FAVORITA STORE SALES PREDICTION

OVERVIEW

About the Business

Favorita is one of the **largest supermarket chains** in Ecuador, known for its extensive selection of **groceries, household items,** and other goods. It has **numerous locations** across the country.

Purpose & Impacts of the Project

- Given past sales data of different categories and stores, **predict future daily sales** of each category in different stores.
- Improve **customer experience**
- Control **business cost**

DATASET AND PREPROCESSING

Data source: Favorita Company

5 Tables

3M rows, 12 Columns

1. Sales
2. Stores
3. Oil
4. Holidays & Events
5. Store Total Sales

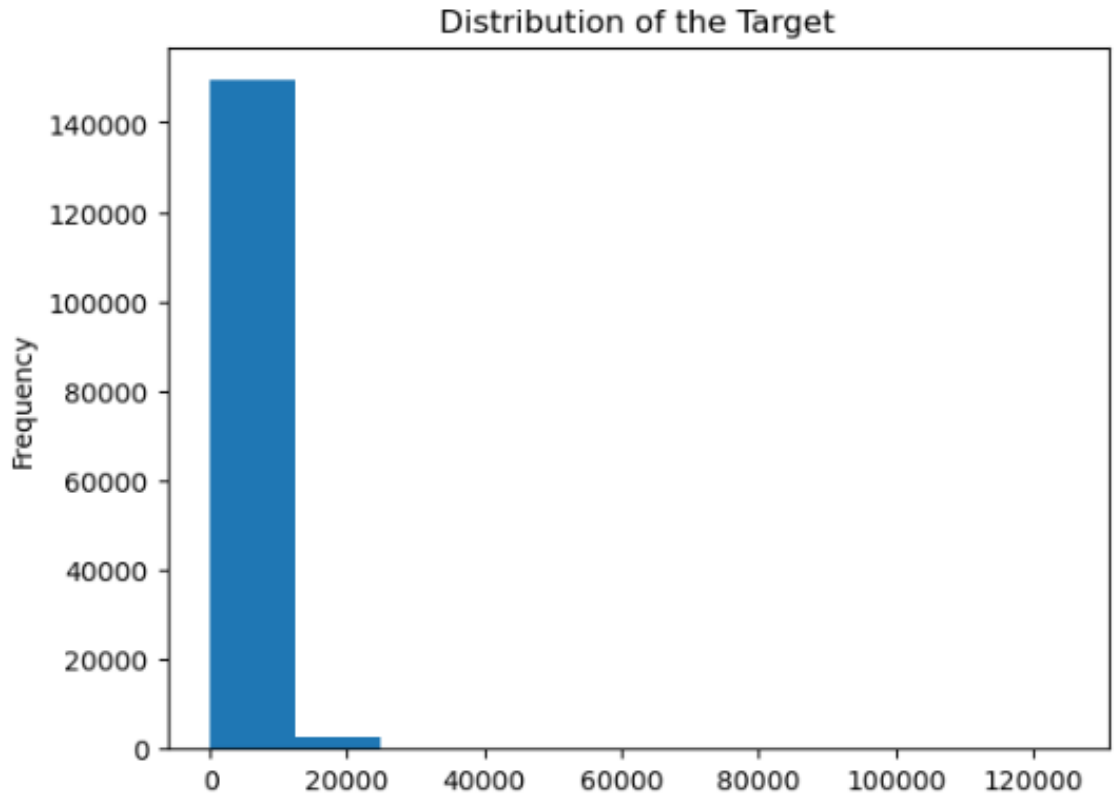
- **Each table's columns**
- **The relationships** between each table
- Duplicates, date format, 43 null values

- **Define project scope**
- **Xs' relationship** with the **target**
- Filled 43 null values
- Joined the five tables

- **Duplicated columns**
- Visualize and fill **nulls**
- Distributions
- Translation
- Dummy variables
- Time series column **feature engineering**

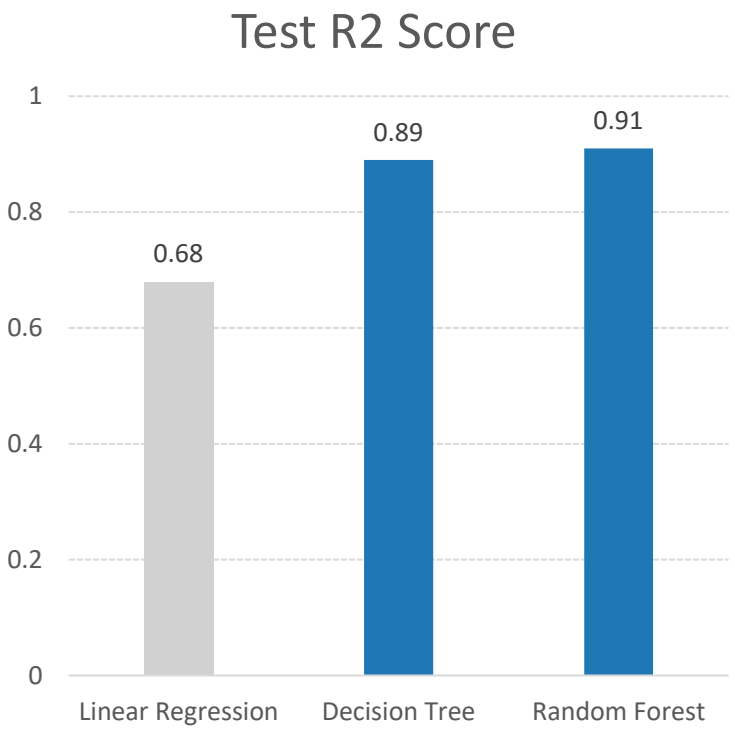
TARGET DISTRIBUTION

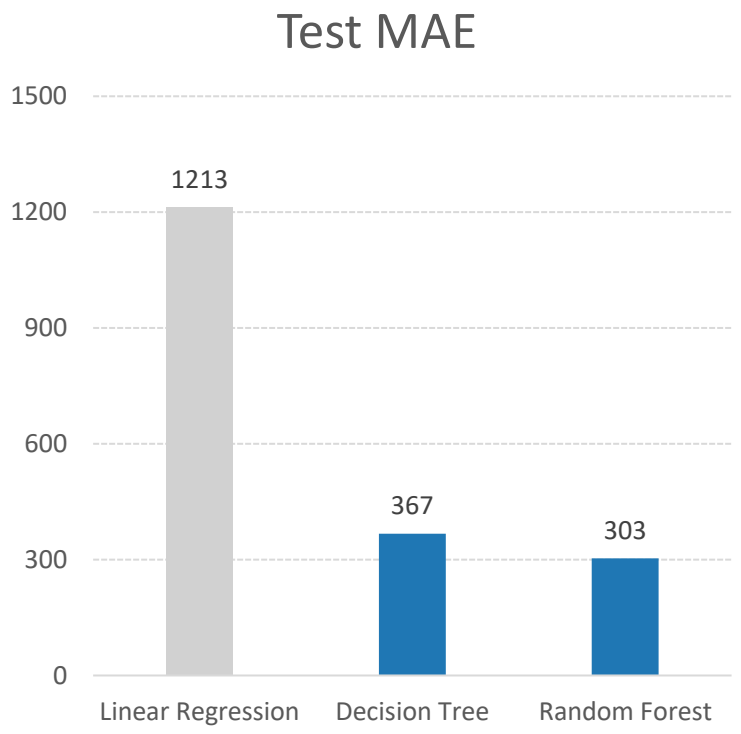
An earthquake caused **GROCERY I's** sales units to spike to a record high, resulting in a right-skewed dataset. Sales units over 40k represent only **0.0043%** of the data, making it challenging for models to capture these abnormally large sales units.

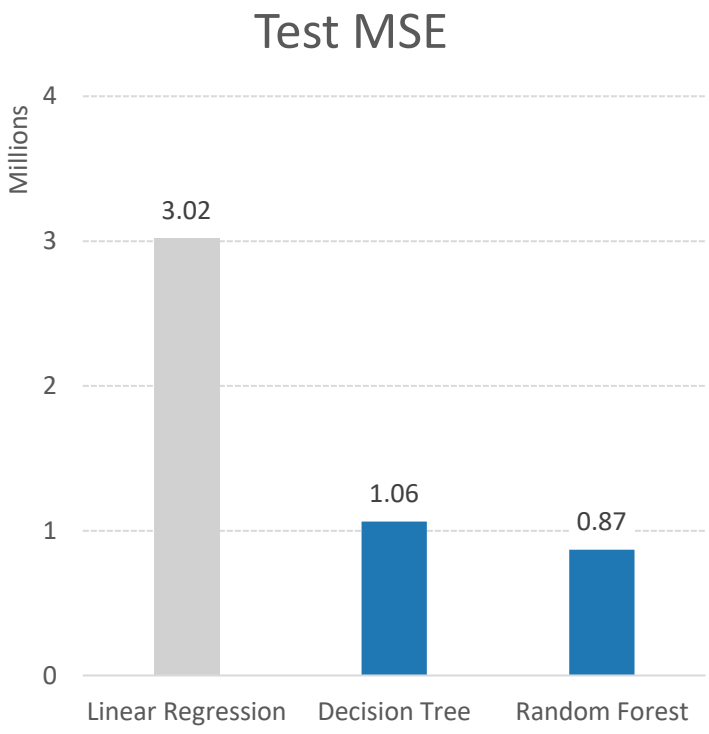


mean	2,872
std	3,062
min	-
25%	842
50%	1,895
75%	3,843
max	124,717

MODEL PERFORMANCE : Compared to LR, DT and RF performed much better across all three scores; however, the extremely large sales units contribute to high MAE and MSE in the advanced models.





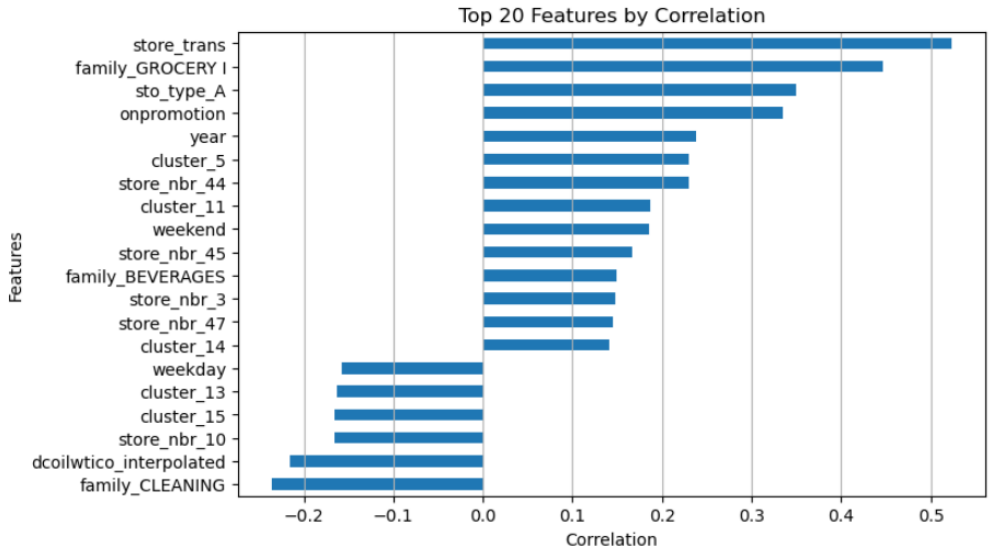


LINEAR REGRESSION EVALUATION

Linearity and Multicollinearity

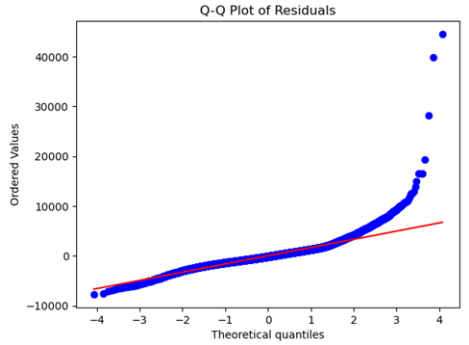
- Out of 116 features, only 24 show high correlations; most features have low correlations with the target.

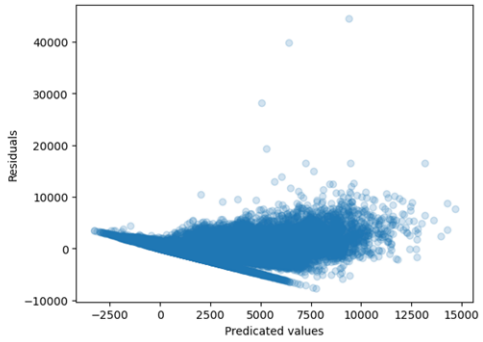
	feature1	feature2	feature_corr
0	sto_type_A	store_trans	0.73
1	cluster_11	store_nbr_45	0.68
2	cluster_11	store_nbr_49	0.69
3	cluster_12	store_nbr_17	1.00
4	cluster_15	store_nbr_10	1.00
5	cluster_5	store_nbr_44	1.00
6	cluster_6	store_nbr_9	0.79
7	cluster_6	sto_type_B	0.77
8	cluster_9	store_nbr_4	1.00
9	description_eng_Foundation of Quito-1	locale_Local	0.71
10	description_eng_foundation of Quito	locale_Local	0.70
11	year	dcoilwtico_interpolated	-0.83
12	weekend	weekday	-0.75



Residuals and Homoscedasticity

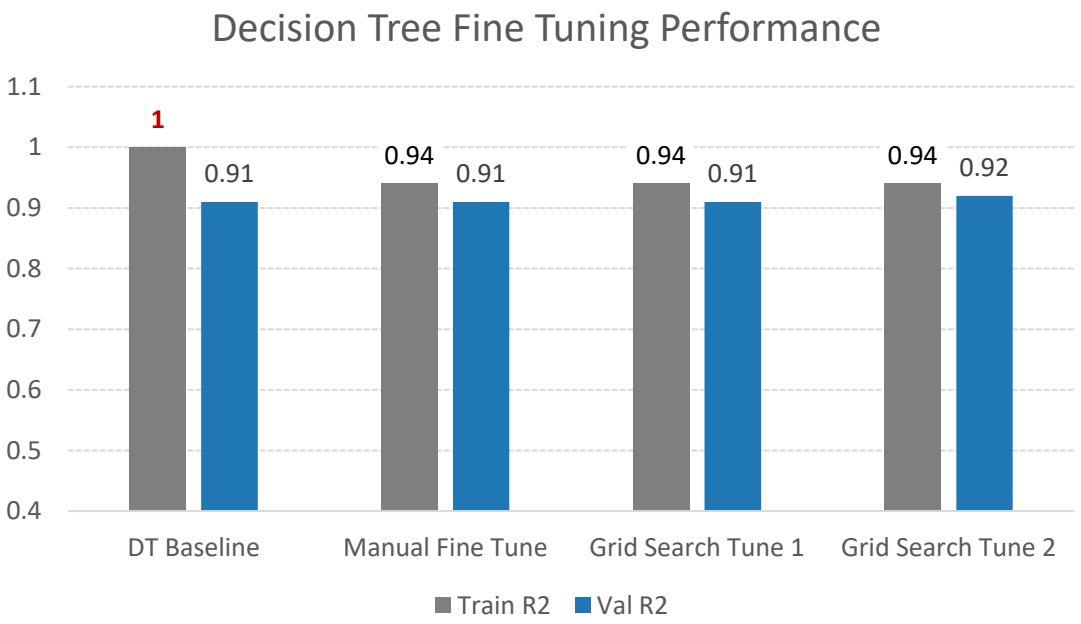
- Large gap at the tail & clear pattern seen between residuals and predicated values
- The model is not reliable

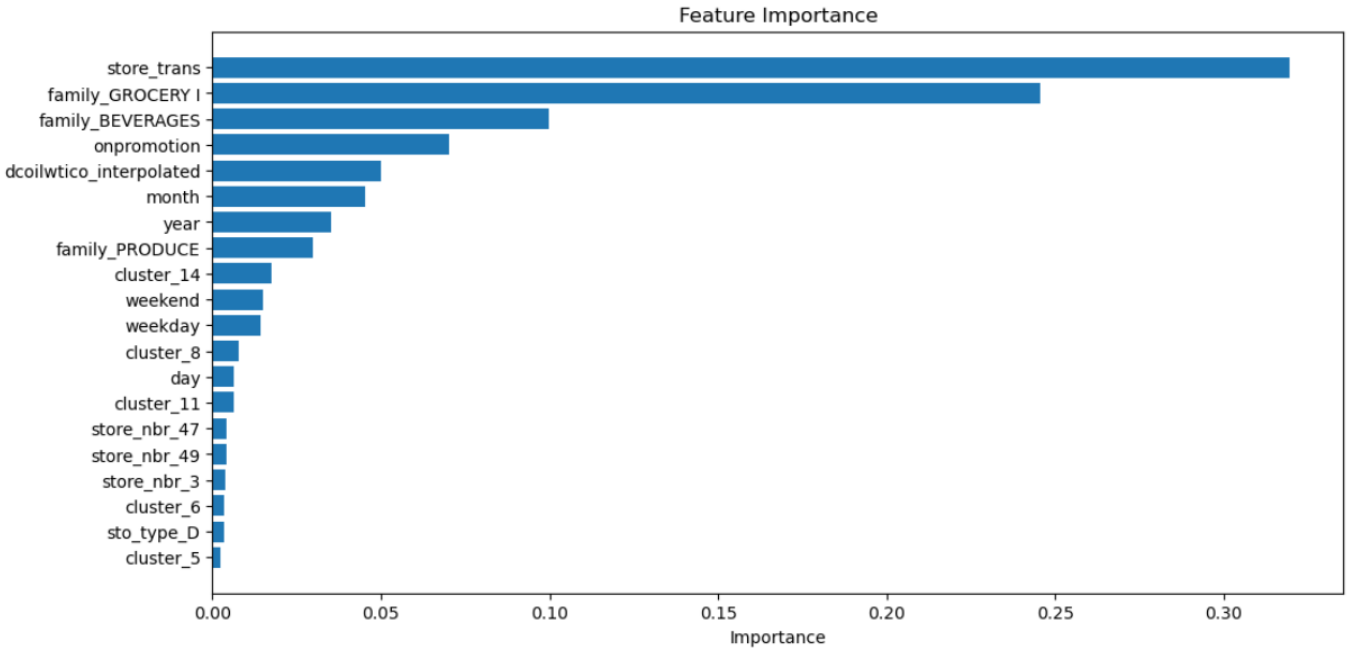




DECISION TREE MODELING:

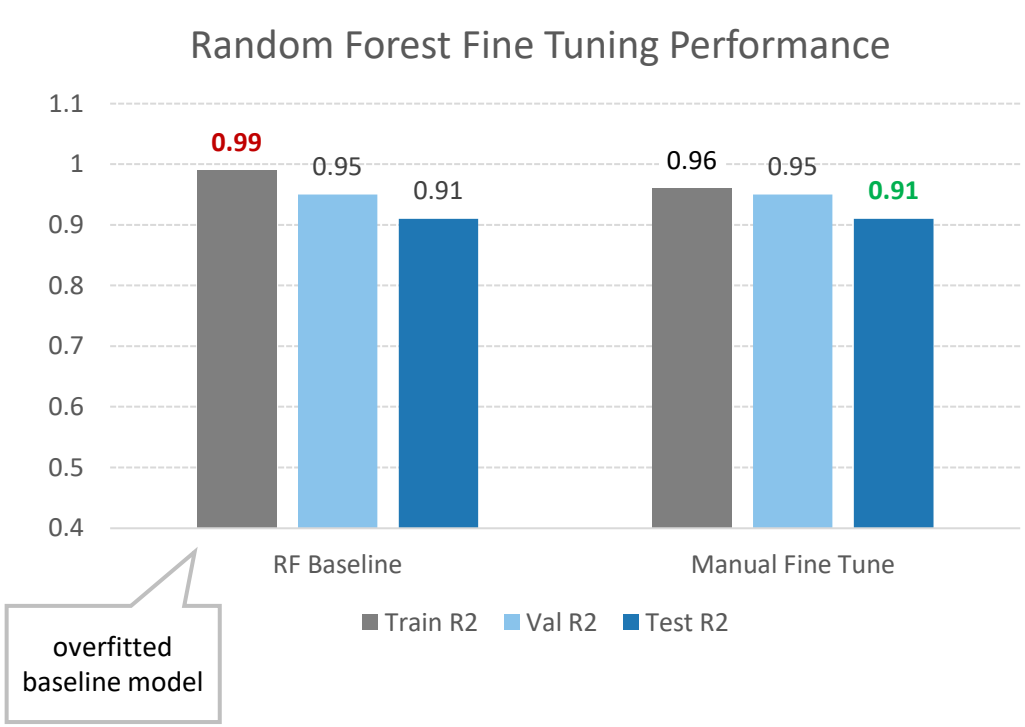
- The baseline model is perfectly fitting. To reduce overfitting, three rounds of fine-tuning were applied.
- The most important features are "store_trans", "GROCERY I", "BEVERAGES", "onpromotion", “dcoilwtico”, " month“, “year”.

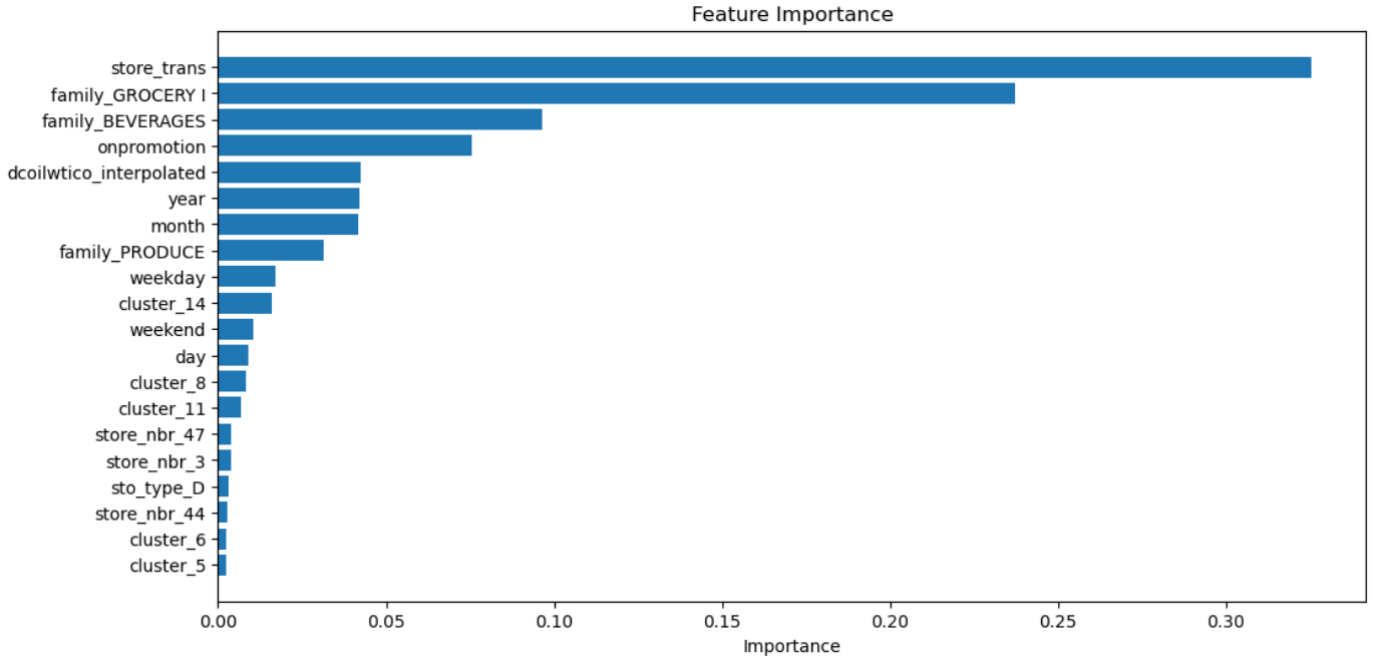




RANDOM FOREST MODELING:

- RF outperformed all other models with the least effort in fine-tuning.
- LR's top features are store_trans, Grocery I, **store type A**, onpromotion, year, **cluster5**, **Cleaning**, and daily oil price.
- DT and RF list store_trans, Grocery I, **Beverages**, onpromotion, daily oil price, year, **month**, and **Produce** as top features.






THANK YOU