Cloud versus Non-Cloud Market Segmentation

Capstone Team 27

Edward Eustachon, Jeremy Friedman, Sachin Balakrishnan, Tushar Gupta

Overseen by Professors Tejwansh Anand and Dan Mitchell and NetApp's Ryan Maas

Executive Overview

Over the past year, NetApp has observed that many of its customers are beginning to move on from traditional on-source data infrastructures in favor of harnessing the power of cloud data storage. Thus, the goals of this project are to create a descriptive profile of recently migrated companies as well as predict the companies most likely to migrate next. By creating a descriptive profile of such companies, the marketing team would be able to create campaigns and outreach programs that target and cater to companies that specifically fit such descriptions. On the other hand, our predictions would enable the sales division to target the specific companies likely to move next rather than reaching out to all companies which would greatly decrease the costs associated with their efforts. Overall, both departments would be able to target a much smaller pool of companies who are more likely to be successful conversions in cross-selling and upselling efforts which would result in lower costs as well as higher success rates in their efforts.

In order to complete our project, we were provided with two data sources. First, a third-party provider of company level characteristics and data usage statistics such as employee count, total revenue, internal and external IT expenditure, as well as which major cloud provider they worked with. The second set of data consisted of internal NetApp product statistics which tracked the NetApp products and features present in a company as well as the usage of NetApp storage usage in various facets such as storage spent and storage remaining.

We applied this data to predictive models that would calculate the probabilities of switching to the cloud as well as clustering methods that would group together companies into market segments based on the similarities of the companies. Our analysis concluded that many companies on-cloud and likely to be on-cloud are larger companies in terms of employee size, revenues, and storage expenditure. Moreover, many companies on-cloud are often heavy users of NetApp's products with almost double the total amount of NetApp products present in the company.

Out of the 4,479 companies that were not on the cloud in our dataset we were able to identify less than 25% of such companies had characteristics similar to those already moved to the cloud (see Figure 1 below). Overall, we calculate a lift value of 1.46 which translates to an almost 50% higher rate of success than if we were to select companies at random. In terms of

business value, our project creates a salient pool of potential customers for NetApp to target in their sales and marketing efforts leading to a reduction of costs and manpower needed to conduct business compared to if NetApp had devoted their resources in reaching out to all 4,479 companies.

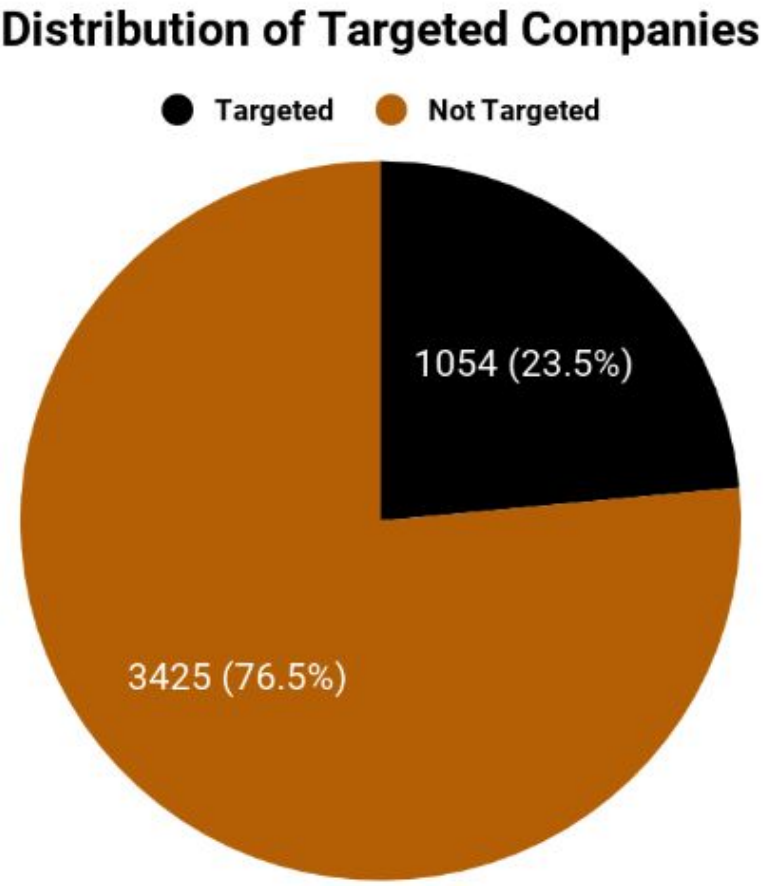

Figure 1: Distribution of targeted companies

Articulating the Business Outcomes

Business context

NetApp is a Fortune 500 company that positions itself as a “hybrid cloud data service” provider or more simply, an enterprise-level storage company[1]. This should not be confused as a

[1] NetApp. (2016, June 27). Retrieved from https://fortune.com/fortune500/2016/netapp/

cloud storage provider that will be a competitor to the big three of Microsoft Azure, Amazon Web Services (AWS), or Google Cloud Platform (GCP). Rather, NetApp specializes in helping firms take advantage of their data in order to foster IT transformation that continues into digital transformation.

In short, NetApp helps companies create and leverage personalized data storage solutions that create a competitive advantage for the firm. For instance, NetApp helps customers interface with their preferred form of public cloud storage be it Azure, AWS, or GCP through the use of NetApp integrated applications[2]. Moreover, NetApp helps firms build and integrate public and private data infrastructure as well as assist companies in running their current data fabric more efficiently. Thus, the status quo: NetApp is currently seeing a lot of customers move their data storage from traditional on-premise solutions to the cloud via public providers.

As mentioned before, NetApp has products that can assist customers in their move to the cloud but first must identify those who are expected to do so. If NetApp is able to successfully identify those likely to make the move they can further develop their relationship with existing customers. Furthermore, successful prediction/identification of those willing to switch to the cloud allows NetApp to develop descriptive customer profiles that can identify preferable market segments to target.

Business questions

Prior to the winter break, we had a preliminary meeting with Ryan Maas, our point of contact at NetApp, regarding introductions between the two groups as well as expanding upon the details of the project as per the prospectus. The meeting concluded with Ryan intending to finalize the project scope and assemble the data for the start of the Spring semester. When we initially received the dataset from Ryan it came with some corruption that Ryan needed to patch up, so we spent our first meeting of the semester discussing the audience of our analysis as well as how our findings would be consumed.

The business question/problem comes two-fold under the umbrellas of marketing and sales. The marketing question: what traits and features of a company are relevant in identifying

[2] About NetApp. Netapp.com. (2020).

those who are likely to switch to the cloud soon? Then the sales question: what users do we predict to be likely to move to the cloud? The questions almost go hand-in-hand in this context. NetApp would want to know which users are likely to require assistance in cloud migration. To do so, NetApp would want to also know what the key indicators of a company being likely to make the switch to the cloud are.

Business outcomes

Upon conclusion of the project, NetApp should have gained further understanding of their customers for both those likely and unlikely to move to the cloud soon. This would emerge from both the descriptive profile of various customer types as well as the actual prediction of which customers are most likely to do so. Additionally, NetApp would now have predictive models regarding this matter that they could productionize as well as fine-tune or even retool for different purposes.

Overall, NetApp should be able to target existing customers with a high propensity to move to the cloud and assist them in creating a data storage solution that fits their needs. Such business outcomes are further extensions of the firm's strategy in helping customers create efficient data storage architecture unique to their needs.

Integrating predictive power into their marketing and sales efforts allows NetApp to develop a further competitive advantage over other cloud migration services in the form of a first mover's advantage and potentially allow NetApp to capture a larger market share among data storage service firms. Another consequence of NetApp's success in helping customers move is that NetApp could further develop their partnerships with the big three cloud providers.

Due to the nature of the project, our data was assembled in such a way that many characteristics of companies as well as the NetApp products they utilize are hidden. Because of this, it is tough to quantify the ROI or expected business value of the project. However, when we brought up this issue to Ryan he assured us that as the project went further underway he would be able to tell us what such features or products maybe since it would aid us in creating a strong presentation before NetApp and the class. When we receive such information, we expect to be able to quantify and measure the impact of our project in a much more concrete manner.

Explore Source Data

NetApp has provided us with a collection of tables that record various characteristics of both companies and NetApp storage products in use. This data comes from two separate data sources: first, the tables containing data about NetApp storage solutions contain data about storage capacity and usage and the features/products being used by NetApp storage products. This data is internal data collected directly by NetApp. They also provided a data table of companies and their storage expenses in the past year. Otherwise, company data was collected by a 3rd party known as HG Insights. This data contains numerous aggregate statistics of the firm such as country, region, industry, and employee size. HG Insights also collected data about the company's internal and external IT budgets and expenses. Most importantly, they provide information about whether the company is currently using any of the big three cloud storage providers (AWS, Google Cloud, and Microsoft Azure).

Overall, the data provides information about companies and how they use NetApp storage solutions if they do at all. This includes the type of NetApp products/features as well as storage usage statistics. Because this data is being used to identify companies with a high propensity to switch to the cloud and why they switch it also provides numerous company level characteristics of the data such as employee size, total revenue, country, industry, and even IT budgets/expenses. This is important because the decision by a company to migrate data fabrics is a company-wide decision that requires consideration of the company's overall status.

When initially receiving the data there were a few corruptions in the tables that came as a result of converting the data into different formats for consumption. These issues were addressed by Ryan, our client lead, in a prompt fashion. Aside from some minor data corruption issues, the data came in an easy to ingest manner that required only slight data manipulation before it was ready to be run through the wringer. The only piece of work left for us before beginning our analysis was to find an adequate way of consolidating the tables since some tables had company identifiers while others used unique identifiers of NetApp storage products (likely hardware IDs) that belonged to a company.

There is definitely information useful in answering the problems set out by NetApp. For instance, the inclusion of geographic data allows us to consider whether cloud usage may be related to a regional bias. Furthermore, because of how NetApp's product line is structured, we

would expect users more invested in NetApp's bundle offerings to have a higher likelihood of cloud migration than those who are less or not at all invested in NetApp products and features. Additionally, perhaps cloud migration is catered towards specific industries or company sizes; such data is also provided to us by NetApp. The only other issue with the data, at this point, is that much of the features have had their true meanings and identities masked for the sake of security. However, as we move forward in the project we have been reassured by Ryan that the meaning or identities of the most important features would be provided to us further down the line in advance of preparing our final presentation and report.

The first issue that needed to be addressed by the team was how to reconcile some of the data being identified at a company level while other data was identified at a unique identifier level. We believed that in our business context it was more practical to aggregate the data at a company-wide level. Our reasoning is this: because we are most interested in identifying companies likely to migrate as well as the descriptive profile of such companies then it is best to aggregate at a company-level rather than product-level. After rolling up the data at the company level, we found that there were some duplicate data entries, which had to be dealt with. Feature_1, Feature_2, Feature_3, and Feature_4 files each had 29 duplicate entries while Product and Storage files had 2 duplicate entries. Since these duplicates were not significant in number, we omitted them from our dataset without any further investigation.

After we aggregated our data we were then left with the tasks of feature engineering and data imputation to a satisfactory level. While merging our data we had decided to create a number of columns that would aggregate much of the unique identifier information under one company. For example, we had four separate tables that recorded information about the presence of a feature or product in a NetApp storage device. Because not all devices will have the same feature or product offerings it is important to retain such information while aggregating the data. Therefore, we created "sum" columns for the numerous product and feature traits to keep track of such information even at the company-level. However, an issue with using the sum is that it is insensitive to the magnitude of NetApp devices in use at the company. For instance, some companies should be expected to have larger sum columns simply because they are larger and therefore use more NetApp storage. In response to this, we created percent columns as well that would track the product/features' relative presence within the company. Furthermore, we engineered a binary indicator "flag" that would mark if the product or feature was present in the company at all. Doing so allows us to control for the

importance that the mere existence of a product may have rather than its relative or total presence within the company. By creating a sum, percent, and binary variable we were able to measure a company's "stickiness" to NetApp in three different dimensions allowing us to precisely identify the importance of a product.

Another feature engineering consideration was how we would encode the numerous countries as a feature in the data. NetApp's business spans across the globe with activities involved across all continents. We therefore had to think of a way to control for the possibility of a geographic effect at play. During our EDA we discovered that the five most prevalent countries accounted for 64% of the companies in our dataset. Consequently, we decided to create indicator variables for these five countries as a way to "control" for the possible signal in geographic data.

The final issue to address was the presence of missing values in our data and how we would impute such data. We opted to proceed with a K-nearest neighbors approach in our missing value imputation[3]. This seemed rational to us since our predictive models and clustering algorithms would be comparing the various traits of a company with those around them to identify them as cloud or non-cloud users. Furthermore, we assumed that companies similar in size, in terms of revenue, employee count, and IT expenditure, would be likely to have similar values in other aspects. Therefore, we believed it to be practical that any missing values of a company be imputed by assigning its value to the average value of the 10 companies it was most similar to. In our imputation, the mean was calculated where we weighed each of the 10 companies uniformly and used Euclidean distance as a measure of similarity among the feature space.

Analyses (Modeling, simulation, and optimization)

We divided the analyses into two parts:

- Supervised Learning Approach
- Unsupervised Learning Approach

[3] Obadia, Y. (2020). The use of KNN for missing values. Medium.

Supervised Learning Approach

We initially began analyzing the data from a supervised learning approach. This involved framing the business problem as a question of classification i.e. construct a model that will be able to predict whether the company uses cloud storage or does not with the given traits and features in our dataset. We believed that this framing was proper for the problem since NetApp was interested in identifying companies likely to move to the cloud as well as the important variables that would push them to consider cloud storage. In summary, we believed that a supervised approach was appropriate for the business problem since we were not only interested in making predictions but also finding what variables were important in making such predictions.

To this end, we turned towards using ensemble tree methods since we could calculate feature importance as well as estimate the numerical probabilities of cloud movement for each company. At the time we conducted this analysis, we thought we might be able to obtain data about the costs and revenues associated with marketing and successful conversions. Using the calculated probabilities and such data, we could actually calculate the expected profit of each company and create a more refined targeting campaign as opposed to simply reaching out to companies that have switching probabilities exceeding 0.50.

After training multiple models including Random Forest, Gradient Boosted Trees, XGBoost, and tuning hyper-parameters using the grid-search approach, Random Forest gave the best AUC ROC. Random Forest grid-search results are mentioned below:

	max_depth	**ntrees**	**model_ids**	**logloss**
0	10	800	rf_grid2_model_9	0.57225756
1	10	500	rf_grid2_model_5	0.57259712
2	15	500	rf_grid2_model_6	0.57264618
3	10	1000	rf_grid2_model_13	0.57269262
4	15	800	rf_grid2_model_10	0.57270742
5	10	200	rf_grid2_model_1	0.57271359
6	15	200	rf_grid2_model_2	0.57290246
7	15	1000	rf_grid2_model_14	0.57329577
8	20	800	rf_grid2_model_11	0.57460769

9	20	500	rf_grid2_model_7	0.5746655
10	25	500	rf_grid2_model_8	0.57501988
11	20	1000	rf_grid2_model_15	0.57503356
12	25	800	rf_grid2_model_12	0.57503548
13	25	1000	rf_grid2_model_16	0.5753502
14	20	200	rf_grid2_model_3	0.57627545
15	25	200	rf_grid2_model_4	0.57663146

Figure 2: Grid search cross-validation results for our Random Forests

The best model gave us an AUC ROC of 0.74. The confusion matrix can be seen below:

Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.3749952957034111:

		0	1	Error	Rate
0	0	534.0	468.0	0.4671	(468.0/1002.0)
1	1	146.0	715.0	0.1696	(146.0/861.0)
2	Total	680.0	1183.0	0.3296	(614.0/1863.0)

Figure 3: Random Forest confusion matrix

As we can see in the confusion matrix, the classification approach gave us a lot of false positives (type 1 errors). This meant that many companies that were not on cloud showed characteristics quite similar to the ones on the cloud, which made our model misclassify them as positives. This is why we decided to scrap this approach. At the same time, we were advised by Professor Dan Mitchell to consider an unsupervised approach to the data. This made sense as well since if we could identify the dimensions that made companies more likely to use cloud solutions then we could simply target companies that had similar dimensions or geometry. However, we did use the variable importance provided by the Random Forest model in the unsupervised approach.

Unsupervised Learning Approach

With our unsupervised learning approach, we initially began by attempting to use Principal Components Analysis to see if we could reduce the dimensionality of the dataset. We had roughly 250 variables after conducting our previously mentioned feature engineering. When

using PCA on the entire feature set, we wanted to capture 80% of the data's variance with fewer dimensions but were unsuccessful in doing so. PCA still gave us around 100 components to work with. We chose to abandon pursuing PCA since we believed that it was important to retain interpretation given we would be presenting to people that were not versed in data science methods.

That being said, we believed 250 features to be far too many still and decided to apply the variable importances calculated from our Random Forest model as a means of variable selection. We considered only the 50 most important variables and proceeded forward with trying to cluster the data around these columns.

The 10 most important variable can be seen below:

variable	relative_importance	scaled_importance	percentage
internal_total	8270.704	1	0.045269
total_revenue	8060.357	0.974567	0.044118
US	7886.864	0.95359	0.043168
external_budget_by_total	7393.417	0.893928	0.040468
capacity_utilization	6686.136	0.808412	0.036596
external_vs_internal	6680.139	0.807687	0.036563
it_budget_total	6171.915	0.746238	0.033782
total_features_per_identifier	6055.726	0.73219	0.033146
total_employees	5825.788	0.704388	0.031887
total_200_features	5596.074	0.676614	0.03063
max_free_capacity_tb	5547.119	0.670695	0.030362

Figure 4: Top 10 feature importance from our Random Forest model

Something of note is that we noticed an extremely disproportionate distribution in countries. The US accounted for nearly 70% of companies in the data. To control for this imbalance, we decided it appropriate to only cluster US companies with other US companies. Also, the US variable was coming out to be the most important variable in the Random Forest model. Thus, we conducted 2 clustering analyses, one for US companies and another for non-US companies to see if there would be a significant difference in how the clusters are composed in each set.

When clustering the data, we considered a variety of clustering methods. This ranged from K-means clustering, multiple flavors of hierarchical clustering[4] and spectral clustering[5], Gaussian mixture clustering as well as DBSCAN clustering[6]. This was done for two different reasons: first, we wanted to see if the clustering between each method varied greatly and secondly, if the clustering was consistent we reasoned that the clustering did well in discovering variable importance and was, therefore, a valid approach to the marketing and sales campaign for NetApp.

We have now established and justified the two approaches we took with the data and can now discuss their results and findings. In this section, we will depart from the heavily technical language and focus on the business implications of our analysis. First, we will discuss the business implications themselves and then try to rationalize why our findings are valid and worth consideration.

As we mentioned previously, while we initially began with two approaches, we scrapped our supervised modeling approach in favor of unsupervised clustering. This is because we were unable to tune our models to have neither strong accuracy nor strong class separation. Thus, our business implications emerge from the analysis of our clustering algorithms.

When interpreting our clustering, we noticed the most important dimensions were as follows: internal IT spending, external versus internal IT spending, total employees, “stickiness” to NetApp products, and the percentage of software using a feature whose definition is masked from us. Based on these features, many companies that were recorded as cloud storage users often had high IT spending although there were mixed signals on whether it was important the spending was external or internal IT spending. Next, many cloud users were associated with higher counts of employee size and had more invested in NetApp products and services than those not on the cloud. Unfortunately, for the last few important variables, the meaning of these exact features is masked from us.

Many of our most important dimensions are fairly intuitive which should signal confidence in our findings. One would not need a data science team to think of these relationships although a data science team could concretely support such findings through

[4] 2.3. Clustering — scikit-learn 0.22.2 documentation. Scikit-learn.org. (2020).
[5] Doshi, N. (2020). Spectral clustering. Medium.
[6] Chauhan, N. (2020). DBSCAN Clustering Algorithm in Machine Learning

similar methods of analysis as ours. Overall, we can see that the companies that most often move to the cloud are businesses that spend higher than average on their IT infrastructure. Additionally, such companies often have larger than average total employee counts suggesting that the option to move to the cloud may be associated with the amount of data they ingest. Unsurprisingly, many companies on the cloud are also already invested more than normal in NetApp products and services which makes sense given the company values itself upon providing a friendly enterprise experience with data storage solutions.

In summary, we have created tables to define the clusters that we created using the approach mentioned above. These are the clusters for US companies:

Cluster	# Companies	Definition	Internal Total	External vs Internal	Total Employees	Total 200 Features	Emp2	Ind9
0	753	SMEs	mod internal total	low external vs internal	mod total employees	mod total 200 features	does not belong to emp2	does not belong to ind9
1	126	SMEs with High External Budget	low internal total	high external vs internal	mod total employees	low total 200 features	does not belong to emp2	does not belong to ind9
2	379	Start-ups	low internal total	low external vs internal	low total employees	low total 200 features	belongs to emp2	does not belong to ind9
3	178	Manufacturing	mod internal total	low external vs internal	mod total employees	low total 200 features	does not belong to emp2	belongs to ind9
4	82	Brand Ambassadors	high internal total	low external vs internal	high total employees	high total 200 features	does not belong to emp2	does not belong to ind9
5	1	Outlier	extreme internal total	low external vs internal	extreme total employees	low total 200 features	does not belong to emp2	does not belong to ind9

Figure 5: US company cluster definitions

Cluster	# Companies	Definition	Internal Total	External vs Internal	Total Employees	Total 200 Features	Emp2	Per Feature 307
6	929	Non-Manufacturing SMEs	low internal total	mod external vs internal	mod total employees	low total 200 features	does not belong to emp2	high per feature 307
7	316	Manufacturing SMEs	low internal total	mod external	mod total employees	low total 200 features	does not belong to	high per feature

				vs internal			emp2	307
8	341	Start-ups	low internal total	mod external vs internal	low total employees	low total 200 features	belongs to emp2	high per feature 307
9	96	Large companies	high internal total	low external vs internal	high total employees	mod total 200 features	does not belong to emp2	mod per feature 307
10	74	Pro-cloud Large Manufacturing	mod internal total	high external vs internal	high total employees	mod total 200 features	does not belong to emp2	mod per feature 307
11	33	Pro-NetApp	mod internal total	low external vs internal	mod total employees	high total 200 features	does not belong to emp2	mod per feature 307
12	92	High Potential	mod internal total	mod external vs internal	mod total employees	low total 200 features	does not belong to emp2	zero per feature 307
13	106	High External Spenders	low internal total	extreme external vs internal	mod total employees	low total 200 features	does not belong to emp2	mod per feature 307
14	1	Outlier	extreme internal total	high external vs internal	low total employees	mod total 200 features	does not belong to emp2	mod per feature 307
15	1	Outlier	high internal total	high external vs internal	extreme total employees	mod total 200 features	does not belong to emp2	mod per feature 307

Figure 6: Non-US company cluster definitions

Results and Findings: Cluster Performance

Now that we have completed tagging the cloud companies into different clusters, it is time to evaluate how well the clustering performs when exposed to unseen data. We will make use of the data from the holdout set for this. There are 1000 companies in the holdout set in total which we will try to cluster and predict their propensity to cloud migration.

Note: The below table is for illustration purposes only. The cloud labels were not used at the time when the companies in the holdout set were clustered.

Category	Cloud	Non Cloud	Total
US	220	86	306
Non-US	275	419	694
			1000

Figure 7: Holdout set - US vs Non-US companies

Our clustering assigned 71 of the 306 US companies to one of the 5 clusters, predicting a high propensity for cloud usage for those. After validation, we found out that 52 of those 71 companies were actually on cloud, which gave us a prediction accuracy (Precision) of 73%. Here is the precision rating (percent_TP field) for the different clusters on the US holdout set.

cluster	total_count	percent_TP	median_distance
0	32	81.25	1.6420
1	5	100.00	4.6895
2	18	44.44	4.1330
3	11	72.73	4.1170
4	5	100.00	7.0255

Figure 8: Cluster performance - US holdout set

Coming to the 694 non-US companies in the holdout set, our algorithm tagged 47 of those to one of the cloud clusters. Out of the 47 identified, 28 companies actually turned out to be on cloud, which gave us a Precision of 60% for non-US companies. Here is the precision rating (percent_TP field) for the different clusters on the non-US holdout set.

cluster	total_count	percent_TP	median_distance
6	15	60.00	1.5830
7	9	66.67	3.1060
8	8	25.00	4.1250
9	5	80.00	5.7300
10	2	100.00	6.1900
11	2	50.00	10.5100
12	1	100.00	9.5400
13	5	60.00	11.3875

Figure 9: Cluster performance - Non-US holdout set

Having gotten a respectable precision rate amongst our holdout set, we took a collection of non-cloud companies and mapped them to the clusters. Of the 4,479 non-cloud companies we were provided with, 1,054 were successfully mapped to one of the clusters. This subset contains ~23.5% of the total market, a small enough collection of companies to successfully reduce exorbitant marketing expenses while big enough to offer a wide selection for the sales team to consider. Given the precision rates found above, companies belonging to this subset can be expected to have a lift of 1.46 when considering their propensity for migrating to the cloud.

Descriptive Profiling

The second objective of our Capstone project was to build a descriptive profile of the companies that are on cloud. We conducted quite a bit of exploratory data analysis on the data to understand how the distribution of the predictor variables are varying across cloud and non-cloud companies. The results we got were fairly consistent with the feature importance from our Random Forest model.

Large-sized companies in terms of budget allocation, employee count, and total revenue prefer cloud to non-cloud. We have 4 different feature series (100, 200, 300, and 400) consumed by NetApp customers. Our analysis results suggest feature 200 series seem to be the most decisive one when it comes to distinguishing between cloud and non-cloud companies. We can

see from the fourth plot below, the usage of series 200 features for cloud companies is twice compared to non-cloud.

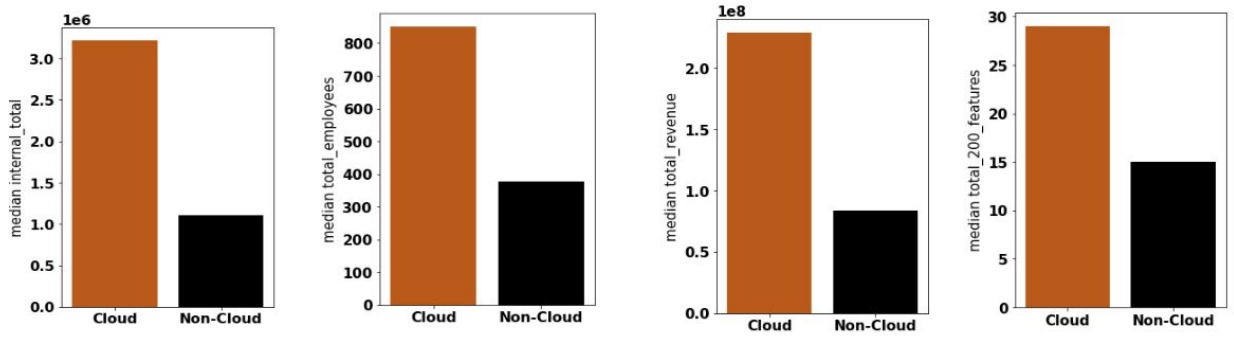

Figure 10: Differences between cloud and non-cloud companies on select features

Recommendations, Operational Execution, and Change Management

Despite currently being non-cloud, the subset of 1,054 companies we have identified possess an above-average propensity to becoming users. As such, it is our recommendation that NetApp focus on these companies in future marketing efforts, as they can expect an almost 50% higher number of sales than if they selected their targets at random. When considering potential customers that were not in the provided dataset, we further suggest NetApp to compare each to the thirteen company profiles we have specified, which should similarly improve their sales volume.

A more specific plan is as follows: the sales team should be sent the list of companies we have identified as ideal candidates for their own review. They should then proceed to market cloud services to these companies, beginning with whichever ones they feel would be the most profitable. While our research identified which clients were more likely to buy, NetApp's sales team undoubtedly has a better understanding of which companies, if made cloud customers, would yield the greatest returns. We thus encourage them to use their discretion when choosing from our provided list. By doing this, NetApp's sales are certain to improve quickly because the targeted companies are more likely to buy. Instructing the sales team to use their own judgement when prioritizing which companies to start with also facilitates change management, as people are less opposed to sudden change when they have some measure of control over what happens next. They can follow our recommendations on their own terms.

The sales and marketing teams would also be sent the descriptive profiles of the on-cloud companies we built. For the marketing team, this information would prove useful when reaching out to companies NetApp does not currently provide new services for. When tailoring the sales pitch to these potential clients, knowing whether they match a profile would help determine how much additional emphasis they should put on describing their cloud services. A company that meets the profile is more inclined to use cloud, so bringing it up in the initial meeting could improve the odds that they become a new customer. Should the company initially decline cloud but still become a client, the sales team will be able to mark them early on as good targets for later convincing.

Conclusions

The project set forth by NetApp was to find patterns in company characteristics and NetApp product usage between companies on the cloud and those not. We then applied these patterns to companies not on the cloud to predict those who would have a higher propensity to migrate to cloud usage. We were provided with third party data of such companies that included company-level statistics such as IT expenditure, total revenue, cloud usage, and employee size. Additionally, we used NetApp collected data that tracked the presence and usage of numerous NetApp products and services in order to research whether NetApp's product line affected company decisions to move to the cloud.

Initially, we attempted supervised learning methods like ensemble decision trees to both identify important signals in the data as well as predict whether or not companies were to switch to the cloud. However, due to an unsatisfactory AUC score and severe false positive rate we had to approach the data from another perspective. To this end, we conducted unsupervised clustering analyses on our training data to group together similar companies based on their data and cloud usage. We then mapped non-cloud companies in our testing data to our newly created clusters to identify the various types of clusters that had a majority of data points situated on the cloud.

Note that although we conducted two separate clustering procedures based on US/non-US designation of the companies our clusters were very similar in definition despite the differences in test set performance. Our cluster definitions indicated that many of the companies on the cloud were often companies larger in size as evidenced by higher revenues,

employee count, and IT expenditures. Moreover, we also discovered that NetApp's feature 200 series was used two times as much by companies using cloud storage solutions than those not.

Based on our clustering, we were able to identify 1,054 of the 4,479 non-cloud companies in our dataset that had a high propensity to move to the cloud and ought to be targeted in NetApp's cross-selling and upselling efforts. By using our targeted selection we expect NetApp to have an almost 50% increase in sales and marketing efforts compared to if they reached out to companies at random.

References

2.3. Clustering — scikit-learn 0.22.2 documentation. Scikit-learn.org. (2020). https://scikit-learn.org/stable/modules/clustering.html.

About NetApp. Netapp.com. (2020). https://www.netapp.com/us/company/about-netapp/index.aspx.

Chauhan, N. (2020). DBSCAN Clustering Algorithm in Machine Learning - KDnuggets. https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html.

Doshi, N. (2020). Spectral clustering. Medium. https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7.

NetApp. (2016, June 27). Retrieved from https://fortune.com/fortune500/2016/netapp/

Obadia, Y. (2020). The use of KNN for missing values. Medium. https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637.

Schelling, B., Plant, C. Dataset-Transformation: improving clustering by enhancing the structure with DipScaling and DipTransformation. Knowledge and Information Systems 62, 457–484 (2020). https://doi.org/10.1007/s10115-019-01388-5

Acknowledgments

Thank you to Professors Tejwansh Anand, Daniel Mitchell, and Jim Griffin for their oversight during the course of our Capstone project. Without your comments and feedback we would never have considered a clustering approach as the solution let alone been able to present our findings and insights as well as we did. Thank you to Ryan Maas and the rest of the NetApp team that we met during the project. Being able to present and dry-run our thought process throughout the project was a boon to our success. It was a gift having such a solid working relationship as well as the guidance provided throughout the semester, and we hope that you continue to sponsor the MS programs' capstones for years to come.