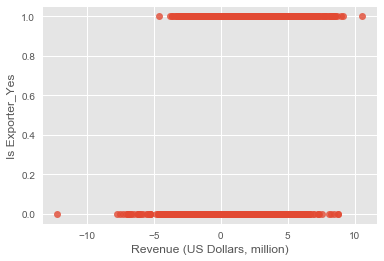
**Wisconsin Export Report**

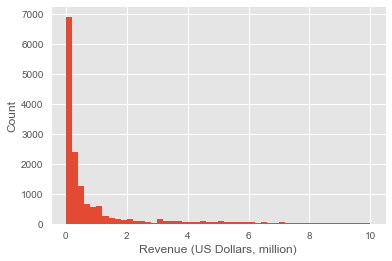
**Approach**

The Economic Development Council of Wisconsin wants 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. They collected a large dataset on over 20,000 companies to see if a model could be created to predict which companies were exporters. With this in mind, companies that fit the criteria but are not exporting can be targeted.

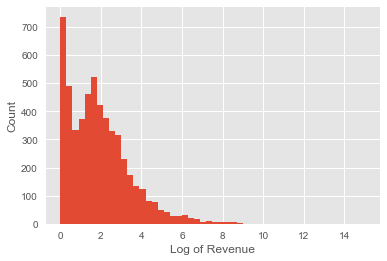
There were 200 features in this data set and over 20,000 rows. The first step was selecting which features were important. Some had so much missing data that they were excluded (anything with more than 80% NaN was excluded outright). Eventually 10 were chosen for close analysis. By looking at visualizations 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. An example of the new distribution:



There is still a spike on the very small edge(many companies reported near 0 revenue), but the overall shape is more normalized.

To examine categorical features, frequency tables were created and analyzed for dramatic frequency discrepancies. For example, around 10% of companies were women-owned, and about 10% 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. For example, note that more than half of exporters are importers, although only about 7% of all companies are importers.

|  | **Not Importer** | **Is Importer** | **rowtotal** |
| --- | --- | --- | --- |
| **Not Exporter** | 17801 | 1047 | 18848 |
| **Is Exporter** | 566 | 646 | 1212 |
| **coltotal** | 18367 | 1693 | 20060 |

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.

**Findings**

Initially a logistic regression was performed using the non-transformed continuous variables. While there was fairly high accuracy (around 90%), the confusion matrix showed that the model was simply predicting everything to be a non-exporter. Since about 90% of the companies aren’t exporters, 90% accuracy. However, when the log-transformed data was used the accuracy increased to around 93% and there were actually predictions being made. Many of the exporters were still labeled as non-exporters (false negatives), but there were also now a good chunk of false positives – which are actually the most interesting case.

Random forest and XGBoosted RF both showed similar results to the logistic regression on the log-transformed data. Similar accuracy and confusion matrices were found. So it appears that with the current feature selection this is near the highest accuracy that can be achieved.

**Next Steps and Recommendations**

Further exploration with features would be warranted. Measures of productivity were used (revenue per employee), but just including revenue could also be helpful. Further, gross income was rejected because of significant missing data, but perhaps it would be worth considering with imputing the missing data.

With the current results, the following is recommended:

1. Some companies were included multiple times as they were subsidiaries. It would be interesting to analyze if being connected to another company accounts for some of the false positives or negatives.
2. An analysis of false positives should be conducted. Especially in light of the model’s bias toward predicting negatives, any false positives probably share many characteristics with exporters. Thus, these companies are probably ready to export in terms of their productivity. This is where resources should be focused for intervention.
3. Further data should be collected year over year to allow for time analysis and see both what features predict the transition and to see what effects beginning to export has on a company.