# Applied Data Science Capstone Project – Churn

## 1. Introduction/Business Problem

Machine learning techniques like K Nearest Neighbors and Support Vector Machines are often used to predict the likelihood of customer loss (or churn) for various businesses such as credit card issuers or cell phone carriers. By identifying customers most likely to switch brands, businesses can potentially develop strategies for retaining those customers. Since the cost of acquiring new customers is typically greater than the cost of retaining existing customers, predictive analytics can be a cost-effective tool.

These machine learning techniques potentially can be used to analyze various aspects of the public sector (i.e., government) economy as well. Government could conceivably provide services more cost-effectively or to a larger group of constituents if it had a clearer picture of where and when to intervene.

Many municipalities in the U.S. require businesses operating within their city limits to register for a business license, which is a combination of taxes, fees, and regulatory requirements. The specific regulations and costs vary by city.

New businesses start up all the time. While some fail, a few relocate to neighboring cities to find cheaper rent or a more lucrative customer base, and others survive for years. Those businesses that fail or move away often leave behind unpaid debts such as taxes or fines to the city. Using predictive analytics, cities could potentially intervene earlier, identifying debtor businesses that are “flight risks” or businesses that could potentially be persuaded to stay in town and support the local economy. Churn analysis is therefore potentially of interest to both finance and economic development officials at any city hall.

## 2. Data

For this project, I will use a subset of business license data from the City of San Rafael, a municipality in Marin County, California. More specifically, I will look at restaurants operating in San Rafael in 2019 to predict whether they would disappear in 2020 (regardless of whether they closed their doors, moved out of town, or sold to a new owner).

Included in the dataset will be business name, address, district name associated with the address, form of ownership (sole proprietor, corporation, etc.), and number of years in business. Some of the restaurants have been “red tagged;” that is, they were temporarily shut down by county officials for health code violations.

Only non-confidential aspects of the business license data will be used. Regrettably, this may greatly weaken the accuracy of the churn predictions, as financial results are presumably a key factor in a business’s survival.

Instead of financials, I will use the techniques from this capstone course to retrieve Foursquare ratings for each business. After joining the Foursquare information to the dataset, I will test the hypothesis that low customer ratings are a sign of impending business failure.