December 17 2024

Predicting Store Sales

Team Texas:

Data Architect: Bethun Bhowmik, Jong Hur
Data Engineer: Michael Cavallaro, Edwin Tembo,
Data Scientist: Imran Naskani, Davar Jamali,
BI Analyst: Pranab Nepal

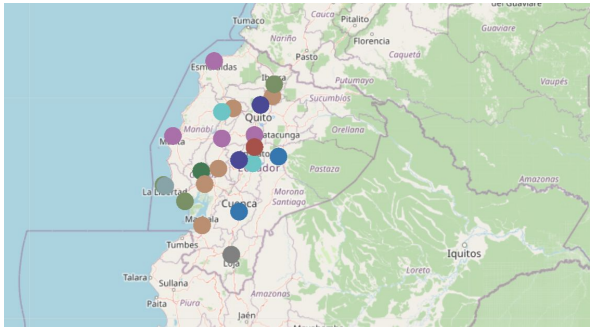
CSCI E-103 **Data Engineering for Analytics to Solve Business Challenges**
Fall 2024
Harvard Extension School
Harvard University

Group 5 Information - Team Texas

Name	**Role**	**Tasks**
Bethun Bhowmik	**Data Architect**	ERD / Partitioning
Jong Hur	**Data Architect**	DR / CI/CD Pipeline
Michael Cavallaro	**Data Engineer**	Ingest bronze data/ clean silver / merge gold
Edwin Tembo	**Data Engineer**	Streaming CSV to bronze / Workflows
Imran Naskani	**Data Scientist/ Team Leader**	Developed Model / ML Flow
Davar Jamali	**Data Scientist**	Developed Model / ML Flow
Pranab Nepal	**BI Analyst**	Developed Dashboard

Team Texas met on November 30th, December 7th, 14th-16th. Individual break out groups also met; Data Engineering met on December 4th

Problem Statement



01 Business Use Case

- Assist **Corporación Favorita**, a large Ecuadorian grocery retailer, by leveraging a **data lakehouse** to consolidate multiple data sources and provide actionable insights for sales forecasting

02 Problem

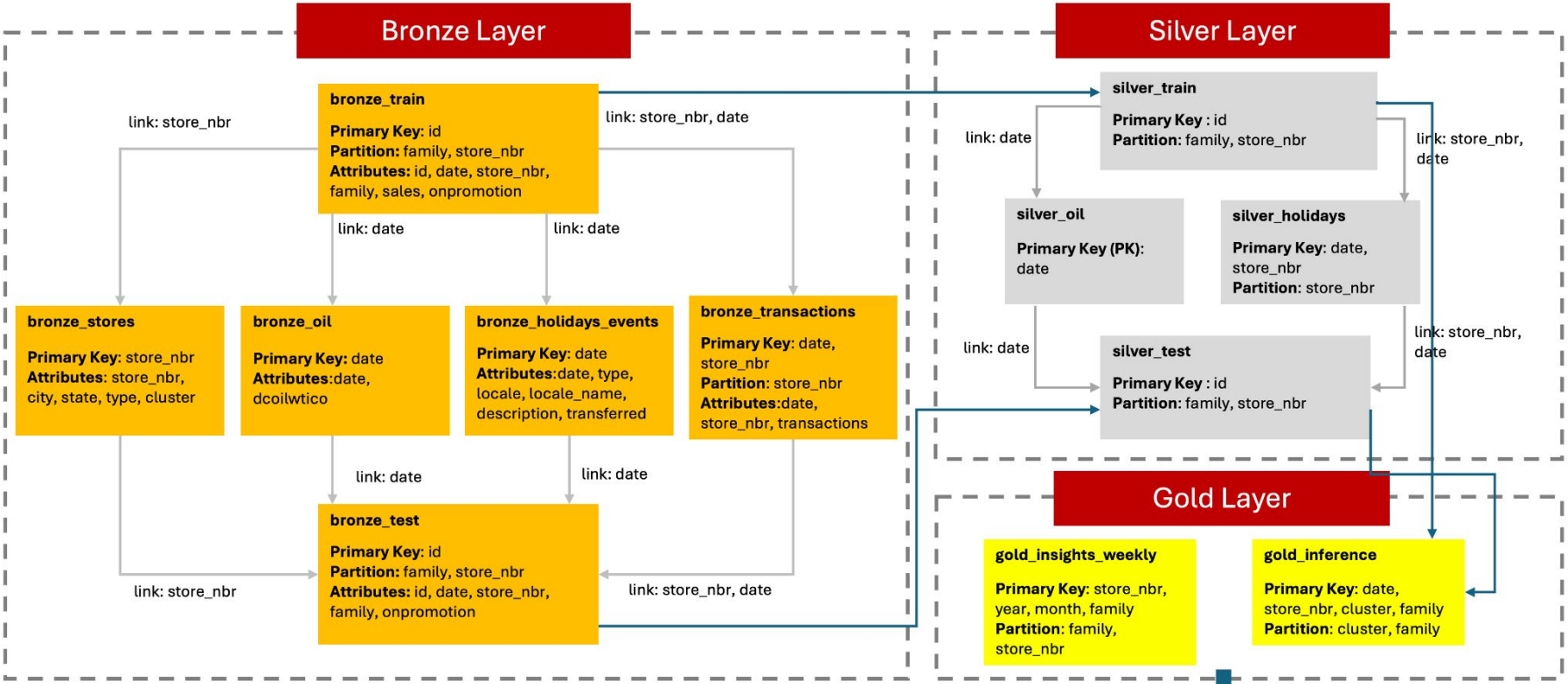
Retail grocery sales fluctuate due to promotions, holidays, and macroeconomic factors such as oil prices. Managing these fluctuations is critical to:

- minimizing stockouts
- minimizing shelf life and maximizing revenue
- understanding consumer behavior across stores, clusters, and product categories

The goal is to build a unified data lakehouse that:

- ingest data from multiple sources
- cleanse and transform data into a structured format
- enable sales forecasting and visualization on dashboards

Entity relationship diagram for Store Sales dataset

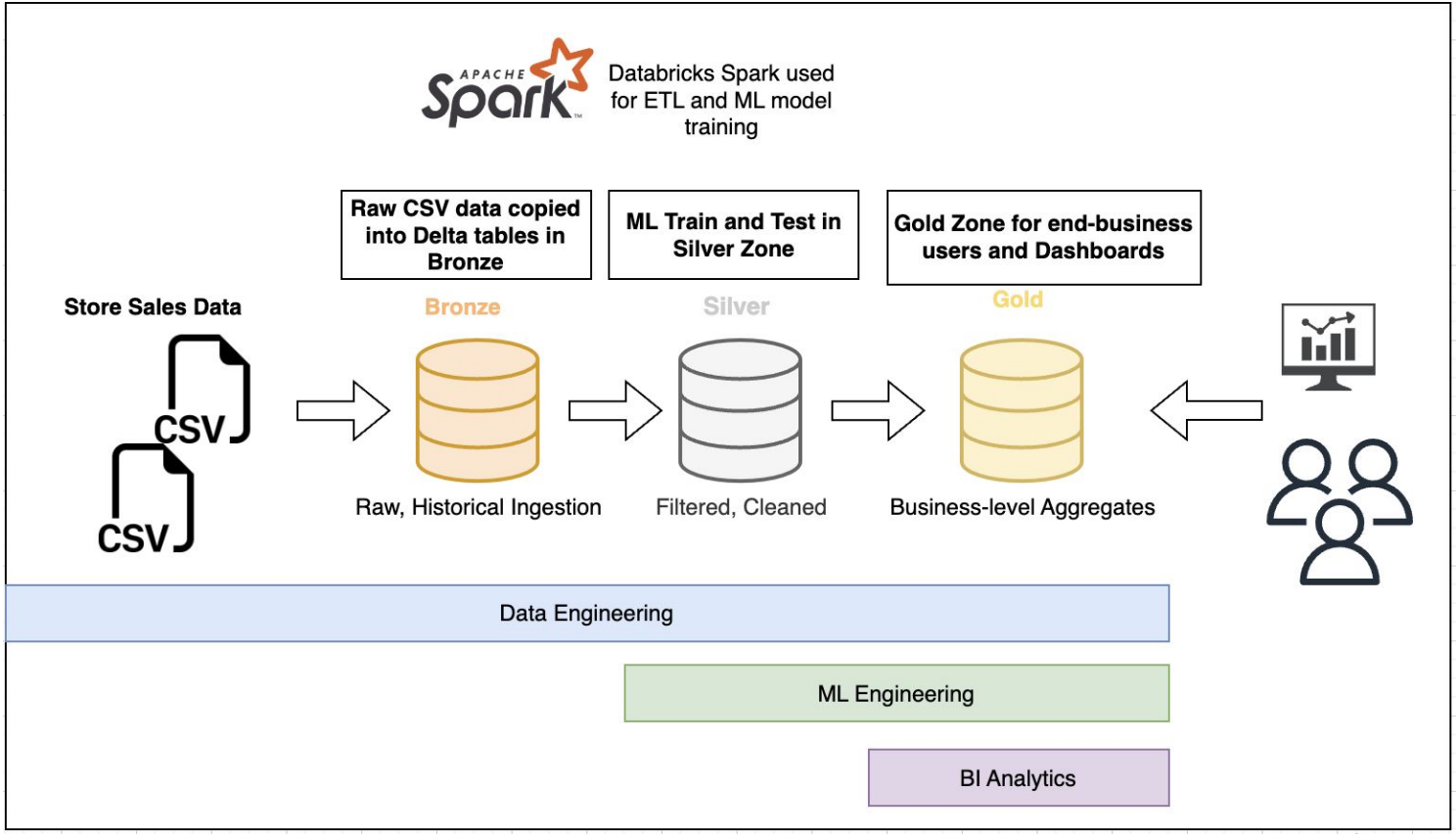


Relationships

- **bronze_train ↔ silver_holidays**: date, store_nbr
- **bronze_train ↔ silver_oil**: date
- **bronze_train ↔ bronze_stores**: store_nbr
- **bronze_test ↔ silver_holidays**: date, store_nbr
- **bronze_test ↔ silver_oil**: date
- **bronze_test ↔ bronze_stores**: store_nbr
- **gold_insights ↔ silver_train**: store_nbr, year, month, family

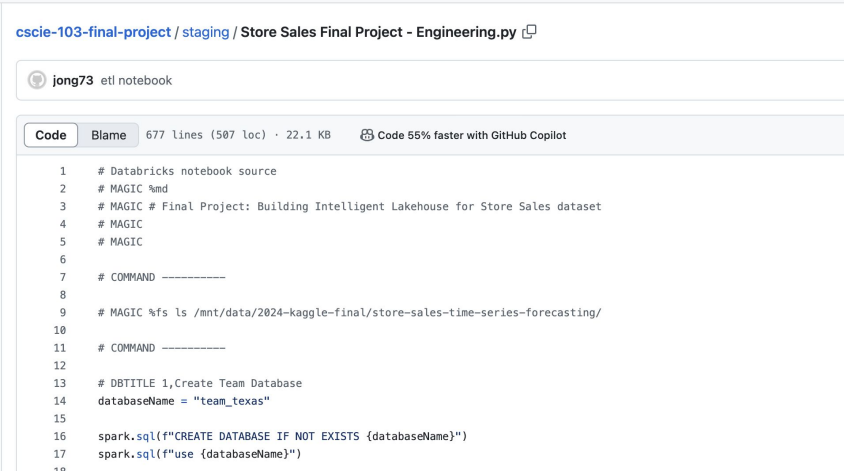
Data Visualization on dashboard

Data Architecture - Data Flow Diagram



Data Architecture - CI/CD

- **Databricks Asset Bundles** to deploy resources across environments and regions
- All jobs, pipelines, notebooks and resource definitions will be committed to **Github** for version control
- Separate environments defined for testing and development



Data Architecture - Disaster Recovery

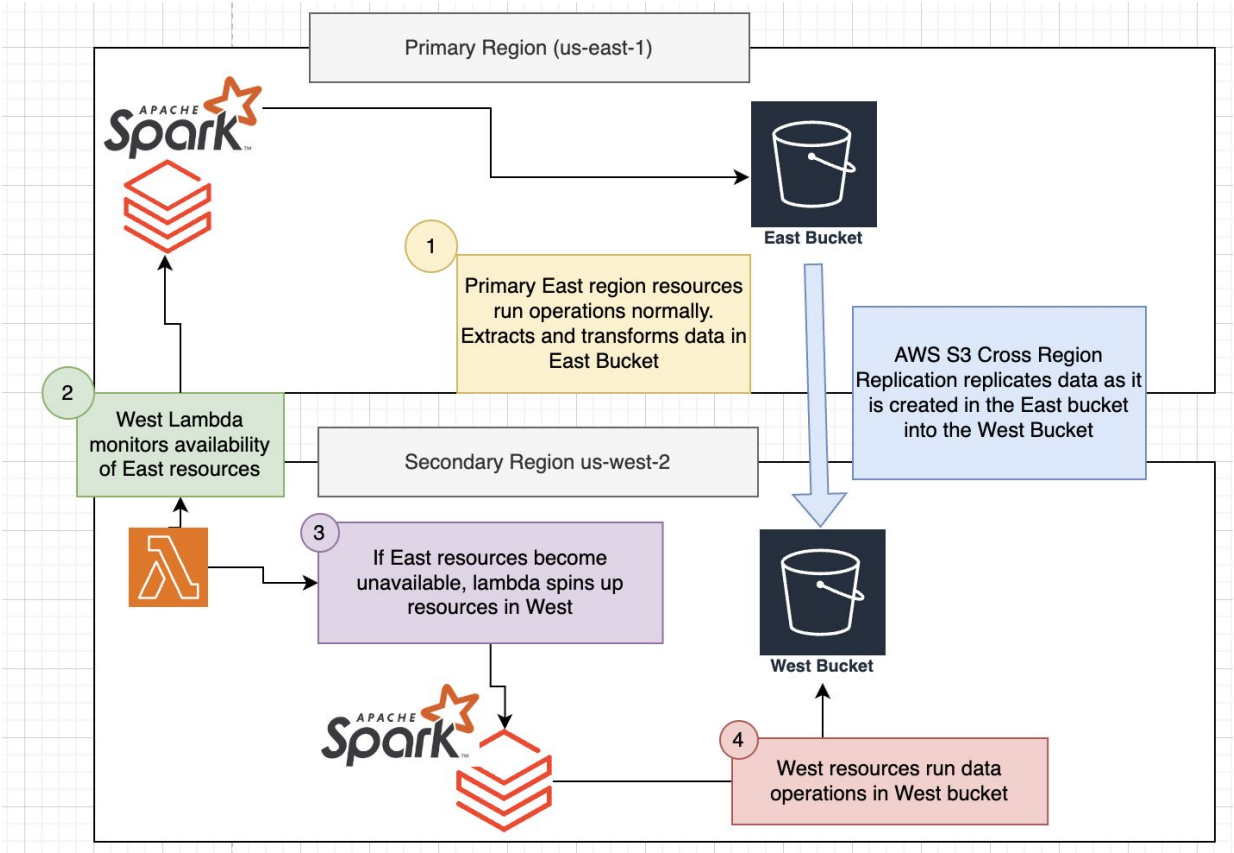
Recovery Point Objective (RPO)

- Utilize S3 Cross Region Replication to minimize Recovery Point Objective (RPO)
- **S3 Cross Region Replication** can minimize the RPO to 15 minutes

Recovery Time Objective (RTO)

- Commit code to Github for version control
- Utilize **Databricks Asset Bundles** to deploy resources anywhere
- Set up **AWS Lambda** function to monitor availability of Primary Region resources. If resources become unavailable, deploy resources to Secondary Region

Data Architecture - Disaster Recovery Diagram



Data Architecture - Real Time Streaming

- Set up a **Kafka** consumer in Databricks to stream incoming JSON messages containing store sales data.
- Use **Spark Structured Streaming** to process the data in real time, applying transformations like parsing and flattening JSON fields. Store this data in Silver Zone.
- Store the raw streaming data in Delta Lake's Bronze zone for future reference.
- Output the aggregated results to the Gold zone for advanced analytics or dashboarding.

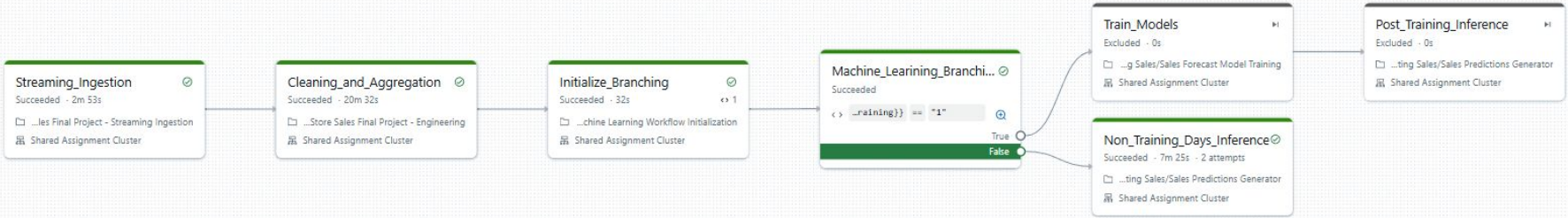
Streaming Upsert to Bronze & Silver Data: Getting Usable Data

Stream Bronze	CSV data streamed to bronze tables using custom upsert functions. New data inserted, existing data updated. No duplicated data.

Oil Prices **Goal**: Join to test and train without resulting in any missing oil price values.	**Problem:** **43** missing oil prices in bronze_oil Cartesian join to distinct dates in union of train and test results in **528** missing obs (oil does not have weekend data). **Solution:** Using LOCF logic to interpolate missing data (LOCB for first value).

Holidays Events **Goal**: Join to test/train by date and store_nbr.	**Problem(s):** Transferred holidays (not holidays) mingled with true holidays “Work days” shown as holidays Local, Regional, and National Holidays all in the same table. **Solution:** Find all transferred holidays corresponding with true holidays. Use true holidays’ date Filter out work days Split data by type. Local joined to stores data by ‘city’, regional by ‘state’ and national is applied to all stores

Data Engineering - Workflow



Daily Ingestion and Aggregation

New **csv** files received once a day

readStream from csv folder for each table

writeStream Triggered Once Upsert *w. microbatch* function

Cleaned Data to **Silver Tables**

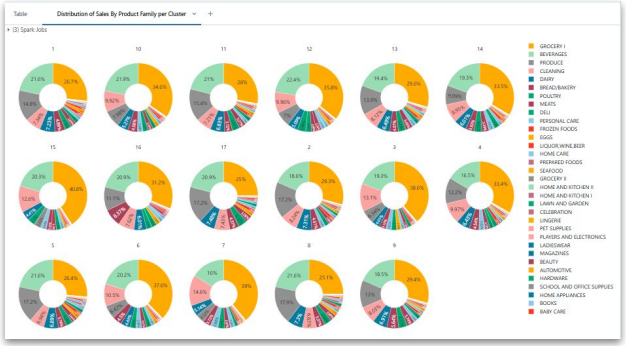
Gold Insights, Gold Inference

ML Orchestration

- **Job Parameters** used for **Conditional Branching** *(comma separated list of days)*
- **Daily Inference** on new data
- **Model Retraining** Every 1st and 16th of the month

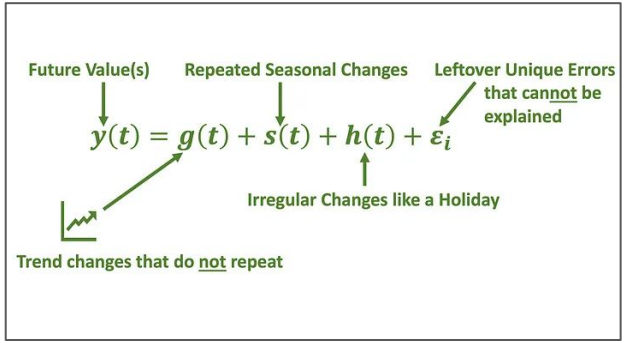
Data Science / EDA, Model Selection and Feature Engineering

- 34 unique family of products and 54 stores. Resulting in 1836 models.
- Selected 17 store clusters and highest selling family products for modeling. **Total #models 17x6 =102.**
- **Prophet Forecasting Model**: Automatic seasonality detection, robust to missing data, customizable holiday effects, scalable for large datasets, intuitive and provides interpretable results with proven accuracy in handling noisy time series.
- Added temporal features like days & months similarity based on sine and cosine.



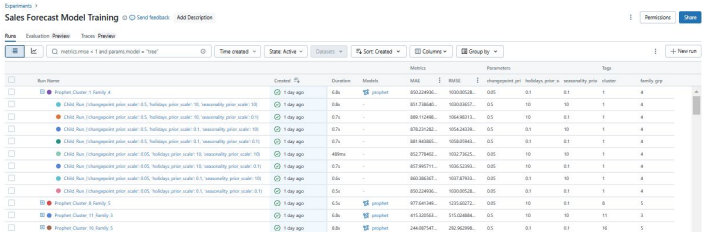
Family products per cluster **(Before)**

Family products per cluster **(After)**



Facebook Prophet Forecasting Model

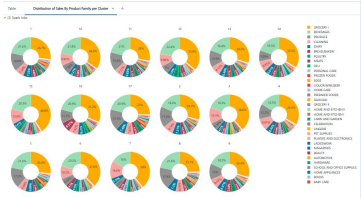
Data Science / Modeling Flow



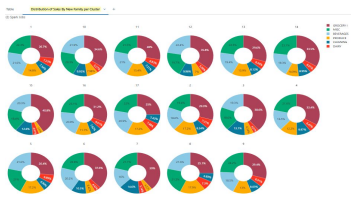
Sales Forecast Model Training

Silver Train

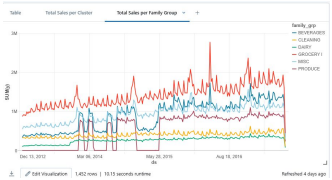
EDA:
Check the trend per store, cluster and family to identify patterns



Feature Engineering:
Recategorize Family
Add Temporal features
Apply String Indexer,



Pre-Processing
Apply aggregation on Cluster and New Family grouping.



Model Training with MLFlow Tracking Using Train and Validation split

Training: 2015-06-01
Validation: 2017-07-01
Utilize Pandas UDF
Hyperparameter tuning
Log metrics & parameters for parent & child models
Choose best Model based on RMSE and MAE

Register best model
One for each cluster and family group combination.
Total 102 models

Log parameter values and metrics from the best Model

Archive older versions before registering any new version

Run Inference on Entire Training data.
Utilizing Pandas UDF

Save Training Inference in gold_inference.
Use Overwrite mode
Run every 15 days

Sales Predictions Generator

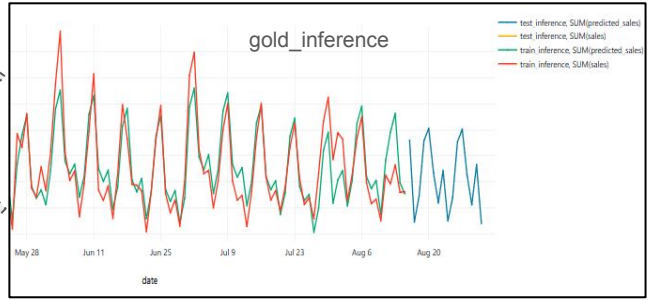
Utilize Pandas UDF
Load registered models using Model names & uri

Silver Test

Apply same feature engineering and preprocessing as train data

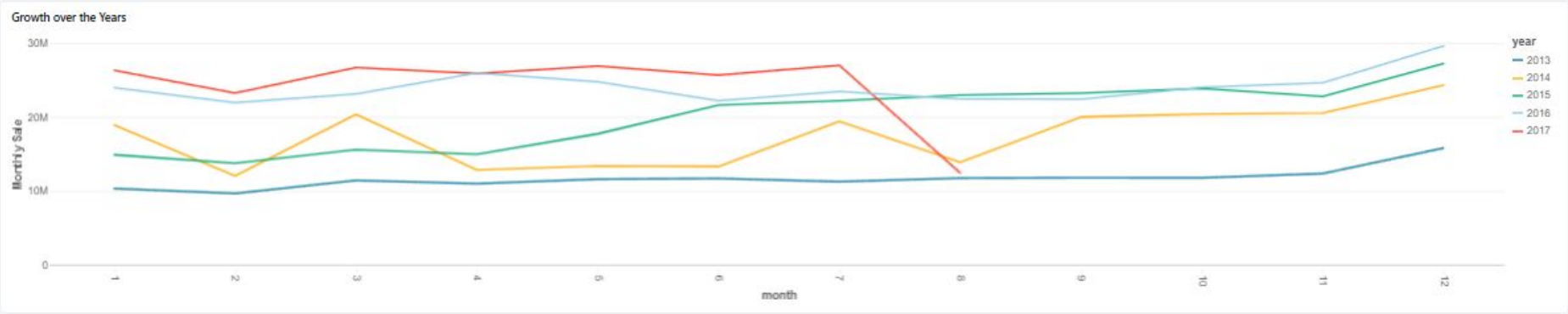
Run Inference on Entire Test data.

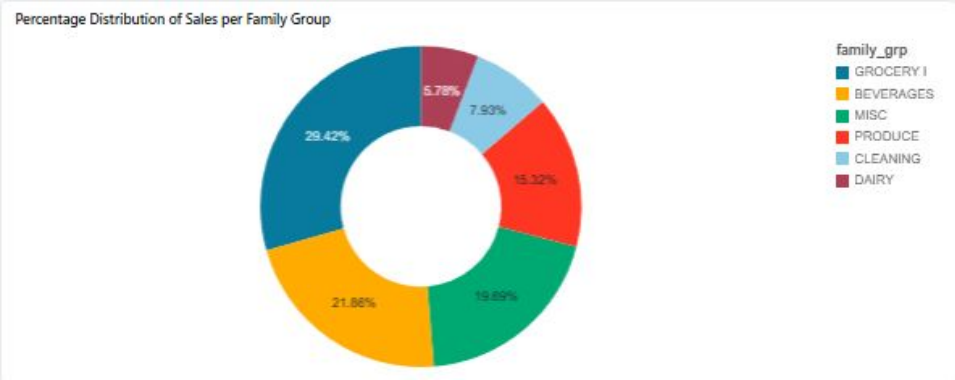
Save Testing Inference in gold_inference.
Use append mode
Run every days



Dashboard - Walkthrough

Please click on this link to [Dashboard]





Thank You from Team Texas!

Appendix

Please click on this link to Dashboard

Sample snapshot from Experimentation for logged models with metics and hyperparameters

Experiments >

Sales Forecast Model Train

Best logged model per cluster & family based on lowest RMSE

Permissions | Share

Runs | Evaluation Preview | Traces Pre

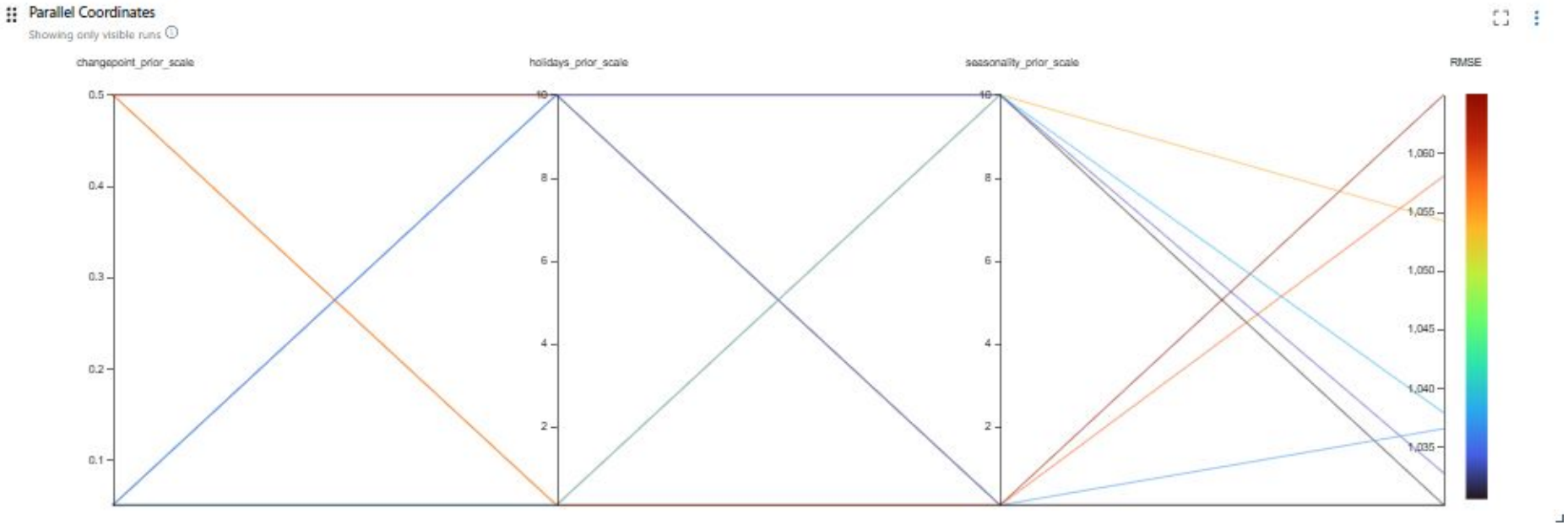
metrics.rmse < 1 and | Time created | State: Active | Datasets | Sort: Created | Columns | Group by | New run

				Metrics		Parameters			Tags	
Run Name	Created	Duration	Models	MAE	RMSE	changepoint_pri	holidays_prior_s	seasonality_prio	cluster	family_grp
Prophet_Cluster_1_Family_4	1 day ago	6.8s	prophet	850.224936...	1030.00528...	0.05	0.1	0.1	1	4
Child_Run_{'changepoint_prior_scale': 0.5, 'holidays_prior_scale': 10, 'seasonality_prior_scale': 10}	1 day ago	0.8s	-	851.738640...	1030.03657...	0.5	10	10	1	4
Child_Run_{'changepoint_prior_scale': 0.5, 'holidays_prior_scale': 10, 'seasonality_prior_scale': 0.1}	1 day ago	0.7s	-	889.112498...	1064.98313...	0.5	10	0.1	1	4
Child_Run_{'changepoint_prior_scale': 0.5, 'holidays_prior_scale': 0.1, 'seasonality_prior_scale': 10}	1 day ago	0.7s	-	878.231282...	1054.24339...	0.5	0.1	10	1	4
Child_Run_{'changepoint_prior_scale': 0.5, 'holidays_prior_scale': 0.1, 'seasonality_prior_scale': 0.1}	1 day ago	0.7s	-	881.943865...	1058.05943...	0.5	0.1	0.1	1	4
Child_Run_{'changepoint_prior_scale': 0.05, 'holidays_prior_scale': 10, 'seasonality_prior_scale': 10}	1 day ago	489ms	-	852.778402...	1032.73625...	0.05	10	10	1	4
Child_Run_{'changepoint_prior_scale': 0.05, 'holidays_prior_scale': 10, 'seasonality_prior_scale': 0.1}	1 day ago	0.7s	-	857.995711...	1036.52393...	0.05	10	0.1	1	4
Child_Run_{'changepoint_prior_scale': 0.05, 'holidays_prior_scale': 0.1, 'seasonality_prior_scale': 10}	1 day ago	0.6s	-	860.386367...	1037.87933...	0.05	0.1	10	1	4
Child_Run_{'changepoint_prior_scale': 0.05, 'holidays_prior_scale': 0.1, 'seasonality_prior_scale': 0.1}	1 day ago	0.5s	-	850.224936...	1030.00528...	0.05	0.1	0.1	1	4
Prophet_Cluster_8_Family_5	1 day ago	6.5s	prophet	977.641349...	12[illegible]				8	5
Prophet_Cluster_11_Family_3	1 day ago	6.8s	prophet	415.320563...	515.024884...				11	3
Prophet_Cluster_16_Family_5	1 day ago	8.8s	prophet	244.087547...	292.962998...				16	5

Child models per cluster & family for hyperparameter tuning

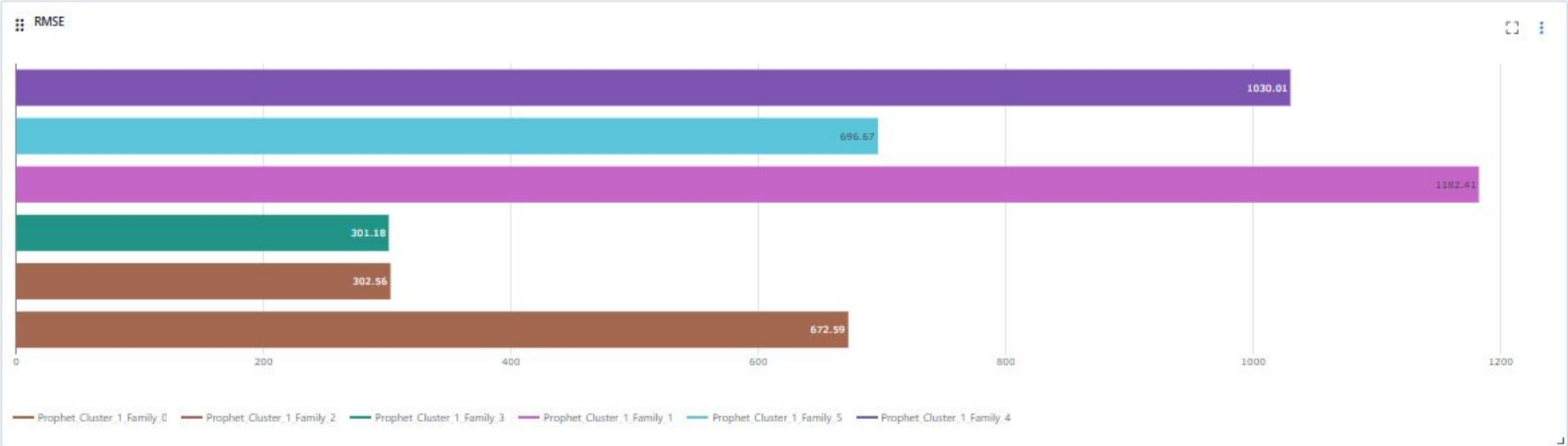
Sample snapshot from Experimentation

Sample snapshot from parallel coordinates from hyper parameter tuning of a model for one cluster and family combination



Sample snapshot from Experimentation

- Sample snapshot of error (RMSE) comparison of models of different families in the same cluster.
- This shows which family models need further improvement within a cluster.



Sample snapshot from Registered Model

Sample snapshot of deployed model for cluster 1 - family 0 from registered model

Registered Models >

prophet_best_model_cluster_1_family_0

Details | Legacy serving

Notify me about: All new activity

Created Time: 2024-12-15 22:38:57 | Last Modified: 2024-12-15 22:39:07 | Creator: naskani_imran@yahoo.com

Description Edit

Prophet Model for cluster 1 and family 0

Tags

Name	Value	Actions
model_type	Prophet Model for cluster 1 and family 0	
owner	Imran Naskani - Team Texas	

Name | Value | Add

Versions All | Active 1 | Compare

Version	Registered at	Created by	Stage	Pending Requests	Description
Version 1	2024-12-15 22:38:58	naskani_imran@yahoo.com	Production	-	Prophet Model for cluster 1 a...