Machine Learning Engineer Nanodegree

Capstone Project

Craig Bath

June 2nd, 2019

Definition

Project Overview

Domain Background

A problem facing Marketing Agencies in the Business-to-Business (B2B) space is "who" and "when" to target companies for enterprise level sales on behalf of paying clients.

Agencies run “Sales Campaigns” with the objective being to increase the sales of particular products and/or solutions for a given period of time. Multiple marketing channels, such as Inside Sales and email, are utilized in attempts to increase the likelihood that sales will increase and as a result highlight their value as a marketing agency. However, proving that a given campaign can be attributed to 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 due to random spikes in activity that could just as easily be due to pure chance.

Targeting a company with a sales campaign when there is no intent to purchase wastes time and resources for a campaign, making justification to use the agency much harder. It can also result in long term damaging effects outside of a single campaign, as every time the agency runs a campaign it risks being classified as a source of spam itself, and therefore reduces the effectiveness of future campaigns.

This project aims to highlight, based on a company’s online activity, what company segments can be derived that exhibit distinct behavioral patterns behaviors that will help better inform marketers, so that their sales operations can be more personalized and streamlined, driving increased engagement and therefore sales.

Data-sets and Inputs

The dataset utilized in this project, which provides a source of sales of intent, is based on the frequency IP addresses visiting certain URLs (aka websites) from within a given universe of URLs. The data has been gathered from what is known as a "Demand Based Platform" (DSP) in the Marketing industry of online display adverts.

Behind every website on the public internet that contains digital advertisements there are “Ad Exchanges” which, in essence, are real time bidding platforms (similar to the stock exchange) where companies/people bid for the right to serve a person, based on IP address, a digital advertisement of their choice.

For example, when I visit amazon.com they store my details and activity on their website and have data showing I am interested in a new video game. When I go onto a different website, e.g. stackoverflow.com, my IP address will appear on the Ad exchange and amazon will likely attempt to win the auction in order to display an advert for the video game I was recently viewing in an attempt to lure me into a sale.

DSP are the companies which conduct the bidding side on a large scale, on behalf of other companies. Conducting this bidding at scale results in a vast dataset of IP address and the websites they have visited. Along with the dates and frequencies of these visits. Activity/Intent can be defined as "page views", “hits” and “unique views” 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. E.g. let’s say Microsoft Philadelphia owns the IP block 1000 to 2000, we can attribute any activity from any IP address between 1000 and 2000 to someone at Microsoft in Philadelphia.

To further enrich the dataset, we can also include firmographic details of the company to be used as discrete labels to help further understand possible segments. E.g. Number of Employees.

Data Set Creation

The sample dataset created for this project was generated from a query on a production database of DSP, which I have I access to explore from my employer. Details, with certain table names redacted, on the queries used to create this dataset are found in the following Q Jupyter notesbooks:

1. “Base Data Creation - Safe.ipynb” - Raw database queries used to sample IP tracking data for the month of March and aggregate data up to an account level
2. ”Creating Derived Dataset - Safe.ipynb” – Combines data from above with firmographic data along with pivoting the dataset so that metric counts are broken out by categories as columns

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, are highlighted to marketing executives so that they can prioritize who and when to contact companies in a more informed, and data driven, manner.

As an outcome of a more informed priority list, the probability of generating more leads should increase in a given timeframe, which will greatly aid the marketing agency in demonstrating their value and direct causal correlations in the actions of their sales executives and any uplift in sales for the client.

The online activity of companies will be collected for a random, full, month in 2019 with activity group by “IAB Category”. This is an online standard which helps categorize websites into distinct groups based on their content e.g bbcnews.com would be classified as under “IAB12” which corresponds to news websites. The full list of categories can be found here for reference https://support.aerserv.com/hc/en-us/articles/207148516-List-of-IAB-Categories

An unsupervised clustering model will then be created, with the aim of creating multiple distinct segments of companies based on their activity which can help provide actionable insight into the expected behaviors of a company based on their on the URLs which they visit.

Multiple clustering techniques, such as k-means and DB scan, will be investigated with each model being evaluated based on the distinction between the clusters they generate. A numeric score will be generated, Silhouette Score, so that we can objectively compare results and the optimal method, along with optimized hyperparameters, can be selected.

Once a clustering model has been selected, details on the number of clusters and logic on how to associate new data points with a cluster will be provided so that we can correctly segment accounts based on their behavior.

The information generated should be presentable to a marketer so that he/she can perform actionable tasks to optimize a marketing sales campaign and prioritize companies by their likelihood to engage with sales content. No prior knowledge of machine learning or statistics should be required to understand the actions provided.

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.

The Silhouette Score measure, between -1 and 1, to how close datapoints are to it's assigned cluster as well as any neighboring clusters. It allows us to evaluate the distinction between clusters, with a score close to 1 highlight that the datapoints are close only to it's assigned cluster in N-dimensional space.

With a set range between -1 and 1, it allows us to find the optimal number of clusters, as well clustering model, by simply iterating through the variations of models and hyperparameters, e.g. test k equal to all values between 1 and 10 for k-means, and simply noting the highest score returned.

Analysis

Data Exploration

Dataset Creation

The dataset used as part of this report "sample_model_data_protected_60k.csv" was generated from querying a kdb+ database and combining IP activity data and company level labelled data. The names of the companies have been removed for privacy reasons.

Full query details can be found in the Q jupyter session "Base Data Creation - Safe.ipynb" and "Creating Derived Dataset - Safe.ipynb"

Schema Details

Column Name	Data Type	Description
acct_id	object	Unique Indentifier of an account, defined by a company + location
city	object	City of an account
region	object	Region of an account
countryCode	object	ISO2 Country Code of an account
revenue_mil_usd	int64	Account revenue in USD millions
total_employees	int64	Account's total employees
usageType	object	Classifation of the account's website e.g. ISP
datesCount	int64	Number of unique days of online activity seen for the account
domainsCount	int64	Number of different domains visited by the account
hitsSum	int64	Number of times an account is seen on domains
pageViewsSum	int64	Number of times an account visits websites
uniqueViewsSum	int64	Unique user sessions an account is seen. Based on a sessionID
clicks	int64	Number of times an account has been seen to click of an online display advertisment
clickDates	int64	Number of unique days an account has clicked on a online display advertisment
hits_iabCat_IAB_1	int64	Hits on domains classed as the IAB category = Arts & Entertainment
hits_iabCat_IAB_2	int64	Hits on domains classed as the IAB category = Automotive
hits_iabCat_IAB_3	int64	Hits on domains classed as the IAB category = Business
hits_iabCat_IAB_4	int64	Hits on domains classed as the IAB category = Careers
hits_iabCat_IAB_5	int64	Hits on domains classed as the IAB category = Education
hits_iabCat_IAB_6	int64	Hits on domains classed as the IAB category = Family & Parenting
hits_iabCat_IAB_7	int64	Hits on domains classed as the IAB category = Health & Fitness
hits_iabCat_IAB_8	int64	Hits on domains classed as the IAB category = Food & Drink
hits_iabCat_IAB_9	int64	Hits on domains classed as the IAB category = Hobbies & Interests
hits_iabCat_IAB_10	int64	Hits on domains classed as the IAB category = Home & Garden
hits_iabCat_IAB_11	int64	Hits on domains classed as the IAB category = Law, Gov't & Politics
hits_iabCat_IAB_12	int64	Hits on domains classed as the IAB category = News
hits_iabCat_IAB_13	int64	Hits on domains classed as the IAB category = Personal Finance
hits_iabCat_IAB_14	int64	Hits on domains classed as the IAB category = Society
hits_iabCat_IAB_15	int64	Hits on domains classed as the IAB category = Science
hits_iabCat_IAB_16	int64	Hits on domains classed as the IAB category = Pets
hits_iabCat_IAB_17	int64	Hits on domains classed as the IAB category = Sports
hits_iabCat_IAB_18	int64	Hits on domains classed as the IAB category = Style & Fashion
hits_iabCat_IAB_19	int64	Hits on domains classed as the IAB category = Technology & Computing
hits_iabCat_IAB_20	int64	Hits on domains classed as the IAB category = Travel
hits_iabCat_IAB_21	int64	Hits on domains classed as the IAB category = Real Estate
hits_iabCat_IAB_22	int64	Hits on domains classed as the IAB category = Shopping
hits_iabCat_IAB_23	int64	Hits on domains classed as the IAB category = Religion & Spirituality IAB23-1 Alternative Religions
hits_iabCat_IAB_24	int64	Hits on domains classed as the IAB category = Uncategorized
hits_iabCat_IAB_25	int64	Hits on domains classed as the IAB category = Non-Standard Content
hits_iabCat_IAB_26	int64	Hits on domains classed as the IAB category = Illegal Content

Sample Rows (Pivoted)

Columns	Row 1	Row 2	Row 3	Row 4
acct_id	1-1000AUX	1-1000JEX	1-1000KMO	1-1000SPZ
hq_id	1-1000AUX	1-1000JEX	1-1000KMO	1-15PE2WH
company	REDACTED	REDACTED	REDACTED	REDACTED
userDomain	REDACTED	REDACTED	REDACTED	REDACTED
city	Chippewa Falls	Madison	Rochester	Moorpark
region	Wisconsin	Wisconsin	New York	California
countryCode	US	US	US	US
revenue_mil_usd	6.6806	2.9769	24.9513	141.8804
total_employees	64	85	75	775
naic3	423	722	541	333
naic6	423110	722513	541810	333314
isISP	0	0	0	0
usageType	COM	COM	COM	COM
datesCount	5	5	14	25
domainsCount	5	5	20	34
hitsSum	9	8	61	88
pageViewsSum	9	8	109	263
uniqueViewsSum	9	8	65	111
clicks	0	0	0	0
clickDates	0	0	0	0
hits_iabCat_IAB_1	0	1	6	6
hits_iabCat_IAB_2	1	1	0	1
hits_iabCat_IAB_3	0	0	0	0
hits_iabCat_IAB_4	0	0	0	6
hits_iabCat_IAB_5	0	0	1	0
hits_iabCat_IAB_6	0	0	0	0
hits_iabCat_IAB_7	0	0	0	0
hits_iabCat_IAB_8	0	0	0	0
hits_iabCat_IAB_9	0	1	2	2
hits_iabCat_IAB_10	0	0	0	2
hits_iabCat_IAB_11	0	0	0	0
hits_iabCat_IAB_12	1	1	15	24
hits_iabCat_IAB_13	0	0	2	5
hits_iabCat_IAB_14	0	0	0	3
hits_iabCat_IAB_15	0	0	0	2
hits_iabCat_IAB_16	0	0	0	0
hits_iabCat_IAB_17	0	0	5	5
hits_iabCat_IAB_18	1	0	1	1
hits_iabCat_IAB_19	5	3	29	18
hits_iabCat_IAB_20	0	0	0	0
hits_iabCat_IAB_21	0	0	0	6
hits_iabCat_IAB_22	0	0	0	7
hits_iabCat_IAB_23	0	0	0	0
hits_iabCat_IAB_24	0	0	0	0
hits_iabCat_IAB_25	0	0	0	0
hits_iabCat_IAB_26	1	1	0	0

Dataset Summary

As highlighted below by the summary statistics, we can see immediately that this data is going to need some cleaning. Key points:

- Numerous feature columns where the standard deviation (STD) is significantly higher than the value for the 75th percentile, e.g. revenue_mil_usd has an STD of 2,603,505 with the 75th percentile at 448. Outlier analysis, transformations and even column removal will have to be investigated.
- Hits for each IAB category appear to be extremely skewed to the right. Data transformations will be required.

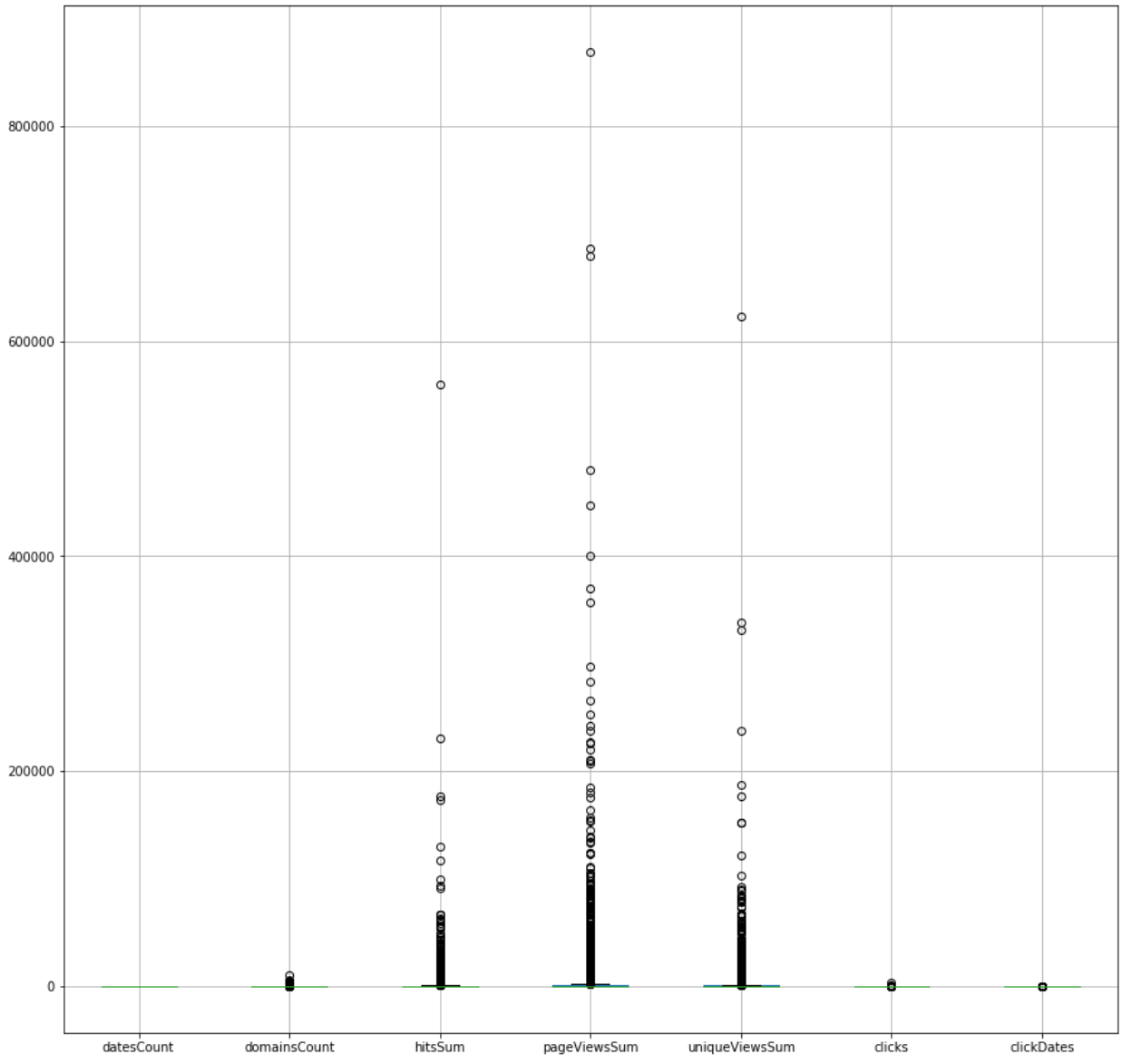
	count	mean	std	min	25%	50%	75%	max
revenue_mil_usd	37116	25264.80171	2603505.991	0	9.288425	50	448.9327	500000000
total_employees	37116	15529.65613	59039.63484	0	64	265	2300	2300000
datesCount	37116	13.86200022	9.895018342	0	5	12	22	31
domainsCount	37116	72.65518914	203.9609131	0	5	17	58	10366
hitsSum	37116	711.1817545	5429.031726	0	11	50	229	559606
pageViewsSum	37116	2316.942397	16254.11036	0	20	119	615	868332
uniqueViewsSum	37116	1059.101897	7549.64227	0	14	74	338	623298
clicks	37116	0.334006897	38.69142592	0	0	0	0	6748
clickDates	37116	0.028101088	0.427017343	0	0	0	0	31
hits_iabCat_IAB_1	37116	41.98243345	342.9405187	0	0	2	12	33061
hits_iabCat_IAB_2	37116	51.229362	360.3542933	0	0	2	18	22518
hits_iabCat_IAB_3	37116	11.07859144	79.44189667	0	0	0	2	6048
hits_iabCat_IAB_4	37116	6.872022847	55.43018934	0	0	0	1	4792
hits_iabCat_IAB_5	37116	36.26452204	374.2843827	0	0	0	4	30931
hits_iabCat_IAB_6	37116	1.234723569	10.07864145	0	0	0	0	869
hits_iabCat_IAB_7	37116	9.952904408	79.43054053	0	0	0	3	5755
hits_iabCat_IAB_8	37116	1.513067141	10.81818521	0	0	0	0	748
hits_iabCat_IAB_9	37116	51.98973489	409.6876364	0	0	2	13	31538
hits_iabCat_IAB_10	37116	7.572152172	52.82373032	0	0	0	3	3273
hits_iabCat_IAB_11	37116	0.530067895	9.45447973	0	0	0	0	1647
hits_iabCat_IAB_12	37116	145.426366	1533.430682	0	1	9	46	230481
hits_iabCat_IAB_13	37116	49.52680785	407.5051328	0	0	2	12	35870
hits_iabCat_IAB_14	37116	3.690753314	29.8913153	0	0	0	1	2219
hits_iabCat_IAB_15	37116	34.27198513	473.1445977	0	0	0	3	28269
hits_iabCat_IAB_16	37116	0.144331286	1.197395095	0	0	0	0	75
hits_iabCat_IAB_17	37116	16.06662895	112.684504	0	0	1	5	10424
hits_iabCat_IAB_18	37116	15.47602112	138.6982642	0	0	1	5	14082
hits_iabCat_IAB_19	37116	155.6612782	1071.758205	0	1	9	50	69477
hits_iabCat_IAB_20	37116	1.383554262	10.46862317	0	0	0	0	627
hits_iabCat_IAB_21	37116	8.919711176	95.08134605	0	0	0	2	9027
hits_iabCat_IAB_22	37116	51.46042138	660.7754846	0	0	0	8	105466
hits_iabCat_IAB_23	37116	0.327379028	2.688733497	0	0	0	0	154
hits_iabCat_IAB_24	37116	4.861676905	33.9004382	0	0	0	1	2028
hits_iabCat_IAB_25	37116	0.000350253	0.021398905	0	0	0	0	2
hits_iabCat_IAB_26	37116	3.744907856	27.7157804	0	0	0	0	1702

Exploratory Visualization

Summary Activity Metrics

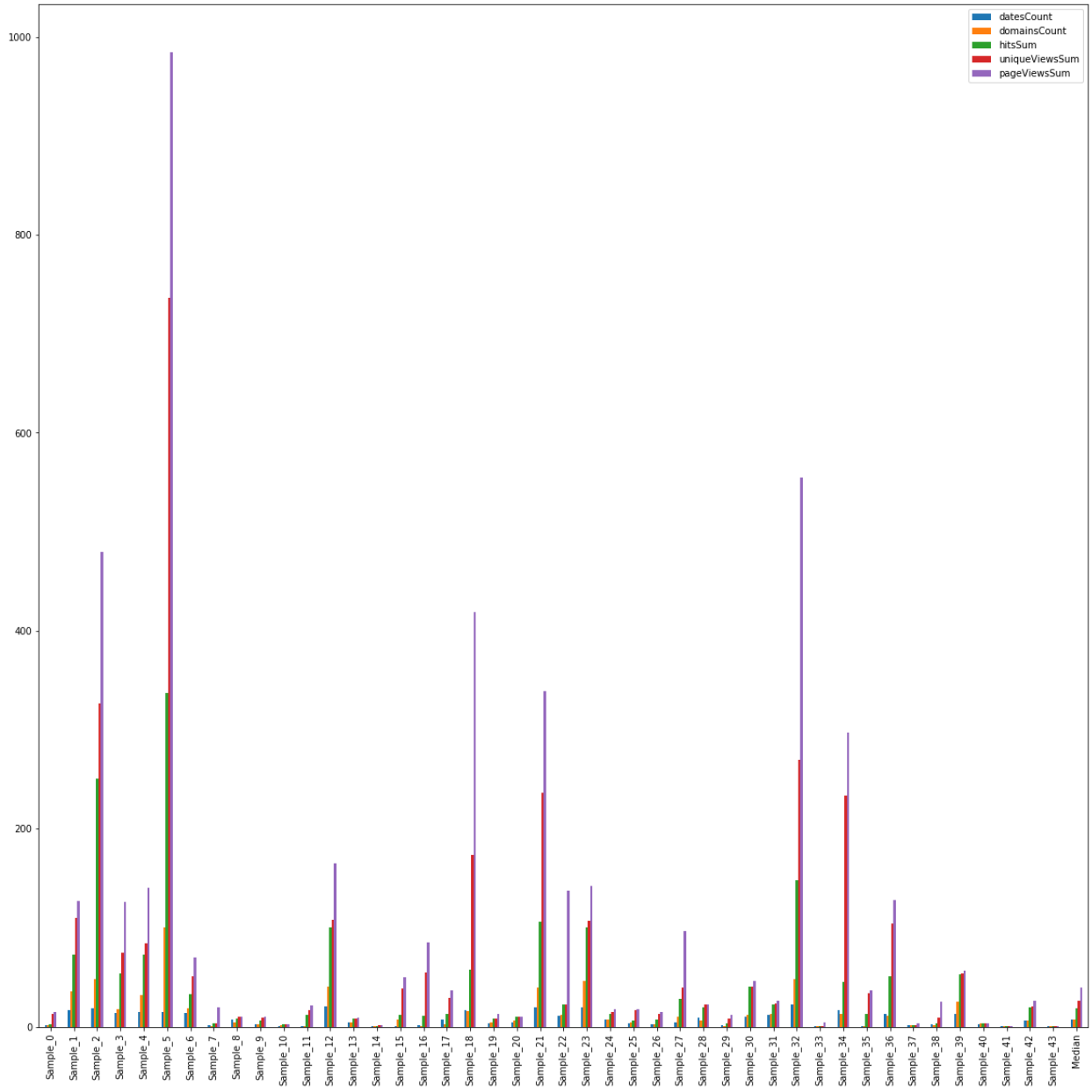
This skewness in the data is further highlighted by the boxplot, *figure 1* below, which shows numerous outliers for each metrics based on the interquartile range with a high density of values close to zero but also a large number of values mugh larger zero, resulting in a heavily right skewed dataset.

Figure 1:



Taking random samples of the dataset, and plotting at the account level, see *figure 2 below,* it is consistently shown that some accounts have significantly more online activity. This will hopefully provide some useful clusters and highlight some distinctions between accounts behaviours

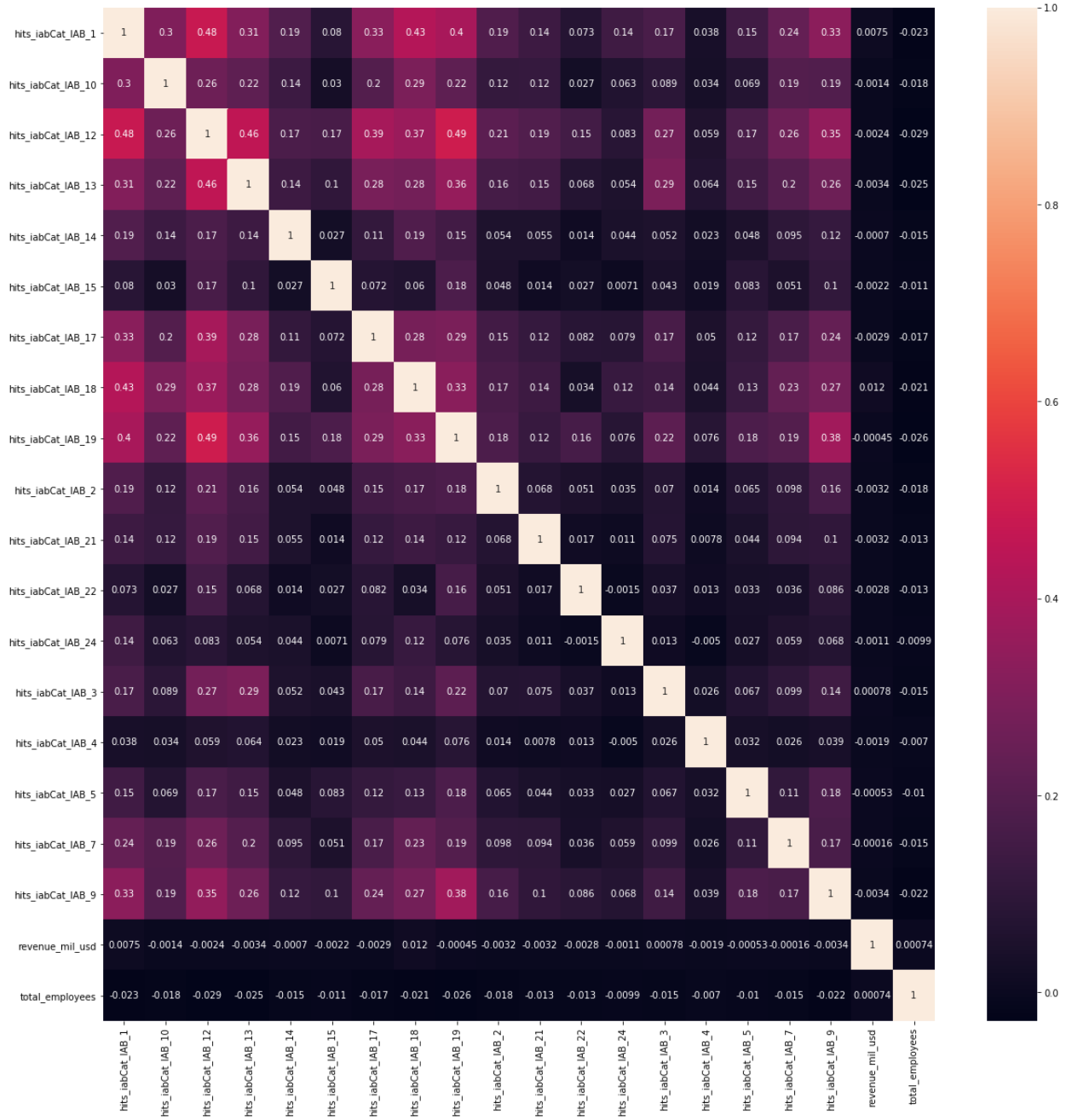
Figure 2.



Looking at a correlation heat map for all IAB category hits, see *figure 3* below, we see that there are only some occasional, and weak, correlations between activity in one category to others, suggesting that that there will likely be poor predictive power between the features. Which is useful for unsupervised learning.

Also note that revenue and employee size have extremely poor correlations to anything and more worryingly there is no correlation between employee size and revenue, which smells like a serious data quality issue.

Figure 3.



Algorithms and Techniques

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

As a result, the project will focus on unsupervised learning to help better understand what discrete groups of accounts 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.

Principal Component Analysis (PCA)

Due to the large number of features, 42 here, that are available it is essential that we attempt to reduce the features down in the dataset to only those likely to provide unique signals prior to fitting the dataset to any clustering model. Reducing the number of input features is hugely beneficial in machine learning as it:

- Reduces the dimensionality and therefore less computation needed for each model
- Reduces the risk of modeling noise due to unnecessary features being part of our modelling process

PCA is a frequently used technique to achieve this. It is a systemized way to transform input features into principle components (PCs) and uses these principal components as new features as the inputs into unsupervised learning models.

PCs are the defined as the directions of the data with the maximum variance (minimize information loss) when you compress data points down along them . You can have multiple, rank-able, PCs. Higher variance = higher ranks . This helps in removing noise from that dataset as well as reducing the computational cost of apply ML techniques as fewer inputs are used.

It also allows for visualization of higher dimensionality data.

Clustering Models

k-means

Base on a set number of groups (K) you wish to split your dataset into, the K-means algorithm will move each K group until the average distance of points associated to it is minimized for all groups.

Optimizing k-means involves minimizes the quadratic distance of all the points to their assigned clusters

Pros:

1. Computationally very cheap so we can avoid risk of maximizing to local minima by initializing the algorithm numerous times using ensemble methods to reduce the risk derived from random placement of centroids.
2. Works very well with circularly clustered data

Cons

1. K-means is a "hill-climbing" algorithm and is very dependent on where the initial centroids are placed. Which can be problematic when they are placed randomly and if we have a small amount of data and cannot iterate numerous times
2. This can also make it very vulnerable to optimizing to local minima during the assignment step of centroids e.g. 2 centroids out of 3 initialized extremely close together and 1 very far a away
3. k-means uses distance to centroid, and as a result is only really useful for circular (or hyper spherical) clusters. As a result it would not be able to identify the below as 3 long and skinny clusters

Hierarchal Clustering

Like K-means but for when you do not know the number of groups you wish to split your data into. Instead you define a distance from a group where you wish to stop clustering your data points. It works iteratively through your data points selecting the two closest data points and creating a new group for them or adding them to an existing group. Once the next nearest points are beyond your set distance, the algorithm finishes.

There are multiple hierarchal clustering methods and the main distinguishing features are how they measure the distance between clusters. This is heavily linked to the shape of clusters each method has an affinity for.

"Single link" for examples measures from the closet point in a cluster and tends towards elongated clusters as a result

In general, groups are a hierarchally created, with the most granular level being a cluster for each data point, then in pairs and so on until there is a single cluster encompassing the entire data set like below.

Pros:

1. Resulting Hierachal representation can be very informative
2. Provides an additional ability to visualize the data
3. Especially potent when there are hierarchal relationships in the data e.g. evolutionary biology

Cons:

1. Computationally expensive
2. Sensitive to noise and outliers and require a high level of data cleanliness

Decision Trees

A very basic concept that will continually split the data on a true/false condition (branches). The condition is determined by the question that will split the dataset into distinct group most effectively. In theory you could make as many branches until a dataset is split in all unique outputs but this will be prone to error on future datasets (over trained).

It will primarily be used in this investigation to highlight whether an unsupervised learning methodology is appropriate or not based on the predictive power between features.

Benchmark

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. But more importantly the result will be relayed back to the business to truly justify if there is any useful insight provided by splitting companies into n clusters

Methodology

Data Preprocessing

To ensure any results found are as optimized as possible, the first stage will involve pre-processing of the data using the following several techniques designed to reduce noise and data quality issues within the data as well as preparing the data in way likely to improve any ML models applied to the dataset

Outlier Analysis

The aim of outlier analysis is to highlight, and remove where necessary, any datapoint which fall significantly outside of the distribution of the rest of the dataset and as result will cause skewing of the results by disproportionality affecting the distribution and average of the dataset.

To class a data point as an outlier I will, for each column individually, identify any values which are larger than the (3rd quartile + 1.5* Inter Quartile Range) or smaller than the (1srd quartile - 1.5* Inter Quartile Range).

To ensure I am not removing datapoints which may have genuinely large activity, I will remove any datapoints for the modelling dataset which are flagged as outliers for 2 or more features

Principal Component Analysis

When using PCA, one of the main goals is to reduce the dimensionality of the data — in effect, reducing the complexity of the problem. Dimensionality reduction comes at a cost: Fewer dimensions used implies less of the total variance in the data is being explained. Because of this, the cumulative explained variance ratio is extremely important for knowing how many dimensions are necessary for the problem.

I will first apply PCA to the dataset utilizing all the features in order to return a breakdown of the cumulative explained variance of each dimensions. From this I will select the number of dimensions which explains ~90% of the explained variances, N, and then apply PCA to the dataset again to reduce down to N dimensions.

A biplot will be then be used visually highlight the dimensionality reduction. A biplot is a scatterplot where each data point is represented by its scores along the principal components. The axes are the principal components. In addition, the biplot shows the projection of the original features along the components. A biplot can help us interpret the reduced dimensions of the data, and discover relationships between the principal components and original features.

Normalization

The distribution of each continuous feature will be evaluated and transformed to as close to a gaussian distribution as possible in order to help the predictive power of the ML models.

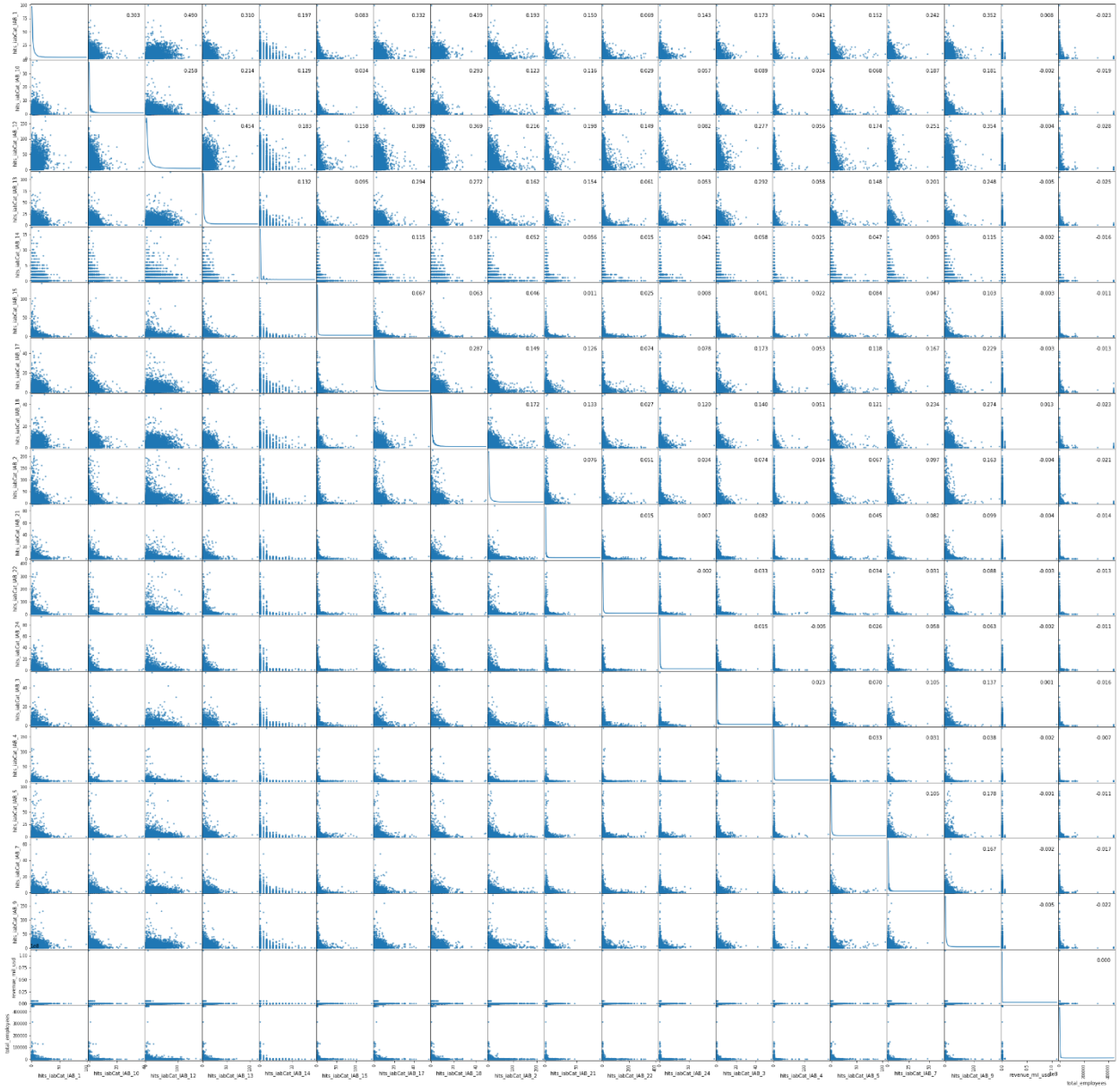
Results

Due to data distributions being so far away from a gaussian distribution with large skewness to the right, the following features were removed completely from any modelling due to unreliability and all data points being flagged as outliers:

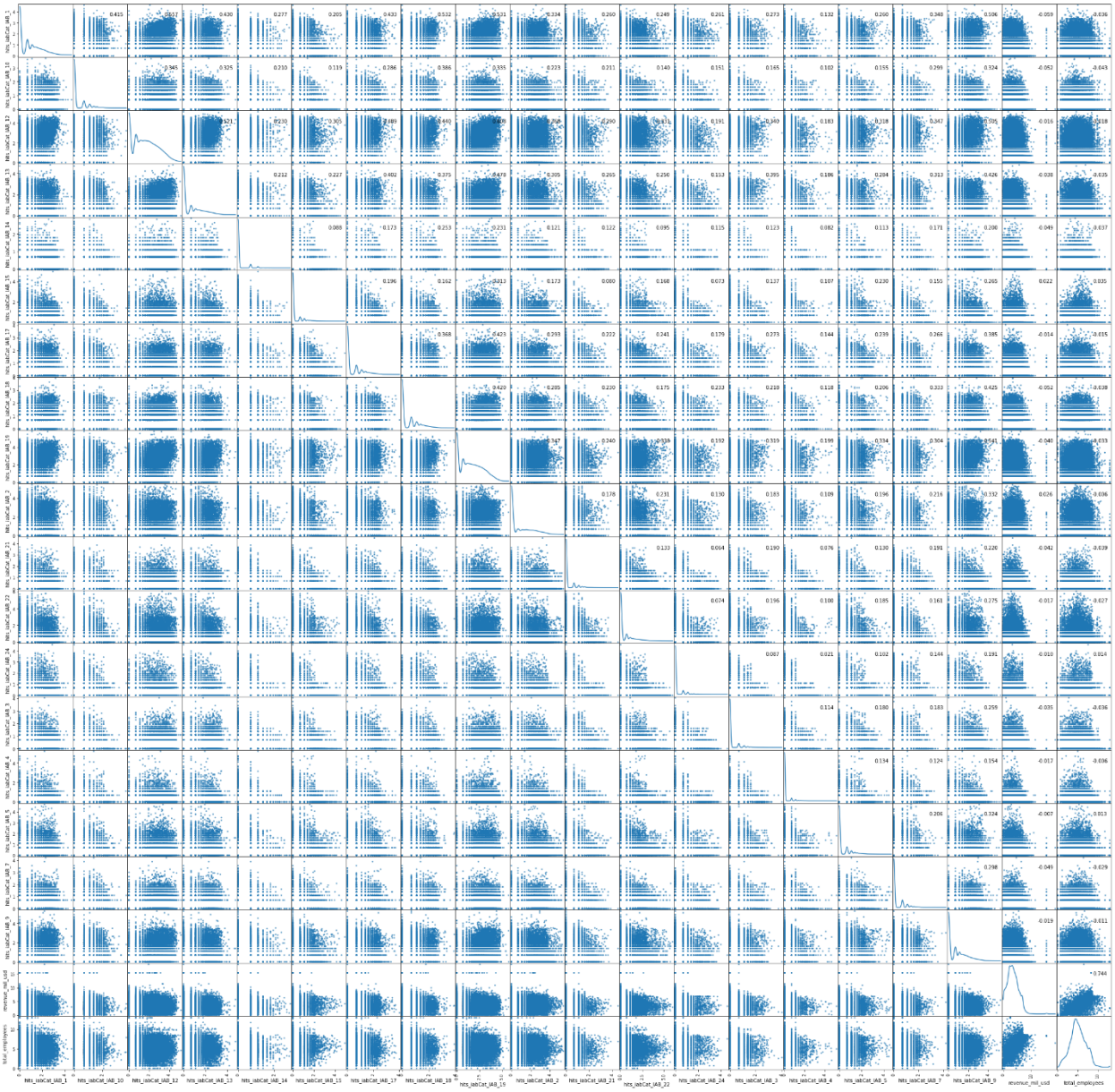
- clicks
- clickDates
- hits_iabCat_IAB_6
- hits_iabCat_IAB_8
- hits_iabCat_IAB_11
- hits_iabCat_IAB_16
- hits_iabCat_IAB_20
- hits_iabCat_IAB_23
- hits_iabCat_IAB_25
- hits_iabCat_IAB_26

The “zeroFillColumn” function was also necessary to null fill both continuous and discrete values to remove issues around missing data which result in modelling operations failing due to “NaN” errors.

Plotting all IAB category hits against each other, we can see again that there are minimal, if any, correlations between the categories. The distribution of all of the metrics also remain highly skewed



We see some improvement when transforming the data to a logarithmic scale, and as such clustering will be based on the log transformed data:



Implementation

After the data preprocessing steps, which were fairly intensive due to the nature of the dataset used, the following machine learning techniques were applied to the cleaned dataset in attempt to derive some significant inferences.

Decision Tree Regression

A decision tree regression model was applied to the dataset to help highlight that there was no significant predictive power between the IAB category activity metrics, as suspected by the initial visualizations.

The continuous features of the dataset were taken and the data was split into training and testing sets (25% for testing) and the "Decision Tree Regressor" model was applied in an attempt to predict the amount of activity in IAB category 19 based on the activity in all other IAB categories.

Using "fbeta_score" as the scoring metric for model evaluation, a score was consistently returned close to zero (last run was 0.3). From this I could conclude that there was no significant predictive power from activity in one IAB category to another and that unsupervised techniques should be investigated in order to help drive future analysis by finding a starting point for what to look for in as a useful source of predictions.

Dimensionality Reduction

Due to the large number of features available, 26 IAB categories, principal component analysis (PCA) was applied to the dataset in order to capture as much variance in the dataset as possible in the least number of dimensions, to both aid in reducing modelling noise but also to help computationally.

Applying PCA to the dataset I found the following breakdown of variances:

Dimension	Explained Variance
Dimension 1	0.3709
Dimension 2	0.6207
Dimension 3	0.6778
Dimension 4	0.7289
Dimension 5	0.7659
Dimension 6	0.8005
Dimension 7	0.8299
Dimension 8	0.8561
Dimension 9	0.879
Dimension 10	0.8952
Dimension 11	0.9104
Dimension 12	0.9247
Dimension 13	0.9381
Dimension 14	0.9498
Dimension 15	0.9605
Dimension 16	0.9706
Dimension 17	0.9804
Dimension 18	0.9882
Dimension 19	0.9955
Dimension 20	0.9999

As result I decided to reduce the dataset down to 10 dimension so that ~90% of the variance found in the dataset was represented but also reduced the name of dimension in half, 20 down to 10.

Clustering

Utilizing the reduced dimensionality dataset generated from PCA, along with the log transformations, I began trying to apply clustering algorithms to the dataset to highlight any potential groupings which could be treated with distinct actions based on unique collection of activity behaviors.

Based on the Silhouette score, I have found that best fit possible for the data was 3 clusters created using the k-mean algorithm. The best score generated being ~0.25. Breakdown by Clusters:

Clusters	Silhouette score
2	0.209323431
3	0.245920294
4	0.189204729
10	-0.203943332

Results

Model Evaluation and Validation

Using k-means to try and determine if there are distinct cluster of companies to be found within this dataset, based on their web activity on website of different categorizations, as per the IAB taxonomy, there appears to be no distinct behavior of companies. A silhouette score of 0.25 highlights this along with the clustering visualization above which clusters with large overlaps.

Justification

Next Steps for the Business:

- There are clearly some serious data quality issue here. The business needs to decide if it is worth investing in monitoring any of the blacklisted categories or a decision should be made to stop tracking activity on blacklisted IAB categorized website
- Click data needs to be increased in order to drive any significant machine learning inferences going forward. Right now we not have significant data accross all accounts. I'd suggest we invest in serving display ads to all companies in our database to be building up a baseline for activity and help drive future analysis.
- There are some serious outliers, which should be taken into consideration with the existing scoring models, as they could skew the entire accounts list. Highly likely they will always bubble to the top of client accounts

Further Analysis:

- Can any inferences be made at a high level of granularity other than company, e.g. at the country level?

Conclusion

Free-Form Visualization

This was the output of the "best" model. Along with the breakdown of the centroids by feature value.

	Segment 0	Segment 1	Segment 2
hits_iabCat_IAB_1	1	1	4
hits_iabCat_IAB_2	1	1	4
hits_iabCat_IAB_3	0	0	1
hits_iabCat_IAB_4	0	0	0
hits_iabCat_IAB_5	0	0	1
hits_iabCat_IAB_7	0	0	1
hits_iabCat_IAB_9	1	0	4
hits_iabCat_IAB_10	0	0	1
hits_iabCat_IAB_12	2	2	17
hits_iabCat_IAB_13	0	0	3
hits_iabCat_IAB_14	0	0	0
hits_iabCat_IAB_15	0	0	1
hits_iabCat_IAB_17	0	0	2
hits_iabCat_IAB_18	0	0	2
hits_iabCat_IAB_19	2	2	18
hits_iabCat_IAB_21	0	0	1
hits_iabCat_IAB_22	0	0	2
hits_iabCat_IAB_24	0	0	0
revenue_mil_usd	218	7	22
total_employees	1024	42	122

Based on the visualization we can see that all of the accounts appear in one giant blob with no distinguishing features. The centroids do highlight that there are just accounts that have significantly higher activity overall, e.g. segment 3 has incredibly large centroid for IAB categories 12 and 19, which at least could be useful for the marketing agency to prioritize these accounts. However this inference could have been derived from more simple calculations, even in excel.

Reflection

I think the main reflection for me personally on this project is the amount of time and energy that goes into gathering and cleaning the dataset at hand. Also, that it is okay if no useful machine learning models can be derived as long as the issues leading towards them are highlighted clearly to the business so that development time can be allocated in resolving them. Furthermore, it is also useful to highlight that what may have been believed by the business anecdotally, e.g. IAB activity from one category can help predict another, is not actually correct and we need to go and focus our energy potentially else if we want real predictive power.

Improvement

In terms of improvements, outside of data cleaning and restructuring, I think I could improve by utilizing quicker analysis techniques such as visualizations and summary statistics to help fail quickly. I.e. Highlight early on if the project is likely to hit a dead end so that efforts can be focused on potentially more lucrative investigations.

Having said that, all the issues worked through due this project will greatly aid in any future work as the techniques seem extremely extendable to the vast majority of datasets I will encounter in my day to today, with function used here already applicable.