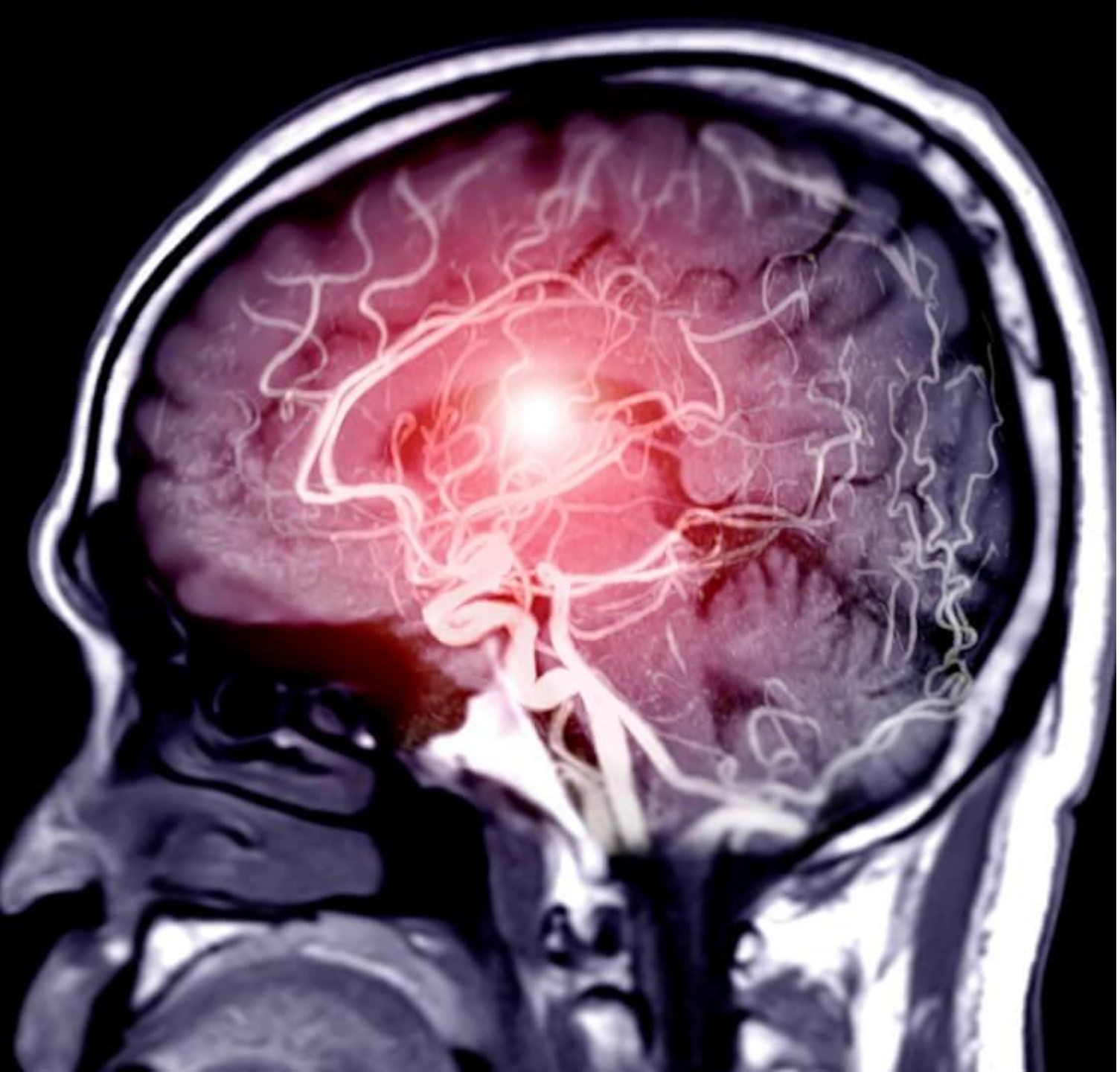
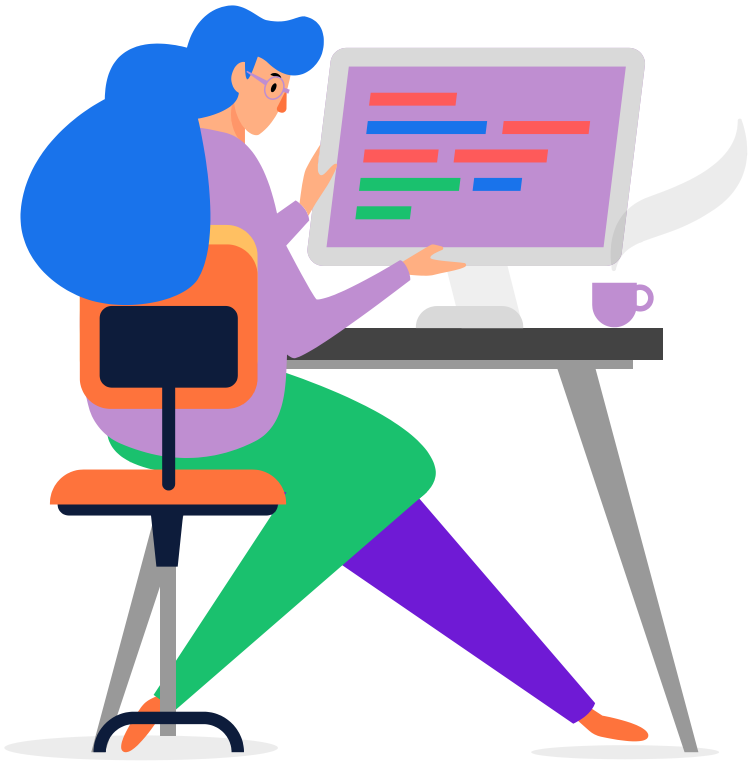


Stroke Prediction Pipeline (Azure)

Feby Hadayani
Monica Ko
Sirinda Leesuravanich
Sarthak Dhanke

Content



What risk are we trying to detect, and why ?

- **Problem:** Stroke is a critical medical event requiring timely intervention. Effective **risk stratification and early triage** are essential to save lives and improve patient outcomes.
- **Business Question:** Can we leverage this routine, non-invasive patient data to accurately **triage (prioritize)** individuals for immediate or early comprehensive screening?
- **Objective**: The primary objective is to develop and evaluate a suite of machine learning models to **accurately rank and prioritize** patients based on their predicted risk of stroke.

Dataset is overall clean; BMI is the only variable with missing values

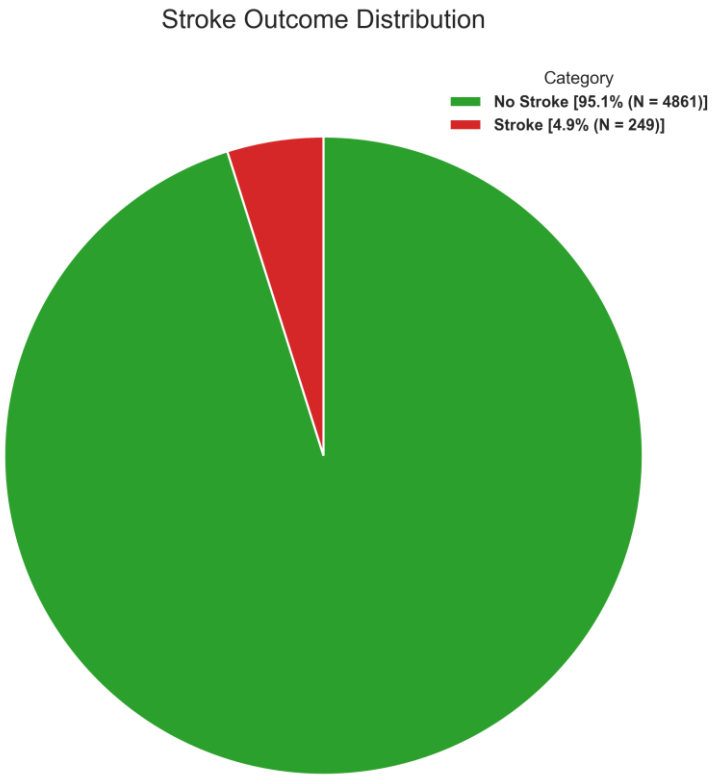
Column	MissingCount	MissingPercent
bmi	201	3.93
age	0	0.00
gender	0	0.00
hypertension	0	0.00
heart_disease	0	0.00
work_type	0	0.00
ever_married	0	0.00
residence_type	0	0.00
avg_glucose_level	0	0.00
smoking_status	0	0.00
stroke	0	0.00

Type	Count
Numeric	6
Categorical	5

- Only BMI has missing values (~4%); all other fields complete
- 6 numeric, 5 categorical predictors
- ~5000 unique individuals

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Stroke events are rare; the model must operate under severe class imbalance



- Total records: 5110 patients
- Only ~1 in 20 patients had a stroke
- All subsequent modeling and evaluation must account for this class imbalance (e.g., stratified splits, recall-focused metrics).

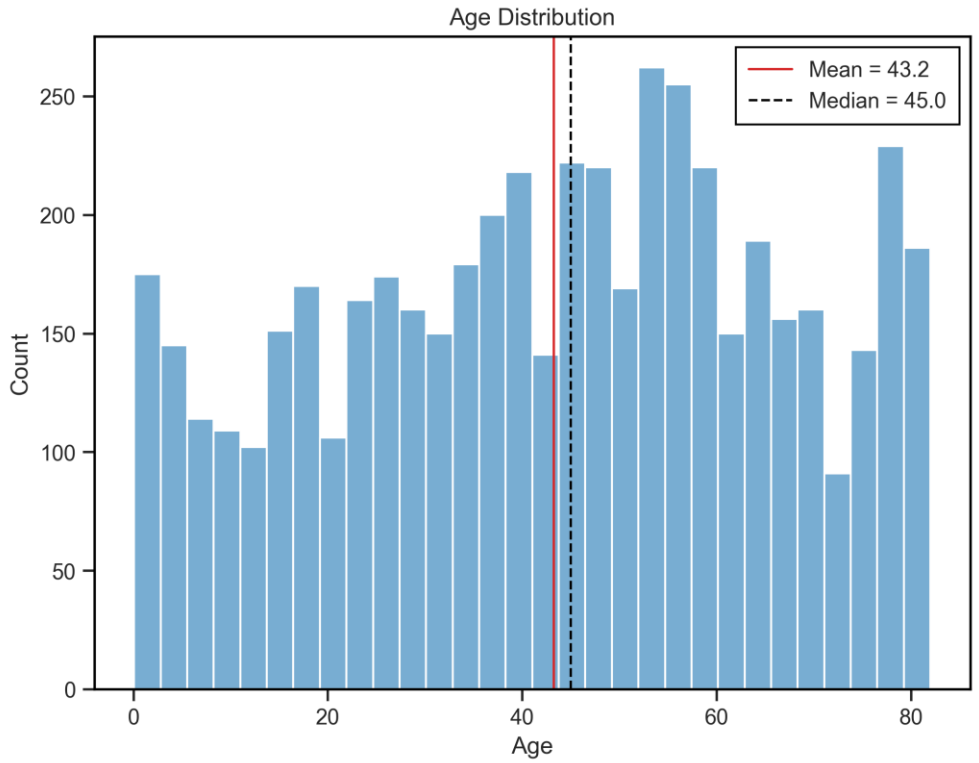
Clinical intuition suggests age, glucose, BMI, and comorbidities drive stroke risk

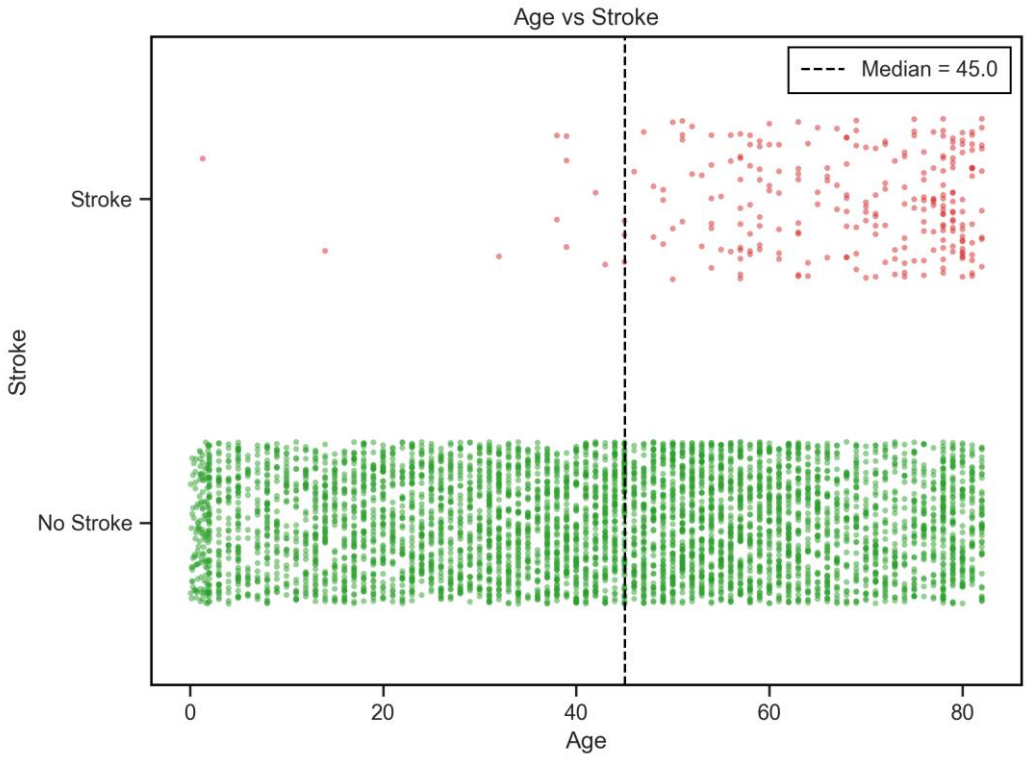
Variable	Type	Expected Relationship
age	Numeric	Older age increases stroke risk
avg_glucose_level	Numeric	Higher glucose increases stroke risk
bmi	Numeric	Higher BMI increases stroke risk
smoking_status	Categorical	Current smokers more likely to have stroke
hypertension	Categorical	Hypertension increases stroke risk
heart_disease	Categorical	Heart disease increases stroke risk

These expectations guide both EDA and feature engineering.

Stroke cases concentrate at older ages; age is the strongest single signal

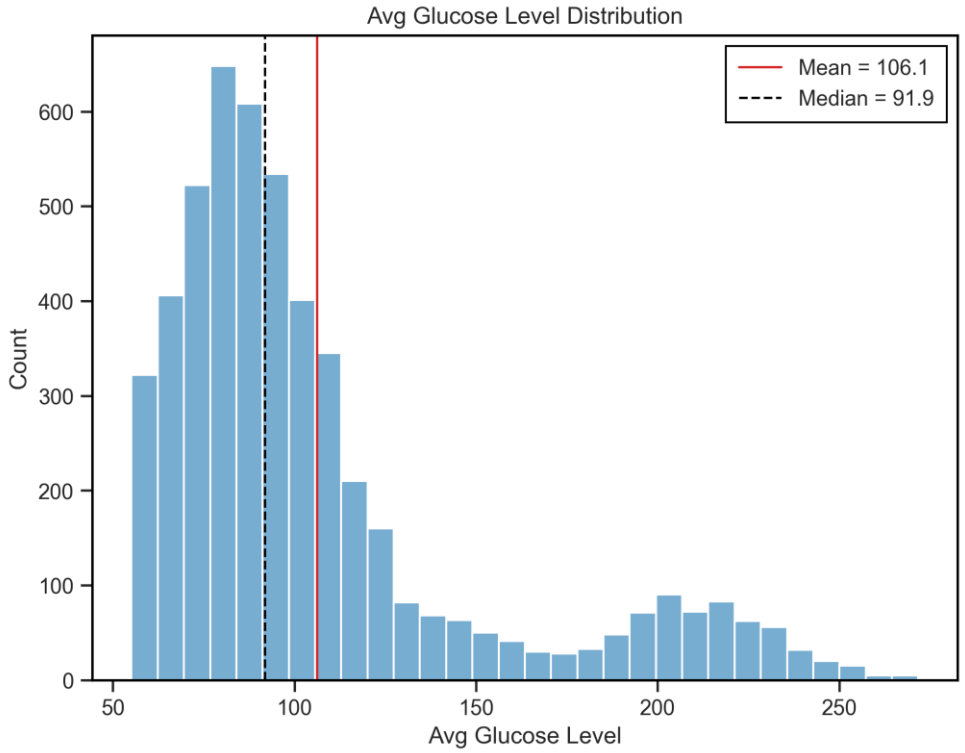
- Almost **no strokes under ~40**; risk rises steeply with age.
- Feature decision: create **age_group**, **age_over_60**, **age_over_80**

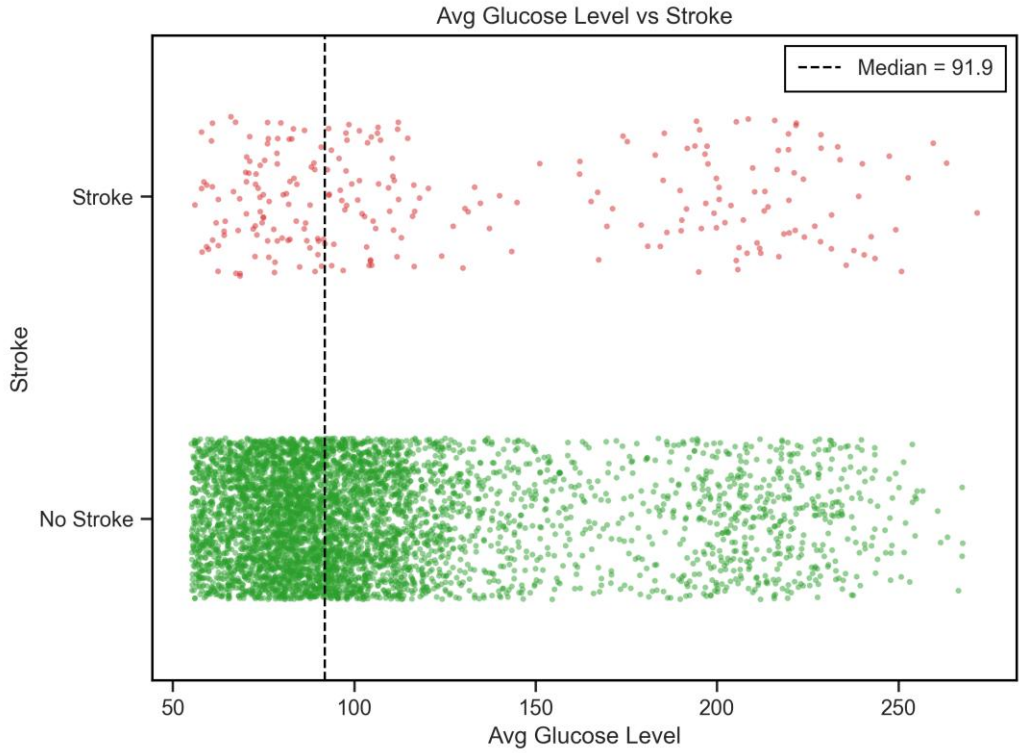




Higher average glucose levels shift stroke risk to the right tail

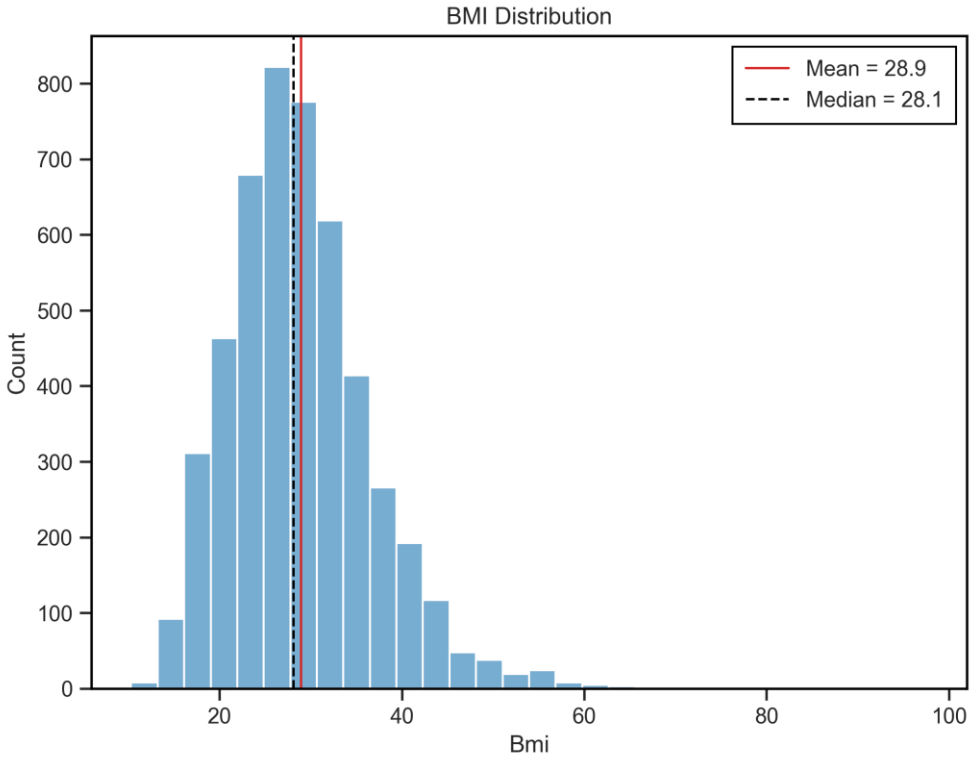
- Stroke patients cluster at higher glucose levels
- Feature decision: **clip extreme values (< 250)**; create **glucose_group**, **glucose_above_150**, **glucose_above_250**

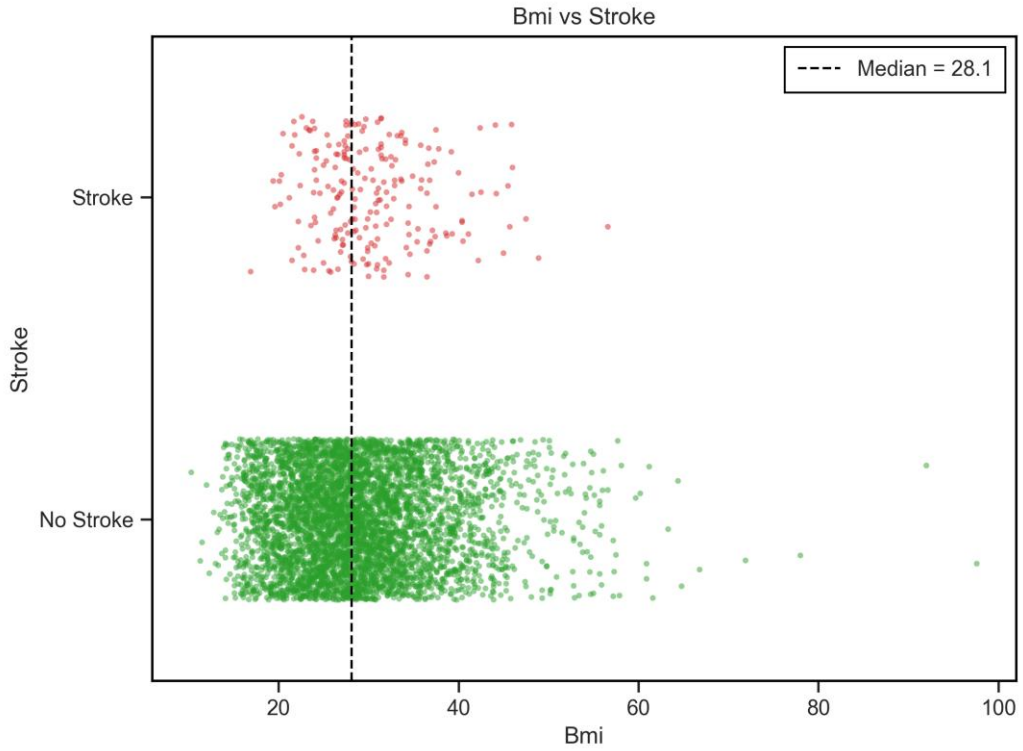




Higher BMI associates with stroke, but extreme values are likely data errors

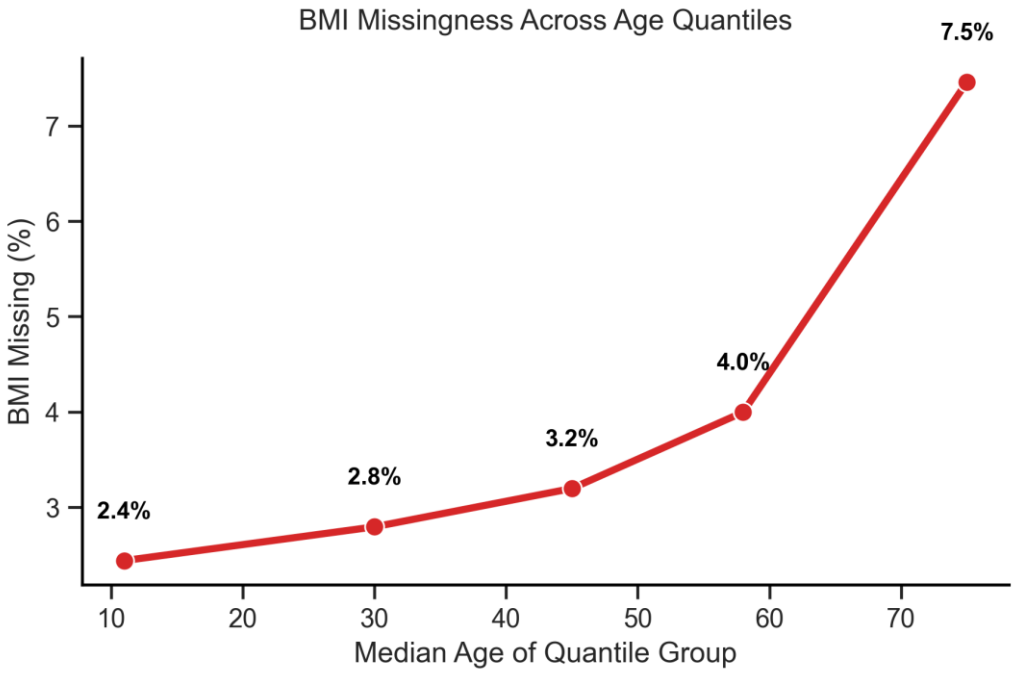
- Majority between 15-45; tail above ~60 is implausible
- Feature decision: **clip BMI to 10-60**; create **bmi_outlier_flag**

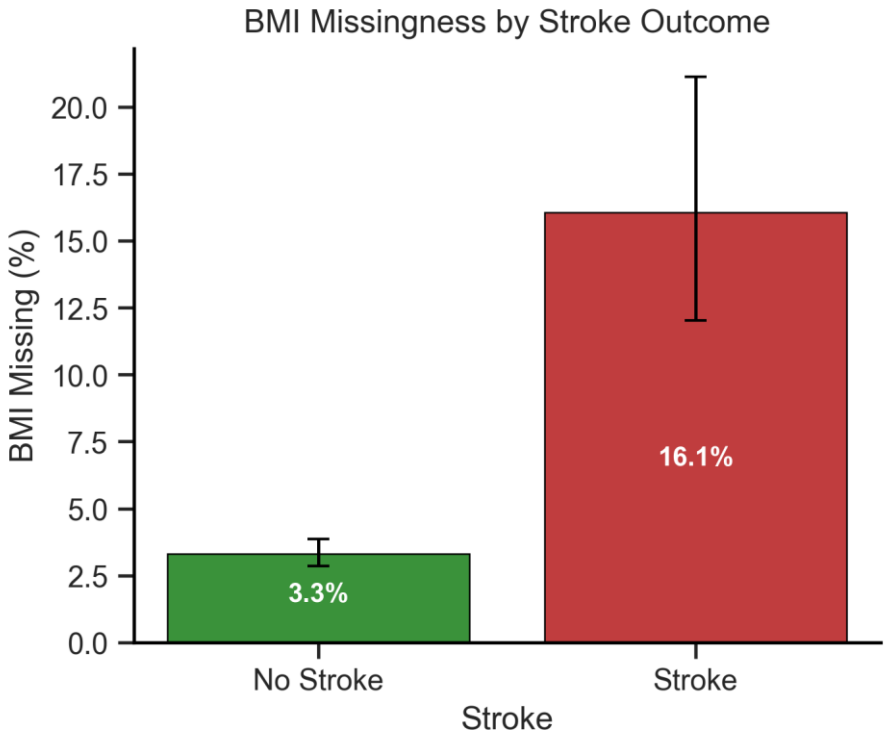




BMI is not missing at random; missingness increases with age and stroke

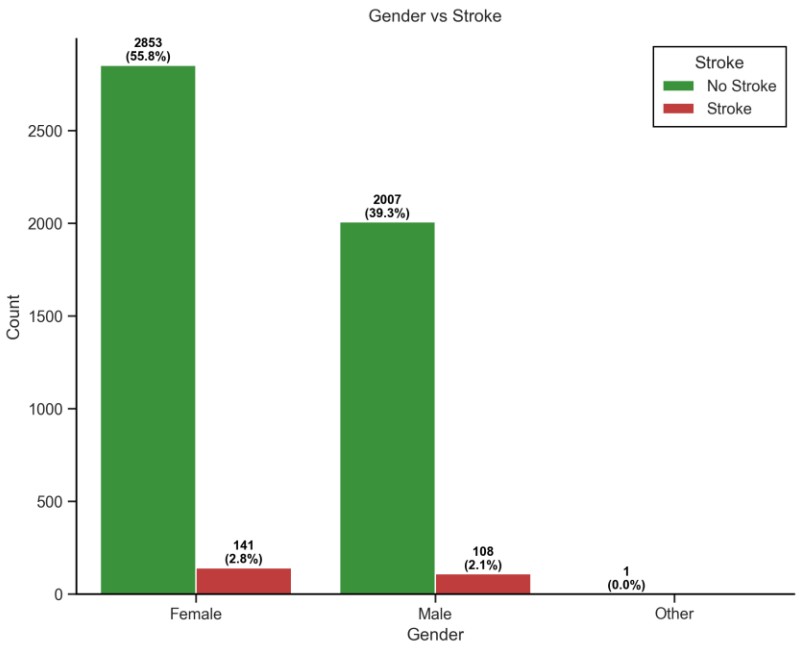
- Missingness BMI rises steadily with age and is higher among stroke cases
- Feature decision: keep rows, add **is_bmi_missing** flag and use **iterative imputation**.



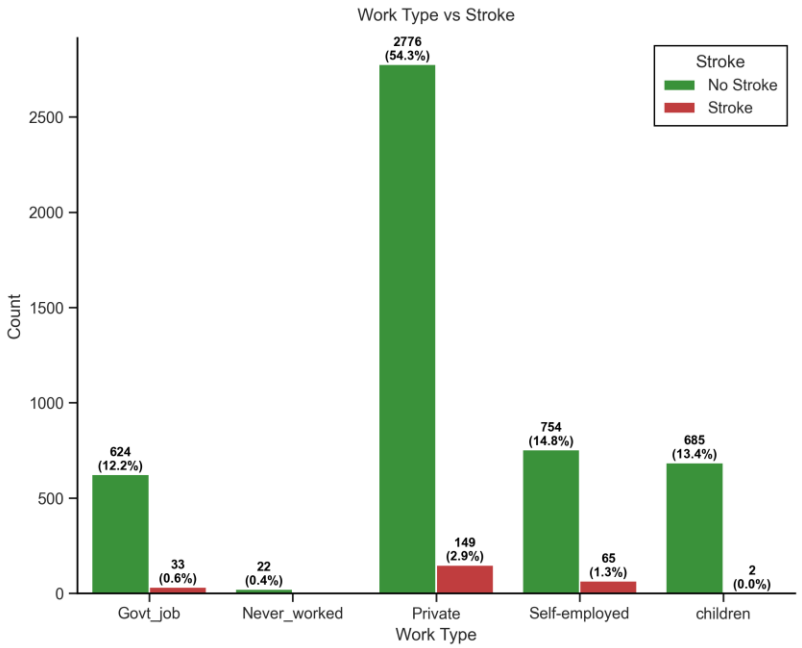


Most patients fall into a few dominant categories; rare levels are consolidated and “Unknown” is explicitly flagged

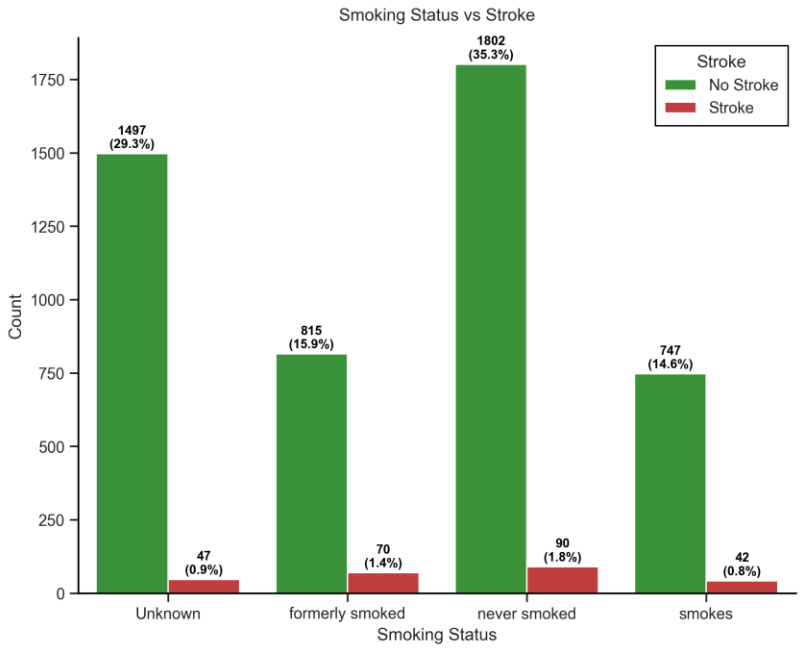
Gender distribution is balanced; 1 ‘Other’ case merged into Female to avoid a spurious level



Private dominates; ‘children’ and ‘never_worked’ have very few records and are collapsed into ‘Other’ to stabilize splits



All four smoking categories have enough samples; we keep them but add an indicator for ‘Unknown’ and for current smokers



Comorbidities and marital status meaningfully increase stroke risk

Variable	Category	Total	Stroke Cases	Stroke Rate (%)
hypertension	No	4612	183	4.0
	Yes	498	66	13.3
heart_disease	No	4834	202	4.2
	Yes	276	47	17.0
ever_married	No	1757	29	1.7
	Yes	3353	220	6.6
Residence_type	Rural	2514	114	4.5
	Urban	2596	135	5.2

- Stroke rate more than **triples** with hypertension (4.0% → 13.3%).
- Heart disease patients have the **highest stroke rate** in the dataset (~17%).
- Ever-married and urban patients show moderately higher stroke rates, so we keep these variables but do not engineer extra flags.

EDA decisions translate into four versioned datasets for modeling

V0: Raw → V1: Clipped → V2: Dropped-missing → V3: Fully engineered & imputed

Data Cleaning

- Outlier flagging: Unrealistic BMI, avg glucose level
- Imputation: Impute BMI using age, Impute numerical variables with median & categorical variables with most frequent

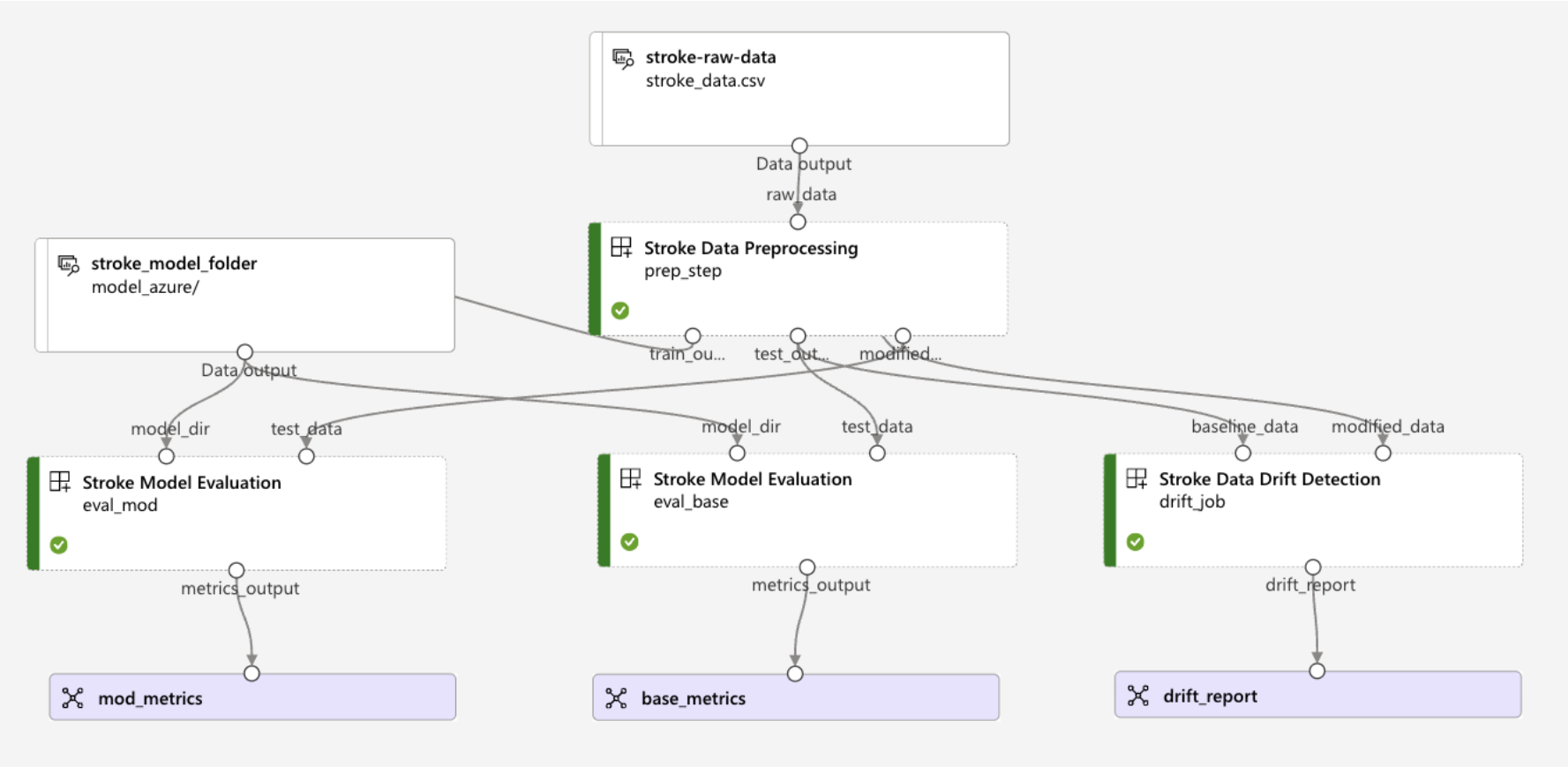
Key engineered features grouped by source

- **Age**: age_group, age_over_60, age_over_80.
- **Glucose**: glucose_group, glucose_above_150, glucose_above_250, glucose_outlier_flag.
- **BMI**: bmi_category, is_overweight, is_bmi_missing, bmi_outlier_flag.
- **Smoking**: is_smokes, is_smoking_unknown.
- **Categorical cleanup**: gender merge, work_type rare-level collapse.

Key Performance Indicator

- The project will be measured on its ability to maximize Recall (catching true stroke cases) while maintaining a strict clinical minimum of Precision (avoiding false alarms).
- If prevalence is 5%, a very high precision would require the model to filter out nearly all patients, resulting in unacceptably low recall.
- **Goal**: Maximize Recall (Sensitivity) subject to a minimum Precision of 10% (for Class 1: Stroke).
- **Evaluation Metrics**: Compare models based on AUCPR (Area Under the Precision-Recall Curve—the primary sorting metric) and AUROC (Area Under the Receiver Operating Characteristic curve).
- **Framing**: This is a risk triage tool, where a False Negative (missing a stroke case) has a higher cost than a False Positive.

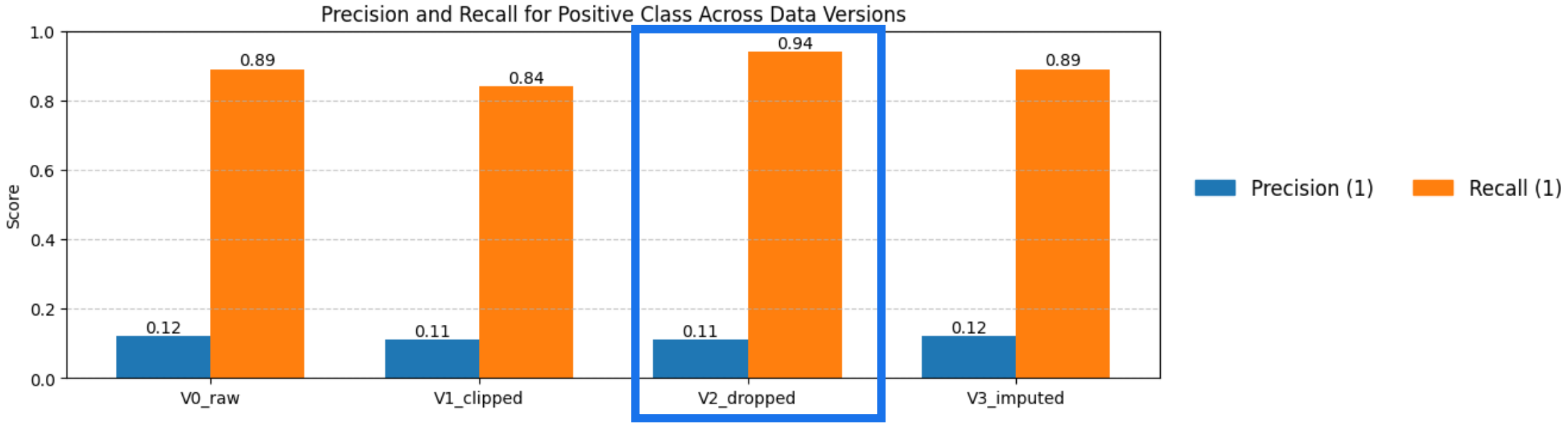
Pipeline Design



Model Experimentation

- Ran **H2O AutoML** across four data versions
- Fine-tuned the top models and selected a decision threshold to **maximize recall** while keeping **precision above 0.10**
- Compared recall and overall performance across all four data versions to identify the best approach
- Used **MLflow** to track experiments, log parameters, and record performance metrics

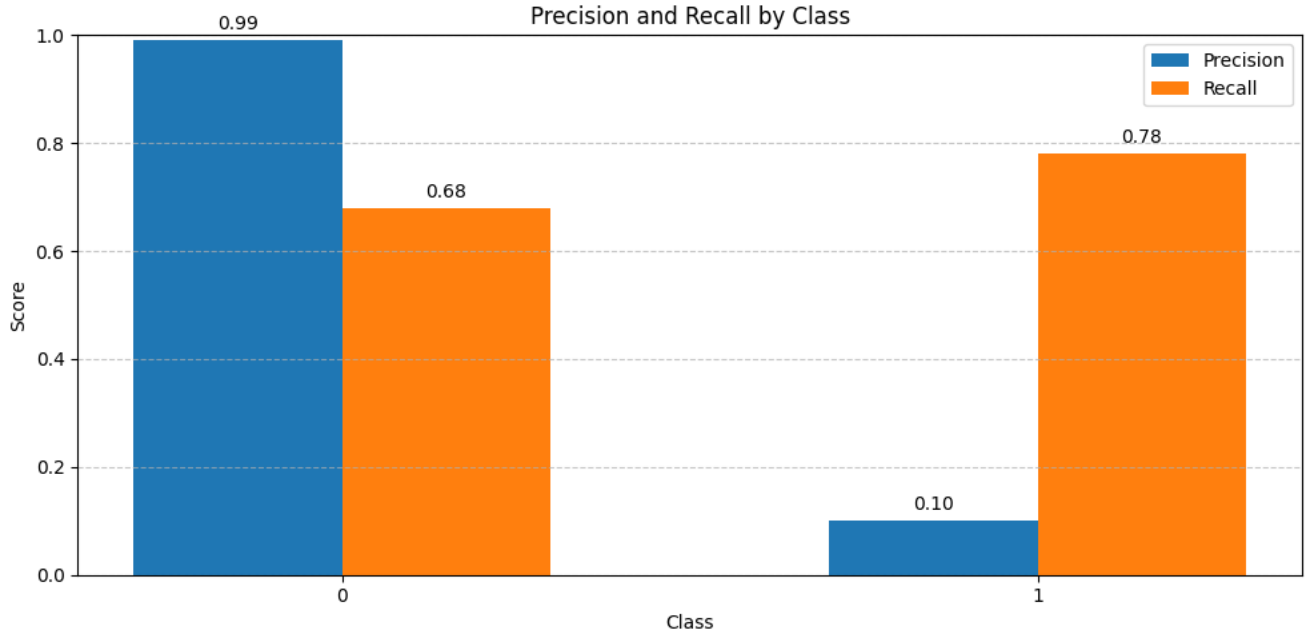
Model Performance on Validation Set

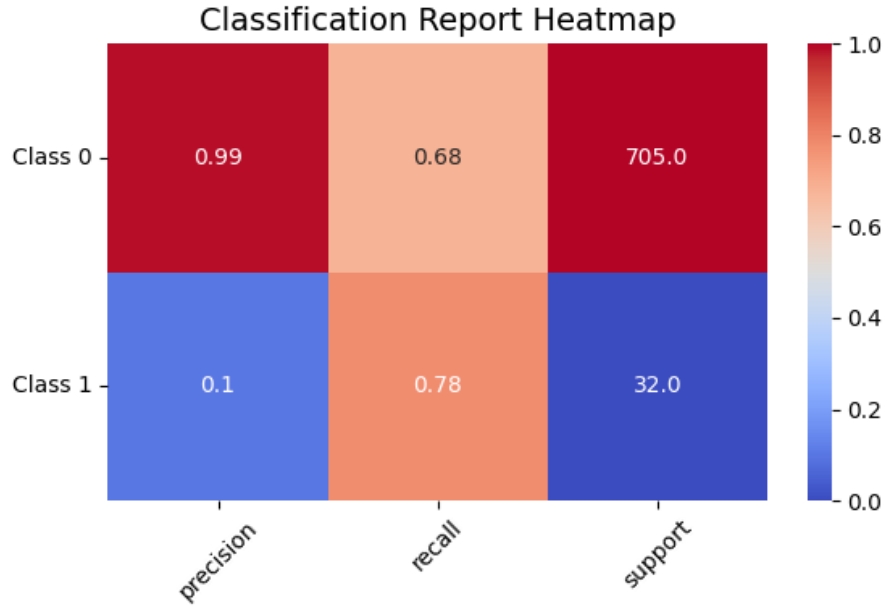


Model Evaluation

Final Model Configuration: Model: XGBoost (tuned), Data Version: V2_dropped, Decision Threshold: 0.4304

Model Performance on Test Set

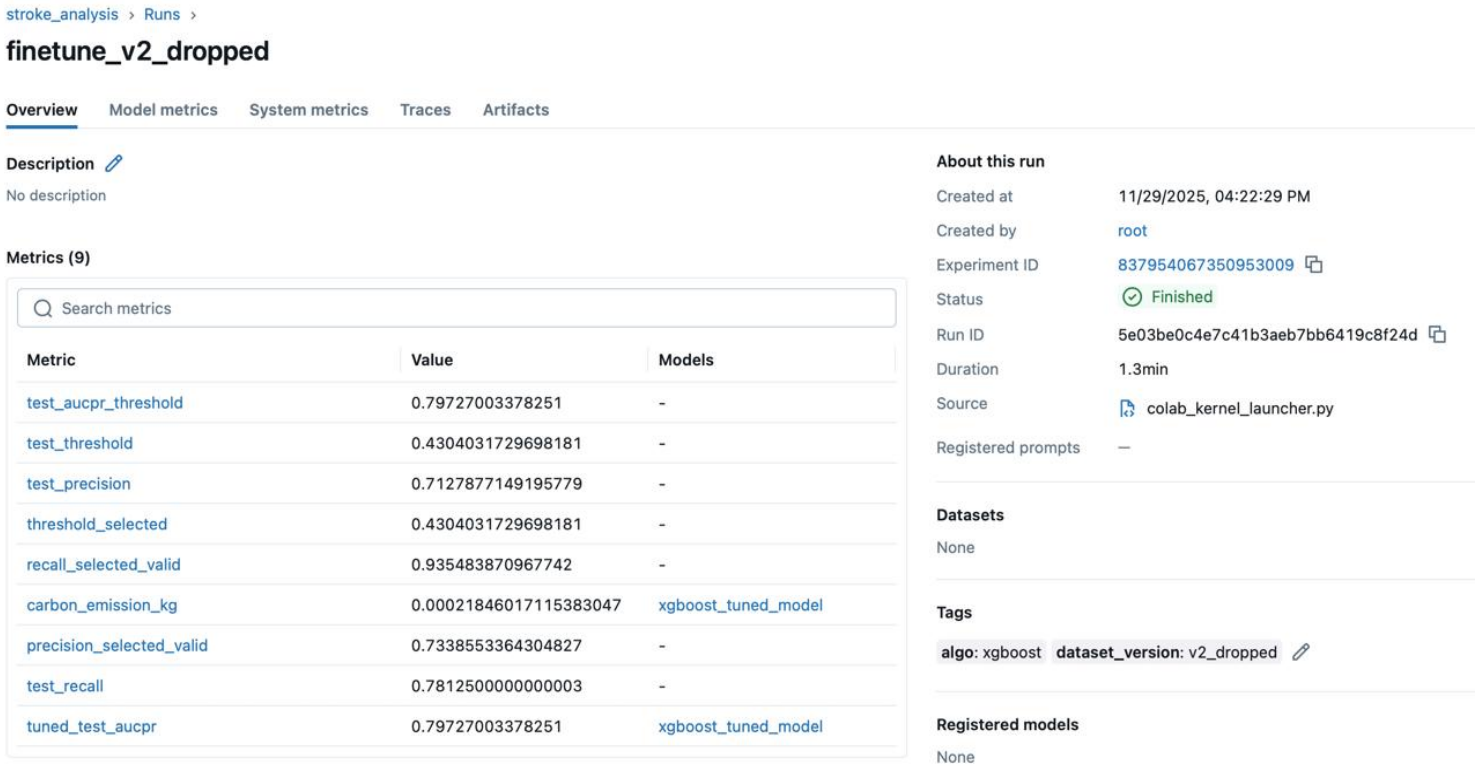




- The XGBoost model achieves high sensitivity (recall) for stroke cases, successfully identifying most positive cases despite class imbalance.
- Precision for the positive class remains low due to the very small number of stroke cases.

AutoML Result

Final Model Configuration: Model: XGBoost (tuned), Data Version: V2_dropped, Decision Threshold: 0.4304



stroke_analysis › Runs ›

finetune_v2_dropped

Overview | Model metrics | System metrics | Traces | Artifacts

Description

No description

Metrics (9)

Search metrics

Metric	Value	Models
test_aucpr_threshold	0.79727003378251	-
test_threshold	0.4304031729698181	-
test_precision	0.7127877149195779	-
threshold_selected	0.4304031729698181	-
recall_selected_valid	0.935483870967742	-
carbon_emission_kg	0.00021846017115383047	xgboost_tuned_model
precision_selected_valid	0.7338553364304827	-
test_recall	0.7812500000000003	-
tuned_test_aucpr	0.79727003378251	xgboost_tuned_model

About this run

Created at	11/29/2025, 04:22:29 PM
Created by	root
Experiment ID	837954067350953009
Status	Finished
Run ID	5e03be0c4e7c41b3aeb7bb6419c8f24d
Duration	1.3min
Source	colab_kernel_launcher.py
Registered prompts	—

Datasets

None

Tags

algo: xgboost | dataset_version: v2_dropped

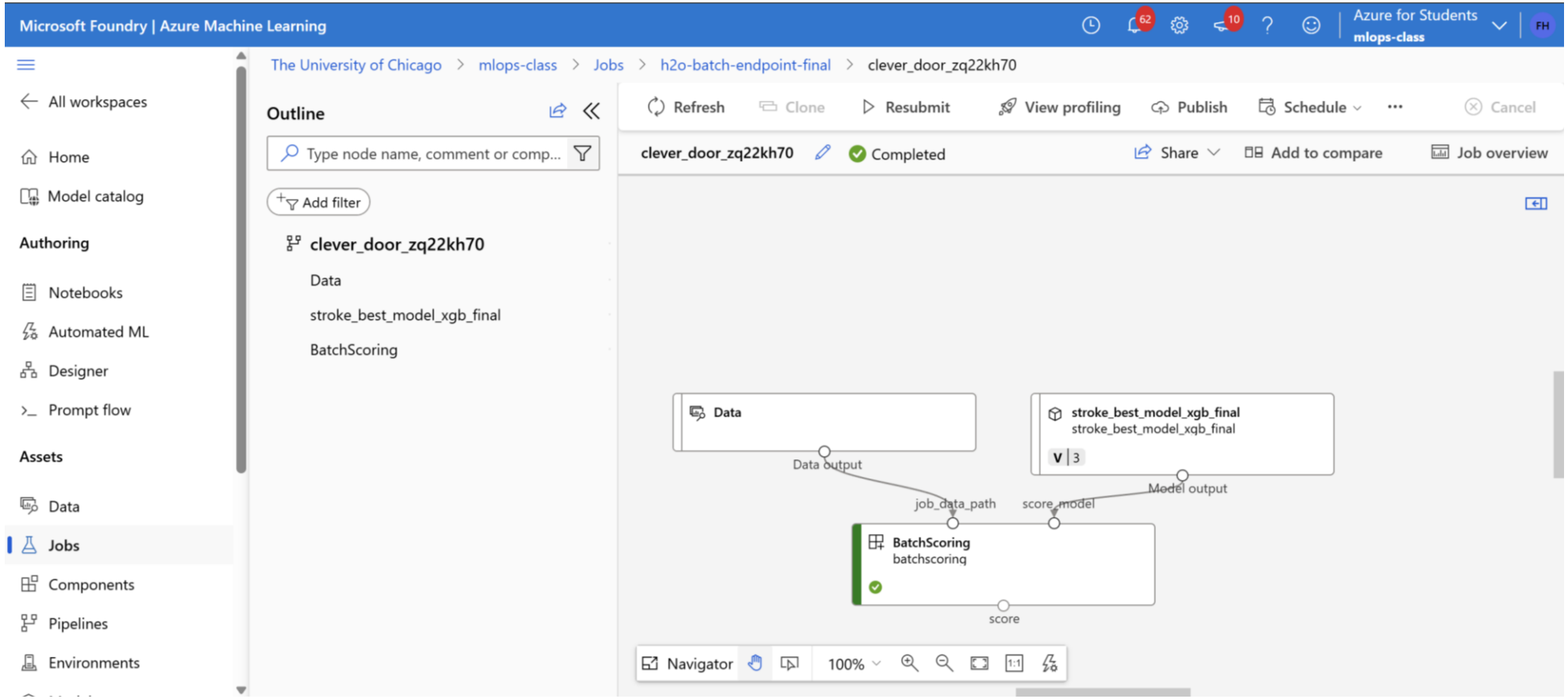
Registered models

None

Deployment Steps

1. Retrieve best model from H2O MLFlow and save it to fine tune: XGBoost model
2. Upload trained H2O model (MLflow format) and register it in Azure ML
3. Create custom environment that match H2O requirement
4. Prepare test dataset for batch inference
5. Create and run scoring script to load H2O model & threshold then generate predictions
6. Create batch deployment yaml: deployment configuration
7. Run batch inference for both original & modified dataset (age+10 years & average glucose +50%)
8. Compare results and create model monitoring

Deployment Result

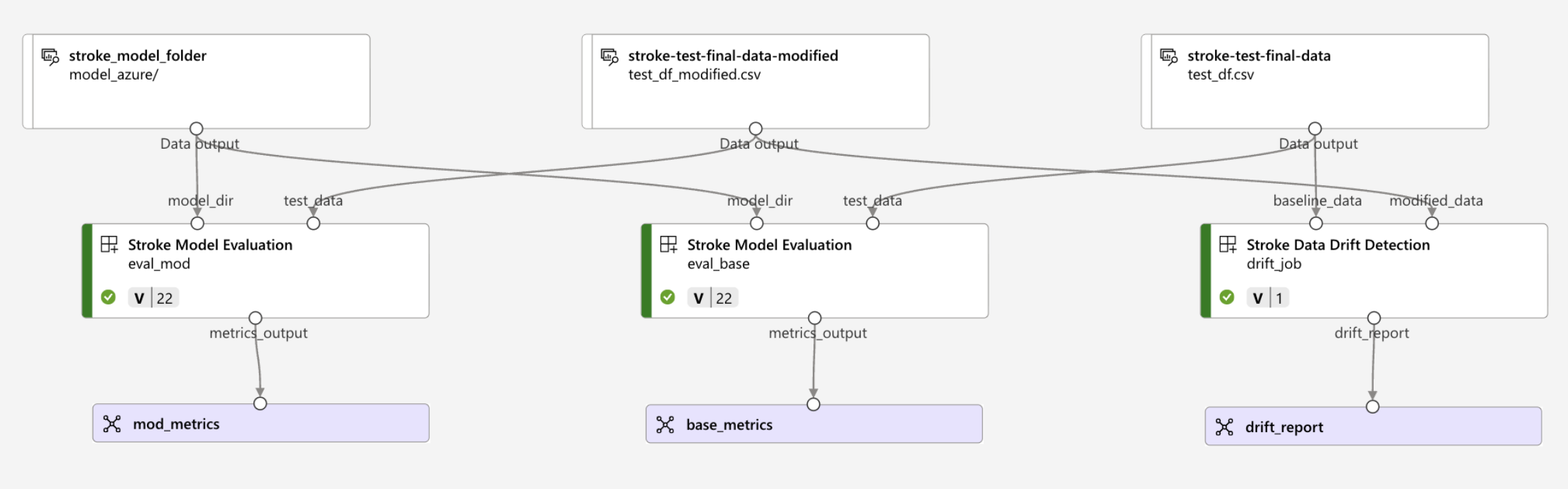


Deployment Result

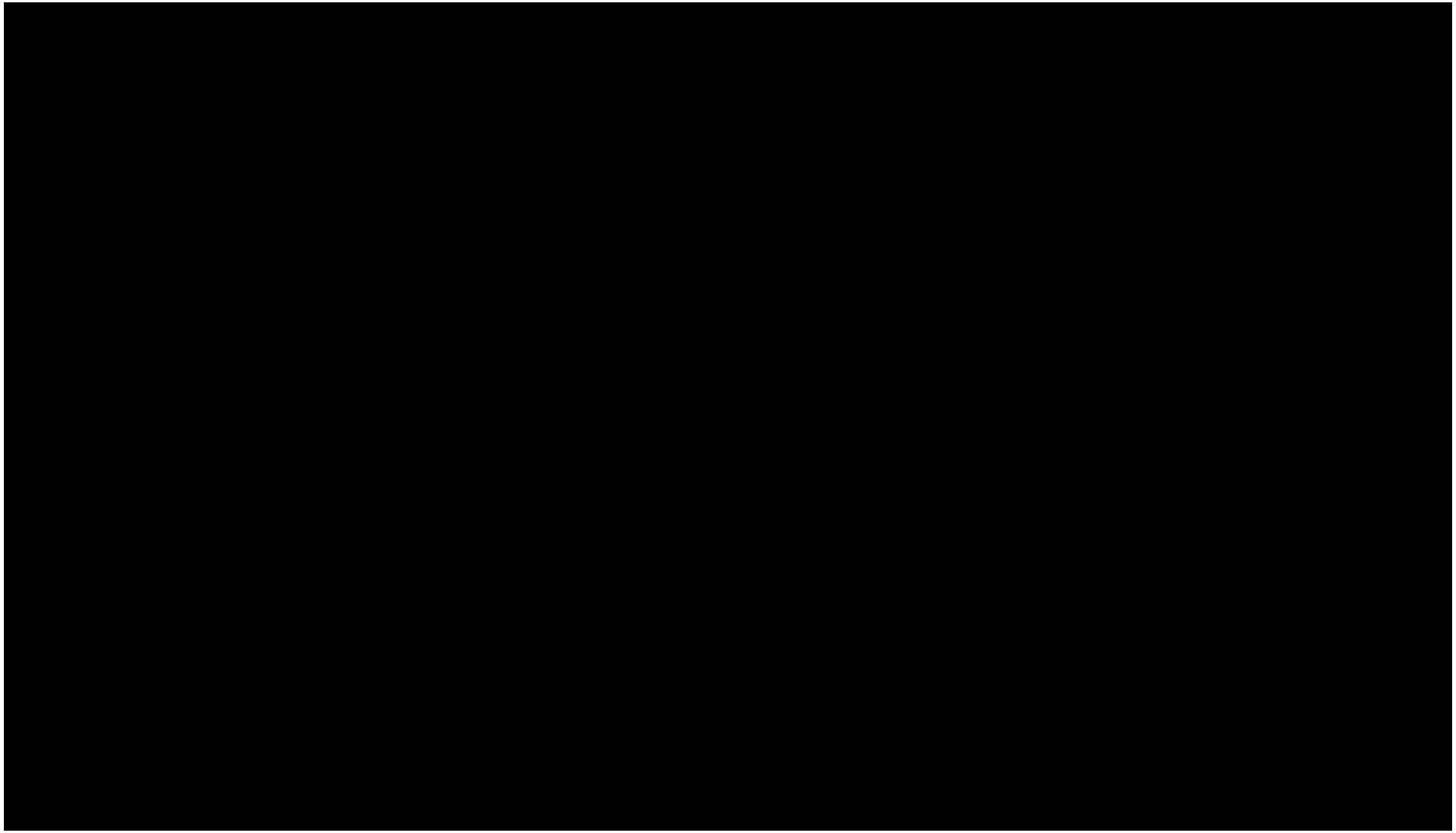
```
================================================================================
PREDICTION COMPARISON: ORIGINAL vs MODIFIED DATA
================================================================================

Original predictions: 737 rows
Modified predictions: 737 rows

================================================================================
PREDICTION DISTRIBUTION
================================================================================

Original Data:
  No Stroke (0): 486 (65.94%)
  Stroke (1):    251 (34.06%)

Modified Data:
  No Stroke (0): 353 (47.90%)
  Stroke (1):    384 (52.10%)

Change in Stroke Predictions:
  Absolute: +133 predictions
  Percentage: +18.05% points

================================================================================
STROKE PROBABILITY ANALYSIS
================================================================================

Original Data Probabilities:
  Mean: 0.2917
  Std:  0.2858
  Min:  0.0169
  Max:  0.8858
```

```
================================================================================
MODEL DEPLOYMENT & MONITORING REPORT
================================================================================
Generated: 2025-12-07 03:04:30
================================================================================

1. DEPLOYMENT STATUS: ✓ SUCCESS
   - Model: stroke_best_model_xgb_final:3
   - Deployment: h2o-model-script
   - Endpoint: h2o-batch-endpoint-final
   - Environment: h2o-batch-env-java:1 (with Java 11)

2. ORIGINAL TEST DATA RESULTS:
   - Total predictions: 737
   - Stroke predictions: 251 (34.06%)
   - No stroke predictions: 486 (65.94%)

3. MODIFIED TEST DATA RESULTS:
   - Modifications applied:
     * Age: +10 years
     * Glucose level: +50%
   - Total predictions: 737
   - Stroke predictions: 384 (52.10%)
   - No stroke predictions: 353 (47.90%)

4. IMPACT ANALYSIS:
   - Change in stroke predictions: +133
   - Change in stroke rate: +18.05% points
   - ⚠️ SIGNIFICANT INCREASE in stroke risk predictions
   - Model correctly identifies higher risk from age & glucose changes

5. MODEL MONITORING:
   - Feature drift detected: Age & glucose distribution shifted
   - Model response: Appropriate increase in stroke predictions
   - Recommendation: Model is performing as expected

================================================================================
CONCLUSION: Deployment successful. Model monitoring active.
Model correctly responds to risk factor changes.
================================================================================
```

Monitoring

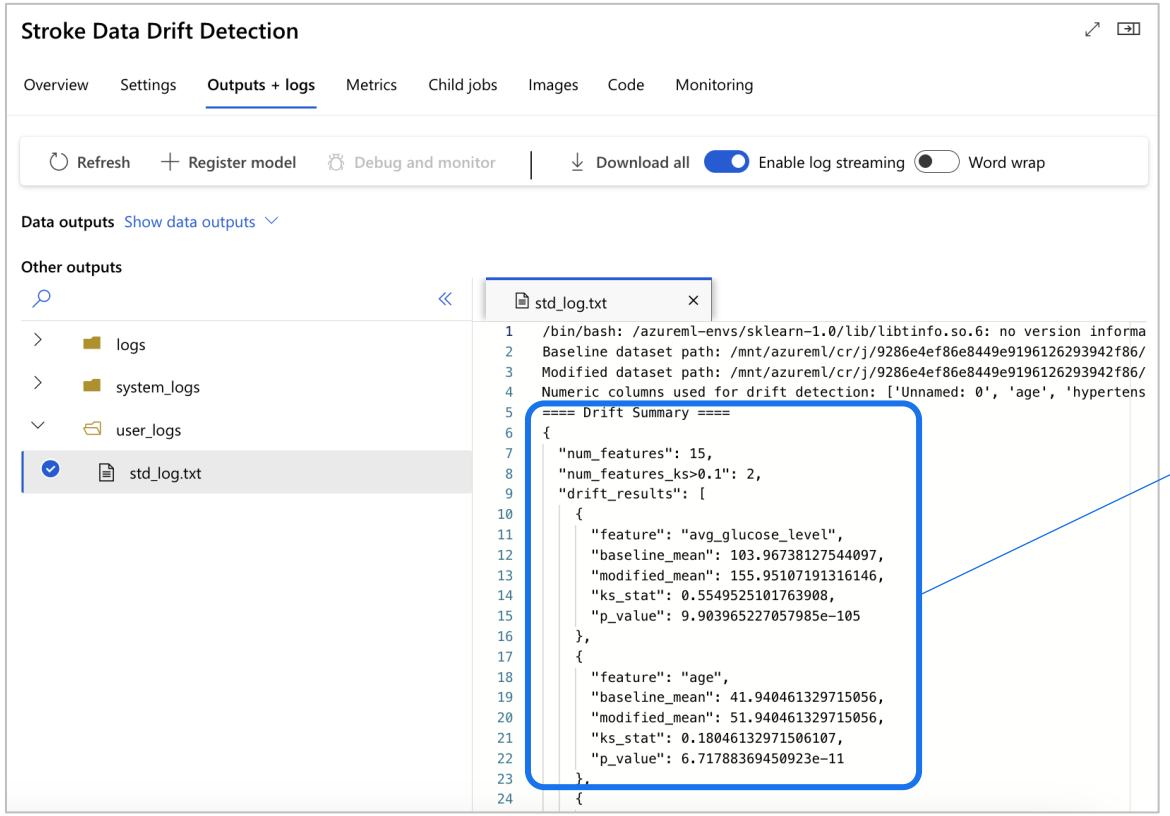

stroke_model_folder
model_azure/
stroke-test-final-data-modified
test_df_modified.csv
stroke-test-final-data
test_df.csv
Data output
Data output
Data output
model_dir
test_data
model_dir
test_data
baseline_data
modified_data
Stroke Model Evaluation
eval_mod
V 22
Stroke Model Evaluation
eval_base
V 22
Stroke Data Drift Detection
drift_job
V 1
metrics_output
metrics_output
drift_report
mod_metrics
base_metrics
drift_report

Demo



Data Drift Observation

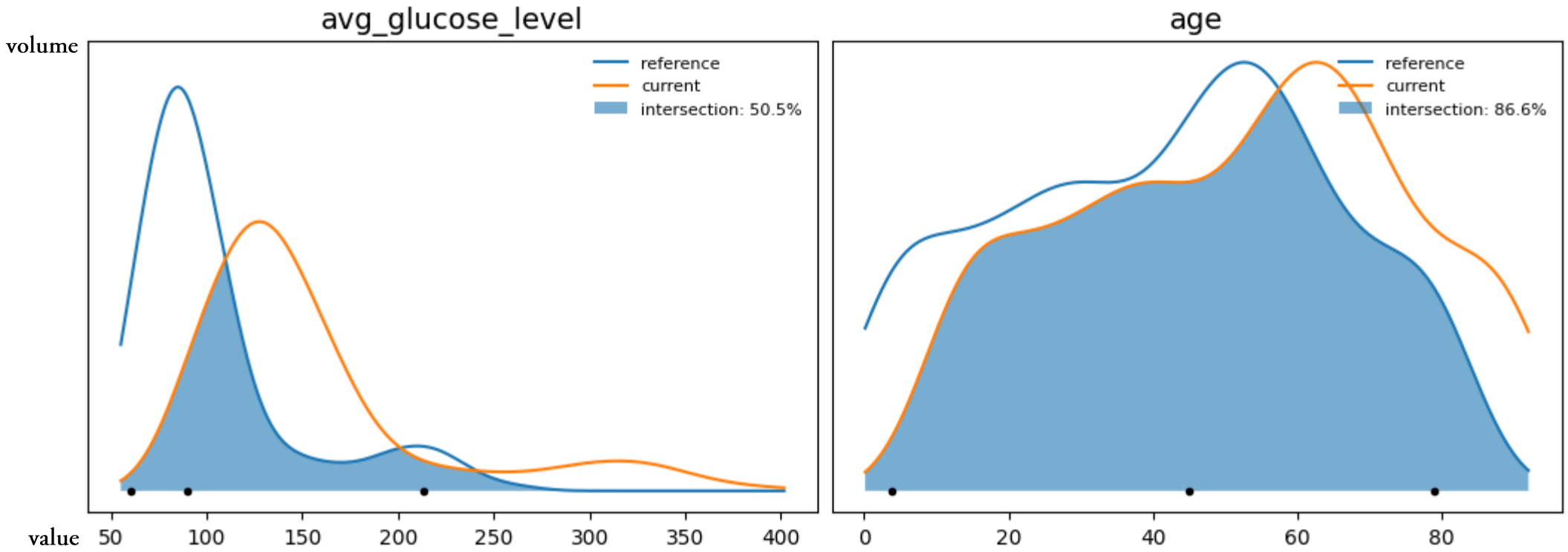
The result shows that 2 feature values changed → original & modified dataset (age+10 years & average glucose +50%)



==== Drift Summary ====
{
"num_features": 15,
"num_features_ks>0.1": 2,
"drift_results": [
{
"feature": "avg_glucose_level",
"baseline_mean": 103.96738127544097,
"modified_mean": 155.95107191316146,
"ks_stat": 0.5549525101763908,
"p_value": 9.903965227057985e-105
},
{
"feature": "age",
"baseline_mean": 41.940461329715056,
"modified_mean": 51.940461329715056,
"ks_stat": 0.18046132971506107,
"p_value": 6.71788369450923e-11
}

Data Drift Observation

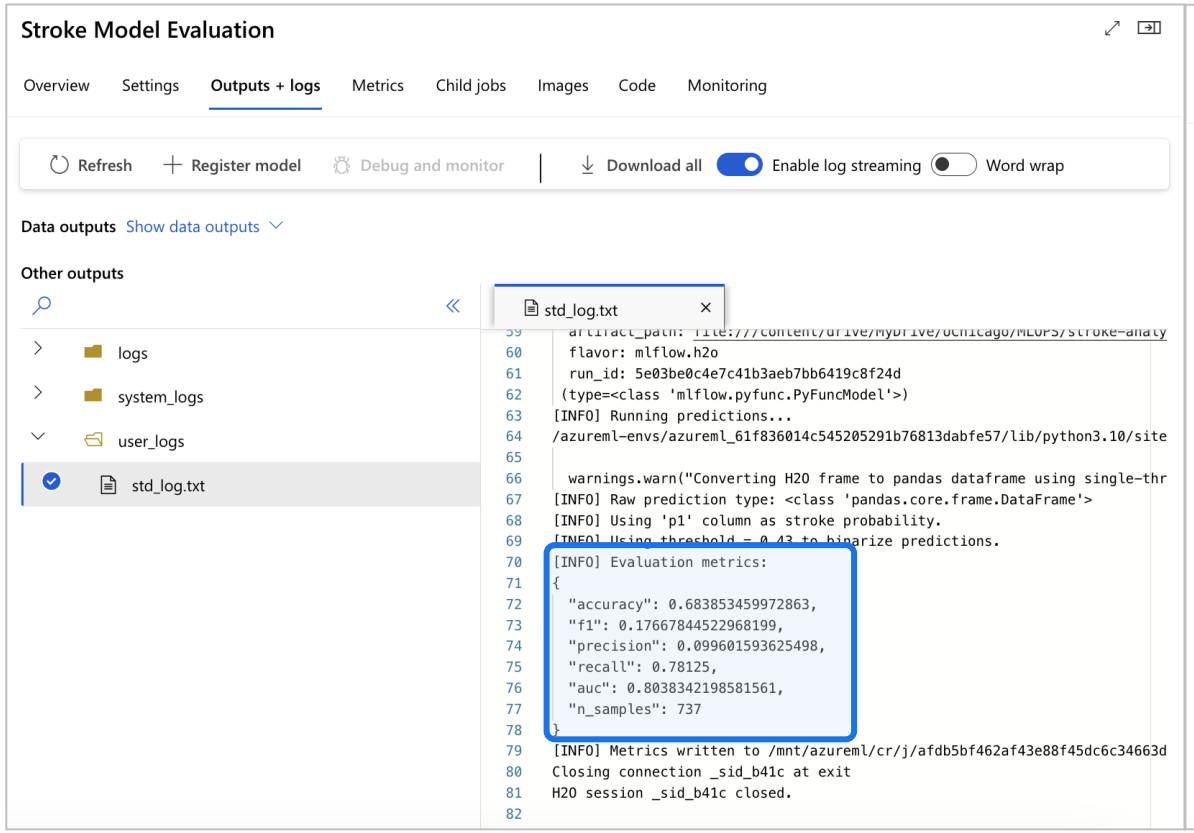
Feature Distribution drift (reference vs current)



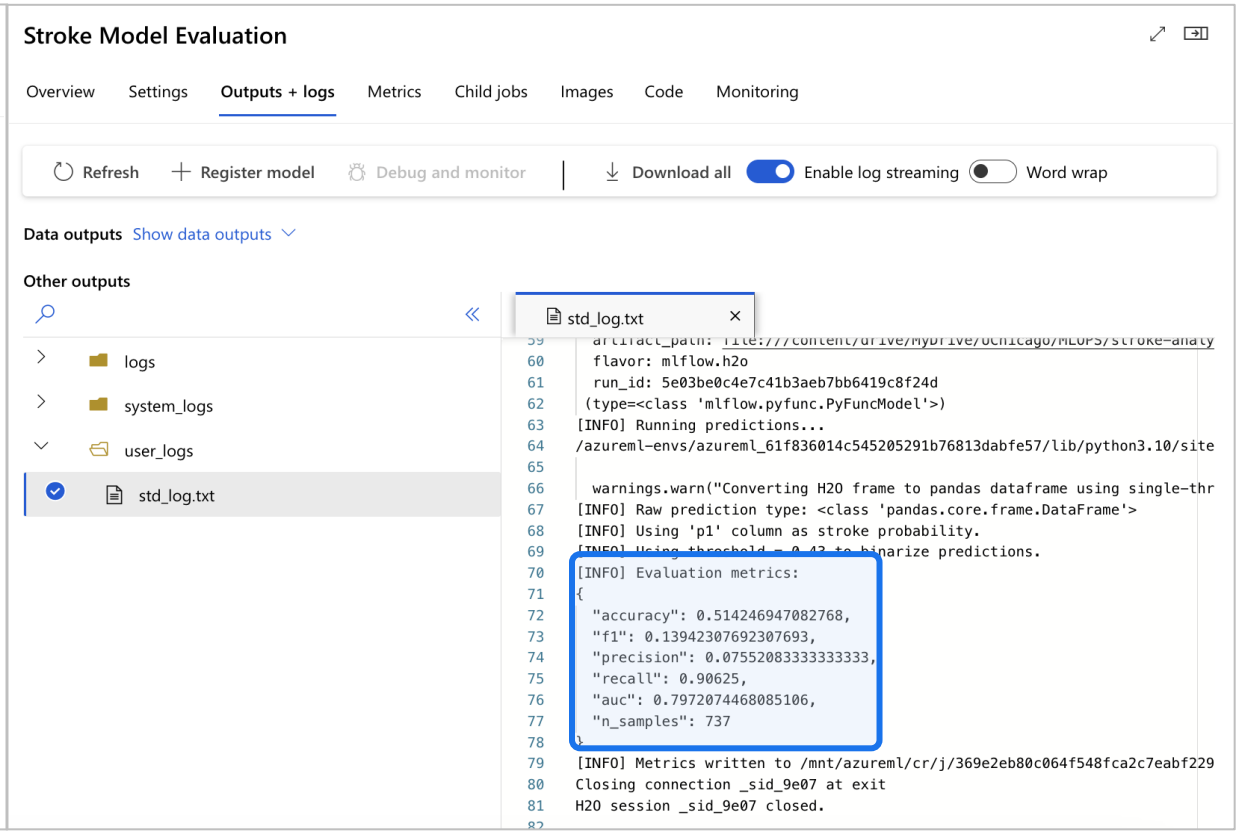
The modified test data show substantial shifts in average glucose level and age distributions compared to the reference set, with reduced overlap, indicating feature drift that may impact model reliability.

Model Evaluation (original v.s. modified test data)

Evaluation metrics (original test data)



Evaluation metrics (modified test data)



Model Evaluation (original v.s. modified test data)

	Original	Modified	% Change
Recall	0.78	0.90	+15.4%
Precision	0.10	0.08	–20%
Accuracy	0.68	0.51	–25.0%
F1	0.18	0.14	–22.2%
AUC	0.80	0.80	**0.0%**

The modified test data indicate a decrease in accuracy, precision, and F1 score, while recall increases due to the model's heightened sensitivity, leading to more false positives.

Conclusion

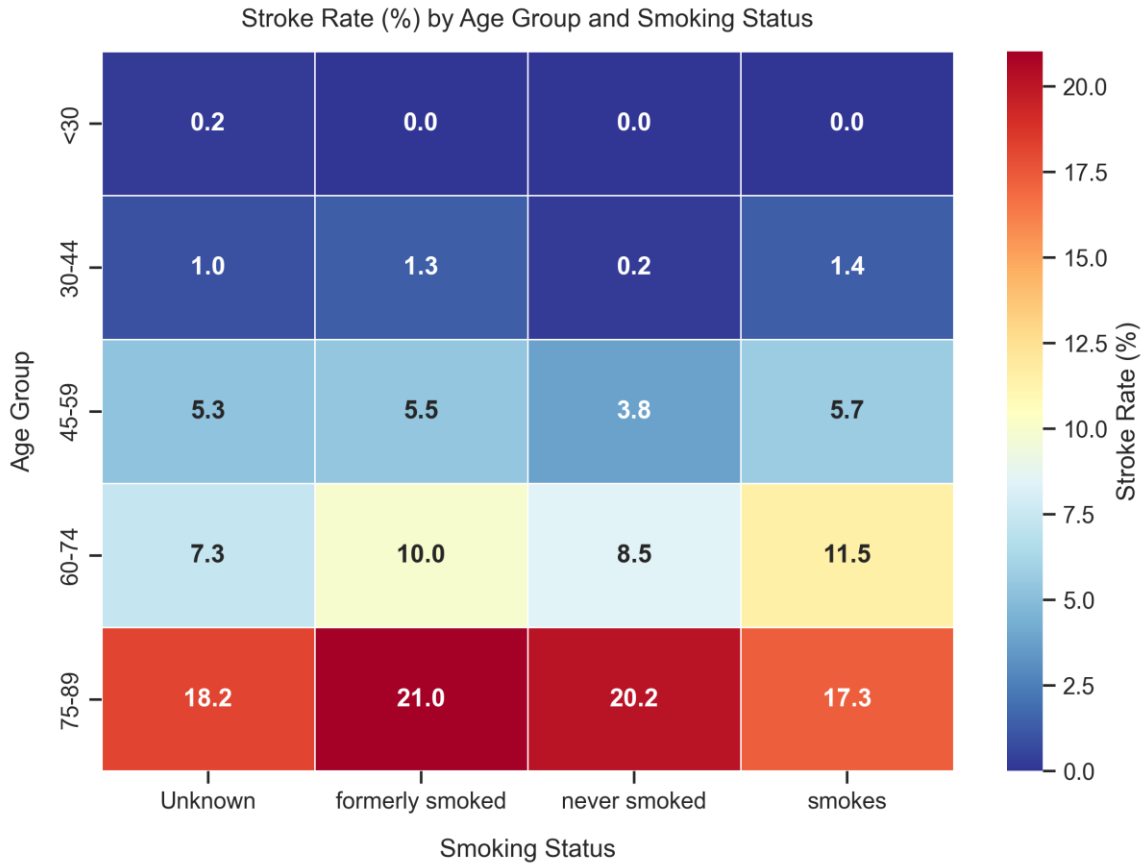
- We set out with the objective of creating a risk triage model using routine patient data and found that due to low prevalence we had to make a trade off with low precision to maintain high recall.
- Created features grounded in clinical intuition and data statistics to guide the model better.
- We used AutoML (H2O) for finding a baseline model and improving upon the baseline model.
- We deployed our model on Azure Cloud and it is being continuously monitored for data drift.

Team Contribution

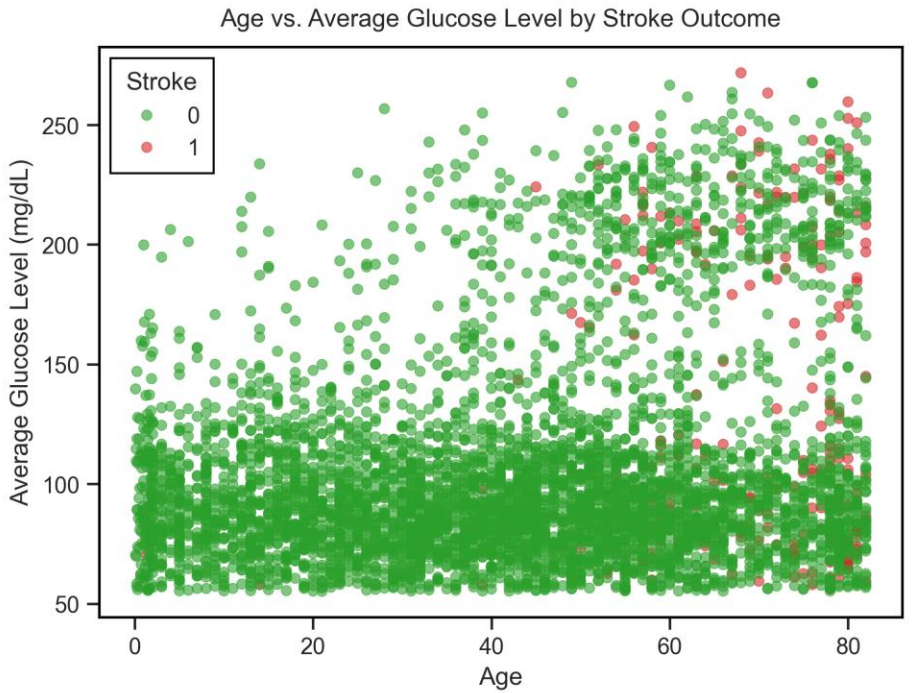
- Sarthak Dhanke: EDA & Feature Engineering & Baseline Model
- Sirinda Leesuravanich: Model Development & Model Evaluation
- Feby Hadayani: Model Deployment & Preprocessing
- Monica Ko: Model Monitoring & End-to-end Pipeline

Github link https://github.com/Sarztak/analysis-and-prediction-of-stroke-using-ml/tree/master

Appendix


Stroke Rate (%) by Age Group and Smoking Status
Age Group
<30
30-44
45-59
60-74
75-89
0.2
0.0
0.0
0.0
1.0
1.3
0.2
1.4
5.3
5.5
3.8
5.7
7.3
10.0
8.5
11.5
18.2
21.0
20.2
17.3
Unknown
formerly smoked
never smoked
smokes
Smoking Status
Stroke Rate (%)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

Correlation



Stroke cases cluster among higher age and glucose values, suggesting a joint risk pattern.

Model Evaluation

Final Model Configuration: Model: XGBoost (tuned), Data Version: V2_dropped, Decision Threshold: 0.4304

Model Performance on Test Set

Class	Precision	Recall	F1-score	Support
0	0.99	0.68	0.80	705
1	0.10	0.78	0.18	32
Accuracy			0.68	737
Macro avg	0.54	0.73	0.49	0.12
Weighted avg	0.95	0.68	0.78	737

Azure AutoML

- Applied Azure AutoML to the raw dataset to optimize **weighted AUC**

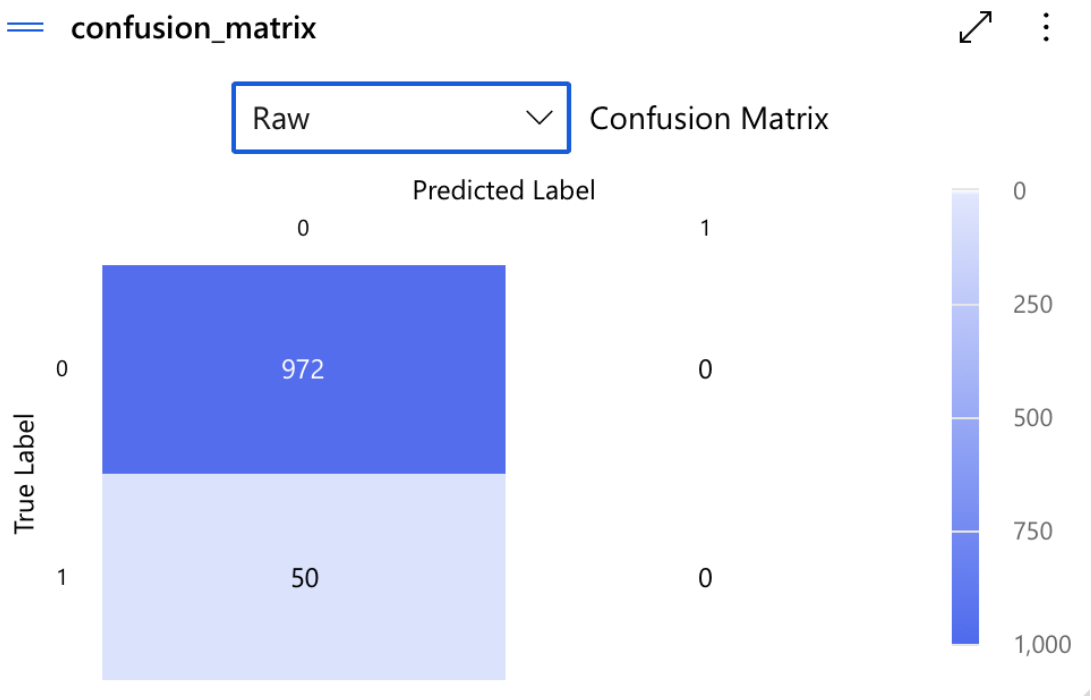
Overview | Data guardrails | **Models + child jobs** | Outputs + logs | Child jobs

Refresh | Deploy | Download | Explain model | View generated code

Search

Algorithm name	Responsible AI	AUC weighted ↓
VotingEnsemble		0.85859
MaxAbsScaler, LightGBM		0.85705
MaxAbsScaler, LightGBM		0.85654
MaxAbsScaler, LightGBM		0.85516
StandardScalerWrapper, LightGBM		0.85512
MaxAbsScaler, LightGBM		0.85507
StandardScalerWrapper, LightGBM		0.85478

Model Performance on Test Set



Data Description

- https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (~5000 rows, 11 columns)

Variable	Description
Id	Unique patient identifier
Gender	Male, Female
Age	
Hypertension	0: No hypertension, 1: Has hypertension
Heart_disease	0: No heart disease, 1: Has heart disease
Work_type	Children, Govt job, Never worked, Private or Self-employed
Residence_type	Urban, Rural
Avg_glucose_level	Avg glucose level in blood
BMI	Body Mass Index
Smoking_status	Formerly smoked, Never smoked, Smokes, Unknown