Fraud Prediction with AWS Cloud

MSBA 6330 Group 7

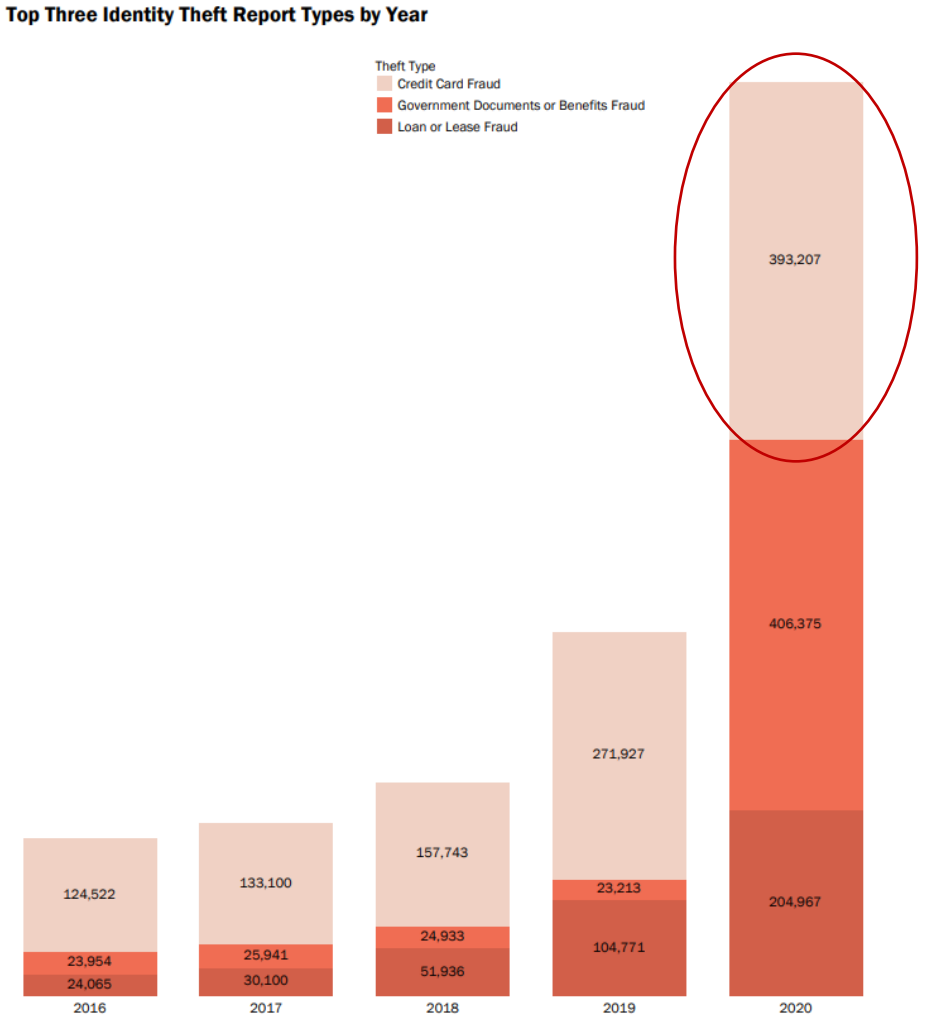
Grant Barland, Jacob Braun, Ravi Hasyagar, Shuyun Liu, Huiyuan Xiong, Jiayi Wang

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Credit Card Fraud in the US

- 393,207 instances of credit card fraud were reported in the US in 2020 alone
- Victims and banks lost over $300M
- This is an increase of 48% year over year

- Banks and credit card companies employ predictive modeling to detect card fraud

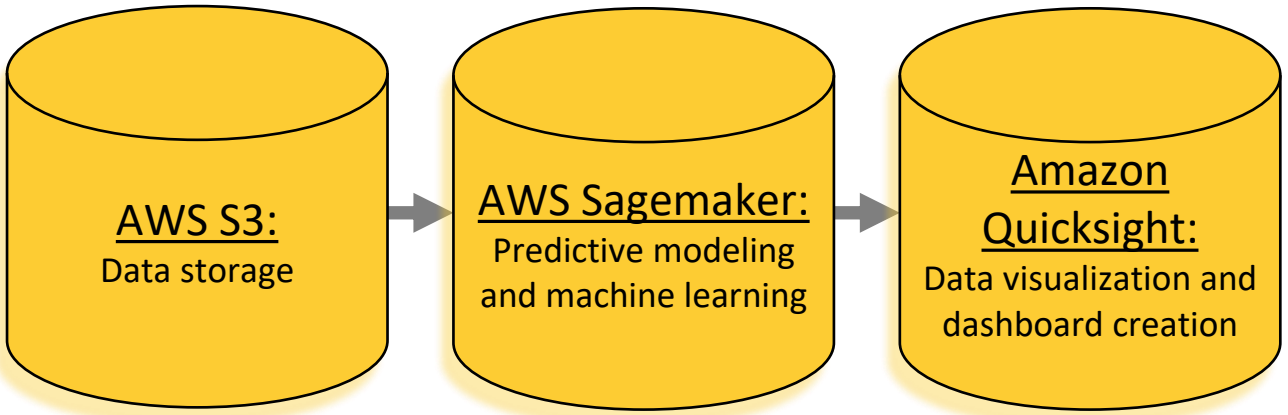


Source: FTC Annual Data Book
https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2020/csn_annual_data_book_2020.pdf

Data Processing Workflow

Customer Transaction Data:
- From Kaggle
- 1.5M records
- Fraud/Not Fraud labels

Data Pipeline:



Outcomes:

✓ Real-time fraud detection and prediction

✓ Executive dashboard

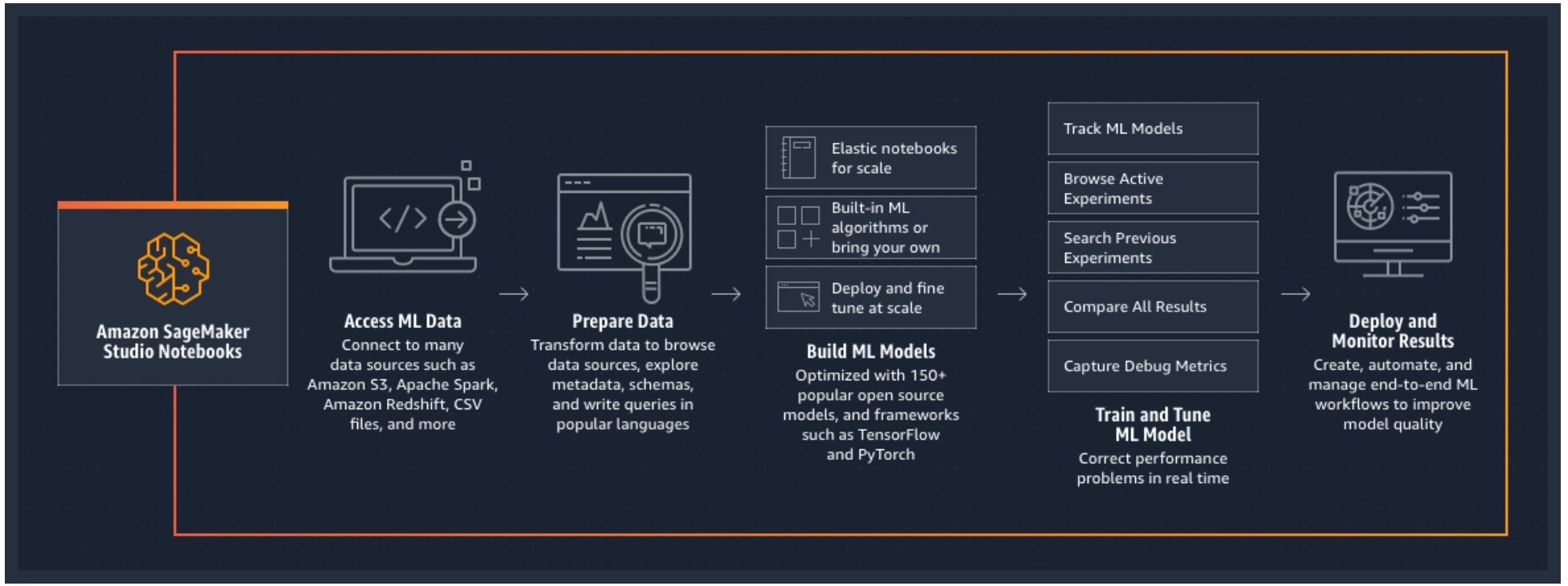
✓ Deeper understanding of customer groups

AWS S3

- Amazon Simple Storage Service (S3)
- Benefits: Scalability, security, access controls, backup and recovery
- We uploaded our dataset as a CSV to an S3 bucket to be accessed by the AWS suite

Source: AWS S3
https://aws.amazon.com/s3/?nc2=h_ql_prod_fs_s3

AWS Sagemaker



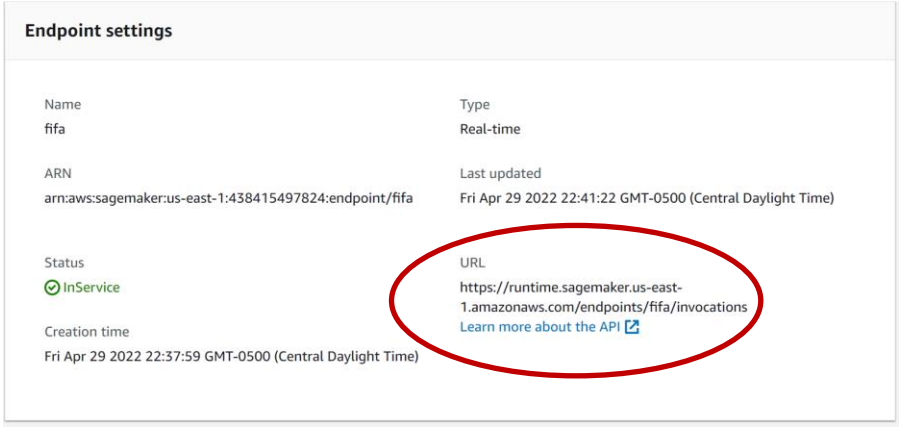
- Cloud-based machine learning solution
- Develop models in Jupyter using Python or use SageMaker's autopilot to fine-tune automatically
- Create shareable endpoint to connect additional data services

Source: AWS Sagemaker
https://aws.amazon.com/sagemaker/data-scientist/

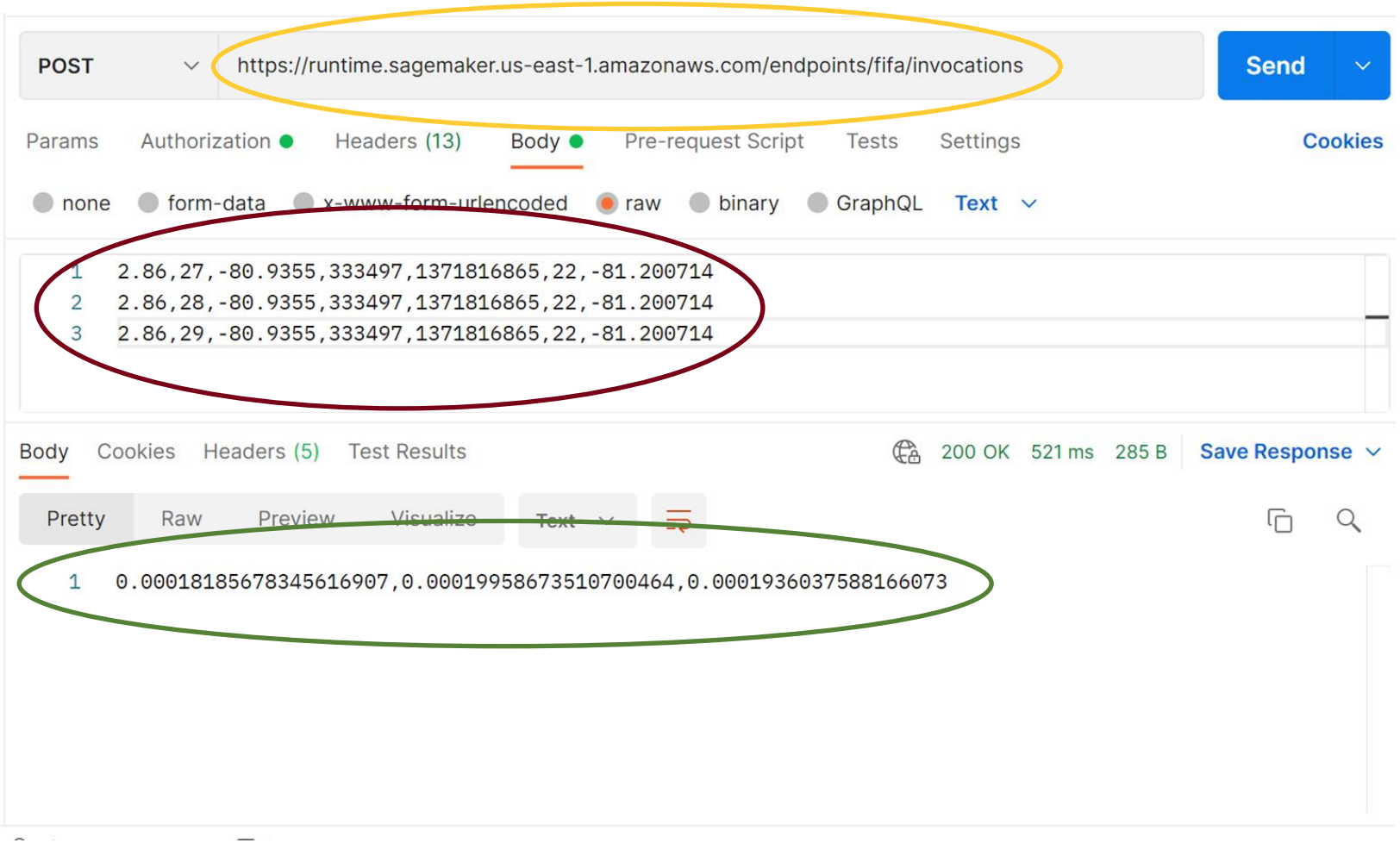
Our AWS Sagemaker Workflow

```
In [5]: xgb = sagemaker.estimator.Estimator(container,
                                    role,
                                    train_instance_count=1,
                                    train_instance_type='ml.m4.xlarge',
                                    output_path= 's3://trends-marketplace-data/fraud-detection/modelDir/model.joblib/',
                                    sagemaker_session=sess)
        xgb.set_hyperparameters(max_depth=10,\
                                objective='binary:logistic', \
                                eta =0.2,\
                                gamma=0,\
                                min_child_weight=1,\
                                max_delta_step=0,\
                                subsample=0.85, \
                                colsample_bytree=0.7,\
                                colsample_bylevel=1,\
                                alpha=0,\
                                scale_pos_weight=1,\
                                num_round = 100,\
                                seed=42)

        xgb.fit({'train': s3_input_train})

train_instance_count has been renamed in sagemaker>=2.
See: https://sagemaker.readthedocs.io/en/stable/v2.html for details.
train_instance_type has been renamed in sagemaker>=2.
See: https://sagemaker.readthedocs.io/en/stable/v2.html for details.

2022-04-30 03:16:30 Starting - Starting the training job...ProfilerReport-1651288590: InProgress
...
2022-04-30 03:17:13 Starting - Preparing the instances for training.........
2022-04-30 03:18:54 Downloading - Downloading input data...
2022-04-30 03:19:16 Training - Downloading the training image......
2022-04-30 03:20:17 Training - Training image download completed. Training in progress.INFO:sagemaker-containers:Imported fr
amework sagemaker_xgboost_container.training
INFO:sagemaker-containers:Failed to parse hyperparameter objective value binary:logistic to Json.
Returning the value itself
INFO:sagemaker-containers:No GPUs detected (normal if no gpus installed)
INFO:sagemaker_xgboost_container.training:Running XGBoost Sagemaker in algorithm mode
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Determined delimiter of CSV input is ','
[03:20:22] 555719x7 matrix with 3890033 entries loaded from /opt/ml/input/data/train?format=csv&label_column=0&delimiter=,
INFO:root:Single node training.

In [11]: xgb_predictor = xgb.deploy(initial_instance_count=1, instance_type='ml.t2.medium', endpoint_name = 'fifa')

---------!
```

- XGBoost model developed using Python hosted in AWS Jupyter Notebook instance
- Specified model endpoint allows users to connect to the trained model and make predictions on new data



Our AWS Sagemaker Workflow



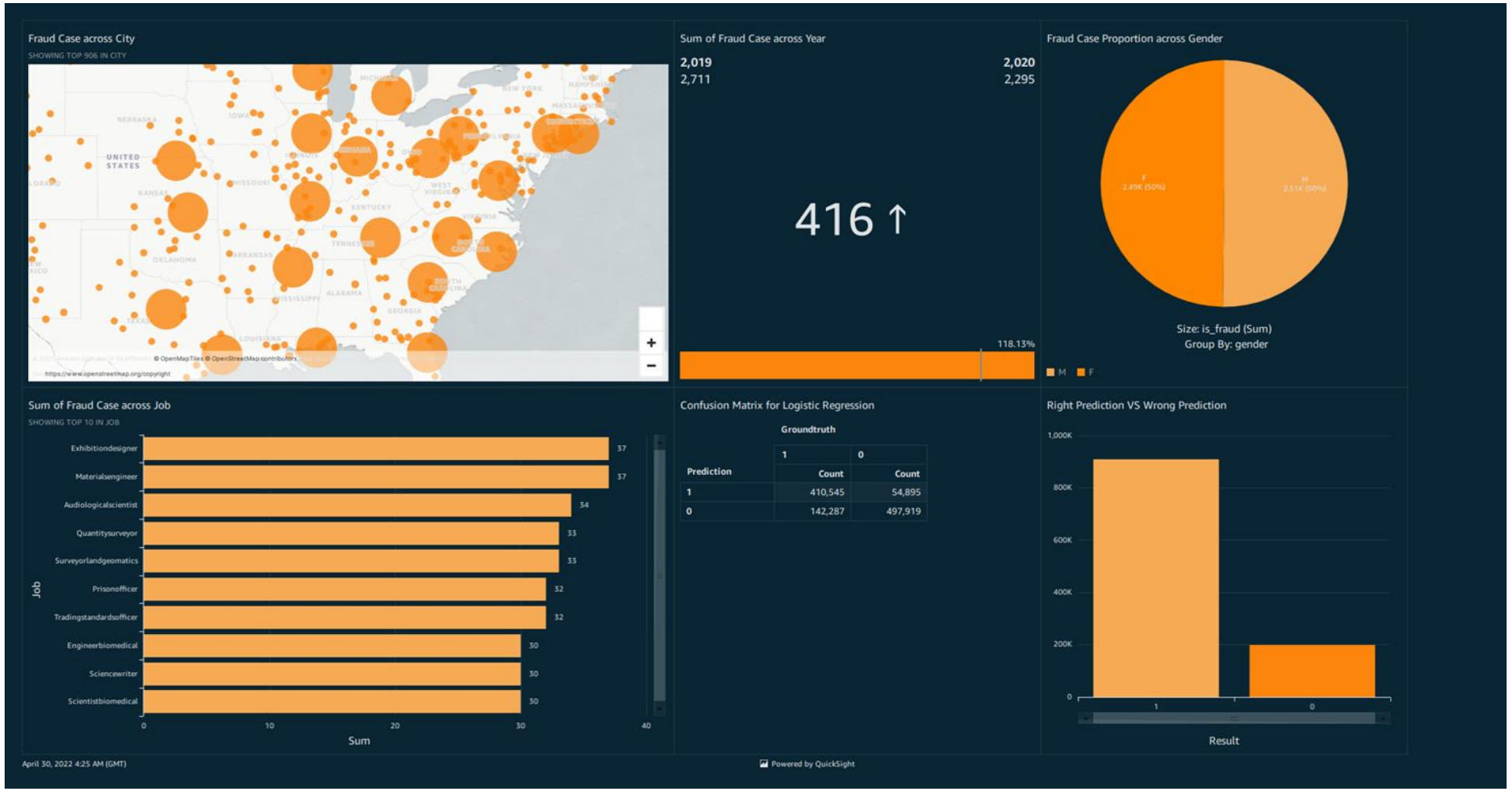
- Using the model's endpoint, we can feed new input data and quickly receive new predictions
- Quicksight can connect and update real-time

Amazon Quicksight

- Cloud-based BI service

- Benefits:
  - Intuitive GUI
  - Easy connection to S3, Sagemaker, and more
  - Auto-scaling compute resources
  - Servers managed by Amazon for easy maintenance
  - Built-in security and access controls

Our Amazon Quicksight Dashboard



- Visuals update in real time as data changes
- Showcase and monitor predictive performance
- Different views and access for different users in an organization

Benefits of AWS Ecosystem

	Public Cloud: Flexible Storage and Compute	**Machine Learning Focus:** Features and support for ML	**All-in-One Corporate Suite:** Non-data tools	**Breadth of Services:** Flexible Storage and Compute	**Customer Profile:** Volume of high-profile users
AWS	✓	✓		✓	✓
Microsoft Azure	✓	✓	✓		
Google Cloud	✓	✓			

- AWS offers over 175 services covering storage, compute, networking, developer and management tools, security, and more
- Trusted by the industry – users include Netflix, Dow Jones, AirBnB, Nike, Pfizer

Thank you!

Questions?

Grant Barland, Jacob Braun, Ravi Hasyagar, Shuyun Liu, Huiyuan Xiong, Jiayi Wang