Time Series Forecasting for Retail Sales

Armaan Dhanda - 906310394
Georgy Jacob - 506309986
Nithin Srivatsa - 506310386

Executive Summary

This time series forecasting project focused on predicting future sales across three product categories: Furniture, Technology, and Office Supplies using three different models: SARIMA, Prophet, and LSTM.
The dataset, derived from a fictional Superstore, contained detailed information about orders, customers, shipping methods, and products.

The project utilized three distinct forecasting models, each chosen for its unique strengths in time series analysis. SARIMA, a traditional statistical model, was selected for its ability to handle seasonality and trends in linear data. However, its limitations became evident when faced with more complex, non-linear sales patterns. Prophet, a model developed by Facebook, was chosen for its flexibility in modeling both linear and non-linear trends, as well as its ability to handle missing data and sudden trend changes (also known as changepoints). Finally, the deep learning model LSTM was employed to capture long-term dependencies in the sales data, a feature that can be especially useful when patterns in the data stretch over extended periods.

After training and evaluating each model, it was clear that Prophet performed the best overall, with the lowest or second lowest Root Mean Squared Error (RMSE) in all three categories. Prophet excelled in capturing the non-linear trends and seasonal variations present in the sales data for both Furniture and Technology. LSTM, while computationally more intensive and requiring careful tuning to avoid overfitting, performed ahead of Prophet in the Technology category. SARIMA, although the least effective model for non-linear trends, performed best for the Office Supplies category, where sales followed a more stable and linear pattern, achieving the lowest RMSE for this category.
In conclusion, Prophet emerged as the most effective forecasting model due to its flexibility and ability to handle complex, non-linear patterns in the sales data. LSTM was a close second, particularly in handling the Technology category's sales, where long-term dependencies played a significant role. SARIMA, while useful for simpler, linear trends, was less effective overall but demonstrated its strengths in categories with more consistent sales patterns like Office Supplies.

This project highlighted the importance of model selection and flexibility, along with respecting different models and not making assumptions about the nature of the data.

Dataset Overview and Key Insights

We're working with a Superstore dataset that captures detailed sales transactions across multiple dimensions. The dataset includes:

- Transactional Data: Unique order IDs and their corresponding dates
- Logistics Information: Shipping details and delivery methods
- Customer Insights: Detailed customer segmentation and geographic distribution
- Product Portfolio: Hierarchical categorization of products
- Financial Metrics: Key performance indicators including Sales, Quantity, Discount rates, and Profit margins

What makes this dataset particularly interesting for time series analysis is its rich temporal structure, with sales patterns varying across different product categories (Furniture, Office Supplies, and Technology). The granular nature of the data, with 21 columns capturing various aspects of each transaction, provides an excellent foundation for building robust forecasting models.

Data Exploration and Preprocessing

The exploratory data analysis (EDA) revealed several crucial insights that guided our modeling decisions:

1. Temporal Processing: We transformed the string-formatted dates into datetime objects, which is crucial for capturing temporal dependencies.

2. Weekly Aggregation Strategy: Rather than working with daily data, which often contains too much noise, or monthly data, which might miss important patterns, we chose weekly aggregation (W-MON). This decision was driven by:
    a. The need to capture weekly business cycles
    b. Sufficient data points for model training
    c. Reduction of daily volatility while preserving meaningful patterns
    d. Ability to detect month-end spikes and seasonal trends

3. Category-specific Analysis: Breaking down the analysis by product categories revealed distinct sales patterns:
    a. Furniture: Shows strong seasonal patterns with high variance
    b. Office Supplies: Exhibits more stable, predictable trends.
    c. Technology: Demonstrates complex patterns with both trend and seasonality

Modeling Approach

Our model selection strategy focused on three powerful but fundamentally different approaches to time series forecasting:

SARIMA (Seasonal ARIMA)

- Traditional powerhouse in time series analysis
- Particularly strong at handling:
- Clear seasonal patterns
- Linear trends
- Autoregressive relationships
- Key limitation: Assumes linear relationships between variables
- Best suited for the Office Supplies category due to its more linear nature

Prophet

Facebook's model, it is a modern approach to time series forecasting

- Automatic changepoint detection
- Robust handling of missing data
- Flexible seasonal modeling
- Built-in holiday effects
- Particularly effective for retail data due to its ability to handle:
- Multiple seasonality patterns
- Irregular events
- Trend changes
- Non-linear patterns

LSTM (Long Short-Term Memory)

Deep learning approach to sequence modeling

- Capturing complex non-linear patterns
- Learning long-term dependencies
- Handling multiple input features
- Particularly strong with the Technology category due to its ability to capture complex relationships

This approach was taken because SARIMA is great at Linear relationships, whereas Prophet is adept at seasonality, and LSTM is strong when it comes non-linear data. This allows us to explore different aspects of the data and understand it on a deeper level.

Implementation Details and Technical Insights

- SARIMA:
    - Parameter selection through grid search and information criteria (AIC/BIC)
    - Diagnostic checking using:
    - ACF/PACF plots
    - Residual analysis
    - Ljung-Box test for autocorrelation
    - Notable observation: While SARIMA showed decent performance for Office Supplies, it struggled with more complex categories. It did well in Office Supplies because it was a more stable dataset as compared to furniture or tech, whereas more complex models like LSTM and Prophet did better.

- Prophet:
    - Implemented with custom seasonality settings
    - Utilized changepoint_prior_scale tuning for optimal trend flexibility
    - Incorporated holiday effects
    - Achieved best overall performance
- LSTM:
    - Sequence length optimization
    - Feature scaling using MinMaxScaler
    - Architecture:
    - Multiple LSTM layers
    - Dropout for regularization
    - Dense output layer
    - Strong performance on Technology category, showing the value of deep learning for complex patterns

Performance Analysis and Model Comparison

The performance metrics revealed interesting patterns:
Category-wise Analysis:

- Furniture:
    - Prophet led with RMSE.
    - Higher error due to inherent volatility in furniture sales
    - LSTM performed second-best, handling non-linear patterns well

- Technology:
    - Prophet and LSTM showed comparable performance
    - LSTM was slightly better overall.
    - LSTM's ability to capture complex patterns particularly valuable here

    - Office Supplies:
        - Surprisingly, SARIMA performed best.
        - Indicates more stable, predictable patterns in this category

Model	**Furniture RMSE**	**Technology RMSE**	**Office Supplies RMSE**
SARIMA	598.95	1003.17	428.04
Prophet	264.449	822.67	432.57
LSTM	3014.69	716.81	495.77

Critical Evaluation and Future Improvements

- Current Strengths:
    - Robust weekly aggregation strategy
    - Comprehensive model comparison
    - Strong performance across different product categories

- Areas for Enhancement:
    - Consider incorporating external features:
    - Macroeconomic indicators
    - Competitor pricing data
    - Marketing campaign information
    - Weather data for seasonal products

- Additional Model Considerations:
    - Ensemble methods combining multiple models
    - XGBoost with time-based features

- Neural Prophet for combining Prophet's strengths with neural networks
- Vector AutoRegression (VAR) for capturing inter-category relationships

Technical Recommendations

1. Data Pipeline Optimization:
    - Implement automated anomaly detection
    - Develop robust feature engineering pipeline
    - Create automated model retraining schedule

2. Model Enhancement:
    - Explore hybrid models combining statistical and ML approaches
    - Implement hierarchical forecasting for category-subcategory relationships
    - Consider quantile forecasting for uncertainty estimation

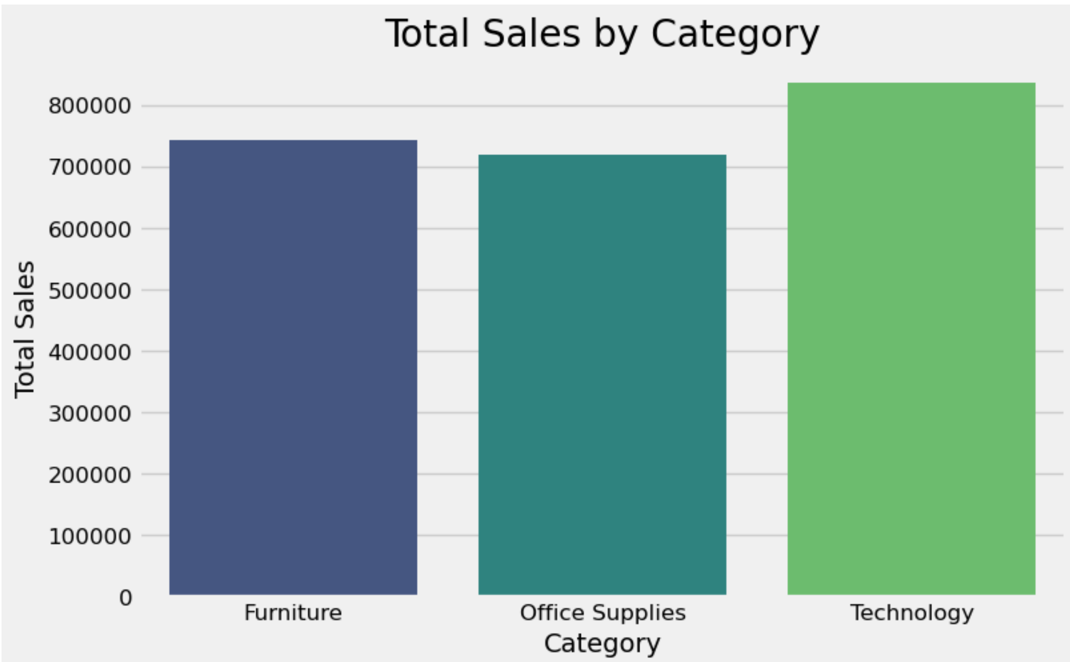
3. Production Considerations:
    - Model versioning and monitoring
    - Performance tracking across different time horizons
    - Automated retraining triggers based on forecast deviation

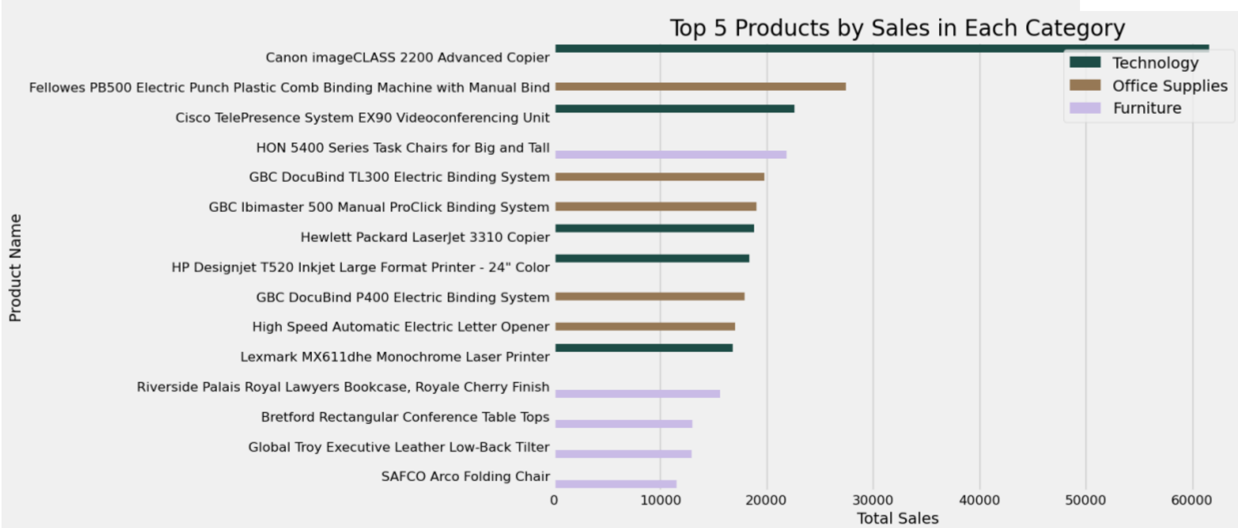
Conclusion

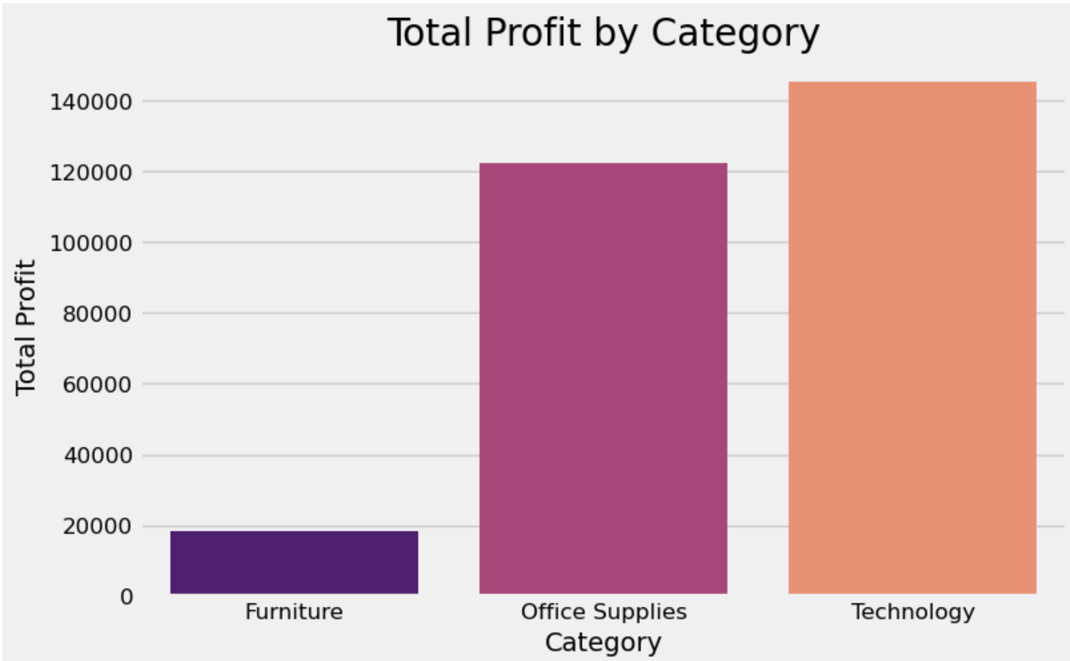
This analysis demonstrates the importance of matching model selection to data characteristics. Prophet's superior performance highlights the value of flexible, modern approaches to time series forecasting, while LSTM's strong showing in the Technology category validates the use of deep learning for complex patterns. The surprising effectiveness of SARIMA for Office Supplies reminds us that simpler models shouldn't be discounted when the data patterns align with their assumptions. This highlights that our approach in model selection was quite successful, since it showcased the different strengths and weaknesses of all 3 models.

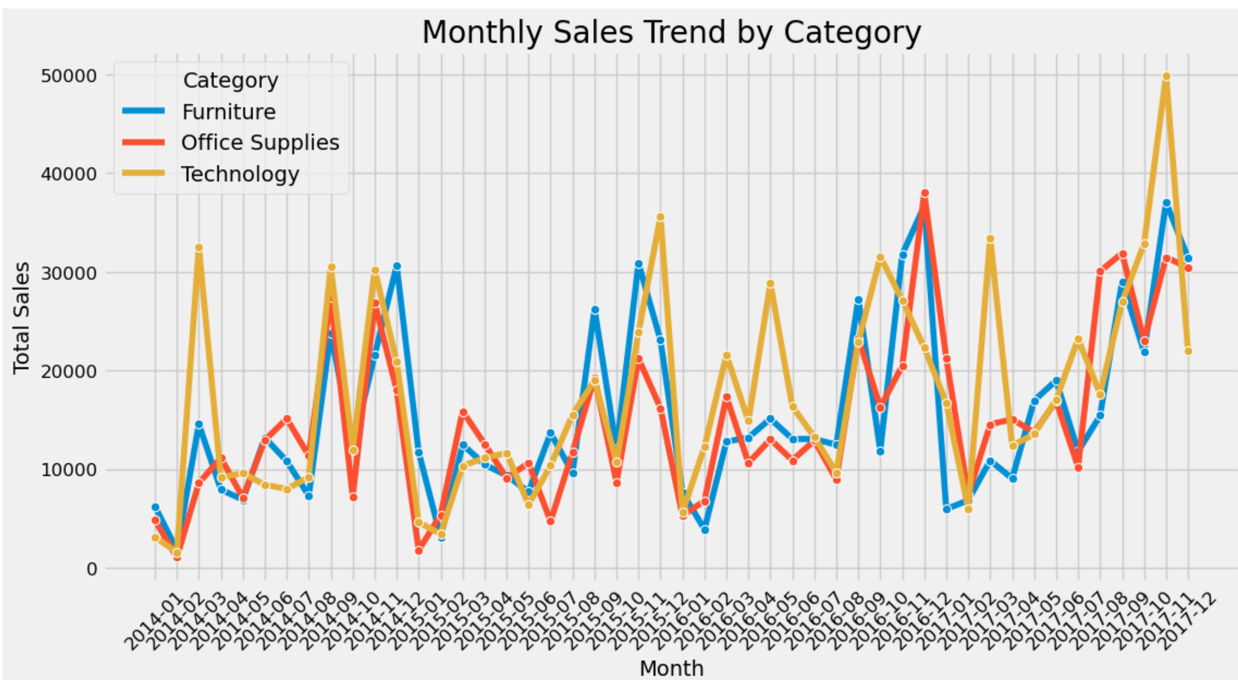
The project successfully delivered accurate forecasts across different product categories, with each model contributing unique insights. The combination of traditional statistical methods (SARIMA), modern algorithms (Prophet), and deep learning (LSTM) provided a robust framework for understanding and predicting sales patterns in retail data.

Appendix

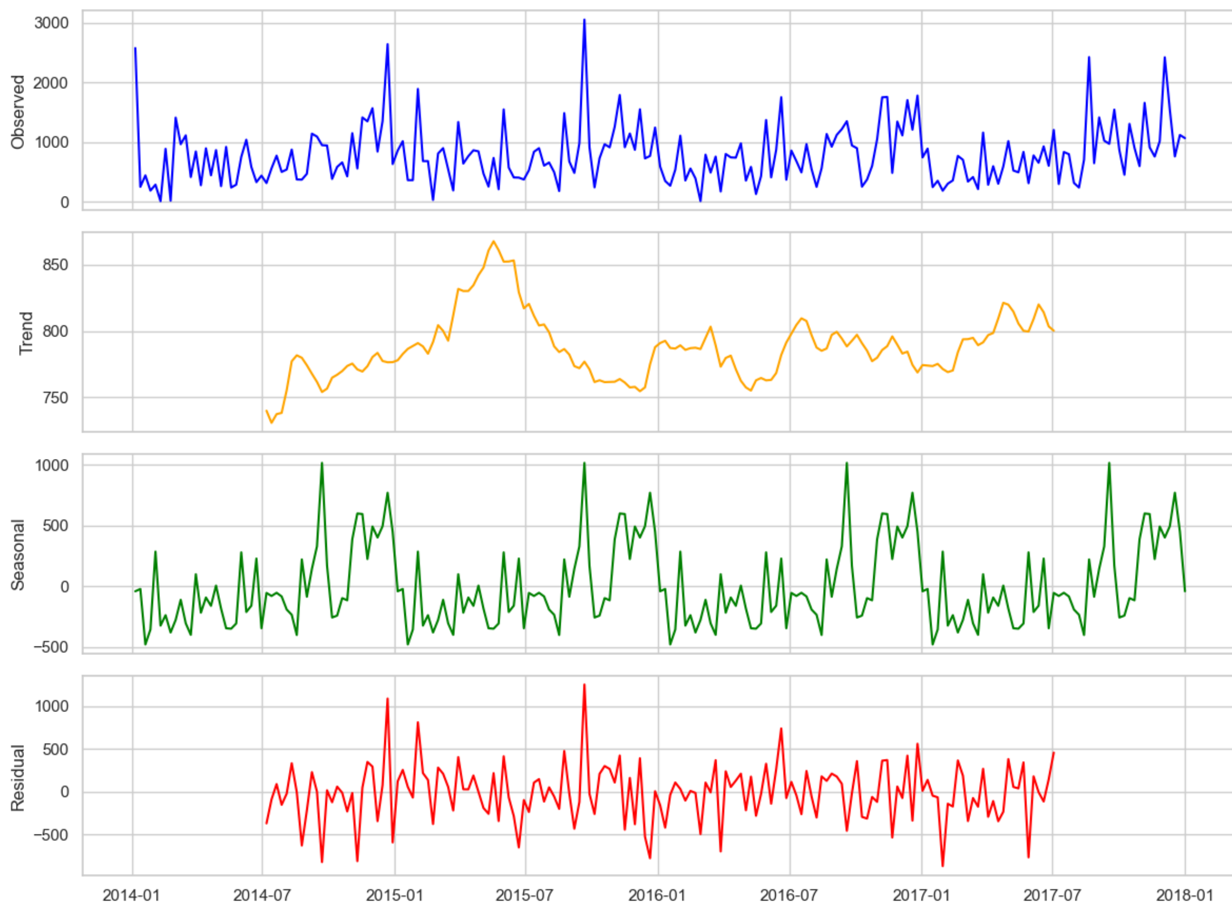

Total Sales by Category
800000
700000
600000
500000
400000
300000
200000
100000
0
Total Sales
Furniture
Office Supplies
Technology
Category


Top 5 Products by Sales in Each Category
Canon imageCLASS 2200 Advanced Copier
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind
Cisco TelePresence System EX90 Videoconferencing Unit
HON 5400 Series Task Chairs for Big and Tall
GBC DocuBind TL300 Electric Binding System
GBC Ibimaster 500 Manual ProClick Binding System
Hewlett Packard LaserJet 3310 Copier
HP Designjet T520 Inkjet Large Format Printer - 24" Color
GBC DocuBind P400 Electric Binding System
High Speed Automatic Electric Letter Opener
Lexmark MX611dhe Monochrome Laser Printer
Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish
Bretford Rectangular Conference Table Tops
Global Troy Executive Leather Low-Back Tilter
SAFCO Arco Folding Chair
Product Name
Technology
Office Supplies
Furniture
0
10000
20000
30000
40000
50000
60000
Total Sales

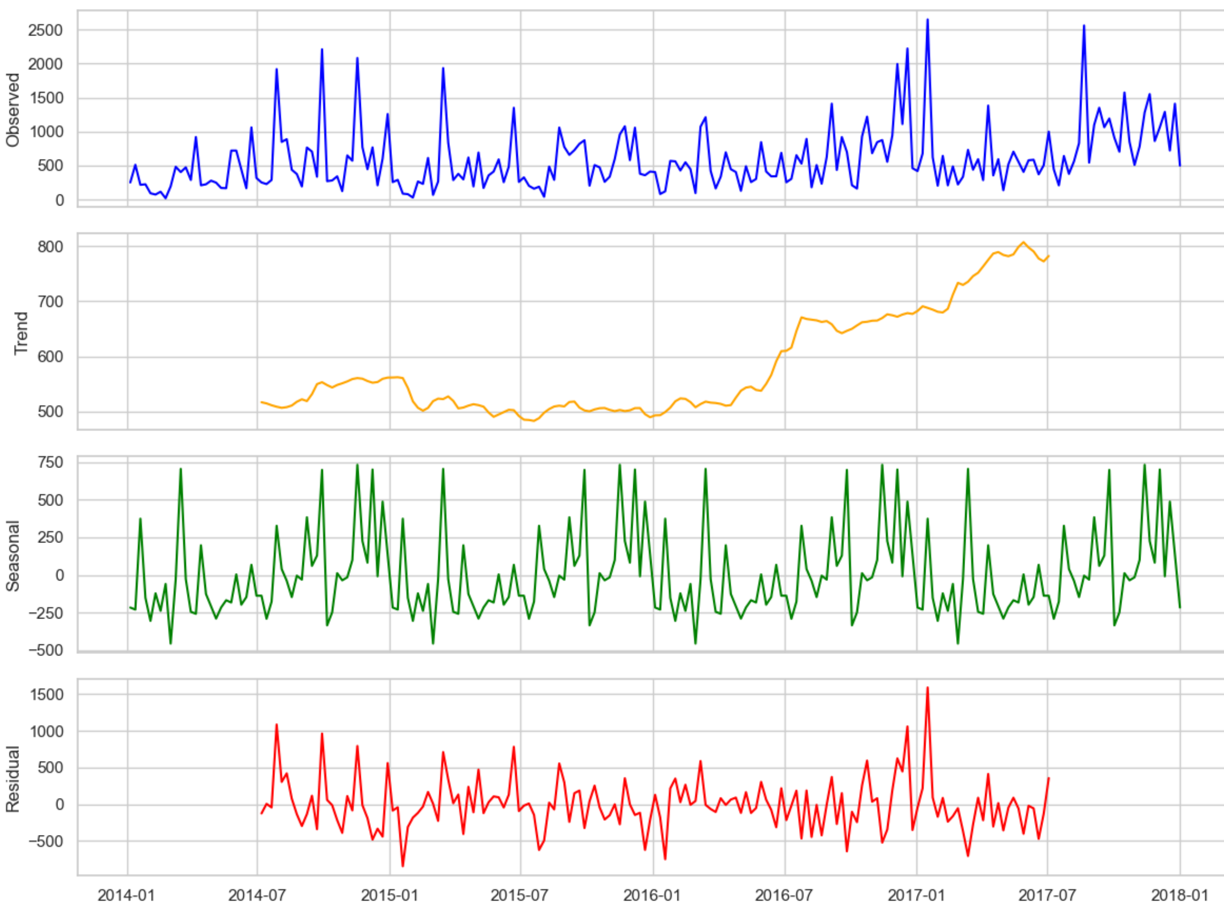

Total Profit by Category
140000
120000
100000
80000
60000
40000
20000
0
Total Profit
Furniture
Office Supplies
Technology
Category

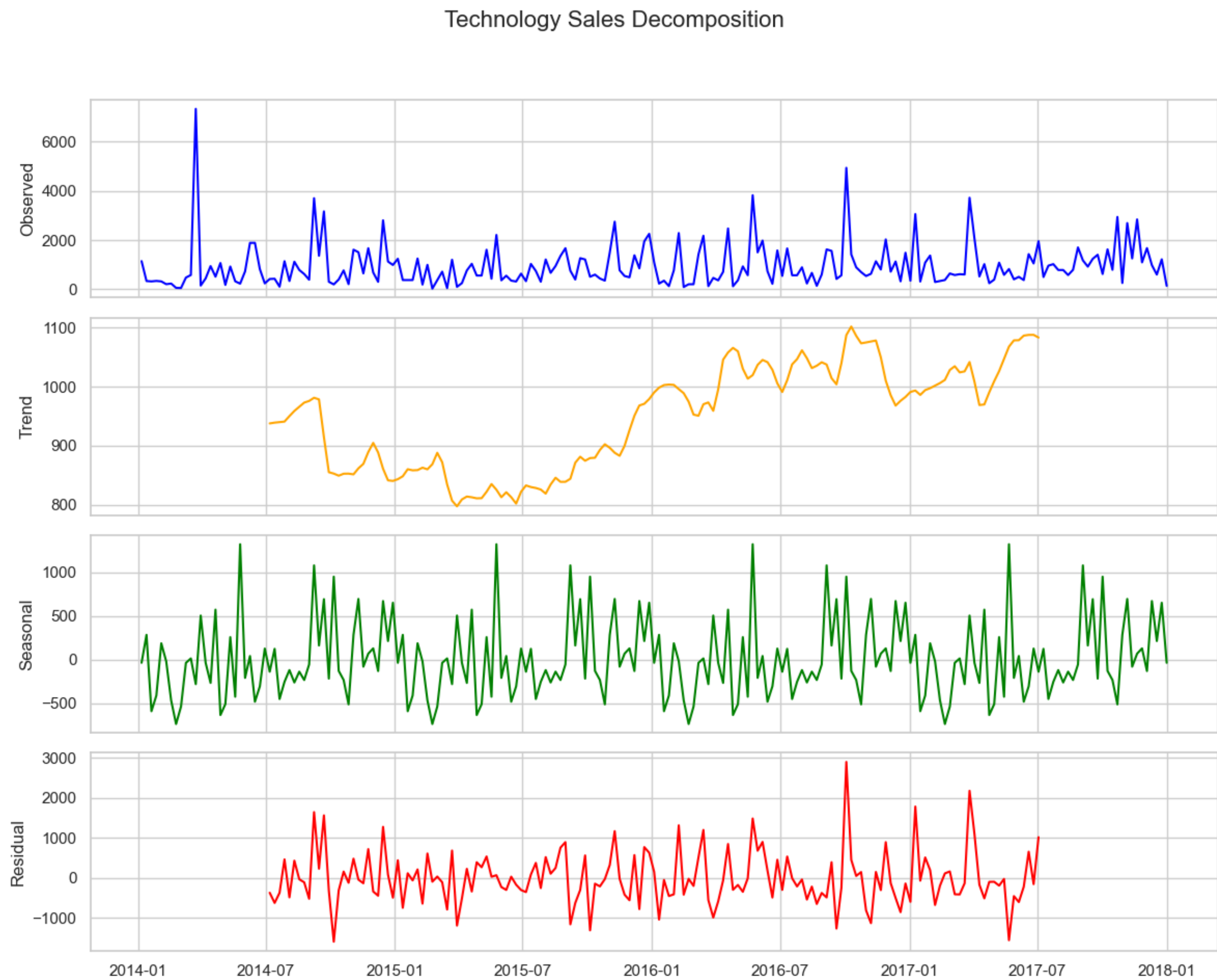

Monthly Sales Trend by Category
Category
Furniture
Office Supplies
Technology
50000
40000
30000
20000
10000
0
Total Sales
2014-01
2014-02
2014-03
2014-04
2014-05
2014-06
2014-07
2014-08
2014-09
2014-10
2014-11
2014-12
2015-01
2015-02
2015-03
2015-04
2015-05
2015-06
2015-07
2015-08
2015-09
2015-10
2015-11
2015-12
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
2017-12
Month

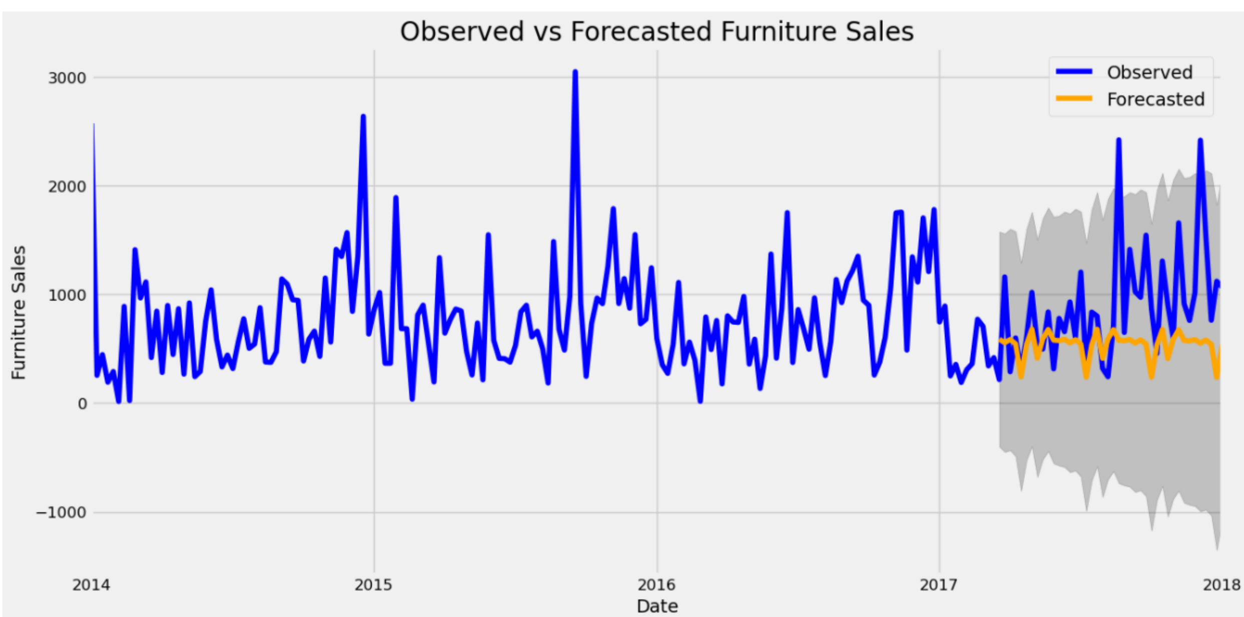
Furniture Sales Decomposition

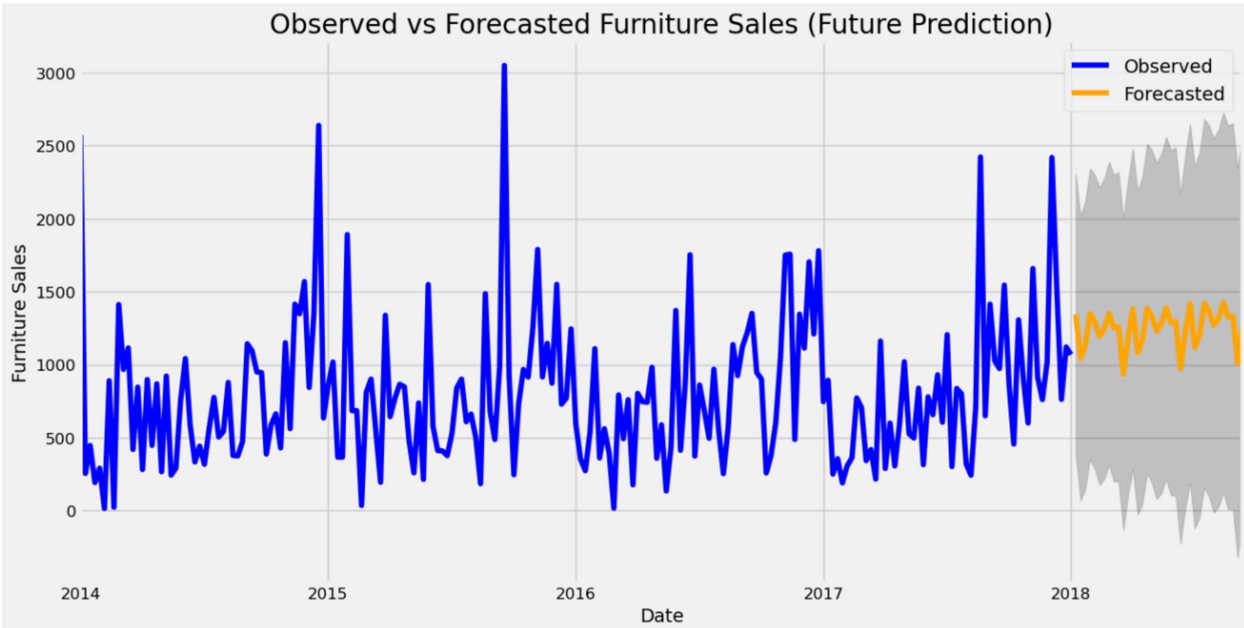


Office Supplies Sales Decomposition

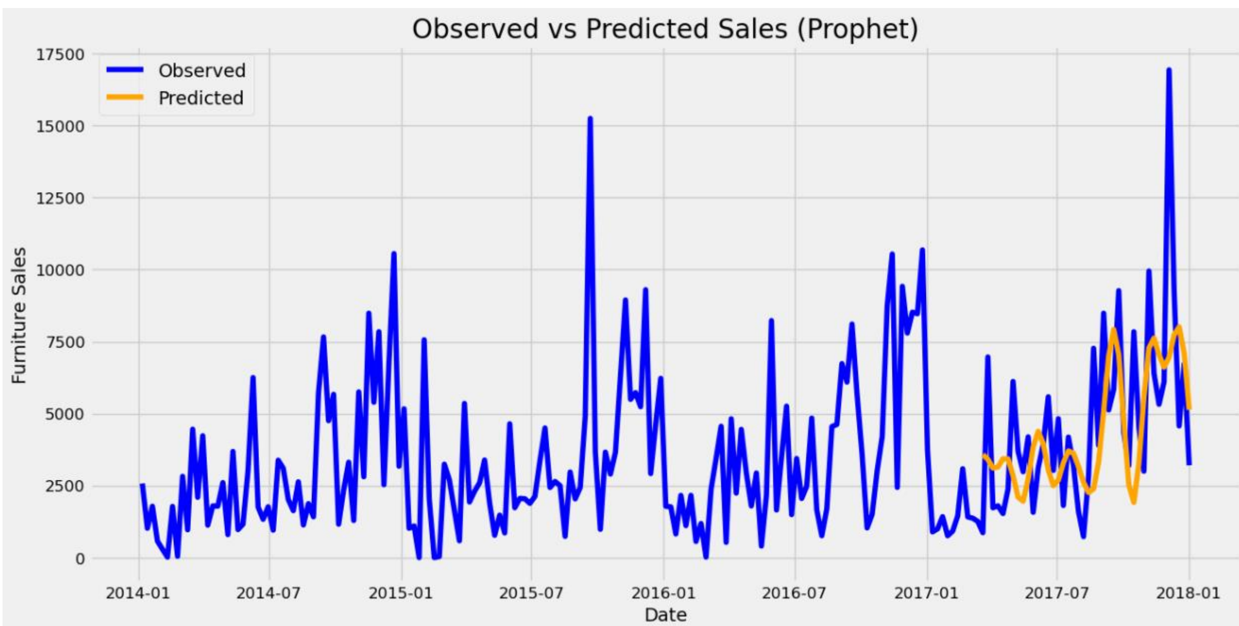



Technology Sales Decomposition
Observed
Trend
Seasonal
Residual
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01

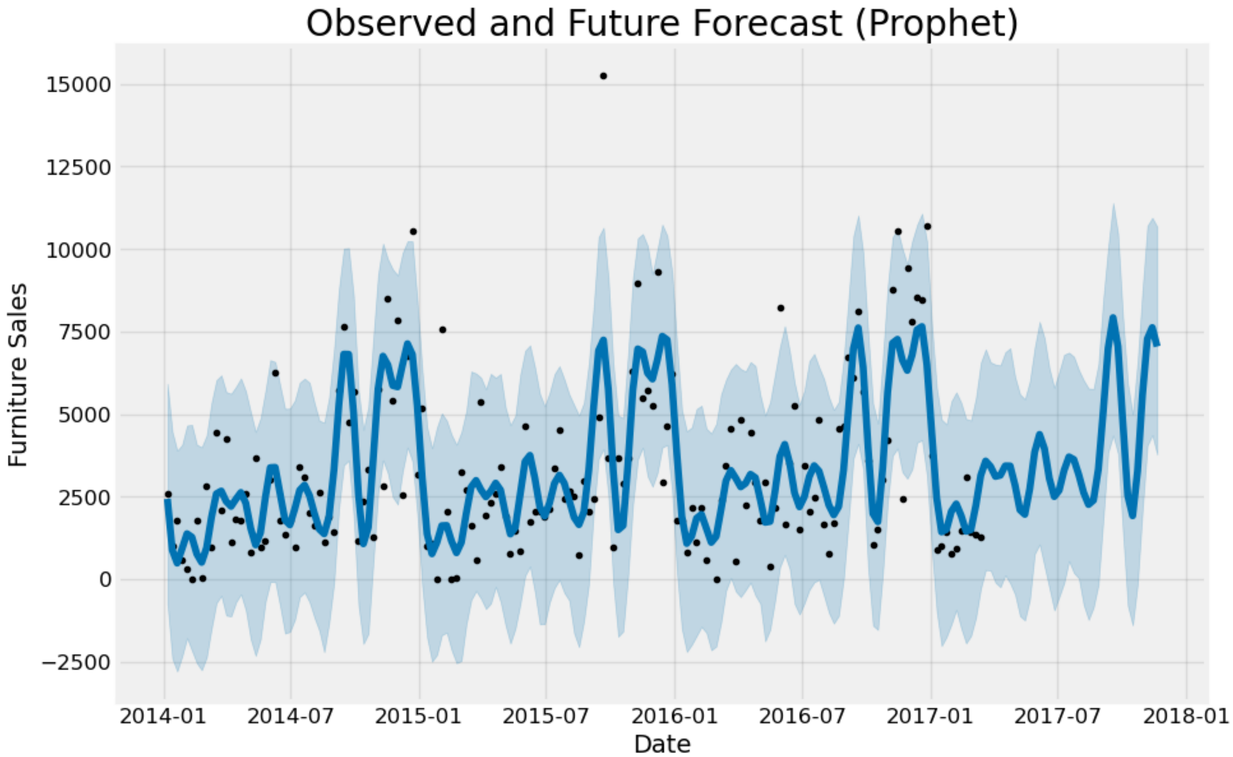

Observed vs Forecasted Furniture Sales
Observed
Forecasted
Furniture Sales
Date
2014
2015
2016
2017
2018

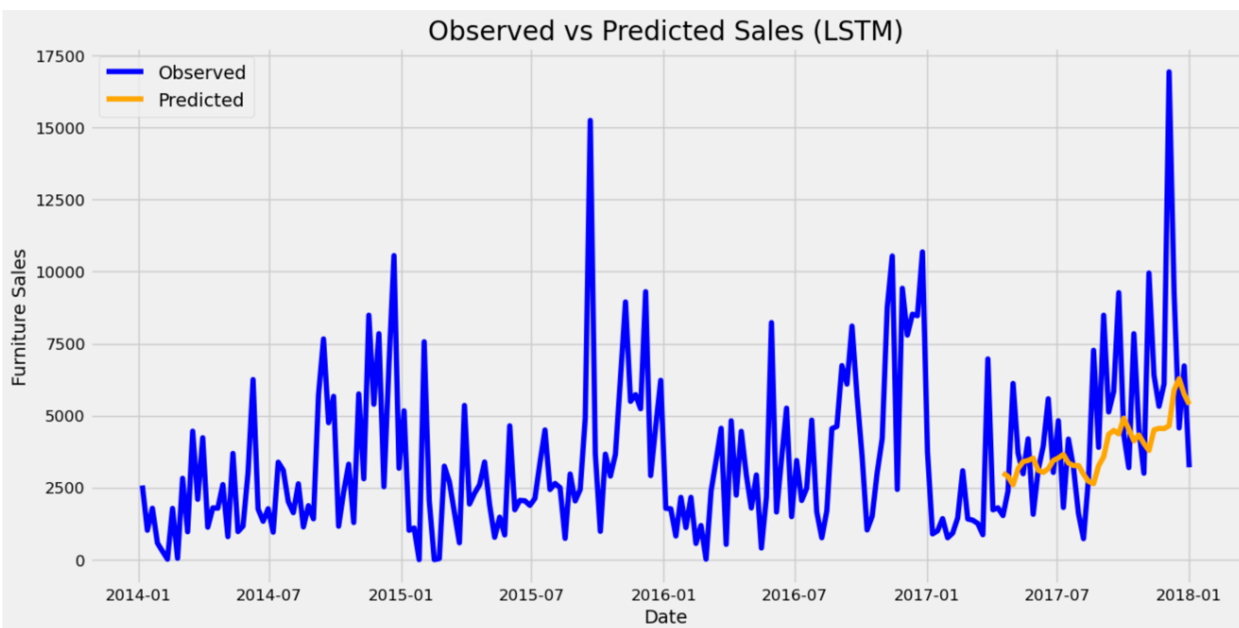


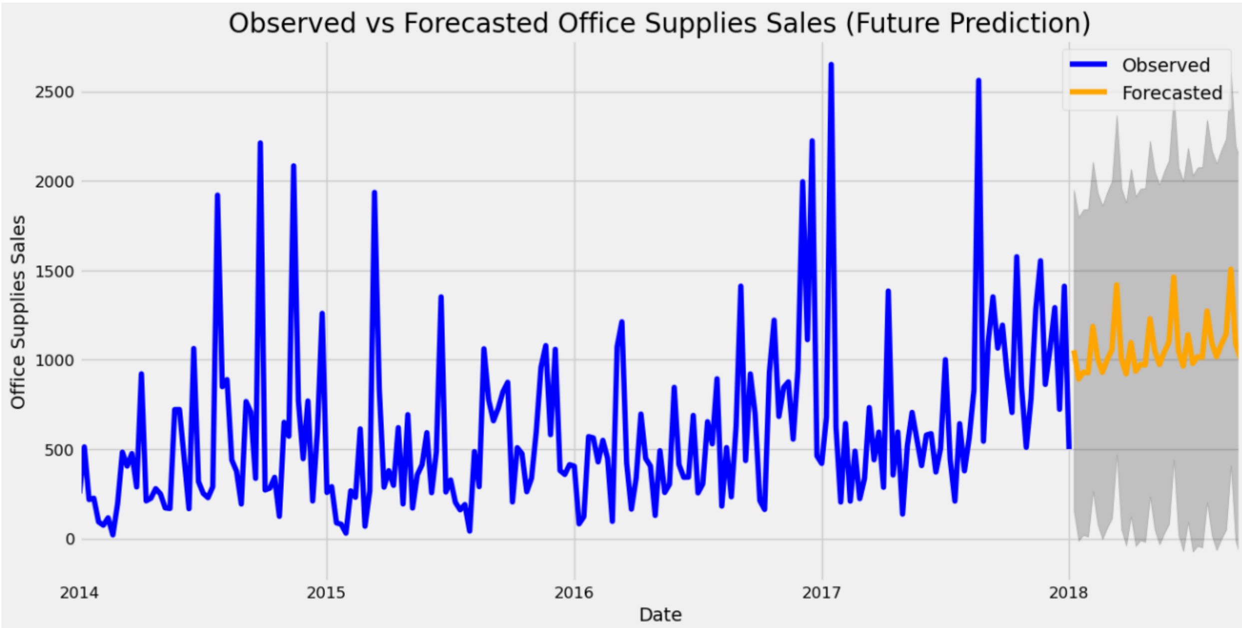
Observed vs Forecasted Furniture Sales (Future Prediction)
Observed
Forecasted
Furniture Sales
3000
2500
2000
1500
1000
500
0
2014
2015
2016
2017
2018
Date

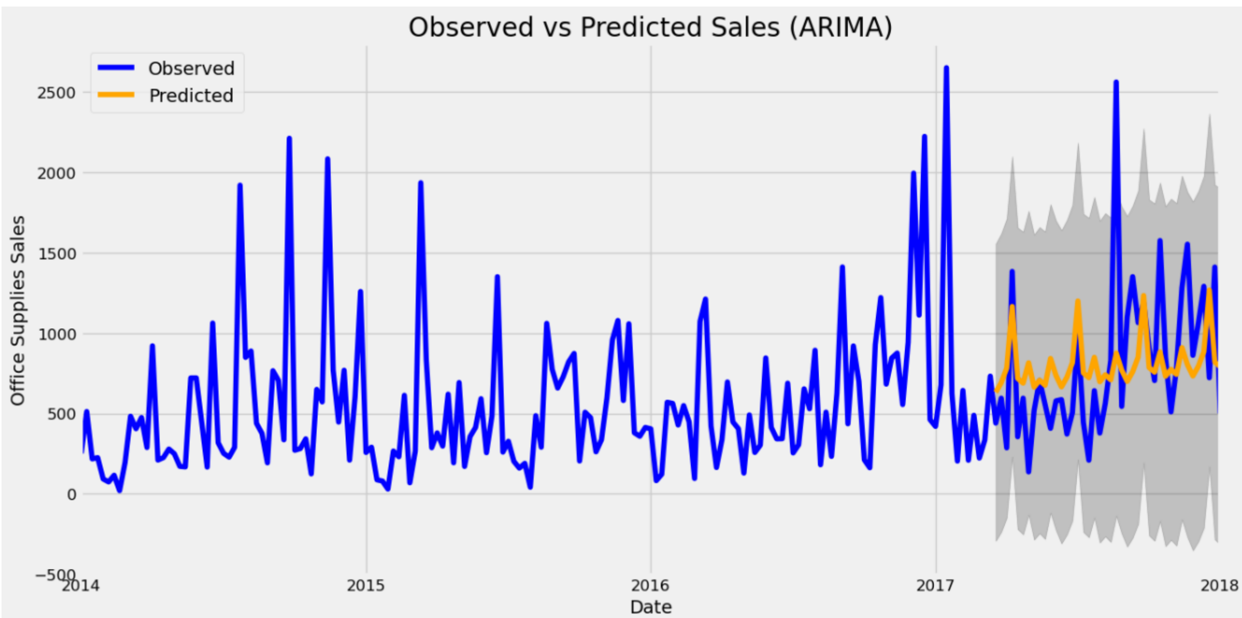


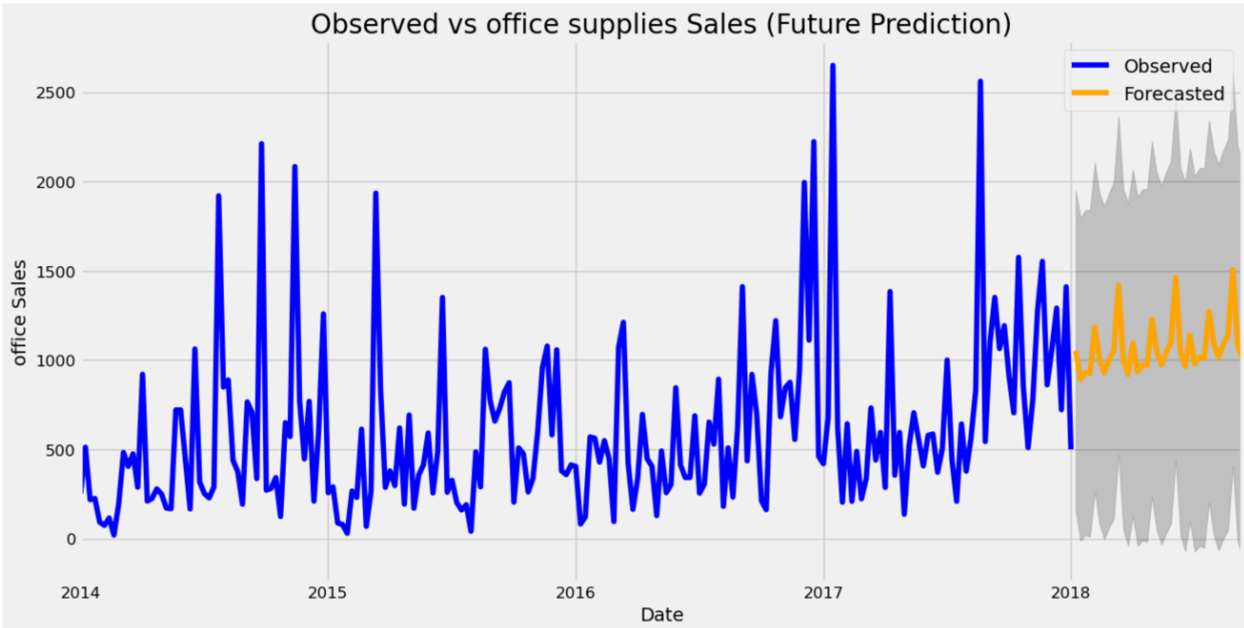
Observed vs Predicted Sales (Prophet)
Observed
Predicted
Furniture Sales
17500
15000
12500
10000
7500
5000
2500
0
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Date

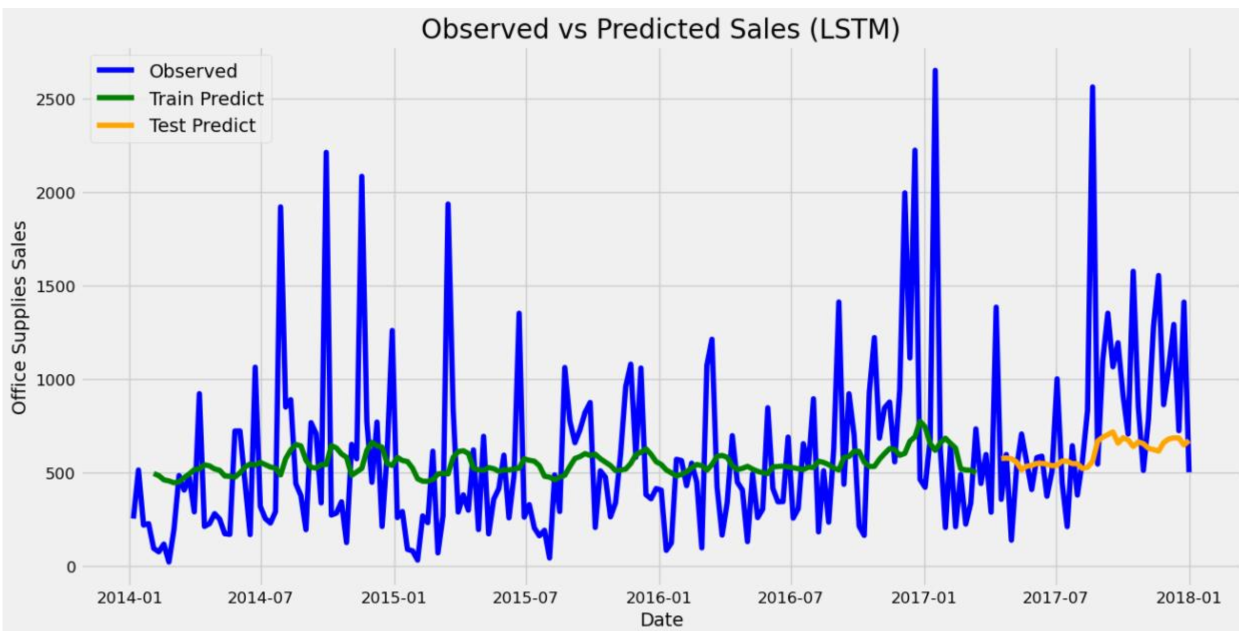

Observed and Future Forecast (Prophet)
Furniture Sales
Date
15000
12500
10000
7500
5000
2500
0
−2500
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01

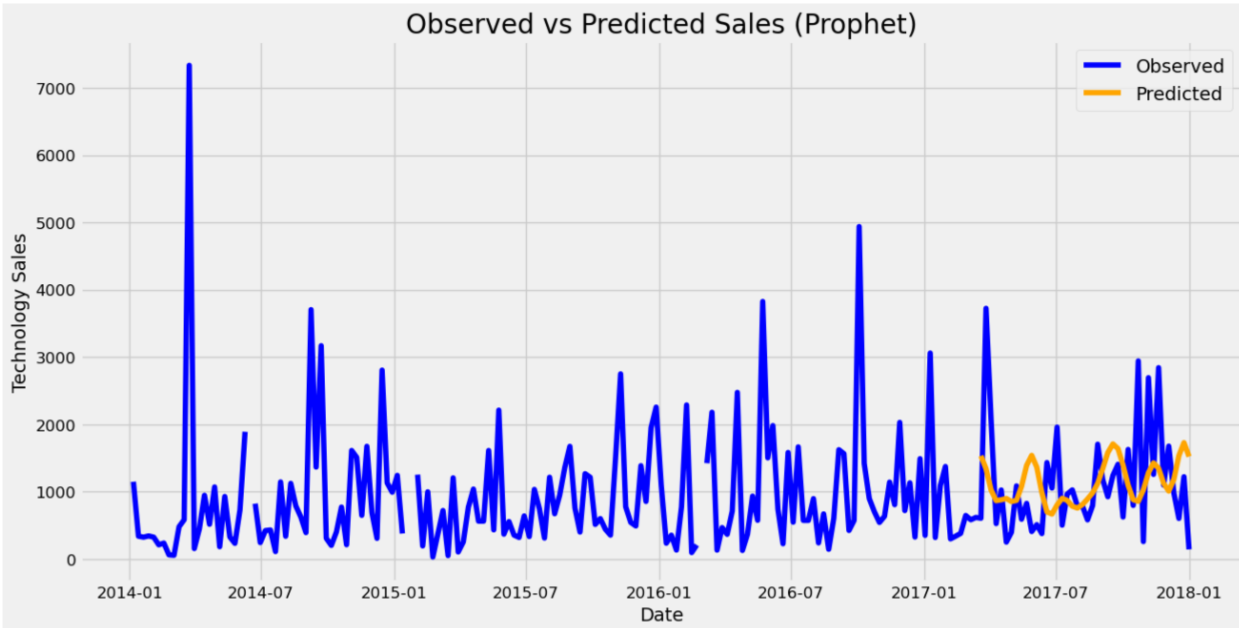

Observed vs Predicted Sales (LSTM)
Observed
Predicted
Furniture Sales
Date
17500
15000
12500
10000
7500
5000
2500
0
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01

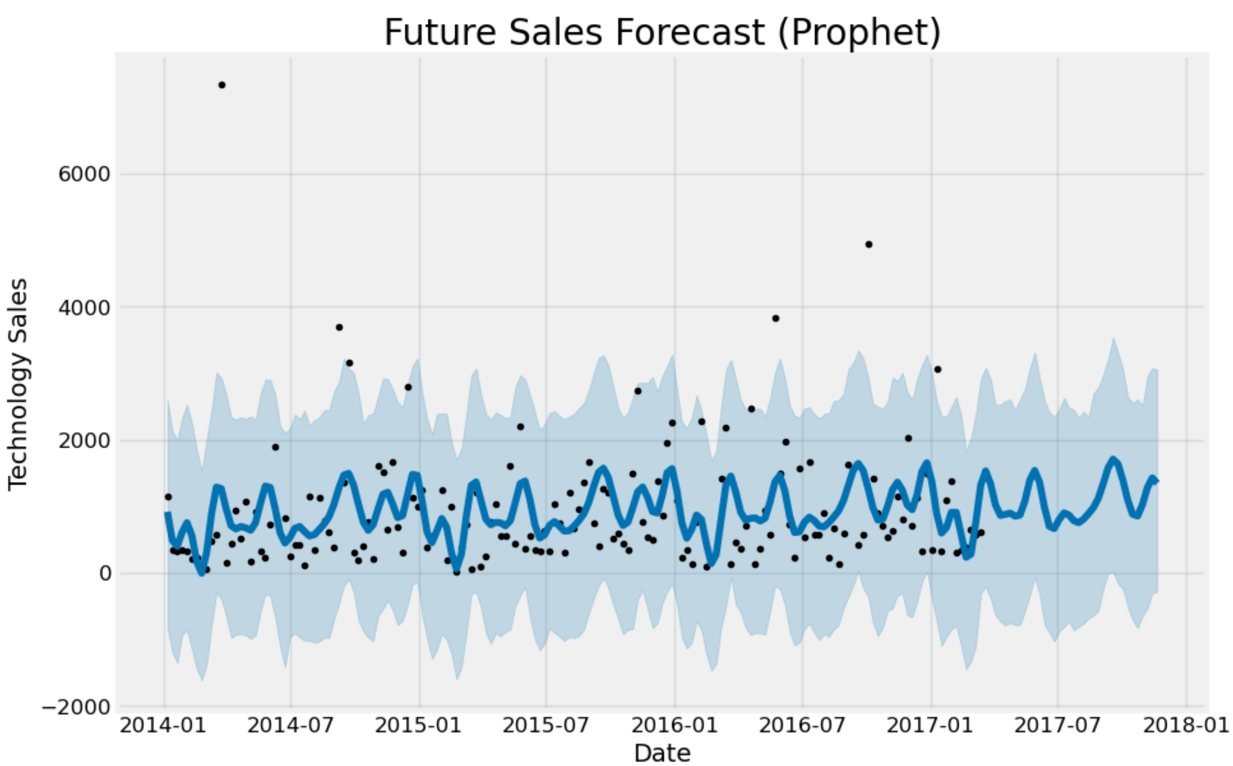

Observed vs Forecasted Office Supplies Sales (Future Prediction)
Observed
Forecasted
Office Supplies Sales
0
500
1000
1500
2000
2500
2014
2015
2016
2017
2018
Date

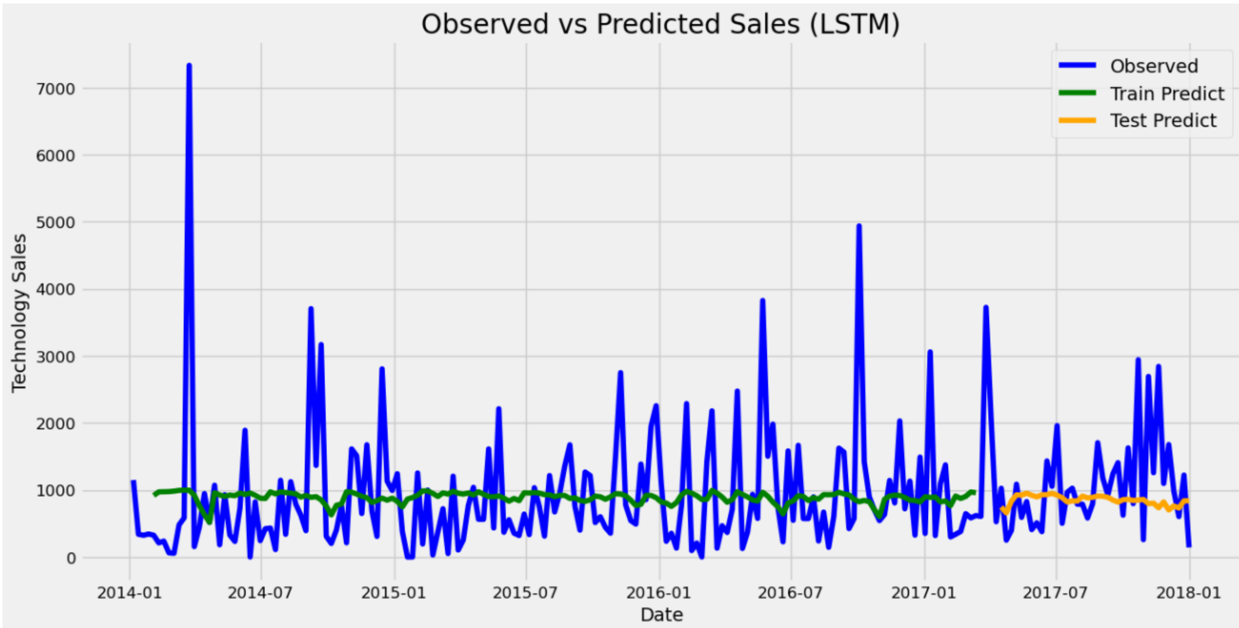

Observed vs Predicted Sales (ARIMA)
Observed
Predicted
Office Supplies Sales
−500
0
500
1000
1500
2000
2500
2014
2015
2016
2017
2018
Date

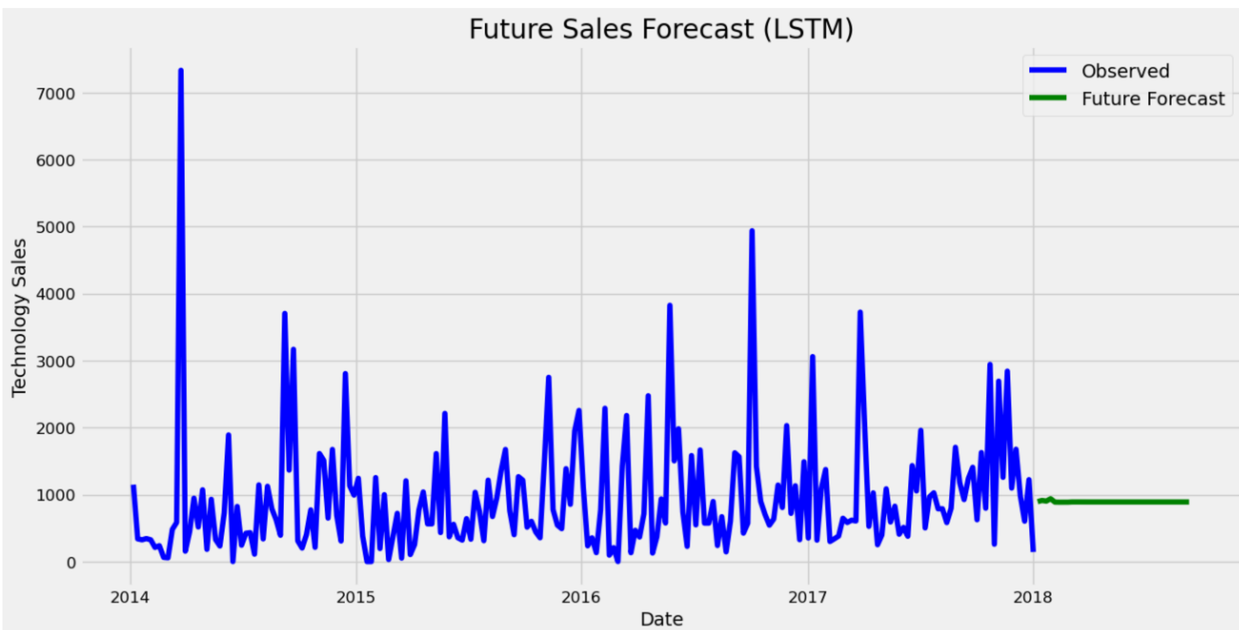

Observed vs office supplies Sales (Future Prediction)
Observed
Forecasted
office Sales
Date
2014
2015
2016
2017
2018
0
500
1000
1500
2000
2500

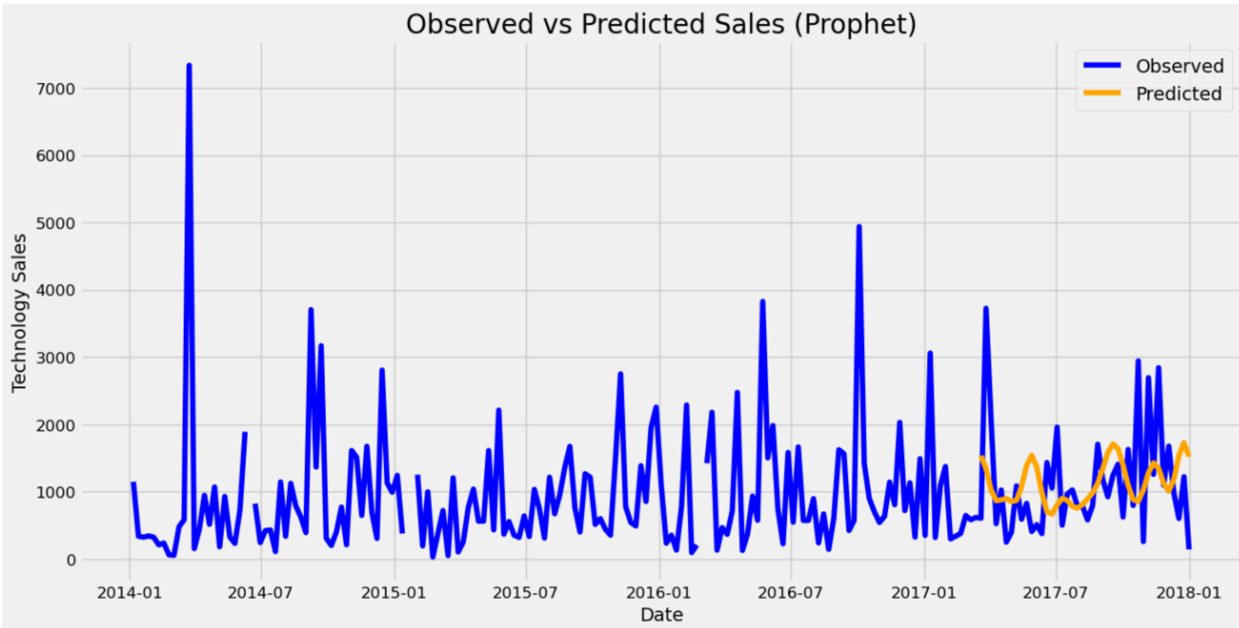

Observed vs Predicted Sales (LSTM)
Observed
Train Predict
Test Predict
Office Supplies Sales
Date
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
0
500
1000
1500
2000
2500

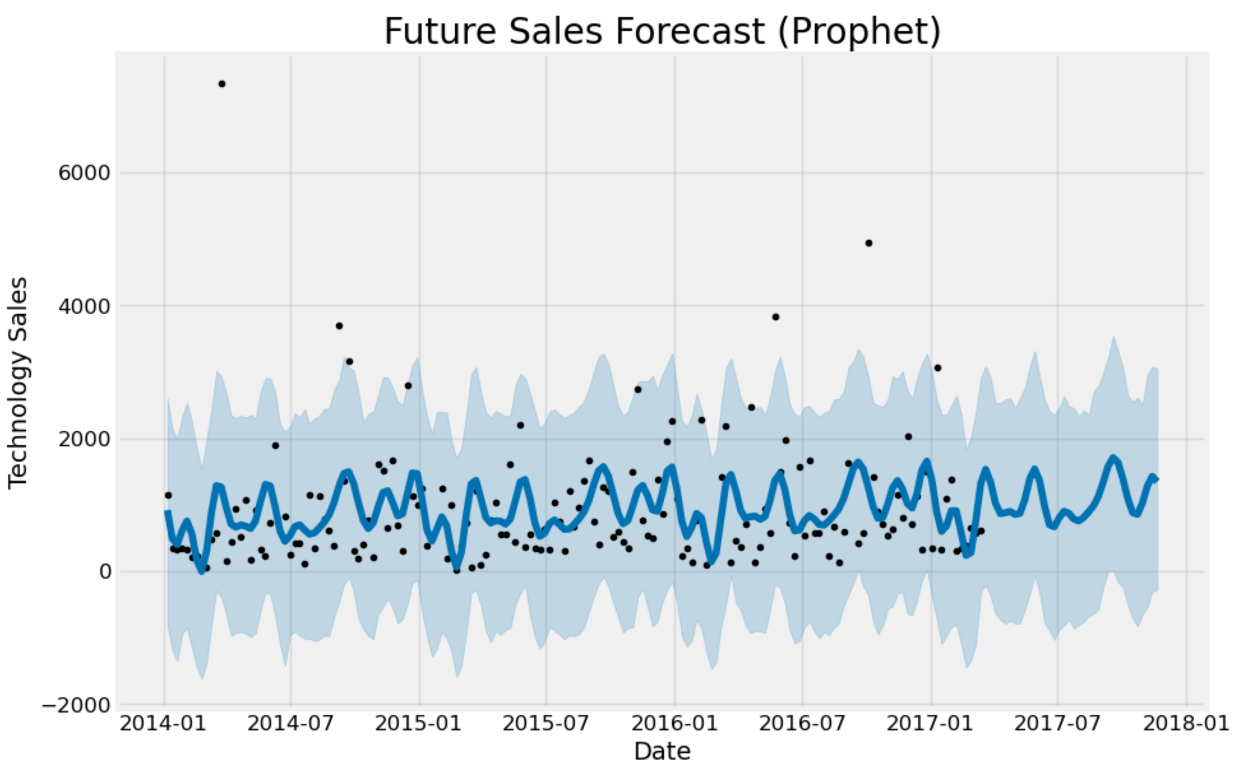

Observed vs Predicted Sales (Prophet)
Observed
Predicted
Technology Sales
Date
0
1000
2000
3000
4000
5000
6000
7000
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01

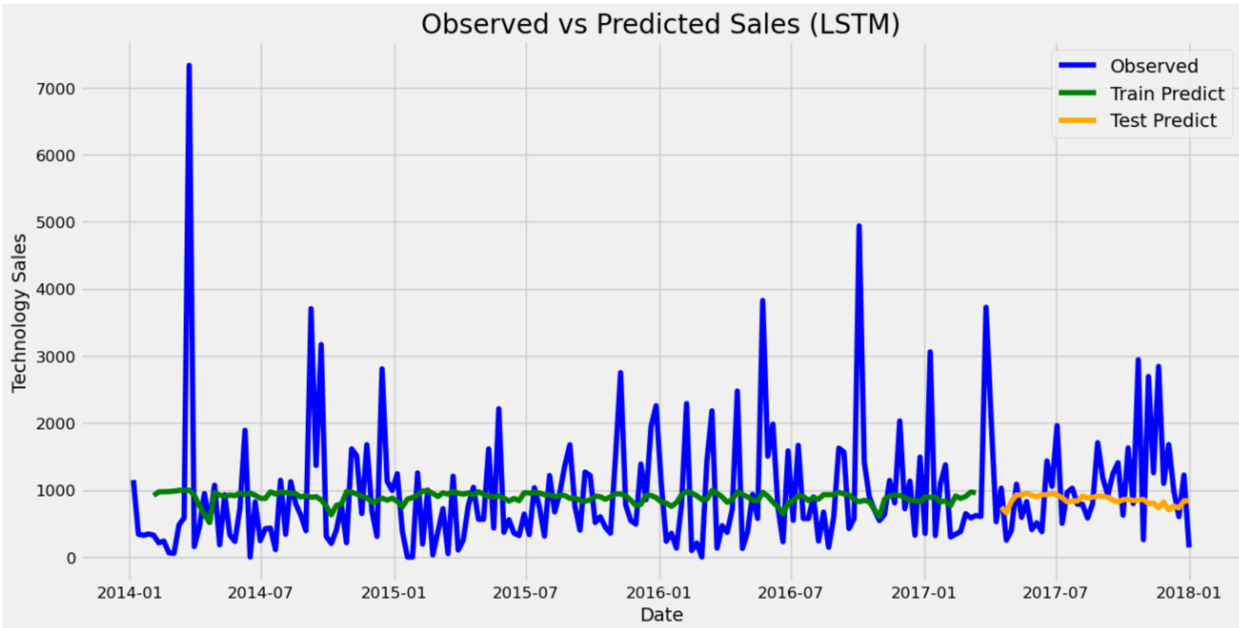

Future Sales Forecast (Prophet)
Technology Sales
Date
−2000
0
2000
4000
6000
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01

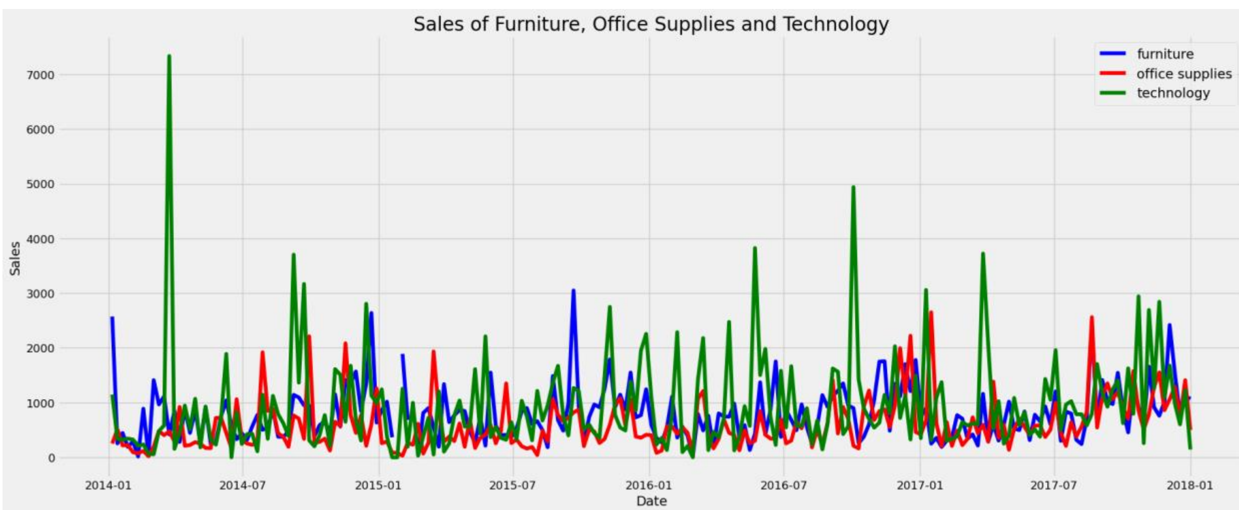

Observed vs Predicted Sales (LSTM)
Observed
Train Predict
Test Predict
Technology Sales
7000
6000
5000
4000
3000
2000
1000
0
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Date

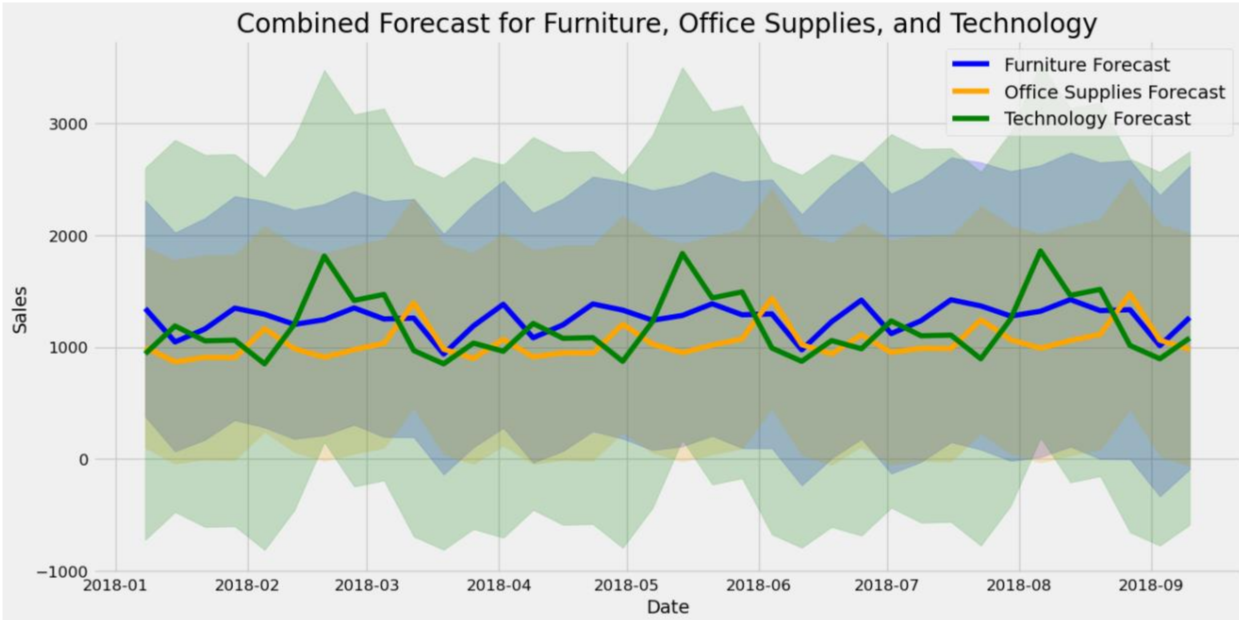

Future Sales Forecast (LSTM)
Observed
Future Forecast
Technology Sales
7000
6000
5000
4000
3000
2000
1000
0
2014
2015
2016
2017
2018
Date

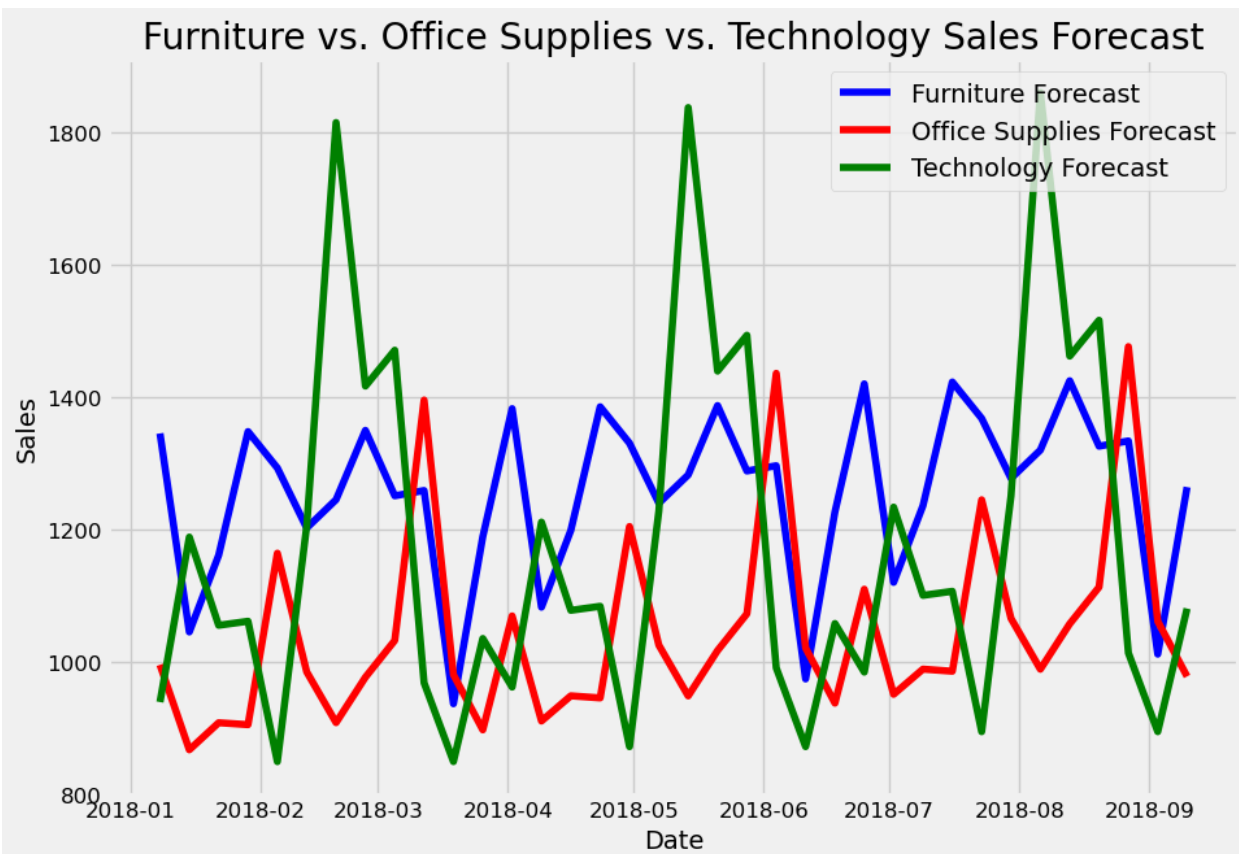

Observed vs Predicted Sales (Prophet)
Observed
Predicted
Technology Sales
Date
0
1000
2000
3000
4000
5000
6000
7000
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01


Future Sales Forecast (Prophet)
Technology Sales
Date
−2000
0
2000
4000
6000
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01


Observed vs Predicted Sales (LSTM)
Observed
Train Predict
Test Predict
Technology Sales
7000
6000
5000
4000
3000
2000
1000
0
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Date


Sales of Furniture, Office Supplies and Technology
furniture
office supplies
technology
Sales
7000
6000
5000
4000
3000
2000
1000
0
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
Date


Combined Forecast for Furniture, Office Supplies, and Technology
Furniture Forecast
Office Supplies Forecast
Technology Forecast
Sales
Date
3000
2000
1000
0
−1000
2018-01
2018-02
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08
2018-09


Furniture vs. Office Supplies vs. Technology Sales Forecast
Furniture Forecast
Office Supplies Forecast
Technology Forecast
Sales
Date
1800
1600
1400
1200
1000
800
2018-01
2018-02
2018-03
2018-04
2018-05
2018-06
2018-07
2018-08
2018-09