**Store Sales: Time Series Forecasting**

**Project Description:**

Our project is inspired by a Kaggle competition that focuses on predicting daily sales for products sold at Favorita grocery stores in Ecuador. Accurate sales forecasting is crucial for optimizing inventory management, reducing waste, and improving customer satisfaction. In this project, we analyze historical sales data to identify trends, integrate external factors such as holidays and oil prices, and develop predictive models for sales forecasting. Using insights from this dataset, we aim to address real-world challenges in retail demand forecasting.

**Data Overview:** The datasets include:

* Store Metadata: Store locations, types, and clusters.
* Sales Data: Historical sales records by product family and items on promotion.
* External Data: Oil prices, local and national holidays, and major events.

**Key Performance Indicators:** The competition’s evaluation metric is Root Mean Squared Logarithmic Error (RMSLE), which balances outliers’ impact and penalizes underestimations more heavily.

**Project Goals:**

* Achieve RMSLE < 0.5.
* Maintain training efficiency (<2 hours runtime).
* Improve RMSLE >10% through feature engineering.

**Model Selection and Evaluation:** We tested multiple models and evaluated their performance:

1. **Mean by Year (RMSLE = 0.902):** Simple averaging of sales on corresponding calendar days across years. Captures basic trends but lacks complexity.
2. **Rolling Average (RMSLE = 0.46141):** Predicts based on the rolling average over the past month. Captures seasonality but overlooks intricate patterns.
3. **Random Forest (RMSLE = 0.51442):** An ensemble decision tree model capturing non-linear relationships. Performed inconsistently on low-sales product families.
4. **Prophet Model (RMSLE = 0.48433):** Designed for seasonal and trend data. Performs reasonably well but struggles with chronic underestimations.
5. **SARIMAX (RMSLE = 0.47600):** Classical time-series model that incorporates seasonality, trends, and covariates. Requires extensive computation for independent store-family combinations.
6. **Hybrid Model:** To address individual model limitations, we developed a combined model using weighted averages of Rolling Average, Prophet, and SARIMAX predictions. This hybrid model achieved the best performance with RMSLE = 0.42913, placing us in the top 100 on the competition leaderboard.

**Insights:**

* **Mean by Year**: Poor performance due to limited complexity but highlights weekend peaks.
* **Rolling Average**: Strong performance but overestimates weekday sales due to weekend peaks.
* **Random Forest**: Captures general trends but struggles with low-sales families.
* **SARIMAX:** Performs well overall but demands significant computational effort.
* **Prophet:** Reasonable results but prone to underestimations for certain product families.
* **Hybrid Model:** Combining Rolling Average, Prophet, and SARIMAX predictions addressed underestimation issues and model-specific weaknesses, delivering robust and accurate forecasts.

**Conclusion:**

Our hybrid approach demonstrates the value of combining models to tackle retail forecasting challenges. The project outcomes highlight practical solutions for inventory optimization and waste reduction in real-world applications.