US Census

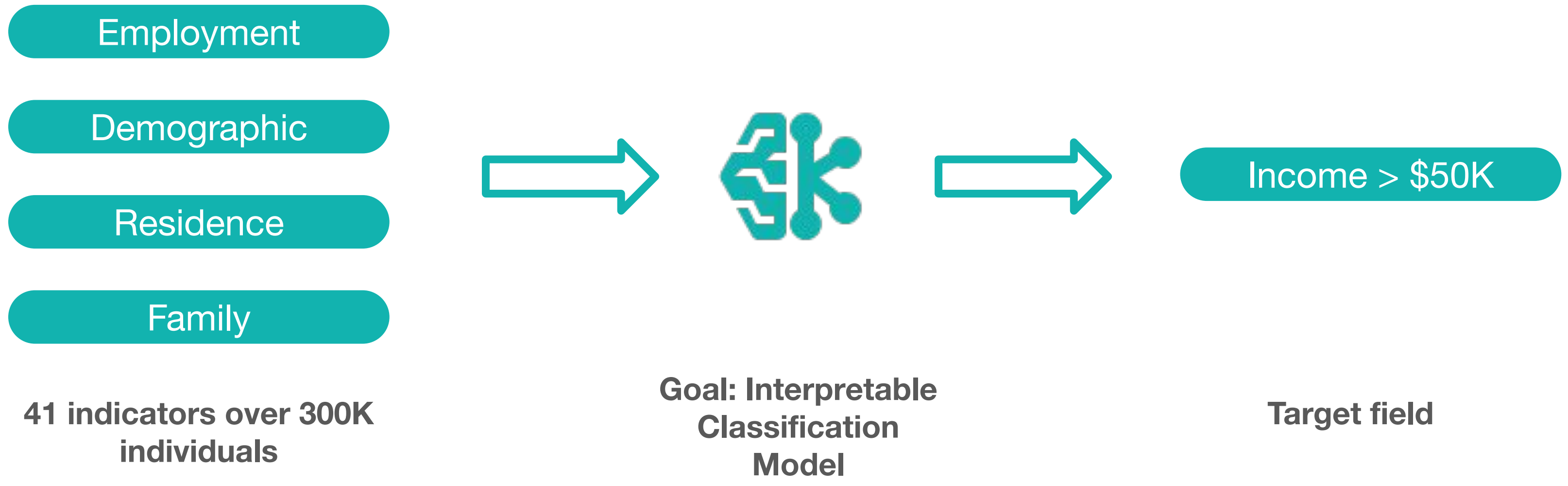
Income Prediction

Daniel Soukup

2025.11.03

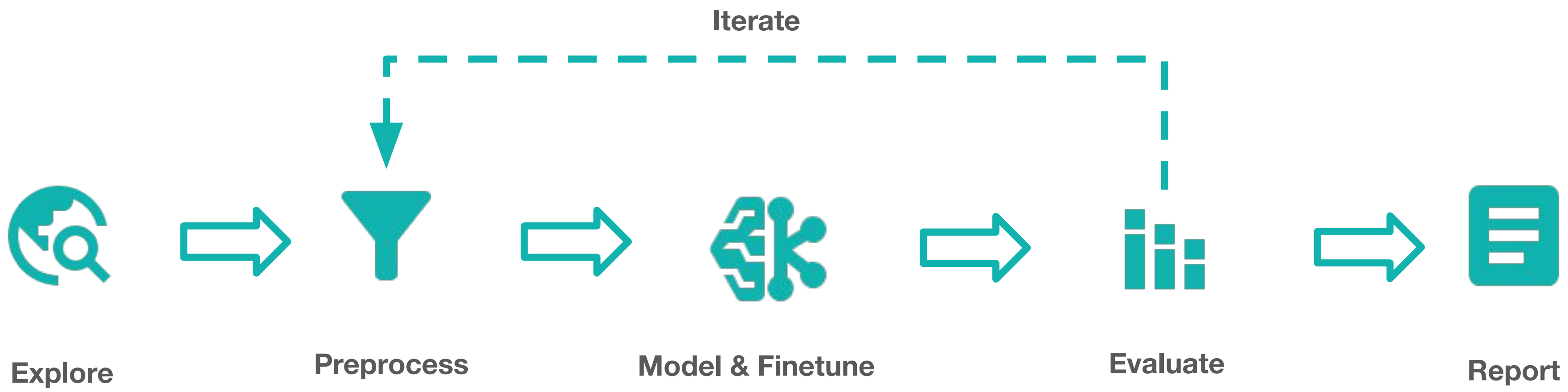
The Challenge

Our goal is finding identifying characteristics that distinguish between a person making more or less than $50,000 per year based on a sample dataset from the US Census archive.


Employment
Demographic
Residence
Family
41 indicators over 300K individuals
Goal: Interpretable Classification Model
Income > $50K
Target field

Solution Framework

In addressing this problem, we followed an iterative process. Guided by our initial data exploration, the data was preprocessed and used to create competing classification models. Informed by subsequent evaluation, we refined the preprocessing and modeling steps to achieve higher performance and in-depth insights.



Data Exploration

Our first step was to thoroughly review the data set, gaining statistical insights to the distribution of each field as well as uncovering any data quality concerns. This step informs the following preprocessing and feature engineering before modeling.

Data Quality

Key concerns uncovered during EDA:

- **Duplicate** rows and **missing** values
- Highly **skewed distributions** with **extreme outliers**
- **High cardinality** categorical columns

8% high income

The **high class imbalance** of the target will affect model tuning and evaluation.

Statistical Indicators

Confirming intuition, we found that **education & occupation** fields have significant relationship with having high income.

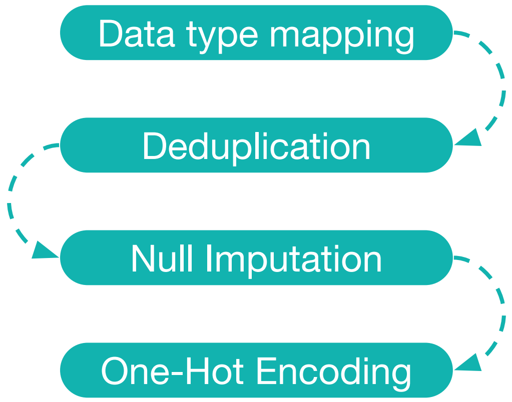
Preprocessing

This step in our solution ensures that the raw data is best prepared for the modeling step. We addressed the data quality issues and encoded the categorical features adequately for our selected tree-ensemble model.

Data Pipelines

We created a **single, modular data processing pipeline** which captures multiple, custom data transformation steps. This approach ensures

- **consistency**: the same operation applied to train & test data sets,
- **no data leakage:** information from the test set is not used to create the pipeline, and
- **reduces tech debt** in the long run.

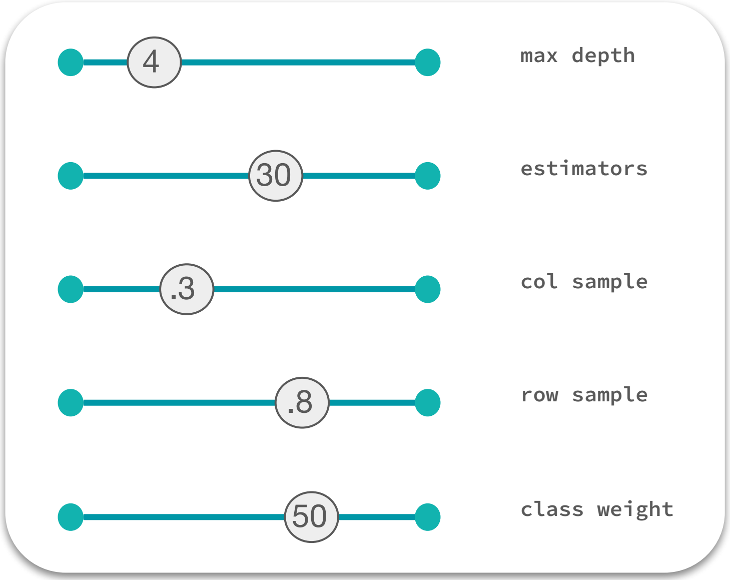


Modeling & Fine Tuning

After the preprocessing of a mix of numeric and categorical features, we optimized an **XGBoost binary classification** model fine-tuning over a number of hyperparameters over 40 model variations. This model type, which successively combines simple decision trees to form an ensemble, is know to handle high dimensional, complex data sets well, while being **computationally efficient** and **robust to data skew** and outliers.

Tuning Objectives

- **Cross validation** to reduce variance of loss estimates.
- Increase **boosting rounds** to create complex models but limit tree depth and lower column/row sampling rates to reduce overfitting.
- Optimize area **under the precision-recall curve** to account for class imbalance.



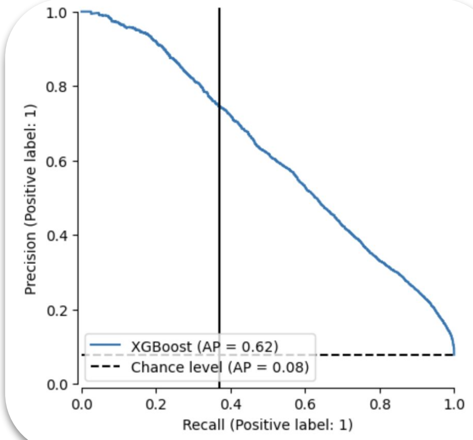
4	max depth
30	estimators
.3	col sample
.8	row sample
50	class weight

Evaluation & Interpretation

Analysing feature importance highlighted **sex**, **education** and **employment** indicators as most significant, the former reaffirming a well-know **bias** in this historical dataset. Our predictions correctly **identified 35% of high income** earners (recall) in the test set while overall the high income **predictions** being **77% of the time correct** (precision).

.35 Recall

.77 Precision



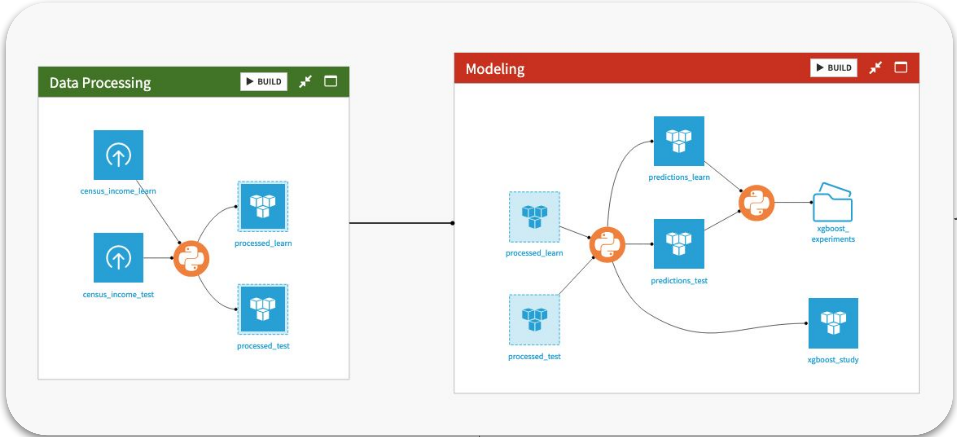
Trade-offs

Our model significantly outperformed naive baselines.

The **recall/precision trade-off** can be further tuned by adjusting the predicted probability cutoff threshold.

Implementation on Dataiku DSS

The complete solution framework was implemented on Dataiku DSS Cloud using Python recipes ran on a custom project environment with 2-CPU cores and 16Gb memory.



The recipe notebooks are synced with Github.

Data stored on DSS managed AWS S3 storage.

Next Steps

Our next steps will be ultimately informed by the key (business) drivers that initiated the project e.g. to understand the effect of individual predictors vs to optimize predictive power.

Data Enrichment

Improve the input data quality & quantity:

- **Balance the classes** using up/down sampling or synthetic data generation.
- **Feature engineering &** alternative encodings.
- Adjust **outliers** and **normalize.**

MLOps & Evaluation

More in-depth **experiment logging** and refined **analysis of the model errors** could reveal model limitations and inform future iterations.

Feature importances and effects on individual samples can be better understood by model agnostic methods such as SHAP or LIME.

Alternative Models

A more thorough search through a **larger hyperparameter space** can further improve model performance.

Alternative model types can be easily tested, such as CatBoost which are built for high-dimensional categorical data.

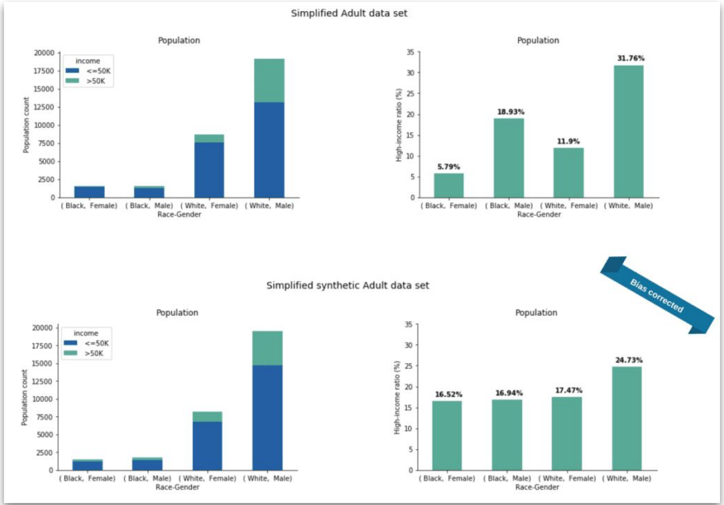
Q & A

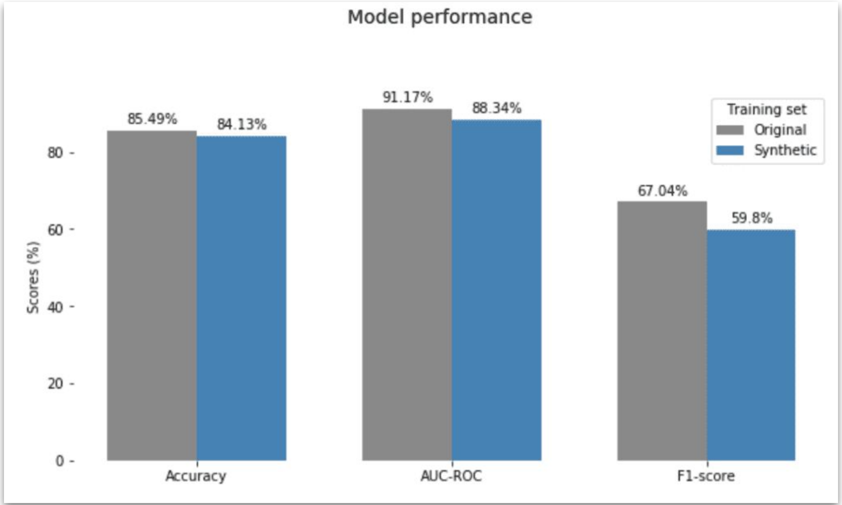
Project: Github

Contact: LinkedIn

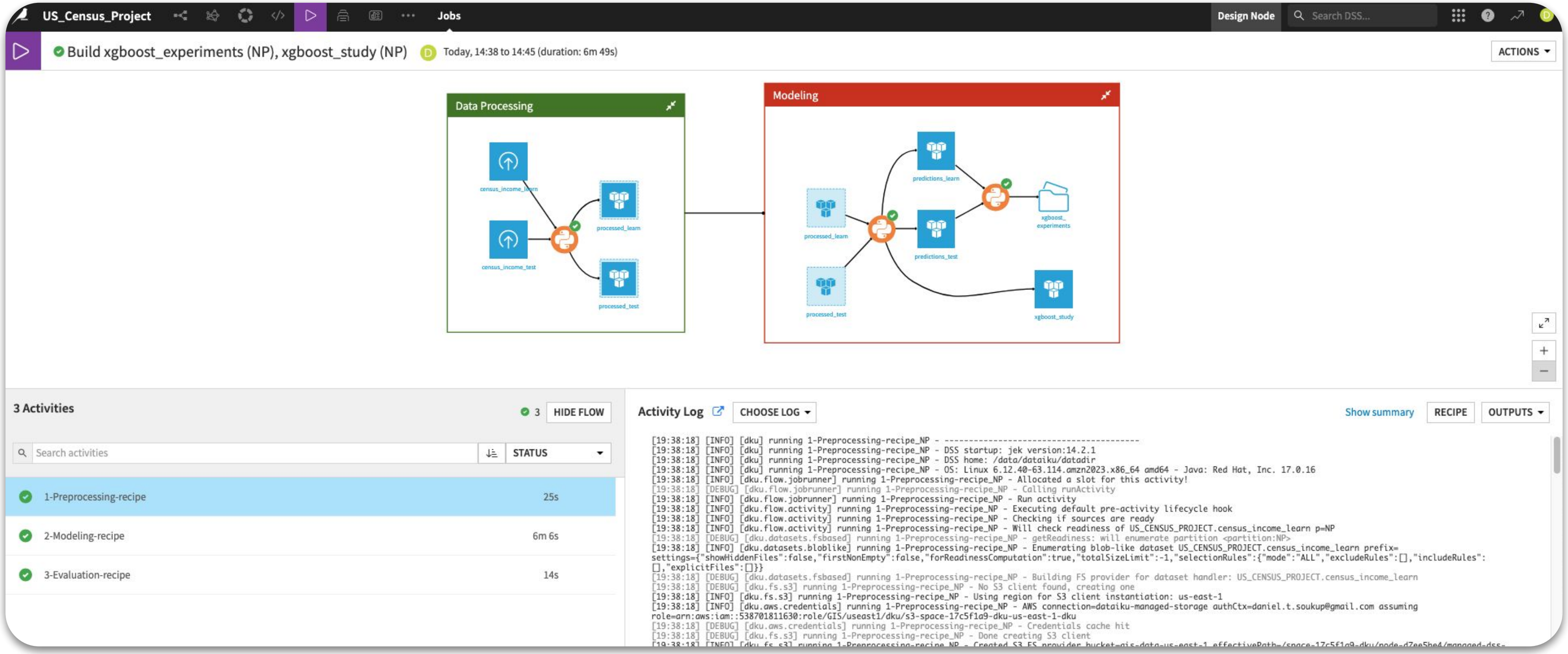
Appendix: Bias Mitigation

Previously, I've implemented privacy and fairness risk mitigation with bias-corrected synthetic data on the Adult Census dataset: this allows for privacy-preserving data sharing and also aids the fair treatment of customers (data subjects) in downstream analysis and machine learning tasks. See more here and in our paper accepted at an ICLR 2021 workshop.





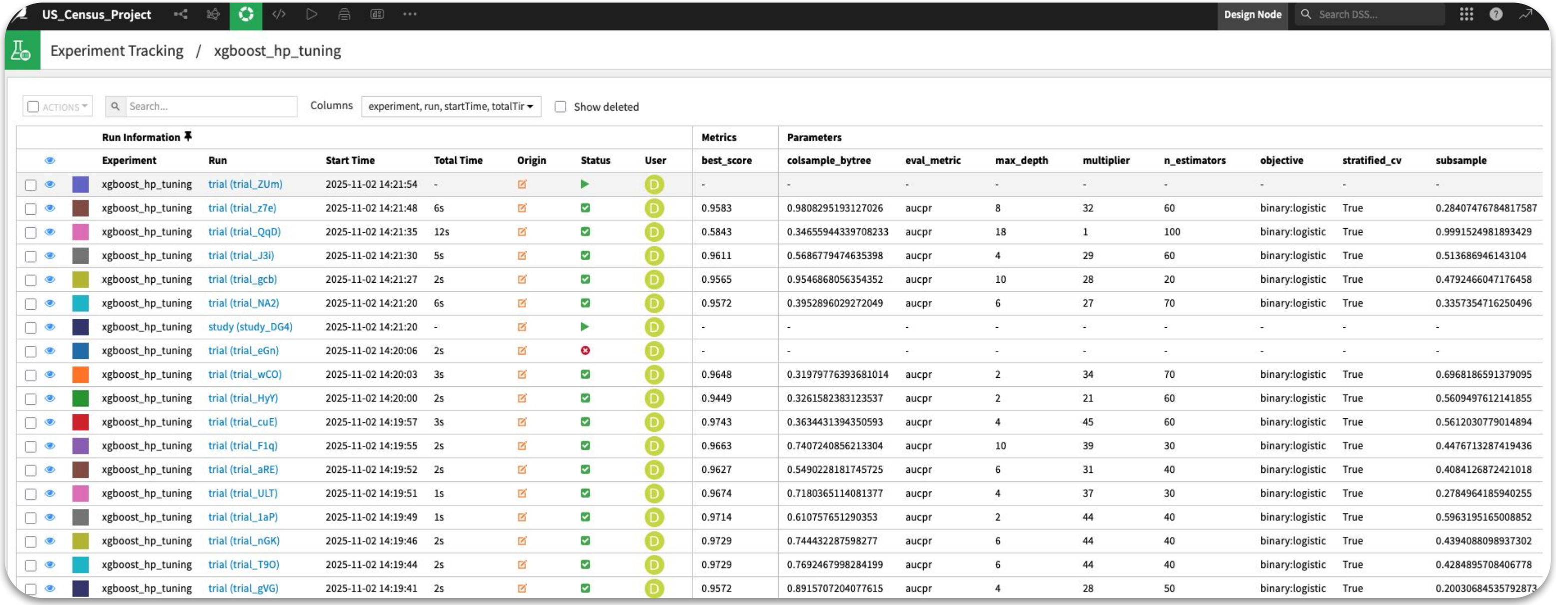
Appendix: End-to-End Flow Build



1-Preprocessing-recipe	25s
2-Modeling-recipe	6m 6s
3-Evaluation-recipe	14s

```
[19:38:18] [INFO] [dku] running 1-Preprocessing-recipe_NP - ----------------------------------------
[19:38:18] [INFO] [dku] running 1-Preprocessing-recipe_NP - DSS startup: jek version:14.2.1
[19:38:18] [INFO] [dku] running 1-Preprocessing-recipe_NP - DSS home: /data/dataiku/datadir
[19:38:18] [INFO] [dku] running 1-Preprocessing-recipe_NP - OS: Linux 6.12.40-63.114.amzn2023.x86_64 amd64 - Java: Red Hat, Inc. 17.0.16
[19:38:18] [INFO] [dku.flow.jobrunner] running 1-Preprocessing-recipe_NP - Allocated a slot for this activity!
[19:38:18] [DEBUG] [dku.flow.jobrunner] running 1-Preprocessing-recipe_NP - Calling runActivity
[19:38:18] [INFO] [dku.flow.jobrunner] running 1-Preprocessing-recipe_NP - Run activity
[19:38:18] [INFO] [dku.flow.activity] running 1-Preprocessing-recipe_NP - Executing default pre-activity lifecycle hook
[19:38:18] [INFO] [dku.flow.activity] running 1-Preprocessing-recipe_NP - Checking if sources are ready
[19:38:18] [INFO] [dku.flow.activity] running 1-Preprocessing-recipe_NP - Will check readiness of US_CENSUS_PROJECT.census_income_learn p=NP
[19:38:18] [DEBUG] [dku.datasets.fsbased] running 1-Preprocessing-recipe_NP - getReadiness: will enumerate partition <partition:NP>
[19:38:18] [INFO] [dku.datasets.bloblike] running 1-Preprocessing-recipe_NP - Enumerating blob-like dataset US_CENSUS_PROJECT.census_income_learn prefix=
settings={"showHiddenFiles":false,"firstNonEmpty":false,"forReadinessComputation":true,"totalSizeLimit":-1,"selectionRules":{"mode":"ALL","excludeRules":[],"includeRules":
[],"explicitFiles":[]}}
[19:38:18] [DEBUG] [dku.datasets.fsbased] running 1-Preprocessing-recipe_NP - Building FS provider for dataset handler: US_CENSUS_PROJECT.census_income_learn
[19:38:18] [DEBUG] [dku.fs.s3] running 1-Preprocessing-recipe_NP - No S3 client found, creating one
[19:38:18] [INFO] [dku.fs.s3] running 1-Preprocessing-recipe_NP - Using region for S3 client instantiation: us-east-1
[19:38:18] [INFO] [dku.aws.credentials] running 1-Preprocessing-recipe_NP - AWS connection=dataiku-managed-storage authCtx=daniel.t.soukup@gmail.com assuming
role=arn:aws:iam::538701811630:role/GIS/useast1/dku/s3-space-17c5f1a9-dku-us-east-1-dku
[19:38:18] [DEBUG] [dku.aws.credentials] running 1-Preprocessing-recipe_NP - Credentials cache hit
[19:38:18] [DEBUG] [dku.fs.s3] running 1-Preprocessing-recipe_NP - Done creating S3 client
[19:38:18] [INFO] [dku.fs.s3] running 1-Preprocessing-recipe_NP - Created S3 FS provider bucket=gis-data-us-east-1 effectivePath=/space-17c5f1a9-dku/node-d7ee5be4/managed-dss-
```

Appendix: Experiment Tracking



US_Census_Project

Design Node

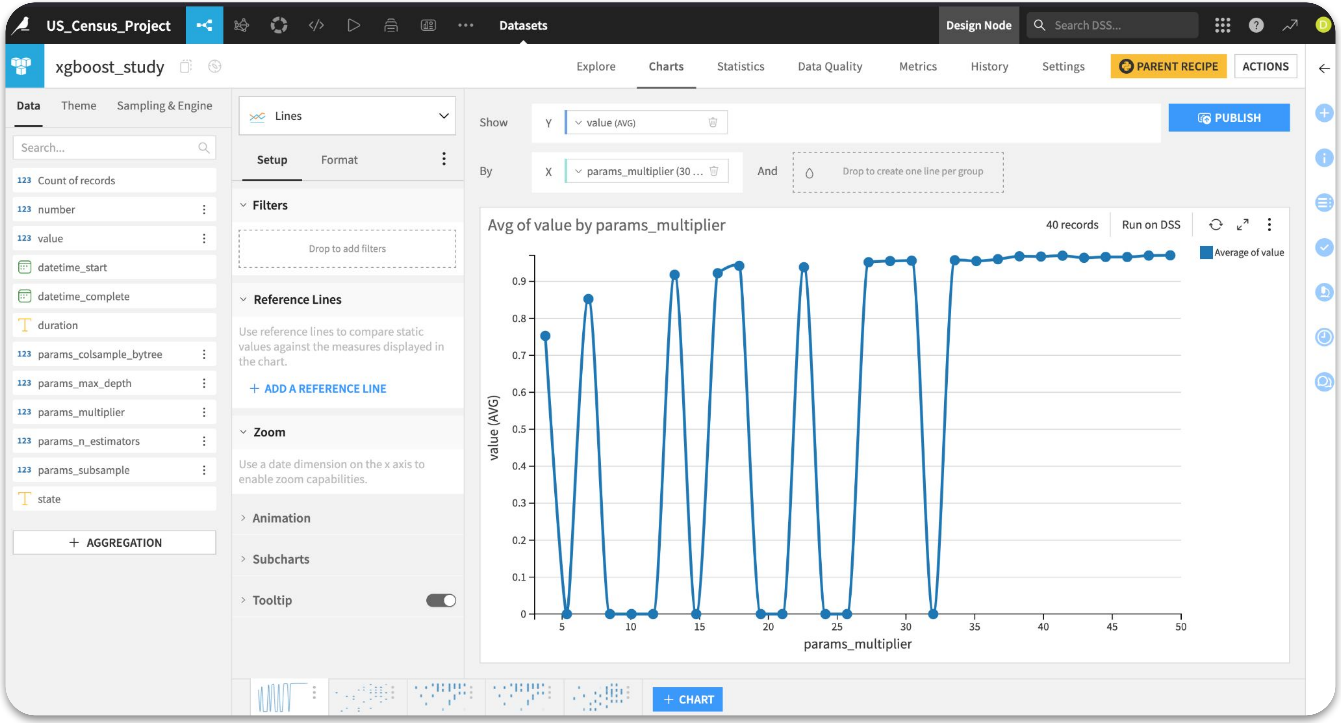
Experiment Tracking / xgboost_hp_tuning

Columns: experiment, run, startTime, totalTir

Show deleted

Run Information							Metrics	Parameters							
Experiment	Run	Start Time	Total Time	Origin	Status	User	best_score	colsample_bytree	eval_metric	max_depth	multiplier	n_estimators	objective	stratified_cv	subsample
xgboost_hp_tuning	trial (trial_ZUm)	2025-11-02 14:21:54	-	✎	▶	D	-	-	-	-	-	-	-	-	-
xgboost_hp_tuning	trial (trial_z7e)	2025-11-02 14:21:48	6s	✎	✓	D	0.9583	0.9808295193127026	aucpr	8	32	60	binary:logistic	True	0.28407476784817587
xgboost_hp_tuning	trial (trial_QqD)	2025-11-02 14:21:35	12s	✎	✓	D	0.5843	0.34655944339708233	aucpr	18	1	100	binary:logistic	True	0.9991524981893429
xgboost_hp_tuning	trial (trial_J3i)	2025-11-02 14:21:30	5s	✎	✓	D	0.9611	0.5686779474635398	aucpr	4	29	60	binary:logistic	True	0.513686946143104
xgboost_hp_tuning	trial (trial_gcb)	2025-11-02 14:21:27	2s	✎	✓	D	0.9565	0.9546868056354352	aucpr	10	28	20	binary:logistic	True	0.4792466047176458
xgboost_hp_tuning	trial (trial_NA2)	2025-11-02 14:21:20	6s	✎	✓	D	0.9572	0.3952896029272049	aucpr	6	27	70	binary:logistic	True	0.3357354716250496
xgboost_hp_tuning	study (study_DG4)	2025-11-02 14:21:20	-	✎	▶	D	-	-	-	-	-	-	-	-	-
xgboost_hp_tuning	trial (trial_eGn)	2025-11-02 14:20:06	2s	✎	✗	D	-	-	-	-	-	-	-	-	-
xgboost_hp_tuning	trial (trial_wCO)	2025-11-02 14:20:03	3s	✎	✓	D	0.9648	0.31979776393681014	aucpr	2	34	70	binary:logistic	True	0.6968186591379095
xgboost_hp_tuning	trial (trial_HyY)	2025-11-02 14:20:00	2s	✎	✓	D	0.9449	0.3261582383123537	aucpr	2	21	60	binary:logistic	True	0.5609497612141855
xgboost_hp_tuning	trial (trial_cuE)	2025-11-02 14:19:57	3s	✎	✓	D	0.9743	0.3634431394350593	aucpr	4	45	60	binary:logistic	True	0.5612030779014894
xgboost_hp_tuning	trial (trial_F1q)	2025-11-02 14:19:55	2s	✎	✓	D	0.9663	0.7407240856213304	aucpr	10	39	30	binary:logistic	True	0.4476713287419436
xgboost_hp_tuning	trial (trial_aRE)	2025-11-02 14:19:52	2s	✎	✓	D	0.9627	0.5490228181745725	aucpr	6	31	40	binary:logistic	True	0.4084126872421018
xgboost_hp_tuning	trial (trial_ULT)	2025-11-02 14:19:51	1s	✎	✓	D	0.9674	0.7180365114081377	aucpr	4	37	30	binary:logistic	True	0.2784964185940255
xgboost_hp_tuning	trial (trial_1aP)	2025-11-02 14:19:49	1s	✎	✓	D	0.9714	0.610757651290353	aucpr	2	44	40	binary:logistic	True	0.5963195165008852
xgboost_hp_tuning	trial (trial_nGK)	2025-11-02 14:19:46	2s	✎	✓	D	0.9729	0.744432287598277	aucpr	6	44	40	binary:logistic	True	0.4394088098937302
xgboost_hp_tuning	trial (trial_T9O)	2025-11-02 14:19:44	2s	✎	✓	D	0.9729	0.7692467998284199	aucpr	6	44	40	binary:logistic	True	0.4284895708406778
xgboost_hp_tuning	trial (trial_gVG)	2025-11-02 14:19:41	2s	✎	✓	D	0.9572	0.8915707204077615	aucpr	4	28	50	binary:logistic	True	0.20030684535792873

Appendix: HP Tuning Analysis



123 Count of records
123 number
123 value
datetime_start
datetime_complete
T duration
123 params_colsample_bytree
123 params_max_depth
123 params_multiplier
123 params_n_estimators
123 params_subsample
T state

Show Y value (AVG)

By X params_multiplier (30 ... And

Avg of value by params_multiplier