This is certainly not a unique idea, but load origination is a classic data science example. Broadly speaking, a loan business will take in various pieces of data about an individual and then put that data through a model which will output the likelihood of that individual defaulting on a loan. That information is then used to offer a specific interest rate to the individual. This is a business where the most accurate model wins. If the model consistently underestimates the default likelihood, the company may go bankrupt. If the model consistently overestimates, other companies with better models will be able to undercut the offered interest rates.

All seven steps of the DS pipeline will be utilized, though some steps are more important than others. Discovering data could be as simple as polling previous customers’ credit scores and combining it with self-reported information (current income for example). Wrangling and profiling the data could present some issues. Data fields may be added or removed over time, and the model will need to account for this in some way. Another issue is that the formula for credit scores is regularly updated. Meaning that a 650 score in 2024 is not the same as a 650 score from 2005. For the model itself, there are several approaches that a company could take, though the simplest versions will likely use some form of regression analysis. Visualizing and reporting should be fairly straightforward. Identifying how accurate is the model and are new iterations performing better than older ones will be critical to improving future versions of the model.

I’ve mostly discussed a regression model that would attempt to predict default rates, but another option which could use the same dataset would be a clustering model. This model could identify categories of customers which are likely to be interested in specific services, allowing the company to direct marketing resources towards those customers.