Patterns and Forecasted Behavior of Small and Medium Enterprises

Data Science Capstone Project with Capital One

Team: Alexandra Sudomoeva, DongGu Kim, Haotian Zeng, Chengzhang Xu, Yang Gao
Industry Advisors: Gerry Song, Isabel Xiao
Faculty Advisor: Vincent Dorie

Data Science Institute
COLUMBIA UNIVERSITY

Introduction: Why SMEs?

SME (small and medium enterprises with total employee count below 500 people) growth fluctuations represent an exciting and relevant research subject for a number of reasons:

1. SME performance can be used to define the national financial health
   - Drive 48% of the job market in the US
   - Account for approximately 52% of net job growth
2. SME deaths can have serious negative consequences on the economy
   - Only 50% of all new small businesses survive after the first 4 years
   - SMEs generate a significant amount of innovation

Focus: Discover and understand the drivers behind SME formation, growth, and dissolution.
Goal: Forecast and measure SME growth by state utilizing external economic, financial, and geopolitical factors suggested by previous research.

Data Description & Methodology

The data collection process was strongly guided by the things learned from related research. We have aggregated our pull on an annual basis spanning for 1992-2016.

Five variables were selected as targets for modeling:

Name	Description	Variable Type	Source
expand_establish	Number of expanding SMEs	continuous	Bureau of Labor Statistics
contract_establish	Number of contracting SMEs	continuous	
open_establish	Number of opening SMEs	continuous	
end_establish	Number of closing SMEs	continuous	
net_change	Net change in SME count	continuous	

The explanatory variables were categorized into two categories:

Category Name	Variable examples	Sources
Geopolitical	State taxes, medical assistance, reginal avg. rent	National Science Foundation; Bureau of Labor Statistics
Economic	GDP, personal income/consumption, inventory	Bureau of Economic Analysis; United Census Bureau

Modeling

Guided by the project goal outlined above, we chose to look at a variety of modeling techniques:

1. Regression – used to explore the causal relationship between targets and exogenous variables

Decision Tree for Contracting SMEs as target:

Feature	Importance
Gross domestic product (GDP) by state (lag2)	0.77
Computer Manufacturing	0.12
Worker's Compensation (lag2)	0.05
Cattle Production (lag2)	0.01
PCE: Gasoline (lag2)	0.01

Test Accuracy (lag1): 92%
Test Accuracy (lag2): 94%

Decision Tree for Closing SMEs as target:

Feature	Importance
Gross domestic product (GDP) by state (lag1)	0.72
Computer Manufacturing	0.15
PCE: Clothes (lag1)	0.04
PCE: Food (lag1)	0.02
Cattle Production (lag1)	0.01

Test Accuracy (lag1): 98%
Test Accuracy (lag2): 95%

- The modeling was built exclusively with lagged variables (a single and a two-year variable lag)
- OLS, Decision Tree, and XGBoost Regressors were fitted on the dataset
- Most impactful features across all models were centered around economic indicators (specifically GDP)

2. Classification – used to develop a binary growth/decline forecast for each state

Class Definition Opening – Closing:

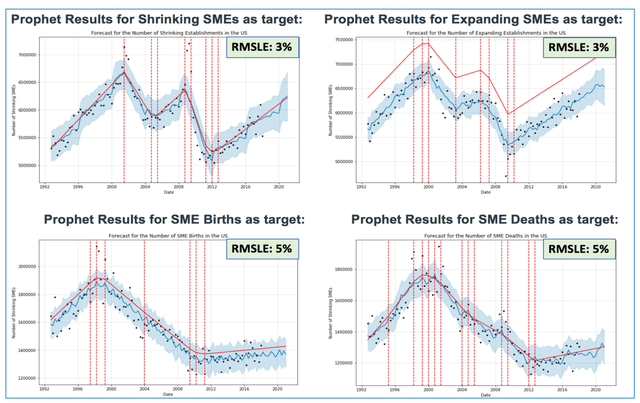
Model	Accuracy	Weighted F1
SVM	96%	94%
Random Forest	87%	90%
GXBoost	93%	93%

Class Definition Expanding – Contracting:

Model	Accuracy	Weighted F1
SVM	87%	85%
Random Forest	92%	88%
GXBoost	90%	88%

- To define the target, each observation was assigned a binary value depending on whether the number of dying/contracting SMEs was less than opening/expanding (1) or vice versa (0)
- The highest performing model was used to generate a state level forecast for defined target for the following year

3. Time Series Analysis – used to forecast SME growth based on historical trend and seasonality



- Utilized the Prophet model developed by Facebook for time series analysis and forecasting
- Model performance was evaluated using root mean squared log error (RMSLE)
- Forecast was generated on a quarterly basis covering the time period of 2018 Q3 - 2020 Q4
- Also generated a number of SME growth forecasts by industry using NAICS codes
- The state level forecasts were used in building the interactive time series tool

Model Productionalization

We chose to develop two interactive tools that summarize the model findings and help users get a closer look into SME growth

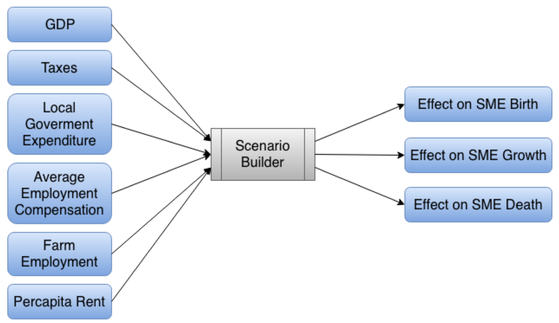
Scenario Builder

Given a percentage change in selected features, scenario builder returns expected effect of individual feature on targets as well as total change. The tool uses the coefficients of logged features in its calculation.

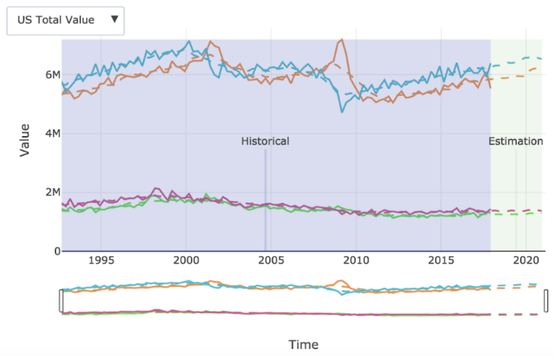
Interactive Time Series

Graphical interpretation of the prophet forecast for defined targets over time. The tool has capabilities around: choosing time range, filtering by state of interest as well as specific target

SME Scenario Builder Logic



Interactive Time Series Screenshot



References

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.

Brock, William A., and David S. Evans. "Small Business Economics." Small Business Economics, vol. 1, no. 1, 1989, pp. 7–20.

Douglas B., G. and Morrison, E., R. "Serial Entrepreneurs and Small Business Bankruptcies." Columbia Law Review, vol. 105, no. 8, 2005, pp. 2310–2368.

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.

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.