A new approach to election forecasting
Providentia enables campaign strategists and political analysts to *uncover new insights* by leveraging machine learning to *forecast elections*.
The Providentia Difference

Providentia offers a non-polls based approach to election forecast that is cost-effective and generates insights with quicker turnaround time.
Most election forecasts use polls.
Yet, there remains a large number of undecided voters who can swing the election

Source: KFF - Data Note: A Look At Swing voters Leading Up To The 2020 Election
Particular focus on swing states dynamics

We believe undecided voters vote by issues, not along ideological lines.

Our Value

Actionable insights for devising campaign plans

User friendly visualizations to show how and where to focus your resources to swing undecided voters.
Overview of Modeling Approach

Datasets
1. Census Bureau demographic and economic data:
   - Age groups by county
   - Education levels
   - Population by ethnicity
   - Unemployment % by county
2. Past election results from 2000 to 2016

Undecided Voter Signals
Factors we considered:
- Twitter, headline news, and google trends sentiment for each candidate
- Voter turnout impacts
- Trending issues in each state
  - Generated from election's enter and exit polls (Healthcare, Economy, Immigration, Climate)

Random Forest Classifier & Linear Regression Model
Modeling approach:
1. Predict winner by county
2. Aggregate county results to infer winner by state
3. Aggregate electoral votes to infer winning candidate

Sentiment Analysis - RNTN & Capsule Neural Network
Modeling approach:
1. Sentiment Score on key topics and candidates
2. Mix popularity index in different parts of the country

Challenges:
1. Overfit due to small dataset
2. Imbalanced dataset
3. Cleaning inconsistent tweets

Parameter Tuning:
1. Grid search
2. Cross validation
3. Neural Network learning rate
4. Optimal tweet size

Final Feature Set:
- Total Population
- Total Votes (~Voter Turnout)
- Unemployment Rate
- Midterm elections outcome
Integration of Regression and NLP Sentiment Analysis

- Predicted Total votes by County
- Aggregated Predictions By State
- Weights (Hyper Parameter)
- Integration of Regression & NLP Model Outputs
  - Votes are adjusted based on sentiment scores by topic specific to state, and by how candidates are trending amongst voters
- Final Electoral Vote Prediction to determine election outcome

Sentiment Scores by Candidates and Topics
Regional Trends and Top Issues By State
Model Evaluation: Accuracy and Backtesting On Past 2 Elections

*Accuracy* defined as: # of correct state predictions/ all states

<table>
<thead>
<tr>
<th>Year</th>
<th>Democratic Electoral Votes</th>
<th>Republican Electoral Votes</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012 (Romney vs. Obama)</td>
<td>Actual: 332</td>
<td>Predicted: 316</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>Actual: 206</td>
<td>Predicted: 222</td>
<td></td>
</tr>
<tr>
<td>2016 (Trump vs. Clinton)</td>
<td>Actual: 233</td>
<td>Predicted: 229</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td>Actual: 305</td>
<td>Predicted: 309</td>
<td></td>
</tr>
</tbody>
</table>
PRELIMINARY 2020 ELECTION FORECAST & PRODUCT DEMO

Target Audience: campaign strategist for a political consultancy

Use Case: where does campaigns need to focus its attention and resources in order to secure victory?
We vs. Them

538’s 2016 Election Forecast

Flipped winning party due to NLP/Trend adjusted sentiment score

Accurate prediction against 2016 actual results

**Providentia’s Predicted 2016 Election Results**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>1,282,494</td>
<td>1,168,931</td>
<td>988,758</td>
<td>1,102,320</td>
<td>Democrat</td>
<td>Democrat</td>
<td>Democrat</td>
</tr>
<tr>
<td>FL</td>
<td>4,474,914</td>
<td>4,066,054</td>
<td>3,702,272</td>
<td>4,111,131</td>
<td>Democrat</td>
<td>Republican</td>
<td>Republican</td>
</tr>
<tr>
<td>IA</td>
<td>614,413</td>
<td>685,742</td>
<td>812,179</td>
<td>740,849</td>
<td>Republican</td>
<td>Republican</td>
<td>Republican</td>
</tr>
<tr>
<td>MI</td>
<td>2,591,647</td>
<td>2,360,214</td>
<td>2,036,997</td>
<td>2,268,429</td>
<td>Democrat</td>
<td>Democrat</td>
<td>Republican</td>
</tr>
<tr>
<td>MN</td>
<td>1,284,561</td>
<td>1,157,777</td>
<td>1,251,116</td>
<td>1,377,899</td>
<td>Democrat</td>
<td>Republican</td>
<td>Democrat</td>
</tr>
<tr>
<td>NV</td>
<td>697,714</td>
<td>642,456</td>
<td>407,445</td>
<td>462,702</td>
<td>Democrat</td>
<td>Democrat</td>
<td>Democrat</td>
</tr>
<tr>
<td>NH</td>
<td>415,165</td>
<td>379,871</td>
<td>290,697</td>
<td>325,990</td>
<td>Democrat</td>
<td>Democrat</td>
<td>Democrat</td>
</tr>
<tr>
<td>NC</td>
<td>2,144,849</td>
<td>1,930,743</td>
<td>2,137,256</td>
<td>2,351,361</td>
<td>Democrat</td>
<td>Republican</td>
<td>Republican</td>
</tr>
<tr>
<td>OH</td>
<td>2,741,897</td>
<td>2,474,911</td>
<td>2,597,810</td>
<td>2,864,795</td>
<td>Republican</td>
<td>Republican</td>
<td>Republican</td>
</tr>
<tr>
<td>PA</td>
<td>3,151,683</td>
<td>2,855,089</td>
<td>2,780,181</td>
<td>3,076,774</td>
<td>Democrat</td>
<td>Republican</td>
<td>Republican</td>
</tr>
<tr>
<td>VA</td>
<td>1,806,485</td>
<td>1,619,038</td>
<td>1,942,441</td>
<td>2,129,887</td>
<td>Democrat</td>
<td>Republican</td>
<td>Democrat</td>
</tr>
</tbody>
</table>
Key Technical Takeaways

1. **NLP to Process Twitter Data:**
   - Twitter data **lacks strict grammatical constructs**
     - Fortunately, it was correlated to the size of the tweets
   - **Removal of short tweets** improved model quality
   - **Stop words** play critical roles in training Sentiment Analysis for text with deep learning
   - **Aspect based sentiment score** is critical when a tweet or headline contains multiple candidates and/or multiple topics

2. **Geo-Mapping in Bokeh:**
   - **Limited number of tools** available for geo-mapping
   - Bokeh is a great choice if the visualization warrants limited interactivity
     - Optimizing Bokeh performance and **integration with HTML/Javascript** was challenging

3. **Data Engineering:**
   - Complexity integrating various sources of data
   - Limited experience with GCP, including integration of Bokeh; **dockerized the app environment** as a workaround
Importance of Clear and Concise Communication & Storytelling

Well crafted stories can communicate complex and abstract ideas that encourage understanding and value connection.

Diversity of Experience, Perspectives, and Skill Sets

Team composition and interdisciplinary interaction helped build on efficiencies and quality, while creating a unique learning experience.
## Future Applications

<table>
<thead>
<tr>
<th>Main Topic</th>
<th>Category</th>
<th>Insights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy Impact</td>
<td>In-depth analysis of COVID impact</td>
<td>Analysis on voter turnout. States with higher number of cases prior to the election are likely to have lower voter turnouts. Some states may not allow vote by mail which could be a determining factor</td>
</tr>
<tr>
<td>Proactively Surface</td>
<td>Electoral vote optimization</td>
<td>Optimization tool on which swing states to invest resources and maximize electoral votes. Enabling scenario analysis for different combinations</td>
</tr>
<tr>
<td>Insights (Pilot Sandbox)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Parse out undecided voters’ voting trends to enable actionable and targeted campaign tactics in battleground states.
Q&A Session