A Bad Time or a Bad Business? Identification of Profitable Firms with Short Run Problems

Diego Escobar
Juan Ignacio Vila

April 30^{th}, 2019

1 Data

We started by downloading the data files that we considered as high priority for the predictive analysis. This includes downloading and further processing the business licenses data. We have constructed several features based on this and adapted the variables to the format required to be used in our pipeline. We also downloaded data from ACS and geographically matched the business license data with the census blocks from the ACS. This match will also allow us to use other sources of data that vary geographically. The upcoming steps now include extending this by adding sector level data and interactions of data from different sources (e.g. number of past closures in the area of businesses from the same sector).

2 Analysis

We implemented a basic model within the framework of our pipeline. The prediction of the current model are definitely not what we expect them to be. This is however because we used a very limited set of variables, while focusing mainly in having a working version of the code. Although we are currently able to produce predictions based on these small amount features, these are absolutely not reliable and we constructed them exclusively to test our pipeline.

3 Changes in the Project

The main guidelines for the analysis remain unchanged. One extension that we intend to include is using k-mean clustering to generate geographical clusters of the businesses. The reason to do this is that neighborhoods or wards are not necessarily the correct unit of analysis, since several business areas cross neighborhoods and wards.

4 Next Steps

During this week we plan to focus on two main tasks: 1) adding more data sources, and 2) creating more features for the model. Now that our pipeline is working, we expect this to be relatively

fast. A final step will then be to add more features to the model. Once this is done it will be straightforward to experiment with different models and parameters, since our pipeline from the homework is designed to do that.