Machine Learning in Email Marketing

- **Introduction**
- **Reviewed papers**
- **Proposed Approach**

Introduction:

Email marketing is an important tool used by almost all established companies today as a communication mechanism to target their customers with specific content. Sending emails is fast, cheap and highly targeted and enables companies to push the content they want to their customer base. One big advantage in this approach is that related statistics can be gathered directly concerning user responses to these emails such as how many opened and followed through to a link, which users remained idle and who opted to quit the mailing list. These statistics are important as they reveal quantifiable metrics such as conversion rate and unsubscription ratios. Email marketing is prevalent across all industries and is not confined to any particular market hence there has always been a need to improve its efficiency.

The goal of this research project is to focus on various Machine learning and Reinforcement learning methods which can help improve the efficacy of email marketing. We strive to find an optimal solution using the gathered data of user responses to propose most favorable time and content the emails should be send with in order to increase the conversion rate.

Reviewed papers:

Concurrent Reinforcement Learning from Customer Interactions - [Silver, David et al 2014]

Used a variant of temporal-difference learning algorithm to learn from online from partial interaction sequences, so that information acquired from one customer is efficiently assimilated and applied in subsequent interactions with other customers. The difference in their approach from traditional reinforcement learning is that the agent interacts with many customers concurrently. They have used a simulator to compare the method with MC, traditional TD and Contextual Bandit algos and have demonstrated that it works better.

Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming - [Cui, Geng et al 2006]

They applied Bayesian networks learned by evolutionary programming to model responses to direct marketing. They show that this approach of predictive modeling in direct marketing works better than other benchmark methods, including neural networks, classification and regression tree (CART), and latent class regression.

They used a data set from The Direct Marketing Education Foundation of a company that sells multiple product lines of general merchandise. The company sends regular mailings to its list of customers, and this particular data set contains the records of 106,284 consumers. Each customer record contains 361 variables, including purchase data from recent promotions and the customer’s purchase history over a 12-year period. A recent promotion, achieved a 5.4% response rate, which represents 5,740 customers who made purchases from emailed catalog.

Empirical Comparison of Various Reinforcement Learning Strategies for Sequential Targeted Marketing [Naoki, Abe et al. 2002]

In this paper various reinforcement learning methods performance is evaluated for sequential targeted marketing. They have worked with "batch" and "simulation based" methods. Direct or batch reinforcement learning attempts to estimate
the value function Q(s, a) by reformulating value iteration as a supervised learning problem. On the other hand, Indirect. or simulation-based methods of reinforcement learning first build a model of Markov Decision Process by estimating the transition probabilities and expected immediate rewards, and then learn from data generated using the model and a policy that is updated based on the current estimate of the value function. Their findings are that where modeling is possible indirect methods works better whereas in realistic situations the performance degrade. They also show that semi-direct methods are effective in reducing the amount of computation necessary to attain a given level of performance, and often result in more profitable policies.

Proposed Approach:

The problem at hand is to find the most optimal time when to send the email to a client. The data concerning this involves historical interaction behaviour which includes whether the email was opened, unsubscribed or left idle. Since Daniyal is already working on Reinforcement Learning approaches, I propose a ML solution which tackles the issue as a classical classification problem.

Labelling:

Before choosing a model, firstly we can label the data such that for email opened = 1, idle = 0 and unsubscribed = 2. Then we can check the class balance in the data. Intuitively the idle must have a higher ratio than others but still this needs to be explored. We can then add class weights and/or use SMOTE oversampling and other such pre-processing techniques.

Features:

We have a times series and we can divide that into half hourly or hourly slots (or something similar) so that we can treat them as a categorical variable in feature list. We can also add whatever other features are available.

Model:

Since we are treating it as a classification problem, we have lot of options to play around with and select the one that fits best to our need. It can be decision trees such as Random Forests or XGB or even simpler logistic regression. We can define metrics of precision/recall/accuracy to measure how good the model is depending on our finer objective.

However I think that exploring the data descriptively and visually is of utmost importance here as it will give out patterns which will help us understand the problem better. For example looking at the counts and percentages of emails opened vs not opened by year, month, day, and time of day will give us insights that will be imperative in modeling the problem solution.