# **DIVE Journey: Store Sales Prediction with Product Family Granularity**

### **D: Surface Discovery**

* Initial modeling efforts focused on predicting **daily sales at the store level** using basic store metadata (store number, type, cluster) and temporal features (day of week, day of month).
* The first model achieved a **moderate R² of 0.35** but exhibited a **high mean absolute error (~$5,537)**, limiting its operational usefulness.
* This suggested that aggregated sales data lacked sufficient granularity to capture the complexity of sales drivers across diverse product categories.

### **I: Deeper Investigation Findings**

* Recognizing the limitation, I **expanded the dataset to include product family as a categorical feature**, transforming the problem to predict **daily sales at the store + product family level**.
* I engineered a new feature table joining sales with store metadata and temporal variables, then trained a refined linear regression model.
* The updated model showed substantial improvement:
  + **R² increased to ~0.68**, nearly doubling explained variance.
  + **Mean absolute error decreased sharply to ~$281**, indicating much more precise predictions.
  + Median absolute error also dropped, signaling consistency.
* These findings confirm that **product family is a critical driver of sales variation** and that finer-grained data enables more accurate forecasting.

### **V: Model Limitations and Risks**

* Despite improvements, **about 32% of sales variance remains unexplained**, indicating missing variables and model complexity limits.
* Current features do not capture external factors such as:
  + Holidays, local events, and seasonality effects
  + Weather impacts and regional economic conditions
  + Historical sales trends (lagged/time-series data)
  + Store-specific competitive or demographic factors
* The model’s **linear regression approach may miss nonlinear relationships and complex interactions** present in retail data.
* Risk exists that relying solely on this model for daily operational decisions could lead to suboptimal inventory or staffing outcomes without supplementary insights.

### **E: Strategic Recommendations**

* **Incorporate Additional Features:** Enrich the dataset with external factors (holidays, weather, promotions) and time-series variables (lagged sales) to capture more variance.
* **Explore Advanced Modeling Techniques:** Implement gradient boosting, random forests, or recurrent neural networks to better model complex relationships and temporal dependencies.
* **Develop Category-Level Dashboards:** Use model predictions to empower store managers with granular forecasts, enabling optimized inventory management and promotional planning.
* **Implement Continuous Monitoring and Retraining:** Set up a feedback loop to track model performance, capture concept drift, and retrain models regularly with fresh data.
* **Pilot and Integrate with Business Processes:** Combine model insights with domain expertise to guide decisions and gradually automate forecasting workflows, reducing operational risks.