Classifying Industry at Signup

Author: Annika Sougstad

Claps: 54

Date: Sep 11

Klaviyo helps accounts personalize communication with their customers at scale. Onboarding asks this same question internally: How do we (Klaviyo) personalize communication with *our* customers? To deliver personalized onboarding experiences, we need information about our new signups.

I joined Klaviyo as a data scientist last year, right out of college. My first project was to figure out if we could classify accounts by industry as soon as they signed up. (This was before ChatGPT, which is relevant because youâ€™ll see below that the tool we turned to was not a large language model.)

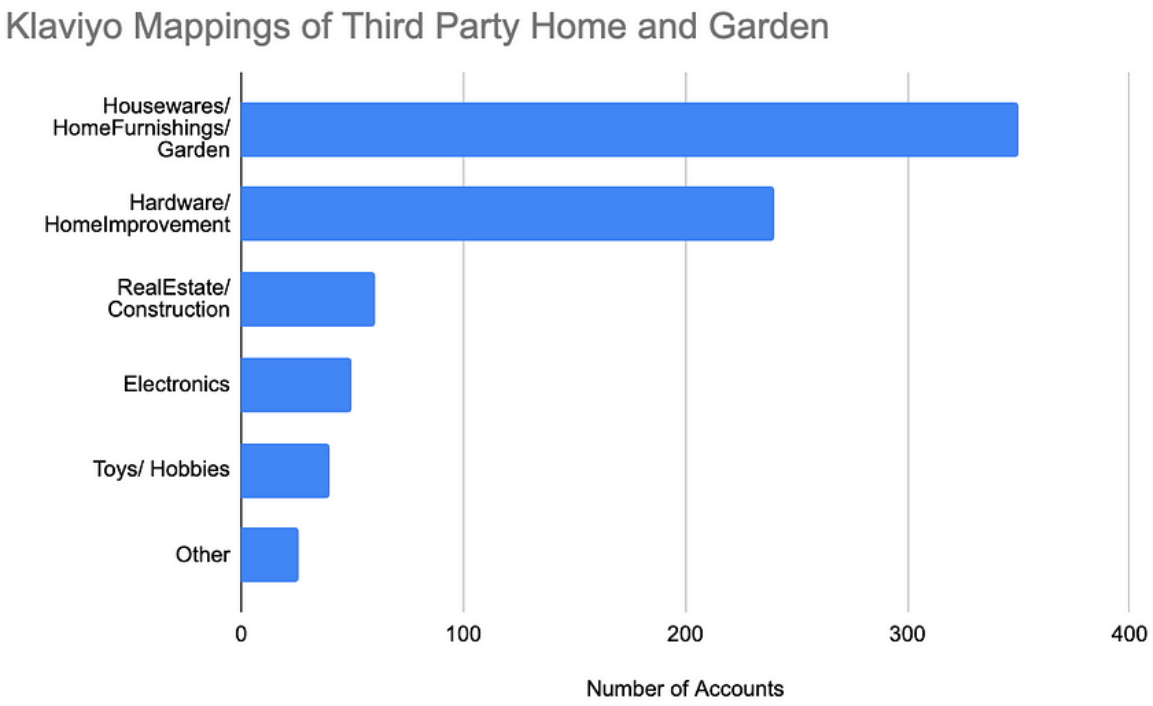
When I started on the project, we had about 10,000 labeled data points. The majority (7,000) of these came from accounts self selecting their industry, supplemented by some hand labeled accounts. I did not appreciate how nuanced categorizing businesses was until I started looking at examples. Accounts frequently did not cleanly fit into one category â€” often many categories were appropriate.

We also did not believe all of our data was labeled perfectly, as there can be multiple â€œcorrectâ€ ways to label an account. This meant that achieving near perfect accuracy as measured by our labeled data didnâ€™t make sense. Our initial goal was to achieve â€œreasonableâ€ accuracy so we could do analysis by industry across all of our accounts (where we could tolerate imperfect classifications) and sort industries for users to choose from.

Third-Party Data

We had the option to use a third-party commercial database already employed by our sales and marketing teams. Initially, this seemed like a really simple solution â€” we could just map their industry categories to ours. However, there was significant lack of industry coverage and inconsistency from this outside source.

Importantly, the third-party data almost exclusively represented medium to large ecommerce stores; but many of our accounts are newer stores or in non-ecommerce industries. Furthermore, their categorizations did not align especially well with our Klaviyo mappings. For example, here are the Klaviyo classifications for a sample of sites categorized as Home & Garden by the third-party source:



So the problem wasnâ€™t as simple as looking up companies in a database.

We decided the best approach was to build a machine learning model, and to treat the third-party data as one of many features. (If not familiar with models, a feature is an input variable.) We ended up using XGBoost and since it handles nulls well, we didnâ€™t need special treatment for sites that werenâ€™t in the database.

Website Features

How would a human classify a business or organization? I, and Iâ€™m sure most people, would visit the homepage and look at the pictures and text. Thatâ€™s where we started with our model. We scraped text from the homepage including alt text for images.

To vectorize this text for our model, we compared word2vec, TF-IDF, and sentence embeddings. We found that old, trusty TF-IDF (Term Frequency - Inverse Document Frequency) vectorization performed best.

TF-IDF is a measure for how relevant a word is to a document (website) within a larger set of documents (websites). Specifically, it combines how many times a word appears in a document (term frequency) and how unique the word is across a set of documents (the inverse document frequency). In our model, we generate a single vectorizer of unigrams (single words) and bigrams (two consecutive words) from the corpus of site text. We decided to include 5,000 features from these embeddings based on performance plateauing.

We noticed some surprising misclassifications when testing within a particular category: *Education*. Some websites were easily classified as they were clearly offering learning materials and had a lot of education related words on their site. However, some universities (e.g. mit.edu) were classified as *Events/Entertainment* because they were promoting events going on around campus for a majority of their homepage.

This gave us the idea to include the top level domain of the website as a feature. If a website is using .edu, we can be pretty certain they are in the education industry. This particularly improved recall for education, government, and software industries.

Other features from the website are â€œtechnologiesâ€ detected on the site (e.g. javascript snippets), as these can give a signal as to what industry the website is in (ChowNow, for example, is almost exclusively found on sites for restaurants or food stores). We check for ~400 of the most common technologies, and use these as additional features to the model.

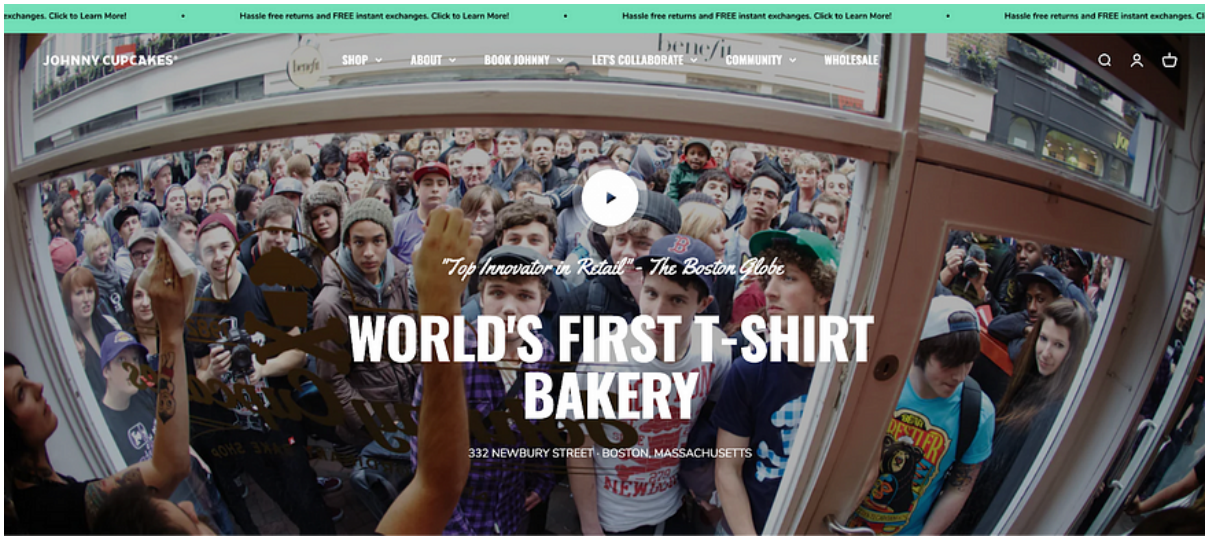
We also considered using OCR to extract more information from website images. However, this increased processing time by a factor of 10 and only marginally improved performance, so it was not included.

Our Predictions

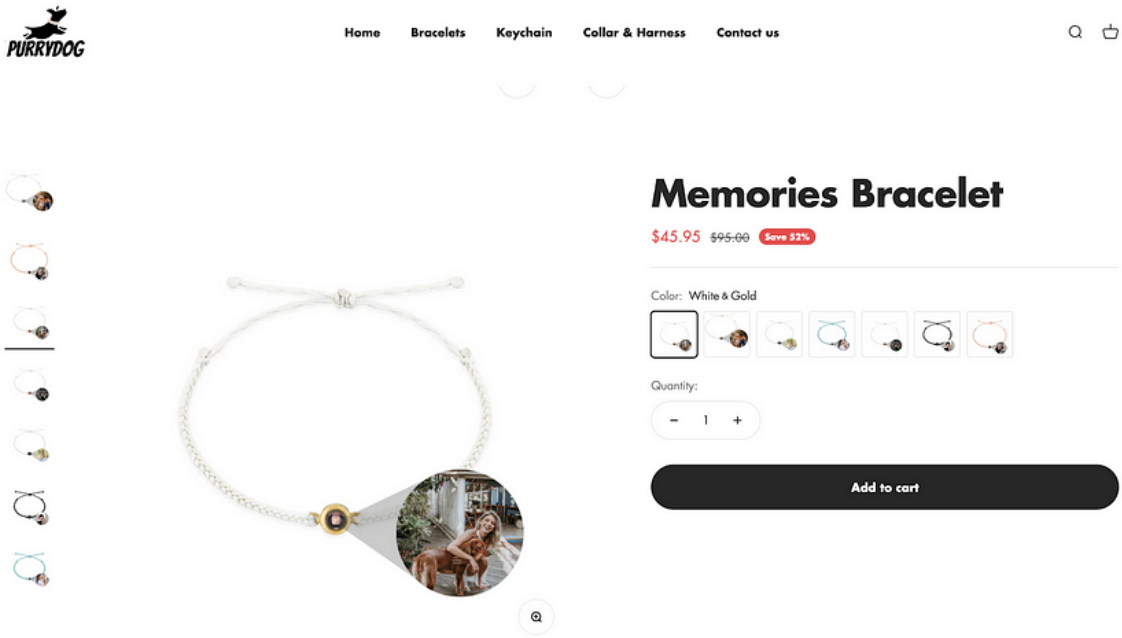
The current version of our model has 5,453 features: 5,000 for the TF-IDF vectorizations of the homepage, and 453 from the common top level domains, technologies detected, and third-party data. Overall, we were pleased with the performance of the classifier at over 70% accuracy, especially given, as mentioned above, the nuances of classifying an account into an industry. We store the top three predictions for a site to give us a fuller picture of its industry.

As I worked on the model, I looked at lots of sites (some Klaviyo accounts, some not) that it classified correctly and incorrectly. Here are a few examples.

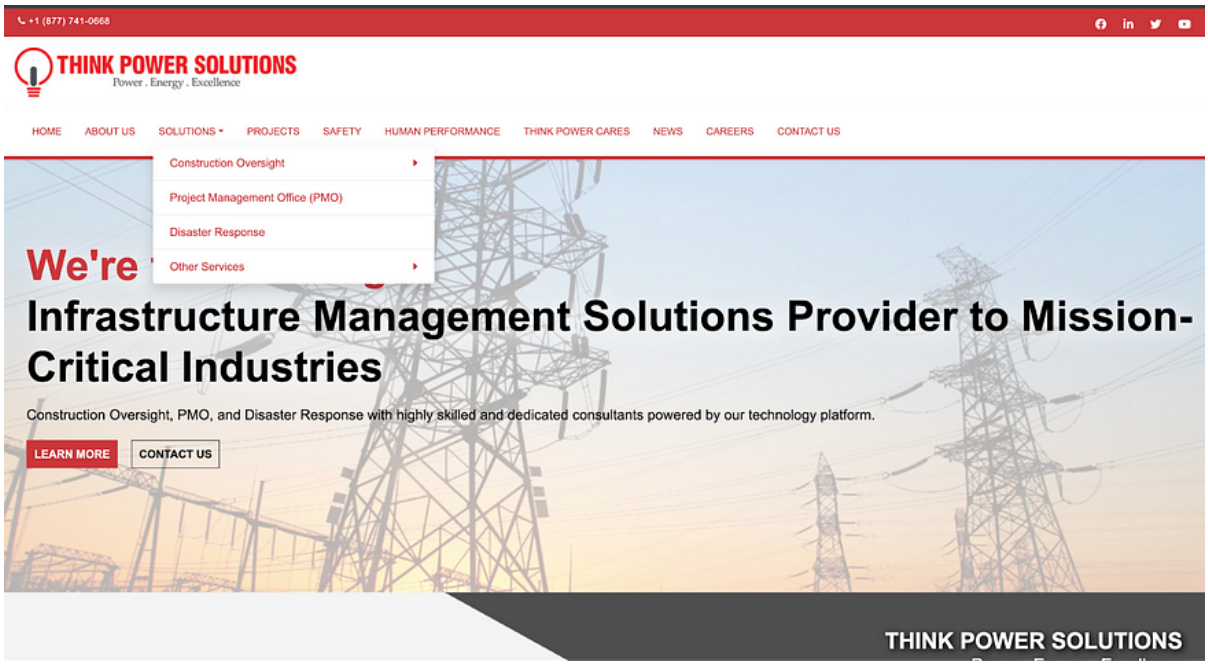
If youâ€™ve walked down Newbury Street in Boston, you may have passed by Johnny Cupcakes, and stopped in hoping for a dessert â€”only to find that itâ€™s a t-shirt store! Testing this tricky case on our classifier, we actually get it right. Weâ€™re quite confident (74%) that itâ€™s an *Apparel/Accessories* store, despite the fact that words like â€œcupcake,â€ â€œbaked goods,â€ and â€œbakeryâ€ are all over the page.



Another tricky classification is PurryDog. This business sells items for pets, but also many bracelets and keychains for owners. Their bracelets are highlighted on their homepage, and the first link in their top navigation. We predict that they are a *Jewelry* store, but since most of their products are for pets, *Specialty (pet)* store is probably the more appropriate classification.

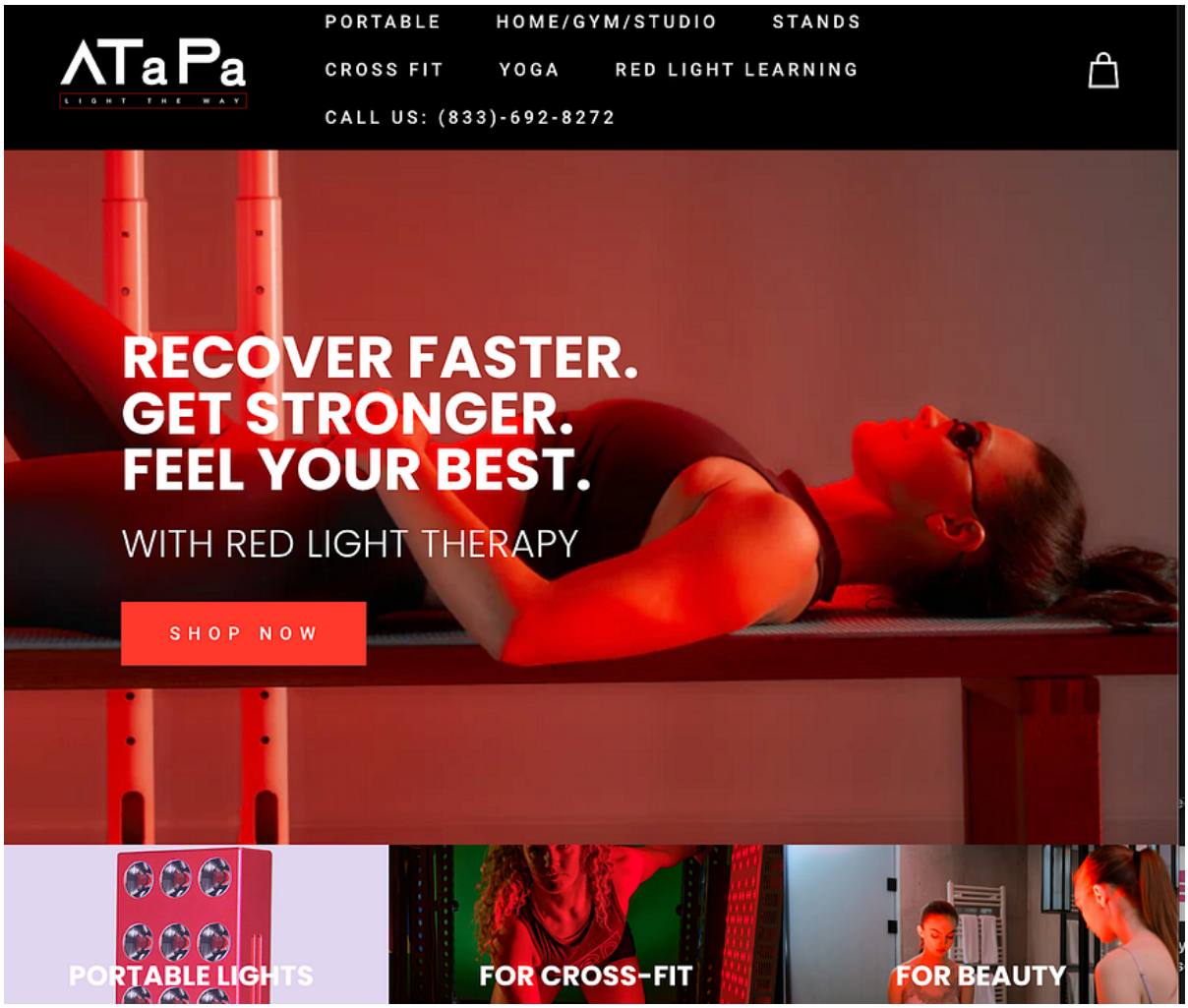


Below is a screenshot of the Think Power Solutions homepage. They provide construction oversight, PMO, and disaster response via consultants and a technology platform. This site has elements of a construction, consulting, and SaaS. Our classifier predicts *Real Estate/Construction* with 50% confidence, *Software/SaaS* with 17% confidence, and *Agency/Marketing/Consulting* with 11% confidence. Our top classification is reasonable, but obviously this business does not fall into any one category, and is a good example of why storing our top three predictions provides a fuller picture.



We â€œincorrectlyâ€ label World Rugby Shop above as a *Sporting Goods* store (itâ€™s labeled as *Apparel/Accessories* in our source of truth), however, this incorrect prediction seems completely reasonable. We do include *Apparel/Accessories* in the top three most likely categories.

Below is another example of an â€œincorrectâ€ prediction. This site was labeled Specialty in our source of truth, but we classify it as *Health/Beauty*. To my eyes, *Health/Beauty* and *Specialty* are equally correct.



Going Live

Before this model was implemented, we knew only a tiny fraction of our accountsâ€™ industries, and this was not known until significantly after an account signed up. Now, whenever a new account is created, we have an immediate prediction!

Future Ideas

We live in a big, rich, creative world and after this project, itâ€™s clear to me that it would be impossible to come up with a set of boxes such that every business and organization would cleanly fit into one. Still, down the road, I think it will be worth investigating additions and tweaks to our industry taxonomy. It also seems that while different taxonomies could be best for different purposes (onboarding best practices, benchmarking, internal business intelligence, internal sales/marketing), I doubt the complexity of that outweighs having a uniform Klaviyo standard.

Revisiting this work in a post-ChatGPT world would also be interesting, and Iâ€™d like to see how an LLM-based approach compares to our current model. Related, Iâ€™d like to revisit our labeled data and account for examples that are ambiguous (e.g. in the simplest case by excluding them), which will give us a more reliable accuracy metric, which would be nice to have before comparing an LLM approach to our current model.

Predicting industry is just the start for onboarding as we better understand our customers. We want to keep unlocking opportunities to provide more personalized user journeys and reveal relevant, tailored guidance to new signups.