Final Project, DAT\_19\_SF Colby Schrauth

**Customer Segmentation with K-Means Clustering and Principal Component Analysis**

I work for a company called Lithium Technologies. We’re headquartered in San Francisco, and sell forum software to businesses. We commonly refer to our customer forums as communities. For example, the [GoPro Community](https://gopro.com/news/tags/community) is built on top of Lithium technology.

Lithium technology is software-as-a-service (SaaS), which means our success is dependent on our ability to keep the customers we have. If a customer stops using our technology, we refer to them as ‘churn’. This project focuses on the following question:

*What do churn customers have in common, and how do these similarities contrast to customers that consistently renew their contracts?*

My hypothesis is that there is not a magic formula to completely eliminate churn from our business. Not yet at least. However, I do believe this implementation will represent our customer data in a way that’s never been done before—helping our Customer Success Team have more provocative conversations about the state of an account.

For this project I gathered the following data on each of our customers:

* Case Volume – number of support cases opened
* # of Languages – number of languages the forum supports
* Account Age – age of the customer
* # of Features – number of additional products purchased
* Geographic – area in which the customer is headquartered
* Support Tier – support package purchased
* Executive Sponsor – whether or not a Lithium executive sponsored the customer
* SSO Enabled – is there a connection between their forum and CRM

Most of this data was collected from our CRM system (Salesforce). For the most part, each data point lived in a different reporting schema within Salesforce. Custom reports we’re built, and internal analytic systems we’re used to maintain a unique identifier (account\_id) for each extract. At a macro-level, the workflow went as such:

*Reporting System > CSV Export > Excel Manipulation > SQL Import > Table Join(s) > CSV Export*

I chose these features to provide the algorithm with a breadth of account information. This feature-set includes location, purchase, usage and integration data. I had a handful of other features to bring in, however there would have been a massive sacrifice in data integrity had I done so. For example, contract price was a feature containing over 50% null values.

In exploring the dataset, I learned that ‘Account Age’ is forever ticking. In other words, this data point is not referencing the day in which the server instance was turned off. An account that has churned two years ago is still ‘live’ and getting older each passing moment. I also learned that the ‘# of Features’ data point was associated to a purchase, versus whether or not the customer was actually using the feature in its production forum.

For all categorical features (Geographic, Support Tier and Executive Sponsor) dummy columns we’re created in Python. Also, any non-binary data (Case Volume, # of Features, # of Languages, Account Age) was normalized using scikit-learn’s StandardScaler. A column denoting churn was also in the dataset, but ignored from the algorithm implementation. This was done so results could be distributed by whether or not the customer had churned.

I chose K-Means to group the customers, and started manually with 4 clusters. Looping through a range (2 – 20) of potential cluster amounts, I ran a visualization via matplotlib to identify the silhouette score for my manually chosen cluster amount, and how it compares to each of the other iterations. Ultimately the algorithm identified 8 clusters as the local maxima.

I then used Principal Component Analysis, forcing the dimensionality down to 2 features, in order to plot the results visually. This was only learned after hours of visualization attempts on all features!

Challenges included data integrity and availability. However, to my knowledge, this is the first time as a company we’ve used our account data to identify broader trends in the business. I experience this as a massive success. In turn, the immediate extension of this project is to transform our logging habits and experiment with results.