Capstone Proposal

Domain Background

A problem facing Marketing/Advertisement agencies in the Business-to-Business (B2B) space is "who" and "when" to target companies for enterprise level sales.

The objective of Marketing Sales Campaigns are to increase the sales of particular products and/or solutions. Multiple marketing channels, such as Inside Sales and email, are utilized to increase the likelihood that sales will increase. However, proving that a given campaign can be attributed with having a positive impact, it is essential that sales campaigns are ran as quickly and efficiently as possible so that clear spikes in activity can be justly linked to the campaign itself, and not just random increases.

Targeting a company with a sales campaign when there is no intent to purchase wastes time and money for a campaign making justification to use the agency much harder but it can also long term damaging effects outside of a single campaign as it risks companies classifying the agency itself as source of spam and therefore reduces the effectiveness of future campaigns.

Problem Statement

If a marketing agency could better understand the current, and expected levels, of intent for a company it would greatly aid in more personalized marketing, where companies that are more likely to engage, and are therefore more likely to generate more leads through optimized sales campaigns.

This project aims to highlight, based on online URL activity, what company segments can be created that have distinct features and activity behaviors which will help better inform marketers so that their sales operations can be more personalized.

Data-sets and Inputs

The data utilized in this project is generated by a "Demand Based Platform" (DSP) which bids on the online ad exchanges in order to place display adverts on webpages. A by product of placing display ads is a database of IP address and the URLs they appeared in which an display advert could of have been bid on. Activity is defined as "pageViews" which is the estimated number of page visits per a unique session ID

This data is enriched by a database of 12 million IP bands and the companies which are registered to those IP bands so that we have some knowledge around the online activity at the company-domain-city-region level. Included also are the firmographic details for these companies to be used as discrete labels to help further understand possible segments.

Solution Statement

An unsupervised clustering model to be created in which distinct segments of companies are generated that provide actionable insight into the expected behaviors of a company based on their firmographic information as well as past activity.

The information generated should be presentable to a marketer so that he/she can perform actionable task to optimise a marketing sales campaign and prioritize companies by their likelihood to engage with sales content.

Benchmark Model

The current benchmark is to treat all companies as being equally likely to engage with all marketing content.

One form engagement which will be used to highlight the success of the model will be changes in "Click-through rates" of online display ads. The industry wide mean is around 0.05%, so any significant variations, positively or negatively, away from this will be deemed as useful and actionable insights.

If i can highlight that the clusters provided significantly vary from this industry benchmark then this will provide value to the business. Comparison against a 12 month average click-through rate for the full data-set combined with some form of confidence interval should suffice.

Evaluation Metrics

Cluster distinction scores, such as the Silhouette Score, will be used to evaluate how unique the segments created are and therefore how useful they are to the business in implementing distinct strategies.

Project Design

The project will be split into two sections, with the the outcomes from the first section to b used to help drive and narrow the focus of the second.

Based on the last 12 months of IP tracking data, the aim is to determine the following:

1. Can we successfully cluster companies into discrete groups based on firmographic, geographical and URL activity.
2. Identify outliers in the data-set and determine if this indicate high/low levels of intent or potential data quality issues
3. Can the number of features typically used in analysis be reduced so that future analysis can become more streamlined

Data Pre-processing/Cleaning

To ensure any results found are as optimized as possible, the first stage will involve pre-processing of the data using the following techniques:

1. Outlier Identification
   - Are there any accounts that should be removed from the training sets?
2. One Hot Encoding
   - Are there any continuous features which could be used as discrete labelled data?
3. Dimensionality Reduction
   - Are there any features provide similar insights and can be reduced into a single dimension to simplify analysis?
4. Normalisation
   - Can the data be represented by a Gaussian Distribution?
   - Should features be transformed to help aid analysis and modeling?
   - Should continuous values be normalized between 0 and 1 for reduce the effect of outlier values?
5. Visual Analysis
   - Are there any noticeable feature relationships in the data that are interesting and are worth investigating further

Any discoveries found from the above will then be summarized and considered prior to starting the stage of the project. e.g. what outliers should be removed and why?

Unsupervised Learning

No prior knowledge is available around the distinct behaviors, if any, between different groups of companies and their respective URL activity and display advert engagement,

As a result the 2nd stage will focus on unsupervised learning to help better understand what discrete groups of companies there may be in and help highlight focus on certain features as well as providing useful results.

Multiple Clustering techniques will be investigated and evaluated base on their success in creating distinct groups with minimal overlaps. Metrics such as the "Silhouette Score" will be used to quantify this so that models can be compared.

Training and testing sets will be create to reduce the chance of over-fitting or under fitting models to the data and ensure the most generic model is selected.

Clustering Techniques:

- K-means
- DB Scan
- Ward

Considerations to be made:

- Distinction of clusters
- Computation performance

Summary of Results

The output of the clustering models should be accompanied by both clear visualizations and actionable insights that can be understood by a marketer without any prior knowledge into Machine Learning.