Victor Vulovic

Radius

Coding Challenge 1 Results

October 5, 2017

Procedural Explanation and Data Findings

This initial coding challenge was presented to me by Samantha Rigan as a technical interview for the Data Scientist role. The purpose of this report is to elaborate on my thought process and to show the findings of the four calculation challenges:

1. **Fill Rate** - capture the initial rate of cells for which there is actually data
2. **True-Valued Fill Rate** - for the cells that are not empty, at what rate are those cells actually filled with relevant/useful data?
3. **Cardinality** - determine the total unique values for all the fields captured
4. **Something Interesting** - draw an interesting insight of my choosing from the data and discuss

Before diving into those results, I'll describe a bit more about the data set I worked with and set up the context for my thought process. Then I'll go ahead and elaborate on the points above and wrap up with some final thoughts.

The data provided to me was a list of 1M businesses delivered by external providers and gave some basic descriptive fields about those businesses: name, location, revenue, etc. My initial goal was to get a sense of the quality of the data that I would be working with— are there mistakes in the entries? How large is the data set? What are some typical occurrences to look out for?

The first task of finding Fill Rate was relatively straight forward, and I wanted to include both raw counts along with proportions to give a better feel for the meaning of those ratios as they can often be misleading when given alone. For example, having 1/2 and 50/100 both equate to a ratio of 0.5, so it's helpful to see raw performance juxtaposed with generalized performance. This raw count also proved to be helpful for comparison with the other calculations made later.

True-Valued Fill Rate was a much more involved task and certainly the most challenging of the four tasks. I decided to use regular expressions to filter out the irrelevant entries in the fields. In other words, I took a more manual approach to examine what kind of patterns in each field could be expected to be "normal" or "relevant". For example, the "zipcode" column has an obvious pattern: relevant entries will strictly be 5 digits long. Addresses, on the other hand, have much more variance in terms of legitimacy. For example, an address could be as simple as "1 Sample Rd", as complex as "8707 New Yorkshire Blvd Ste. # 305-A", and everything in between. Being able to correctly identify relevant addresses was a much more complicated task and required more extensive experimentation for the best pattern.

We obviously want to maximize the amount of useful/relevant data that we gather, so we are incentivized to capture as much as possible. However, getting too much noise will lead to inaccuracy later down the road when we build a predictive model. Therefore, I assumed that entries of "0" for some numerical fields, such as revenue, were irrelevant. This is because the dominating pattern led me to believe that any business with zero revenue would not be one Radius is interested in (irrelevant), and there

was a lower bound category that captured any company earning less than a minimum threshold.

As expected, the True Value Fill Rate is lower than regular Fill Rate. It's no stretch to expect that mistakes will be made in data entry or fields will just be left blank. I was interested to find that the field with the biggest negative impact was the "phone" field. This makes intuitive sense since many businesses either don't want flood themselves with phone calls and/or rely more solely upon internet-based communication channels.

Cardinality was also a straight-forward task to achieve. The purpose of this statistic is to understand the diversity of the dataset; how many unique values exist in each field? The cardinality I calculated was on the original dataset, before the regular expressions were applied for true value counts. Directions for this were unclear, but if cardinality is necessary for the relevant values, I'd be happy to provide the result.

Lastly, for an interesting fact, I wanted to see if there was a hotbed for billionaires in any of the states in particular. Where do all the billion dollar businesses reside? As one might expect, California is in the lead. Many companies in the billion dollar range are tech companies, which mainly are born in Silicon Valley. Add to the mix the entertainment industry in Los Angeles, and it's no surprise that California is in the lead. Also, the financial services industry has had a strangle-hold on the most profitable companies list for several decades, so New York and Illinois (Chicago) are higher on the list. What is surprising, however, is the huge proportion of billion dollar businesses in Texas and Florida. My initial intuition is that Texas can still rely on oil, but Florida is

still surprising— perhaps many real estate, healthcare, or insurance companies are based there?

Moving forward, it would be interesting to see how these values have changed over time. It's possible that we could create a time-series forecasting model to predict which new companies are destined for greatness based on location, current revenue, current headcount, and time-in-business.