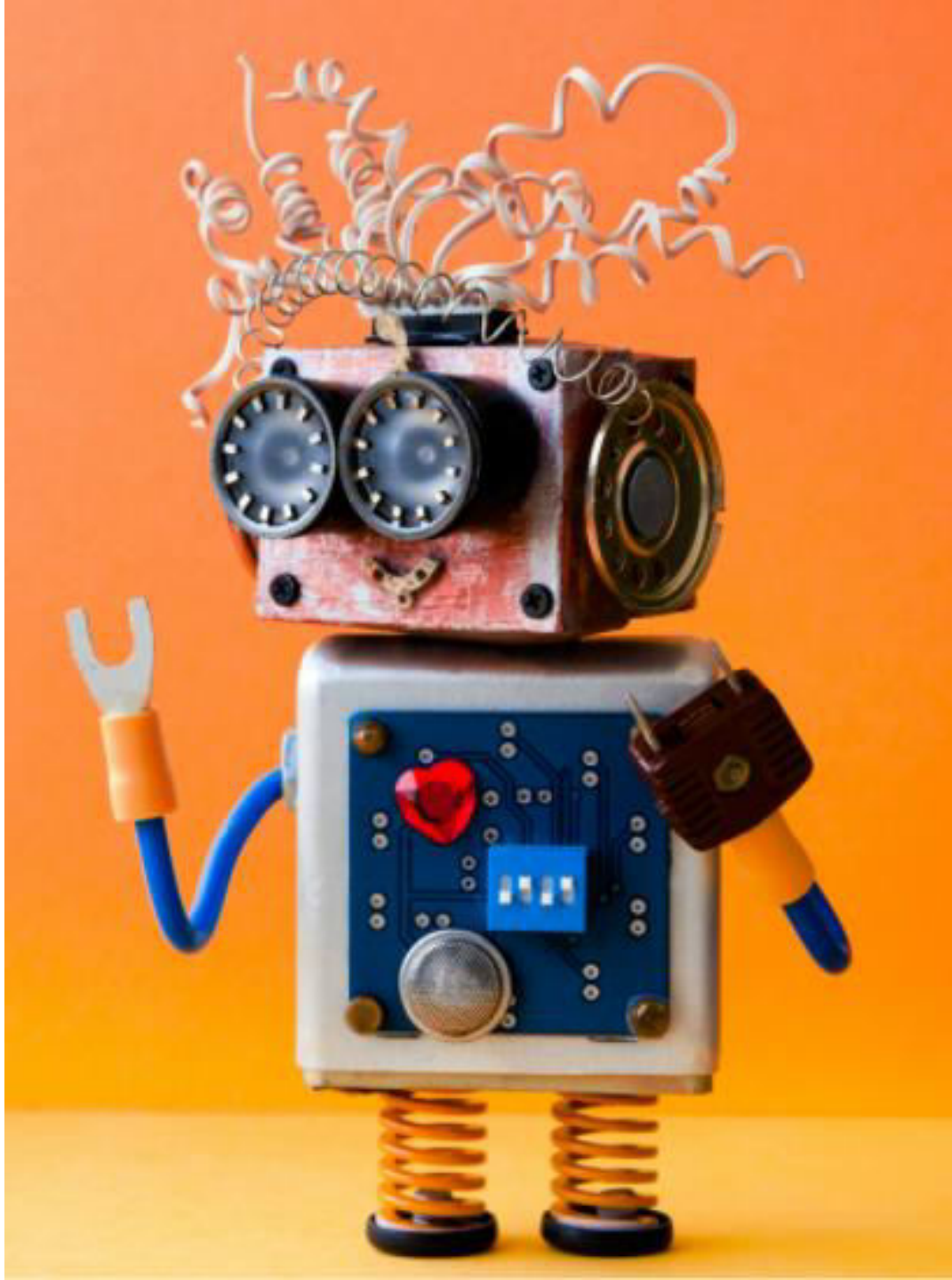
PROJECT CARD

EASY ADVANCED



PROJECT CARD

GOAL:

- *Build, train, test and deploy an AI/ML model to predict which employees are more likely to quit based on their features.*

TOOL:

- *AWS SageMaker Canvas*

PRACTICAL REAL-WORLD APPLICATION:

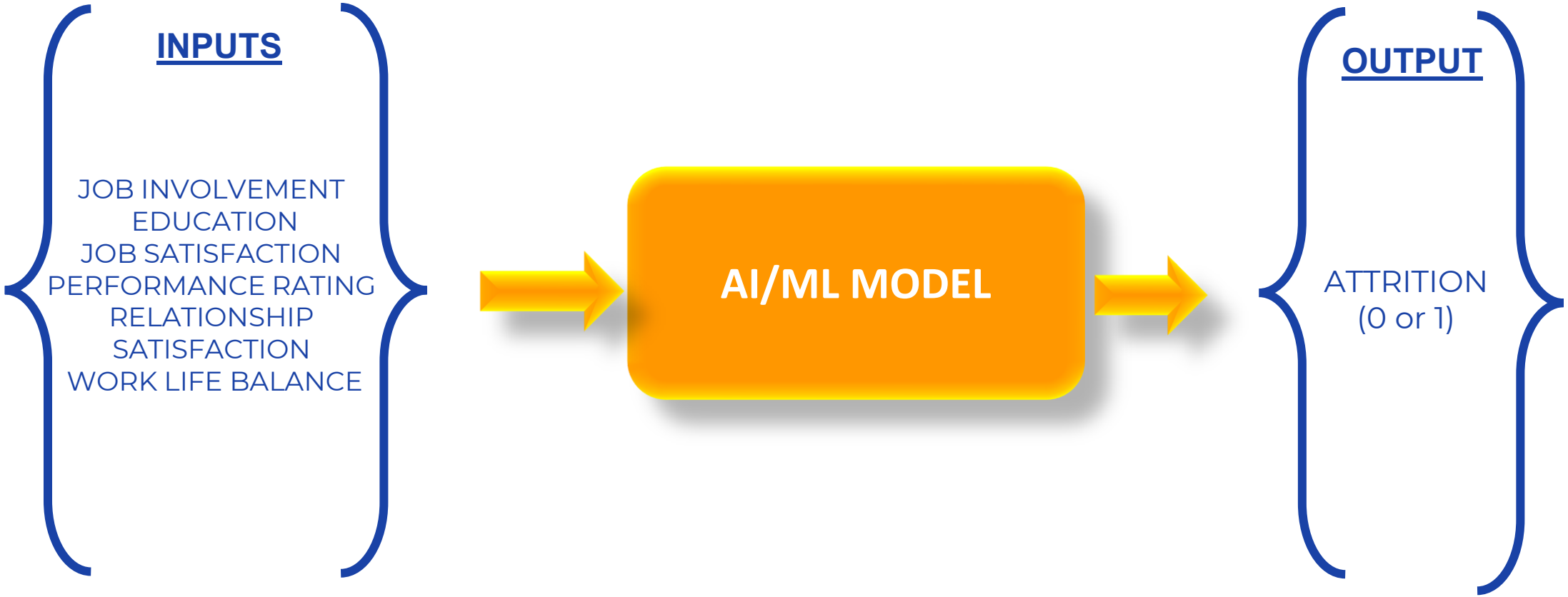
- *This project can be effectively used in Human Resources Departments to predict employee attrition and understand key factors that contribute to it.*

DATA:

- **INPUTS:**
  - *Job Involvement, Education, Job Satisfaction, Performance Rating, Relationship Satisfaction, Work Life Balance*
- **OUTPUT:**
  - *Attrition (binary), i.e.: employee stayed or left*

https://publicdomainvectors.org/en/free-clipart/Employee-search-concept/83802.html
https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

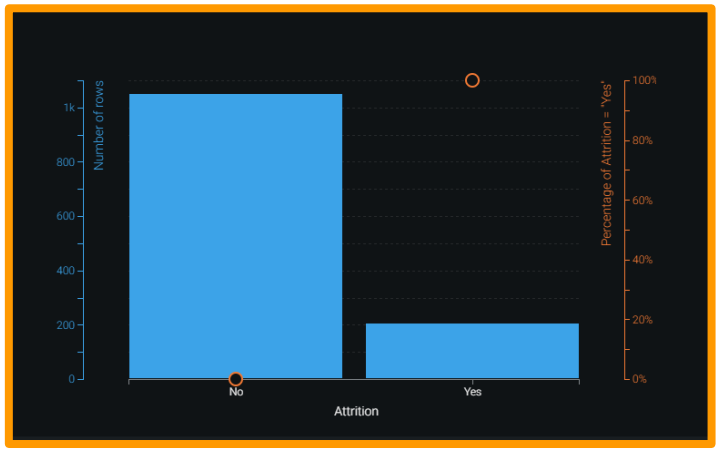
INPUTS AND OUTPUTS



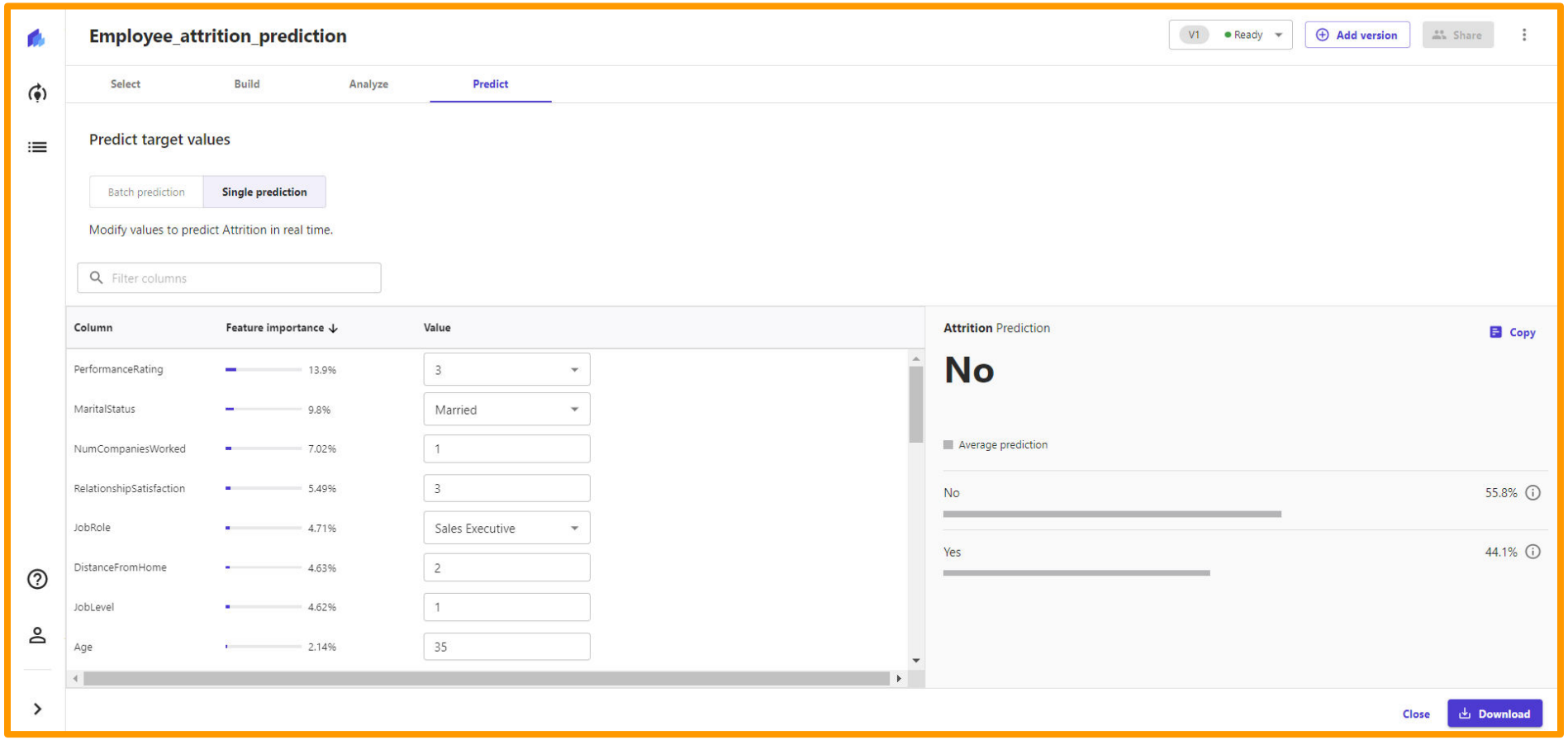
DATA OVERVIEW

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate	NumCompaniesWorked	Over18	OverTime	PercentSalaryHike	PerformanceRating
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	2	Sales Exe	4	Single	5993	19479	8	Y	Yes	11	3
4	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	3	Male	61	2	2	Research	2	Married	5130	24907	1	Y	No	23	4
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	4	Male	92	2	1	Laborato	3	Single	2090	2396	6	Y	Yes	15	3
3	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	4	Female	56	3	1	Research	3	Married	2909	23159	1	Y	Yes	11	3
27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	1	Male	40	3	1	Laborato	2	Married	3468	16632	9	Y	No	12	3
3	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	4	Male	79	3	1	Laborato	4	Single	3068	11864	0	Y	No	13	3
53	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	3	Female	81	4	1	Laborato	1	Married	2670	9964	4	Y	Yes	20	4
3	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11	4	Male	67	3	1	Laborato	3	Divorced	2693	13335	1	Y	No	22	4
38	No	Travel_Frequently	216	Research & Development	23	3	Life Sciences	1	12	4	Male	44	2	3	Manufac	3	Single	9526	8787	0	Y	No	21	4
3	No	Travel_Rarely	1299	Research & Development	27	3	Medical	1	13	3	Male	94	3	2	Healthca	3	Married	5237	16577	6	Y	No	13	3
39	No	Travel_Rarely	809	Research & Development	16	3	Medical	1	14	1	Male	84	4	1	Laborato	2	Married	2426	16479	0	Y	No	13	3
2	No	Travel_Rarely	153	Research & Development	15	2	Life Sciences	1	15	4	Female	49	2	2	Laborato	3	Single	4193	12682	0	Y	Yes	12	3
3	No	Travel_Rarely	670	Research & Development	26	1	Life Sciences	1	16	1	Male	31	3	1	Research	3	Divorced	2911	15170	1	Y	No	17	3
3	No	Travel_Rarely	1346	Research & Development	19	2	Medical	1	18	2	Male	93	3	1	Laborato	4	Divorced	2661	8758	0	Y	No	11	3
20	Yes	Travel_Rarely	103	Research & Development	24	3	Life Sciences	1	19	3	Male	50	2	1	Laborato	3	Single	2028	12947	5	Y	Yes	14	3
21	No	Travel_Rarely	1389	Research & Development	21	4	Life Sciences	1	20	2	Female	51	4	3	Manufac	1	Divorced	9980	10195	1	Y	No	11	3
36	No	Travel_Rarely	334	Research & Development	5	2	Life Sciences	1	21	1	Male	80	4	1	Research	2	Divorced	3298	15053	0	Y	Yes	12	3
21	No	Non-Travel	1123	Research & Development	16	2	Medical	1	22	4	Male	96	4	1	Laborato	4	Divorced	2935	7324	1	Y	Yes	13	3
59	No	Travel_Rarely	1219	Sales	2	4	Life Sciences	1	23	1	Female	78	2	4	Manager	4	Married	15427	22021	2	Y	No	16	3
3	No	Travel_Rarely	371	Research & Development	2	3	Life Sciences	1	24	4	Male	45	3	1	Research	4	Single	3944	4306	5	Y	Yes	11	3
21	No	Non-Travel	673	Research & Development	11	2	Other	1	26	1	Female	96	4	2	Manufac	3	Divorced	4011	8232	0	Y	No	18	3
39	Yes	Travel_Rarely	1218	Sales	9	4	Life Sciences	1	27	3	Male	82	2	1	Sales Re	1	Single	3407	6986	7	Y	No	23	4
34	No	Travel_Rarely	419	Research & Development	7	4	Life Sciences	1	28	1	Female	53	3	3	Research	2	Single	11994	21293	0	Y	No	11	3
2	No	Travel_Rarely	391	Research & Development	15	2	Life Sciences	1	30	3	Male	96	3	1	Research	4	Single	1232	19281	1	Y	No	14	3
3	Yes	Travel_Rarely	699	Research & Development	6	1	Medical	1	31	2	Male	83	3	1	Research	1	Single	2960	17102	2	Y	No	11	3
5	No	Travel_Rarely	1282	Research & Development	5	3	Other	1	32	3	Female	58	3	5	Manager	3	Divorced	19094	10735	4	Y	No	11	3
3	Yes	Travel_Frequently	1125	Research & Development	16	1	Life Sciences	1	33	2	Female	72	1	1	Research	1	Single	3919	4681	1	Y	Yes	22	4
4	No	Travel_Rarely	691	Sales	8	4	Marketing	1	35	3	Male	48	3	2	Sales Exe	2	Married	6825	21173	0	Y	No	11	3
4	No	Travel_Rarely	477	Research & Development	7	4	Medical	1	36	1	Female	42	2	3	Healthca	4	Married	10248	2094	3	Y	No	14	3
4	No	Travel_Rarely	705	Sales	2	4	Marketing	1	38	2	Female	83	3	5	Manager	1	Single	18947	22822	3	Y	No	12	3
3	No	Travel_Rarely	924	Research & Development	2	3	Medical	1	39	3	Male	78	3	1	Laborato	4	Single	2496	6670	4	Y	No	11	3
4	No	Travel_Rarely	1459	Research & Development	10	4	Other	1	40	4	Male	41	3	2	Healthca	4	Married	6465	19121	2	Y	Yes	13	3
3	No	Travel_Rarely	125	Research & Development	9	2	Medical	1	41	4	Male	83	2	1	Laborato	3	Single	2206	16117	1	Y	No	13	3
39	Yes	Travel_Rarely	895	Sales	5	3	Technical Degree	1	42	4	Male	56	3	2	Sales Re	4	Married	2086	3335	3	Y	No	14	3
2	Yes	Travel_Rarely	813	Research & Development	1	3	Medical	1	45	2	Male	61	3	1	Research	4	Married	2293	3020	2	Y	Yes	16	3
42	No	Travel_Rarely	1273	Research & Development	2	2	Medical	1	46	4	Female	72	4	1	Research	3	Divorced	2645	21923	1	Y	No	12	3
5	Yes	Travel_Rarely	869	Sales	3	2	Marketing	1	47	1	Male	86	2	1	Sales Re	3	Married	2683	3810	1	Y	Yes	14	3
35	No	Travel_Rarely	890	Sales	2	3	Marketing	1	49	4	Female	97	3	1	Sales Re	4	Married	2014	9687	1	Y	No	13	3
3	No	Travel_Rarely	852	Research & Development	5	4	Life Sciences	1	51	2	Female	82	2	1	Research	1	Married	3419	13072	9	Y	Yes	14	3
33	No	Travel_Frequently	1141	Sales	1	3	Life Sciences	1	52	3	Female	42	4	2	Sales Exe	1	Married	5376	3193	2	Y	No	19	3
3	No	Travel_Rarely	464	Research & Development	4	2	Other	1	53	3	Male	75	3	1	Laborato	4	Divorced	1951	10910	1	Y	No	12	3
27	No	Travel_Rarely	1240	Research & Development	2	4	Life Sciences	1	54	4	Female	33	3	1	Laborato	1	Divorced	2341	19715	1	Y	No	13	3
2	Yes	Travel_Rarely	1357	Research & Development	25	3	Life Sciences	1	55	1	Male	48	1	1	Laborato	3	Single	2293	10558	1	Y	No	12	3

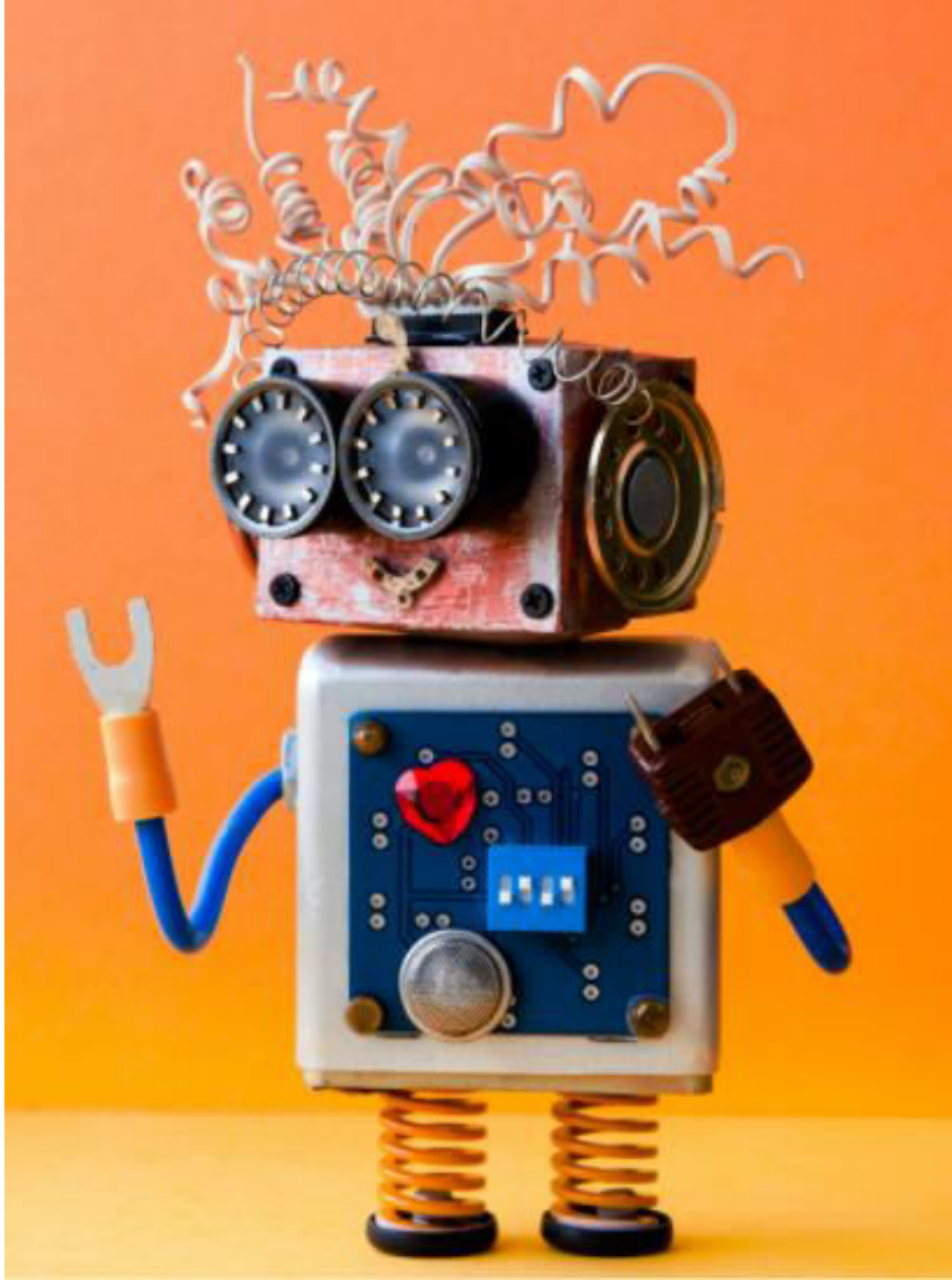
MODEL OUTPUT: ATTRITION (BINARY VALUE) (YES OR NO)



PROJECT DEMO



WHAT IS SAGEMAKER CANVAS?



WHAT IS AWS SAGEMAKER CANVAS?

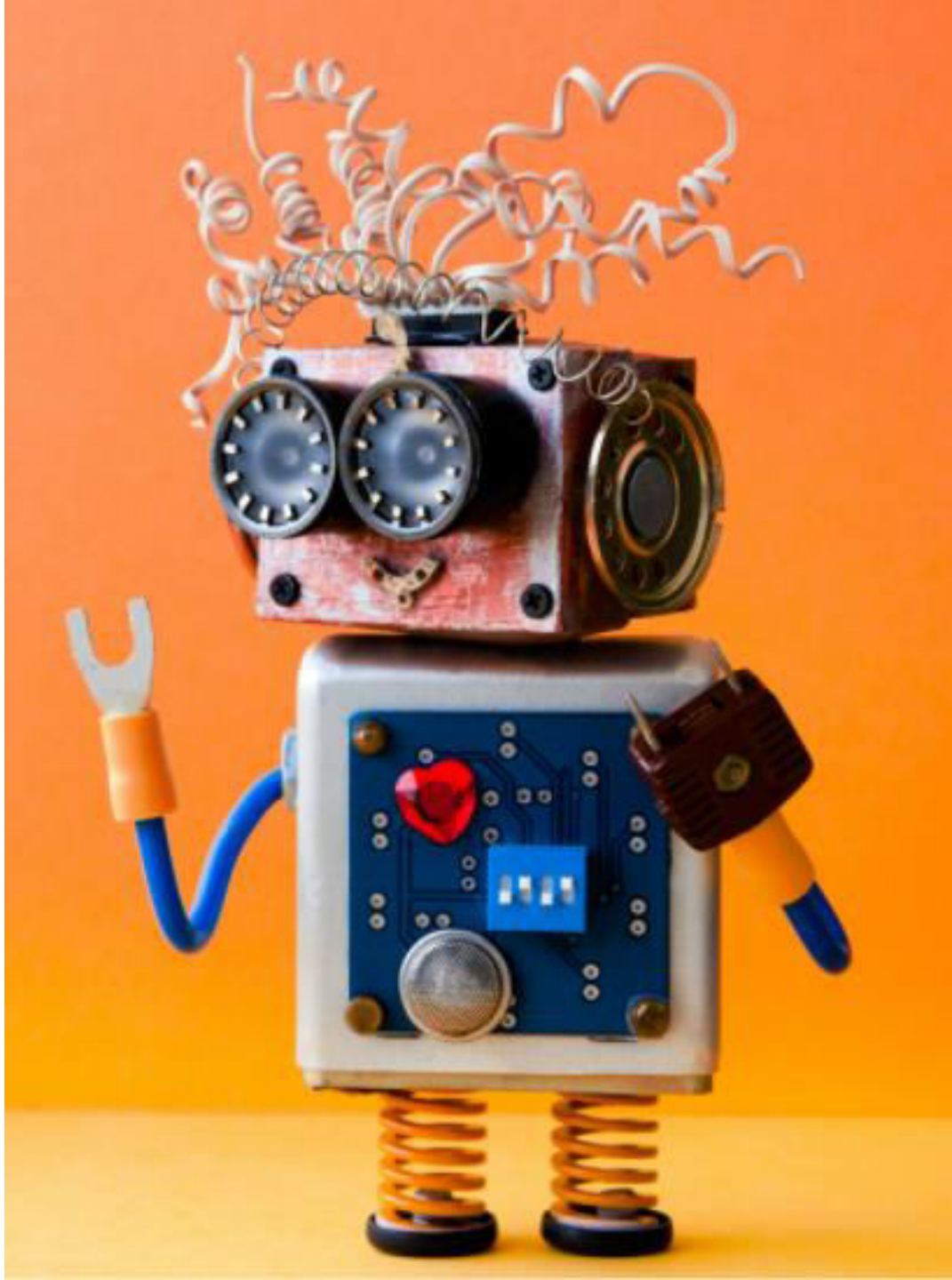
- AWS SageMaker Canvas empowers anyone to build, train and test a machine learning model without writing a single line of code!
- With AWS Canvas, anyone can:
  - Import data from S3 or any other source
  - Build an AI/ML model
  - Assess model performance
  - Perform inference and generate predictions
  - Export model to SageMaker Studio

- AWS SageMaker Canvas Documentation: https://aws.amazon.com/sagemaker/canvas/

SUCCESS STORIES

EASY

ADVANCED



BUSINESS CASE AND SUCCESS STORIES

- Hiring/retaining employees is an expensive process that requires capital, time and skills.
- Average companies lose 1% - 2.5% of their total revenue on the time it takes to bring a new hire up to speed.
- Companies spend 15%-20% of the employee's salary to recruit a new candidate.
- 7 ways AI is reinventing Human Resources: https://www.cmswire.com/digital-workplace/7-ways-artificial-intelligence-is-reinventing-human-resources/

BUSINESS CASE AND SUCCESS STORIES

EXTRACT INFORMATION FROM CANDIDATES RESUMES USING AI

ASSESS EMPLOYEE REFERRALS

GAIN KEY DATA INSIGHTS

IMPROVED ENGAGMENT USING AI-POWERED CHATBOTS

CUSTOMIZED LEARNING JOURNEY FOR EMPLOYEES

PREDICT EMPLOYEES' RETENTION, POTENTIAL AND FATIGUE

ENHANCED WORKPLACE ANALYTICS

https://www.cmswire.com/digital-workplace/7-ways-artificial-intelligence-is-reinventing-human-resources/

READING TIME & QUIZ – HOW AI IS TRANSFORMING HUMAN RESOURCES?

- Please read the article below and answer the following quiz.
  - Link to Article: https://www.business-standard.com/article/jobs/employee-attrition-how-ai-is-transforming-human-resource-practices-118062701396_1.html

10 MINS

5 MINS

Employee attrition? How AI is transforming human resource practices

According to the IT major, the use of the Watson platform has also made talent acquisition much more accurate and seamless

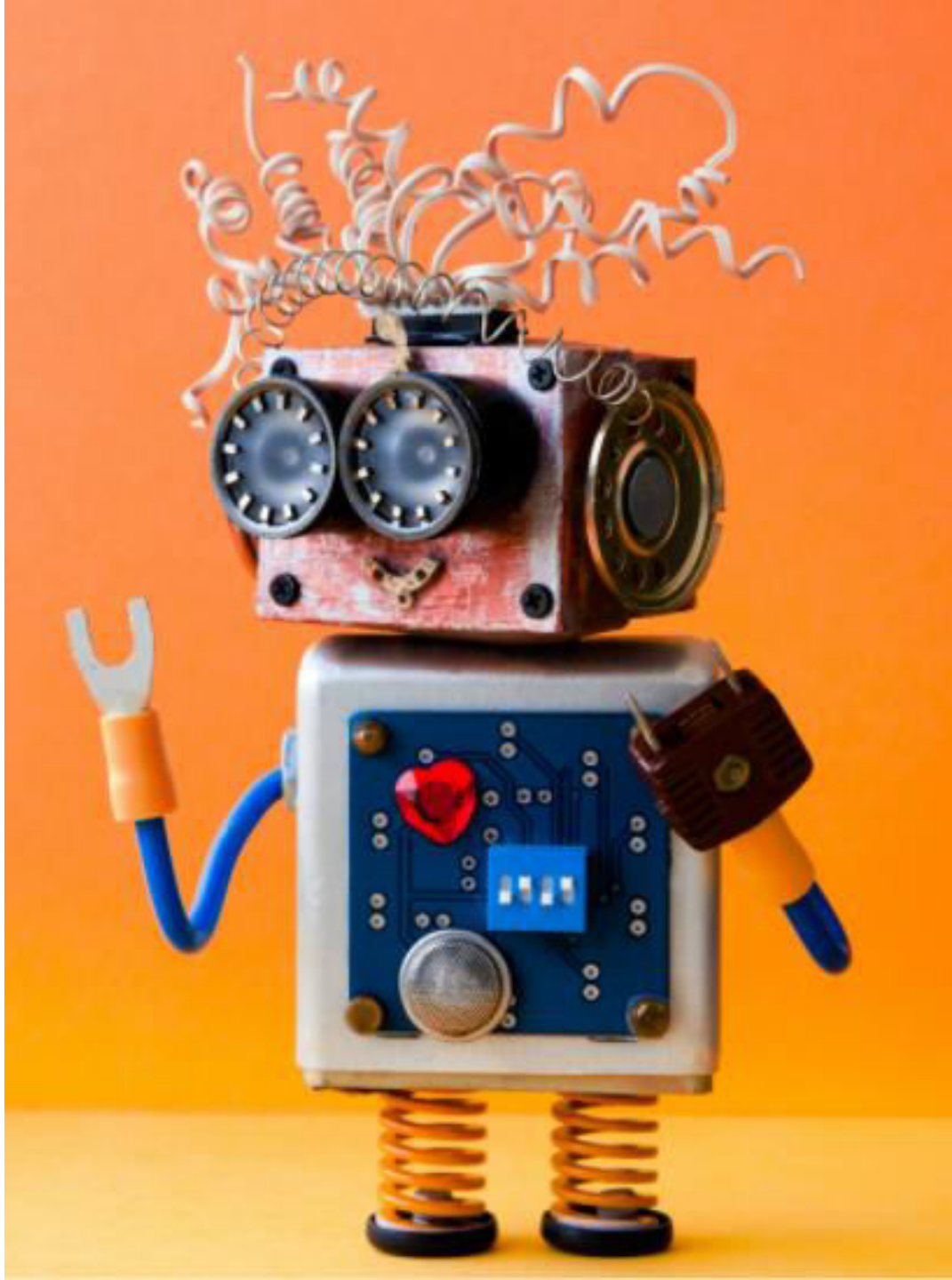
Topics

Automation | Human Resources | Ibm

Bibhu Ranjan Mishra & Debasis Mohapatra

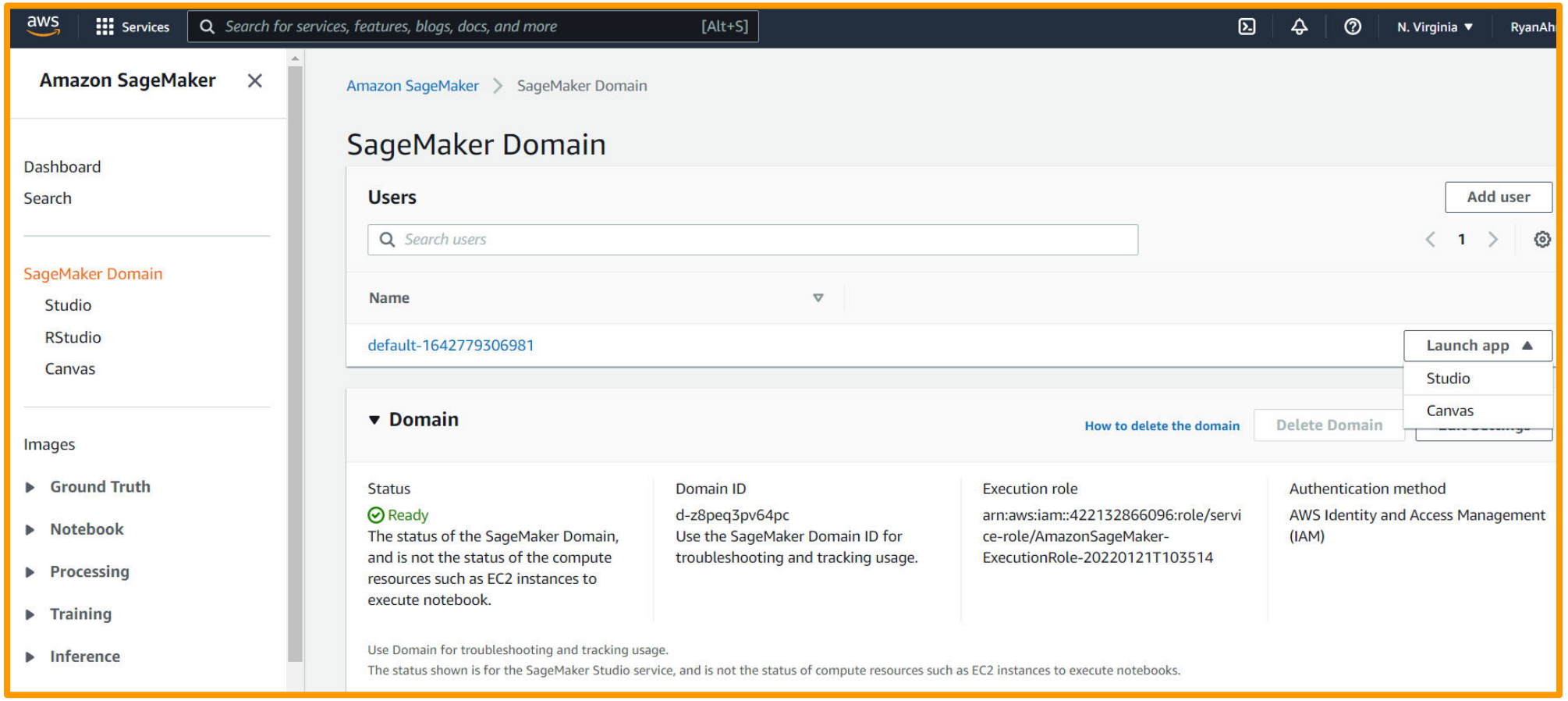
Last Updated at June 28, 2018 09:02 IST

DEMO PART 1: LAUNCH CANVAS & UPLOAD DATA TO S3



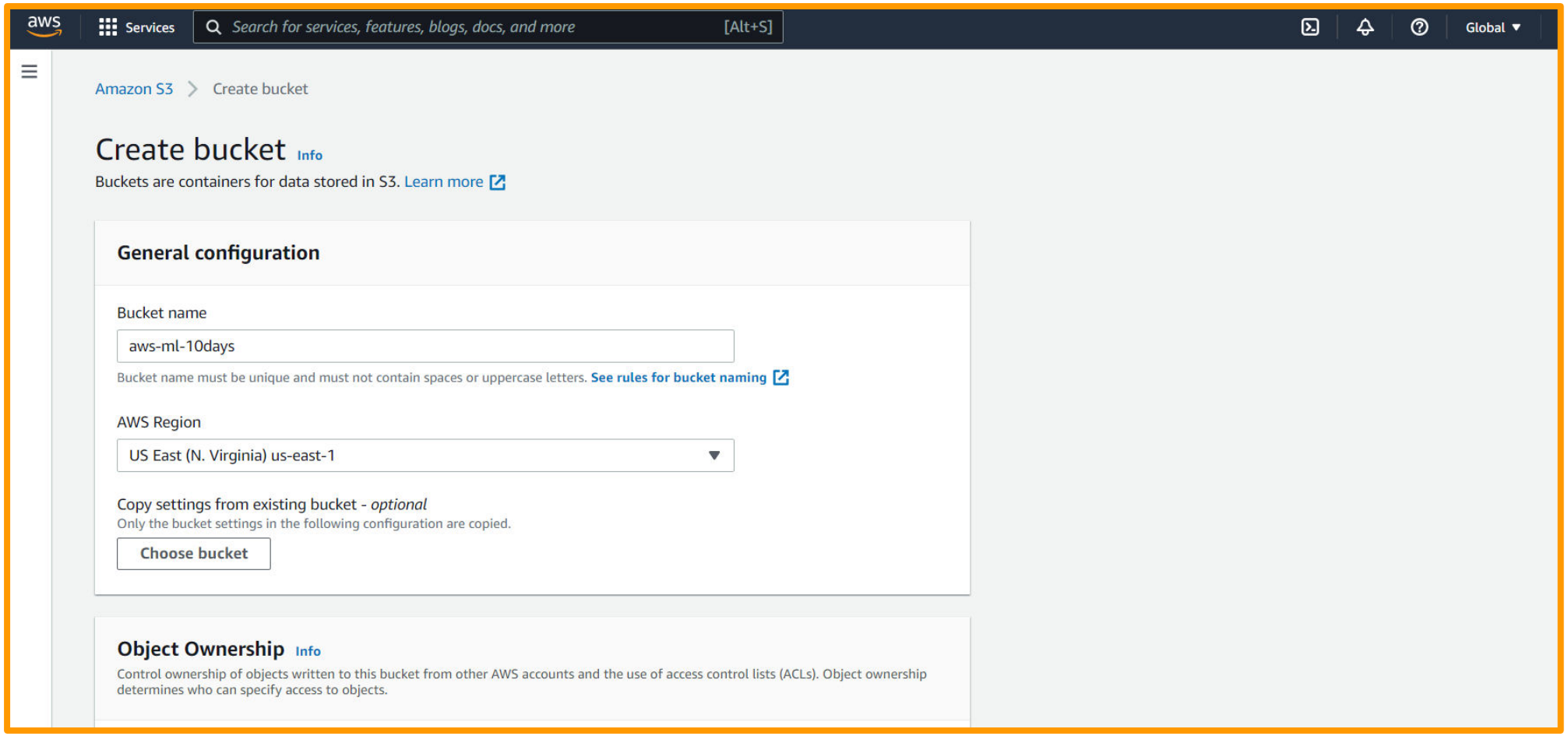
AWS SAGEMAKER CANVAS SETUP

AFTER CREATING A SAGEMAKER DOMAIN, CLICK ON LAUNCH APP AND THEN CANVAS



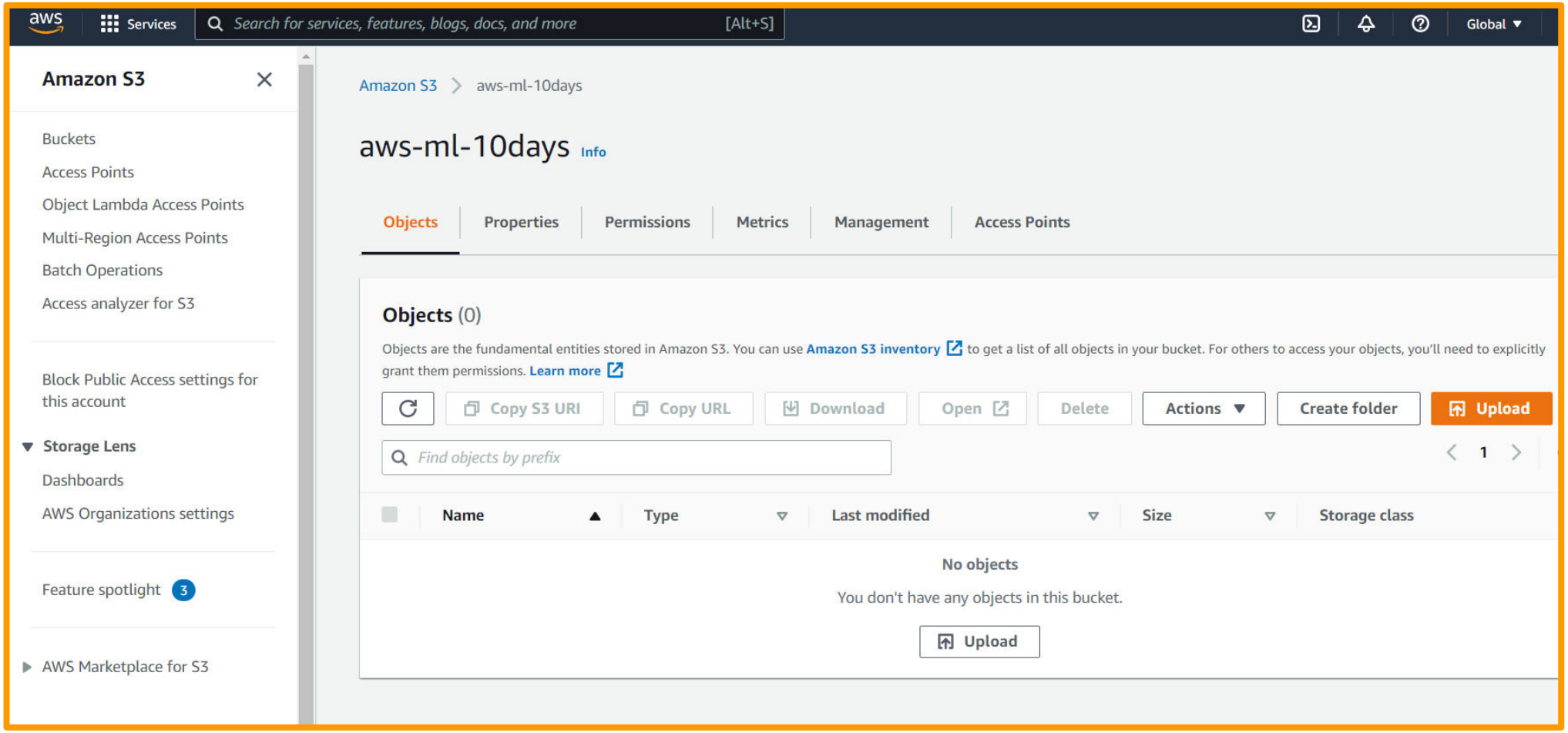
AWS SAGEMAKER CANVAS SETUP

CREATE A NEW S3 BUCKET BEFORE YOU LAUNCH THE AWS SAGEMAKER CANVAS



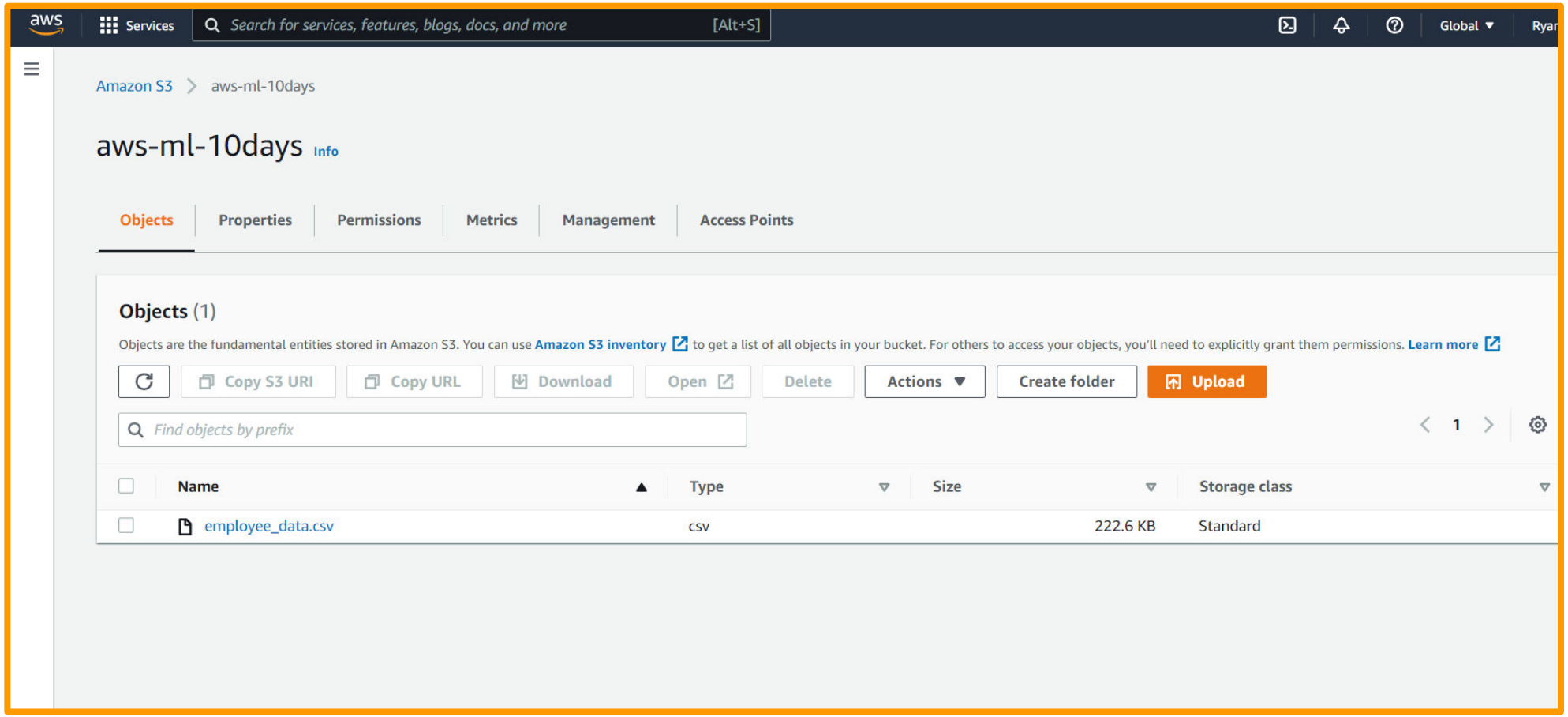
AWS SAGEMAKER CANVAS SETUP

UPLOAD THE DATA TO THE S3 BUCKET

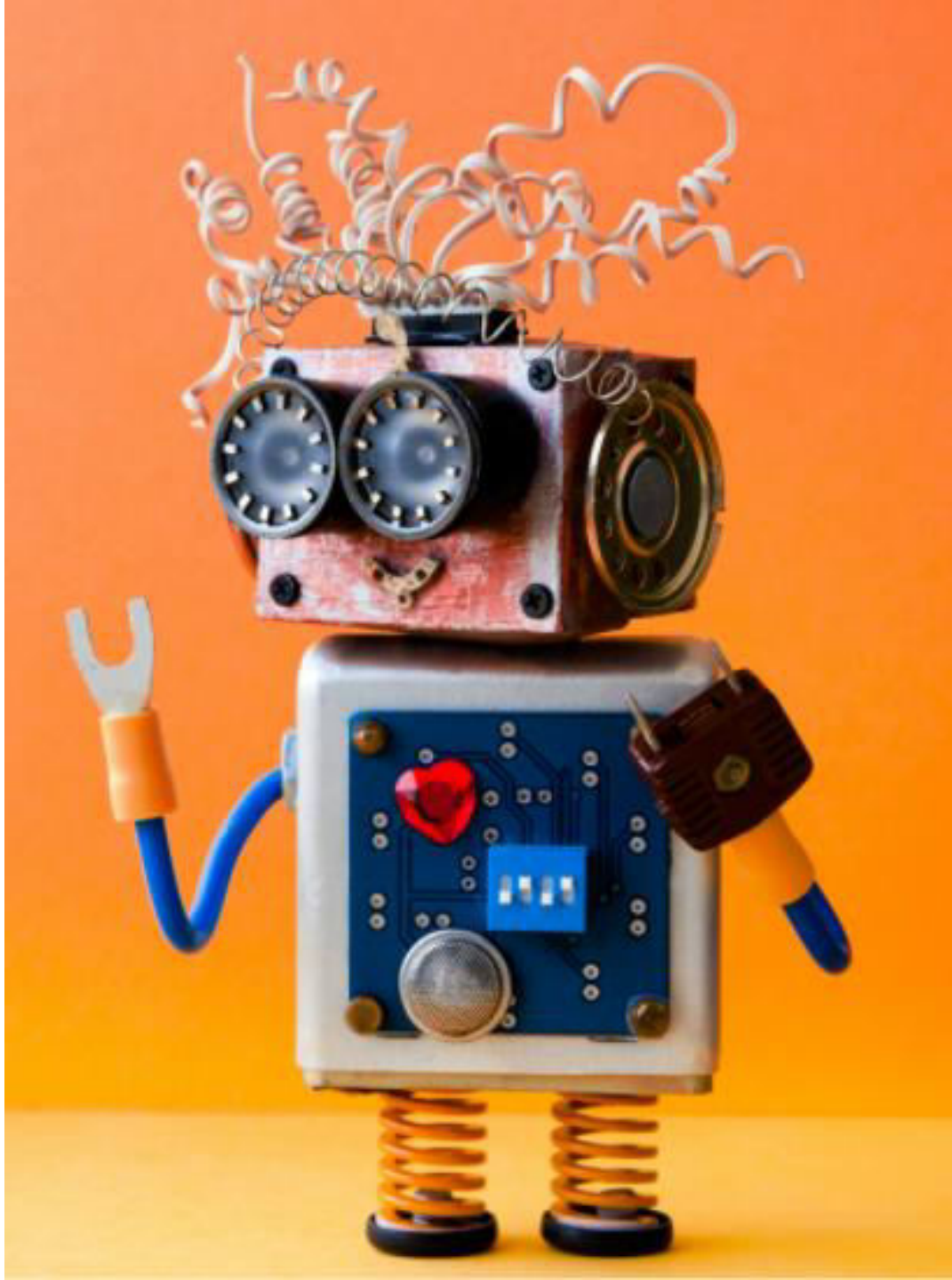


AWS SAGEMAKER CANVAS SETUP

EMPLOYEE_DATA.CSV IS NOW UPLOADED TO THE S3 BUCKET

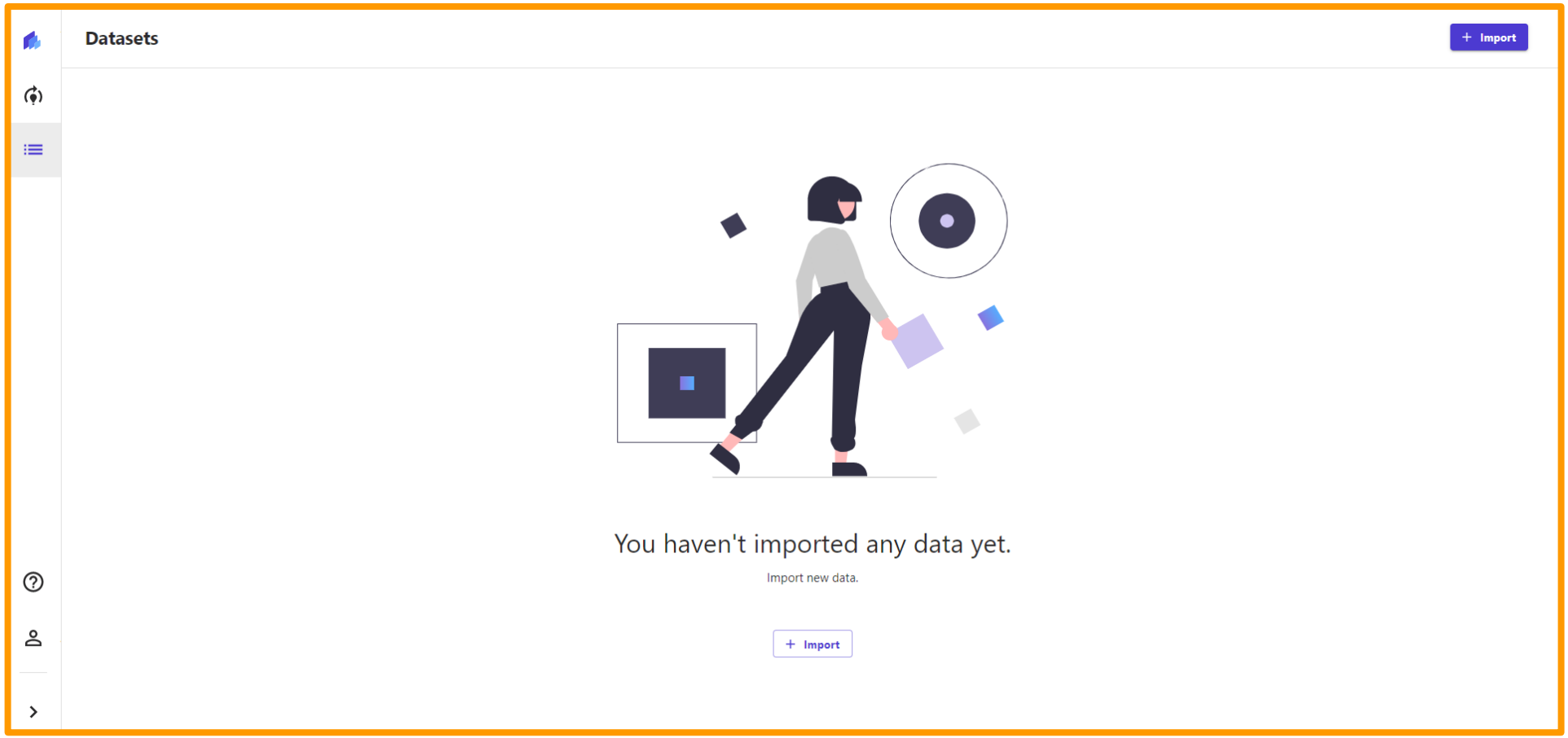


DEMO PART 2: DATA IMPORT INTO CANVAS & MODEL TRAINING



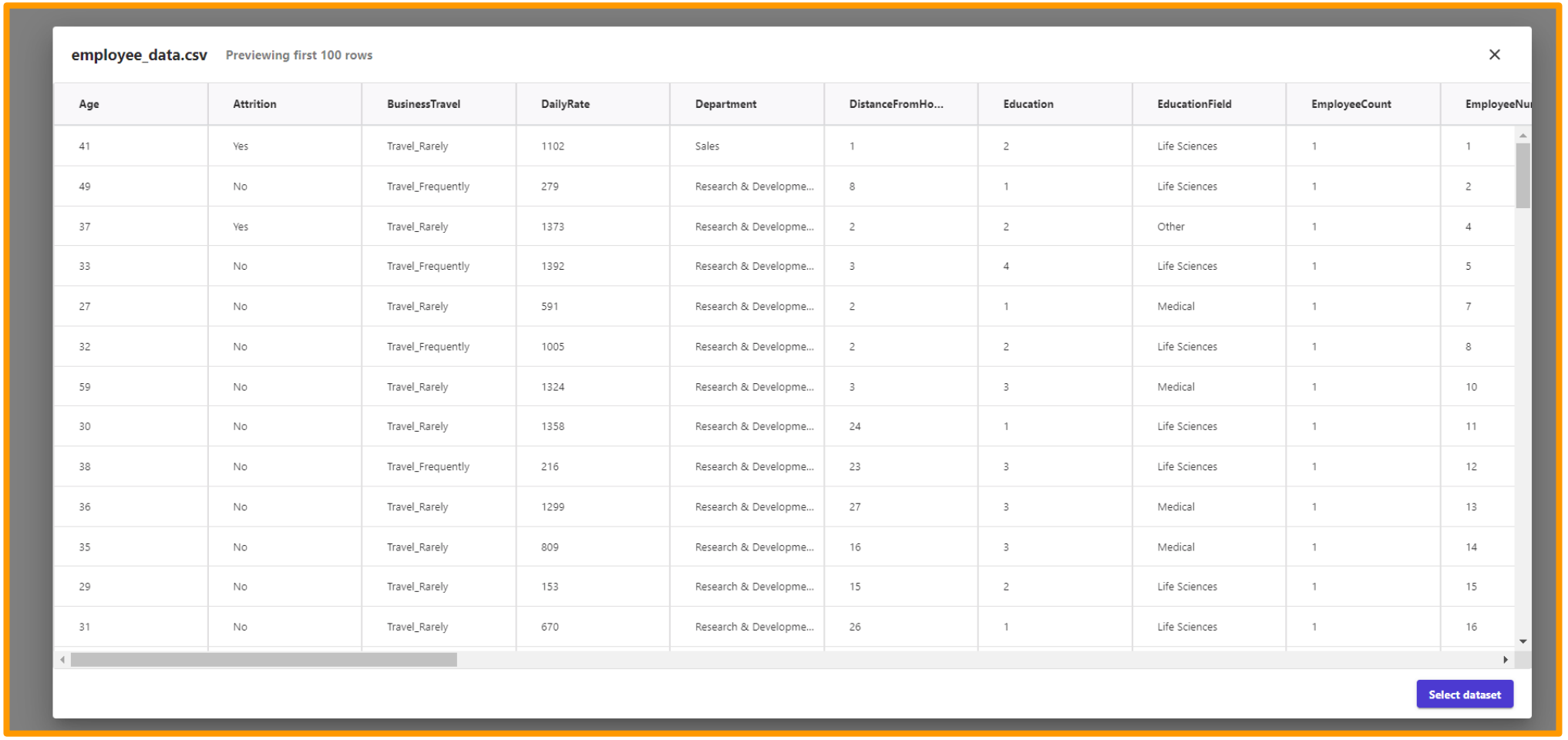
AWS SAGEMAKER CANVAS

FIRST YOU NEED TO IMPORT THE DATASET
FROM S3, CLICK ON IMPORT



AWS SAGEMAKER CANVAS

SELECT THE BUCKET FROM S3. VIEW AND IMPORT THE DATASET



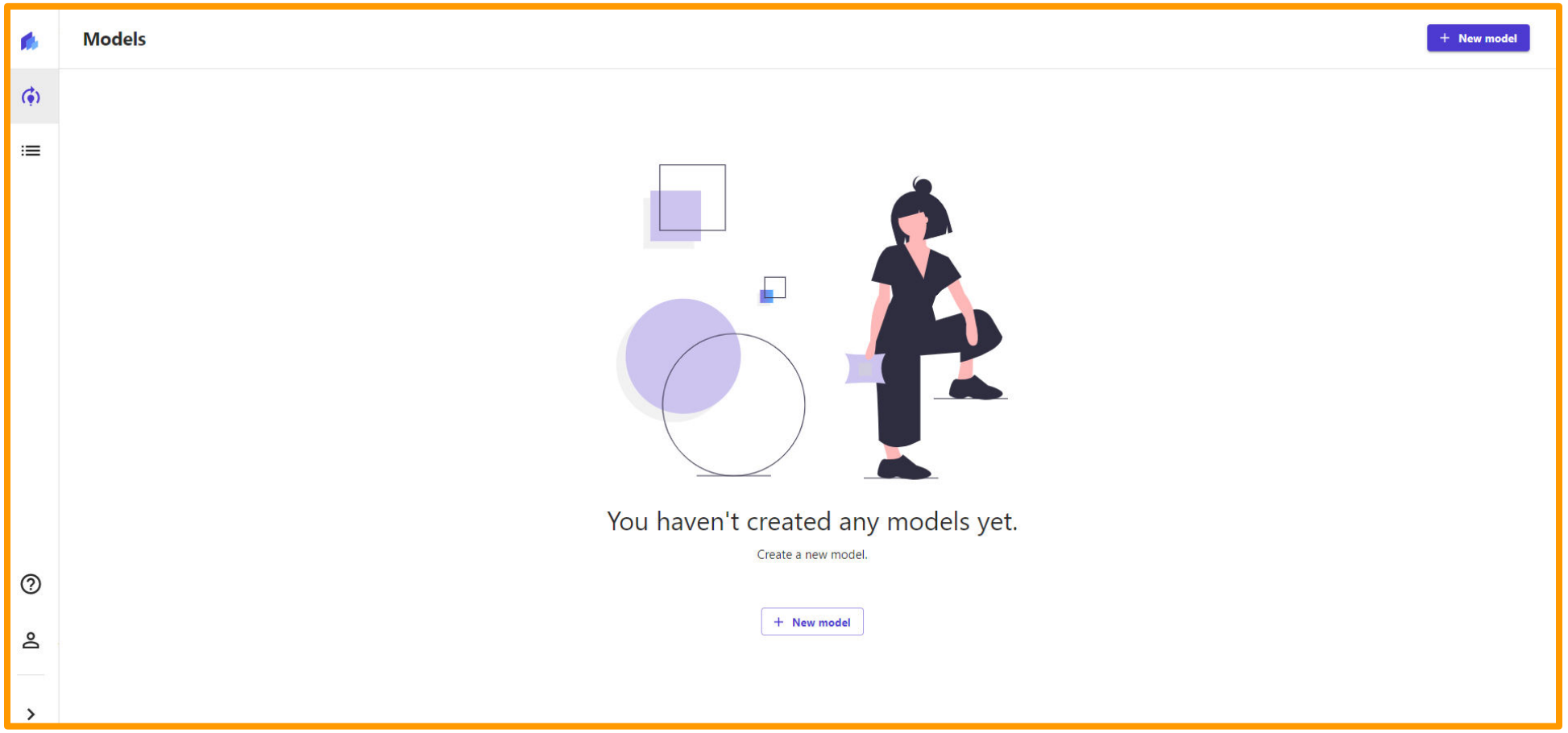
employee_data.csv Previewing first 100 rows

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHo...	Education	EducationField	EmployeeCount	EmployeeNu
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
49	No	Travel_Frequently	279	Research & Developme...	8	1	Life Sciences	1	2
37	Yes	Travel_Rarely	1373	Research & Developme...	2	2	Other	1	4
33	No	Travel_Frequently	1392	Research & Developme...	3	4	Life Sciences	1	5
27	No	Travel_Rarely	591	Research & Developme...	2	1	Medical	1	7
32	No	Travel_Frequently	1005	Research & Developme...	2	2	Life Sciences	1	8
59	No	Travel_Rarely	1324	Research & Developme...	3	3	Medical	1	10
30	No	Travel_Rarely	1358	Research & Developme...	24	1	Life Sciences	1	11
38	No	Travel_Frequently	216	Research & Developme...	23	3	Life Sciences	1	12
36	No	Travel_Rarely	1299	Research & Developme...	27	3	Medical	1	13
35	No	Travel_Rarely	809	Research & Developme...	16	3	Medical	1	14
29	No	Travel_Rarely	153	Research & Developme...	15	2	Life Sciences	1	15
31	No	Travel_Rarely	670	Research & Developme...	26	1	Life Sciences	1	16

Select dataset

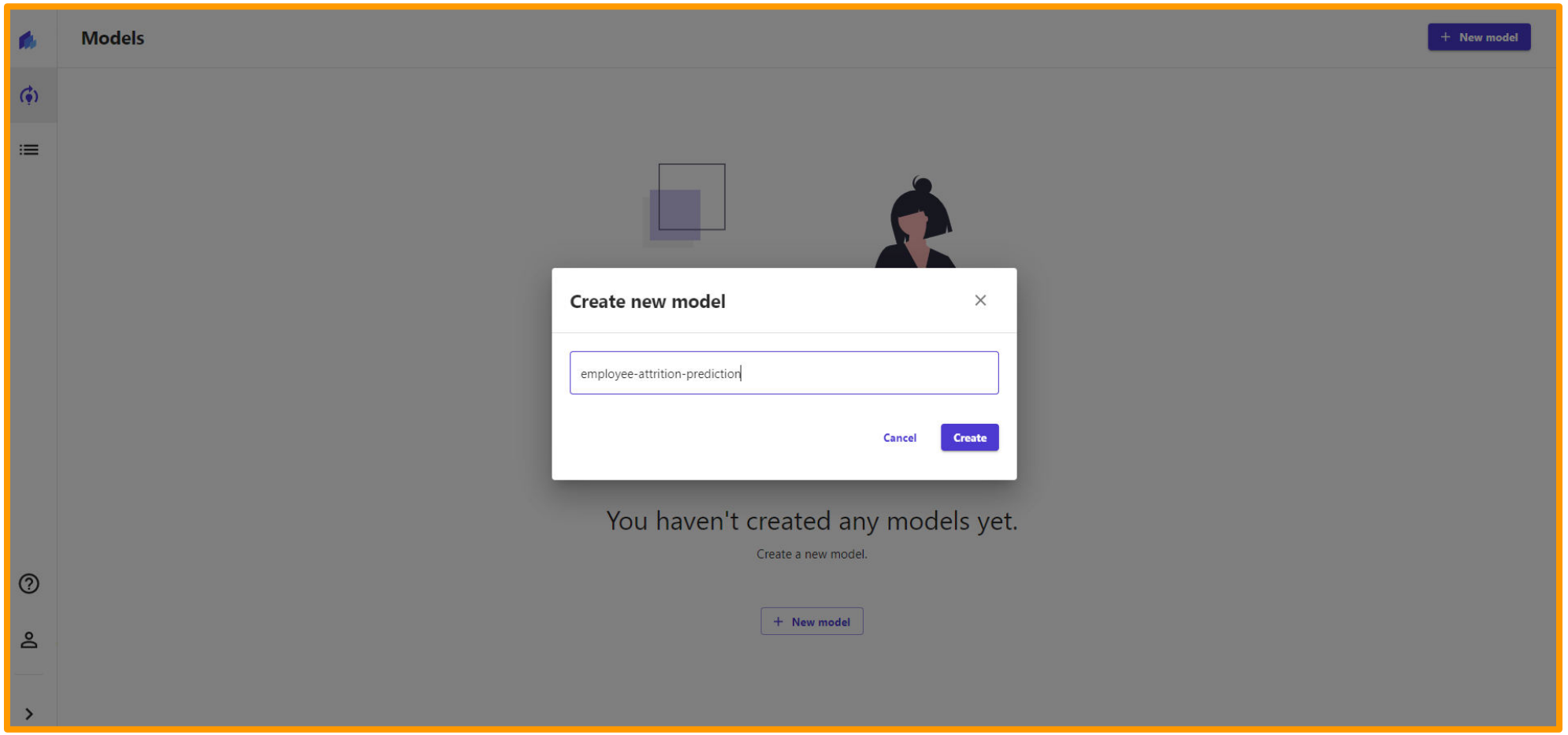
AWS SAGEMAKER CANVAS

CLICK ON MODELS AND CREATE A NEW MODEL AND CLICK CREATE



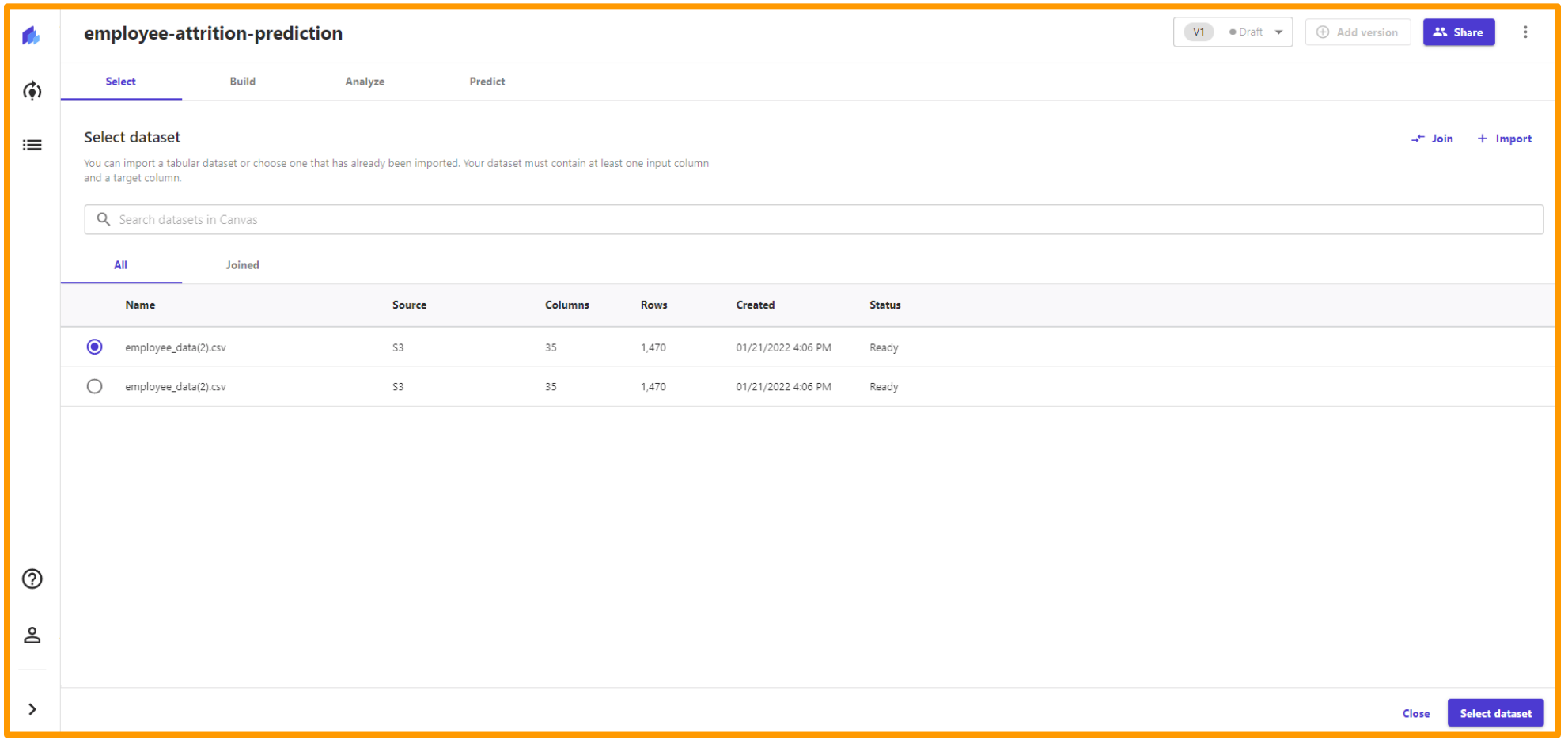
AWS SAGEMAKER CANVAS

GIVE A NAME TO THE MODEL



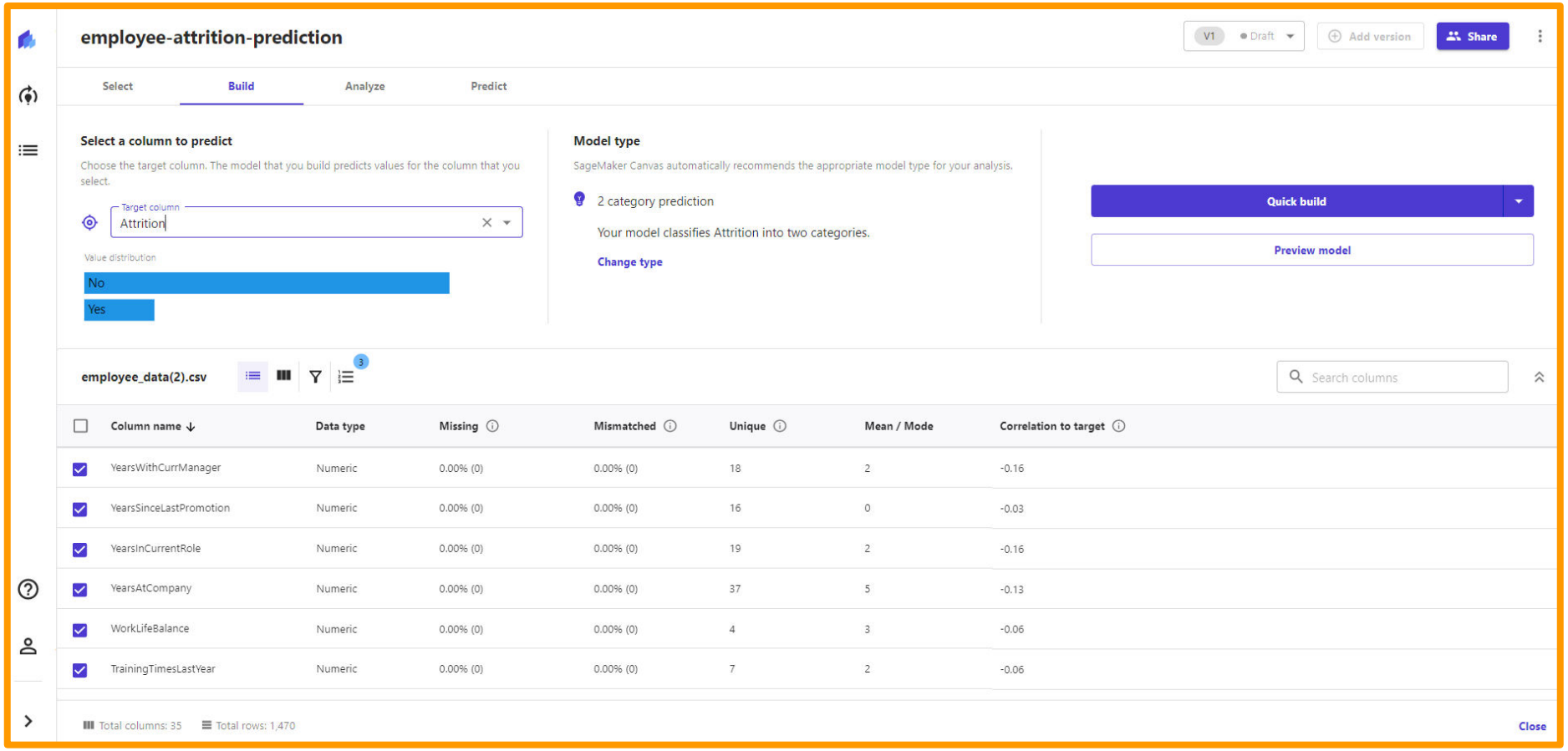
AWS SAGEMAKER CANVAS

SELECT THE DATASET



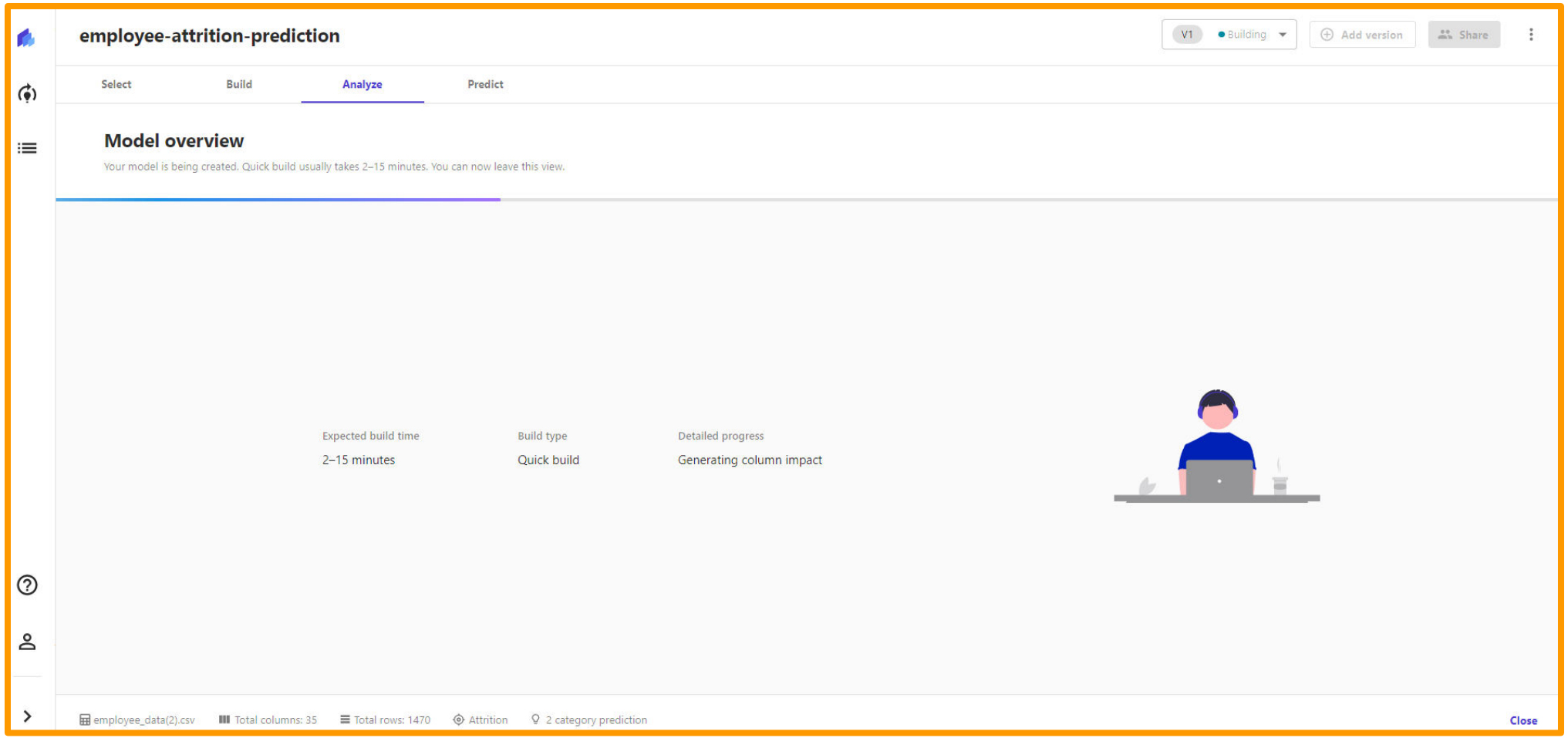
AWS SAGEMAKER CANVAS

SELECT THE TARGET COLUMN “ATTRITION”, CLICK ON “QUICK MODEL”

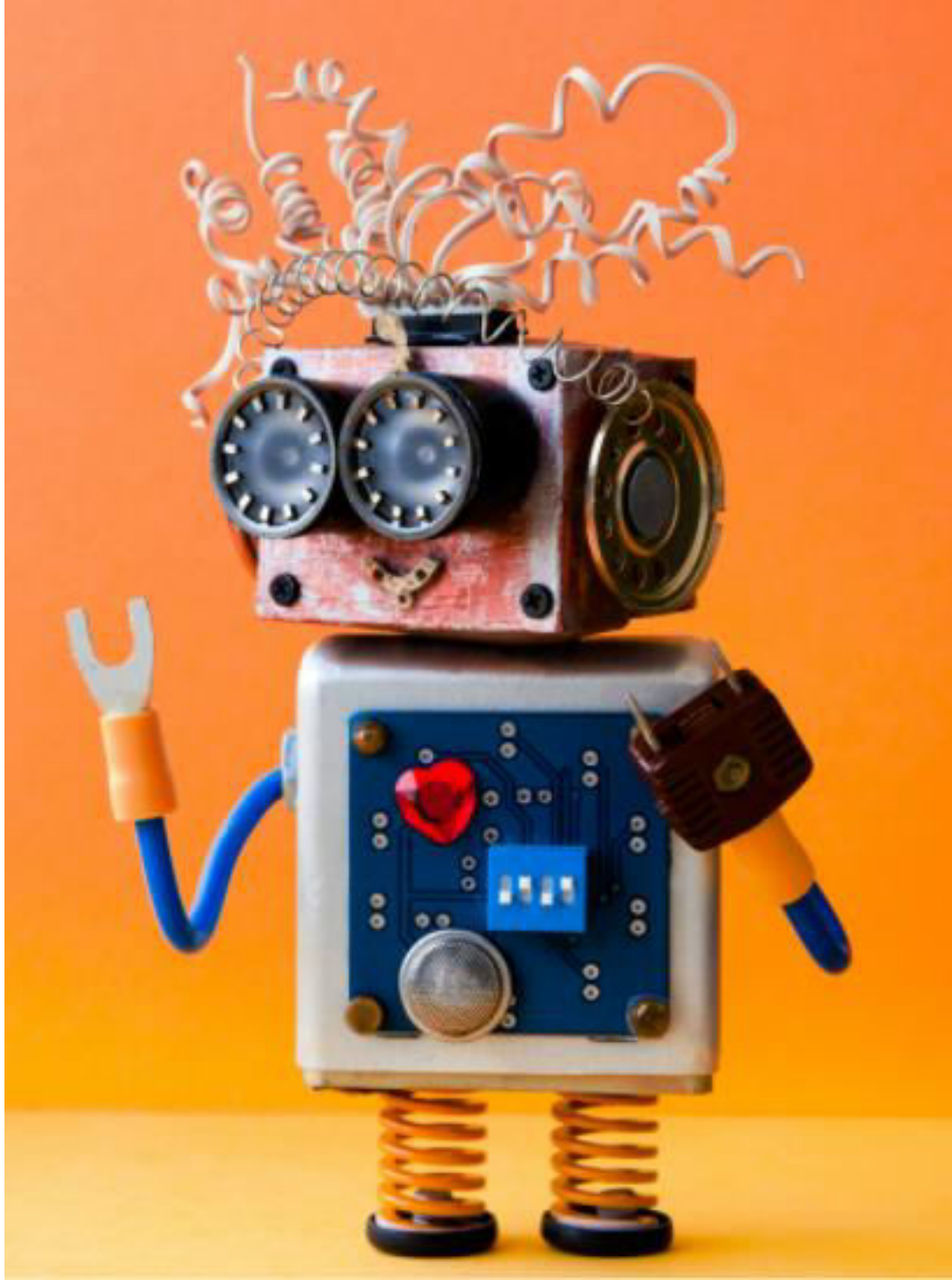


AWS SAGEMAKER CANVAS

QUICK MODEL TRAINING SHOULD TAKE 2-15 MINS TO TRAIN



DEMO PART 3: TRAINED MODEL ASSESSMENT

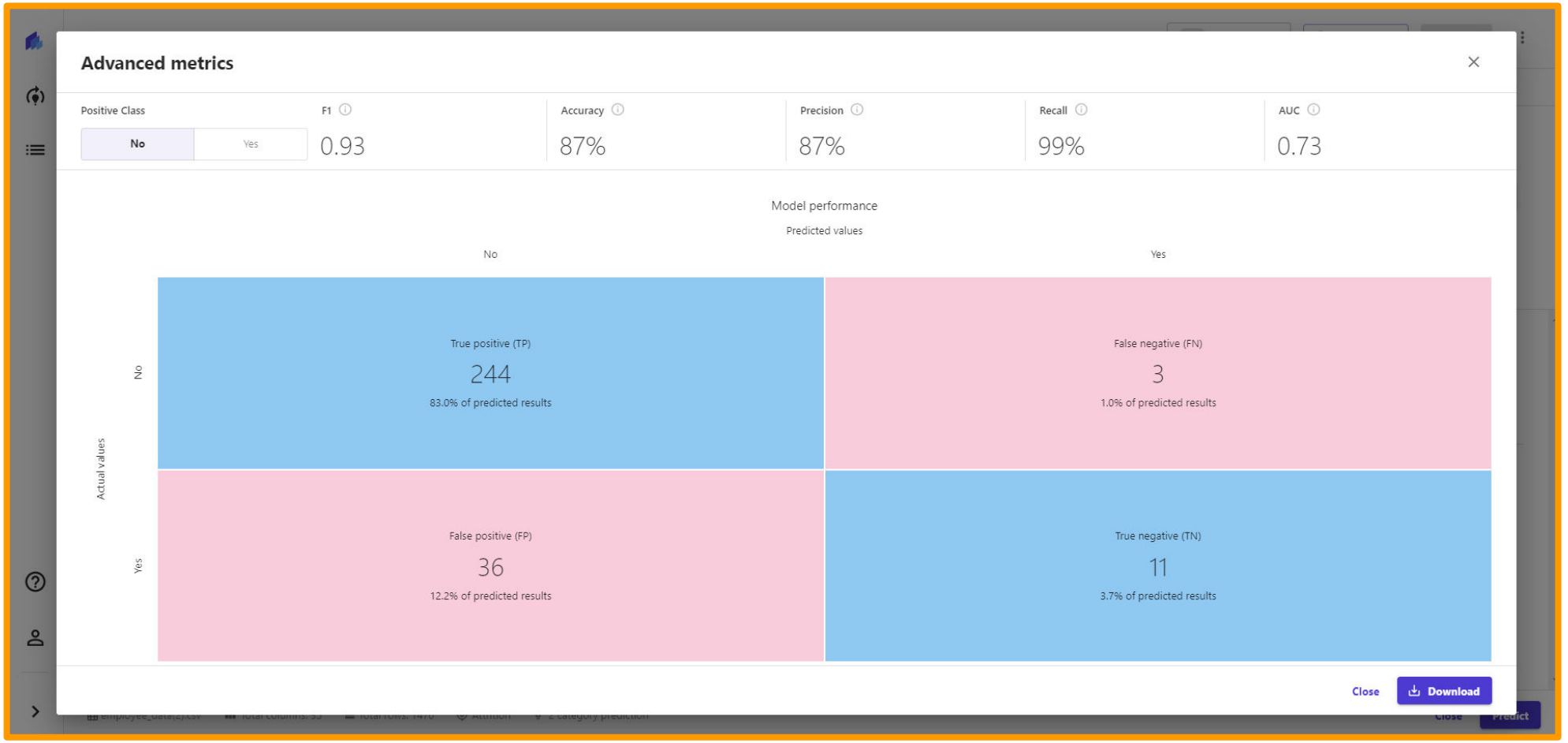


AWS SAGEMAKER CANVAS

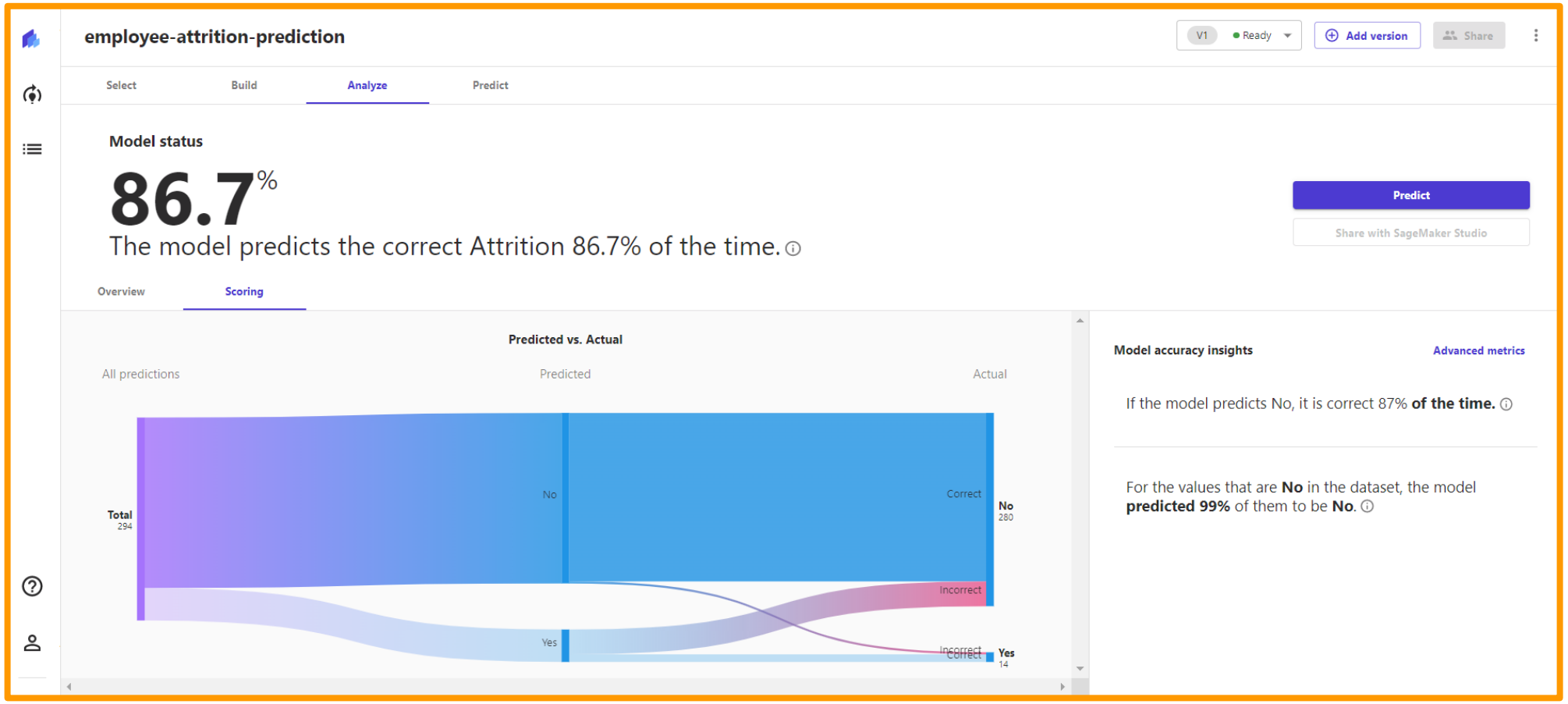
AFTER THE MODEL IS TRAINED, IT'S NOW READY TO BE ANALYZED! CHECK OUT THE FEATURE IMPORTANCE (IMPACT)

AWS SAGEMAKER CANVAS

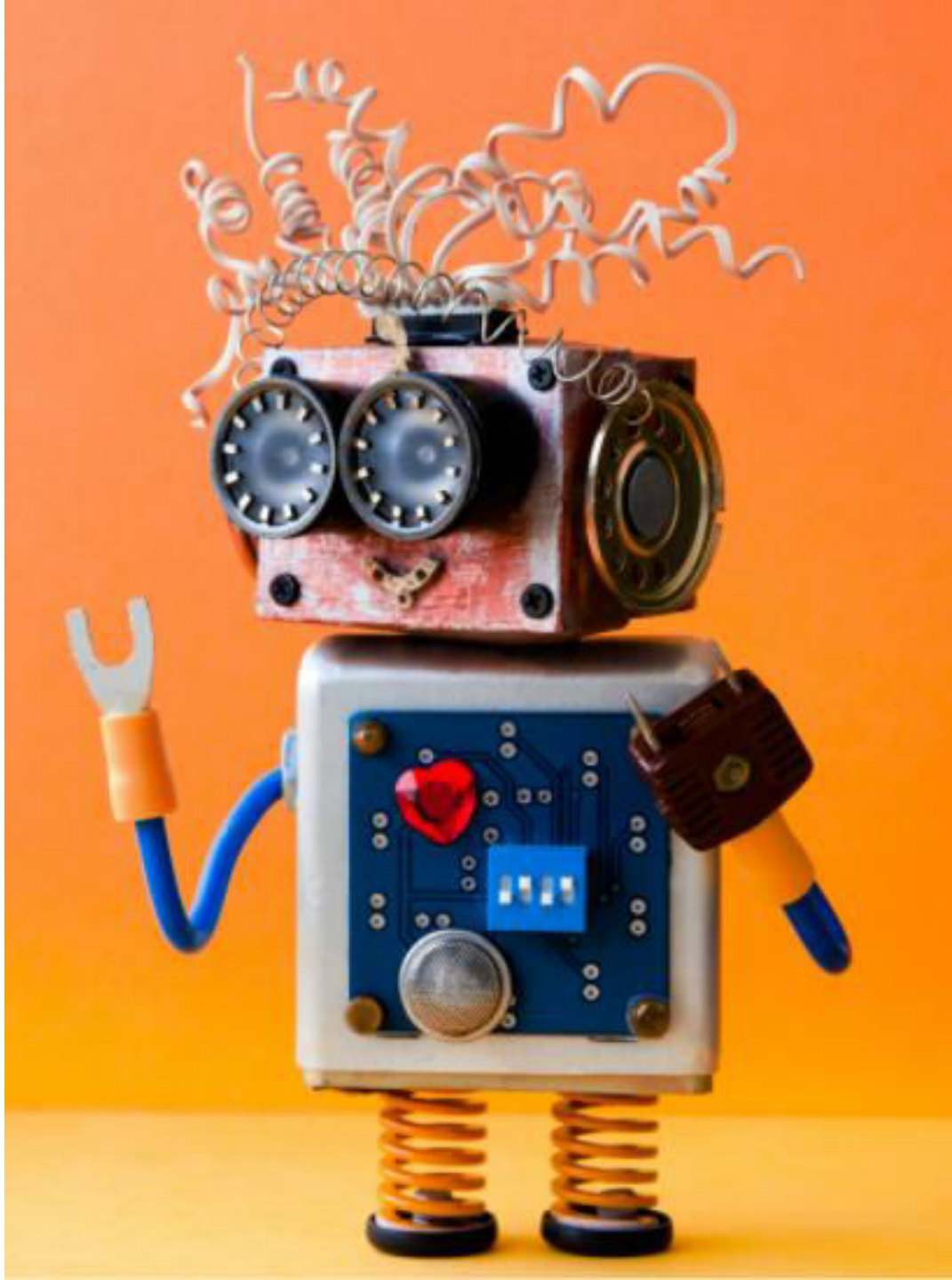
CONFUSION MATRIX



AWS SAGEMAKER CANVAS

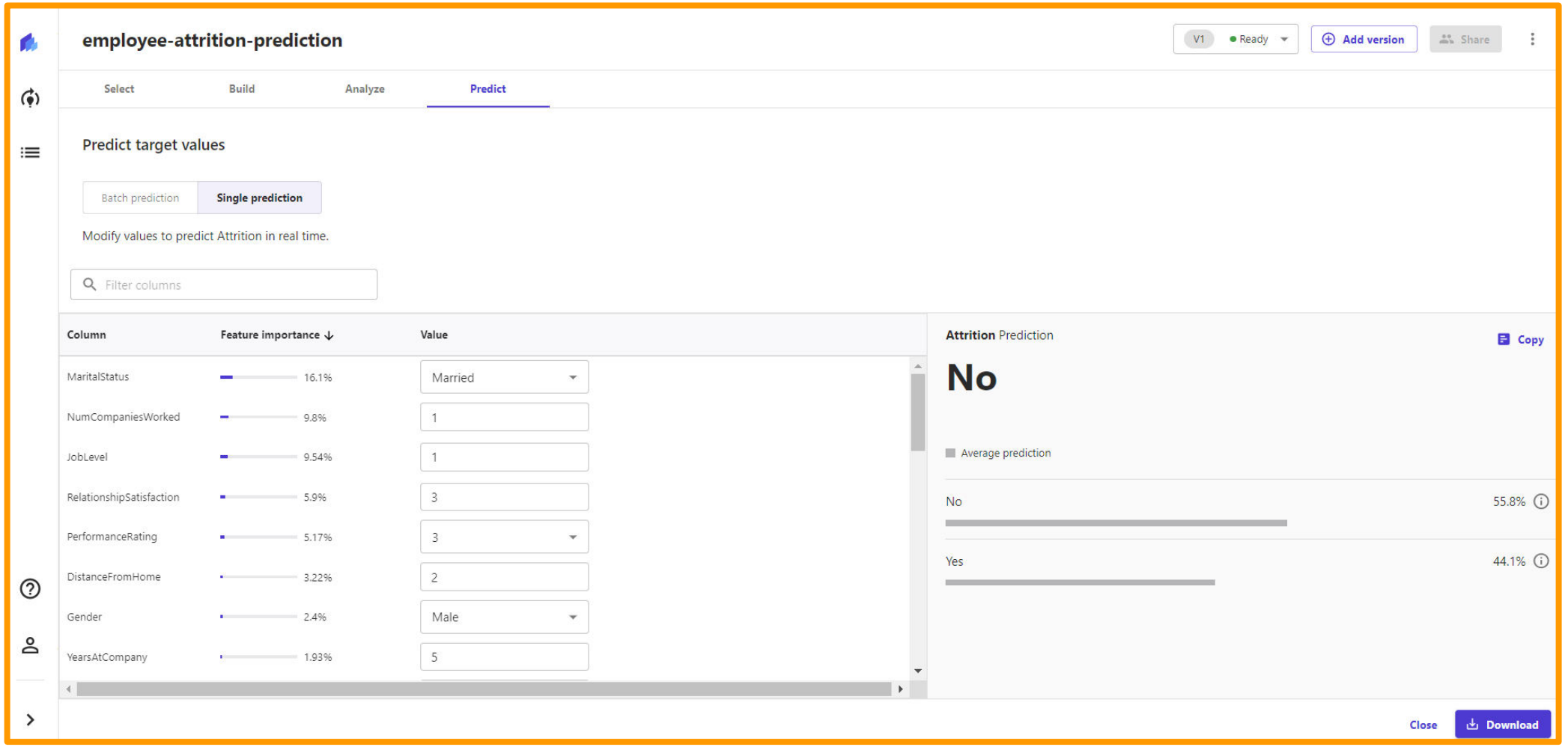


DEMO PART 4: PERFORM MODEL INFERENCE



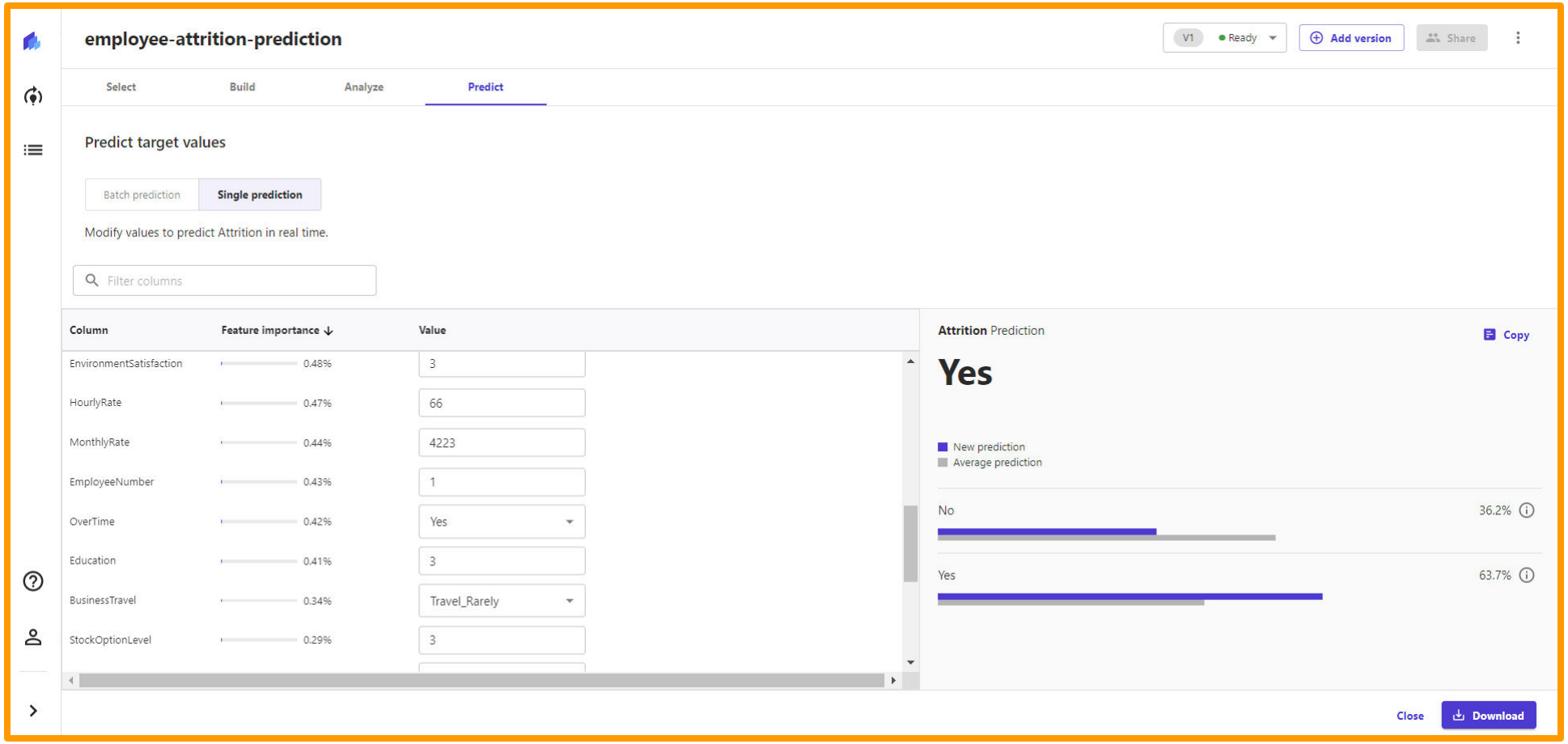
AWS SAGEMAKER CANVAS

SELECT INPUTS AND GENERATE OUTPUT



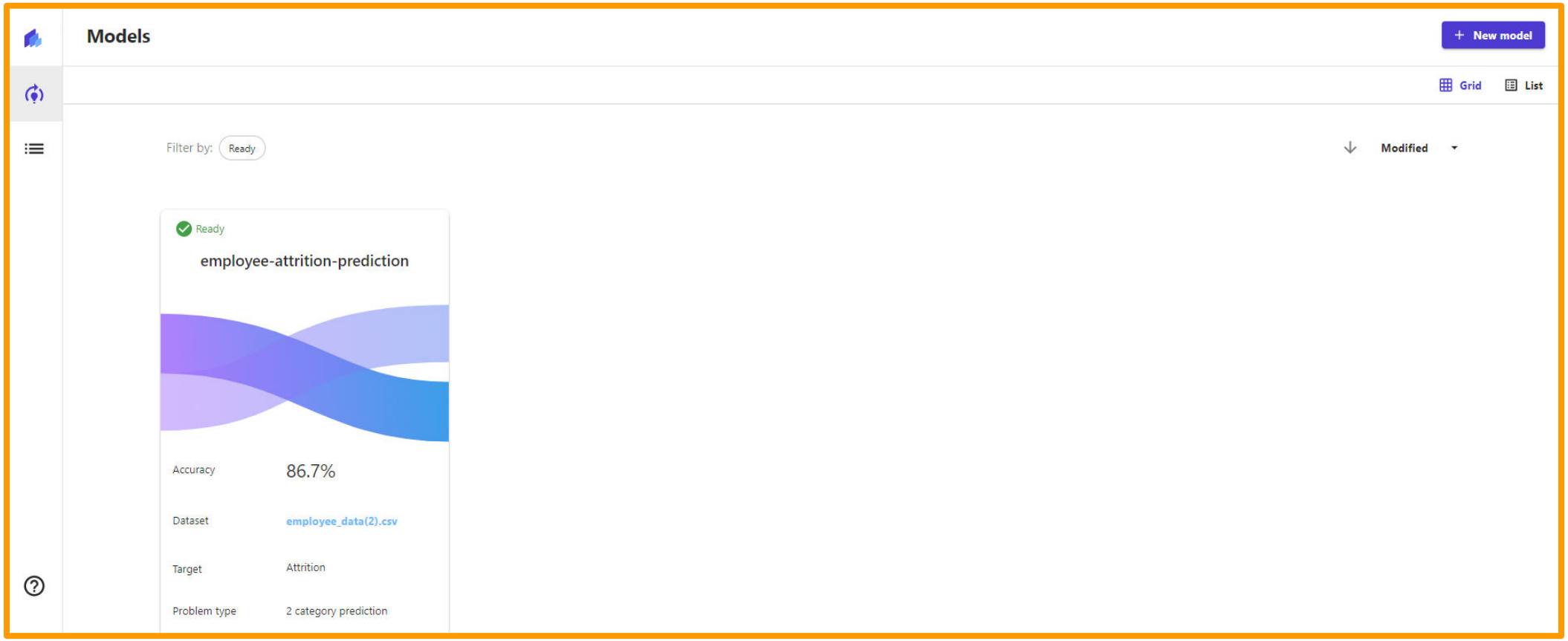
AWS SAGEMAKER CANVAS

UPDATE THE OVERTIME AND CLICK UPDATE



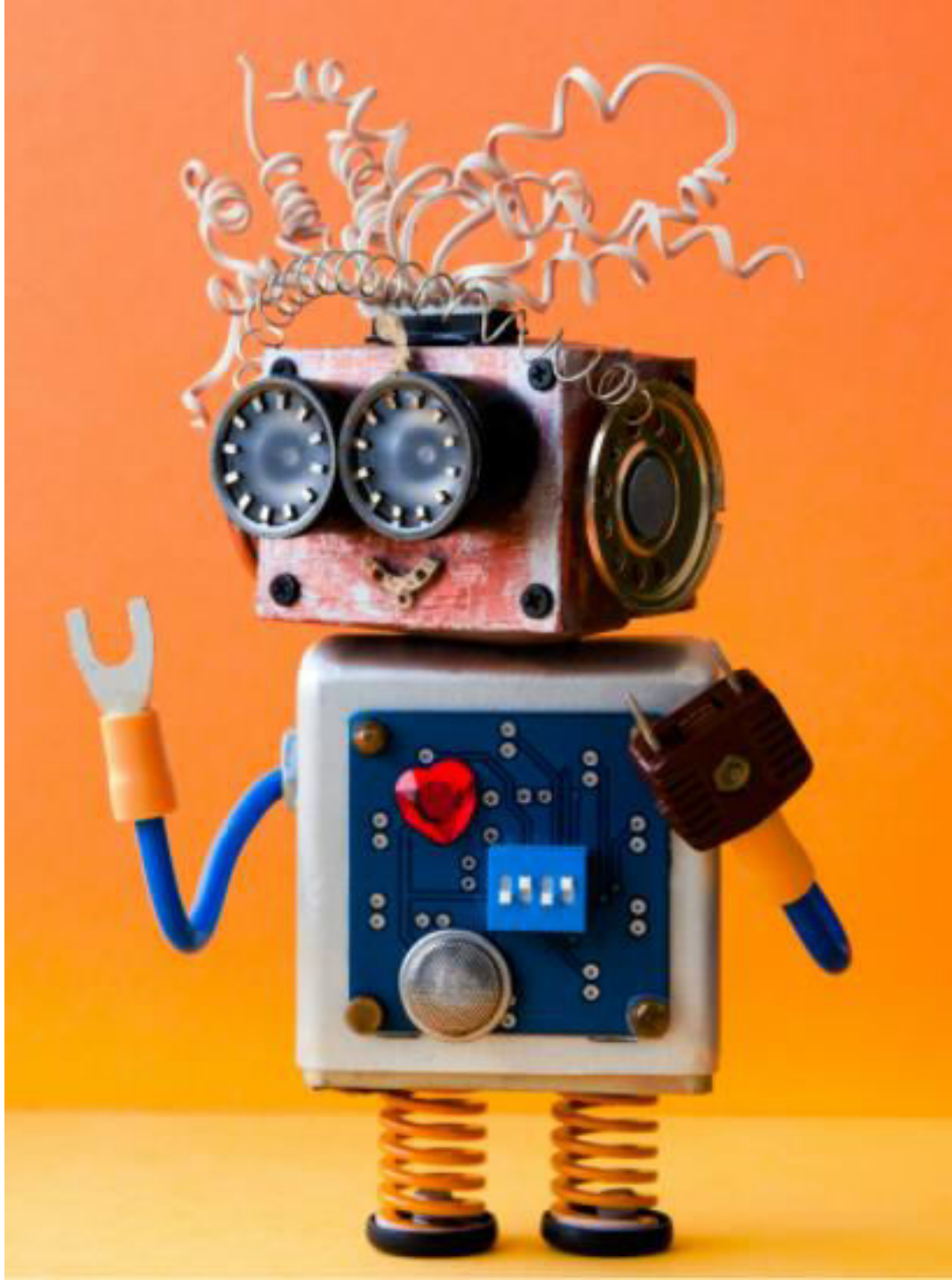
AWS SAGEMAKER CANVAS

YOU SHOULD SEE THE TRAINED MODEL
UNDER THE MODELS TAB



FINAL END-OF-DAY CAPSTONE PROJECT

EASY ADVANCED

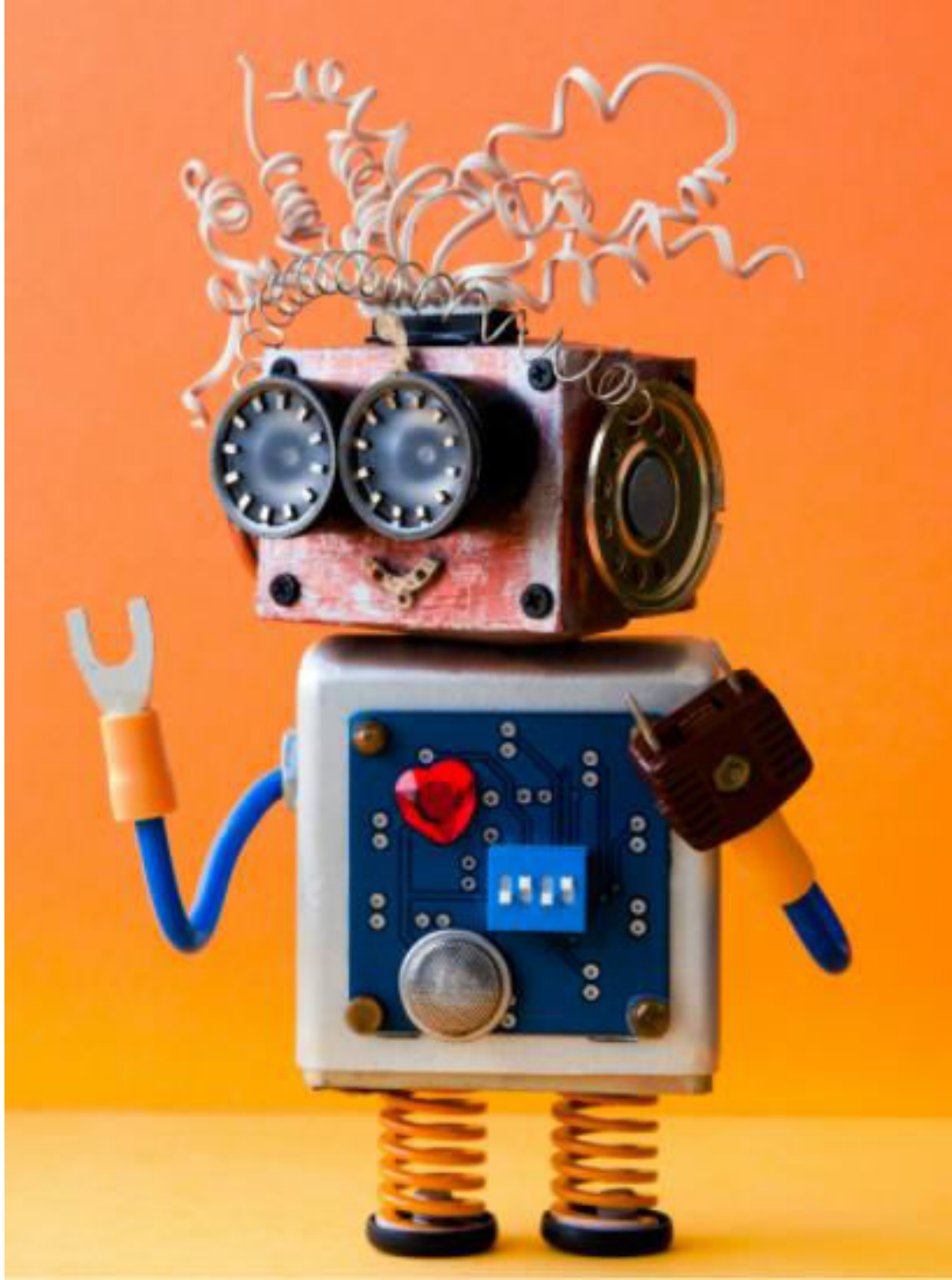


PROJECT

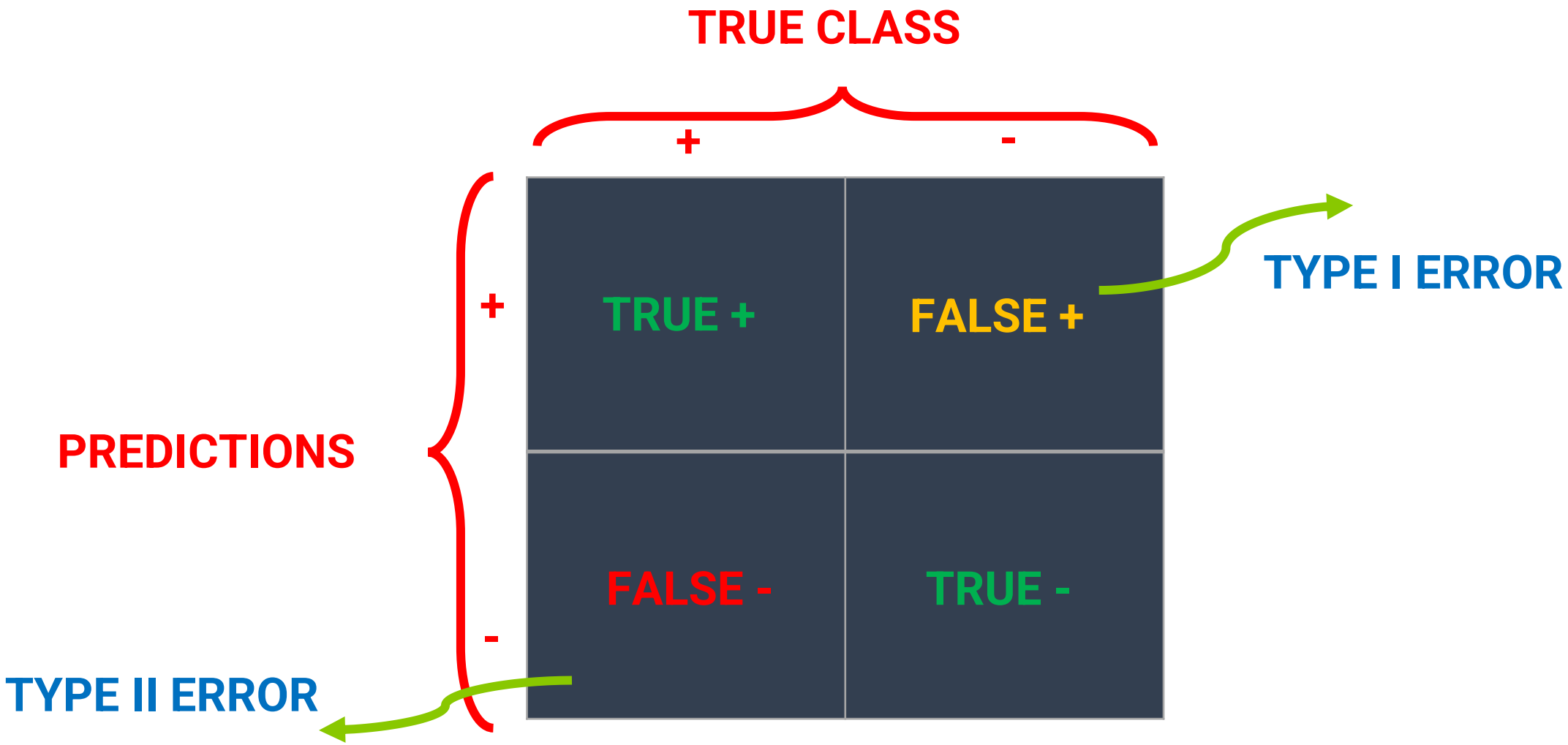
Using AWS SageMaker Canvas, perform the following tasks:

- 1. Upload the "*used_car_price.csv*" dataset to S3 (note that this datasets is a regression-type)
- 2. Perform basic data exploration
- 3. What is the average horsepower and its correlation to target?
- 4. Select the target column (MSRP)
- 5. Click on Preview model and list the estimated RMSE in $ values
- 6. List the 3 most important features in the datasets
- 7. Train a regression model using quick build option
- 8. Test the model using at least 3 single prediction values

CLASSIFIER MODEL KEY PERFORMANCE INDICATORS (KPIs) – REVIEW



CONFUSION MATRIX

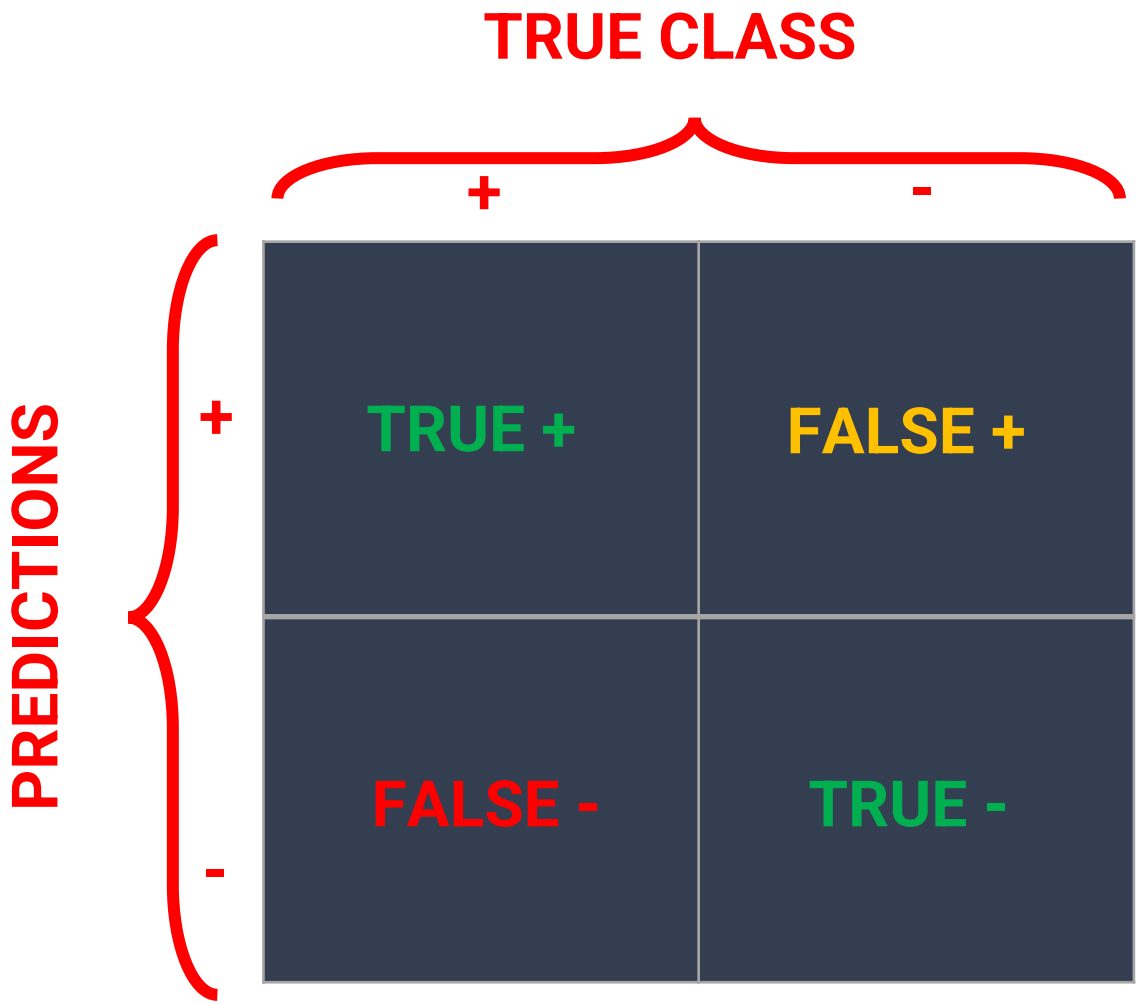


CLASSIFICATION MODEL KPIs

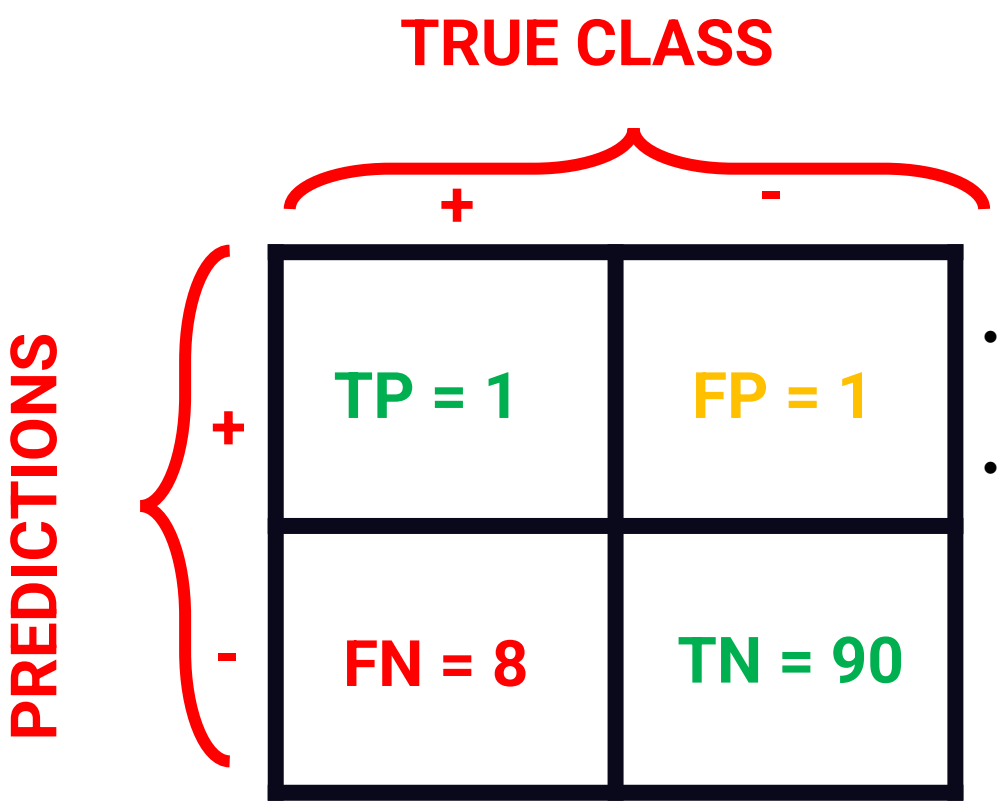
- A confusion matrix is used to describe the performance of a classification model:
    - True positives (TP): cases when classifier predicted TRUE (they have the disease), and correct class was TRUE (patient has disease).
    - True negatives (TN): cases when model predicted FALSE (no disease), and correct class was FALSE (patient do not have disease).
    - False positives (FP) (Type I error): classifier predicted TRUE, but correct class was FALSE (patient did not have disease).
    - False negatives (FN) (Type II error): classifier predicted FALSE (patient do not have disease), but they actually do have the disease
    - Classification Accuracy = (TP+TN) / (TP + TN + FP + FN)
    - Misclassification rate (Error Rate) = (FP + FN) / (TP + TN + FP + FN)

PRECISION Vs. RECALL

- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) (When model predicted TRUE class, how often was it right?)
- Recall = TP/ Actual TRUE = TP/ (TP+FN) (when the class was actually TRUE, how often did the classifier get it right?)



PRECISION Vs. RECALL

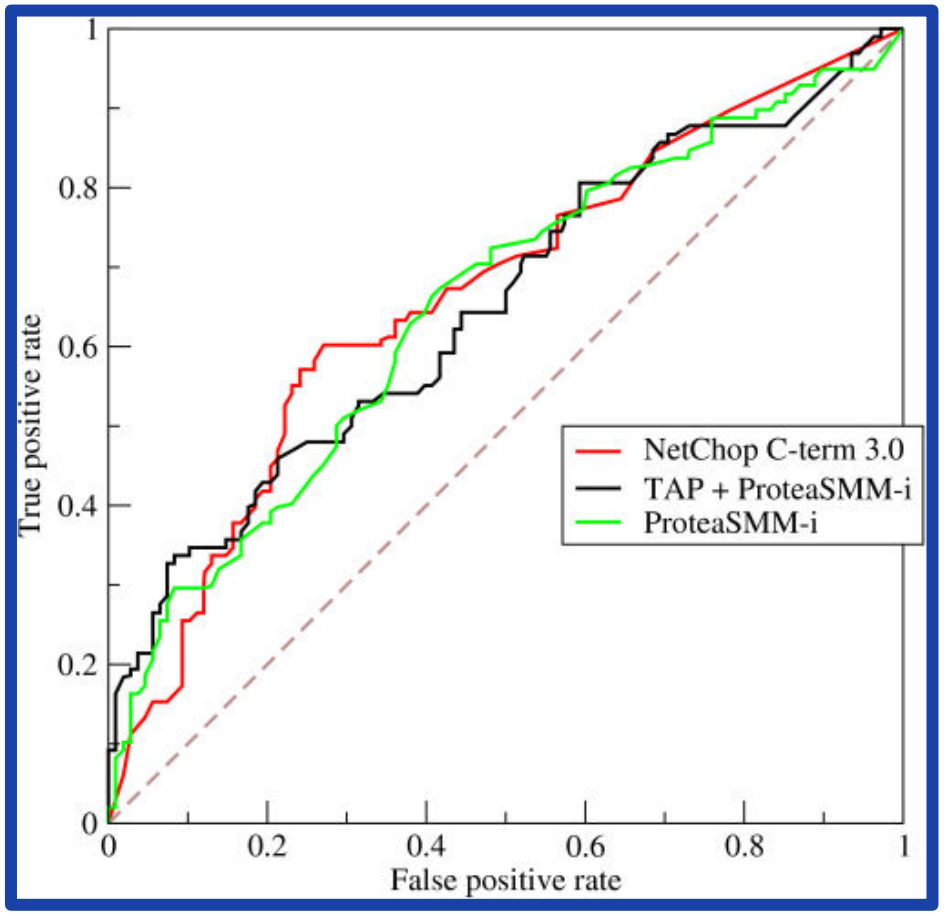


FACTS:
100 PATIENTS TOTAL
91 PATIENTS ARE HEALTHY
9 PATIENTS HAVE CANCER

- Accuracy is generally misleading and is not enough to assess the performance of a classifier.
- Recall is an important KPI in situations where:
  - Dataset is highly unbalanced; cases when you have small cancer patients compared to healthy ones.

- Classification Accuracy = (TP+TN) / (TP + TN + FP + FN) = 91%
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) = ½=50%
- Recall = TP/ Actual TRUE = TP/ (TP+FN) = 1/9 = 11%

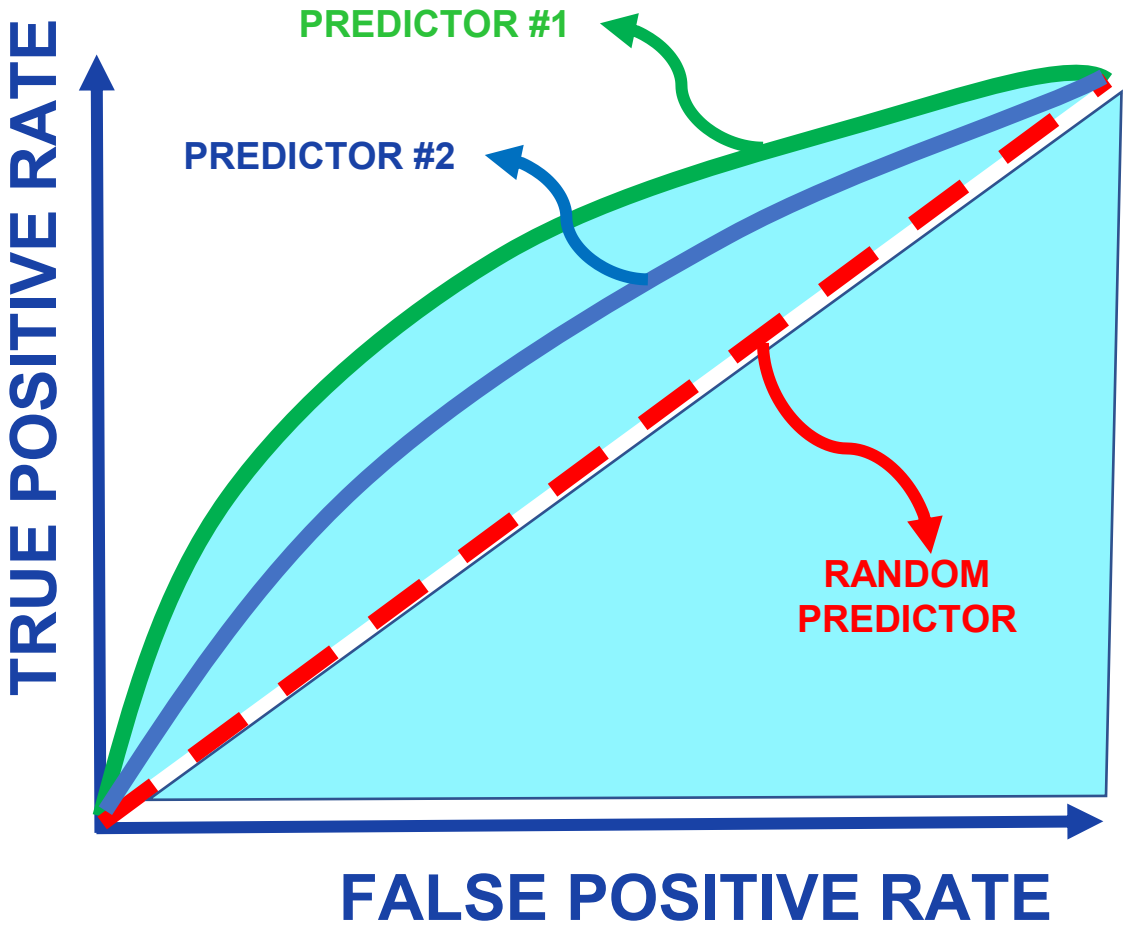
ROC (RECEIVER OPERATING CHARACTERISTIC CURVE)



- ROC Curve is a metric that assesses the model ability to distinguish between binary (0 or 1) classes.
- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.
- The false-positive rate is also known as the probability of false alarm.
- Points above the diagonal line represent good classification (better than random)
- The model performance improves if it becomes skewed towards the upper left corner.

Photo Credit: https://commons.wikimedia.org/wiki/File:Roccurves.png

AUC (AREA UNDER CURVE)



- The light blue area represents the area Under the Curve of the Receiver Operating Characteristic (AUROC).
- The diagonal dashed red line represents the ROC curve of a random predictor with AUROC of 0.5.
- If ROC AUC = 1, perfect classifier
- Predictor #1 is better than predictor #2
- Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.