Classy Data Analysis

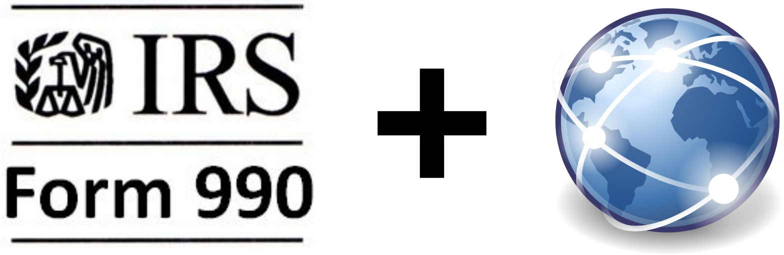
Hansen | Nagda | Pimentel | Reyes | Tai

Challenge

We have a 1-D understanding of NPOs...

- More often than not, the services of an NPO span multiple sectors, e.g Health, Education, etc.
- Financially speaking, small local charities operate very differently from multi-million dollar organizations.

We need a more complex solution to find like-minded organizations.

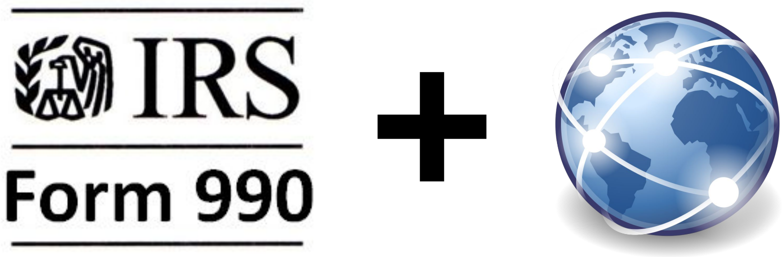


Solution

… but NPOs are multi-faceted.

- Lay out a common “social space”, where the organizations that drive social change and potential donors can connect, find organizations and be offered recommendations, and where discovery of new causes - and events within causes (i.e. fundraisers) - could be facilitated.

- Use combination of government IRS form 990 (returns for nonprofits) data along with external textual information (i.e. social media) to create a robust semantic space.



Agenda

1. Introduction
2. Solution
3. Results and Accuracy
4. Insights
5. Next Steps for modeling
6. Final Product Vision

Results (Text)

Validation / Processing Specs

Preprocessing: Lowercasing / Removal of Numbers + Stopwords + Punctuations / Lemmatization

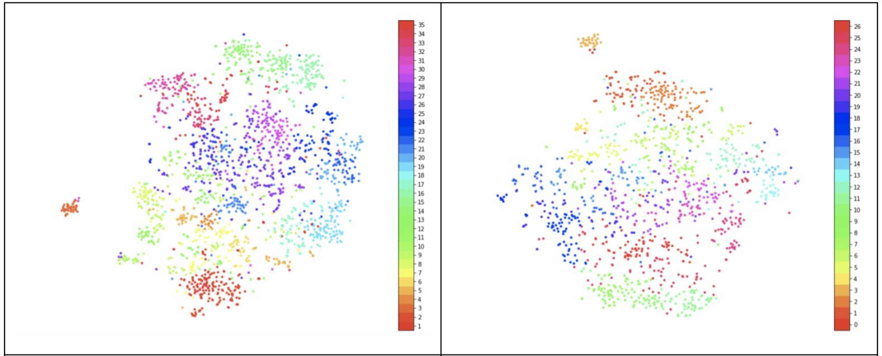
Encoding: Universal Sentence Encoder (512 output)

Dimensionality Reduction: PCA down to 200 features (95% variance) / Down to 70 features (84% variance)

Hierarchical Clustering + Gap Statistic: Linkage matrix generation and gap statistic scanning to find optimal number of clusters

Significance of Results (t-SNE)

The addition of website data did not result in more robust clustering during the initial POC.



No Social Media

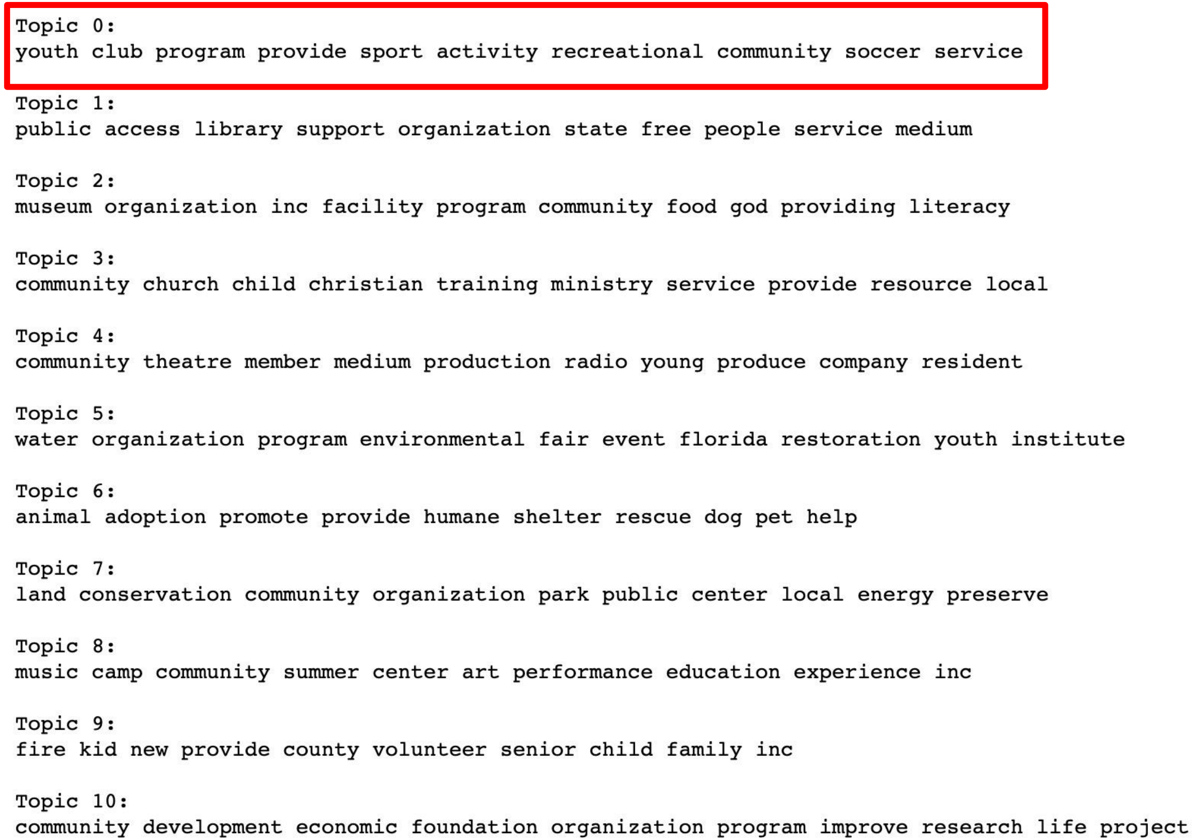
With Social Media

Figure 1: t-SNE Decomposition Scatter Plots

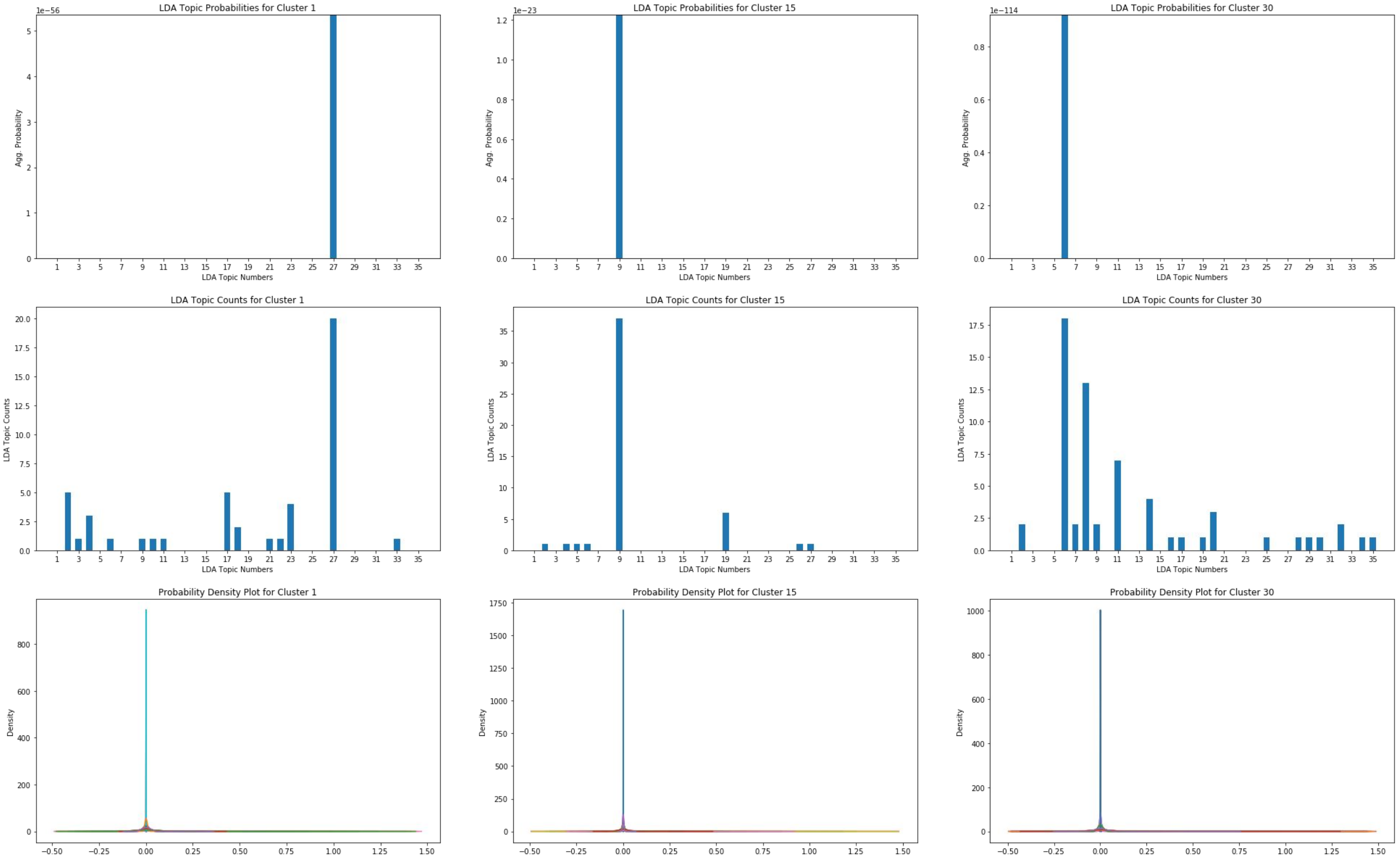
It appears that the vanilla corpus yielded tighter clusters than the website-augmented version. However, this observation is well-aligned with the quality of the web-scraped text. At this stage of the project, our scrapping process remains at a proof-of-concept level.

Significance of Results (LDA)

Topic modeling verified the validity of our text clusters.

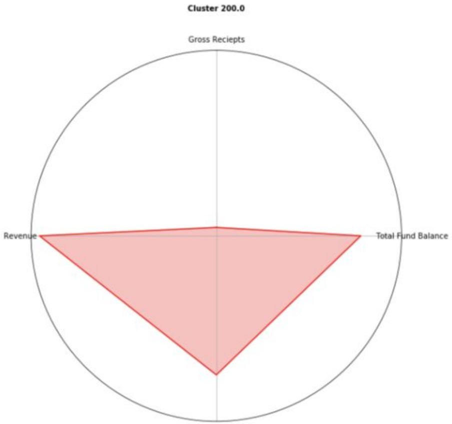


```
Topic 0:
youth club program provide sport activity recreational community soccer service

Topic 1:
public access library support organization state free people service medium

Topic 2:
museum organization inc facility program community food god providing literacy

Topic 3:
community church child christian training ministry service provide resource local

Topic 4:
community theatre member medium production radio young produce company resident

Topic 5:
water organization program environmental fair event florida restoration youth institute

Topic 6:
animal adoption promote provide humane shelter rescue dog pet help

Topic 7:
land conservation community organization park public center local energy preserve

Topic 8:
music camp community summer center art performance education experience inc

Topic 9:
fire kid new provide county volunteer senior child family inc

Topic 10:
community development economic foundation organization program improve research life project
```

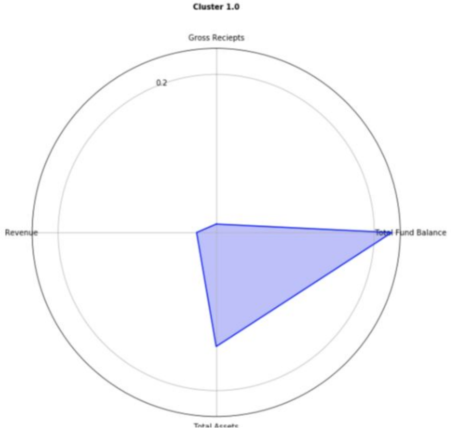

LDA Topic Probabilities for Cluster 1
1e−56
Agg. Probability
0
1
2
3
4
5
LDA Topic Numbers
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
LDA Topic Probabilities for Cluster 15
1e−23
Agg. Probability
0.0
0.2
0.4
0.6
0.8
1.0
1.2
LDA Topic Numbers
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
LDA Topic Probabilities for Cluster 30
1e−114
Agg. Probability
0.0
0.2
0.4
0.6
0.8
LDA Topic Numbers
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
LDA Topic Counts for Cluster 1
LDA Topic Counts
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
LDA Topic Numbers
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
LDA Topic Counts for Cluster 15
LDA Topic Counts
0
5
10
15
20
25
30
35
LDA Topic Numbers
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
LDA Topic Counts for Cluster 30
LDA Topic Counts
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
LDA Topic Numbers
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Probability Density Plot for Cluster 1
Density
0
200
400
600
800
−0.50 −0.25 0.00 0.25 0.50 0.75 1.00 1.25 1.50
Probability Density Plot for Cluster 15
Density
0
250
500
750
1000
1250
1500
1750
−0.50 −0.25 0.00 0.25 0.50 0.75 1.00 1.25 1.50
Probability Density Plot for Cluster 30
Density
0
200
400
600
800
1000
−0.50 −0.25 0.00 0.25 0.50 0.75 1.00 1.25 1.50

Results (Numeric)

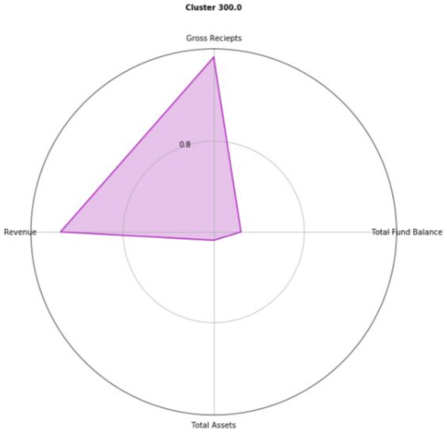
Our aim is to decrease variance within clusters, while maintaining diversity across clusters.



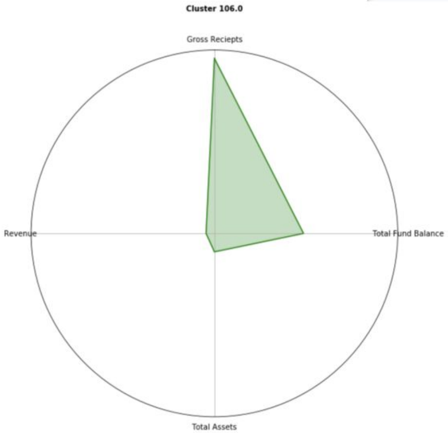
Total Assets & Revenue



Total Assets & Total Fund Balance



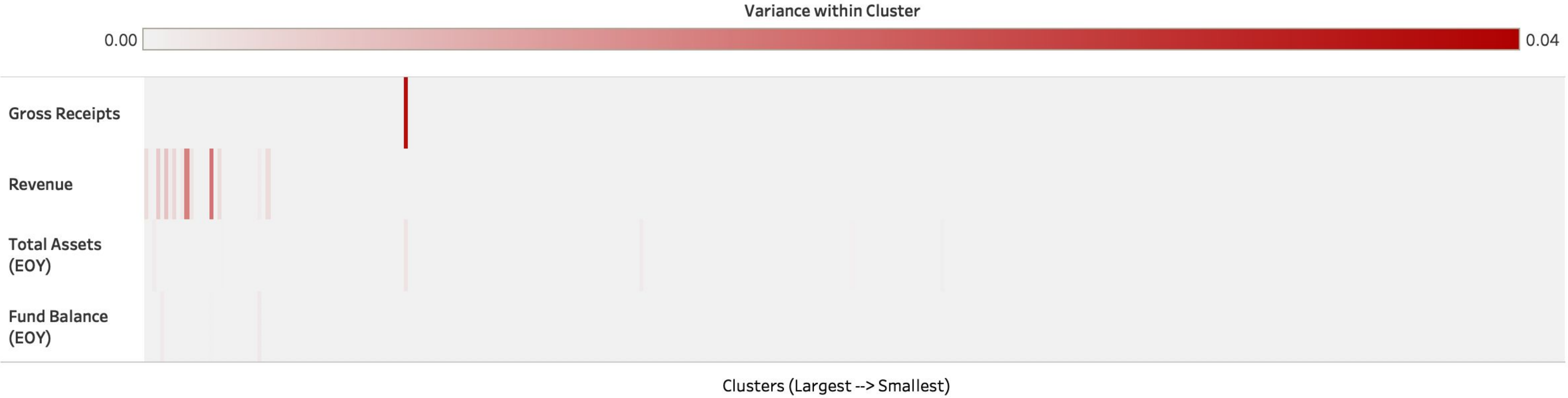
Gross Receipts & Revenue



Radar plots similar to those above allow us to describe individual clusters in a visual way. We can also use these plots to verify the diversity of our clusters.

Significance of Results

Feature engineering aided in model convergence.



The introduction of new features such as % changes from BoY to EoY and funding allocation ratios cut the number of clusters in half and reduced the number of clusters containing only a single nonprofit.

Next Steps for Modeling

Need more robust web text processing - Normalizing terms

- Adding the same text processing methods used for the 990 forms to the text gather from the websites.

- Also looking into normalizing and spell-checking some of the terms of high frequency in the financial forms already made public by researchers investigating the forms.

- Need some heuristics on selecting the best pages and forms to scrape for each website and the alternative to augment with Google descriptions of those websites.

Next Steps for Modeling

Feature engineering and further partitioning has proven to be an effective method for model improvement..

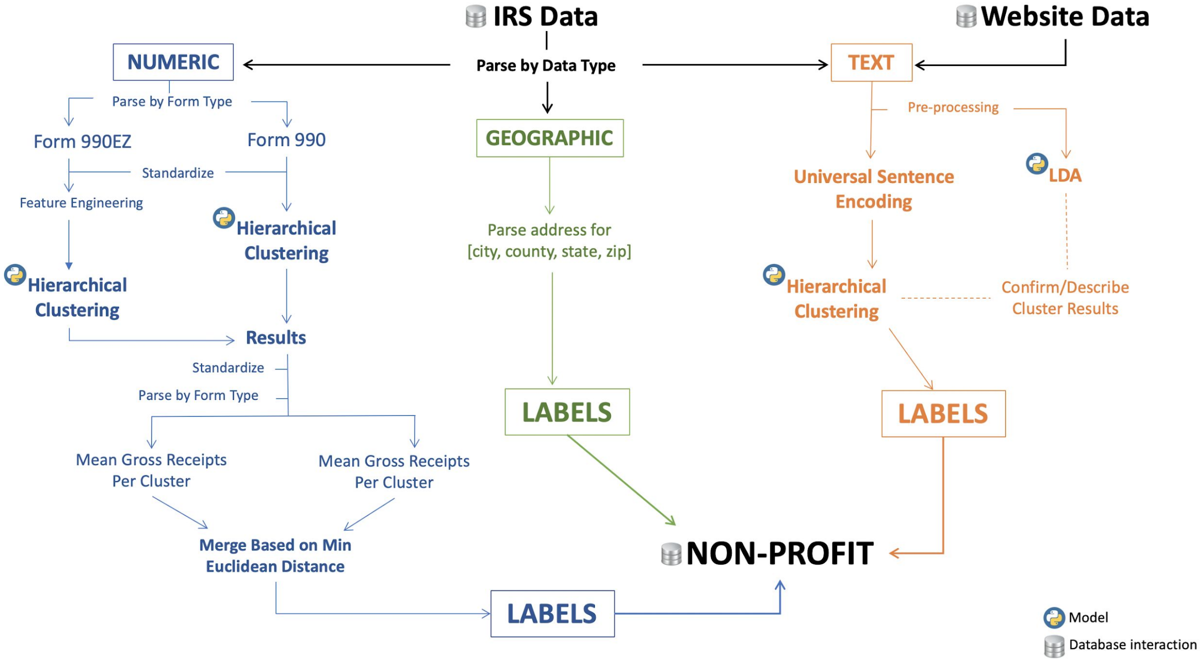
- Simple calculations have greatly improved our model; however, there are several financial features left unexplored at this time.

- We will continue to build new visualizations as a means of describing our results and exploring the features within our model

- Now that a strategic pipeline is in place, we can focus our attention on analyzing the results and integrating with the results from the text model

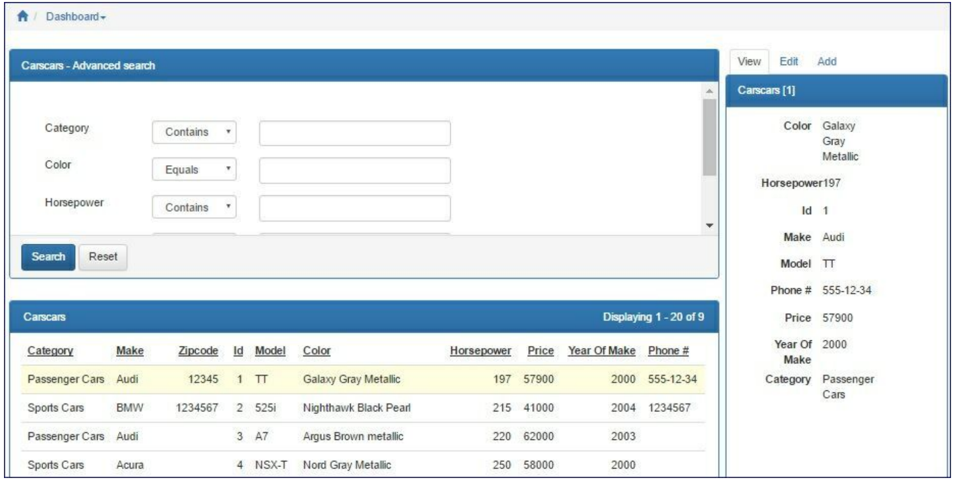
Next Steps for Modeling

Feature engineering and further partitioning has proven to be an effective method for model improvement..



Final Product Vision

- Each org will belong to 3 clusters. User has power to determine how to define “similarity”: financial, mission, geographic, or any combination of the three
- Dashboard interface where user can search for non-profit organizations and view its designated peer group
- Visualizations to understand different aspects of non-profit organization
  - financial metrics as well as descriptions of results from text model



Category	Contains	
Color	Equals	
Horsepower	Contains	

Carscars [1]

Color	Galaxy Gray Metallic
Horsepower	197
Id	1
Make	Audi
Model	TT
Phone #	555-12-34
Price	57900
Year Of Make	2000
Category	Passenger Cars

Carscars — Displaying 1 - 20 of 9

Category	Make	Zipcode	Id	Model	Color	Horsepower	Price	Year Of Make	Phone #
Passenger Cars	Audi	12345	1	TT	Galaxy Gray Metallic	197	57900	2000	555-12-34
Sports Cars	BMW	1234567	2	525i	Nighthawk Black Pearl	215	41000	2004	1234567
Passenger Cars	Audi		3	A7	Argus Brown metallic	220	62000	2003	
Sports Cars	Acura		4	NSX-T	Nord Gray Metallic	250	58000	2000	

insights

Text modeling:

- The USE encodings / hierarchical clustering combo is a robust model
- External web/social data has potential but we need to reduce the noise

Numerical modeling:

- Variance within clusters serves as a solid metric to track model performance
- Merging form type clusters makes sense in most (if not all) cases
- Feature engineering had a greater impact on results than partitioning

Questions?