

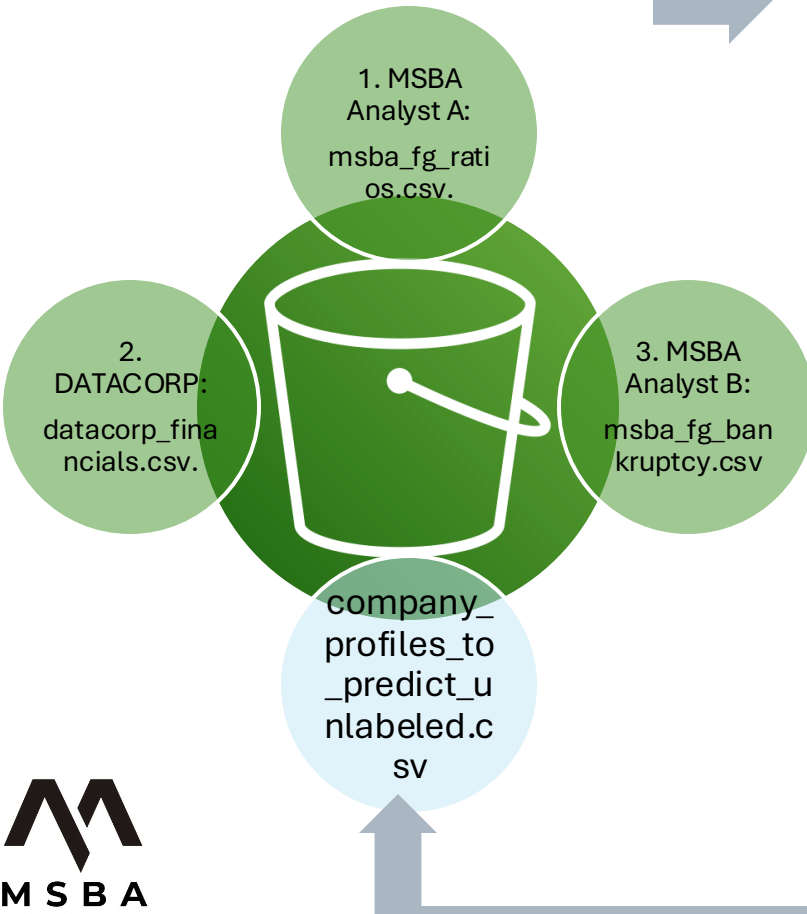


Cloud Native Data Architecture

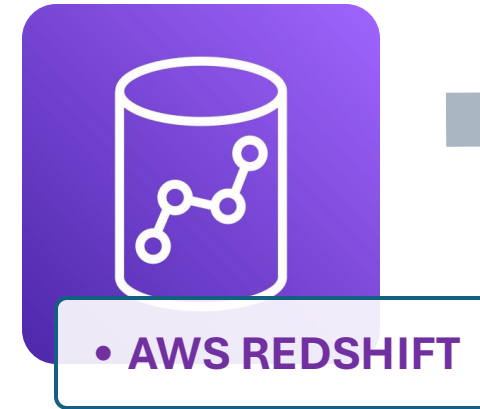


Data Flow Diagram_Cloud Native Data Architecture

1. AMAZON S3 BUCKETS / DATA LAKE



- Perform our Extract, Transform and Load process.
- Make future data ingestion easy to replicate or automate.
- Joining ratios and financials data to a single table.



- “Single source of truth”**
- Data Warehouse
 - 1. Create a Redshift Cluster: msba-fingroup-cluster
 - 2. Use Query Editor to:
 - Create New tables to place data we transformed.



- Create a Model using AWS Canvas.
- Review model performance, insights, analyze metrics.
- Predict probability of bankruptcy for new set of companies.





Data Lake: S3-Bucket

- Our data lake is the landing place for all our data, we can store all types of files, securely. Buckets allow to store data for various data types, sizes, and sources.

Amazon S3 > Buckets > msba-fingroup-232

msba-fingroup-232

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (2) Info

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	data_files/	Folder	-	-	-
<input type="checkbox"/>	prediction/	Folder	-	-	-

Amazon S3 > Buckets > msba-fingroup-232 > data_files/

data_files/

Objects | Properties

Objects (3) Info

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	datacorp_financial_data.csv	csv	October 4, 2024, 19:31:52 (UTC-05:00)	1.2 MB	Standard
<input type="checkbox"/>	msba_fg_bankruptcy.txt	txt	October 4, 2024, 19:31:53 (UTC-05:00)	53.3 KB	Standard
<input type="checkbox"/>	msba_fg_ratio_data.csv	csv	October 4, 2024, 19:31:53 (UTC-05:00)	614.2 KB	Standard

Amazon S3 > Buckets > msba-fingroup-232 > prediction/

prediction/

Objects | Properties

Objects (1) Info

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	company_profiles_to_predict_unlabeled.csv	csv	October 4, 2024, 19:32:15 (UTC-05:00)	3.3 KB	Standard



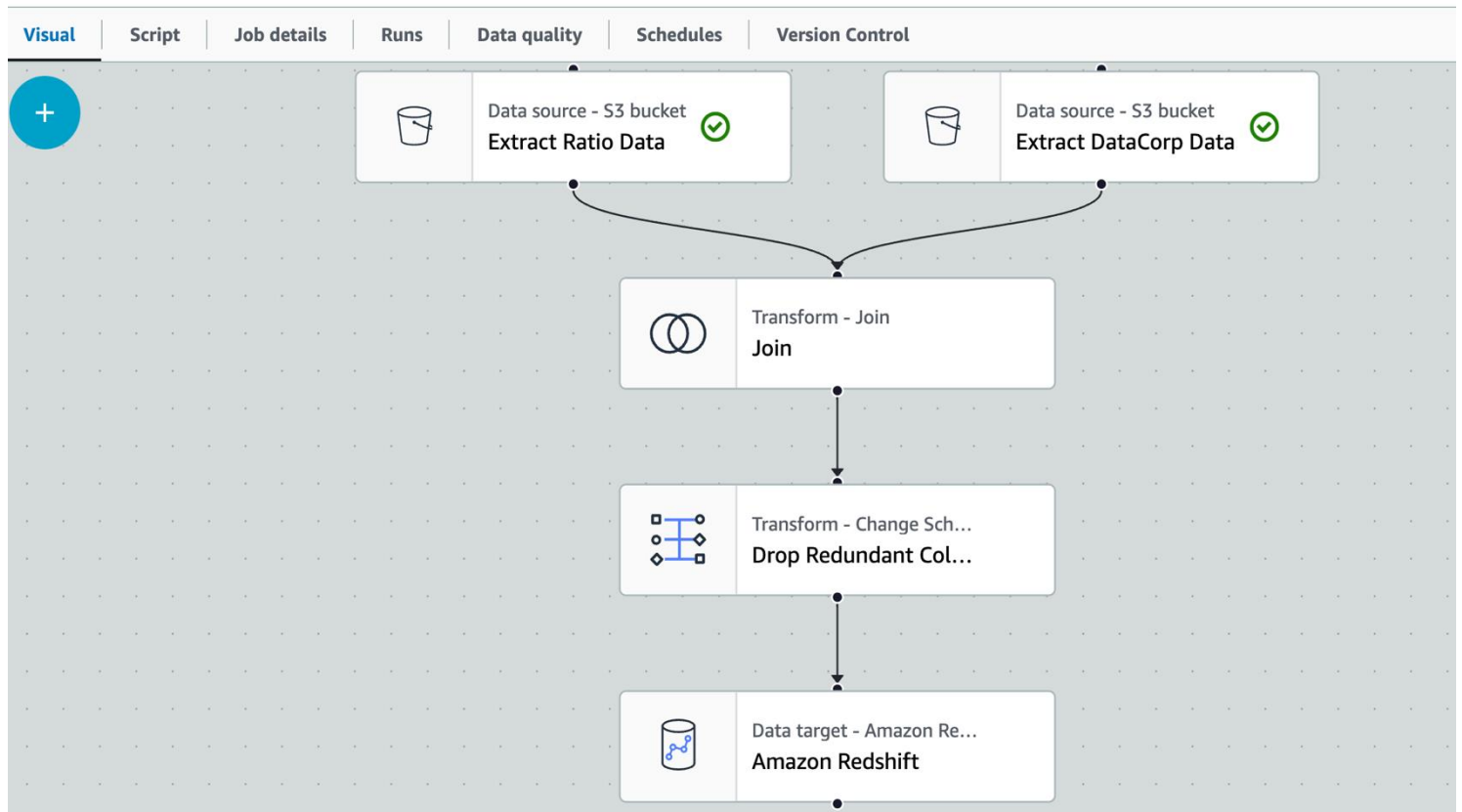
AWS GLUE

- This tool allows us to integrate data.
- For this prototype we want to join data from the files: “msba_fg_ratios.csv” and the “datacorp_financials.csv”
- After joining the tables we will drop redundant columns.
- We will load the final table to our data warehouse, into the table called “financials_combined”.

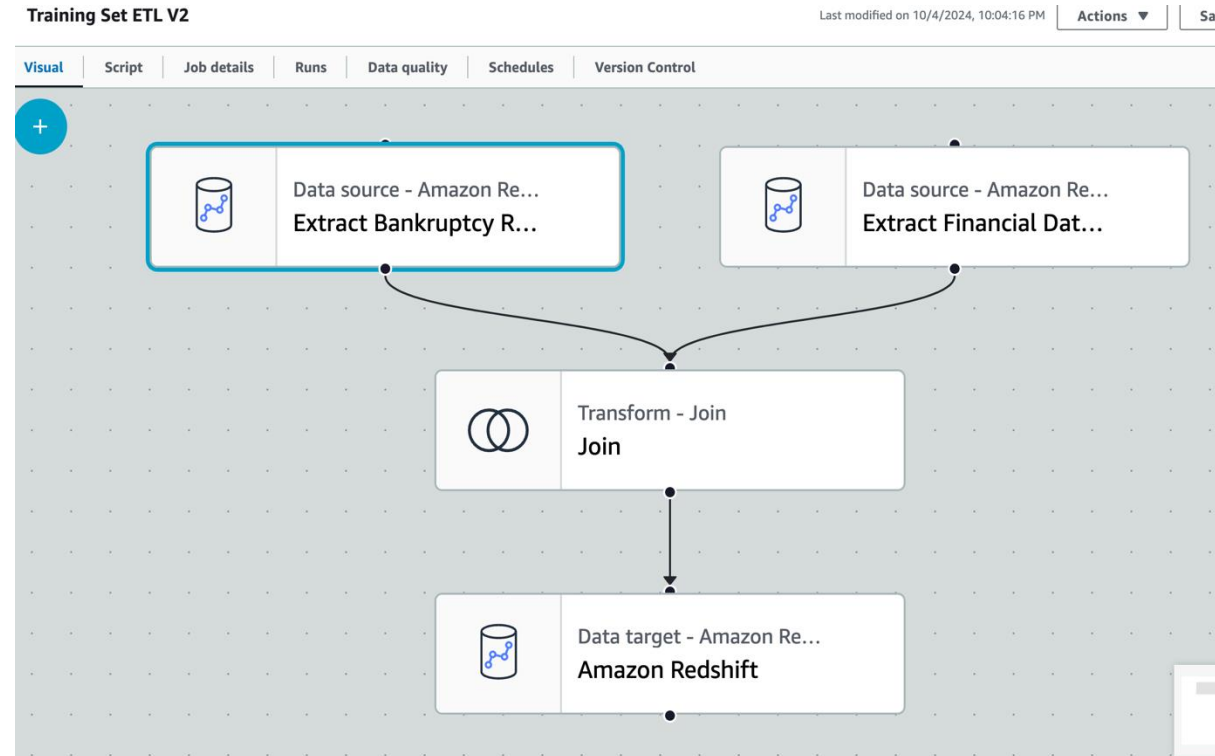
MSBA Fingroup Financial ETL V2

Last modified on 10/4/2024, 9:01:37 PM

Action



- The second joining will be, joining our Bankruptcy table with the financials combined table.
- After performing a left join, we will load the information to our data warehouse, more specifically to a table called training_set.



AWS – Glue - Training Set ETL





Redshift / Data Warehouse

“Single source of Truth”

Redshift / Data Warehouse is a central repository of integrated data.

Pros:

- Provides a consistent view of an organization.
- Integrated view for value added reporting.
- DW in the Cloud benefits: affordable, scalable and availability.
- Can connect to AWS Sagemaker to build, tune and train models.

The screenshot shows the Amazon Redshift console interface. On the left is a navigation menu with options like 'Redshift Serverless', 'Provisioned clusters dashboard', 'Clusters', 'Query editor', 'Datashares', 'IAM Identity Center connections', 'Configurations', and 'AWS Partner Integration'. The main content area is titled 'Connect to Redshift clusters' and includes sections for 'Query data using Redshift query editor', 'Work with your client tools', and 'Choose your JDBC or ODBC driver'. Below these, there's a 'Clusters (1) Info' section with a table listing the cluster 'msba-fingroup-cluster'.

Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ	Storage capacity us...	CPU utili...
msba-fingroup-cluster dc2.large 2 nodes 320 GB	Restoring	ef1b2821-5510-4f11-...	us-east-1e	No	0 %	0 %

- Cluster creation.
- Cluster: computing nodes, used to run a data warehouse. Clusters can be scaled up and can run complex queries.
- Cluster name: msba-fingroup-cluster

Redshift – Query Editor

The Query Editor tool in Redshift allows us to run SQL queries.

We will create three tables in our data warehouse:

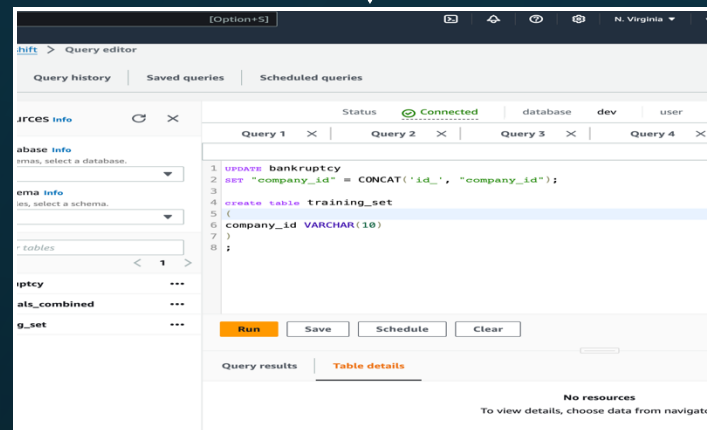
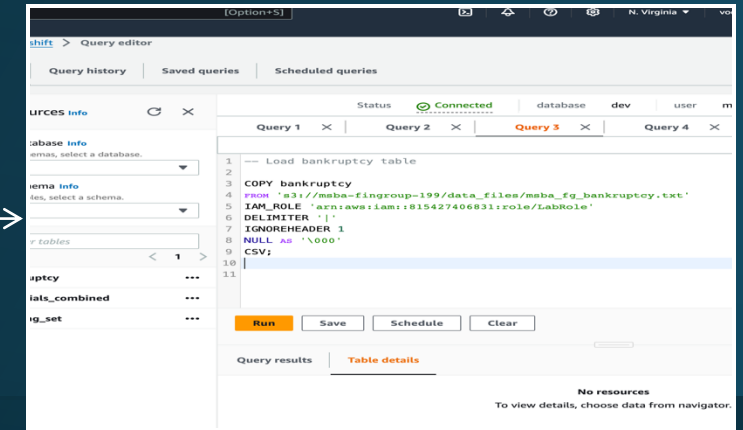
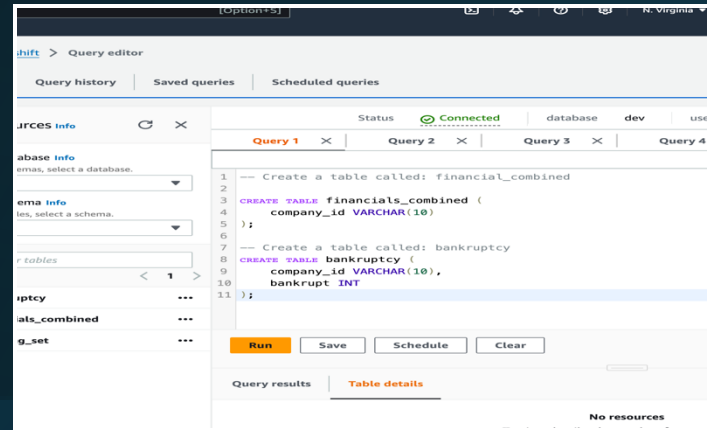
Table 1: “financials_combined”

Table 2: “bankruptcy”

Table 3: “training_set”

Load data:

Into “bankruptcy” table.



Review / Financials Combined

Amazon Redshift

Redshift Serverless [New](#)

Provisioned clusters dashboard

Clusters

Query editor

Query editor v2 [↗](#)

Queries and loads

Datashares

Zero-ETL integrations [New](#)

IAM Identity Center connections [New](#)

Configurations

AWS Partner Integration

Informatica Data Load & Transform

Advisor

AWS Marketplace

Alarms

Events

What's new 8

Select database [Info](#)

To view schemas, select a database.

dev

Select schema [Info](#)

To view tables, select a schema.

public

Filter tables

1

bankruptcy

financials_combined

training_set

Query 1 X

Query 2 X

Query 3 X

Query 4 X

Query 6 X

Query 7 X

+

1 SELECT * FROM financials_combined LIMIT(10);

Run

Save

Schedule

Clear

Send feedback

Query results

Table details

Query [606409](#)

Completed, started on October 04, 2024 at 18:12:12

ELAPSED TIME: 00 m 02 s

Execution

Data

Visualize

Rows returned (10)

Export

Search rows

1

company_id	liability_to_equity	net_income_to_total_assets	working_capital_to_total_assets	net_profit_before_tax_to_paid_in_capital	operating_profit_per_share	net_worth
id_0019	0.280838655	0.806263915	0.766580251	0.175062213	0.102108949	0.8586631
id_0046	0.277884763	0.806666538	0.775479414	0.175929417	0.102678935	0.9057652
id_0096	0.277826354	0.827878672	0.78204296	0.185257522	0.113182966	0.9069746
id_0107	0.287759581	0.796483203	0.729383528	0.169361285	0.099503298	0.8036097
id_0143	0.288862482	0.768874463	0.751484002	0.156172234	0.100724697	0.7982282
id_0144	0.277866374	0.806666538	0.775479414	0.175929417	0.102678935	0.9057652
id_0145	0.277866374	0.806666538	0.775479414	0.175929417	0.102678935	0.9057652
id_0146	0.277866374	0.806666538	0.775479414	0.175929417	0.102678935	0.9057652
id_0147	0.277866374	0.806666538	0.775479414	0.175929417	0.102678935	0.9057652
id_0148	0.277866374	0.806666538	0.775479414	0.175929417	0.102678935	0.9057652



Canva: Model Building

- Using our training_set data we will build our model.
- Sagemaker, will try different models, until it finds the model that will make the best predictions.

Home

Data Wrangler

Datasets

My Models

ML Ops

Ready-to-use

Gen AI

Help

Datasets > msba_financial_prediction V1

+ Create a data flow

Update dataset

+ Create a model

Data

Version history

Auto update

Dataset details

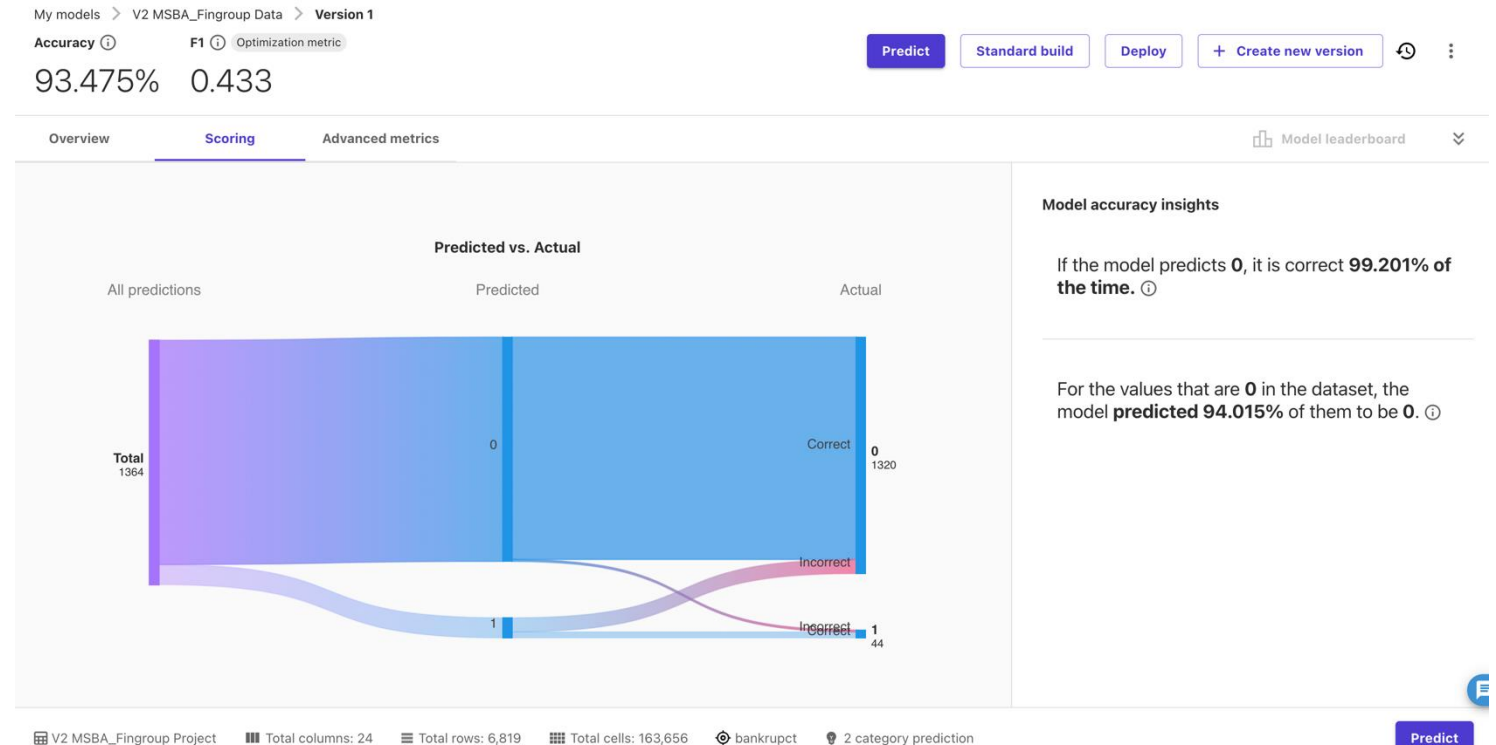
Previewing up to the first 100 rows of msba_financial_prediction

borrowing_depe...	company	company_id	current_liabilitie...	current_liability...	debt_ratio_perc...	equity_to_longt...	liability_to_eq
0.458818609	western corp	id_6988	0.372217526	0.060766259	0.269038909	0.216878232	0.337315248
0.37930429	design solutions	id_7413	0.333345174	0.041219716	0.16186474	0.120812062	0.282763173
0.384998982	innocore	id_8801	0.337392013	0.060765125	0.216101823	0.120561366	0.292504124
0.374219105	pharmasolve	id_9614	0.329803726	0.030201105	0.108202074	0.114507969	0.278607306
0.370253398	ninetech	id_9131	0.328092756	0.021710461	0.058590561	0.110933234	0.276422514
0.37450876	songster inc	id_7102	0.330409488	0.025494302	0.121292741	0.115498978	0.279387519
0.374179962	rogers and sons	id_7012	0.327484654	0.047166348	0.103576503	0.120942627	0.278356432
0.373113046	Hallandall ag.	id_9904	0.328042001	0.033711602	0.094385827	0.116458578	0.277892082
0.377306898	Foster & Kruse	id_6905	0.331114215	0.04351399	0.144662454	0.116955695	0.281113384
0.37315107	Highwood & Hart	id_8039	0.329035251	0.028901672	0.10656952	0.11357108	0.278517753



Canvas: Model Building

- The model gives us an accuracy of 93.475%.
- The data set is unbalanced, that means that it has more non-bankruptcies than bankruptcies. It is a good idea to review other metrics, like AUC-Roc.
- The AUC-ROC shows a 0.922 which means the model can differentiate between the two classes.



F1 ⓘ Optimization metric

43.312%

Accuracy ⓘ

93.475%

Precision ⓘ

30.088%

Recall ⓘ

77.273%

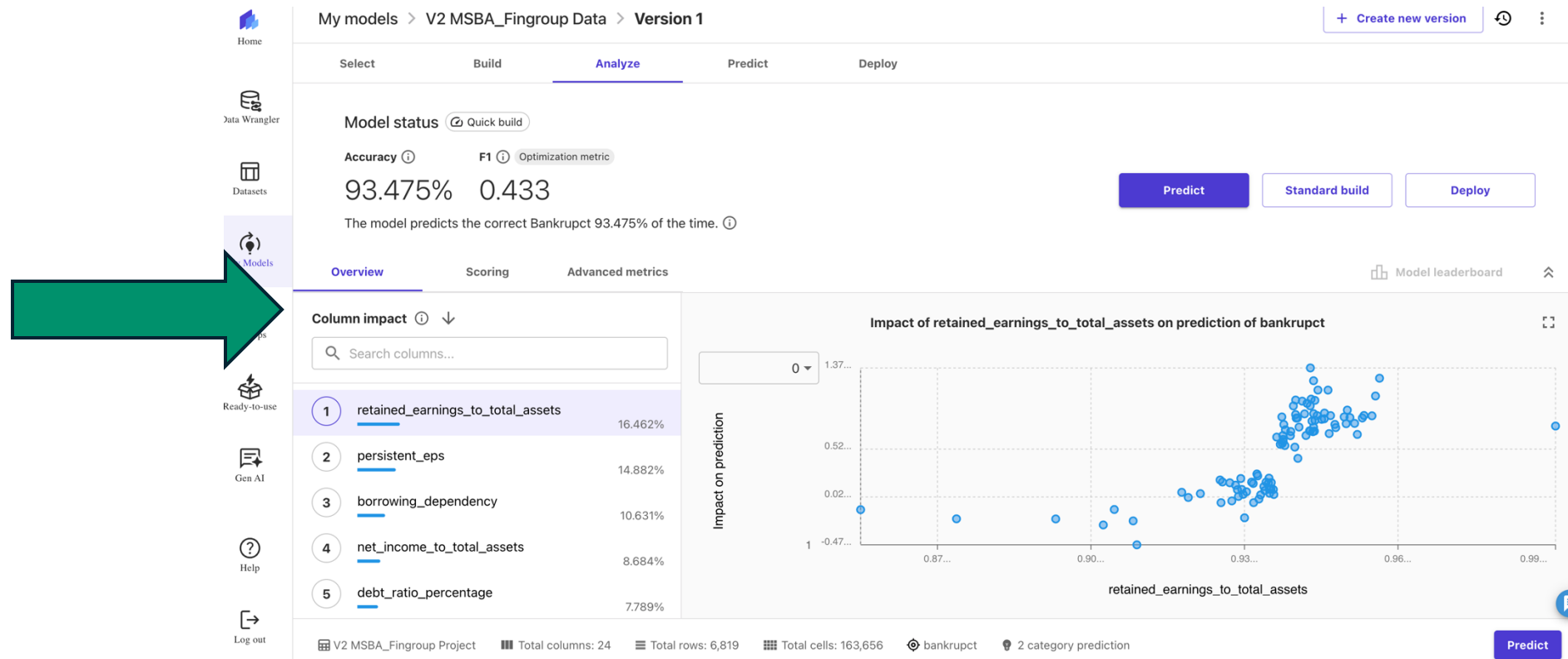
AUC-ROC ⓘ

0.922



Analysis

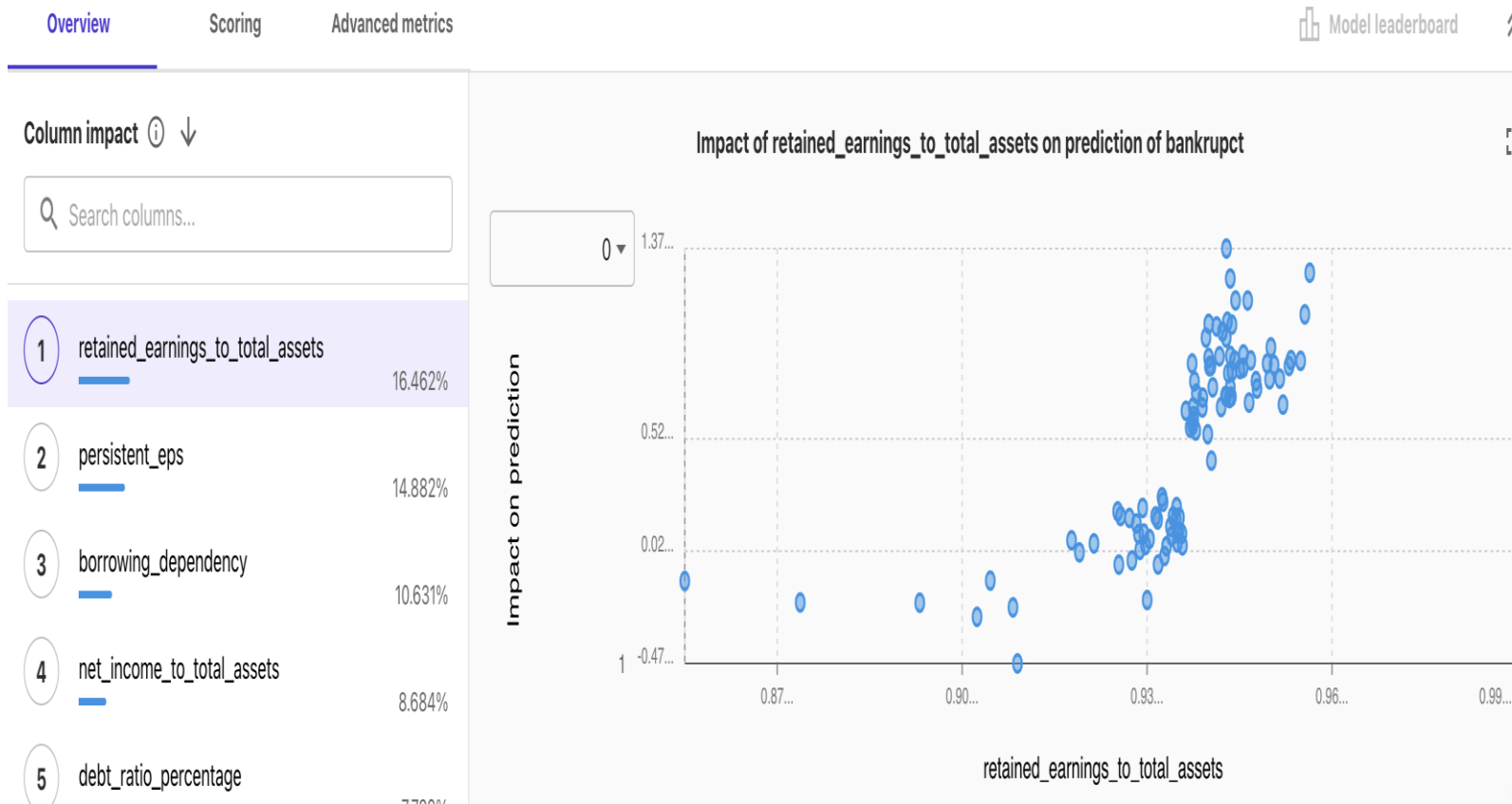
- Sagemaker, shows us which features have the highest impact on the model's ability to predict bankruptcy.





Analysis

- Three most impactful features:



Retained Earnings to Total Assets

This financial metric has a 16.46% impact on our model's ability to predict bankruptcy.

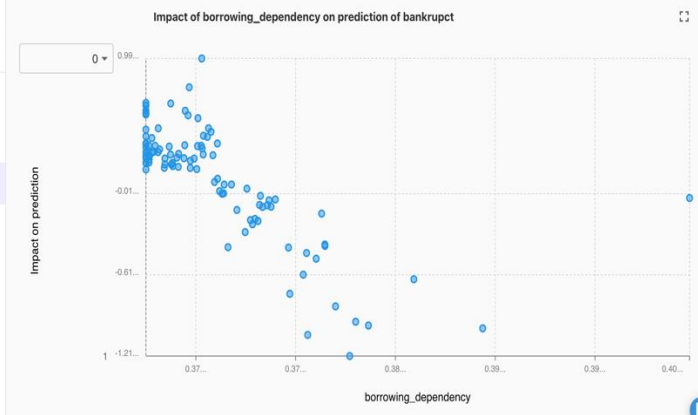
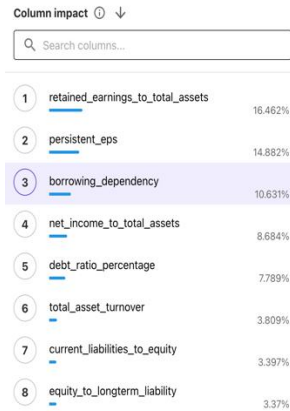
It is a very important financial indicator, because it shows:

- Company's Long Term Profitability.
- Ability of a Company to Internally Finance Operations (makes the company less dependable on external funding).
- Stability.
- Company re-investment.

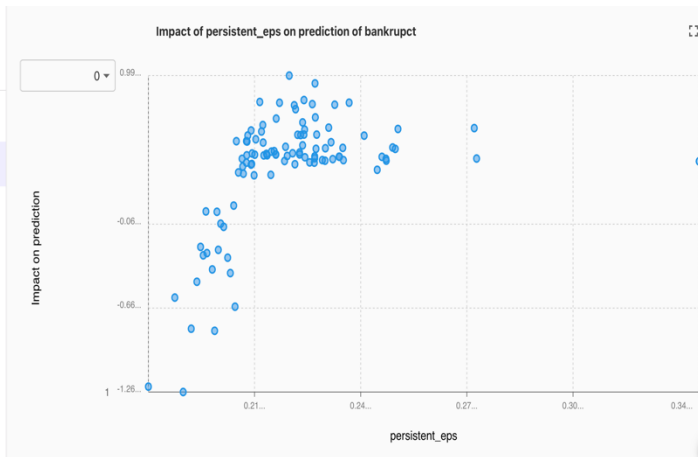
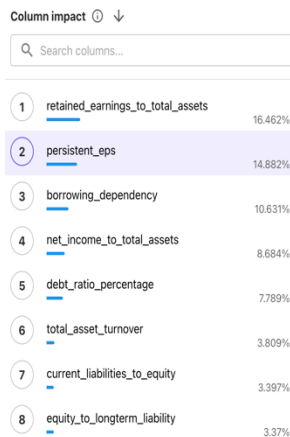
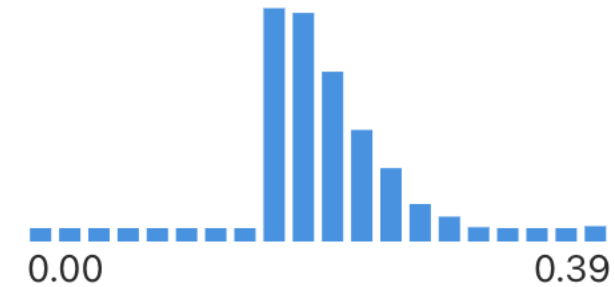


Analysis

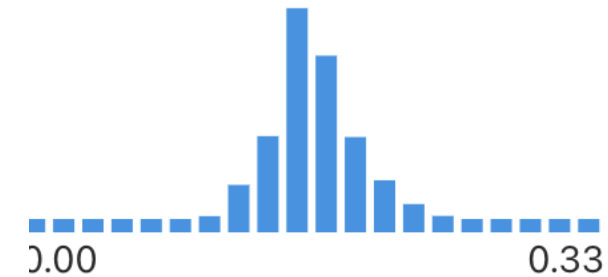
- Three most impactful features:



borrowing_dependency 123



persistent_eps 123

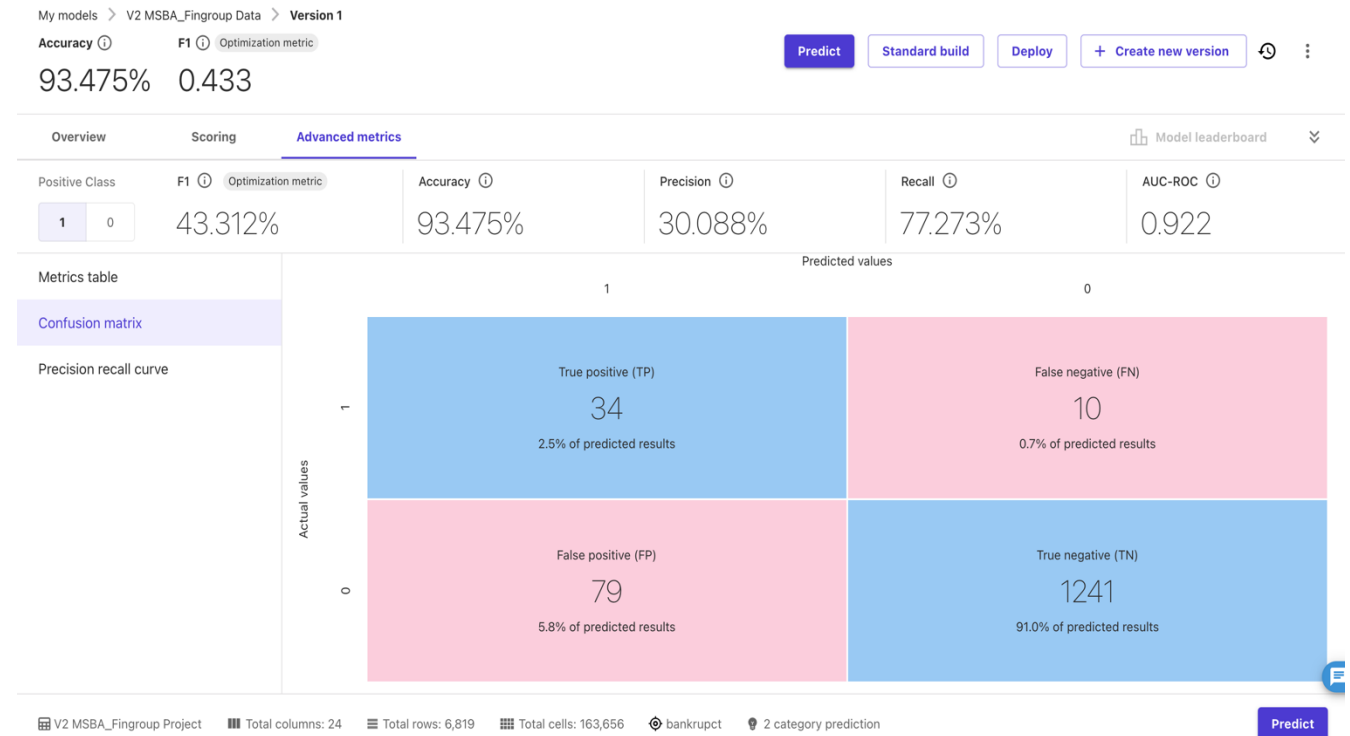




Analysis / Confusion Matrix

- There were 79 false positives, which means that the model predicted 79 companies going bankrupt that did not.
- Also, the model shows 10 false negatives, that means that the model wrongly predicted 10 companies that went bankrupt.
- Our recall metric is at 77.273%, that indicated the model is doing a good job at identifying companies going bankrupt.

*This must be analyzed, as the cost of predicting a company going bankrupt and not occurring (missed opportunity cost) in my opinion is less than investing in a company and having it go bankrupt. False positives might be less damaging than false negatives, as we want to avoid high-risk investments.



Predictions

- Utilizing the model built, we will not predict the probability of bankruptcy for a new set of companies the MSBA is looking to invest in.
- Based on the results is not recommended for MSBA to invest in Western Corp, Design Solutions and Innocore.

bankrupt	probability	borrowing_dependency	company	company_id
1	0.9553555250167850	0.458818609	western corp	id_6988
1	0.7948915362358090	0.37930429	design solutions	id_7413
1	0.8425660729408260	0.384998982	innocore	id_8801
0	0.972851574420929	0.374219105	pharmasolve	id_9614
0	0.9804420471191410	0.370253398	ninetech	id_9131
0	0.9818169474601750	0.37450876	songster inc	id_7102
0	0.9822726249694820	0.374179962	rogers and sons	id_7012
0	0.9725237488746640	0.373113046	Hallandall ag.	id_9904
0	0.9375975131988530	0.377306898	Foster & Kruse	id_6905
0	0.978073000907898	0.37315107	Highwood & Hart	id_8039



Recommendation

- Based on our model's prediction it is recommended to invest in:
 - Rogers & Sons
 - Songster Inc.
 - Pharmasolve

bankrupt	probability	borrowing_dependency	company	company_id	persistent_eps	retained_earnings_to_total_assets
0	0.9822726249694824	0.374179962	rogers and sons	id_7012	0.21896568	0.942828956
0	0.9818169474601746	0.37450876	songster inc	id_7102	0.217831143	0.938289897
0	0.972851574	0.374219105	pharmasolve	id_9614	0.225205635	0.935449309
0	0.9804420471191406	0.370253398	ninetech	id_9131	0.218398412	0.935200209
0	0.9725237488746643	0.373113046	Hallandall ag.	id_9904	0.20468942	0.932955215
0	0.978073001	0.37315107	Highwood & Hart	id_8039	0.212914815	0.930763203
0	0.9375975131988525	0.377306898	Foster & Kruse	id_6905	0.219343859	0.930139635

