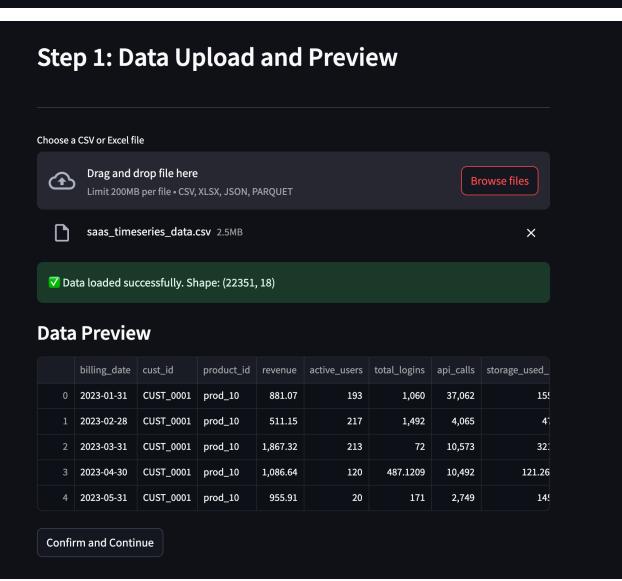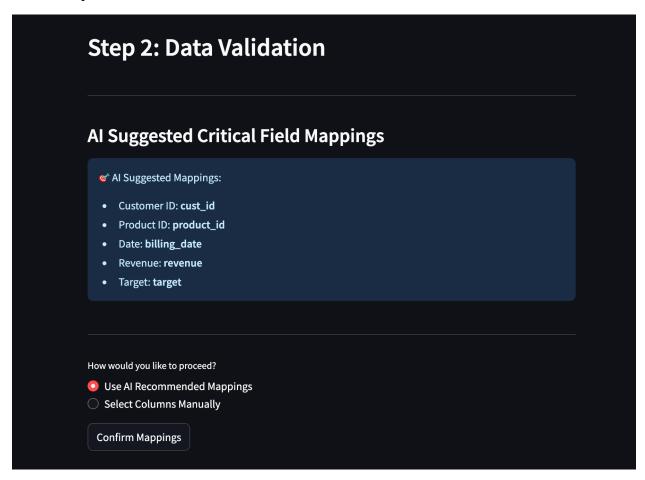# Classification Agent Tutorial

1. Today we will walk you through on how to use StepFunction's Classification Agent. This agent will take your cleaned data and run it through various classification models so you can get your business problem results. Start by uploading your data.

Classification App

Reset

Step 1: Data Upload and Preview

Choose a CSV or Excel file

Drag and drop file here
Limit 200MB per file • CSV, XLSX, JSON, PARQUET

Browse files

saas_timeseries_data.csv 2.5MB

✅ Data loaded successfully. Shape: (22351, 18)

Data Preview

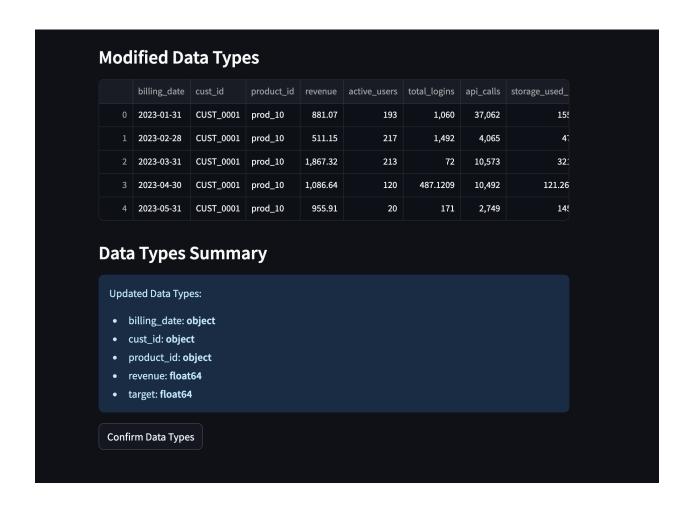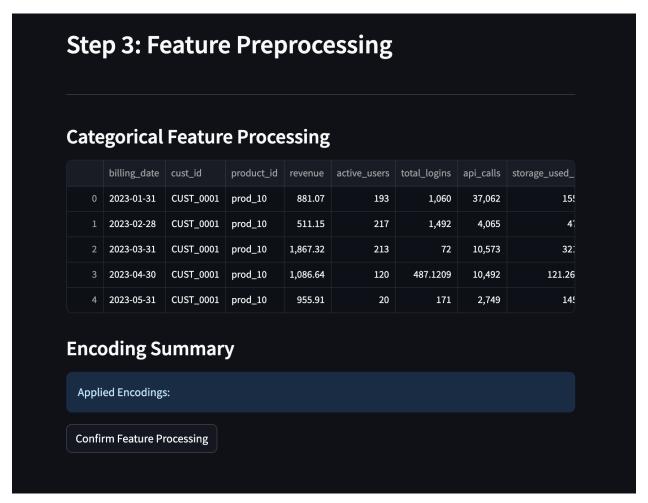|	billing_date	cust_id	product_id	revenue	active_users	total_logins	api_calls	storage_used_
0	2023-01-31	CUST_0001	prod_10	881.07	193	1,060	37,062	15!
1	2023-02-28	CUST_0001	prod_10	511.15	217	1,492	4,065	4;
2	2023-03-31	CUST_0001	prod_10	1,867.32	213	72	10,573	32;
3	2023-04-30	CUST_0001	prod_10	1,086.64	120	487.1209	10,492	121.26
4	2023-05-31	CUST_0001	prod_10	955.91	20	171	2,749	14!

Confirm and Continue

2. Next we will validate your data by aligning it with key mapping fields so that we can align our AI models with the right parameters. The Classification Agent will suggest these mapping fields for you, but if you want to use different ones, you can manually edit them by clicking the manual column selection button. Here, you can manually choose the fields that correspond to the required columns via a dropdown correspond via a dropdown.

Step 2: Data Validation

AI Suggested Critical Field Mappings

🎯 AI Suggested Mappings:

- Customer ID: **cust_id**
- Product ID: **product_id**
- Date: **billing_date**
- Revenue: **revenue**
- Target: **target**

How would you like to proceed?

🔴 Use AI Recommended Mappings
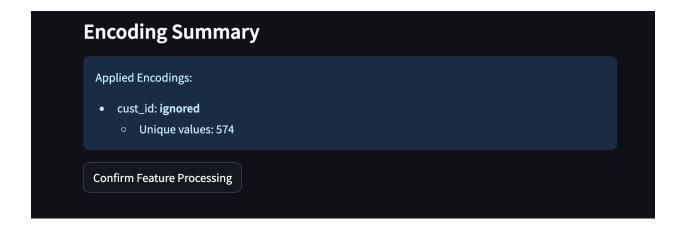⚪ Select Columns Manually

Confirm Mappings

Step 2 continued: After confirming the mappings, we will present you with the datatypes of each of the fields for a final confirmation before proceeding to the next step.

Modified Data Types

	billing_date	cust_id	product_id	revenue	active_users	total_logins	api_calls	storage_used_
0	2023-01-31	CUST_0001	prod_10	881.07	193	1,060	37,062	15!
1	2023-02-28	CUST_0001	prod_10	511.15	217	1,492	4,065	4
2	2023-03-31	CUST_0001	prod_10	1,867.32	213	72	10,573	32
3	2023-04-30	CUST_0001	prod_10	1,086.64	120	487.1209	10,492	121.26
4	2023-05-31	CUST_0001	prod_10	955.91	20	171	2,749	14!

Data Types Summary

Updated Data Types:

- billing_date: **object**
- cust_id: **object**
- product_id: **object**
- revenue: **float64**
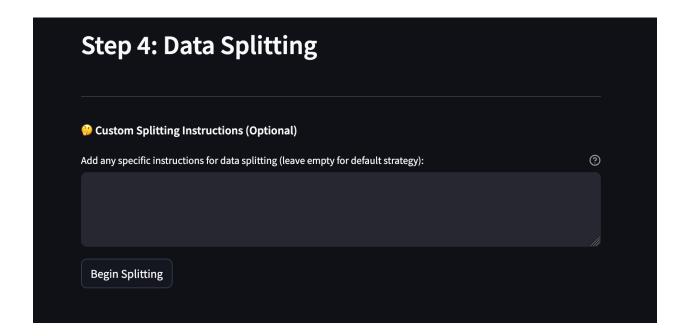- target: **float64**

Confirm Data Types

3. Now, the Classification Agent will perform some data quality checks and clean your data so that it can run through classification models. The individual changes will be presented to you once you confirm.

Step 3: Feature Preprocessing

Categorical Feature Processing

	billing_date	cust_id	product_id	revenue	active_users	total_logins	api_calls	storage_used_
0	2023-01-31	CUST_0001	prod_10	881.07	193	1,060	37,062	15!
1	2023-02-28	CUST_0001	prod_10	511.15	217	1,492	4,065	4'
2	2023-03-31	CUST_0001	prod_10	1,867.32	213	72	10,573	32'
3	2023-04-30	CUST_0001	prod_10	1,086.64	120	487.1209	10,492	121.26
4	2023-05-31	CUST_0001	prod_10	955.91	20	171	2,749	14!

Encoding Summary

Applied Encodings:

Confirm Feature Processing

Encoding Summary

Applied Encodings:

- cust_id: **ignored**
  - Unique values: 574

Confirm Feature Processing

4. Next, your cleaned data will be split into training data and validation data for your model. The Agent will determine the optimal split of your data, but if you want the data split in a custom way, type how you want the data to be split in the text box itself.

Step 4: Data Splitting

🤔 **Custom Splitting Instructions (Optional)**

Add any specific instructions for data splitting (leave empty for default strategy):

Begin Splitting

Step 4 continued: The Agent will show you how it split up your data

Data Split Information

🎯 **AI's Splitting Strategy:**

> For the time series data, the inference set was created using records from the last month, while the validation set included records from the month prior to the last month, covering a total of three months for validation. The training set consisted of all records prior to the validation period.

Training Period:

- Start: 2023-01-31 00:00:00

- End: 2024-09-30 00:00:00

Validation Period:

- Start: 2024-10-31 00:00:00

- End: 2024-12-31 00:00:00

Inference Period:
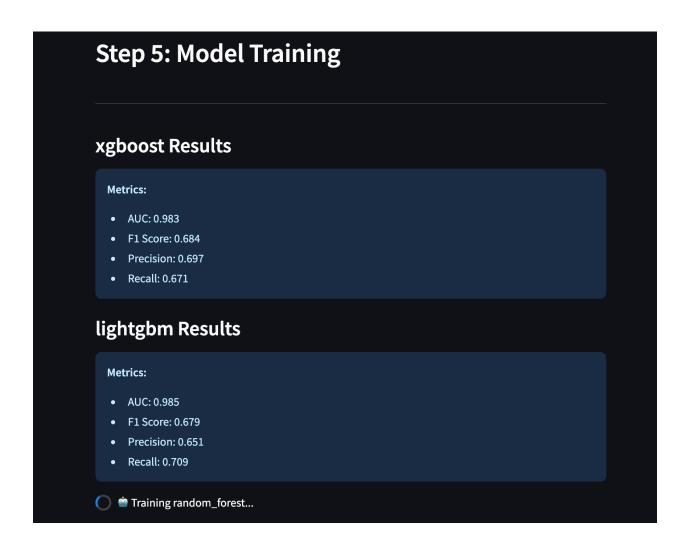
- Month: 2025-01

Sample Counts:

- Training: **18000**

- Validation: **3230**
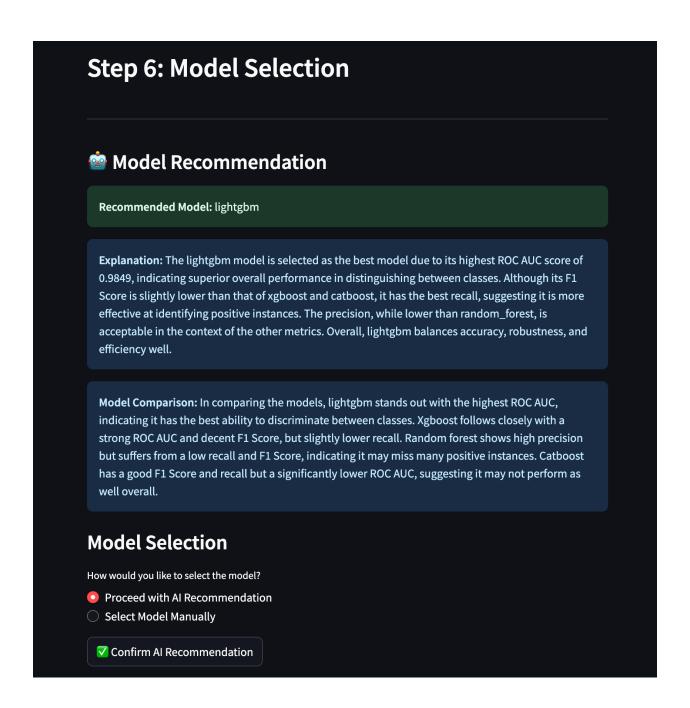
- Inference: **1121**

Confirm Data Split
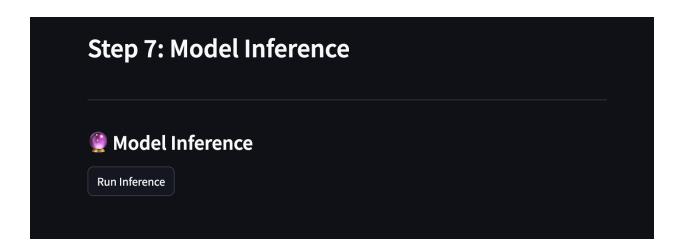
5. Upon confirming the data split in the previous step, multiple classification models will be trained using your data. As they are trained, you will be provided with model health stats that will be used to find the best model for your business task.

Step 5: Model Training

xgboost Results

Metrics:

- AUC: 0.983
- F1 Score: 0.684
- Precision: 0.697
- Recall: 0.671

lightgbm Results

Metrics:

- AUC: 0.985
- F1 Score: 0.679
- Precision: 0.651
- Recall: 0.709

🤖 Training random_forest...

6. After training, the Classification Agent will determine the best model to use to run inference on your data. If you wish to override it and use a different model, you can select the model manually via the button below. Otherwise, you can confirm the Agent's recommendation and move onto the next step.

Step 6: Model Selection

🤖 Model Recommendation

Recommended Model: lightgbm

Explanation: The lightgbm model is selected as the best model due to its highest ROC AUC score of 0.9849, indicating superior overall performance in distinguishing between classes. Although its F1 Score is slightly lower than that of xgboost and catboost, it has the best recall, suggesting it is more effective at identifying positive instances. The precision, while lower than random_forest, is acceptable in the context of the other metrics. Overall, lightgbm balances accuracy, robustness, and efficiency well.

Model Comparison: In comparing the models, lightgbm stands out with the highest ROC AUC, indicating it has the best ability to discriminate between classes. Xgboost follows closely with a strong ROC AUC and decent F1 Score, but slightly lower recall. Random forest shows high precision but suffers from a low recall and F1 Score, indicating it may miss many positive instances. Catboost has a good F1 Score and recall but a significantly lower ROC AUC, suggesting it may not perform as well overall.

Model Selection

How would you like to select the model?

🔘 Proceed with AI Recommendation
⚪ Select Model Manually

✅ Confirm AI Recommendation

7. The Agent will now run inference on the best model and provide you with a summary of the results.

Step 7: Model Inference

🔮 Model Inference

Run Inference

Inference Results Summary:

- Total Records: 1121
- Positive Predictions: 12 (1.1%)
- Average Probability: 0.015

Sample Predictions

	billing_date	cust_id	product_id	revenue	active_users	total_logins	api_calls	storage
96	2025-01-31 00:00:00	CUST_0005	prod_04	403.13	21	645	2,203	
133	2025-01-31 00:00:00	CUST_0006	prod_04	1,356.81	41	1,050	17,057	
158	2025-01-31 00:00:00	CUST_0007	prod_03	1,401.2	20	2,378	16,411	
212	2025-01-31 00:00:00	CUST_0007	prod_09	1,651.34	28	1,597	16,281	
257	2025-01-31 00:00:00	CUST_0009	prod_09	1,193.28	81	1,463	15,670	

✅ Model Inference Complete:

- Processed Records: **1121**
- Positive Predictions: **12** (1.1%)

8. As the last step, download your results, and you have successfully run classification models on your data.

📊 Detailed Results

	billing_date	cust_id	product_id	revenue	active_users	total_logins	api_calls	storag
96	2025-01-31 00:00:00	CUST_0005	prod_04	403.13	21	645	2,203	
133	2025-01-31 00:00:00	CUST_0006	prod_04	1,356.81	41	1,050	17,057	
158	2025-01-31 00:00:00	CUST_0007	prod_03	1,401.2	20	2,378	16,411	
212	2025-01-31 00:00:00	CUST_0007	prod_09	1,651.34	28	1,597	16,281	
257	2025-01-31 00:00:00	CUST_0009	prod_09	1,193.28	81	1,463	15,670	
282	2025-01-31 00:00:00	CUST_0010	prod_02	2,286.96	213	2,516	24,428	
307	2025-01-31 00:00:00	CUST_0010	prod_09	404.32	28	247	4,378	
332	2025-01-31 00:00:00	CUST_0011	prod_01	516.3	84	794	14,796	
380	2025-01-31 00:00:00	CUST_0011	prod_05	1,000.78	193	262	15,293	
405	2025-01-31 00:00:00	CUST_0011	prod_07	1,609.36	138	3,979	13,470	

📥 Download Predictions as CSV