Solving real-world challenges using no-code ML solutions

Aaron Chong
Solutions Architect
Amazon Web Services

27 May 2022

Agenda

- Machine Learning Overview & Challenges
- AWS Low-Code / No-Code ML Overview
- Amazon SageMaker Canvas Features
- Hands-on Lab
- Q&A

The reach of AI/ML is growing

Increased spending

"By 2021, global spending on AI and cognitive technologies will exceed $50 billion"

IDC

From piloting to operationalizing

"By the end of 2024, 75% of enterprises will shift from piloting to operationalizing AI"

Gartner

AI transformation

"57% said that AI would transform their organization in the next three years"

Deloitte

What is it?

Artificial intelligence (AI)

Any technique that enables computers to mimic human intelligence using logic, if-then statements, and ML (including deep learning)

Machine Learning (ML)

Subset of AI that uses machines to search for patterns in data to build logic models automatically

Deep learning

Subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

Common Types of Machine Learning

Supervised Learning

- Classification (Is it a Cat or Dog?)
  - Customer churn prediction
  - Machine failure detection
  - Patient re-admission prediction

- Regression (How many? How much?)
  - House price prediction
  - Demand forecasting

Unsupervised Learning

- Clustering (What is the grouping?)
  - Customer segmentation

Reinforcement Learning

- Self-driving Car

aws

Demystifying Machine Learning

Machine Learning is NOT…

- Put garbage-in and getting knowledge out
- Perform good predictions without data modeling & feature engineering
- Replace business rules – they augment them!

Machine Learning is

- Adapt based on recorded data
- Predict new data based on recorded data
- Extract hidden structure from the data

How ML Drives Value Creation Today

Inventory Forecasting

Churn Prediction Sales Conversion

Prioritized Use Cases

Step	**Business Requirements**	**Data Preparation & Feature Engineering**	**Model Development, Training, and Tuning**	**Model Deployment & Monitoring**
Owner	Business Leads, Domain Experts, and Business Analysts	Data Engineers and Data Scientists	Data Scientists and ML Engineers	ML Engineers

Usually takes from weeks to months

primarily solving for the prioritized use cases

Challenges Analysts Face in Building ML

Analysts lack deep ML expertise, and learning curve is steep

- Need to build understanding for ML concepts across data preparation, model development, and optimization
- Need expertise in choosing the right combination of feature engineering, type of model, and optimization technique
- Learning to write or decipher code is usually needed

Business needs explainability and validation from experts

- Analysts prefer to partner with data scientists in order to learn and build trust in the process, but data scientists time is limited and typically devoted to a few key ML projects
- Analysts need to be able to explain ML model predictions to business executives

Available no-code ML tools tend to lack transparency and have upfront fees

- Many no-code ML options lack code-level transparency making it difficult to inspect and productionalize models
- The UX for analysts and data scientists tends to be the same, requiring analysts to know the ML concepts and jargon
- Frequently, no-code ML tools come with licensing fees, so experimentation requires upfront investment

How Can You Scale ML Value Creation?

1 **Expand Your ML Development Team**
Grow your technical teams in proportion of your needs, **but** ML talent is in high demand

+74% annual compound growth in past 4 years

2x the demand growth of any other emerging job role

2 **Enhance ML Team Productivity**
Leverage low-code / no-code tools that make data science teams more productive

Enable data science teams to experiment faster with low-code / no-code Machine Learning capabilities

3 Democratize ML Innovation
Enable more groups of people, including business analysts to build ML models

Empower business analysts to make smarter decisions with no-code Machine Learning with a dedicated easy-to-use workspace

AWS Offerings on Low-Code / No-Code ML

Business Teams

Amazon SageMaker Canvas

A **dedicated no-code workspace for data analysts** to generate ML-powered predictions

A visual point-and-click interface that allows analysts to generate accurate ML predictions on their own — without requiring any machine learning experience or having to write a single line of code.

+

Data Science Teams

Amazon SageMaker Studio

A **dedicated workspace for data engineers, data scientists and ML Ops teams** to collaborate and bring ML to market faster

Data Wrangler
A faster, visual way to aggregate and prepare data for machine learning

Autopilot
AutoML capability that automatically prepares your data, as well as builds, trains, and tunes the best machine learning models for your tabular and time-series datasets

JumpStart
Pre-built solutions and a model zoo of pre-trained and easily tunable state-of-the-art models for Computer Vision, and Natural Language Processing

Many deployment options

Collaboration

No ML experience required

Amazon SageMaker Canvas expands access to ML by providing business analysts with a visual point-and-click interface, allowing you to generate ML predictions on your own

Use Amazon SageMaker Canvas to prepare data for ML modeling and generate ML predictions

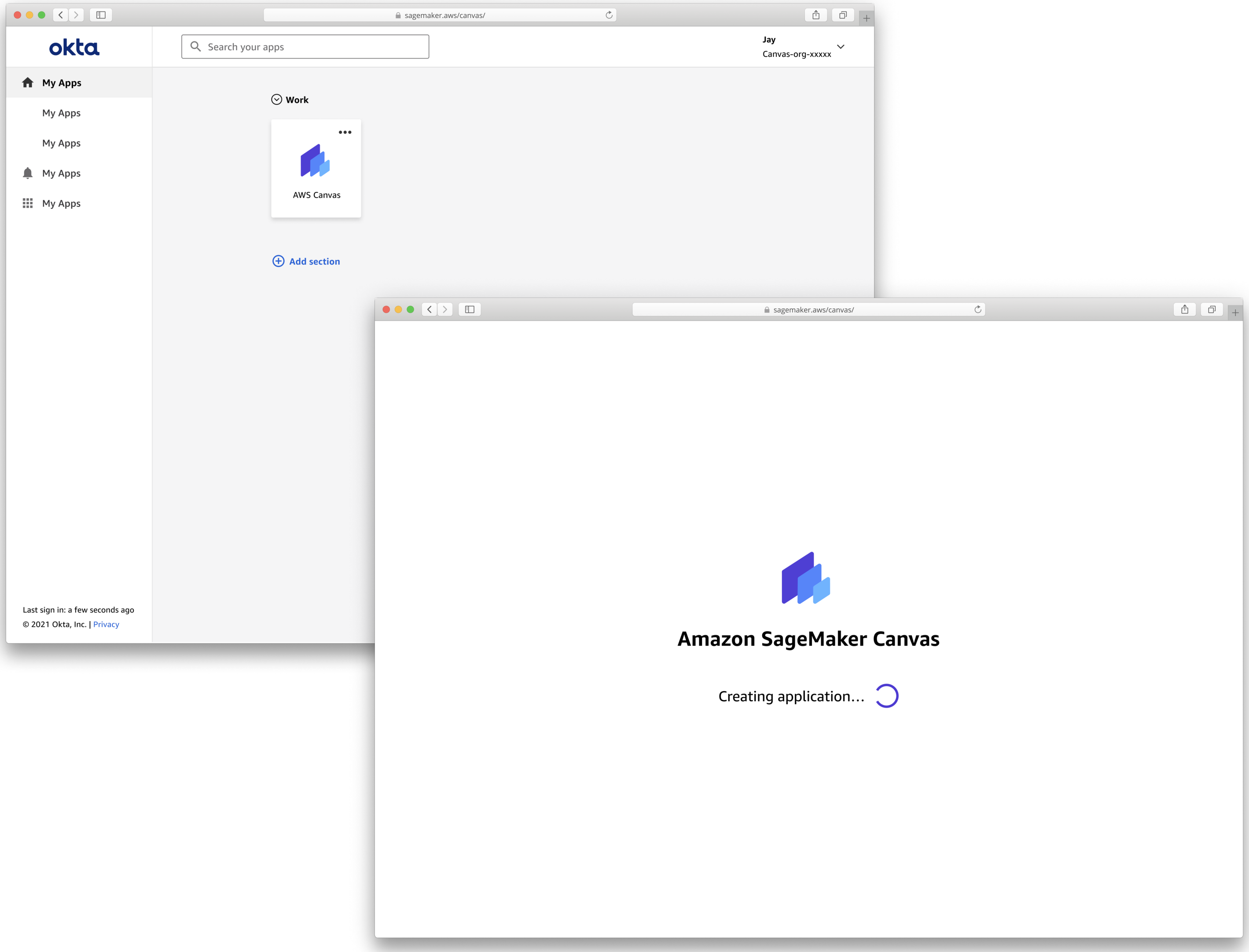
Can quickly **connect to and access data** from disparate sources

Leverages powerful AutoML technology to automatically **train and build models** based on your dataset

Identifies the best model based on your dataset to **generate single or bulk predictions**

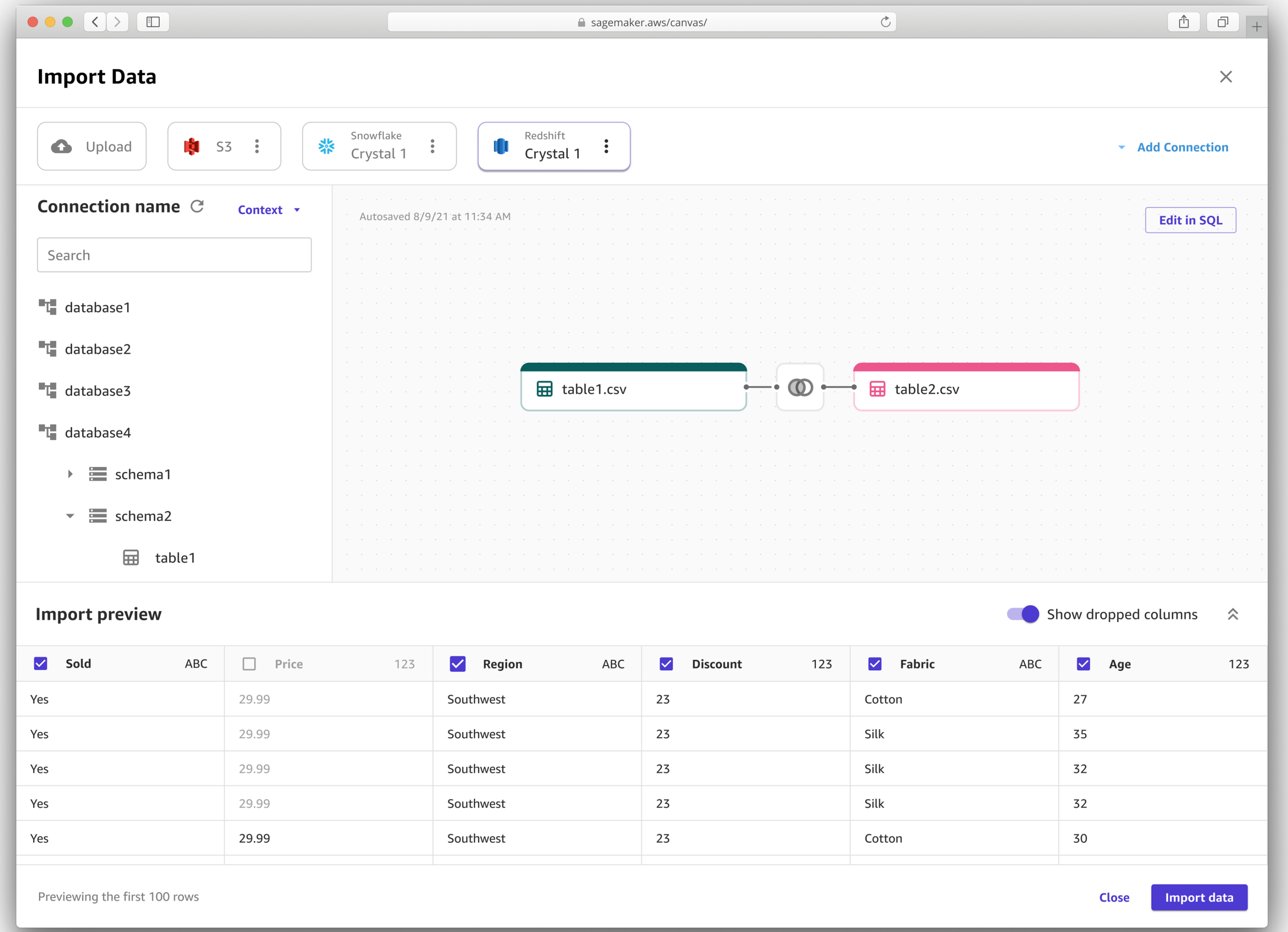
Integrates with SageMaker Studio, making it easy to **share models** with data scientists

Self-service access to a business-friendly tool for Machine Learning, outside of the AWS console

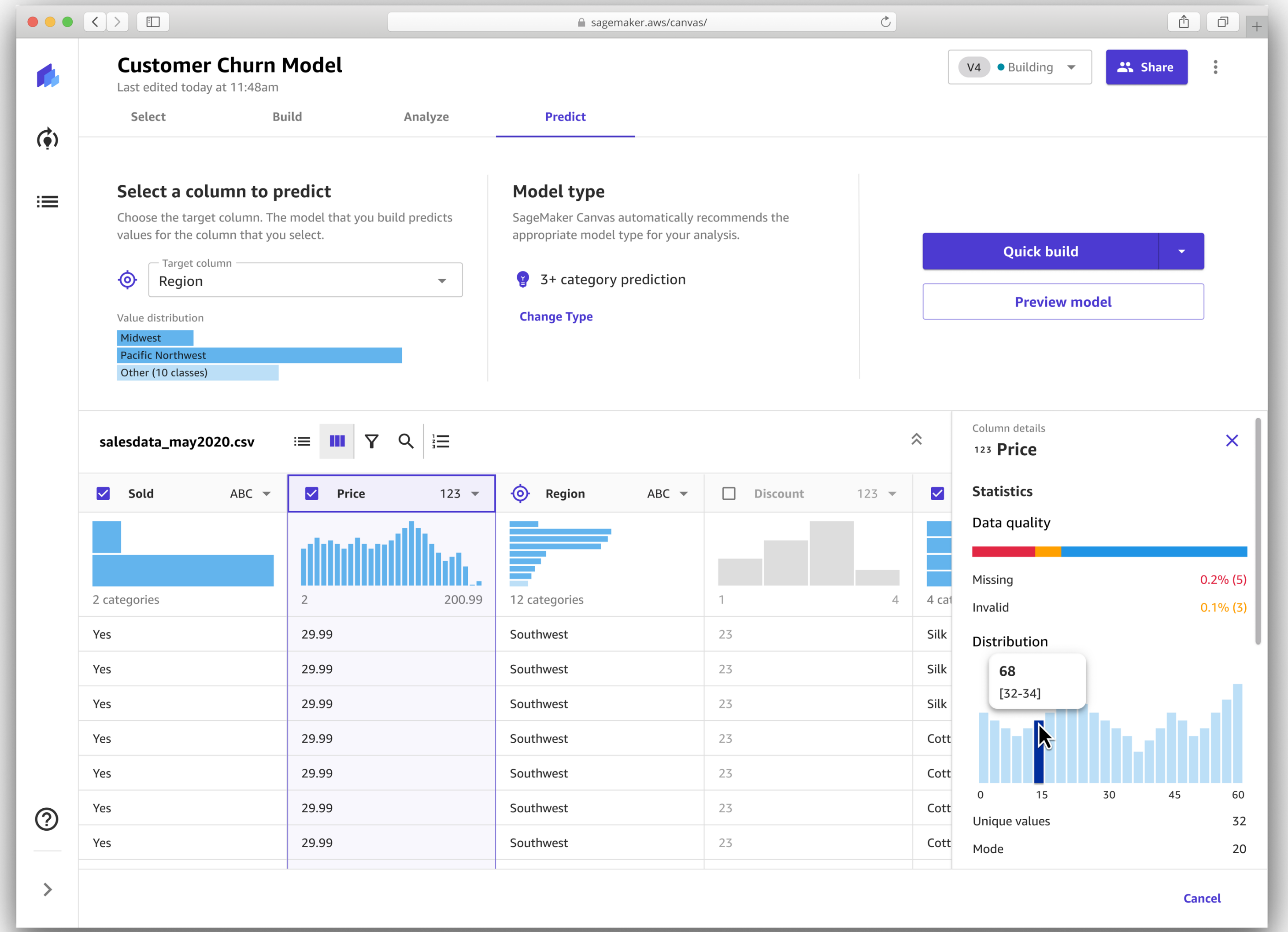


Combine datasets from various sources like local disk, Amazon S3, Amazon RedShift, and Snowflake

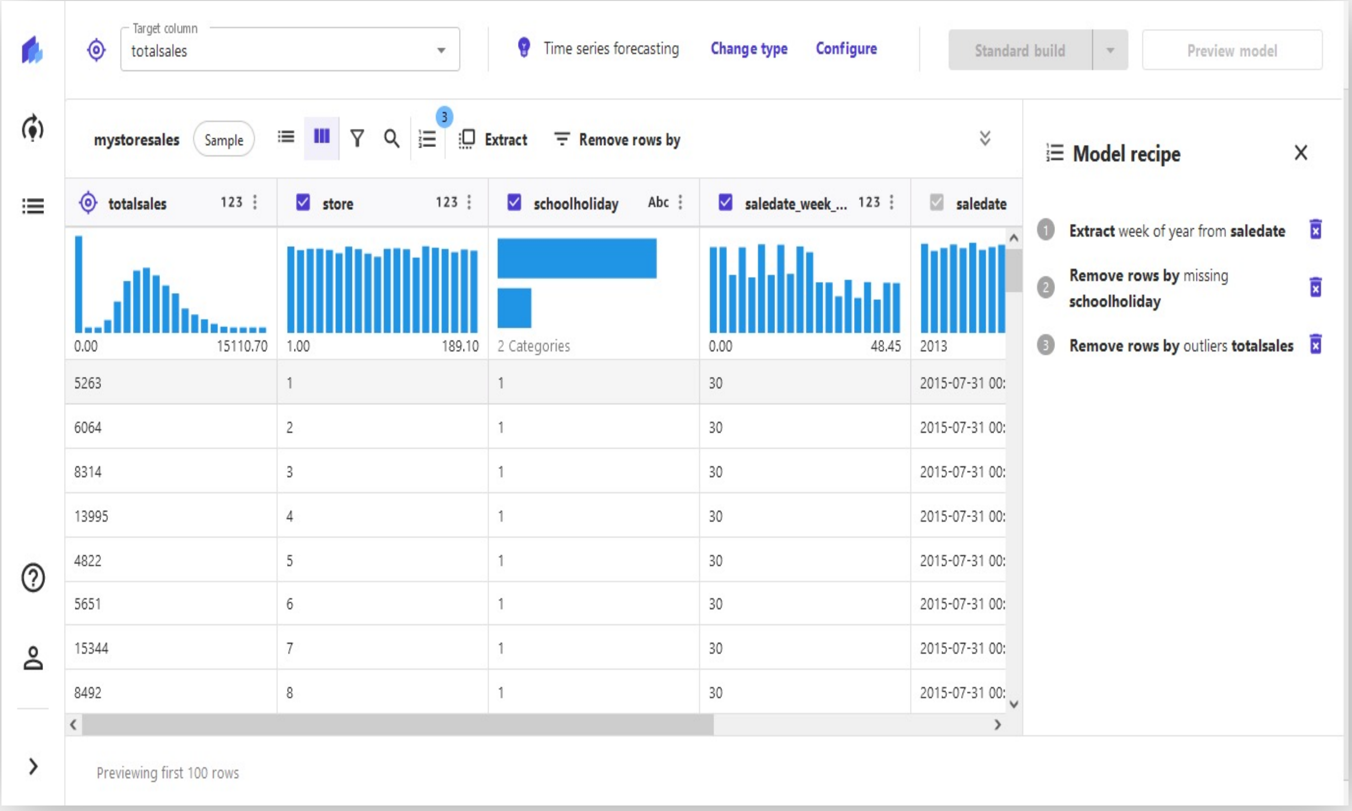
...with others coming soon



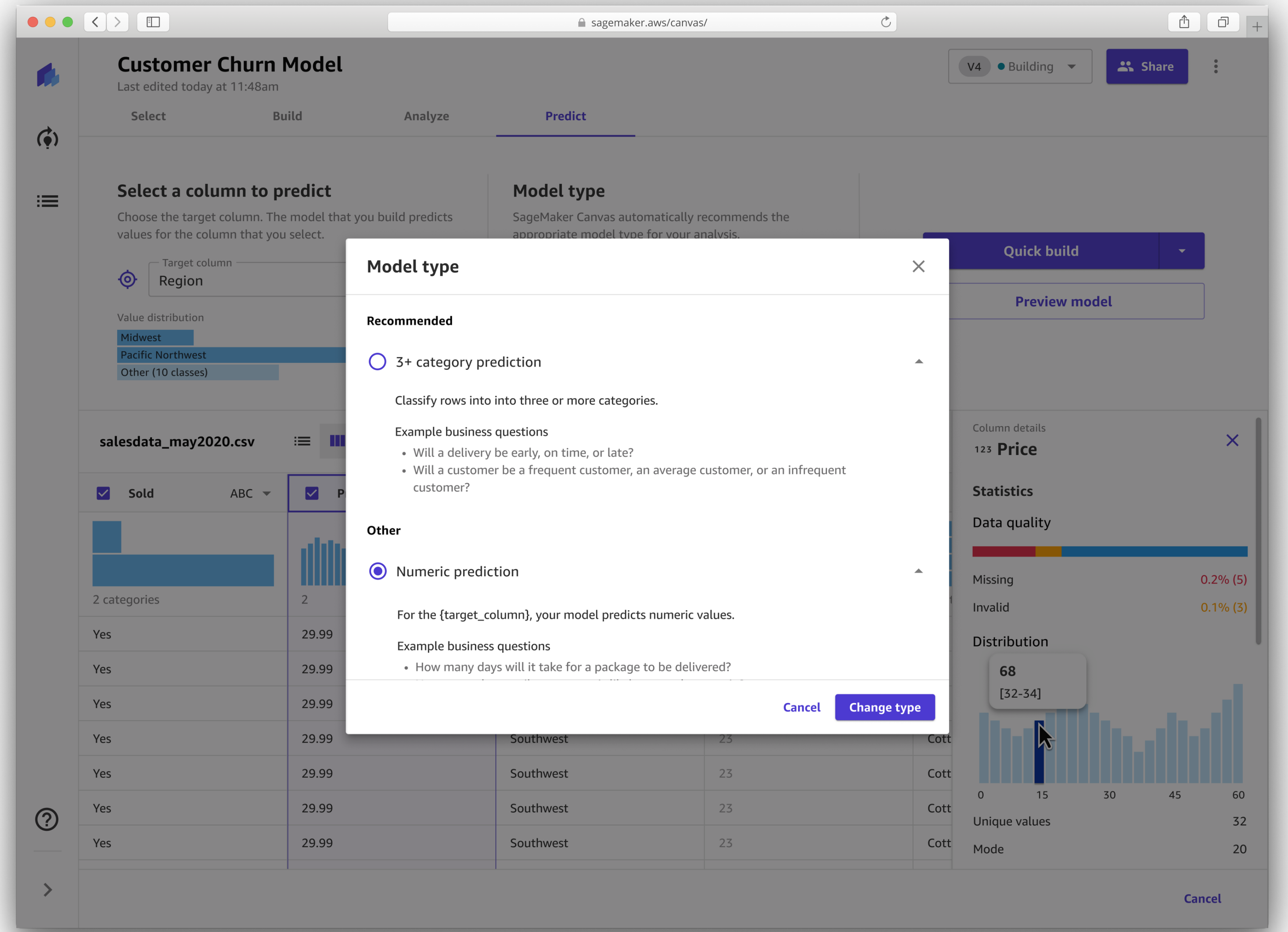
Quickly understand and prepare your data via a visual interface



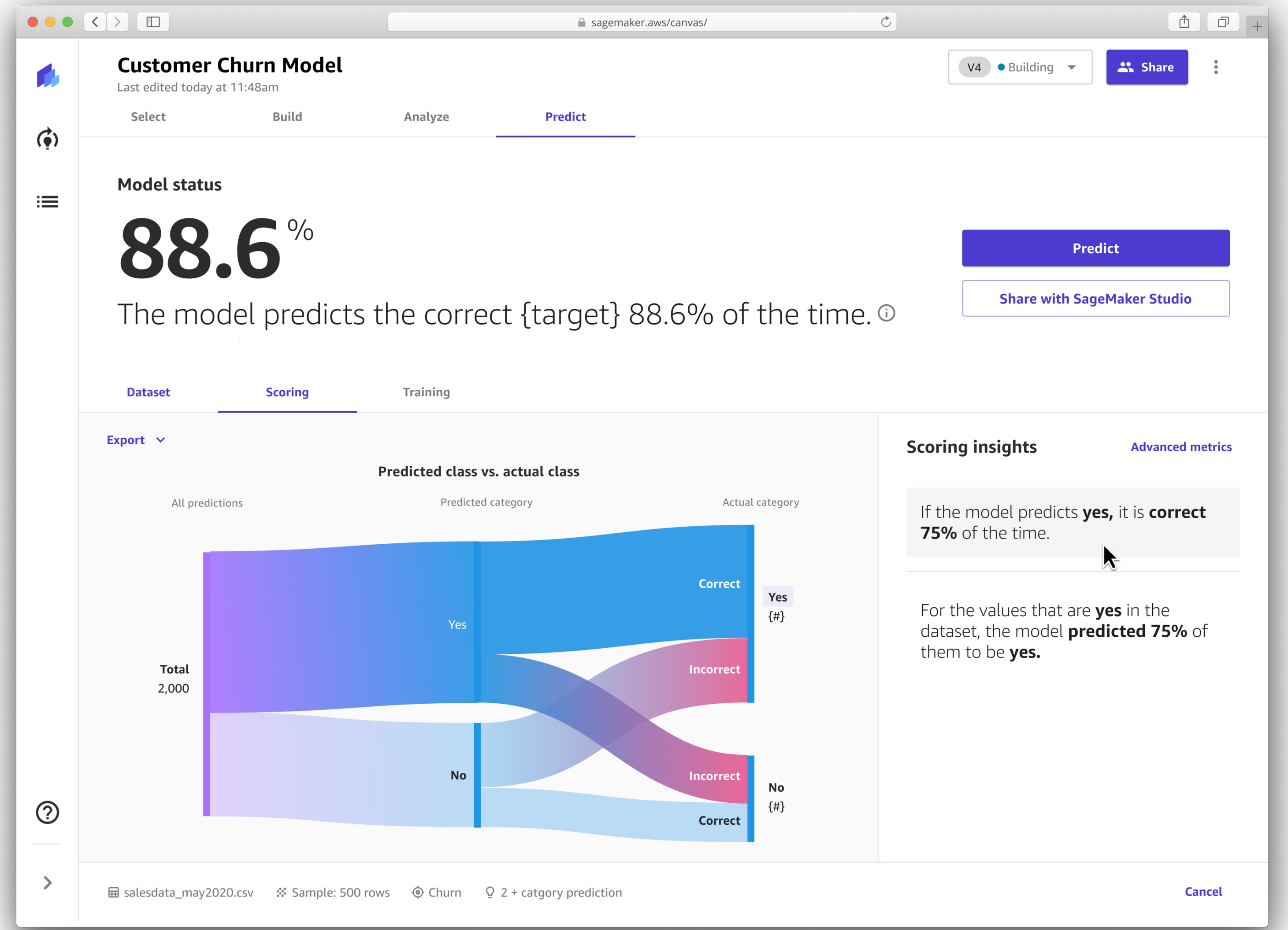
Transform and enrich your data by filtering values and extracting insights to add domain knowledge and improve model accuracy.



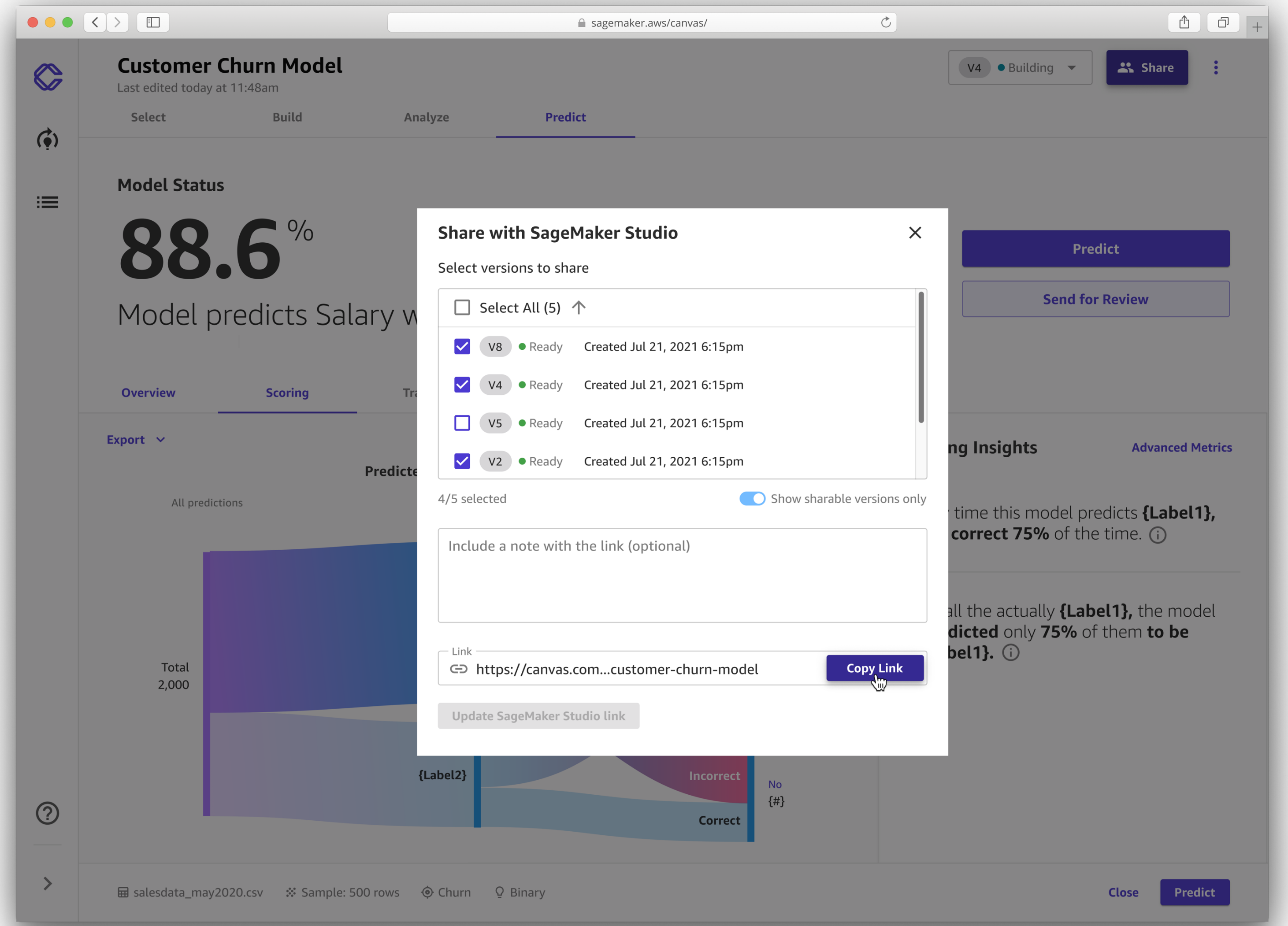
Automatically build an accurate ML model for your dataset



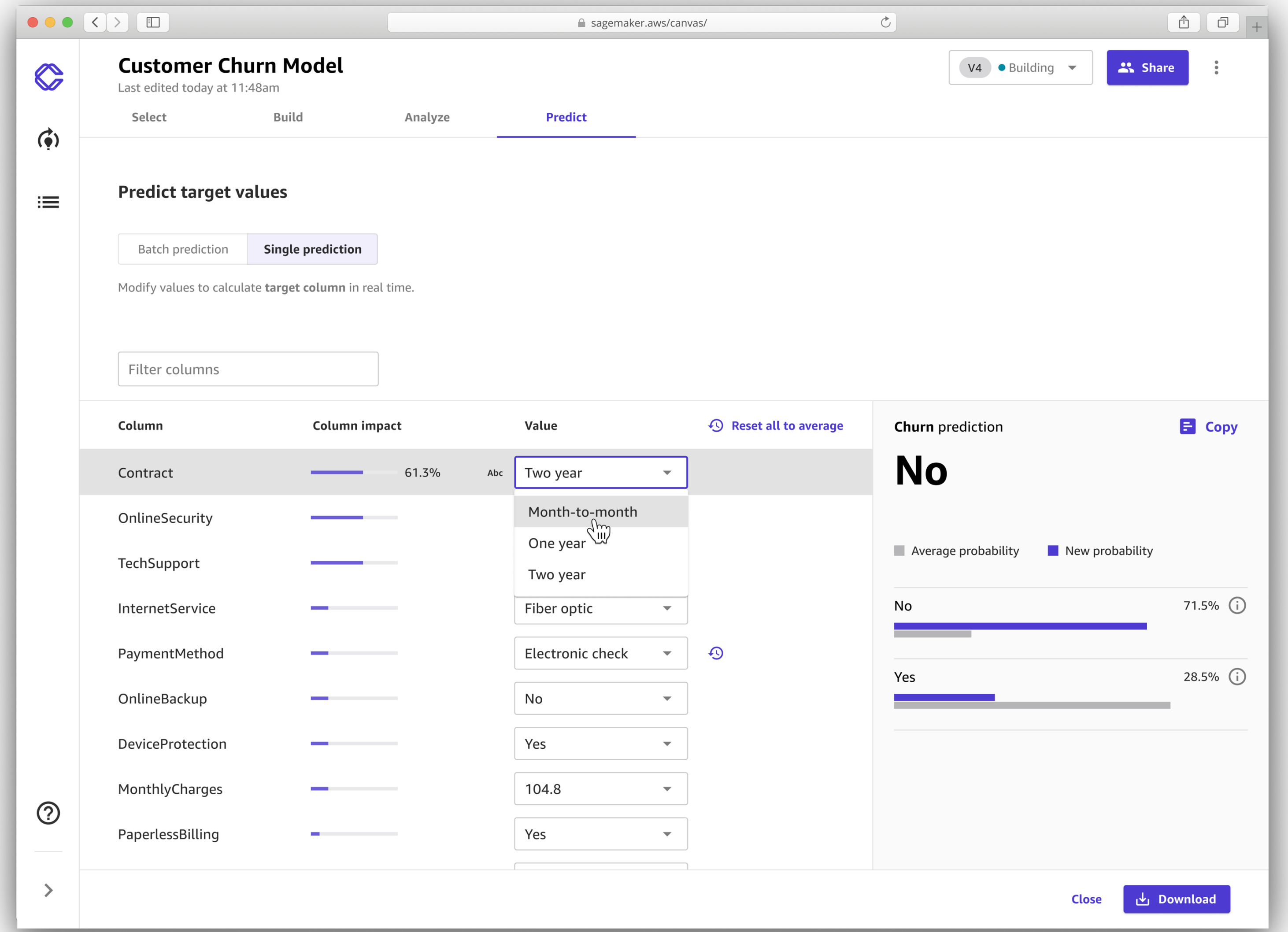
Get the first ML model in minutes. Review advanced metrics and feature importance to understand and explain predictions.



Easily share your models with data scientists to get feedback



Run what-if scenarios, or get predictions on an entire dataset



SageMaker Canvas Use Cases

VAST ARRAY OF USE CASES ACROSS DIFFERENT BUSINESS FUNCTIONS, OR VERTICALS

Sales and Marketing

1. Sales conversion
2. Sales forecasting
3. Propensity to churn
4. Customer lifetime value prediction
5. Marketing mix modeling

Finance and Accounting

1. Credit risk scoring
2. Delayed payments prediction
3. Fraud detection
4. Portfolio optimization
5. Account payables automation

Operations and Logistics

1. Demand forecasting
2. Inventory planning and scheduling
3. Delivery time forecasting
4. Predictive Maintenance

and many more…

aws

SageMaker Canvas Pricing Model

https://aws.amazon.com/sagemaker/canvas/pricing/

Session charges

$1.9 per hour

You pay based on the number of hours for which SageMaker Canvas is used or logged into. A session starts when you launch the SageMaker Canvas application, and ends when you log out (user logs out from inside the interface).

Training charges

You pay based on the number of cells of training data provided to train each model. Multiply your number of columns by your number of rows in your dataset and this will equal your number of cells. For example, if your dataset has 100 columns and 10,000 rows, your number of cells would equal (100 * 10,000), or 1 million.

Number of cells	Price
First 10M cells	$30 per million cells
Next 90M cells	$15 per million cells
Over 100M cells	$7 per million cells

The Amazon SageMaker Canvas **Free Tier** provides a 2-month free tier, which includes **interactive session hours up to 750 hours/month**, and **up to 10 model creation requests/month**, each with **up to 1M cells/model creation requests**.

For example, when running a modeling job with a dataset of 25.5 million cells, the training charges would be based on $30/M x first 10M cells ($300) and $15/M x next 16M cells ($240) for a total of $540.

Amazon SageMaker brings tools for every step of the ML lifecycle under one unified visual user interface

SAGEMAKER STUDIO | RSTUDIO

Label data

Prepare data

Store features

Detect bias

Build with notebooks

Train models

Tune parameters

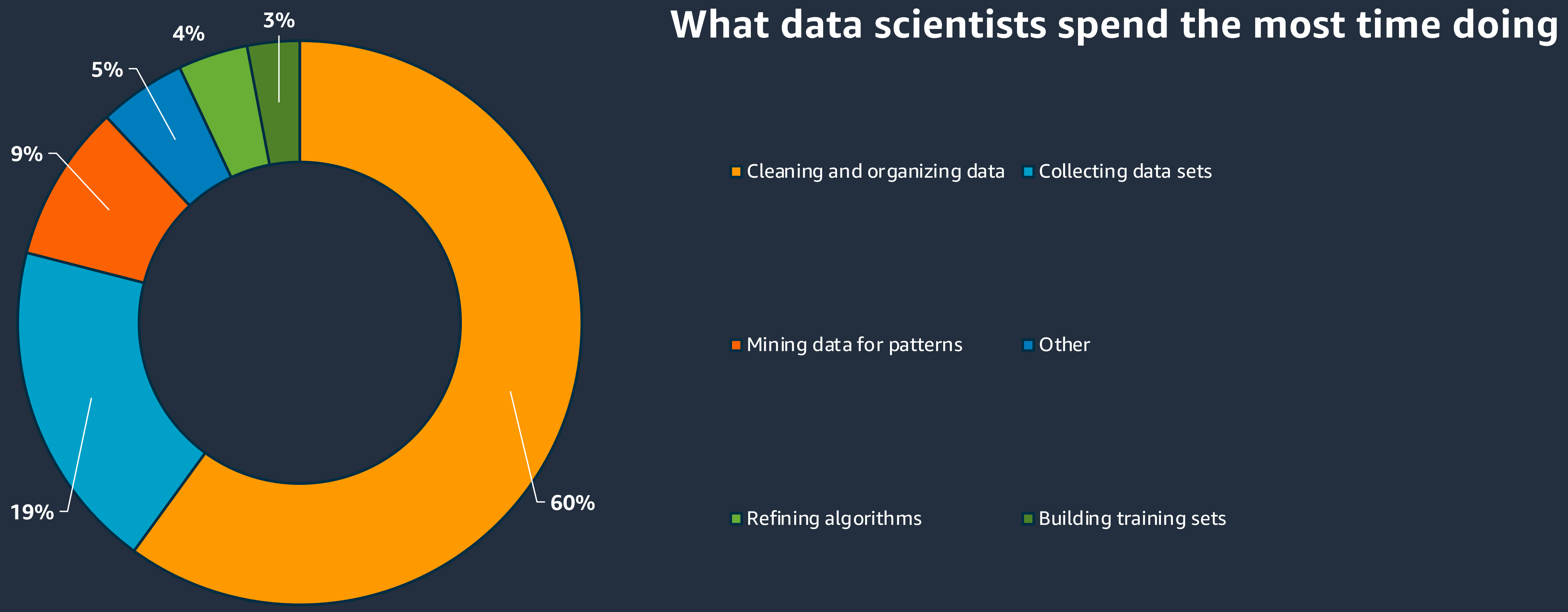
Deploy in production

Explain predictions

Manage and monitor

Manage edge devices

80% of time spent on data prep

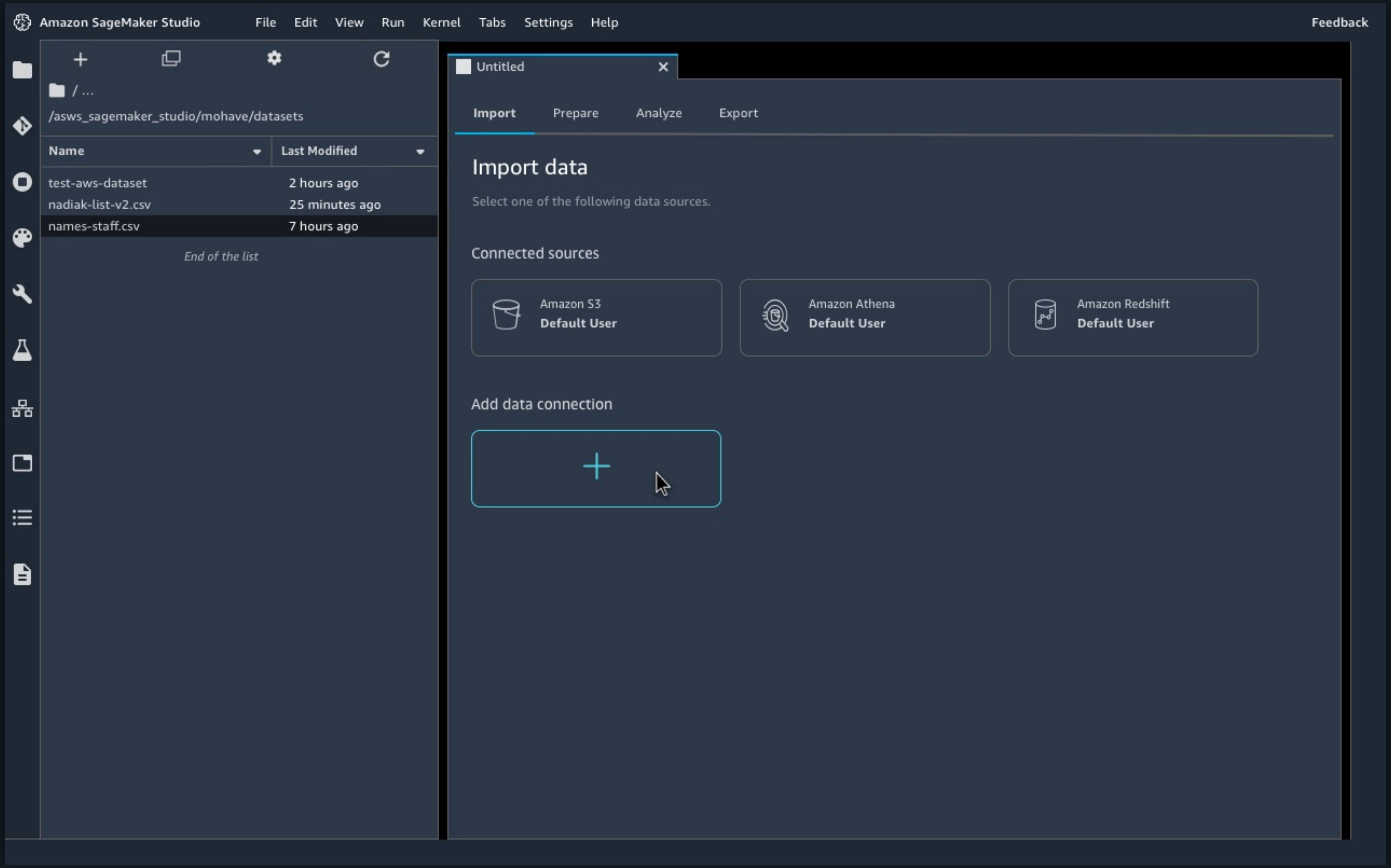


Source: Forbes survey of 80 data scientists, March 2016

Amazon SageMaker Data Wrangler

The fastest and easiest way to prepare data for machine learning

Quickly select and query data



Select data from Amazon Athena, Amazon Redshift, AWS Lake Formation, Amazon S3, and features from SageMaker Feature Store

Write queries for data sources before importing data over to SageMaker Data Wrangler

Import data in various file formats, such as CSV files, parquet files, and database tables directly into Amazon SageMaker

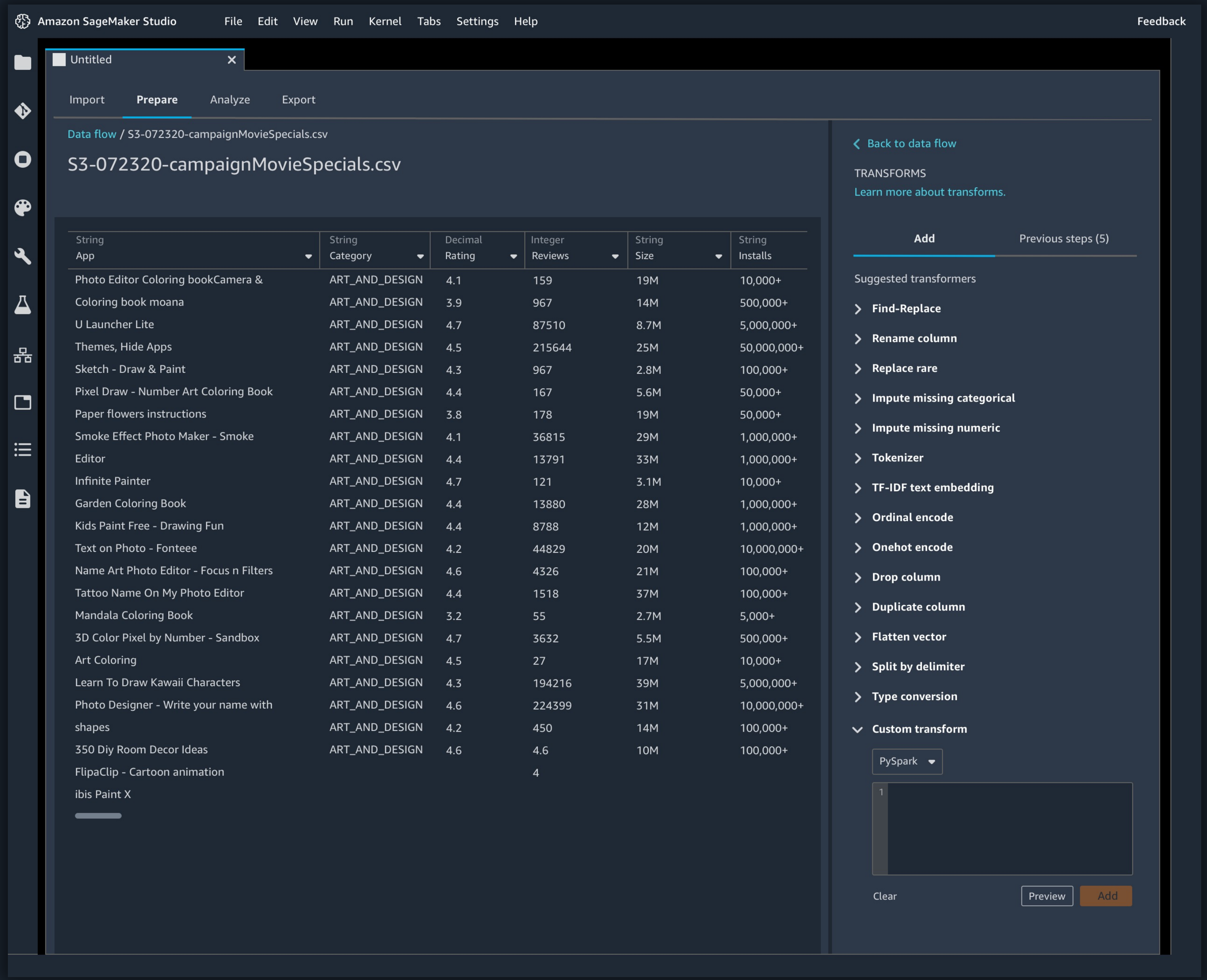
Easily transform data

Transform your data without writing a single line of code using pre-configured data transforms

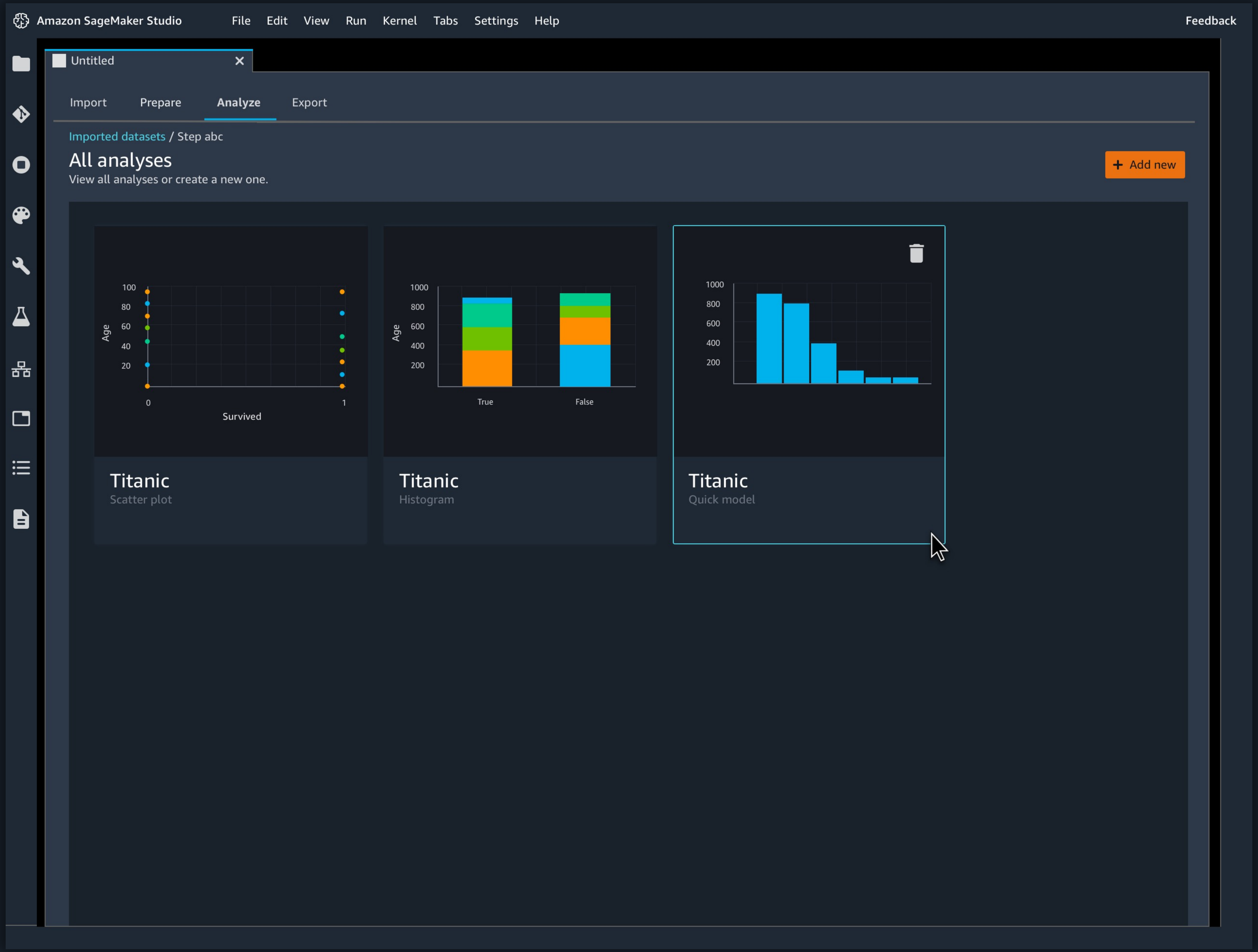
Preconfigured data transforms include convert column type, rename column, and delete column

Author custom transforms in PySpark, SQL, and Pandas

Detect bias and identify dataset imbalance with SageMaker Clarify



Understand your data visually

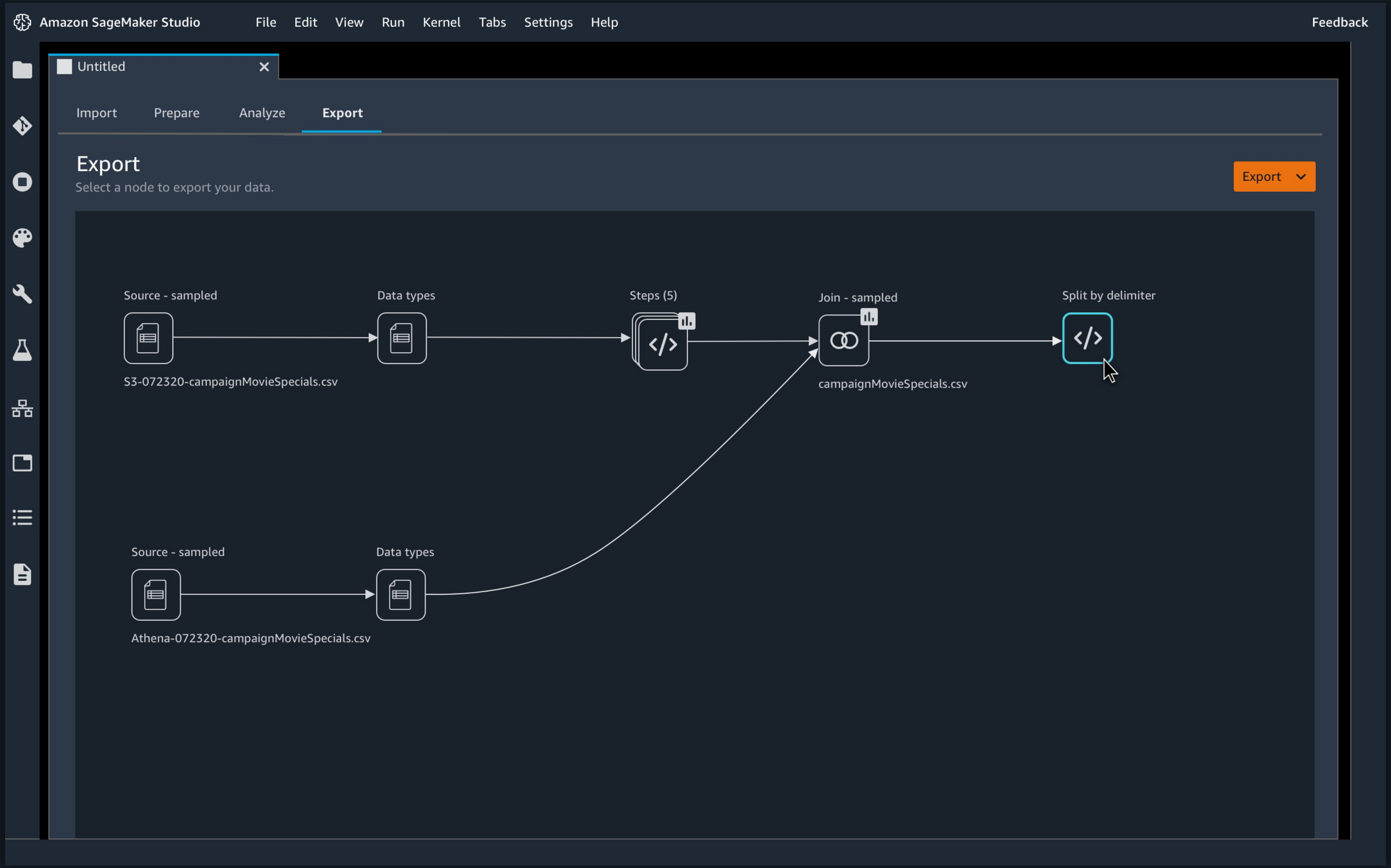


Intuitively understand your data with a set of pre-configured visualization templates

Preconfigured visualization templates include histograms, scatter plots, box and whisker plots, line plots, and bar charts

Interactively create and edit your own visualizations so you can quickly detect outliers or extreme values

Deploy data preparation workflows into production



Export data preparation workflows to a notebook or Python code

Integrate your workflow with SageMaker Pipelines to automate model deployment and management

Publish created features to SageMaker Feature Store for reuse and syndication across teams and projects

Amazon SageMaker Autopilot

Automatic model creation with full visibility and control

Quick to start

Provide your data in a tabular form and specify target prediction

Automatic model creation

Get ML models with feature engineering and model tuning automatically done

Visibility and control

Get notebooks for your models with source code

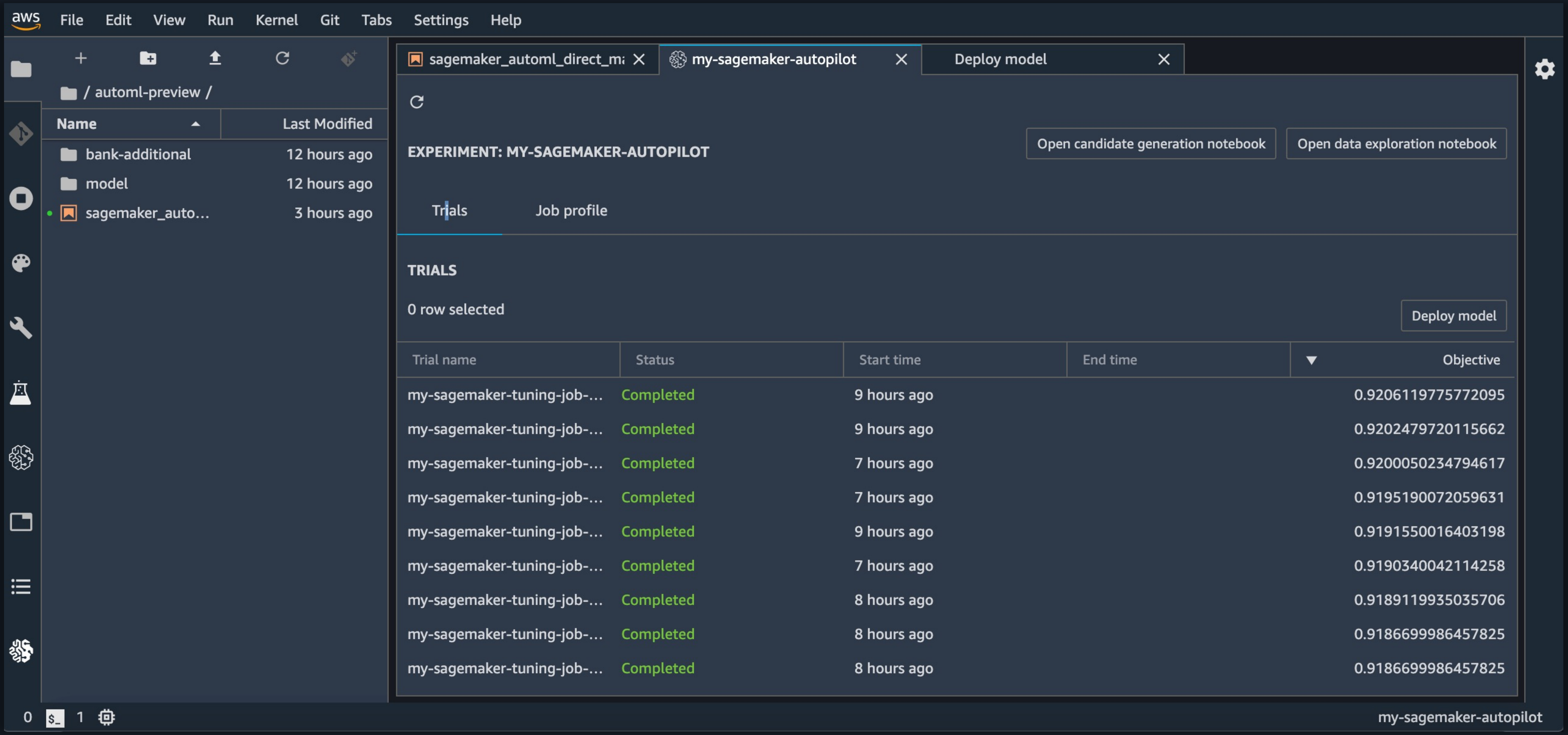
Recommendations and optimization

Get a leaderboard and continue to improve your model

AutoML with Amazon SageMaker Autopilot

1. Upload the dataset to Amazon S3
2. Configure the AutoML job
   - Location of dataset
   - Completion criteria
3. Launch the job
4. View the list of candidates and the auto-generated notebooks
5. Deploy the best candidate to a real-time endpoint, or use batch transform

Use Amazon SageMaker Autopilot to automatically train and tune the best machine learning models



Use Amazon SageMaker Autopilot for data exploration

Dataset exploration notebook

- Dataset statistics – row-wise and column-wise
- Suggested remedies for common dataset problems

Dataset Sample

The following table is a random sample of **10** rows from the training dataset. For ease of presentation, we are only showing **20 of the 21** columns of the dataset.

Suggested Action Items

- Verify the input headers correctly align with the columns of the dataset sample. If they are incorrect, update the header names of your input dataset in Amazon Simple Storage Service (Amazon S3).

	State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	...	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls
0	CO	76	408	412-4185	no	yes	26	214.6	110	36.48	...	87	17.44	134.6	140	6.06	8.1	2
1	NY	104	415	391-1793	no	yes	26	189.1	112	32.15	...	97	15.15	199.3	104	8.97	11.1	4
2	KY	122	408	392-1616	no	yes	27	253.7	84	43.13	...	109	19.48	190.5	123	8.57	9.2	5
3	NH	67	415	355-1113	no	yes	40	104.9	65	17.83	...	93	18.39	217.4	128	9.78	9.6	9
4	WI	153	510	349-3112	no	no	0	159.5	103	27.12	...	90	23.42	176.7	126	7.95	10.1	2
5	NH	146	510	345-2319	no	no	0	115.6	77	19.65	...	100	18.16	218.4	72	9.83	10.7	6
6	WV	63	510	328-9797	no	no	0	261.8	69	44.51	...	135	20.83	202.1	94	9.09	14.7	4
7	NH	90	408	393-7322	no	no	0	140.2	97	23.83	...	102	18.18	120.0	126	5.4	7.1	2

Use Amazon SageMaker Autopilot for model candidates

Fully runnable model candidate notebook

- Data transformers
- Featurization techniques applied
- Override points
    - Algorithms considered
    - Evaluation metric
    - Hyperparameter ranges
    - Model search strategy
    - Instances use

The SageMaker Autopilot Job has analyzed the dataset and has generated **10** machine learning pipeline(s) that use **2** algorithm(s). Each pipeline contains a set of feature transformers and an algorithm.

Available Knobs

1. The resource configuration: instance type & count
2. Select candidate pipeline definitions by cells
3. The linked data transformation script can be reviewed and updated. Please refer to the README.md for detailed customization instructions.

dpp0-xgboost: This data transformation strategy first transforms 'numeric' features using RobustImputer (converts missing values to nan), 'categorical' features using ThresholdOneHotEncoder, 'text' features using MultiColumnTfidfVectorizer. It merges all the generated features and applies RobustStandardScaler. The transformed data will be used to tune a *xgboost* model. Here is the definition:

```
[ ]: automl_interactive_runner.select_candidate({
        "data_transformer": {
            "name": "dpp0",
            "training_resource_config": {
                "instance_type": "ml.m5.4xlarge",
                "instance_count": 1,
                "volume_size_in_gb":  50
            },
            "transform_resource_config": {
                "instance_type": "ml.m5.4xlarge",
                "instance_count": 1,
            },
            "transforms_label": True,
            "transformed_data_format": "application/x-recordio-protobuf",
            "sparse_encoding": True
        }
```

SageMaker JumpStart

Easily and quickly bring machine learning applications to market

15+ pre-built solutions for common ML use cases

Solutions can be used out-of-the-box or can be customized for a specific business problem

Accelerate time to deploy over 150 open source models

Use one-click deployable ML models and algorithms from popular model zoos

Get started with just a few clicks

Easily bring ML applications to market using pre-built solutions, ML models, and algorithms from popular model zoos, and getting started content

SageMaker JumpStart
Use cases

- Autonomous driving
- Predictive maintenance
- Churn prediction
- Computer vision
- Personalized recommendations
- Fraud detection
- Extract data from documents
- Demand forecasting
- Credit risk prediction

Amazon SageMaker JumpStart pre-built solutions

Predictive Maintenance

Predictive maintenance for manufacturing >

Predictive maintenance for vehicle fleets >

Credit Risk Prediction

Explain credit decisions >

Autonomous Driving

Visual perception with active learning >

Demand Forecasting

Demand forecasting with deep learning >

Extract & Analyze Data from Documents

Document summarization, entity, and relationship extraction >

Handwriting recognition >

Filling in missing values in tabular records >

Differential privacy for sentiment classification >

Personalized Recommendations

Entity resolution in identity graphs >

Purchase modeling >

Fraud Detection

Detect malicious users and transactions >

Fraud detection in financial transactions using deep graph library >

Computer Vision

Product defect detection in images >

Churn Prediction

Churn prediction with text >

Learn more about solutions:
https://aws.amazon.com/sagemaker/getting-started/

aws

Benefits with AWS Low-Code / No-Code ML

Business Teams

Amazon SageMaker Canvas

A dedicated no-code workspace for data analysts to generate ML-powered predictions

A visual point-and-click interface that allows analysts to generate accurate ML predictions on their own — without requiring any machine learning experience or having to write a single line of code.

\+

Data Science Teams

Amazon SageMaker Studio

A dedicated workspace for data engineers, data scientists and ML Ops teams to collaborate and bring ML to market faster

Data Wrangler
A faster, visual way to aggregate and prepare data for machine learning

Autopilot
AutoML capability that automatically prepares your data, as well as builds, trains, and tunes the best machine learning models for your tabular and time-series datasets

JumpStart
Pre-built solutions and a model zoo of pre-trained and easily tunable state-of-the-art models for Computer Vision, and Natural Language Processing

Many deployment options

Collaboration

Analysts / Business Teams:

- Get access to a very easy to use no-code environment and get ML-powered predictions without knowing ML, or how to code completely on their own
- Increase trust, as well productionalize models they built using the collaboration between Canvas and Studio

Data Science Teams:

- Do exploratory analyses and build feature engineering pipelines very fast leveraging Data Wrangler's easy to use visual interface that entails 300+ transformations
- Explore 100s of ML pipelines automatically and use the code Autopilot produced in a white-box fashion
- Leverage powerful solutions and models as-is or can fine-tune them for their needs with JumpStart

Low-Code / No-Code ML empowers companies to scale their ML value creation from months to actually hours

Break – 5 minutes

Hands-on Lab – Marketing Churn

Binary Classification

Introduction

Direct marketing, through mail, email, phone, etc., is a common tactic to acquire customers. Because resources and a customer's attention are **limited**, the goal is to only target the subset of prospects who are likely to engage with a specific offer. In the following lab, we will train a model which can be used to predict if a customer will enroll for a term deposit at a bank, after one or more phone calls.

Predicting those potential customers based on readily available information like demographics, past interactions, and environmental factors is a common machine learning problem. You can imagine that this task would readily translate to **marketing lead prioritization** in your own organization.

Dataset Overview

- We have a little over 40K customer records, and 20 features for each customer
- The features are mixed; some numeric, some categorical

Specifics on each of the features:

Demographics:

- `age` : Customer's age (numeric)
- `job` : Type of job (categorical: 'admin.', 'services', ...)
- `marital` : Marital status (categorical: 'married', 'single', ...)
- `education` : Level of education (categorical: 'basic.4y', 'high.school', ...)

Past customer events:

- `default` : Has credit in default? (categorical: 'no', 'unknown', ...)
- `housing` : Has housing loan? (categorical: 'no', 'yes', ...)
- `loan` : Has personal loan? (categorical: 'no', 'yes', ...)

Past direct marketing contacts:

- `contact` : Contact communication type (categorical: 'cellular', 'telephone', ...)
- `month` : Last contact month of year (categorical: 'may', 'nov', ...)
- `day_of_week` : Last contact day of the week (categorical: 'mon', 'fri', ...)
- `duration` : Last contact duration, in seconds (numeric). Important note: If duration = 0 then `y` = 'no'.

Campaign information:

- `campaign` : Number of contacts performed during this campaign and for this client (numeric, includes last contact)
- `pdays` : Number of days that passed by after the client was last contacted from a previous campaign (numeric)
- `previous` : Number of contacts performed before this campaign and for this client (numeric)
- `poutcome` : Outcome of the previous marketing campaign (categorical: 'nonexistent','success', ...)

External environment factors:

- `emp.var.rate` : Employment variation rate - quarterly indicator (numeric)
- `cons.price.idx` : Consumer price index - monthly indicator (numeric)
- `cons.conf.idx` : Consumer confidence index - monthly indicator (numeric)
- `euribor3m` : Euribor 3 month rate - daily indicator (numeric)
- `nr.employed` : Number of employees - quarterly indicator (numeric)

Target variable:

- `y` : Has the client subscribed a term deposit? (binary: 'yes','no')

Our goal

Predict if customer will engage on the offer (y == yes) or not (y == no)

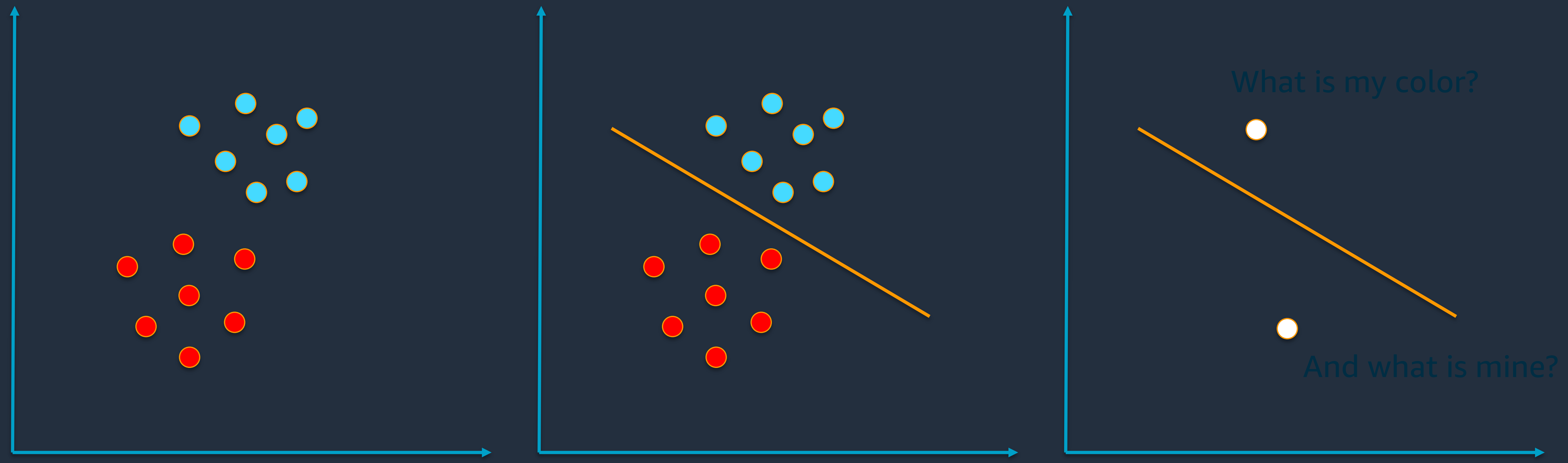
	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41183	73	retired	married	professional.course	no	yes	no	cellular	nov	fri	334	1	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	yes
41184	46	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	383	1	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	no
41185	56	retired	married	university.degree	no	yes	no	cellular	nov	fri	189	2	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	no
41186	44	technician	married	professional.course	no	no	no	cellular	nov	fri	442	1	999	0	nonexistent	-1.1	94.767	-50.8	1.028	4963.6	yes
41187	74	retired	married	professional.course	no	yes	no	cellular	nov	fri	239	3	999	1	failure	-1.1	94.767	-50.8	1.028	4963.6	no

41188 rows × 21 columns

Problem Type: Binary Classification

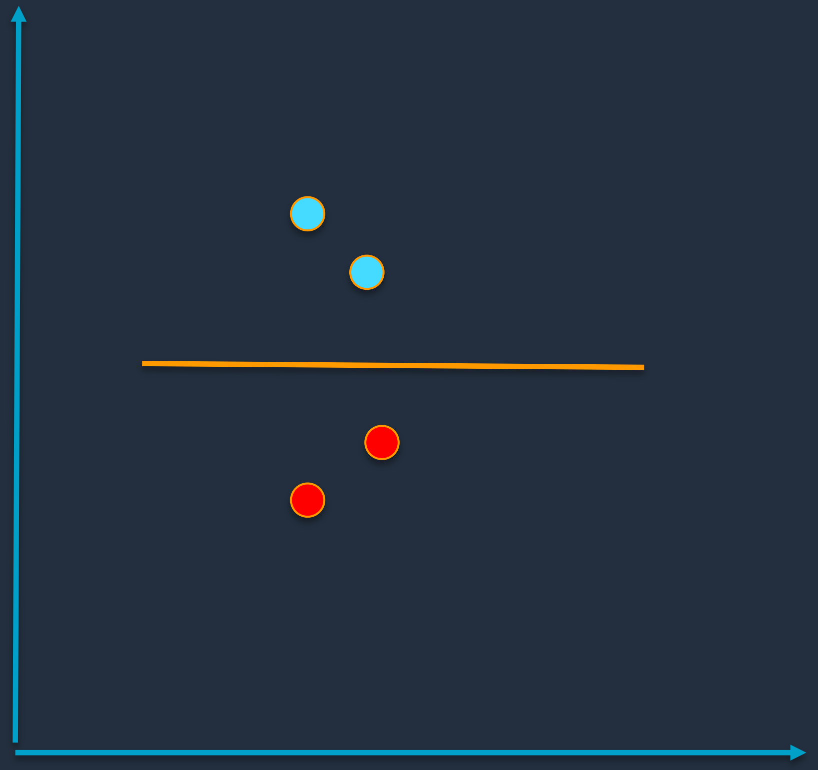
Importance of good training data

ML Model is a function to split the data points

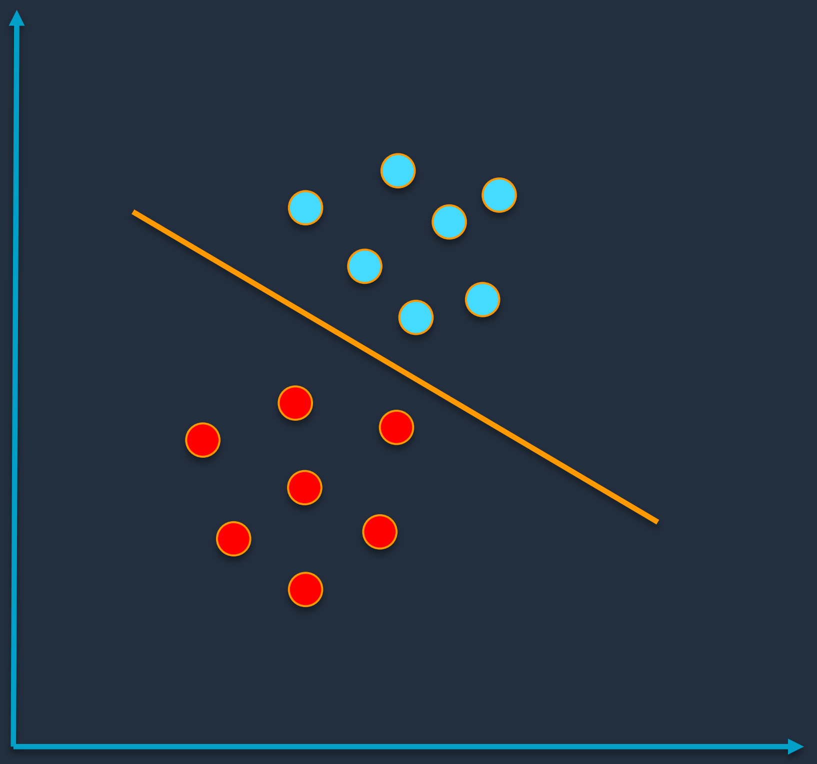


Historical Data | Model Building | Prediction

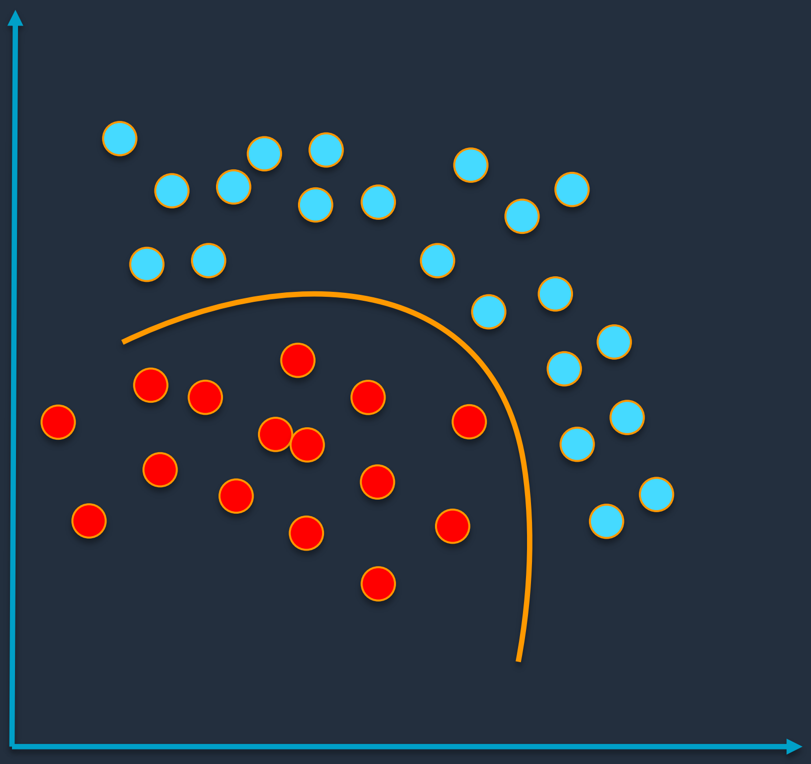
Why is more data better?



Less Data

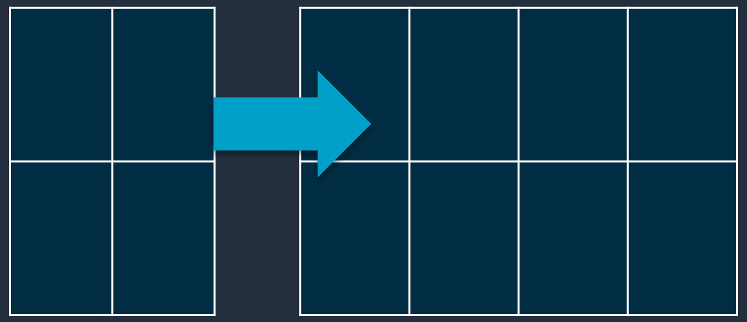


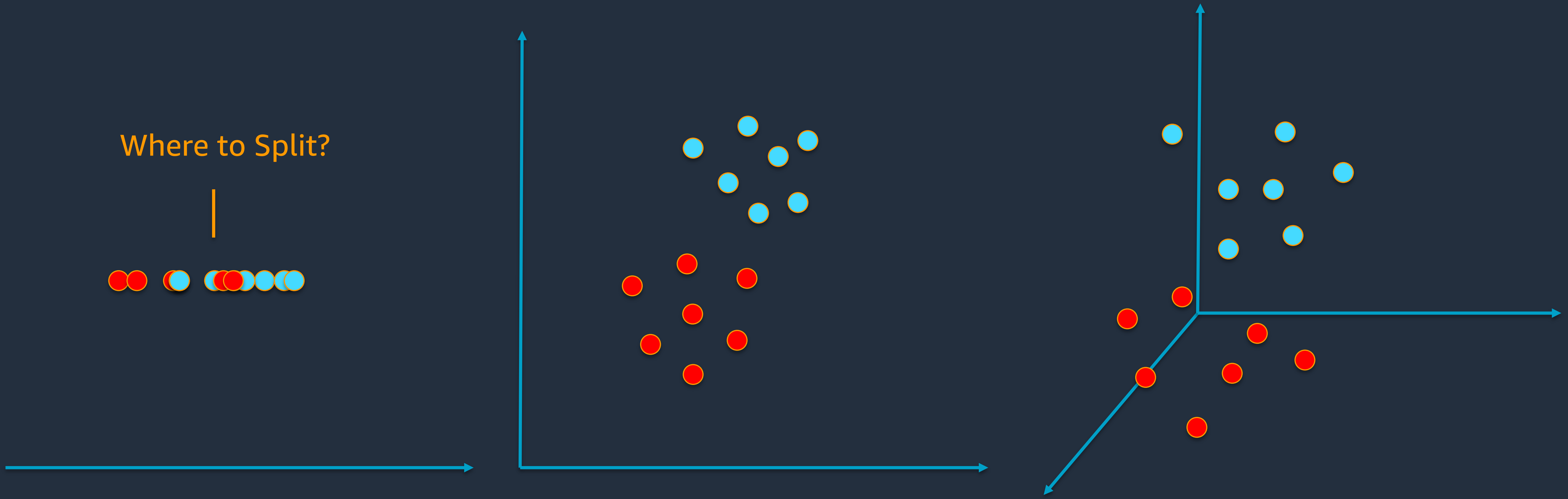
More Data



Even More Data

Why are more attributes better?



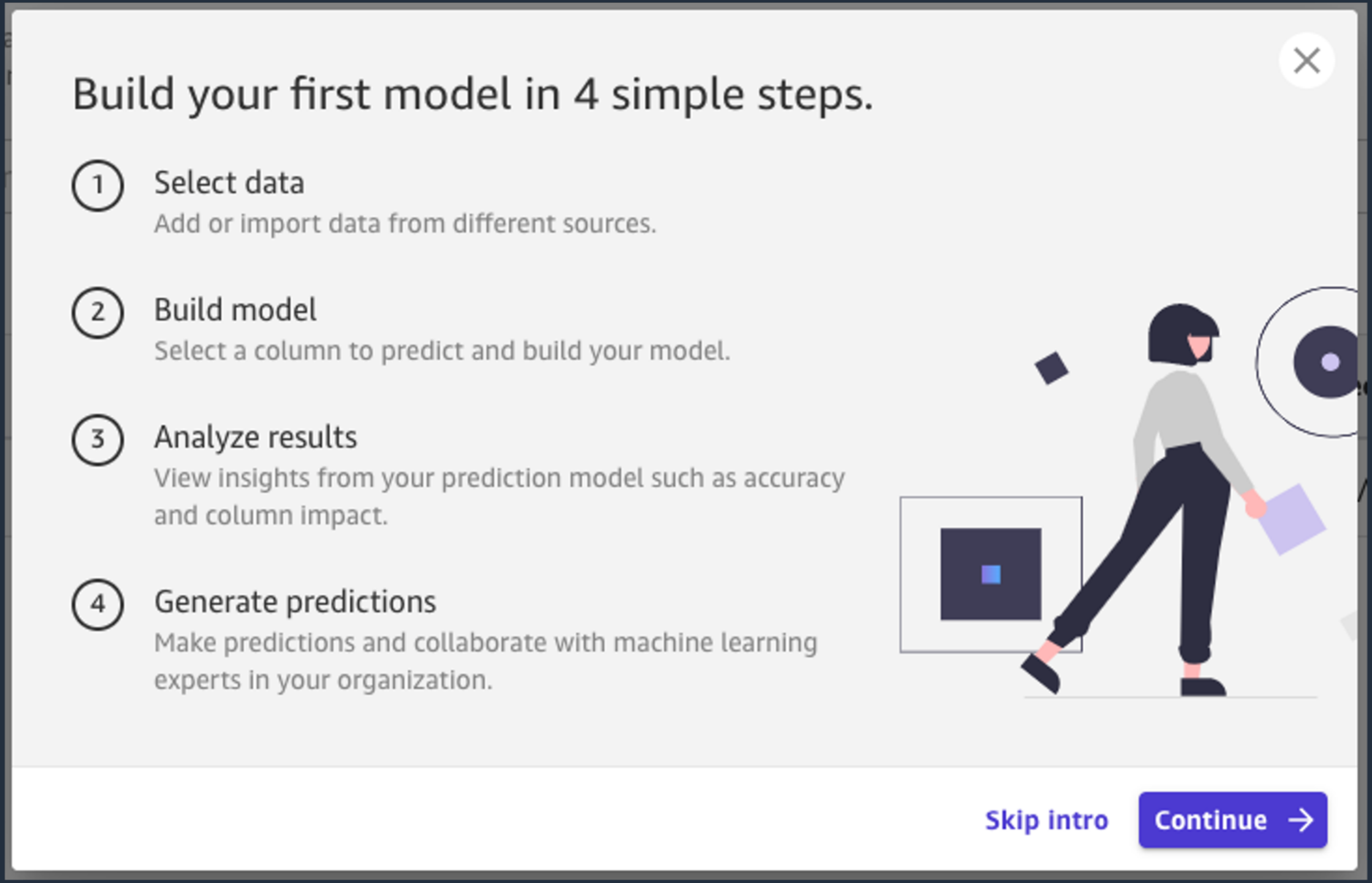


Less Attributes

More Attributes

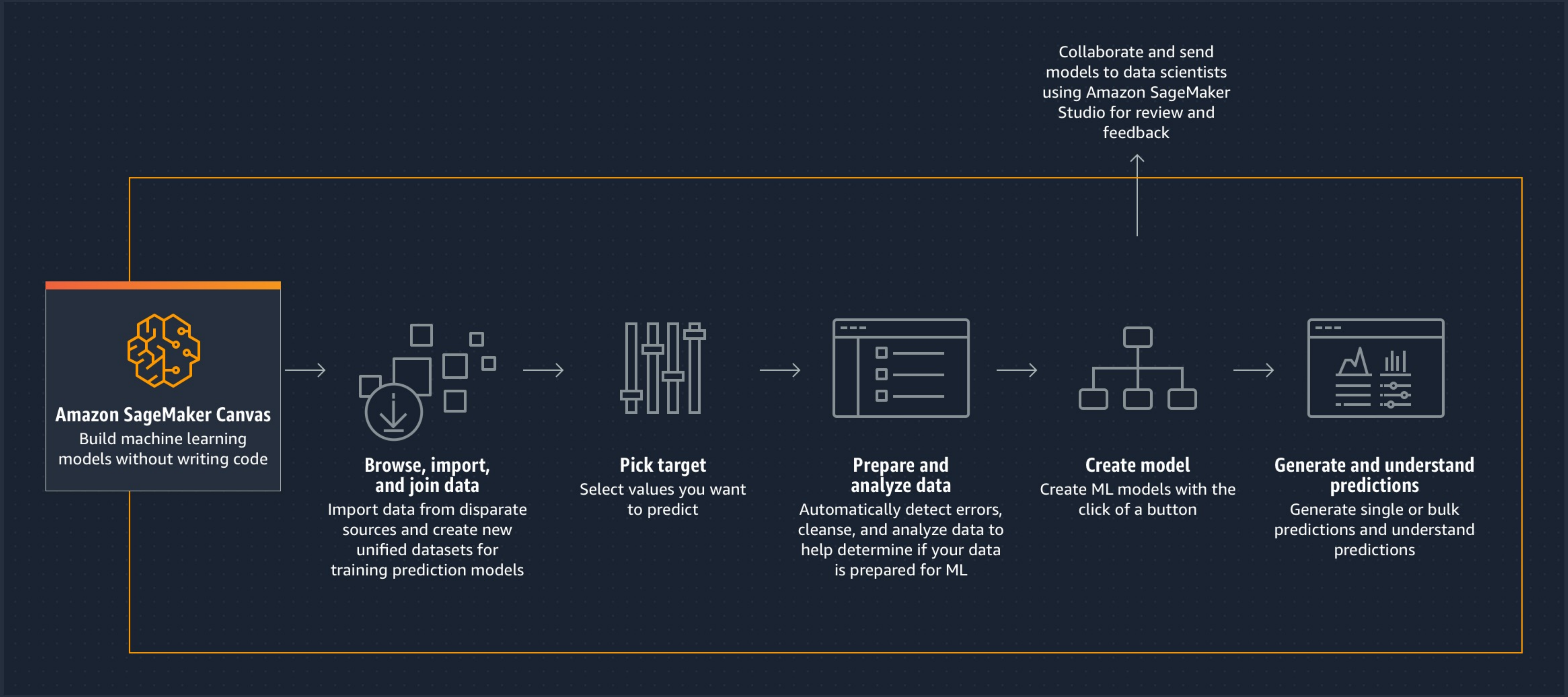
Even More Attributes

Solving the problem with SageMaker Canvas



Summary

Amazon SageMaker Canvas: No-code ML



ML Value Creation From Months to Hours

WITH A TAILOR-MADE WORKSPACE FOR ANALYSTS AND A TAILOR-MADE WORKSPACE FOR DATA SCIENTISTS

Business Requirements

Business Leads, Domain Experts, and Business Analysts

Data Preparation & Feature Engineering

Data Engineers and Data Scientists

Model Development, Training, and Tuning

Data Scientists and ML Engineers

Model Deployment & Monitoring

ML Ops Engineers

SAGEMAKER CANVAS

No-Code ML

Analyst-focused Workspace

Collaboration

SAGEMAKER STUDIO | RSTUDIO

- Prepare data
- Store features
- Detect bias
- Build with notebooks
- Train models faster
- Tune parameters
- Deploy in production
- Explain predictions
- Manage and monitor

Data Scientist-focused Workspace

Getting Started and Documentation

SCALE ML VALUE CREATION

Business Executives

Scale your ML efforts by offering your business teams a way to build ML with no code

BUILD ML WITH NO CODE

Business Analysts & Domain Experts

Learn how to use Canvas and start building ML-powered predictions with no code

ENABLE COLLABORATION

Platform Admins

Onboard analysts in Canvas and data scientists in Studio to enable collaboration

http://aws.amazon.com/sagemaker/canvas

aws

Additional Resources

Want to test on your own?

SageMaker Canvas provides a Free Tier for two months:

- Interactive sessions up to 750 hours/month
- Up to 10 model creation requests/month, each with up to 1M cells/model

Want to know more about Low-Code/No-Code ML at AWS?

- Amazon SageMaker JumpStart
https://aws.amazon.com/sagemaker/jumpstart/

- Amazon SageMaker Data Wrangler
https://aws.amazon.com/sagemaker/data-wrangler/

- Amazon SageMaker AutoPilot
https://aws.amazon.com/sagemaker/autopilot/

Thank You

Appreciate your feedback!

Please help to fill in the survey ☺

AWS Event Survey

https://bit.ly/39QQ1ux

Aaron Chong
aachong@amazon.com
https://www.linkedin.com/in/aaronchong888/