What is the problem you want to solve

Can we predict when a company will begin exporting their product to other countries based on employee productivity, revenue, space, or other measures? Is there a way to predict which companies are ready to make that transition?

Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

My client is the Economic Development Council of Wisconsin. They want to know which Wisconsin companies may begin exporting in the near future so that they could best leverage grants, marketing, and investment in companies needing help making the transition.

What important fields and information does the data set have?

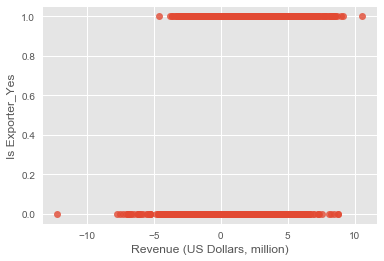
From EDA, the important features appear to be whether a company is an importer and/or manufacturer, if it owns/rents its facility, and measures of productivity: namely revenue per employee and revenue compared to facility size.

What are its limitations i.e. what are some questions that you cannot answer with this data set?

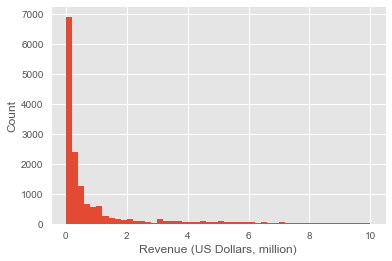
Cause and effect cannot be answered as there is no time component to the data. It is merely a snapshot. Thus, the transition cannot be analyzed.

What kind of cleaning and wrangling did you need to do?

There were 200 features in this data set and over 20,000 rows. The first step was selecting which features were important. Some had so much missing data that they were excluded (anything with more than 80% NaN was excluded outright). Eventually 10 were chosen for close analysis. By looking at visualizations it was determined which features had good correlation to exporting. For example, there was a dramatic shift in terms of revenue for exporters.



However, looking at the distribution of several categories showed dramatic skew.



To combat this log transforms were taken on revenue, facility size, and number of employees.

To examine categorical features, frequency tables were created and analyzed for dramatic frequency discrepancies. For example, around 20% of companies were women-owned, and about 20% of exporters were women-owned, so this did not seem to be an illuminating feature. However, owning facility, being an importer, and being a manufacturer did seem to have correlations.

Finally, measures of productivity were more enlightening than just simple stats like revenue or number of employees, so composite variables were created: revenue per employee and revenue per sq ft. In the end, 5 features were selected: Rev/Emp, Rev/SqFt, Importer, Manufacturer, and Owns/Rents.

Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

With this in mind, a logistic regression will be performed. Random forest (including boosted) will also be performed to see if any improvement can be found.