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

Introduction: Why SMEs?
SMESmall 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.

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
   • 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
   • 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

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:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Variable Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>expand_establish</td>
<td>Number of expanding SMEs</td>
<td>continuous</td>
<td>Bureau of Labor Statistics</td>
</tr>
<tr>
<td>contract_establish</td>
<td>Number of contracting SMEs</td>
<td>continuous</td>
<td>Bureau of Labor Statistics</td>
</tr>
<tr>
<td>open_establish</td>
<td>Number of opening SMEs</td>
<td>continuous</td>
<td>Bureau of Labor Statistics</td>
</tr>
<tr>
<td>close_establish</td>
<td>Number of closing SMEs</td>
<td>continuous</td>
<td>Bureau of Labor Statistics</td>
</tr>
<tr>
<td>net_change</td>
<td>Net change in SME count</td>
<td>continuous</td>
<td>Bureau of Labor Statistics</td>
</tr>
</tbody>
</table>

The explanatory variables were categorized into two categories:

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Variable examples</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geopolitical</td>
<td>State taxes, medical assistance, regional avg. rent</td>
<td>National Science Foundation; Bureau of Labor Statistics</td>
</tr>
<tr>
<td>Economic</td>
<td>GDP, personal income/consumption, inventory</td>
<td>Bureau of Economic Analysis; United Census Bureau</td>
</tr>
</tbody>
</table>

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

References