Machine Learning for Community Forecasting

Christopher Fleisher
Computer Science

Loren Heyns
Computational Science & Eng.

Huyen Nguyen
Bioinformatics

Truc Pham
Computer Science

Trai Tran
Computer Science

Zhengfu Li
Electrical & Computer Eng.

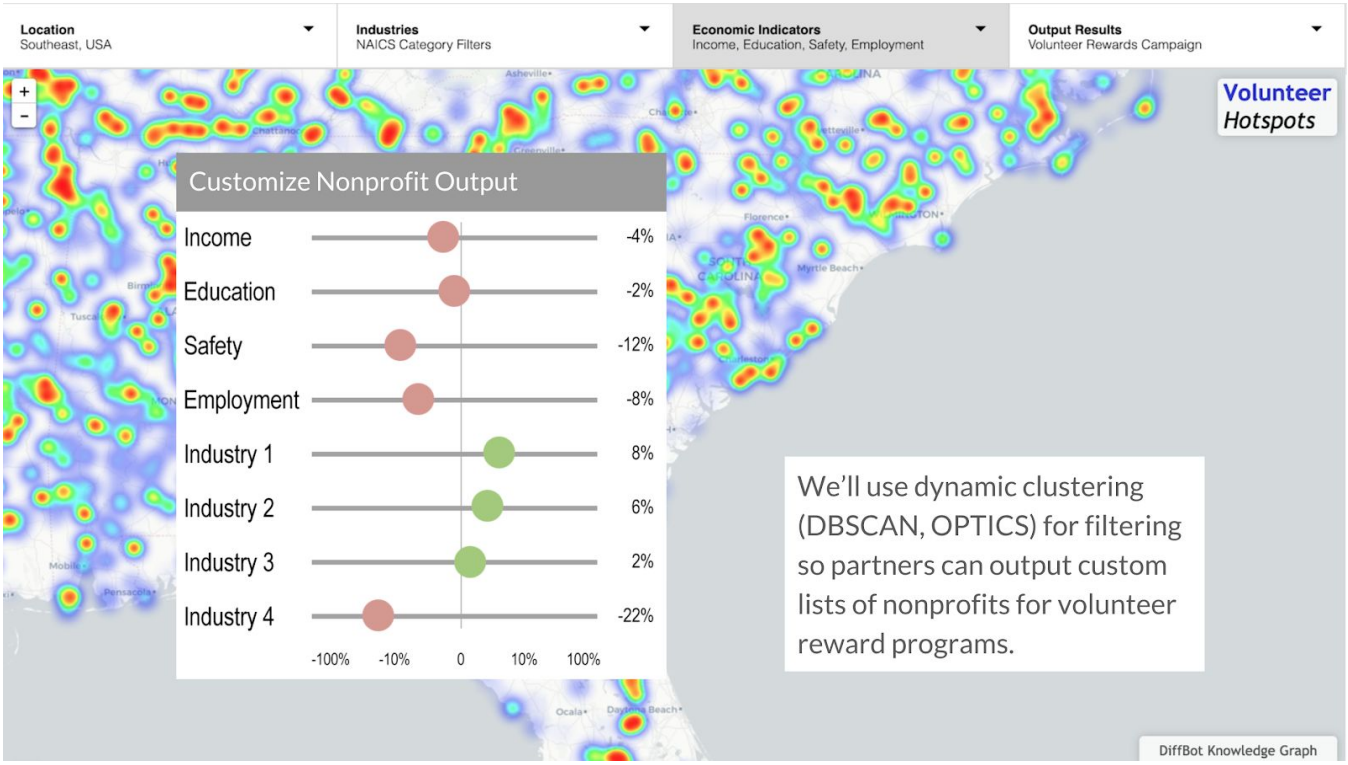

Figure 1. Universal navigation filters and predictive heatmap. Users can adjust settings above the map to filter income, education, employment, industries, crime levels and local nonprofit locations. A timeline slider allows the user to view data from 5 and 10 years prior, which is used to project trends 3-years into the future. Related content is pulled into our browser-based data filters from the DiffBot Knowledge Graph API.

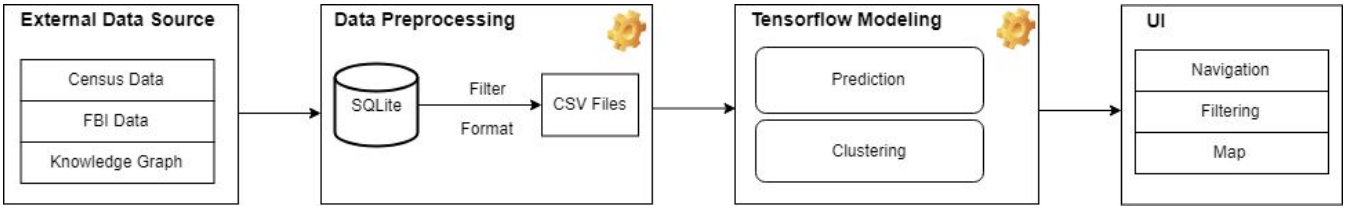

Figure 2. Application Architecture Diagram

ABSTRACT

Browser-based machine learning allows for more accessible predictive analysis tools for use in community planning. Our DataScape model uses spatial data to identify downward economic trends to fine-tune volunteer incentive programs. Combining data from Points of Light, the world's largest volunteer management organization, we compare locations of 6,000 nonprofits within 120 cities with US Census data on income, industries, job growth, education attainment and FBI crime statistics to track the changing economic health of communities across time aid in improving the effectiveness of volunteer reward campaigns.

INTRODUCTION

The DataScape model tackles an analysis need posed by Disney World: How to determine where to focus volunteer incentive programs to have the greatest impact on lower income communities. Disney provides nonprofits with tickets to reward their members for volunteering. To avoid a concentration of reward allocation within more affluent communities, we'll use machine learning to identify nonprofits within declining economic zones so reward providers can filter organization outreach lists by weighted parameters.

RELATED WORKS

In 2018, TensorFlow.js ushered in a new era in browser-based GPU processing [1], adding increased analytical power to client-side analytics and self-contained search [2]. Extensive Knowledge Graphs by DiffBot and Google allow browser-based filtering to be combined with on-demand searches for extra details [3].

Applying data analysis using clustering to analyze public data, including health program participation [4] provides a tool for community transformation and revitalization [5]. Following the expansion of the American Community Survey (ACS) to the local level for the entire country in 2011, [6] regression analysis is increasingly used to predict localized trends.

DATASCAPE MODEL

Clustering

Our model clusters locations with respect to their changing demographic properties, including income, education, NAICS industry levels, employment decline and crime as an indicator of community decline. Combining clustering with multiple year analysis evens out between-year differences. [7] A variety of data mining techniques for crime analysis [8] include detecting vulnerable "hot spots" using cluster analysis [9], machine learning algorithms [10], and a variety of regression methods [11] [12].

To support a large volume of data points and real-time responsiveness within our web app, we'll use the highly efficient clustering algorithm DBSCAN++ [13] which provides performance similar DBSCAN, but with higher time efficiency achieved by computing density of chosen subset each time.

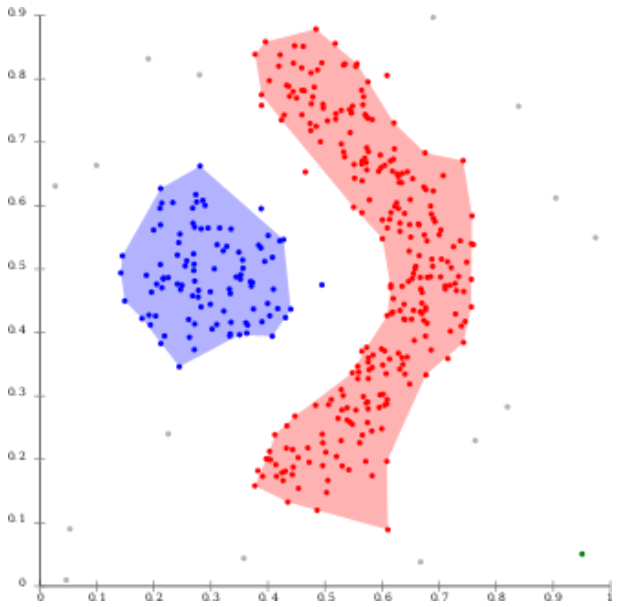

Figure 3. Outliers tend to exist in density-based clustering method, previous work [14] proves that with certain restrictions, interactions from user editing, namely requesting for splitting and merging clusters locally, can lead to desired clustering.

DataScape will allow users to manipulate the weight of factors used for computing dissimilarities in clustering. (See figure 1) Since weighted clustering produces more accurate clustering [15], we'll generate weighting by optimizing the classic K-means algorithm. We'll test using K-means to display the most notable attribute in each cluster to help our user narrow their preferred locations. The OPTICS algorithm will also be implemented as an alternative of density-based clustering. OPTICS can be used for processing and exploring large datasets effectively [16] Another alternative approach is to use semi-supervised clustering, which has been shown [17] to be effective when a small number of datasets are labelled.

Prediction Model

Our economic data will be projected based on a Vector Autoregression Moving-Average (VARMA) model based on the presumption that our time-series exhibits non-causal relationships across local, state and national geographies. VARMA is a weak-stationarity extension of vector autoregression (VAR) incorporating a moving average into the error-term to reflect short-term stochasticity. The moving average error-term has recently been shown to improve regression forecasts across a range of related economic datasets and time horizons [18].

VAR models are a well-known method of multivariate linear regression used in economic forecasting to project outputs based on lagged input of all other variables. Both time lag and variable coefficients are used to capture model relationships [19]. We will vary time lags to optimize both accuracy and precision.

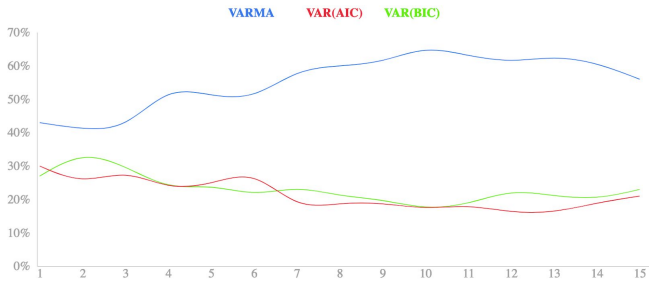

Figure 4. PB counts for |MSFE| for VARMA vs unrestricted VAR selected by AIC and BIC [18]

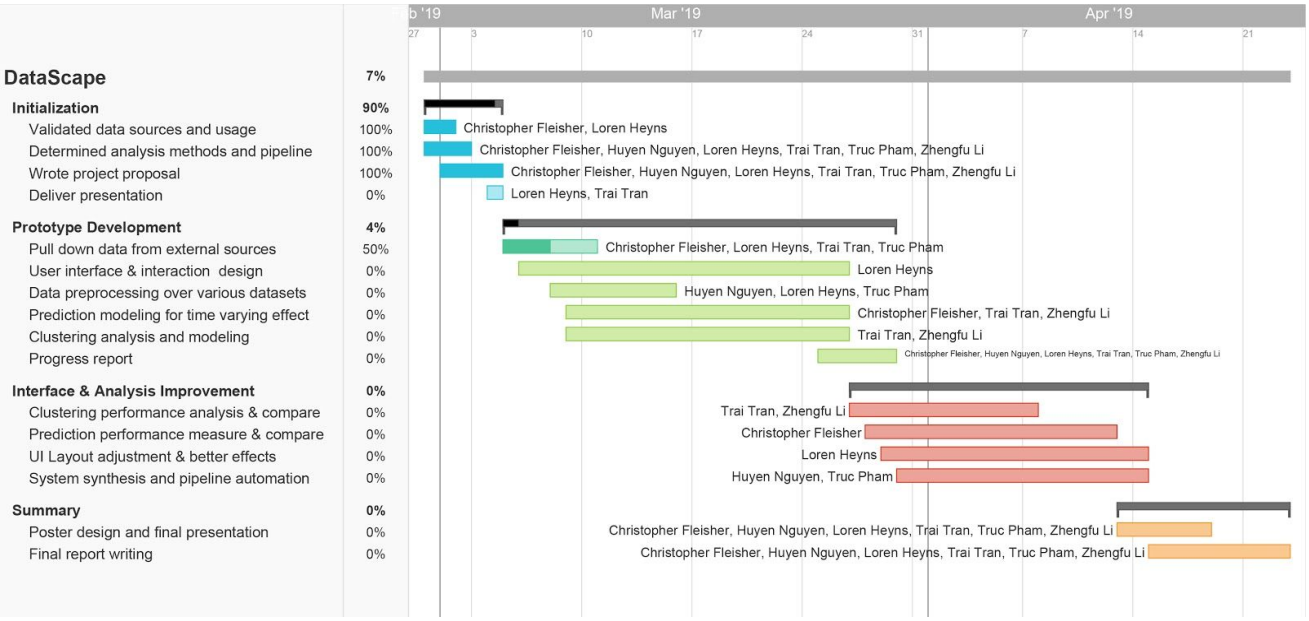

Figure 5. Plan of Activities

One drawback of linear regression analysis is its susceptibility to overfitting. We will explore recent advances with artificial neural nets to try and improve our baseline generalization without impacting the model's near-term predictive power. A Flexible Neural Tree (FNT) may leverage the inherent generalizability of non-linear systems given our problem domain's temporal constraints [20]. Neural network prediction analysis is prevalent in predicting the state of an entity based on historical data. The learning techniques presented in [21] can be used to predict community status based on its properties.

Data Preparation

Data analysis techniques fall into 2 categories, content and thematic [22]. The data preparation for this project will involve joining data from different sources that are related to each other in term of similar trends explaining the health of a community, such as education, income level, crime, and health care issues, which falls under the thematic techniques pinpointing patterns in the data [23].

The datasets are processed by MapReduce programming model that facilitates automatic parallelization and distribution of large-scale computations because it's highly efficient for a big scale fault-tolerant data analysis [24] [25]. The original platform of MapReduce has poor interface for interactive data analysis since users cannot view results faster all jobs are executed, and the two-state data flow's too rigid for different data flow such as joining, projection and filtering. Thus, Pig Latin is proposed to resolve those limitations by combining high-level declarative querying and procedural map-reduce model. The program is easy to use, allowing programmers to see the flow of their data processing and fix the errors with its novel debugging environment [26].

Cost-free Data Analysis

Hosting is provided free on GitHub. CSV files are pre-processed, so interaction occurs quickly in the browser with additional server-side process in Google's free Collab Jupyter notebooks for collaborative Python. Data from the US Census, FBI and DiffBot are also free.

Development Time Span

Project work over 1.5 months with additions continuing in GitHub indefinitely as multiple contributors refine the model.

EXPECTED OUTCOMES

Measuring Progress

Our project's "midterm review" will be conducted with staff members at Points of Light. Our "final exam" will occur via a soft-launch with corporate users managing Disney rewards.

Determining Success

Project success will be measured by changes in reward ticket distribution to lower income nonprofits, and by collecting feedback from users with multiple prompts and GitHub usage stats.

Risks and Payoffs

Risk resides in making inaccurate predictions, hence disclaimers will be visible on all projections. The current economic status for each community will be the default view so the focus is on the most certain data.

CONCLUSION

Increasing access to community economic forecasting models will help volunteer reward programs design local and national initiatives with the greatest impact through the use of both clustering and predictive modeling.

REFERENCES

1. Kahng, Minsuk, et al. "GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation." IEEE transactions on visualization and computer graphics 25.1 (2019): 310-320.

2. Lin, Jimmy. "Building a self-contained search engine in the browser." *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ACM, 2015.

3. Costa, Jose Ortiz, and Anagha Kulkarni. "Leveraging Knowledge Graph for Open-domain Question Answering." *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018.

4. Mueller, Erik, et al. "A Cluster-Based Machine Learning Ensemble Approach for Geospatial Data: Estimation of Health Insurance Status in Missouri." ISPRS International Journal of Geo-Information 8.1 (2019): 13.

5. Wright, Nathaniel S. "Transforming neighborhoods: Explaining effectiveness in community-based development organizations." Journal of Urban Affairs 40.6 (2018): 805-823.

6. Glenn, Ezra Haber. "Estimates with Errors and Errors with Estimates: Using the R'ACS'Package for Analysis of American Community Survey Data." Available at SSRN 2590391 (2015).

7. Siordia, Carlos. "Detecting “real” population changes with American Community Survey data: The implicit assumption of treating between-year differences as “trends”." Journal of Sociological Research 4.2 (2014): 494-509.

8. Thongsatapornwatana, Ubon. "A survey of data mining techniques for analyzing crime patterns." 2016 Second Asian Conference on Defence Technology (ACDT). IEEE, 2016.

9. Grubesic, Tony H., and Alan T. Murray. "Detecting hot spots using cluster analysis and GIS." Proceedings from the fifth annual international crime mapping research conference. Vol. 26. 2001.

10. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyze crime data." *Machine Learning and Applications: An International Journal (MLAIJ)*2.1 (2015): 1-12.

11. Raghavendhar T.V, Joshy Joslin, Mahaalakshmi R, and Soni M Ashutosh. “Crime Prediction and Analysis Using Clustering Approaches and Regression Methods.” 5, no. 4 (2018): 6.

12. Wong, Ka-Chun. "A short survey on data clustering algorithms." *2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI)*. IEEE, 2015.

13. Jang, Jennifer, and Heinrich Jiang. "DBSCAN++: Towards fast and scalable density clustering." *arXiv preprint arXiv:1810.13105* (2018).

14. Awasthi, Pranjal, Maria Florina Balcan, and Konstantin Voevodski. "Local algorithms for interactive clustering." *The Journal of Machine Learning Research* 18.1 (2017): 75-109.

15. Chan, Elaine Y., et al. "An optimization algorithm for clustering using weighted dissimilarity measures." *Pattern recognition* 37.5 (2004): 943-952.

16. Ankerst, Mihael, et al. "OPTICS: ordering points to identify the clustering structure." ACM Sigmod record. Vol. 28. No. 2. ACM, 1999.

17. Basu, Sugato, Arindam Banerjee, and Raymond Mooney. "Semi-supervised clustering by seeding." In Proceedings of 19th International Conference on Machine Learning (ICML-2002. 2002.

18. Athanasopoulos, George, Farshid, Vahid, "VARMA versus VAR for Macroeconomic Forecasting." *Journal of Business & Economic Statistics*, (2008).

19. Lutkepohl, Helmut "Forecasting with VARMA Models." *Handbook of Economic Forecasting*, Vol. 1 (2006)

20. Chen, Y., Yang, B., Dong, J., & Abraham, A. "Time-series forecasting using flexible neural tree model." *Information Sciences*. 174(3–4), 219–235. (2005)

21. Leshno, Moshe, and Yishay Spector. "Neural network prediction analysis: The bankruptcy case." *Neurocomputing 10.2* (1996): 125-147.

22. Namey, Emily, et al. "Data reduction techniques for large qualitative data sets." *Handbook for team-based qualitative research 2.1* (2008): 137-161.

23. Braun, Virginia, et al. "Thematic analysis." *Handbook of Research Methods in Health Social Sciences* (2019): 843-860.

24. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.

25. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: a flexible data processing tool." *Communications of the ACM* 53.1 (2010): 72-77.

26. Olston, Christopher, et al. "Pig latin: a not-so-foreign language for data processing." *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008.