**	**10 days**	**Wed 11/1/17**	**Tue 11/14/17**		**Janell**
13	Code Development	9 days	Wed 11/1/17	Mon 11/13/17		Janell
14	Unit Testing/Validation	0.5 days	Tue 11/14/17	Tue 11/14/17		Janell
15	**Data Mining Technique (2) - Clustering**	**10 days**	**Wed 11/1/17**	**Tue 11/14/17**		**Casey**
16	Code Development	9 days	Wed 11/1/17	Mon 11/13/17		Casey
17	Unit Testing/Validation	1 day	Tue 11/14/17	Tue 11/14/17		Casey
18	**Data Mining Technique (3) - Predictive**	**10 days**	**Tue 11/14/17**	**Mon 11/27/17**	**12**	**Janell**
19	Code Development	9 days	Tue 11/14/17	Fri 11/24/17		Janell
20	Unit Testing/Validation	0.5 days	Fri 11/24/17	Fri 11/24/17		Janell
21	**Data Mining Technique (4) - Classification**	**10 days**	**Tue 11/14/17**	**Mon 11/27/17**	**15**	**Casey**
22	Code Development	9 days	Tue 11/14/17	Fri 11/24/17		Casey
23	Unit Testing/Validation	0.5 days	Fri 11/24/17	Fri 11/24/17		Casey
24	**Post Processing**	**7 days**	**Mon 11/27/17**	**Tue 12/5/17**	**11**	**Casey,Janell**
25	Load Data Mining Results in Hadoop	2 days	Tue 11/28/17	Wed 11/29/17	11	Casey,Janell
26	Data Visualization	5 days	Thu 11/30/17	Wed 12/6/17	25	Casey,Janell
27	Midterm Project	15 days	Tue 10/24/17	Mon 11/13/17		Casey,Janell
28	Poster Session	10 days	Wed 12/6/17	Tue 12/19/17	24	Casey,Janell
29	Final Project	10 days	Wed 12/6/17	Tue 12/19/17	24	Casey,Janell

2 Midterm Progress Report

2.1 Completed Tasks

2.1.1 Understanding Data Set

We have reviewed the data set provided as part of the KDD cup.

The files reviewed were as follows:

- cup98LRN.txt – data file
- cup98VAL.txt – data file
- cup98dic.txt – data dictionary file
- valtargt.txt

By reviewing these files, it gave us a better understanding of the data and how we should approach on applying the techniques that we have proposed.

2.1.2 Preprocessing

We are currently in the process of identifying where data cleaning is required for the data sets provided.

2.2 Next Steps

We will continue with the preprocessing step of data cleaning before we applied the data mining techniques and data visualizations. Due to unforeseen issues, the project is delayed. In order to complete the tasks by mid-December, we have reduced the number of days for each of the data mining technique from 10 days to 6 days.

A revised project plan is shown below.

ID	Task Name	Duration	Start	Finish	Predecessors	Resource Names	% Complete
1	Project Proposal	3 days	Wed 10/11/17	Fri 10/13/17		Casey,Janell	100%
2	Understanding Data Set	2.5 days	Mon 10/16/17	Wed 10/18/17	1	Casey,Janell	100%
3	Implementation	42 days	Wed 10/18/17	Thu 12/14/17			26%
4	PreProcessing	11 days	Wed 10/18/17	Wed 11/1/17	2	Casey,Janell	64%
5	Data Integration - Code Development	6 days	Wed 10/18/17	Wed 10/25/17		Casey,Janell	99%
6	Code Development	5 days	Wed 10/18/17	Tue 10/24/17			100%
7	Unit Testing/Validation	0.5 days	Wed 10/25/17	Wed 10/25/17		Casey,Janell	100%
8	Data Cleansing	12 days	Thu 10/26/17	Sun 11/12/17	5	Casey,Janell	50%
9	Code Development	13 days	Thu 10/26/17	Sun 11/12/17	7		50%
10	Unit Testing/Validation	0.5 days	Sun 11/12/17	Sun 11/12/17	9		50%
11	Data Mining	19 days	Thu 11/9/17	Tue 12/5/17	4	Casey,Janell	0%
12	Data Mining Technique (1) - Association	6 days	Thu 11/9/17	Thu 11/16/17		Janell	0%
13	Code Development	5 days	Thu 11/9/17	Wed 11/15/17		Janell	0%
14	Unit Testing/Validation	0.5 days	Wed 11/15/17	Wed 11/15/17		Janell	0%
15	Data Mining Technique (2) - Clustering	6 days	Thu 11/9/17	Thu 11/16/17		Casey	0%
16	Code Development	5 days	Thu 11/9/17	Wed 11/15/17		Casey	0%
17	Unit Testing/Validation	0.5 days	Wed 11/15/17	Wed 11/15/17		Casey	0%
18	Data Mining Technique (3) - Predictive	6 days	Fri 11/17/17	Fri 11/24/17	12	Janell	0%
19	Code Development	5 days	Fri 11/17/17	Thu 11/23/17		Janell	0%
20	Unit Testing/Validation	0.5 days	Fri 11/24/17	Fri 11/24/17		Janell	0%
21	Data Mining Technique (4) - Classification	6 days	Thu 11/16/17	Thu 11/23/17	15	Casey	0%
22	Code Development	5 days	Thu 11/16/17	Wed 11/22/17		Casey	0%
23	Unit Testing/Validation	0.5 days	Thu 11/23/17	Thu 11/23/17		Casey	0%
24	Post Processing	7 days	Tue 12/5/17	Wed 12/13/17	11	Casey,Janell	0%
25	Load Data Mining Results in Hadoop	2 days	Wed 12/6/17	Thu 12/7/17	11	Casey,Janell	0%
26	Data Visualization	5 days	Fri 12/8/17	Thu 12/14/17	25	Casey,Janell	0%
27	Midterm Project	15 days	Tue 10/24/17	Mon 11/13/17		Casey,Janell	100%
28	Poster Session	10 days	Thu 12/14/17	Wed 12/27/17	24	Casey,Janell	0%
29	Final Project	10 days	Thu 12/14/17	Wed 12/27/17	24	Casey,Janell	0%

2.3 Related Papers (Pros & Cons)

There are many research papers with data mining techniques. We have found a few articles specifically around data mining techniques used in marketing.

2.3.1 Association

2.3.1.1 Pros

1. Find patterns to predict what the customers may be interested in and guide companies to make decisions on different areas such as pricing, selling and business strategies even when very little data is available.

2.3.1.2 Cons

1. Low performance in the algorithms. Although techniques such as Apriori, FP-growth are introduced, it requires many database scans or high memory usage.
2. Quality of the extract rules for comprehensibility.

2.3.2 Clustering

2.3.2.1 Pros

1. Track customers' behavior to create a strategic business initiatives
2. Reduce the number of observations by grouping them into clusters
3. Keep high profit, high value and low risk customers because it will target 10 to 20 percent of the customer base who will create 50 to 80 percent of the company's profits

2.3.2.2 Cons

1. Needs to be validated or may run into risk of a single cluster analysis and take the results as truly informative especially in presence of outliers

2.3.3 Classification

2.3.3.1 Pros

1. Help with customer identification to seek profitable segments of customers through analysis of the customer's characteristics
2. Help with attracting target customer segments especially with direct marketing since it is targeting specific segments of customers.
3. Help with customer retention since it includes one to one marketing and more personalized marketing campaigns which can predict changes in customer behaviors
4. Help with customer development where analysis maximize the customer transaction value by revealing the patterns of behavior
5. Help in identifying frequency, size of purchases and customer groups.

2.3.3.2 Cons

1. One of the main disadvantages of classification is that is can be difficult to not add in bias. One of the main points of classification is to non-subjectively put labels on incoming data. However, depending on how the classifier is built it can have inherent bias from the builder.
    a. If we are trying to classify a customer as consistent or not we may choose to look at how much money they spend and put more weight on that attribute while others may want to look at how often the customer spends money there and place more emphasis on that.
2. Classifiers are subject to the "curse of dimensionality" if the attribute selector is not careful. Meaning spurious classifications can be given to incoming data points.
    a. If we are trying to classify customers based on a random assortment of attributes we may find something odd such as, "During the month of March it looks like customers who wear blue are much more likely to buy a car" and while the data may support that, intuitively it is nonsense.

2.3.4 Linear Regression

2.3.4.1 Pros

1. If both dependent and independent variables are continuous (numerical), then a correlation coefficient (Pearson's) can be derived from the regression line. This is a measure of strength and direction in the relationship between variables.
    a. If the data shows that customers in a certain areas are more likely to spend money during certain times then it save money to only send adds to that area during specific times
    b. Note: In linear regression the dependent variable **must** be continuous, the independent variable can be continuous, binary or categorical
2. Regression lines allow for estimation (prediction). The values of dependent variables can be guessed (approximated) by a regression line given an independent variable value.
    a. The data may show that a combination of independent variables will likely yield a customer to spend over X amount of dollars. Then if we see a new customer with similar traits we can guess how much they will spend
3. Linear regression can be used on univariable problems as well as multivariable problems. Rarely will the dependent variable rely only one a single independent variable.
    a. Especially in marketing or predicting profits the customers come in with many attributes. From an analysis stand point a regression can look at combinations of variables to approximate some outcome.

2.3.4.2 Cons

1. Missing values can cause a lot of problems for linear regression. Handling missing values can change the regression line and correlation coefficient. Simply removing the instances with missing data can be problematic as well. It is possible to run other machine learning models to estimate the missing values, but this adds complexity to the problem.

    a. If customers do not enter their address and those customers are deleted from the data set then a lot of other data is lost because of that.
2. Linear regression is not ideal for picking up subgroups within a population. Therefore, is subgroup exist within a population then they need to be pre-defined, or it's possible that a subgroup can be undetected by linear regression.
    a. In some situations we may have data from children and adults, however, if examined separately they may yield more interesting results.
3. Attribute selection can be tricky for multivariable problems. In some univariable problems a certain independent variable can show a strong effect in the regression analysis. However, if used instead in a multivariable problem with other independent variables the strength of the independent variable can be diminished due to the variables being somehow interdependent.
    a. For example, if we collect information about how many people buy on a certain day and we find that many people buy cupcakes on that day we may choose to send out a lot of cupcake adds during the day. However if we consider a binary variable that shows that that certain day is also "Cupcake day" is may explain away why so many people bought cupcakes and maybe sending ads during that day wouldn't be as helpful since people will buy cupcakes anyway.

2.4 Potential Extensions

If time permits, the following techniques may be implemented.

2.4.1 Sequencing

Time series analysis can be done given the time intervals provided in the data sets to predict subsequent events based on previous behavior of the donors so we can target specific groups.

2.4.1.1 Pros

1. Finding seasonal, cyclical and other trends in the long term.
    a. During the holiday season the data may show that the public is much more likely to donate. It would make sense to ask a higher percentage of people for donations during the end of the year.
2. Finding early warning signs of business slow down.
    a. Based on historical data it might show that when profits trend down based on certain factors that the trend will not end soon. This can give a savy data scientist a clue to change the strategy for a business

2.4.1.2 Cons

1. Historical data may not give a true representation of any information mined from that data to predict future trends.
    a. When looking at data for donations one may find that donations had a large pike around a certain time. However, if when carefully looked at one may find that during that time of the large spike in donations received that there was a large natural disaster. It would be easy to miss something like this when just looking at data without context.

3 References

KDD Cup 1998 - http://www.kdd.org/kdd-cup/view/kdd-cup-1998/Intro

10 Techniques and Practical Examples of Data Mining in Marketing - http://www.egon.com/en/company/news/666-techniques-data-mining-marketing.html

Data Mining for Marketing - http://www.ijritcc.org/download/1426739680.pdf

Drawbacks and solutions of Applying Association Rule Mining in Learning Management systems - http://ceur-ws.org/Vol-305/paper02.pdf

Cluster Analysis for Market Segmentation - https://www.slideshare.net/vishtandel1991/cluster-analysis-for-market-segmentation

Linear Regression Analysis – Part 14 of a Series on Evaluation of Scientific Publications - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/

Data Classification - https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/s10-03-data-classification.html

The Pros & Cons of Trend Analysis in Forecasting - http://smallbusiness.chron.com/pros-cons-trend-analysis-forecasting-58786.html