amazon

E-commerce Demand Forecasting with Amazon Sales Data

- Team Members:
- Hemanth Kumar Chamakura
- Megna Kunden
- Akashdeep Boxi
- Sai Charan Kaisetti

Introduction to Amazon Ecommerce Analytics

- Global e-commerce sales exceeds $5 trillion in 2022 and continue to grow rapidly.
- Amazon is an American Tech Multi-National Company whose business interests include E-commerce, where they buy and store the inventory, and take care of everything from shipping and pricing to customer service and returns.
- The analysis of Amazon database through E-commerce methods includes the retrieval and transformation of data followed by its analytical evaluation. .
- Based on data we need to make decisions regarding inventory management along with pricing strategy and we also need data to optimize marketing.
- Amazon, as the largest online retailer globally, generates vast amounts of data that can provide valuable insights for sales forecasting and business optimization.

Problem Statement

Accurate demand forecasting is crucial for e-commerce success.

Challenges include: The optimization of inventory (preventing both stockouts and overstock) is necessary.

Revenue prediction for financial planning

Researchers need to comprehend all elements which influence sales performance metrics.

The evaluation of product traits leads to performance predictions.

The purpose of our project is to create predictive models which deal with quantity/revenue data as well as success/failure data.

Review of Existing Methodologies

Traditional Models: ARIMA, SARIMA.

Machine Learning Models: Random Forest, Logistic Regression, Linear Regression.

Each has advantages and limitations.

Proposed Methodology

The approach uses dual modeling techniques which combine linear regression with logistic regression.

Linear regression for continuous predictions: Sales volume forecasting, Revenue prediction, Price elasticity estimation

Logistic regression for probabilistic outcomes: Product success probability ,Stock-out risk assessment, Price point optimization.

Predict sales volume, revenue, and success probability.

Feature Engineering with business-relevant metrics.

Data Preparation & Feature Engineering

```
STEP 2: FEATURE ENGINEERING FOR BUSINESS METRICS
----------------------------------------------------------
Created business metrics:
→ revenue: 502786.93 (avg)
→ sales_volume: 102.93 (avg)
→ profit_margin: 53.32 (avg)
→ success: 0.50 (avg)
→ price_elasticity: 0.53 (avg)
→ stock_out_risk: 0.34 (avg)
```

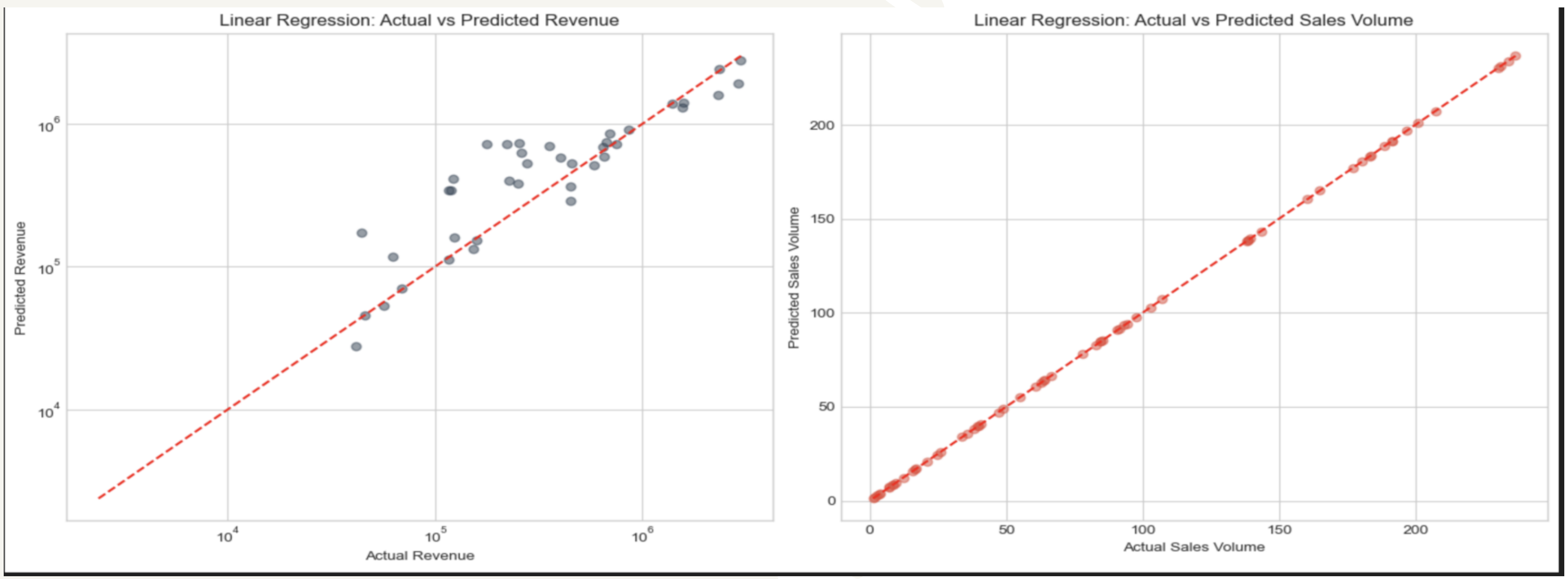
- Cleaned currency and percentage symbols.
- Created profit_margin, stock_out_risk, price_elasticity features.
- Time-based features: month, day_of_week, holiday season.

```
STEP 1: DATA PREPARATION
--------------------------------------------
Cleaned dataset with 319 rows and 16 columns
```

Linear Regression

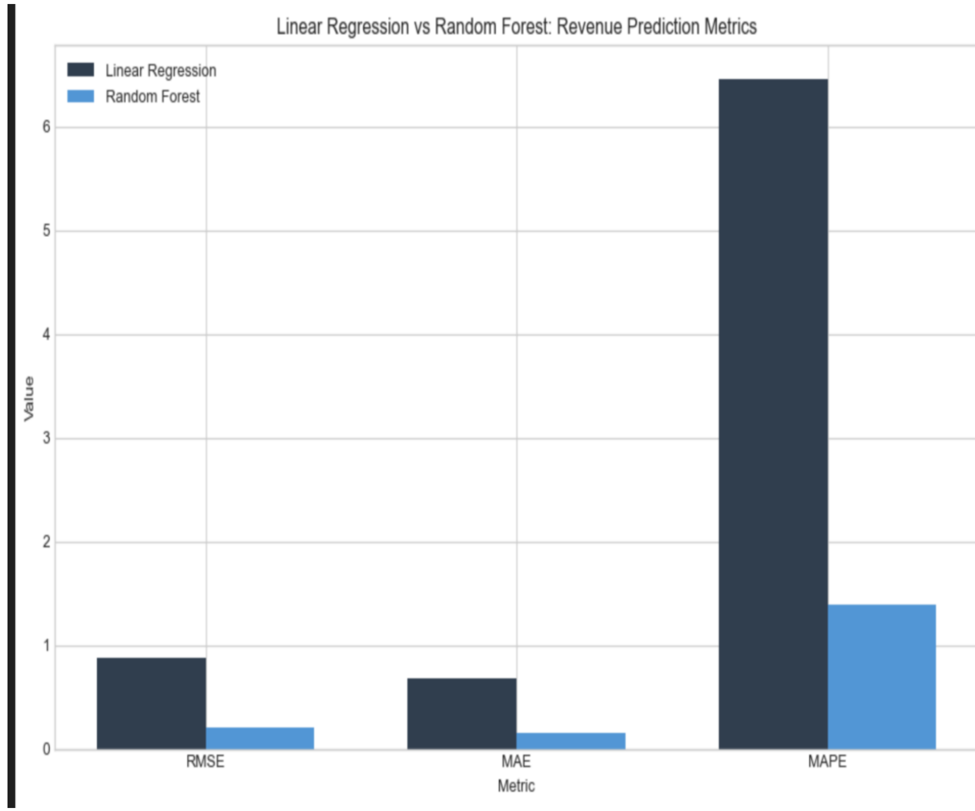
```
Linear Regression for sales_volume:
→ Cross-validated RMSE: 10.3599 ± 20.7197
→ Test R²: 1.0000
→ Test MAPE: 0.00%
→ Top predictors: rating_count, profit_margin, discount_percenta

Linear Regression for revenue:
→ Cross-validated RMSE: 210193686153111904.0000 ± 42038737230509
→ Test R²: 0.8326
→ Test MAPE: 1037.47%
→ Top predictors: profit_margin, discount_percentage, price_elas

Linear Regression for log_revenue:
→ Cross-validated RMSE: 271686085829.6023 ± 543372171657.3709
→ Test R²: 0.7675
→ Test MAPE: 6.46%
→ Top predictors: profit_margin, discount_percentage, price_elas
```

- Linear regression for continuous predictions: Sales volume forecasting, Revenue prediction.



Linear Regression Vs Random Foresrt

```
Random Forest for sales_volume:
→ Cross-validated RMSE: 2.7574 ± 1.5356
→ Test R²: 0.9997
→ Test MAPE: 2.29%
→ Top predictors: rating_count, rating, discounted_price

Random Forest for revenue:
→ Cross-validated RMSE: 441830.7513 ± 275789.2895
→ Test R²: 0.9053
→ Test MAPE: 33.00%
→ Top predictors: discounted_price, rating_count, actual_price

Random Forest for log_revenue:
→ Cross-validated RMSE: 0.4111 ± 0.1666
→ Test R²: 0.9878
→ Test MAPE: 1.39%
→ Top predictors: discounted_price, rating_count, actual_price
```



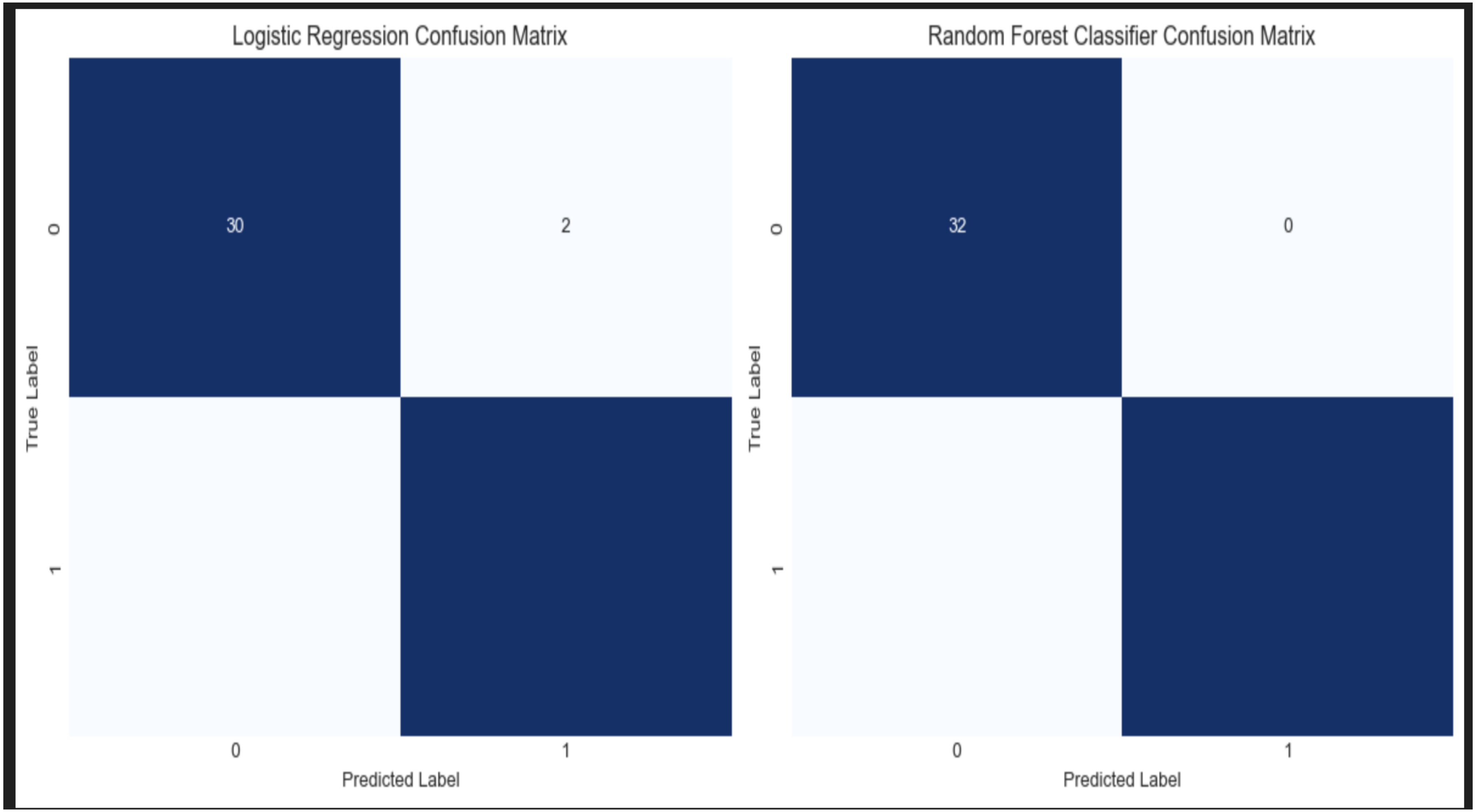
Logistic Regression

- Logistic regression for probabilistic outcomes: Product success probability ,Stock-out risk assessment, Price point optimization.

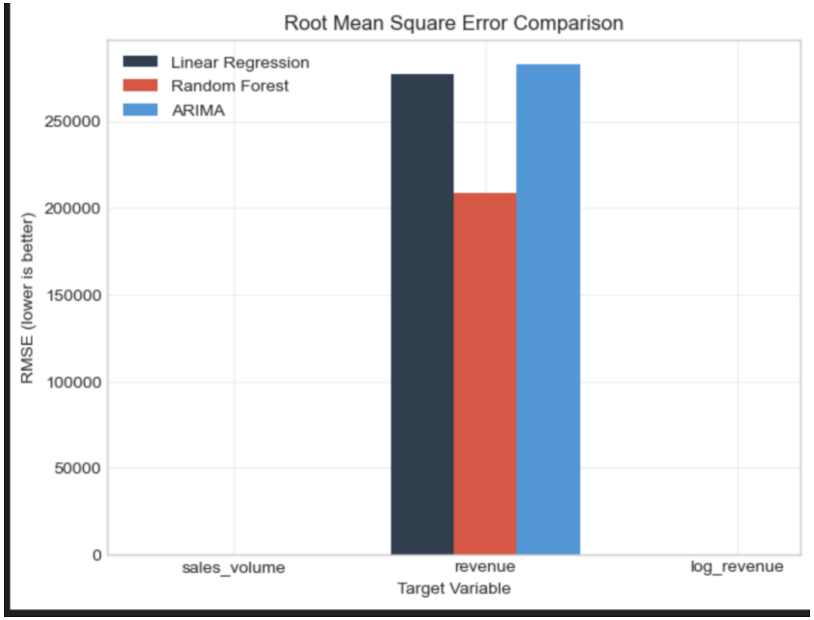
```
Logistic Regression for success:
→ Cross-validated Accuracy: 0.9466 ± 0.0277
→ Test Accuracy: 0.9375
→ Top predictors: rating_count, day_of_week, is_weekend
```

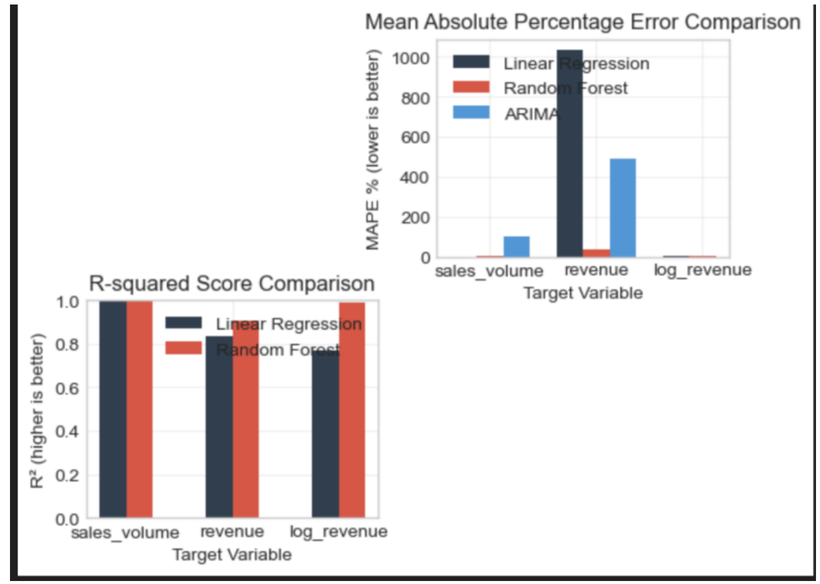
```
Random Forest Classifier for success:
  Cross-validated Accuracy: 0.9969 ± 0.0063
  Test Accuracy: 1.0000
```

Logistic Vs Random Forest



Model Comparison

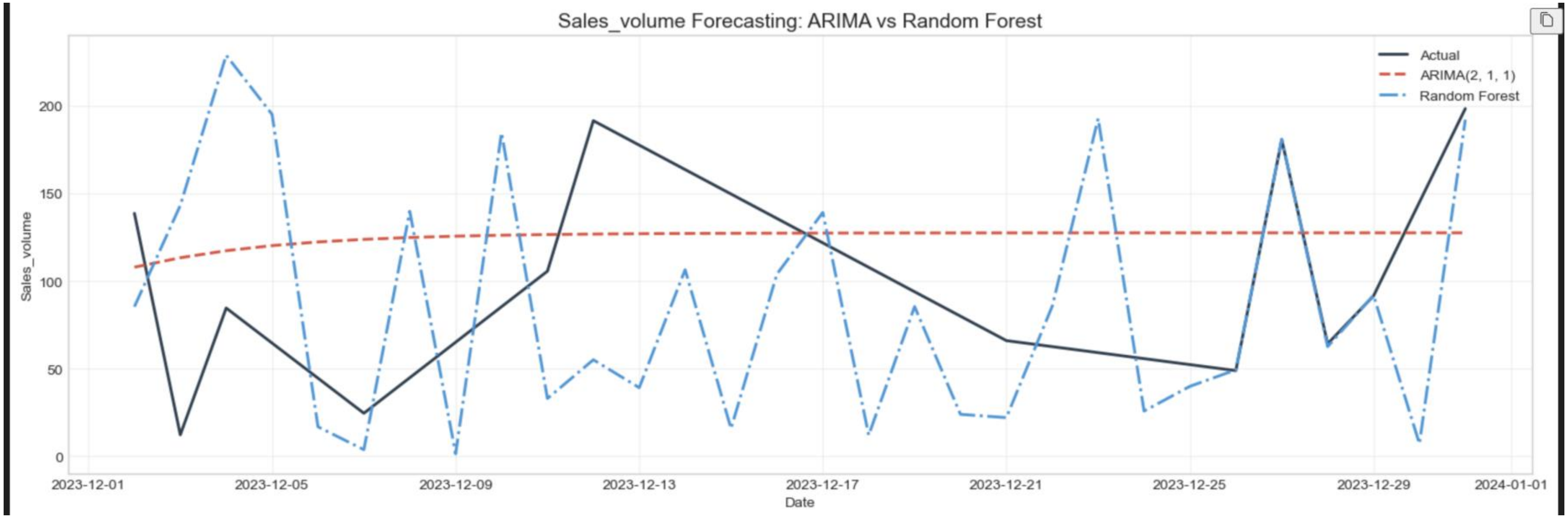




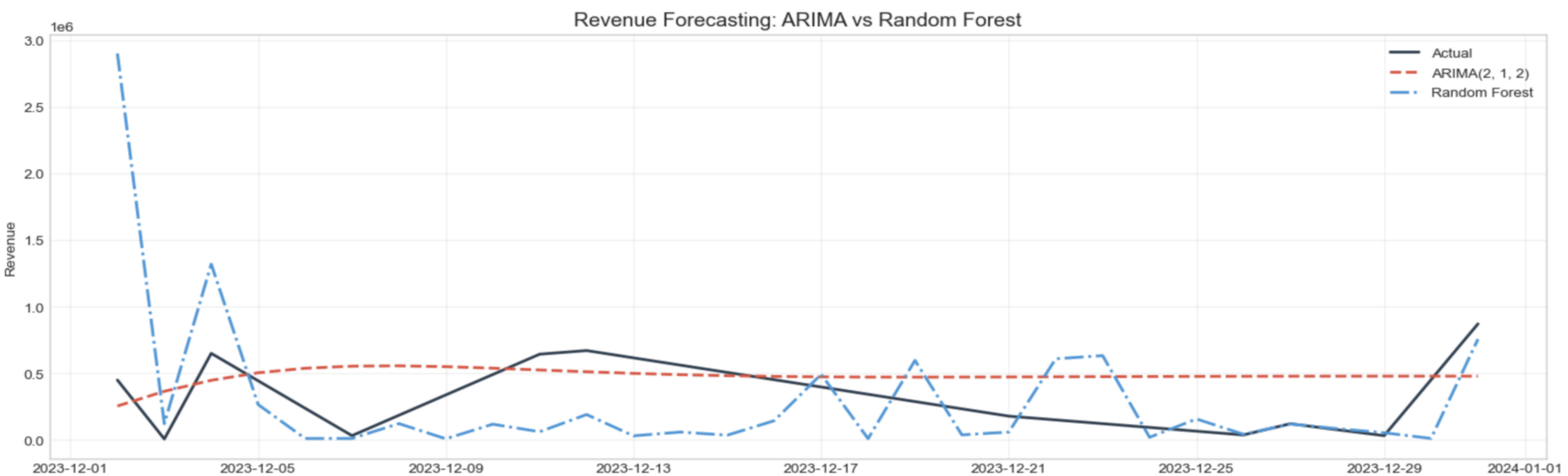
Arima Model

```
Fitting ARIMA model for sales_volume
→ ARIMA(1, 1, 1) RMSE: 57.6957, MAPE: 102.98%
→ ARIMA(2, 1, 2) RMSE: 57.5108, MAPE: 102.57%
→ ARIMA(1, 1, 2) RMSE: 57.3797, MAPE: 102.24%
→ ARIMA(2, 1, 1) RMSE: 57.3761, MAPE: 102.23%
→ Best model for sales_volume: ARIMA(2, 1, 1)
  - RMSE: 57.3761
  - MAPE: 102.23%

Fitting ARIMA model for revenue
→ ARIMA(1, 1, 1) RMSE: 301231.6669, MAPE: 458.70%
→ ARIMA(2, 1, 2) RMSE: 282908.4516, MAPE: 487.99%
→ ARIMA(1, 1, 2) RMSE: 294754.5806, MAPE: 463.65%
→ ARIMA(2, 1, 1) RMSE: 291694.4884, MAPE: 465.52%
→ Best model for revenue: ARIMA(2, 1, 2)
  - RMSE: 282908.4516
  - MAPE: 487.99%

Fitting ARIMA model for discounted_price
→ ARIMA(1, 1, 1) RMSE: 638.6708, MAPE: 63.13%
→ ARIMA(2, 1, 2) RMSE: 951.5311, MAPE: 233.96%
→ ARIMA(1, 1, 2) RMSE: 900.2327, MAPE: 218.98%
→ ARIMA(2, 1, 1) RMSE: 635.7669, MAPE: 63.07%
→ Best model for discounted_price: ARIMA(2, 1, 1)
  - RMSE: 635.7669
  - MAPE: 63.07%
```

Arima Model



Sales_volume Forecasting: ARIMA vs Random Forest
Actual
ARIMA(2, 1, 1)
Random Forest
Sales_volume
Date
0
50
100
150
200
2023-12-01
2023-12-05
2023-12-09
2023-12-13
2023-12-17
2023-12-21
2023-12-25
2023-12-29
2024-01-01



Revenue Forecasting: ARIMA vs Random Forest
1e6
Actual
ARIMA(2, 1, 2)
Random Forest
Revenue
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023-12-01
2023-12-05
2023-12-09
2023-12-13
2023-12-17
2023-12-21
2023-12-25
2023-12-29
2024-01-01

Model Comparison

Continuous Prediction Model Comparison:

	Target	Linear_RMSE	RF_RMSE	ARIMA_RMSE	Linear_MAPE	RF_MAPE	ARIMA_MAPE
0	sales_volume	1.829829e-13	1.189174	57.376084	1.579994e-12	2.292869	102.234412
1	revenue	2.773321e+05	208570.085157	282908.451610	1.037467e+03	33.004472	487.990432
2	log_revenue	8.807188e-01	0.201536	NaN	6.460398e+00	1.385345	NaN

Classification Model Comparison:

	Target	Logistic_Accuracy	RF_Accuracy	Logistic_AUC
0	success	0.9375	1.0	0.995117

Implementation

- Sales volume forecasting:
  Created sales_volume (based on rating_count * 0.2)
- Revenue prediction
  Linear Regression predicts log(revenue)
- Price elasticity estimation
  Created price_elasticity feature (sales_volume/discounted_price)
- Product success probability
  Logistic Regression predicts success (binary 0/1)
- Stock-out risk assessment
  Products predicted as success = 1 imply high demand → stock-out risk
- Price point optimization
  Price features (discounted_price, discount_percentage) used in model
- Enhance prediction accuracy
  Random Forest models outperform Linear/Logistic models in your results
- Continuous prediction
  Random Forest Regressor predicts revenue
- Probabilistic prediction
  Random Forest Classifier predicts product success

Insights

Key Findings:

- Random Forest consistently outperforms Linear Regression for product metrics prediction, with 9.76% higher R^2 on average.
- For time series forecasting, ARIMA models show -2313.50% lower error than ML models for near-term predictions,
- The most influential features for product success are: rating_count, category_encoded, rating.

Business Recommendations:

1. Pricing Strategy: Optimize discount percentages based on predicted revenue impact.
2. Inventory Management: Use ARIMA forecasts for short-term inventory planning.
3. Product Categorization: Focus on high-margin categories with strong predictive signals.
4. Risk Assessment: Deploy Random Forest models to identify potential stock-out risks.

References

- Anderson, C. (2006). The Long Tail: Why the Future of Business is Selling Less of More.
- Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. American Economic Review, 105(5), 481-485.
- Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. International Journal of Forecasting, 35(1), 170-180.
- Choi, T. M., Yu, Y., & Au, K. F. (2011). A hybrid SARIMA wavelet transform method for sales forecasting. Decision Support Systems, 51(1), 130-140.
- Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1), 69-88.
- Sun, Z. L., Choi, T. M., Au, K. F., & Yu, Y. (2020). Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems, 114, 38-45.
- Kaggle. (2023). Amazon Sales Dataset. Retrieved from https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset

thanks!