Data Science Capstone & Ethics (ENGI E4800) *October 22, 2018*
Alexandra Sudomoeva (as5402)
DongGu Kim (dk2983)
Haotian Zeng (hz2494)
Chengzhang Xu (cx2188)
Yang Gao (yg2499)

Capstone Progress Report

Introduction

When talking about business in the US, one typically defaults to big companies such as Apple, Capital One, Walmart etc. Nonetheless, nearly half (48 percent) of the job market in the US is actually driven by small and medium-sized enterprises (SMEs) (Wheat and Farrell). Small businesses represent a substantial and important part of the US economy accounting for a significant share of personal income while driving approximately 52 percent of net job growth (Brock and Evans; Wheat and Farrell). Over the past decade, small businesses have been through periods of rapid competitive change and fluctuation. These fluctuations can be considered a serious factor in defining the national financial health and represent an exciting and relevant research subject.

While SMEs play a crucial role in the US economy, a number of studies show that, on average, only approximately 50 percent of all new small businesses survive after the first 4 years (Hanas and Leatherman). This leads to an inevitable question: what are the possible contributing factors to such a significant and consistent drop-off rate as well as overall SMEs' performance.

Problem Statement

For the purpose of this study, we wanted to focus on discovering and understanding the drivers behind small and medium business formation, growth, and dissolution. In particular, we would like to test a number of factors suggested by previous research and derive a quantitative measure for each of the values found significant. Based on the measures derived, we wish to forecast and measure future changes in the SME distribution based on the quantitative dependency derived in the previous model.

Literature Review

There has been a number of detailed articles and economic research that has helped up form the direction and initial hypothesis for the project. While there are definitely a variety of contributing factors, most research tends to overlap on three main areas as potential drivers of SMEs' performance: economic, financial and regional (geographic).

Some studies suggest that economic fluctuations impact small businesses first and most severely. In their article on small business economics, Broke and Evans discuss the profound implications of economic growth on the number of employment in firms with less than 500 employees. With a number of economic indicators rising, the SMEs' employment also grew 2 percent from 51 to 53 percent in 1984 (Brock and Evans). Other studies, like the one conducted by Kirchhoff, support the notion than the SMEs' performance fluctuations are not random but are rather explained by the economic climate (Kirchhoff). This is supported by the fact that small and medium business "survival rates" do not show much deviation from the 50% benchmark across time except for economy caused fluctuations (Kirchhoff). Lastly, Brown and Batista also derive the mutually dependent relationship between small business survival rate and economic indicators like GDP and unemployment. The researchers particularly outline industry market

competition as one of the two main causes for low surviving rates (De Sousa-Brown and Batista). Looking into particular economic indicators, unemployment was found to have "a negative impact on the firms' chances of survival" as recess in the economy is connected to insufficient demand (De Sousa-Brown and Batista). The market environment also plays a huge role affecting success and failure of small business.

The SME financial sources, barriers, and overall health is another area of consideration when it comes to predicting SMEs' performance. Brown and Batista stress this observation in their study by stating that "firms that intensively use external sources of finance exhibit growth rates much higher than what could be expected with internal finance alone" (De Sousa-Brown and Batista). Small businesses are also often affected by the established barriers in the financial (banking) market. In particular, the credit sources for smaller firms are found to "dry up more rapidly" than those for the larger enterprises (Brock and Evans). Smaller companies are also more likely to be subject to scrutiny when it comes to obtaining capital at market interest rates (Bartik).

In addition to financial and economic drivers, SMEs' business cycles were also found to be associated with regional conditions. This phenomenon can be explained by the fact that many of the small businesses do not have a lot of branches and usually are local. Therefore, such things as state population density, local market demand, and state taxes were often included and found significant in various analyses (Bartik; Hanas and Leatherman). For example, Nicholls and Foreman-Peck found regional productivity divergence to be on great influence on small and medium businesses (Foreman-Peck and Nicholls). In particular, regions with higher productivity seem to encourage positive SME performance. When looking at the effects of the economic crisis, Dachin and Rusei also noticed rather distinct regional differences in the small firm survival rates (Dachin and Rusei). In developing their model, the authors assumed SMEs' "survival capacity is strongly connected to the economic performance of each region/state" supporting the argument of geography being an important driver of performance (Dachin and Rusei).

Regional population density, and the proportion of small size establishments are also found to be important for the performance and initial formation of small companies (De Sousa-Brown and Batista). Other state characteristics like the number of high school graduates is also assumed to have a "highly significant and large positive effect" on small business expansion (a point increase is associated with an average of 3.5 percent SME growth) (Bartik). Looking more granular than the state level, other research also shows how county employment changes and economic conditions become statistically significant indicators of small and medium firm survival (Hanas and Leatherman).

The three driver areas outlined above are not exhaustive at all. There are also a number of other factors that should be considered when trying to predict SME performance. For example, big company takeover, studied by Nicholls and Foreman-Peck, is one of these drivers. With the expansion of established corporations, comes the increased change of SME closing (Foreman-Peck and Nicholls). On the other hand, the presence and strength of unions is another important consideration. Some research suggests that it actually has a negative effect on small business growth (Bartik). Lastly, industry indicators also play an important role in determining the most likely path for SMEs and influence its probability of survival (Forsyth). Driven by unique business environments, various industries allow for different levels of volatility and, therefore, create differences in the SME performance (Hanas and Leatherman).

It is important to mention that, while still successful, all of the studies researched faced a significant number of limitations around data and modeling. For example, some papers produced either contradicting or surprisingly little correlation between firm survival and outlined characteristics. This inconsistency and potential research bias was addressed and explored further by Forsyth. In particular, the researcher has

pointed out the vast multicollinearity between the variables tested in previous research. Nonetheless, despite the possible biases appearing in including the external factors, we believe that with proper modeling and analysis, the risk for false and misleading conclusions can be minimized.

Growth of small and medium business is dependent upon a number of internal and external factors. Internal factors are related to firm itself and external factors are related to market environment surrounding business, and the overall financial conditions. These factors can be characterized as financial opportunities (available interest rates, loan barriers, liquidity), economic indicators (unemployment, GDP, agglomeration), regional characteristics (population density, urban vs rural distribution, local taxation and policies), and other important features (union strength, big company presence, industry breakdown). The importance of these factors is shown by a combination of extensive research. Hence, including them in our analysis is essential for understanding and forecasting the small and medium businesses performance.

Data Collection & Description

The data collection process was strongly guided by the things learned from related research (as outlined in the literature review section). Therefore, we focused on targeting all available datasets that included economic, financial, and regional indicators.

Following the importance of regional impact (geography), we decided to collect all of our data not only on national but also state and county levels. The county descriptive data was predominantly pulled from the Census Bureau's release of annual *County Business Patterns*. Consequently, the state descriptive data was extracted from multiple sources including Bureau of Labor Statistics (union representation), National Science Foundation (education level) and the Census Bureau (area and other descriptives). All of the data had annual basis (1992-2016) and was directly downloaded from the government websites.

In addition to normal regional descriptive statistics around population density, education level, etc. both datasets also included a number of economic indicators around regional unemployment, GDP, personal income, and regional price deflators. These values were pulled on an annual basis for the same time period of 1992-2016 using the Bureaus of Labor Statistics and Economic Analysis (state) and the Census Bureau (county).

Lastly, we have collected a number of financial indicators to help measure overall national financial health and availability. Such reflective indexes as Dow Jones, NASDAQ, and Russell 1000 were pulled using the Cran package. To address the subject of big corporation takeover, we also extracted the stocks of Apple, Amazon, Google, JPMorgan, Facebook, Microsoft, Tesla, and Walmart. We believe that their stock values (growth) could successfully represent such companies' expansion into their HQ state/counties.

To define the target measure around SME performance, we explored the data from *Business Employment Dynamics* released by the Bureau of Labor Statistics. The data pull included annual data (1992-2016) with count of firms in each growth category (birth, death) by company size. This segregation allowed us to establish a reliable and consistent quantitative measure for annual SME growth.

The detailed view of all available features from our joined integrated datafile and corresponding data sources can be found in the appendix for your convenience.

Initial Data Exploration

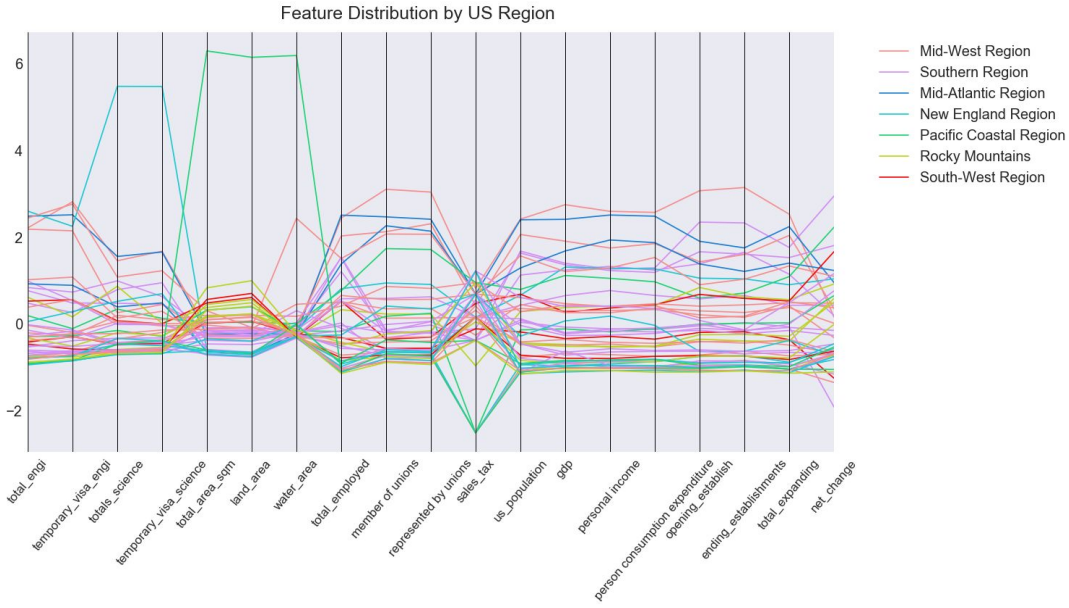
As briefly outlined in the previous segment of the proposal, we developed two separate datasets for county level and state level data. We purposefully chose to prioritize the arguments about the importance of geography in our data structure due to a prevailing number of literature supporting such a level of granularity as well as the rich variety of features available for collection and, hence, testing. On the other hand, other important considerations from the literature review were incorporated inside of the regional framework.

Given the collected datasets, we were able to test some of the initial hypothesis developed in consideration of related literature. The questions we attempted to answer via exploratory data analysis and visualization were the following:

1. What are some of the general characteristics of features chosen to assess SME performance?
2. How is SMEs' growth distributed across various states? What are some of the factors that might skew the distribution (if any)?
3. Which features show the highest possible correlation (normal and lagged) with the growth target for counties and states? And what is the direction of the relationship (positive vs negative)?
4. Is there any multicollinearity between tested SME performance drivers we should consider in further developing the regression model for states and counties?
5. Does big corporation growth has an effect on SME deaths in the corresponding region (corporation's headquarters)?

The results of our analysis and the corresponding answers to these questions are outlined below.

1. What are some of the general characteristics of chosen features?

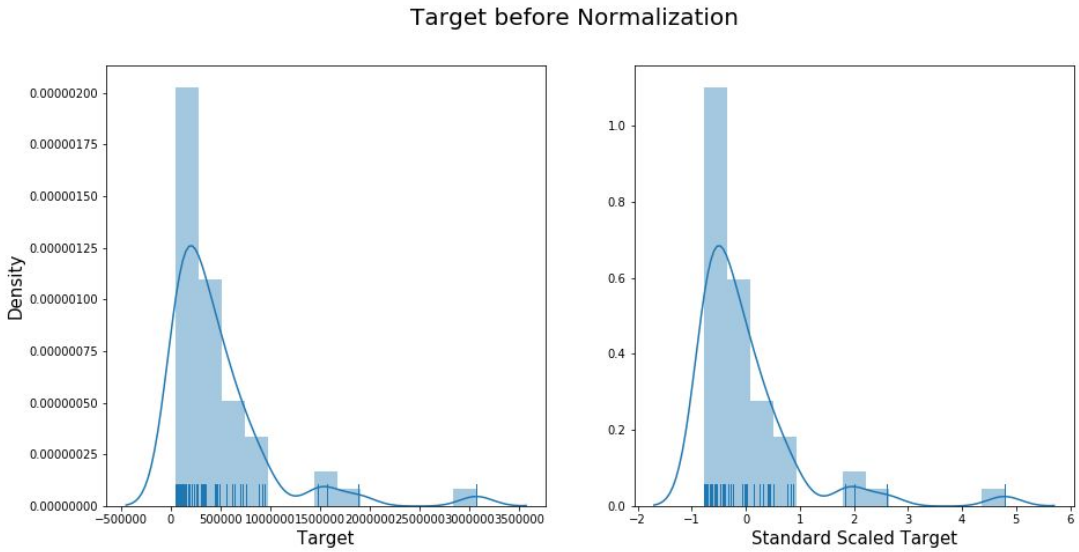


To understand the general characteristics of our variables and the adopted framework, we used parallel coordinates to show each feature distribution for different regions in the US. Looking at the graph,the regions vary in a number of features like GDP, personal income, and education level. The observed variety raises caution for potential biases when developing the model within different regions.

When thinking about the SME performance as a target variable, we were also interested in looking into its distribution within various states. Our analysis showed that some states have very high density of company death (Louisiana, North Dakota, Wyoming, Oklahoma, Arizona, New Mexico). Nonetheless, most states remained on the positive side of the growth scale.

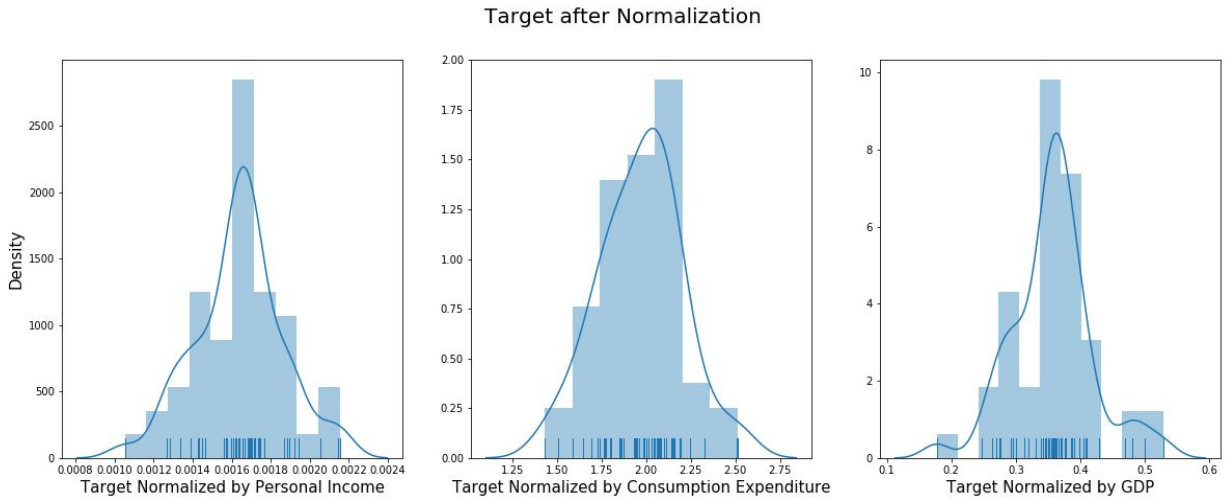
2. How is SMEs' growth (expansion) distributed across various states?

Distribution of raw/standard-scaled target-value appears to be extremely left skewed. It needs to be normalized for better quality of analysis and model training based on probabilistic assumption. We chose to test some of our data features as normalization factors to examine how well each factor performs. We expected to interpret the highest performing factor as the leading driver in the target distribution skew.



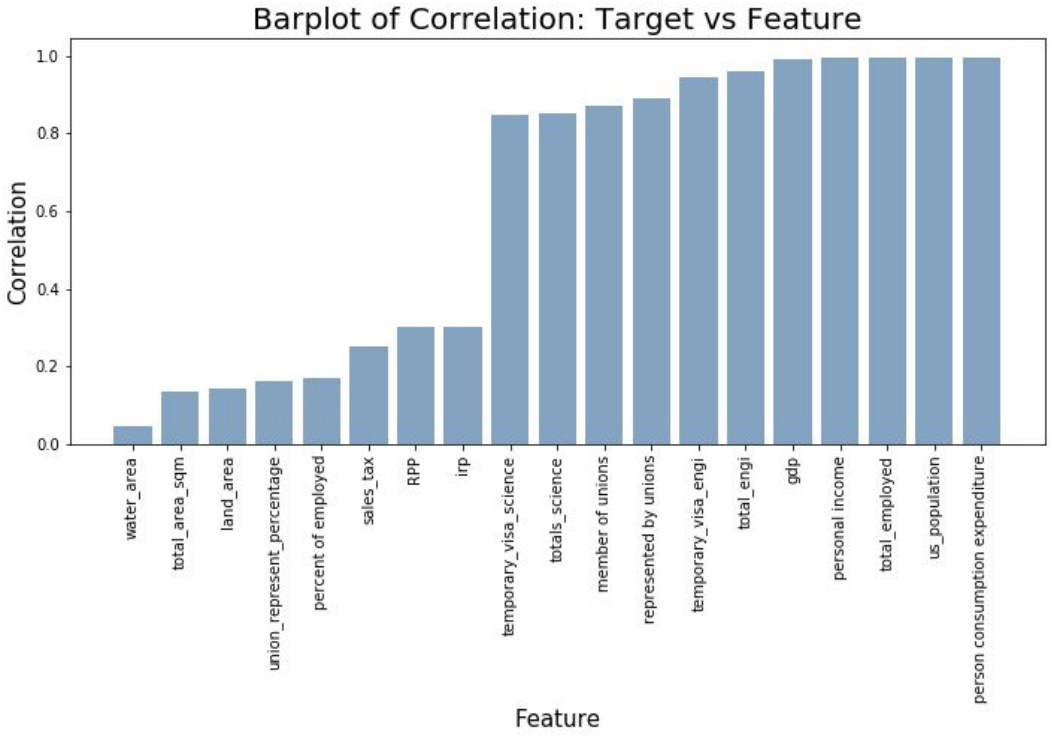
The high skewness is attributed to gap of economic size between different states. As an instance, economic size of New York is much larger than that of Alabama and it's not surprising that the number of SME birth/death of New York is larger than that of Alabama. Hence, we decided to normalize target by dividing it with variables which represent economic size of each state to address skewness. The target value is initially normalized in three different ways: (1) by personal income, (2) by consumption expenditure and (3) by GDP.

By looking at the chart below, the distribution of target normalized by personal income appears to fit better into shape of normal distribution compared to the other cases. Hence, this leads us to an initial hypothesis (to be further tested) that personal income has the strongest impact on SMEs growth for states.

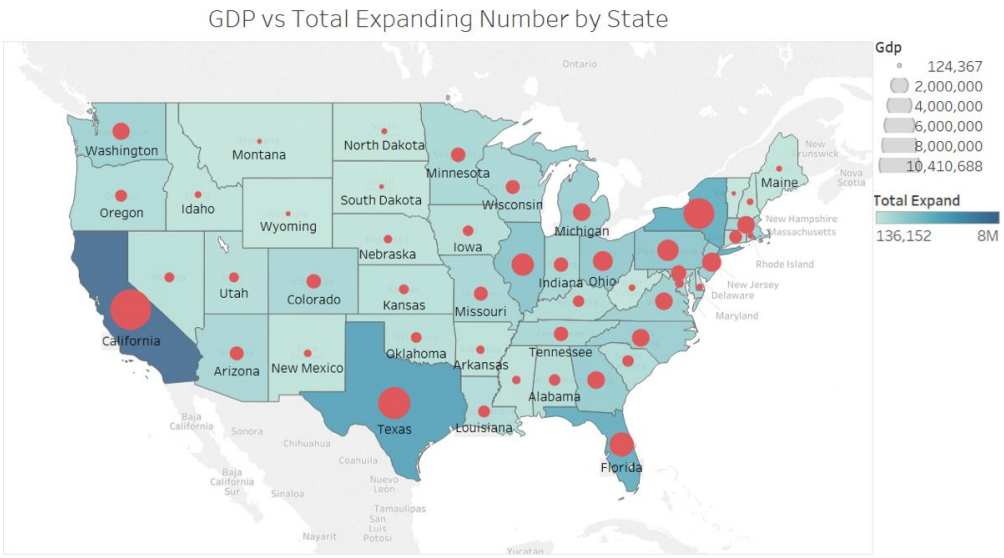


3. Which feature variables show the highest correlation with the growth target?

The plot below illustrates correlation between each feature and the target value before normalization. Looking at the plot, we can see how most features tested have a relatively high correlation with target. Interestingly enough, the correlation of percent of employed is low while that of total employed is high.

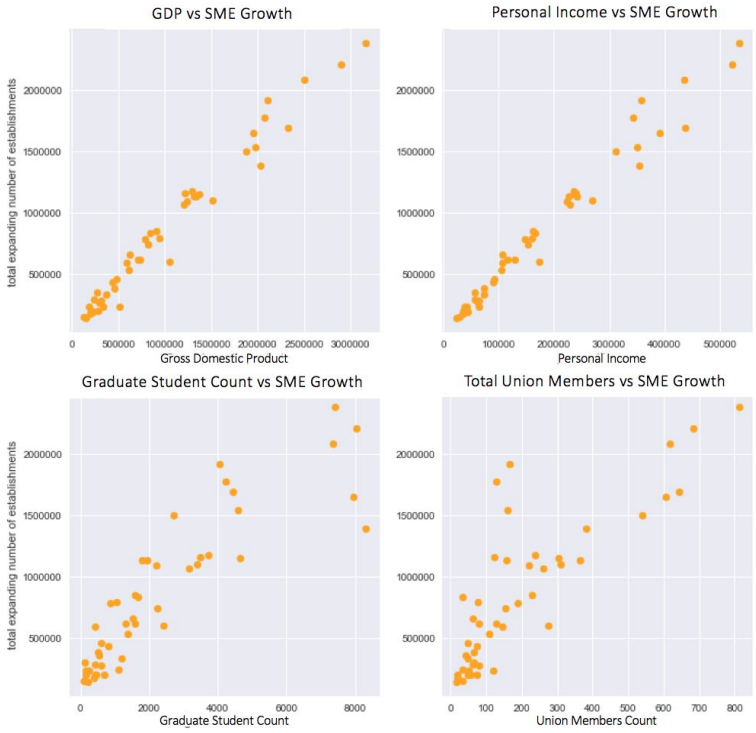


This phenomenon can possibly be explained by the fact that the distribution of target value before normalization is highly skewed by absolute economic size of each state rather than relative records. Addressing the regional framework, the map plot below reveals that states with higher GDP tend to have higher SME net growth.



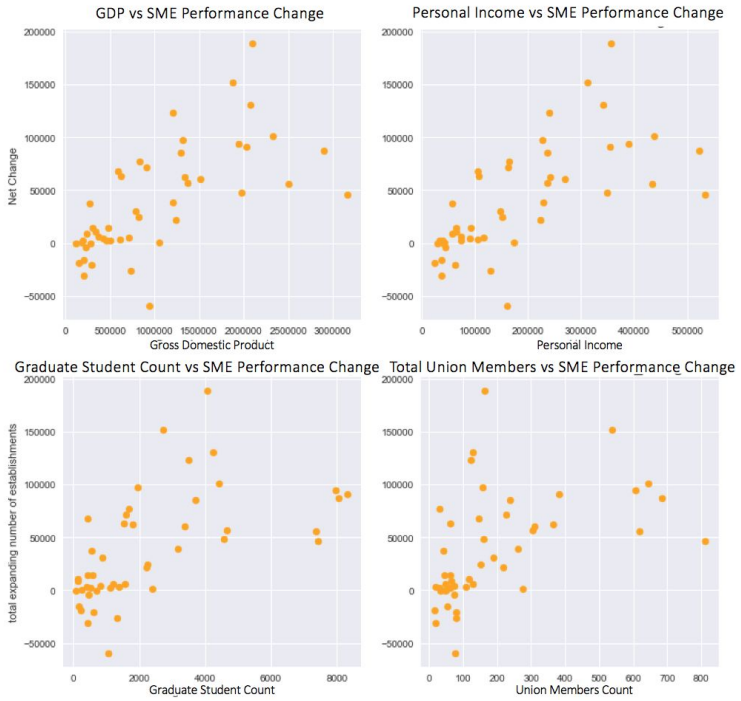
We have picked net change of SMEs performance, and SME growth (expansion) as two focus targets. Looking at the two scatter plot graphs, we can see how the two targets have a slightly different strength (clarity) of the relationship with chosen features. While SME growth is showing a very clear positive relationship close to a line correlation, that of net change is more 'jittered'.

Scatterplot Matrix for SME Growth and Various Features

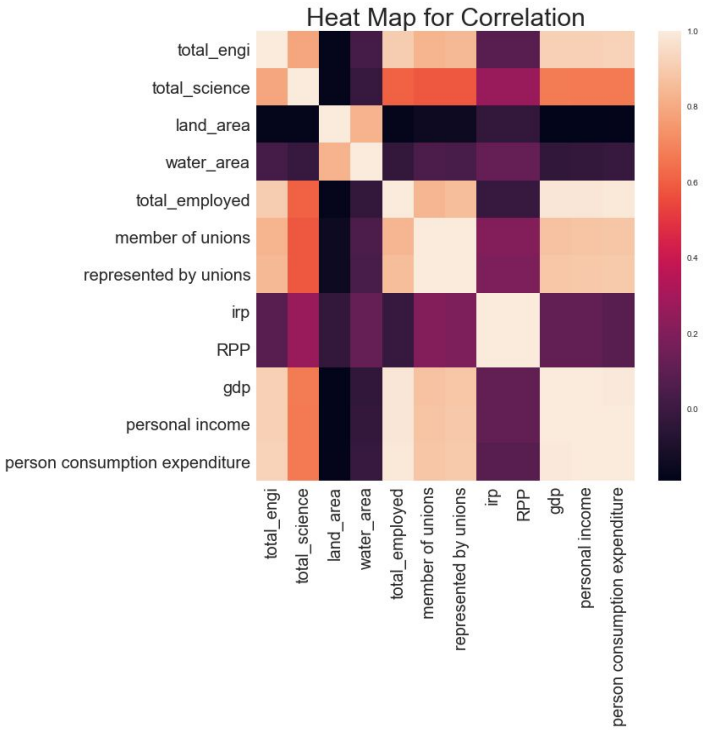


We have also tested other features like price parity, implicit regional price deflator, and real purchase parity. Nonetheless, they did not show significant relationship with either target. Therefore, we will consider testing if lagging those variables would present a different outcome going forward.

Scatterplot Matrix for SME Performance Net Change and Various Features



4. Is there any multicollinearity between tested SME performance drivers?

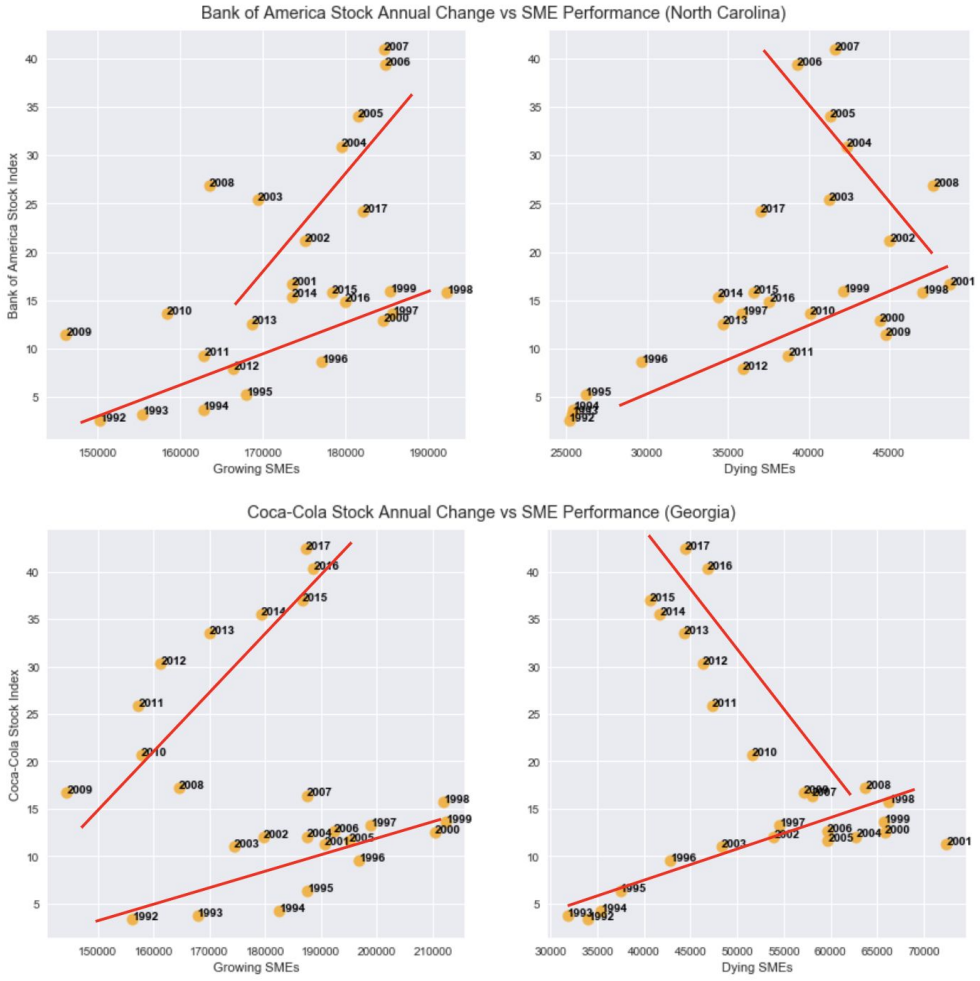


Interdependency between tested variables can create a strong bias in our model. Hence, we decided to test for any possible feature correlations before proceeding into the modeling stages. As seen from the heatmap below, most variables show relatively low levels of correlation. Nonetheless, there are some

variables for which we should be advised not to include in the same model. For example, moving forward it might make sense to drop “total_science” (total number of graduate students) all together due to its high correlation with employment, unions, and economic variables.

5. Does big corporation growth has an effect on SME deaths?

To measure big corporation takeover, we chose to use net change in individual stock prices as a way to determine likely expansion. We have picked Coca-Cola (Georgia) and Bank of America (North Carolina) to explore stock movements within corresponding states’ SME performance changes.



Looking at the resulting scatter plots, the points represent each individual year ranging from 1992 to 2017 in Georgia and North Carolina respectively. The trend between stock change and dying SMEs drastically changes in slope sign after the two major financial crises in 1998 and 2008. Hence, this change created separate clusters with different relationships on the targets. For example, when Bank of America stock index increased, growing SMEs increased as well in North Carolina (counter-intuitively, so did the Dying SMEs between 1992-1998). Nonetheless, the relationship went in the opposite negative direction (normalized) after the year of 2009.

Next Steps

Until this point, our team has worked on performing preliminary research about small business economy. We have gathered the data related to establishment, growth and decline of small businesses as well as potential candidate features. In addition, we have implemented exploratory data analysis to test and define our initial hypothesis on a sample dataset (a subset of the full timeframe).Going forward, we will set SME growth rate (expanding_establish) and SME relative death ratio (number_of_death) as targets of interest to make this analysis more concise and clear.

For the next step, we plan to merge all data found significant in the EDA stage and implement the data preprocessing that we applied to sample dataset. Having merged the master dataset, we plan on starting to train ML models to forecast the target variable, tune parameters and evaluate model performance. In the final stage of model evaluation, we will study critical drivers affecting success and failure of small businesses.

Upon completing the first modeling cycle, we will search for additional features and try different ways of engineering features to improve model performance. Given enough time, we also hope to process newspaper article text data to obtain sentimental score and examine how it is related to small business economy.

Appendix A - Detailed Data Table

Variable Name	Description	Data source
total_eng	Total number of graduate students of nfs in engineering	1
temporary_visa_engi	Total number of graduate students who holds temporary visa of nfs in engineering	1
total_science	Total number of graduate students of nfs in science	1
temporary_visa_science	Total number of graduate students who holds temporary visa of nfs in science	1
total_area	Total area of each state(square miles)	2
land_area	Total land area of each state(square miles)	2
water area	Total water area of each state(square miles)	2
total employed	Total employed people	3
member of unions	Data refer to members of a labor union or an employee association similar to a union.	3
percent of employed	Member union percentage	3
represented by unions	Data refer to both union members and workers who report no union affiliation but whose jobs are covered by a union or an employee association contract.	3
percentage of employed	The represents percentage	3
gdp	Gross domestic product	4
irp	Implicit regional price deflator	4
personal income	Real personal income (thousands of chained (2009) dollars)	4
RPP	Regional price parities	4

person consumption expenditure	Personal consumption expenditures	4
net_change	Change of number of establishments(net_change=total_expanding-total_ contract)	5
total_expand	Total number of opening and expanding establishments total_expanding=expanding_establish+opening_est ablish	5
expand_establish	Number of expanding establishment	5
open_establish	Number of opening establishments	5
total_contract	Total number of contracting and closing establishments total_contract=contract_establish+close_establish	5
contract_estalbish	Number of contracting establishments	5
close_establish	Number of closing establishments	5

Data sources in variable list

1. National Science Foundation
   https://ncsesdata.nsf.gov/gradpostdoc/2016/html/GSS2016_DST_24.html
   https://ncsesdata.nsf.gov/gradpostdoc/2016/html/GSS2016_DST_44.html
2. Census Bureau
   https://www.census.gov/geo/reference/state-area.html
3. Bureau of Labor Statistics:union affiliation
   https://www.bls.gov/news.release/union2.t05.htm
4. Bureau of Economic Analysis
   https://www.bea.gov/data/economic-accounts/regional
5. Bureau of Labor Statistics
   https://www.bls.gov/bdm/bdmstate.htm

The data sources 1 - 4 were directly downloaded from the government website. The data source 5 was scraped by python. The data such as GDP and personal income have sub-categories that were not presented in the table. Nonetheless, will likely be used in the future modeling.

Appendix B - Team Member Contribution

In putting the first progress report together, each team member has contributed to a variety of analysis:

- *Alexandra Sudomoeva* has worked on the introduction, literature review, data collection and description as well as the concept slides
- *DongGu Kim* has also contributed to the literature review and most of the visualization analysis for second and third questions of interest
- *Haotian Zeng* focused his forces on data cleaning and analysis also contributing to the concept slides
- *Chengzhang Xu* was the driving force behind most of the data collection and visualization that focused on answering third and fourth questions of interest
- *Yang Gao* was responsible for collecting the finance data as well as significantly contributing to the analysis of the first and fifth questions in the EDA segment

References

1. Bartik, Timothy J. "Small Business Start-Ups in the United States: Estimates of the Effects of Characteristics of States." *Southern Economic Journal*, vol. 55, no. 4, 1989, pp. 1004–1018. *JSTOR*, JSTOR, www.jstor.org/stable/1059479.
2. Brock, William A., and David S. Evans. "Small Business Economics." Small Business Economics, vol. 1, no. 1, 1989, pp. 7–20. JSTOR, JSTOR, www.jstor.org/stable/40228490.
3. Wheat, C. and Farrell, D. "The Ups and Downs of Small Business Employment: Big Data on Payroll Growth and Volatility." January 18, 2017. JPMorgan Chase & Co. Institute, January 2017. Available at SSRN: https://ssrn.com/abstract=2966135
4. Hanas C., A. and Leatherman, J. C. "Small Business Survival and Sample Selection Bias." *Small Business Economics*, vol. 37, no. 2, 2011, pp. 155–165. *JSTOR*, JSTOR, www.jstor.org/stable/41486124.
5. Douglas B., G. and Morrison, E., R. "Serial Entrepreneurs and Small Business Bankruptcies." *Columbia Law Review*, vol. 105, no. 8, 2005, pp. 2310–2368. *JSTOR*, JSTOR, www.jstor.org/stable/4099396.
6. Dachin, A. and Rusei, A. "Regional Determinants of Small Business Survival during the Crisis in Romania." Acta Universitatis Danubius: Oeconomica. 2013;9(4):200-208. https://doaj.org/article/058df0c2ac114bd58bd3dd960cd04b6d
7. Headd, Brian, and Bruce Kirchhoff. "The Growth, Decline and Survival of Small Businesses: An Exploratory Study of Life Cycles." *Journal of Small Business Management*, vol. 47, no. 4, 2009, pp. 531-550. *ProQuest*, http://ezproxy.cul.columbia.edu/login?url=https://search-proquest-com.ezproxy.cul.columbia.edu/docview/220997774?accountid=10226.
8. Forsyth, G., D. "A Note on Small Business Survival Rates in Rural Areas: The Case of Washington State." *Growth and Change*, vol. 36, no. 3, Summer 2005, pp. 428–440. *EBSCOhost*, doi:http://onlinelibrary.wiley.com.ezproxy.cul.columbia.edu/journal/10.1111/%28ISSN%291468-2257/issues.
9. De Sousa-Brown, S.,Costa Batista. (2008). *County-level analysis of small business and entrepreneurship in west virginia: Impact on rural economic growth* (Order No. 3326474). Available from ProQuest Dissertations & Theses Global. (304447753)
10. Foreman-Peck, J., and Nicholls, T. "SME Takeovers as a Contributor to Regional Productivity Gaps." *Small Business Economics*, vol. 41, no. 2, 2013, pp. 359–378. *JSTOR*, JSTOR, www.jstor.org/stable/43552872.