Project 18: Analytical Exploration for Duke Development

Project Leads: Stephen Bayer, Natalie Spring, Ian Conlon | Project Manager: Billy Gerhard
Undergraduates: Natalie Bui, David Cheng, Cathy Lee

Primary Goal

1. Triage prospects to optimize Duke fundraising

Data

Notes by gift officers documenting interactions (contact write-ups), prospects' lifetime donations to duke, and demographic information for approximately 18,000 prospects (70,000 write-ups total)

Machine Learning

From “Qualify” write-ups, use Naive Bayes Classifier to predict if prospect advances to “Cultivate” or “Drop.” Labels were created with historical data and features were TF-IDF values.

<table>
<thead>
<tr>
<th></th>
<th>Cultivate</th>
<th>Recall</th>
<th>F1</th>
<th>Predicted Cultivate</th>
<th>Predicted Drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.92</td>
<td>0.89</td>
<td>0.91</td>
<td>Actually Cultivate</td>
<td>936</td>
</tr>
<tr>
<td></td>
<td>0.83</td>
<td>0.88</td>
<td>0.85</td>
<td>Actually Drop</td>
<td>79</td>
</tr>
</tbody>
</table>

*Because LDA is not effective for short texts, write-ups were split by text length.

Fig. 1a Precision, Recall, F1 metrics

Fig. 1b Confusion matrix
**Topic Modeling**

**Latent Dirichlet Allocation** (LDAMallet implementation) generates topics that occur across all long contact write-ups, and creates a model that predicts **topic occurrence** for each document. This model can be used to predict the topic distribution of new data:

- 100 topics (see Fig. 2) were manually labeled and grouped into 12 distinct initiatives (see Fig. 3).
- Prospects were assigned “compound scores” by averaging the probabilities of each topic under each initiative.

**Fig. 2** Coherence is a measure of how closely related words are in a topic. While $k=50$ had the best coherence score, $k=100$ had a better balance of specificity and coherence when the topics were manually examined.

**Table 1:**

<table>
<thead>
<tr>
<th>Number of Topics</th>
<th>Coherence Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.40</td>
</tr>
<tr>
<td>40</td>
<td>0.45</td>
</tr>
<tr>
<td>60</td>
<td>0.43</td>
</tr>
<tr>
<td>80</td>
<td>0.44</td>
</tr>
<tr>
<td>100</td>
<td>0.45</td>
</tr>
</tbody>
</table>

$\text{Fig. 4}$ Proportion of potential transformative impact and sports prospects increases as donation tier increases. Proportion of potential arts, energy, and I&E prospects decreases, most likely because these initiatives are newer and attract younger prospects who tend to have less capacity.
**Named Entity Recognition**

Identifies names of organizations, events and/or amount of money mentioned in each of the 2,856 short contact write-ups (SpaCy implementation)

*Write-up: ID 12345 “Donated $50,000 to Bass.”*

- “DONATED $50,000 TO BASS”
- Sentence Segmentation
- [“DONATED $50.00 TO BASS”]
- Tokenization
- [‘DONATED’, ‘$50,000’, ‘TO’, ‘BASS’]
- Part of Speech Tagging
  - ([‘DONATED’, ‘VB’], [‘$50,000’, ‘CD’], [‘TO’, ‘TO’], [‘BASS’, ‘NNP’])
- Entity detection
  - ‘$50,000’
  - ‘BASS’

*Categorization: ID 12345: [Transformative Impact, Bass Connections]*

**Future Research**

The following analytical tools and their visualizations can be used to further understand giving trajectories:

- Sentiment analysis
- Network analysis
- Time series analysis

*Fig. 5* This chord graph visualizes the relationships between the different moves that occur in the donation process. Notably, a similar proportion of prospects move to “Cultivate” and “Drop” from “Qualify.”

**Results**

1. A prospect’s “Qualify” write-up is a strong predictor of whether or not the prospect is dropped (Fig. 1).
   - Prospects classified as “Cultivate” donate significantly more and have more subsequent contacts than those classified as “Drop,” which reflects the data accurately.

2. Topic modeling and named entity recognition grouped more than two-thirds of prospects based on interests.
   - Approximately 12,000 unique prospects categorized by topic modeling
   - Approximately 600 unique prospects categorized by named entity recognition

3. Pipeline was incorporated into a python script that can be applied to new data in the future.