How the Model Predicts-XGBoost

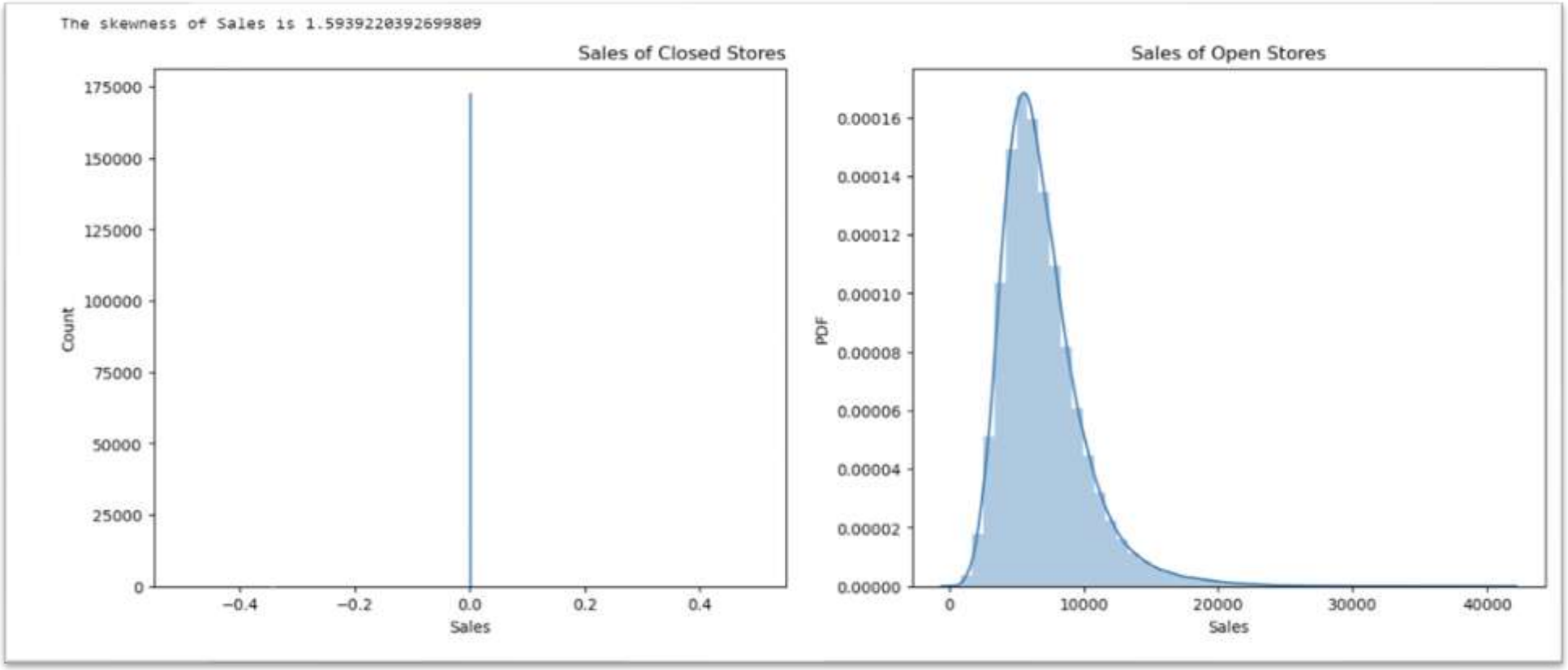
Core XGBoost Objective

$$L^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

- g_i: First-order gradient (direction)
- h_i: Second-order gradient (step size)
- $\Omega(f_t)$: Regularization (prevent overfitting)

Core Idea

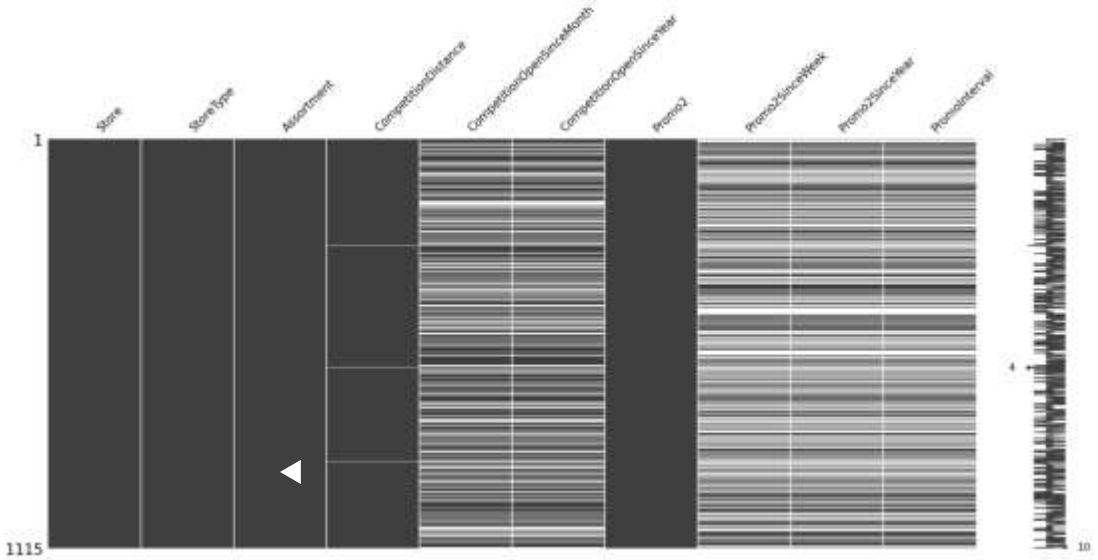
A powerful and scalable machine learning algorithm based on the Gradient Boosting framework.

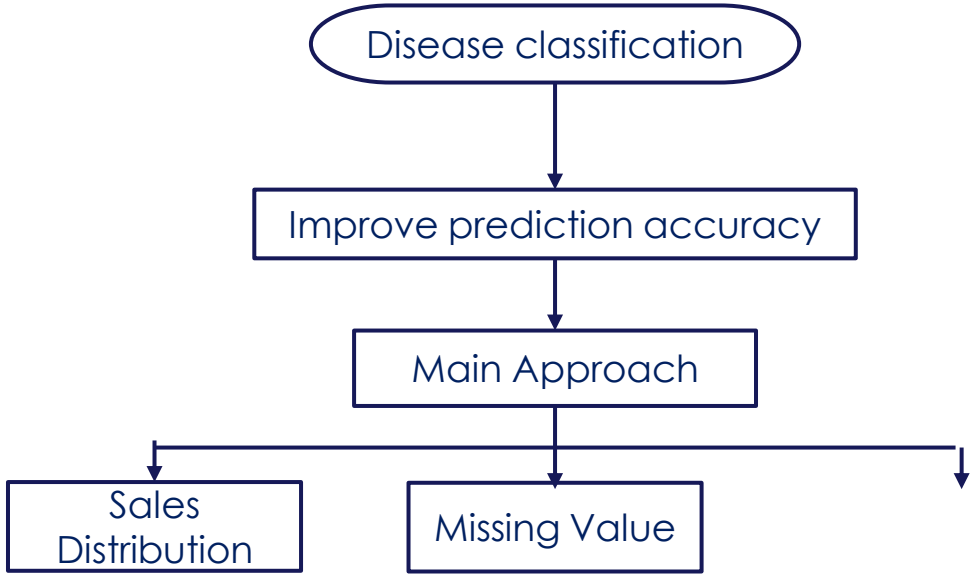


- Sales During Closure
- Modeling Strategy
- Sales Distribution (Open Days)

→ **Sales on open days show a right-skewed distribution.**

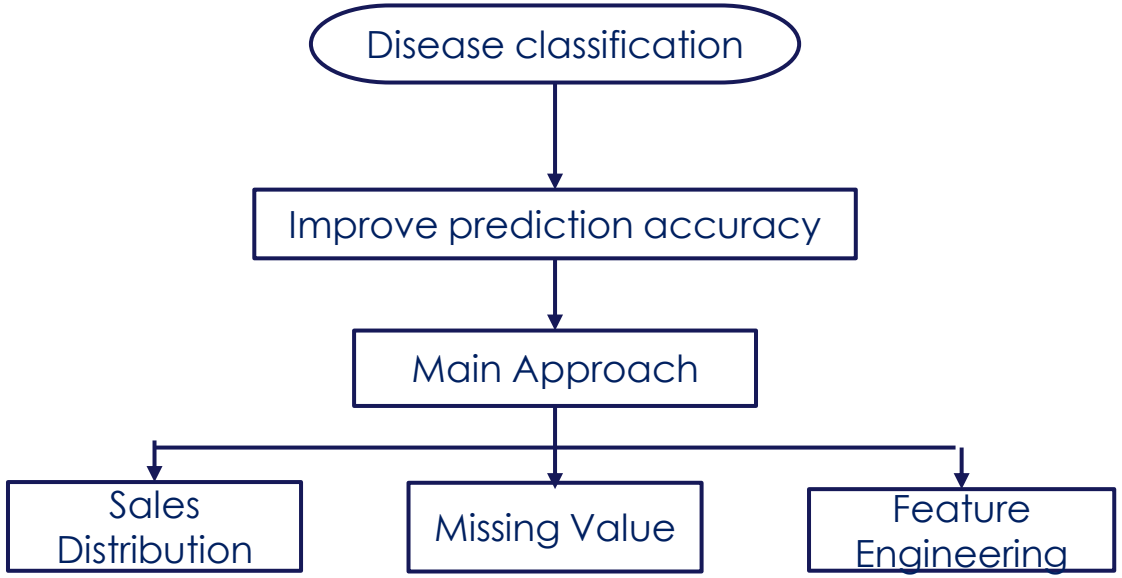
→ **XGBOOST**

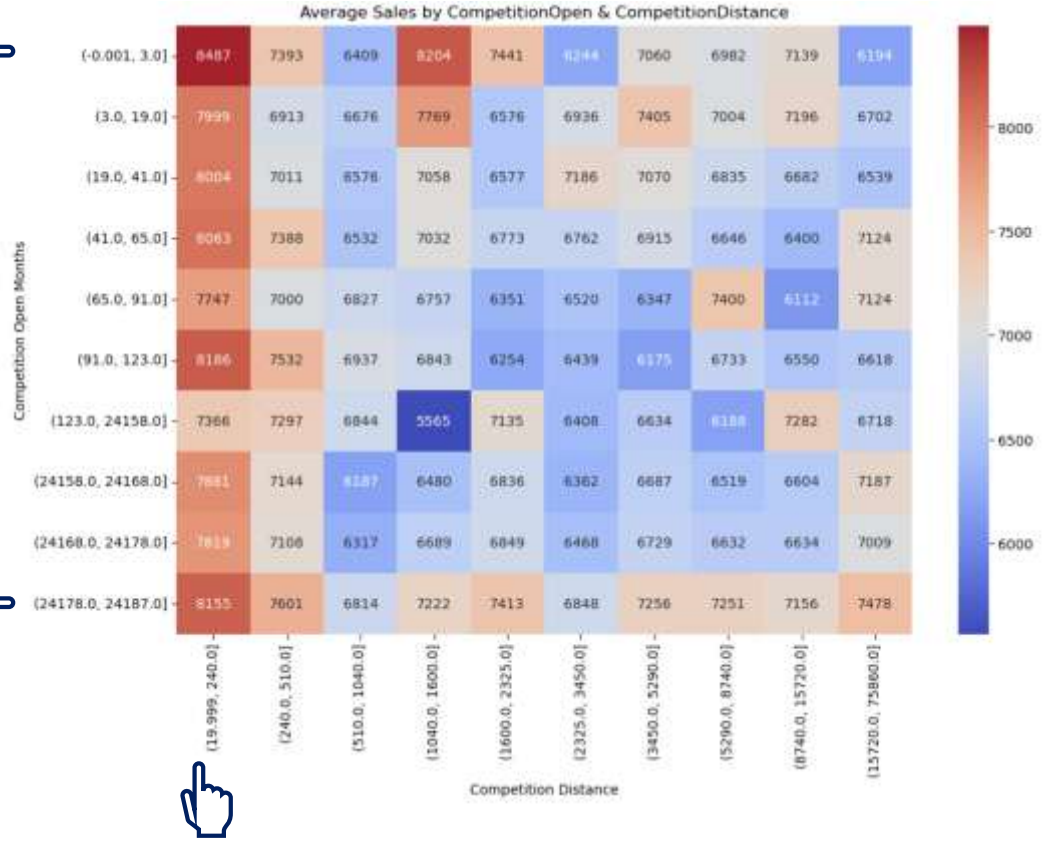




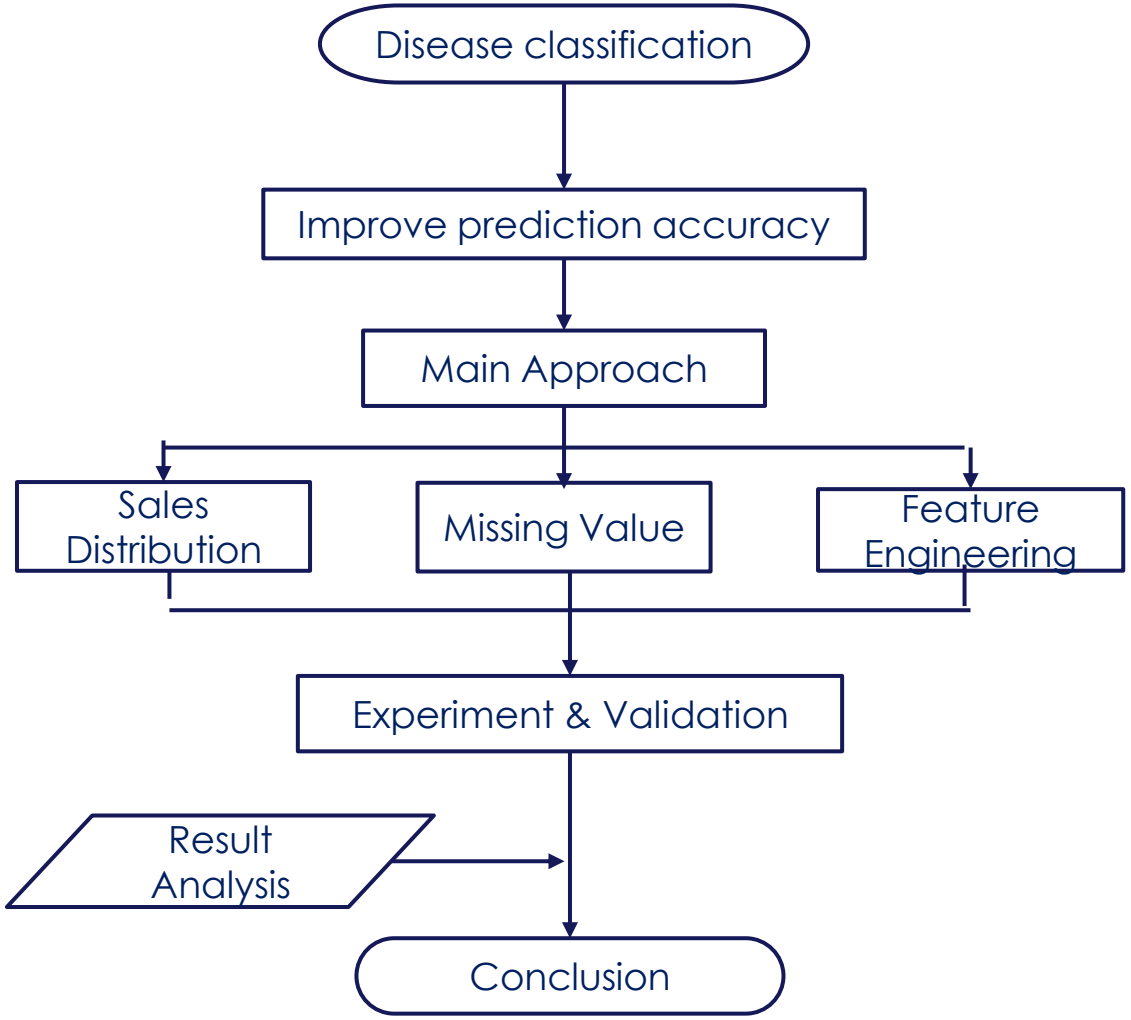
- Test Set: Fill all missing values with 1 for model stability
- CompetitionDistance: Median imputation (robust to right-skew)
- Other Fields: Zero-fill indicates "No Competition" or "No Ongoing Promotion"

1. Vertical Color Bands →
2. Universal High Sales in Specific Weeks →
3. Evidence of a Powerful, Annual Driver (e.g., Christmas) →
4. This Driver is Cyclical & Predictable →
5. Therefore, **WeekOfYear** Encodes This Confirmed Pattern for the Model.

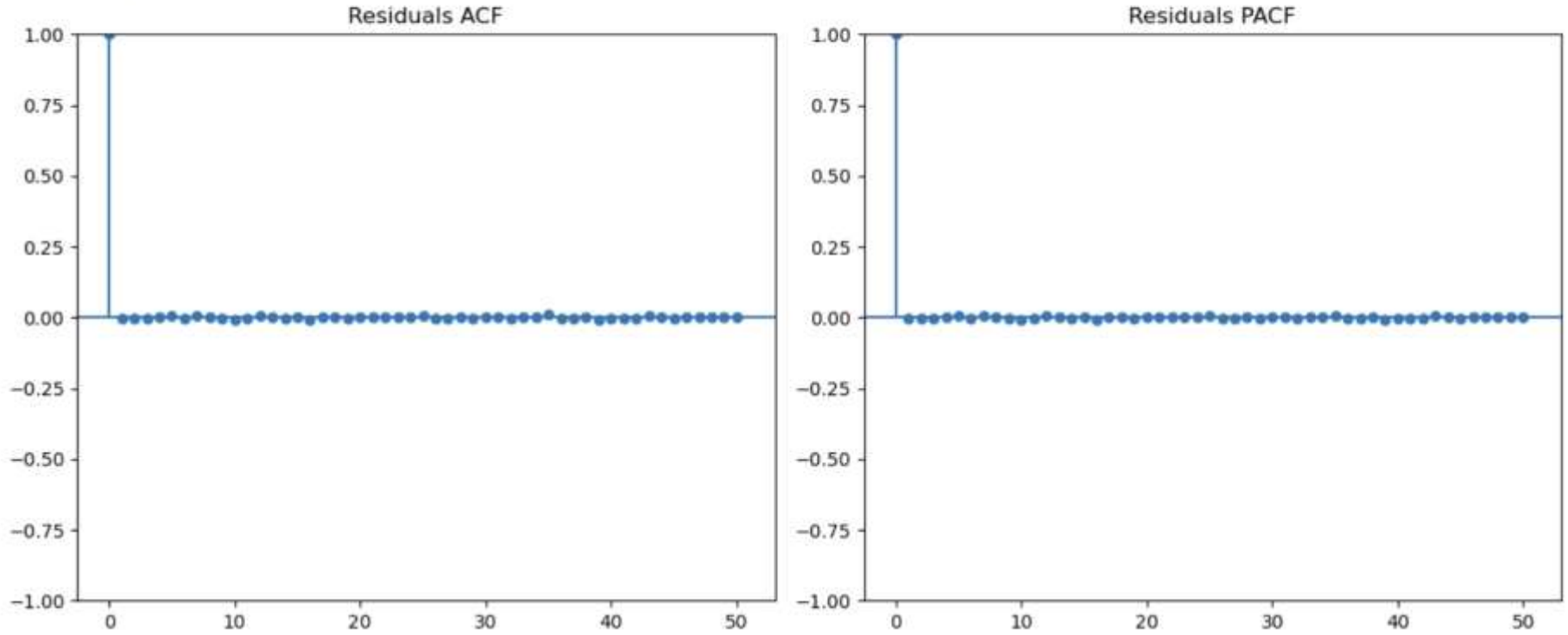




- Key Insight 1: The "U-Shaped" Power of Location
- Key Insight 2: The "New Competitor" Boost

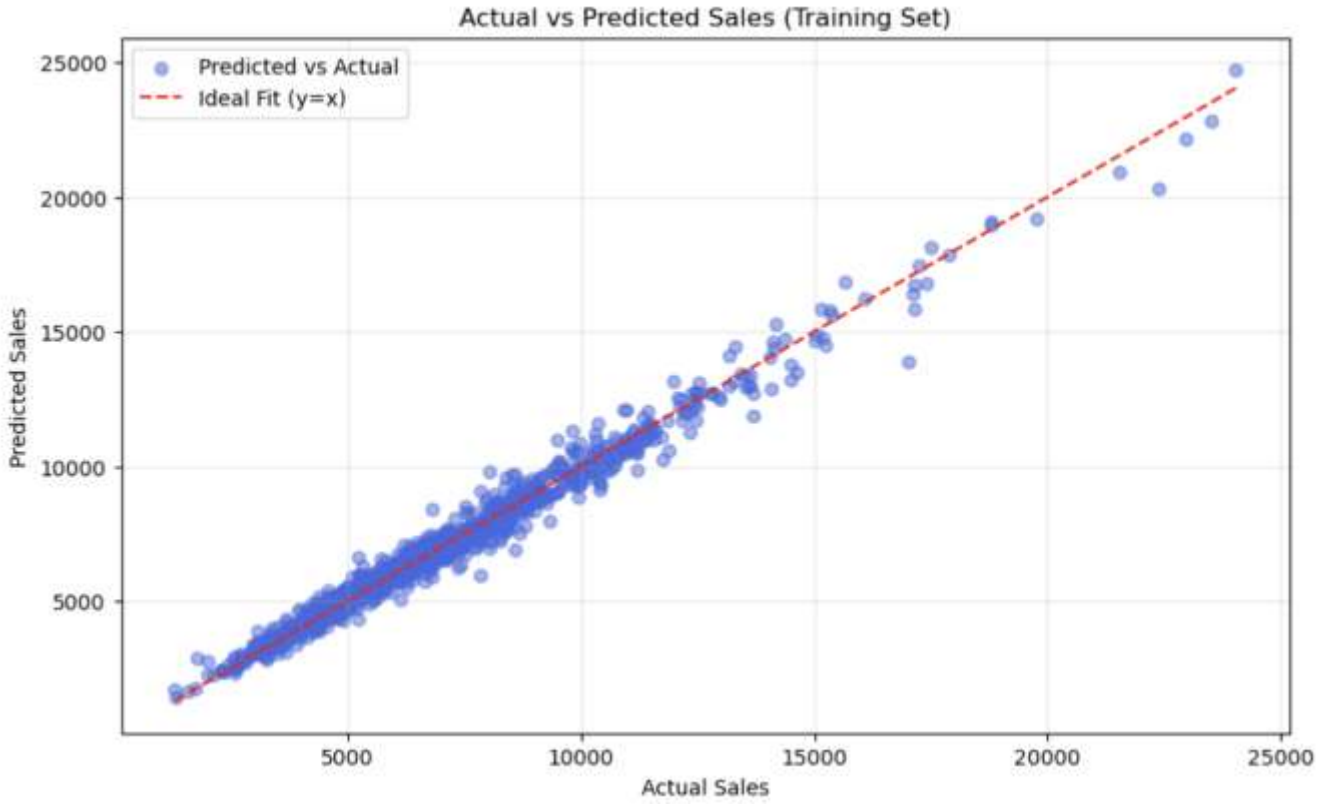


```
MSE: 389813.31
RMSE: 624.35
MAE: 418.42
R²: 0.9595
AIC (approx): 2427789.25
```



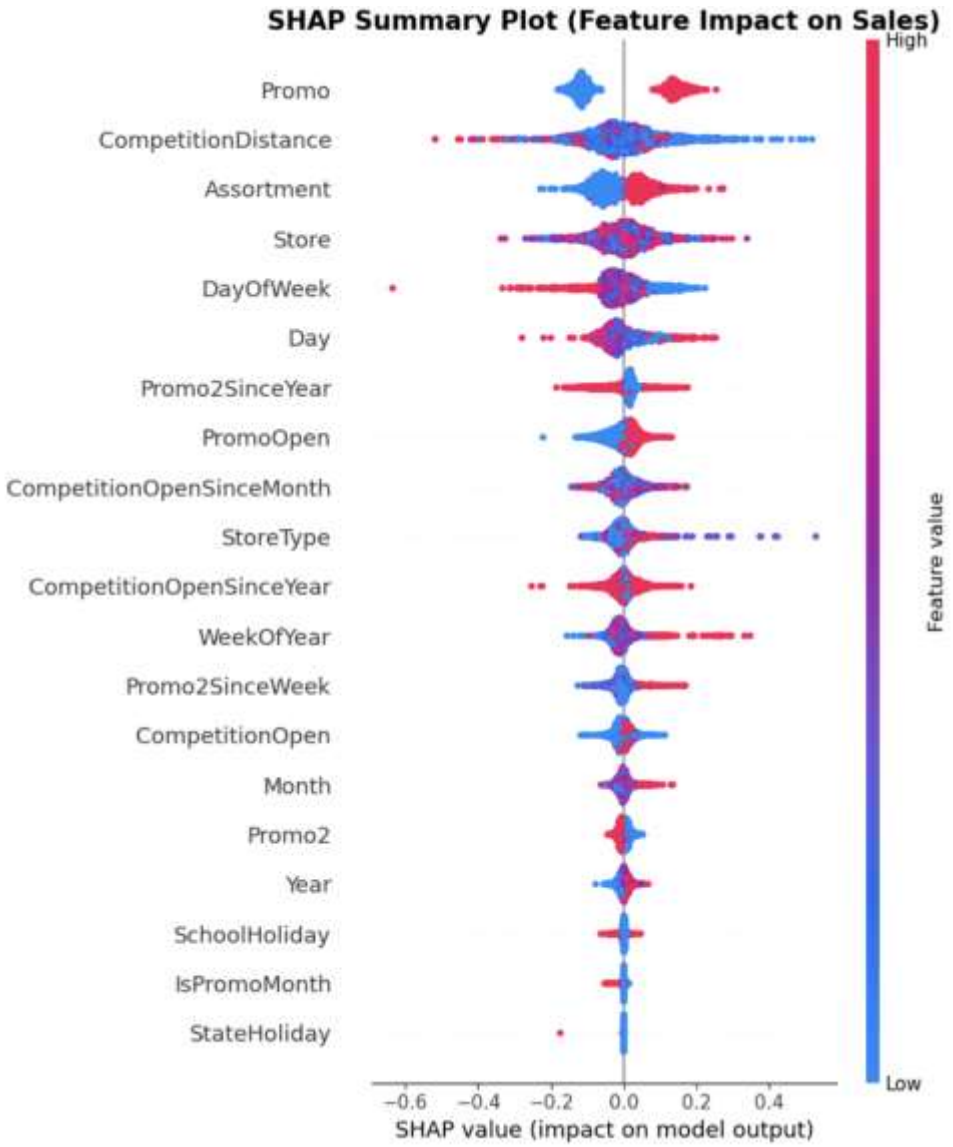
- R-Squared (R^2): 0.9595
- Root Mean Squared Error (RMSE): 624.35
- Mean Absolute Error (MAE): 418.42

→ A robust and highly effective model ready for business application.



- High Accuracy for Routine Operations
- Systematic underestimation of sales during high-demand periods.

→ **A reliable foundation for daily forecasting, with clear next steps for capturing peak demand.**



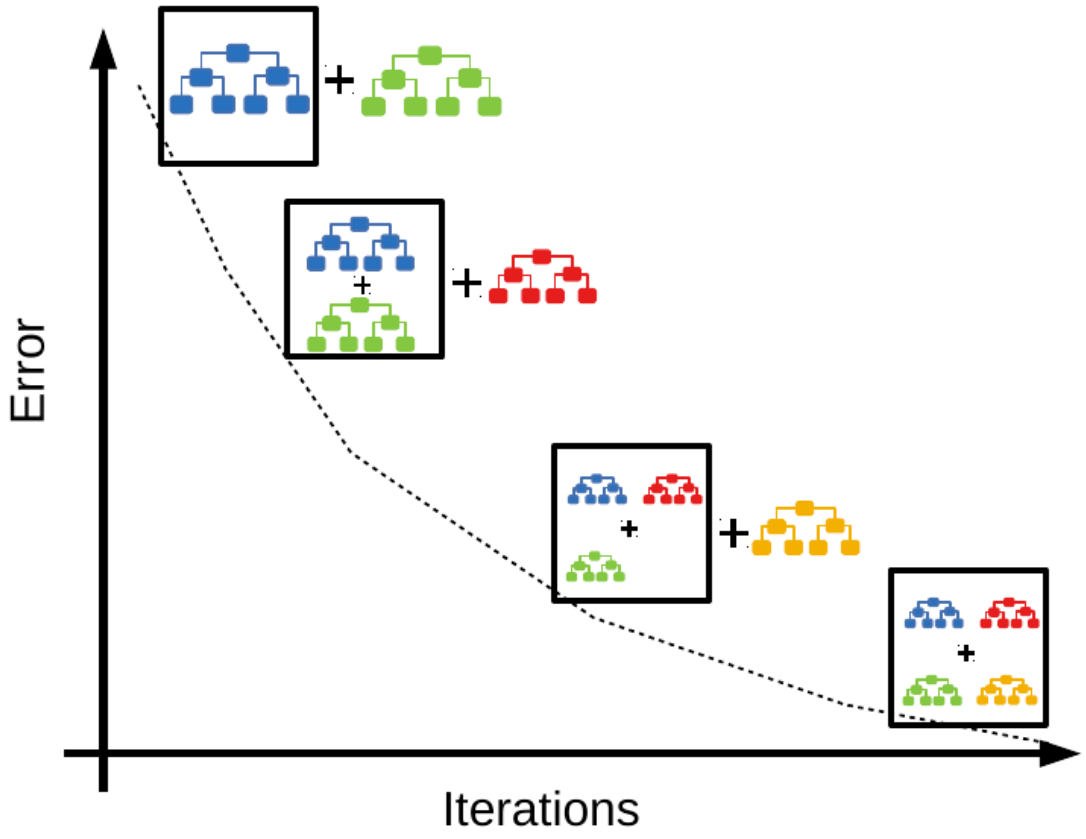
What Drives Sales?

- Top Features: Promo, CompetitionDistance, Assortment, Store, DayOfWeek are the most influential.
- Red = High Value, Blue = Low Value.

Key Insights & Business Implications:

- **Promotions** are Our **Superpower**→ Protect & optimize promotional calendar.
- **Competition is Complex**→ Tailor strategies by competitor proximity.
- Store & Assortment are Key Differentiators→ Localize assortments and share best practices across stores.

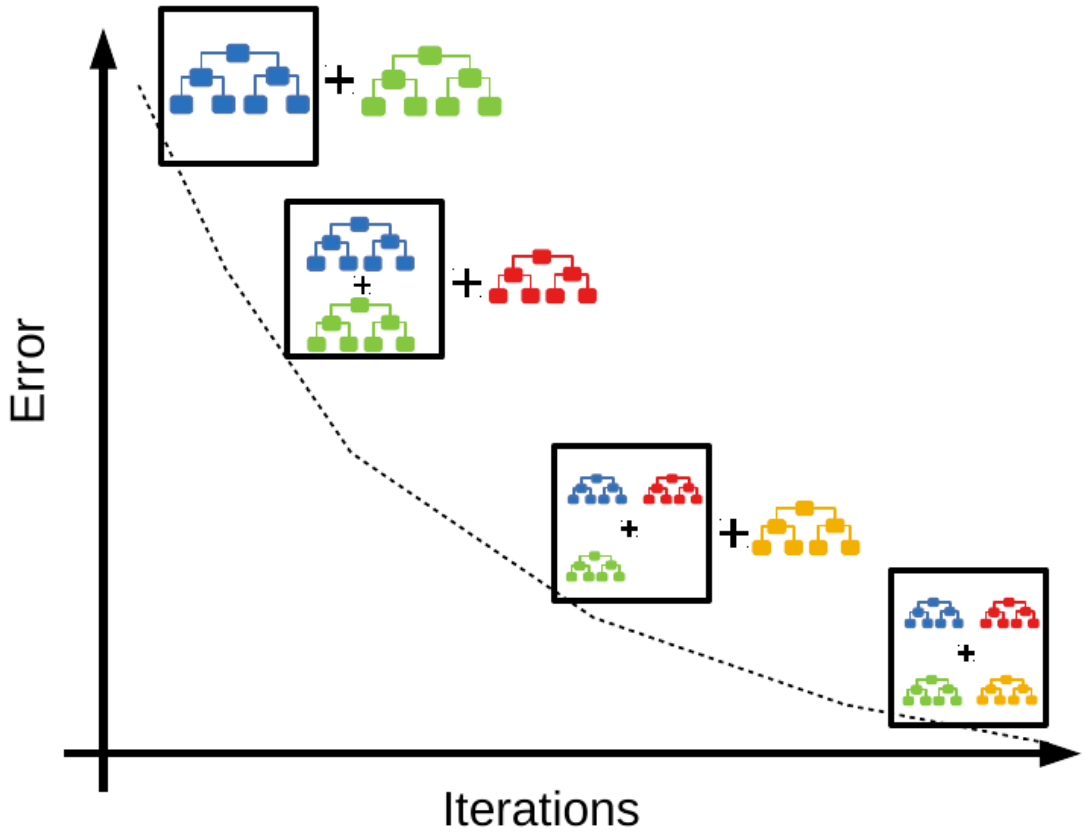
How the Model Predicts-lightGBM



Definition

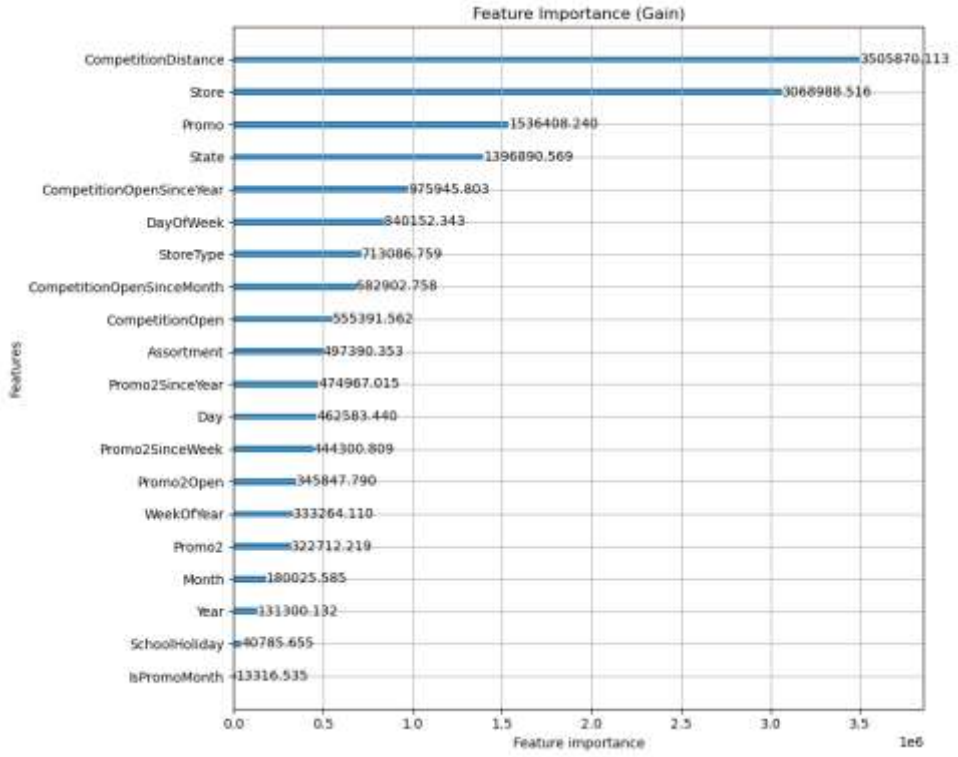
- Advanced ensemble method combining multiple weak decision trees
- Sequentially builds new trees to correct previous errors
- Each iteration refines predictions by focusing on residual errors

How the Model Predicts-lightGBM

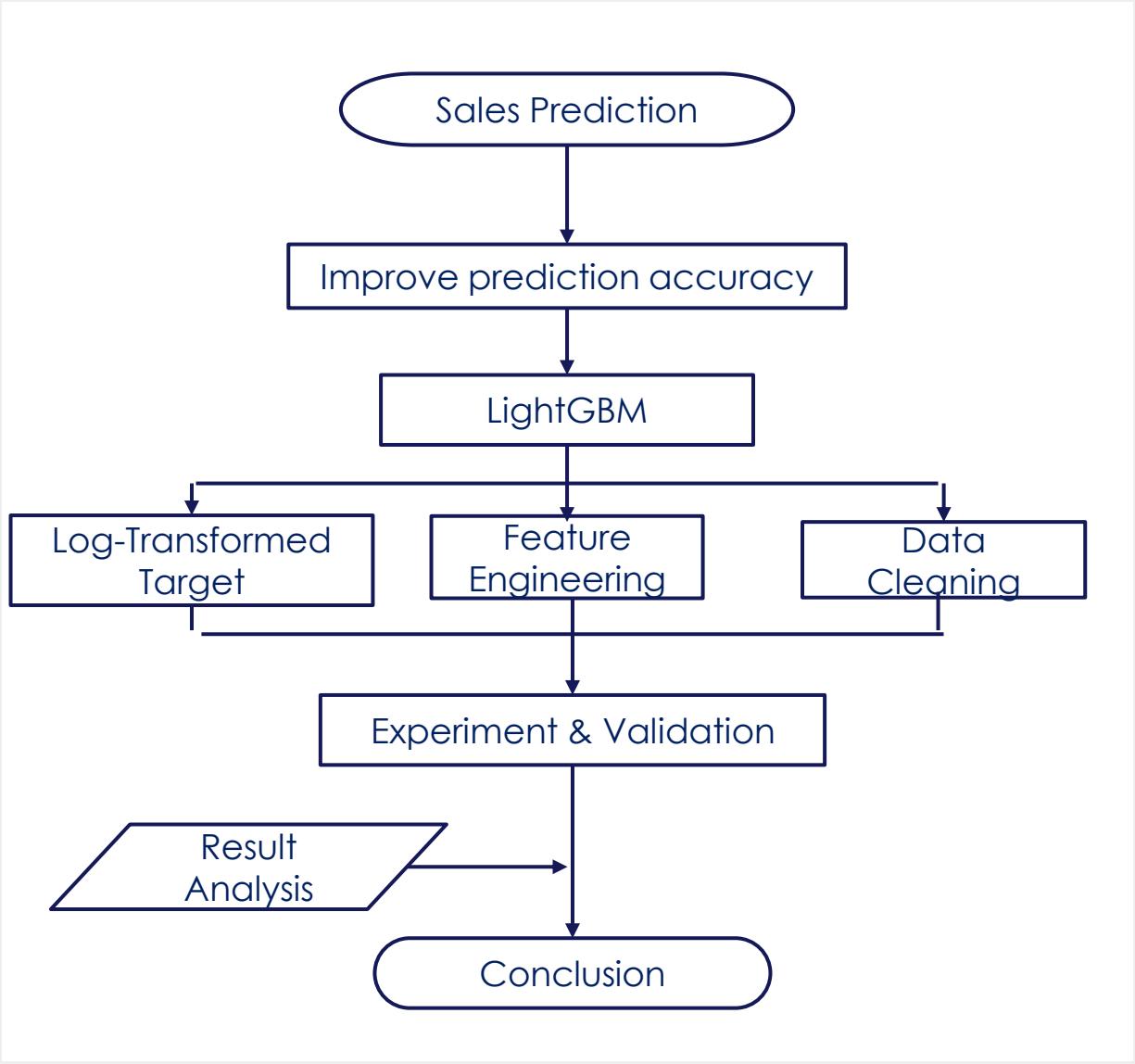


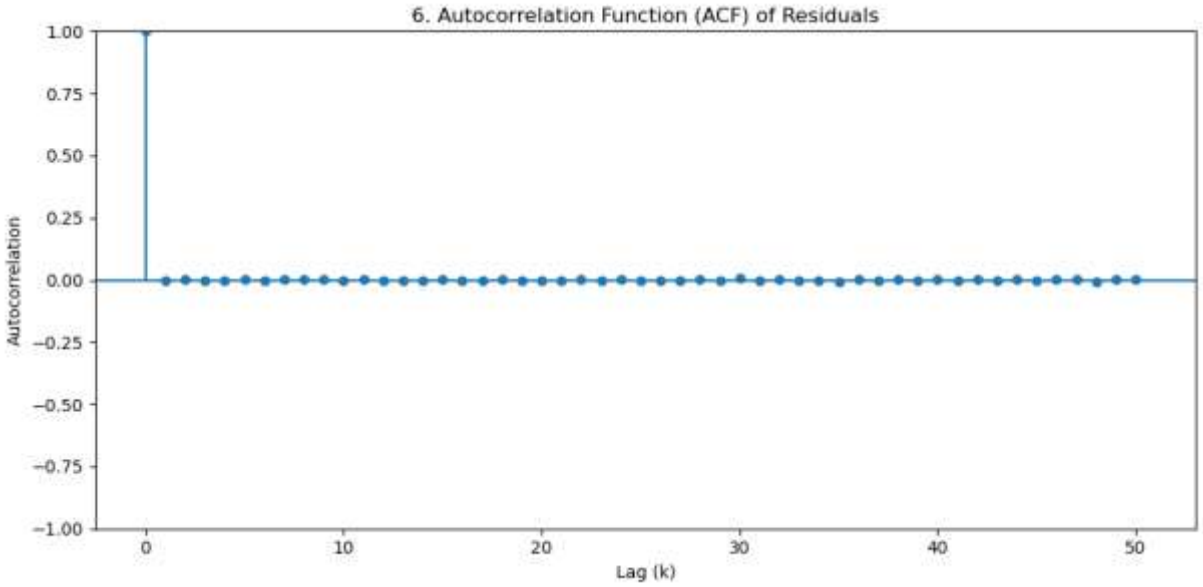
Why LightGBM?

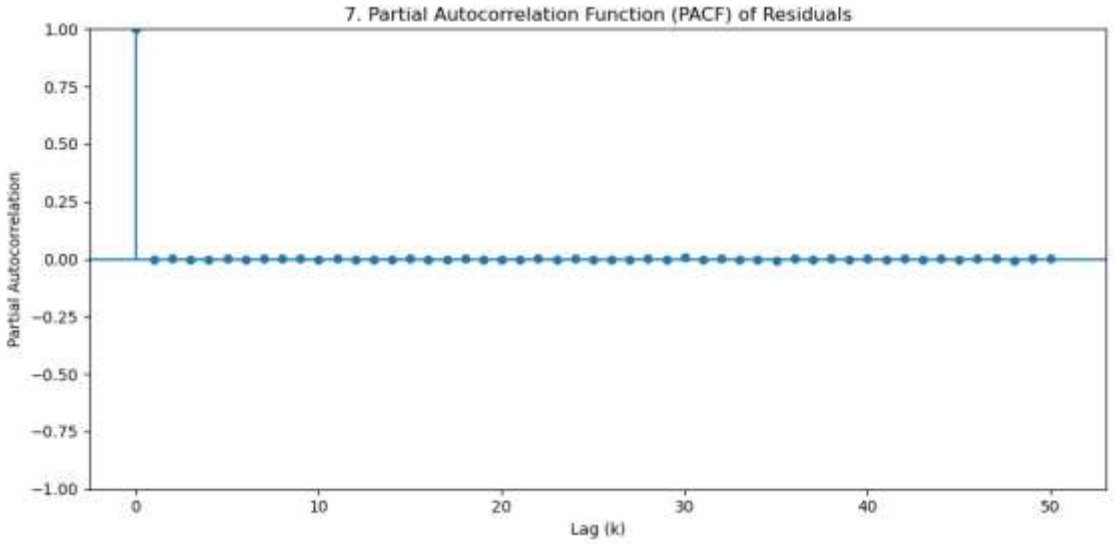
- Excels at capturing complex feature interactions
- Perfect for time series with promotions, seasonality, and competition effects



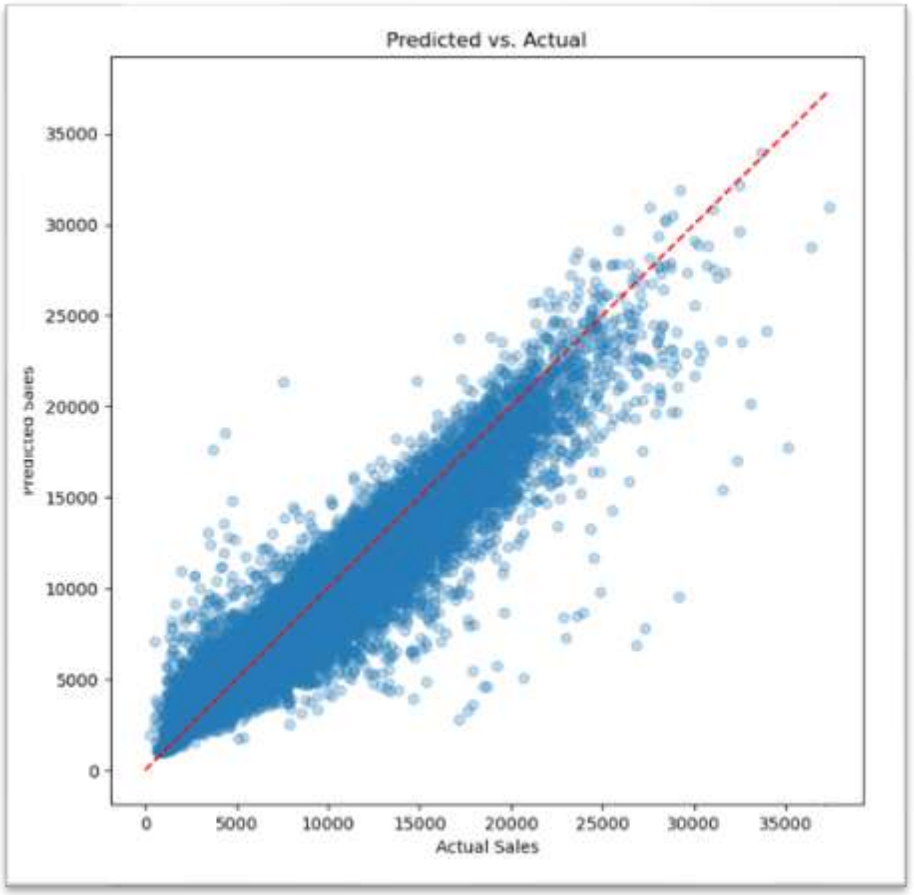
- CompetitionDistance: The #1 factor. Proximity to rivals is critical.
- Store: Individual store performance is a major differentiator.
- Promo: Promotions are our most powerful commercial lever.

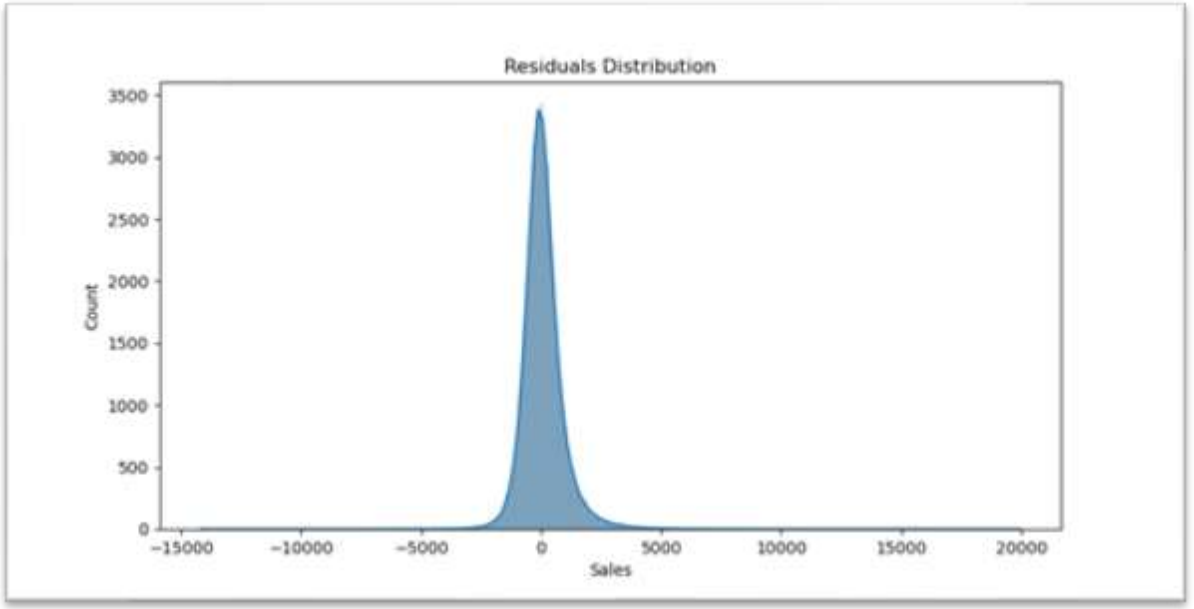






- The LightGBM model has **successfully captured** the temporal dependencies in the data.





- Model captures fundamental patterns effectively
- Residual Distribution centered near zero - indicates minimal bias

→ **Handles normal sales ranges exceptionally well**