Training Machine Learning Models to Characterize Temporal Evolution of Disadvantaged Communities

Milan Jain, Narmadha Meenu Mohankumar, Heng Wan,
Sumitrra Ganguly, Kyle D Wilson, David M Anderson
Pacific Northwest National Laboratory, Richland, WA, USA
{milan.jain, narmadha.mohankumar, heng.wan, sumitrra.ganguli, kyle.wilson, dma}@pnnl.gov

Abstract

Disadvantaged communities (DAC), as defined by the Justice40 initiative of the Department of Energy (DOE), USA, identifies census tracts across the USA to determine where benefits of climate and energy investments are or are not currently accruing. The DAC status not only helps in determining the eligibility for future Justice40-related investments but is also critical for exploring ways to achieve equitable distribution of resources. However, designing inclusive and equitable strategies not just require a good understanding of current demographics, but also a deeper analysis of the transformations that happened in those demographics over the years. In this paper, machine learning (ML) models are trained on publicly available census data from recent years to classify the DAC status at the census tracts level and then the trained model is used to classify DAC status for historical years. A detailed analysis of the feature and model selection along with the evolution of disadvantaged communities between 2013 and 2018 is presented in this study.

Introduction

In 2020, the Department of Energy (DOE) introduced Justice40 initiative, which directs 40% of the overall benefits of certain Federal investments – including investments in clean energy and energy efficiency; clean transit; affordable and sustainable housing; training and workforce development; the remediation and reduction of legacy pollution; and the development of clean water infrastructure – to flow to disadvantaged communities (DACs) (DOE 2022). While using the percentile values of 36 indicators collected from numerous data sources, the initiative proposed a methodology to classify census tracts as Disadvantaged Communities (also referred to as DAC in the rest of the paper). The current version of DOE J40 DAC data (2022c) identifies 15,172 census tracts across the United States as DAC, out of which 262 belong to the state of WA (region of interest for this study).

A deeper understanding of these disadvantaged communities across the nation is crucial for designing equitable and inclusive policies for the communities. However, designing such inclusive and equitable strategies/policies not just requires a good understanding of current demographics, but also a deeper understanding of the transformations that happened in those demographics over the years. But the major hurdle in carrying out such a deep dive into the past is the data availability. The 36 indicator variables used by the Justice40 initiative are collected from numerous data sources, most of which were not even recorded in the past.

Existing studies have have primarily focused on ensuring if the strategies are appropriately designed to respond to critical deficiencies in the DAC communities. Examples include digital-sharing economy, enterprise zone programs, and community development financial institutions and corporations for stimulating employment, income, reciprocity, social interaction, and resource accessibility for the DAC communities (Vidal 1995; Dillahunt and Malone 2015). Akin to that, other studies have also argued that the disparity in spatial accessibility of infrastructure is strongly associated with inequalities among communities and that equitable distribution of public and private sector investments in infrastructure projects is critical (Leyshon and Thrift 1994; Brown and Lloyd-Jones 2014; Mandarano and Meenar 2017; Wiesel, Liu, and Buckle 2018). While these studies exist, the studies that investigate the evolution of DAC status and the transformations of the determinants of DAC status over the years are important but lacking.

In this study, we tackle this challenge by training Machine Learning (ML) models on different combinations of the LODES data (LEHD Origin-Destination Employment Statistics) (Bureau 2022b) and the ACS (American Community Survey) data. For data description, please refer to the Appendix. The trained ML models act as a proxy for those 36 indicators in classifying the DAC status. Once trained, the best trained model is used to classify the DAC status of the census tracts for any time period for which the feature data is available. In this study, we particularly focused on feature selection and model selection to train most accurate model to project DAC for the historical data with limited bias.

Methodology

Data Collection & Preprocessing

For this study, we collected data from three sources: LEHD Origin-Destination Employment Statistics (LODES) Data, American Community Survey (ACS) 5Y Estimates, and DOE Justice40 DAC Data. DOE Justice40 2022c edition

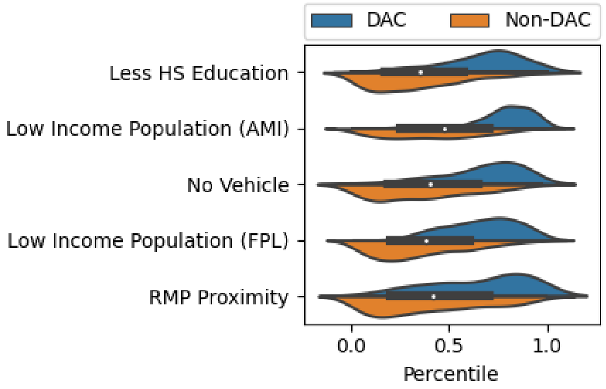


Figure 1: Top 5 indicators (out of 36 DOE J40 DAC indicators) best differentiates DACs from Non-DACs. *Less HS Education* indicates the % of total population, age ¿25, whose reported education is short of a high school diploma. *Low Income Population (AMI)* depicts the % of total population which is considered low income based on area median income (AMI) and *Low Income Population (FPL)* is the % of total population reported at or below 200% of the Federal Poverty Level (FPL). *RMP Proximity* indicates proximity to Risk Management Plan (RMP) facilities.

of the data uses 2018 LODES data and 2019 ACS 5Y estimates. Figure 1 shows the top 5 indicators (out of 36 DOE J40 DAC indicators) that best differentiates DACs from non-DACs. The percentile values shown on the x-axis is the rank of the specified indicator in the national data as a percent of the full dataset. The DAC score, based on which DAC status is decided, is the sum of those percentile values.

It is evident from the plot that the features related to income and education are most important. To capture these indicators directly/indirectly, following five versions of training datasets were prepared for model training:

1. `v1a: LODES(R)` - In this variant, we use all home-area characteristics (see Table 2) of the census tracts from the LODES data in the feature set.
2. `v1b: LODES(W)` - In this variant, we use all work-area characteristics (see Table 2) of the census tracts from the LODES data in the feature set.
3. `v1c: LODES(R+W)` - In this variant, we use both home- and work-area characteristics (see Table 2) of the census tracts from the LODES data in the feature set.
4. `v2a: LI(R)+ACS` - Though demographics (specifically race and ethnicity) are highly correlated with the DAC status, it is hard to intervene demographics through policy design. In this variant, we exclude demographic information (age, sex, race, and ethnicity) and only incorporate number of employed people and employment by industry from the LODES residential-area characteristics data. In addition, LODES income bins have low resolution, are not adjusted for inflation and do not capture the household income. Since income is an important feature (as shown in Figure 1), high-resolution 16-bin household income adjusted for inflation from the ACS data is also included in the feature set.
5. `v2b: LI(R+W)+ACS` - The employment by industry from LODES residential-area characteristics only captures the distribution of industries where people work. To capture distribution of industries that exist in a census tract, we incorporated employment by industry from LODES work-area characteristics (WAC) in this variant, along with existing features from the previous variant.

Prior to training, the data is normalized with total number of employed people for `v1a: LODES(R)`, `v1b: LODES(W)`, and `v1c: LODES(R+W)`, and by total population for `v2a: LI(R)+ACS` and `v2b: LI(R+W)+ACS`.

Model Training

We used H2O.ai for model training and selection. H2O.AI is an AutoML framework (LeDell and Poirier 2020; H2O.ai 2021) extensively used by the community for automating the ML workflow. Training included 30 different variants of 5 key models supported by the H2O.ai library: (1) Distributed Random Forest (DRF), (2) Deep Learning, (3) Gradient Boosting Machines (GBM), (4) Generalized Linear Model (GLM), (5) XGBoost, and (6) Extremely Randomized Trees (XRT). Model details can be found on H2O.AI algorithms page (H2O.ai 2021).

For training, the data was split into 67:33, with 67% data used to train 30 different models and remaining 33% data for model evaluation and selection. The data was standardized by subtracting mean and dividing by standard deviation.

Inference

Though LODES data exists since 2002, a number of features were not included until 2009. Likewise, income information from ACS is only available starting 2013. Therefore, we used the best trained model to infer the DAC status for 5 years: 2013-2017. Lastly, we correlate temporal evolution of estimated DAC status with the most important features, as identified from the trained models.

Evaluation

Table 1 depicts the F1-score of models trained on data from 968 census tracts (165 DACs), when evaluated on the test data which includes 477 census tracts (97 DACs) from the state of WA. Details about the tuned parameters of the best models are provided in Table 4 (on the last page).

Table 1: Feature Engineering and Model Selection

	DRF	DeepLearning	GBM	GLM	XGBoost	XRT
LODES(R)	**0.70**	0.69	0.68	0.68	0.68	0.69
LODES(W)	0.49	0.50	**0.55**	0.52	0.53	0.52
LODES(R+W)	0.68	0.69	0.69	0.71	0.69	**0.74**
LI(R)+ACS	0.72	0.71	0.74	0.75	**0.75**	0.69
LI(R+W)+ACS	0.70	0.72	**0.78**	0.74	0.75	0.70

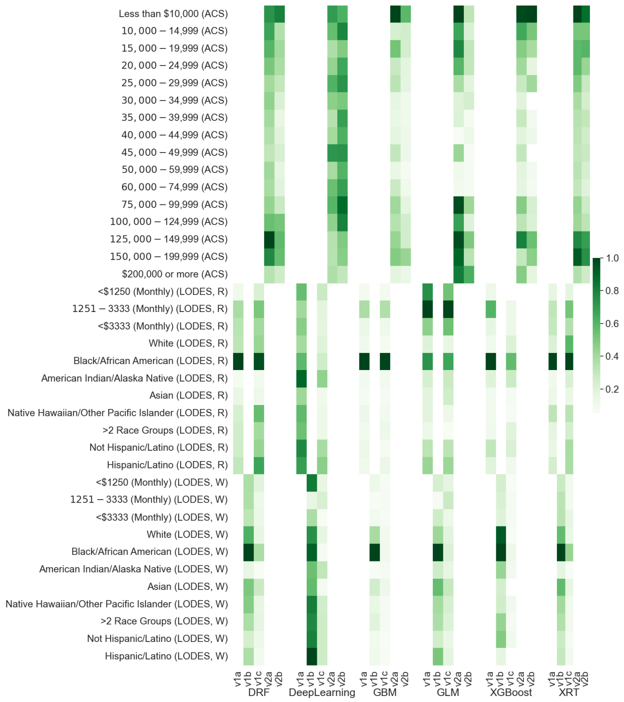


Figure 2: Feature Importance (Race, Ethnicity, and Income): x-axis shows data variants grouped by model. Here, feature importance only from the best variant of the model is shown.

Figure 2 shows the importance of features related to race, ethnicity, and income, and Figure 3 shows the feature importance of employment by industry features for every combination of data variant and ML model (its best version). For GLM, variable importance indicates the coefficient magnitudes. For tree based algorithms (GBM, DRF, XRT, and XGBoost), the variable importance is determined by calculating the relative influence of each variable: whether that variable was selected to split on during the tree building process, and how much the squared error (over all trees) improved (decreased) as a result. Finally, for the Deep Learning model, H2O.AI uses Gedeon method, that considers the weights connecting the input features to the first two hidden layers to compute the variable importance (Gedeon 1997).

Following are some key takeaways:

- Residential-area characteristics alone are better estimator of DAC status than the work-area characteristics.
- Combining residential- and work- area characteristics offer very little improvement over residential-area characteristics alone (comparing `v1a: LODES(R)` with `v1c: LODES(R+W)`).
- Models trained on `v1a: LODES(R)`, `v1b: LODES(W)`, and `v1c: LODES(R+W)` features rely heavily on race and ethnicity distributions (from Figure 2). Dependence on race and ethnicity introduces bias in the model. For instance, the biased model trained on these features sets was projecting census tracts with high African-American population as DAC.

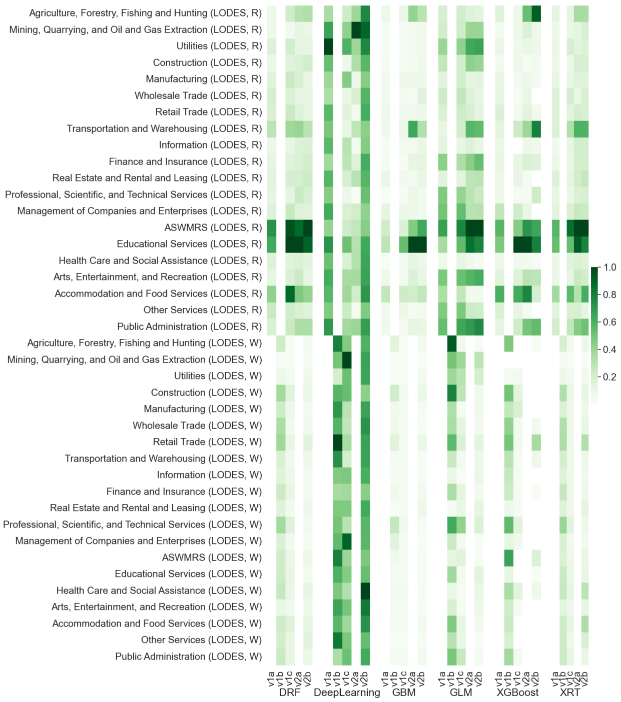
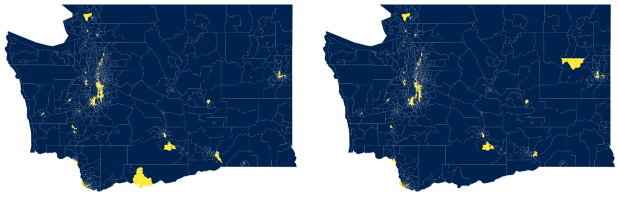


Figure 3: Feature Importance (Industry): x-axis shows data variants grouped by model. Here, feature importance only from the best variant of the model is shown.



(a) Actual

(b) Predicted

Figure 4: Comparing Actual DAC communities with the predicted DAC communities.

- Removing demographic information (race, ethnicity, gender, and age) and only using employment by industry from LODES residential area-characteristics along with high-resolution income bins from ACS (`LI(R)+ACS`) improved the DAC estimation accuracy.
- Three bin income categorization from LODES only covered income of employed people and hardly provided any separation between DAC/non-DAC to the model. Instead, 16-bins of household income (adjusted for inflation) from ACS seems like better feature set for the DAC classification, especially the low- and high-income bins (except for Deep Learning), as shown in Figure 2.
- Industries like Transportation and Warehousing, Educational Services, Administrative and Support and Waste Management and Remediation Services (ASWMRS), Accommodation and Food Services, and Public Admin-

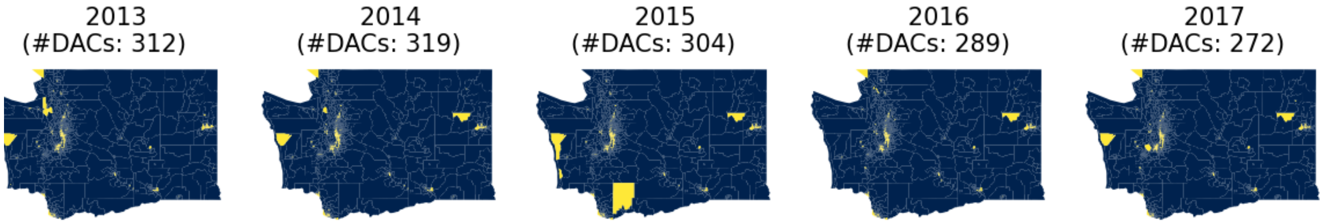


Figure 5: DAC Estimation on Historical Data (2013-2017)

istration come up as important features (Figure 3).

- Employment by industry from residential-area characteristics only capture *where people work* and not *what kind of industries operate in the area*. To address this, we incorporated employment by industry from work-area characteristics in `LI(R+W)+ACS`. However, including this additional information into the feature set offers very limited improvement, if any, across all the models.

It is evident from the analysis that Gradient Boosting Machine (GBM) using `v2b: LI(R+W)+ACS` feature set provides the best DAC classification accuracy of 78%, without any bias towards a particular community. Figure 4 compares actual DACs with the estimated DACs on the map of WA state. While 89% census tracts are correctly identified, there are a few false positives and a few false negatives.

False Negatives (what model missed!): On analyzing false negatives, we noticed that the census tracts that model failed to identify as DAC failed on two key indicators of DOE J40 DAC definition - Low Income Population (AMI) and Low Income Population (FPL). When compared with true positive tracts, low income related indicators are relatively lower for false negative tracts. The trained model missed these tracts because income is a an important feature (see Figure 1). A natural follow-up question is - *why these tracts are DACs then?* Our analysis found that the age of the house (exposure to lead) is a key factor driving the DAC status of these tracts, which is not captured in the feature set.

False Positives (model thinks it's a DAC!): On analyzing the false-positives, we noticed that a number of false positive instances are those census tracts that belong to big cities like Seattle and Spokane. Typically, in such census tracts, industries like Transportation, Waste Management, and Public Services have substantial presence - an important set features for DAC estimation by the trained models, as shown in Figure 3. Besides, these tracts have relatively high percentage of households from both low and high income groups, which is another set of important features of the trained models, as shown in Figure 2. However, these tracts were not identified as a DAC by the DOE J40 DAC definition.

Though the false positive rate can be reduced by incorporating additional features differentiating big cities from relatively smaller cities, they do offer an opportunity to further analyze those communities as potential DAC communities. Given that the DAC definition is still in experimental stage, evaluation of such false positive instances as potential DAC community becomes even more important.

DAC Estimation (Historical Data)

Overall, though the disadvantaged census tracts is distributed very similarly between 2013-2017, the total number of DAC communities seems to have decreased over time. When correlated with important features (from Figure 2 & 3), a decrease in low-income (household) group and increase in high-income (household) group were noticed in the state of WA between 2013 and 2017. One must note here that the household income reported in ACS is already adjusted for inflation. Since income is an importance feature, decrease in low-income group and increase in high-income group explains the reduction in number of DAC communities between 2013 and 2017. However, these correlations doesn't imply causation and a deeper analysis is required to identify true causes. Once identified, those causal links would assist the stakeholders and the decision makers in designing equitable and inclusive policies.

Conclusion

Designing inclusive and equitable strategies not just requires a good understanding of current demographics, but also a deep dive into the transformations that happened in those demographics over years. In this paper, we used AutoML to train several machine learning (ML) models on LODES and ACS data to classify the DAC status at the census tracts level and used the best trained model to classify DAC status between 2013-2017. Our analysis indicates that the Gradient Boosting Machine on features related to employment and income is the most accurate model with no bias towards any community. When used on historical data, we noticed a decline in number of disadvantaged communities between 2013 and 2017. The decline seems to be correlated with reduction in low-income groups and increase in high-income groups, some of the most important features of the trained models. However, one must note here that these correlations doesn't imply causation and further analysis is required to identify the true causes.

Acknowledgments

This research was supported by the Agile Initiative, a multidisciplinary Pacific Northwest National Laboratory (PNNL) initiative. PNNL is operated by Battelle Memorial Institute under Contract DE- AC06-76RL01830.

References

Brown, A.; and Lloyd-Jones, T. 2014. Spatial planning, access and infrastructure. In *Urban Livelihoods*, 211–227. Routledge.

Bureau, U. C. 2022a. *American Community Survey 5-Years Estimates*.

Bureau, U. S. C. 2022b. *Longitudinal Employer-Household Dynamics (LEHD) Survey*. LODES Data.

Dillahunt, T. R.; and Malone, A. R. 2015. The promise of the sharing economy among disadvantaged communities. In *proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2285–2294.

DOE. 2022. *Justice40 DAC Data*. Justice40 DAC Data.

Gedeon, T. D. 1997. Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, 8(02): 209–218.

H2O.ai. 2021. *H2O AutoML*. H2O version 3.32.1.2.

LeDell, E.; and Poirier, S. 2020. H2O AutoML: Scalable Automatic Machine Learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*.

Leyshon, A.; and Thrift, N. 1994. Access to financial services and financial infrastructure withdrawal: problems and policies. *Area*, 268–275.

Mandarano, L.; and Meenar, M. 2017. Equitable distribution of green stormwater infrastructure: A capacity-based framework for implementation in disadvantaged communities. *Local Environment*, 22(11): 1338–1357.

Vidal, A. C. 1995. Reintegrating disadvantaged communities into the fabric of urban life: The role of community development. *Housing Policy Debate*, 6(1): 169–230.

Wiesel, I.; Liu, F.; and Buckle, C. 2018. Locational disadvantage and the spatial distribution of government expenditure on urban infrastructure and services in metropolitan Sydney (1988–2015). *Geographical Research*, 56(3): 285–297.

Data Description

LODES Data: Published by the U.S. Census Bureau, the LODES data (Bureau 2022b) primarily captures the employment statistics at the block-level by demographics and industries. It is organized into three groups: (1) OD – Origin-Destination data associated with transition of employed population between home and work census blocks, (2) RAC – Residence Area Characteristics data by home census block i.e. distribution of people that live there, and (3) WAC – Workplace Area Characteristics data by work census block i.e. distribution of people that work there. Table 2 shows all the demographics features and their corresponding categories available in the LODES data. RAC and WAC columns indicate the availability of a feature in either dataset.

Parameter	Categories	RAC	WAC
Age (in Years)	≤29; 30-54; 55≥	✓	✓
Income (in $/Month)	≤1250; 1250-3333; 3333≥	✓	✓
Industry	20 categories[1]	✓	✓
Race	White; African American; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; Two or More Race Groups	✓	✓
Ethnicity	Not Hispanic or Latino; Hispanic or Latino	✓	✓
Education	Less than high school; High school or equivalent, no college; Some college or associate degree; Bachelor's degree or advanced degree	✓	✓
Gender	Male; Female	✓	✓
Firm Age (in Years)	0-1; 2-3; 4-5; 6-10; 11+		✓
Firm Size	0-19; 20-49; 50-249; 250-499; 500+		✓

Table 2: LEHD Origin-Destination Employment Statistics (LODES) Data

Parameter	Categories

Household Income (Annual)	≤10000;
	10000-14999;
	15000-19999;
	20000-24999;
	25000-29999;
	30000-34999;
	35000-39999;
	40000-44999;
	45000-49999;
	50000-54999;
	55000-59999;
	60000-74999;
	75000-99999;
	100000-124999;
	125000-149999;
	150000-199999;
	≥200000

Table 3: American Community Survey (ACS)

ACS Data: The American Community Survey (ACS) (Bureau 2022a) is a demographics survey program conducted by the U.S. Census Bureau every year and covers a broad range of topics about social, economic, demographic, and housing characteristics of the U.S. population. The 5Y estimates summarize sample data collected from last five years at the block-group level. Over 1Y estimates, 5Y estimates provide increased statistical reliability of the data for less populated areas and small population subgroups. For the purpose of this study, we only used 16-bin household income, adjusted for inflation, at the block-group level from the ACS data (see Table 3).

DAC Data: DAC (DOE 2022) is the U.S. Department of Energy's working definition of disadvantaged communities as pertaining to EO 14008, or the Justice40 Initiative. The DAC data includes thirty-six (36) burden indicators collected at the census tract level and an indicator identifying each census tract as disadvantaged or not disadvantaged. The 36 indicators are taken from various data sources including American Community Survey (ACS), Longitudinal Employer-Household Dynamics (LEHD) Survey, Environmental Justice Screening Tool (EJScreen), among others. For this study, we are using 2022c version of the DAC data. For detailed information, please refer to Justice40 DAC data documentation (DOE 2022).

Model Hyperparameters

Table 4 provides details about the tuned parameters of the best models from H2O.AI.

Table 4: Best Hyperparameters from Grid Search

	v1a:LODES(R)	v1b:LODES(W)	v1c:LODES(R+W)	v2a:LI(R)+ACS	v2b:LI(R+W)+ACS
Gradient Boosting Machine (GBM)					
col_sample_rate	0.7	0.8	0.8	1.0	0.4
col_sample_rate_per_tree	1.0	0.8	0.8	1.0	0.7
learn_rate	0.1	0.1	0.1	0.1	0.1
max_depth	4	15	7	17	6
min_rows	5.0	100.0	10.0	15.0	10.0
min_split_improvement	0.00001	0.00001	0.00001	0.00001	0.00001
ntrees	35	41	37	45	50
sample_rate	0.9	0.8	0.8	0.9	0.5
XGBoost					
booster	gbtree	gbtree	gbtree	gbtree	gbtree
col_sample_rate	0.8	0.8	0.8	0.8	0.6
col_sample_rate_per_tree	0.8	0.8	0.8	0.7	0.8
max_depth	5	5	10	9	9
min_rows	3.0	3.0	5.0	5.0	10.0
ntrees	34	33	35	42	40
reg_alpha	0.0	0.0	0.0	1.0	0.001
reg_lambda	1.0	1.0	1.0	1.0	0.01
sample_rate	0.8	0.8	0.6	0.6	0.6
Generalized Linear Model (GLM)					
alpha	[0.0]	[0.0]	[0.0]	[0.0]	[0.0]
Deep Learning					
epsilon	0.0	0.0	0.000001	0.000001	0.0
hidden	[100, 100]	[10, 10, 10]	[50, 50, 50]	[50]	[100]
hidden_dropout_ratios	[0.1, 0.1]	None	[0.4, 0.4, 0.4]	[0.4]	[0.1]
input_dropout_ratio	0.15	0.0	0.2	0.2	0.15
rho	0.9	0.99	0.95	0.95	0.9
Distributed Random Forest (DRF)					
balance_classes	False	False	False	False	False
ntrees	34	41	40	33	33
max_depth	20	20	20	20	20
col_sample_rate_change_per_level	1.0	1.0	1.0	1.0	1.0
col_sample_rate_per_tree	1.0	1.0	1.0	1.0	1.0
min_split_improvement	0.00001	0.00001	0.00001	0.00001	0.00001
Extreme Random Forest (XRT)					
balance_classes	False	False	False	False	False
ntrees	43	45	35	43	41
max_depth	20	20	20	20	20
col_sample_rate_change_per_level	1.0	1.0	1.0	1.0	1.0
col_sample_rate_per_tree	1.0	1.0	1.0	1.0	1.0
min_split_improvement	0.00001	0.00001	0.00001	0.00001	0.00001