Here are some ideas that I had for my Capstone project.

**Business problem 1:** Score companies based on their propensity for receiving complaints.

**Dataset to be used:** I could start with

<https://catalog.data.gov/dataset/consumer-complaint-database#topic=consumer_navigation>

Hopefully it's possible to merge on data from other sources.

This data set has many different categories of complaints (these categories are called "issues" in the data).  Some of these issues are "communication tactics", "APR or interest rate", and "billing disputes". In addition, each complaint is connected to a type of product offered (e.g., vehicle loans or mortgages).

One idea that I had for the target variable would be a weighted sum of complaints (at the company level) where the weights would be a function of the complaint category or the product. However, there are 95 distinct issues, so I'd probably have to think of some way to group the issues. There are only 12 distinct products, so I probably won't have to group those. In addition, the weighted sum of complaints would have to be scaled by the size of the company. Some ways to measure a company's size are market capitalization, sales, total assets, and enterprise value.

One nice thing about this data set is its size (4100 companies, 536,345 records total). The set is at complaint level. In addition, the company name will be a good merge key.

**Business problem 2:** Predicting which banks will fail within 5 years

**Dataset to be used:** I would start with

<https://catalog.data.gov/dataset/fdic-failed-bank-list>

This has 547 banks that have failed since 2000. To get a complete list of targets, I would have to append a list of banks that didn't fail. This data set has few variables so most of the predictors would have to be merged on by bank name from other sources.

**Business problem 3:** Predicting lawsuit results

**Dataset to be used:** Unknown

This is an important problem in the Legal Financing business. The way that legal financing works is this: The company makes loans (with interest) to carefully selected customers with pending lawsuits or lawsuits in process. The nice thing about this loan (for the plaintiff) is that if the plaintiff loses, then he doesn’t have to repay the loan. There is a big demand for this type of service, because lawsuits take a long time, and often the litigants can’t wait long for their money.

We would use parameters of the lawsuit for predictor variables. Probably one of the most powerful predictors would be the track record of the plaintiff’s attorney. Other predictors may be the zip code of the jurisdiction (since jury sympathy levels vary by geography), plaintiff age and gender, the industry of the defendant (auto, construction, medical, etc.), type of injury or loss suffered (is a bodily injury victim more likely to win than somebody who’s OK but whose car was wrecked?)

This problem may require several models, since there is such a huge variety of lawsuit types, including auto accident, Workers’ Compensation, civil rights violations, construction negligence, general negligence, pedestrian injury, premises negligence (slip and fall), wrongful death, etc.