

```

In [2]: # =====
# AMAZON SALES PREDICTION FOR NEXT 6 MONTHS
# =====

import pandas as pd
import numpy as np
from lightgbm import LGBMRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import timedelta

# ----- LOAD FILES -----
category = pd.read_csv('category.csv')
inventory = pd.read_csv('inventory.csv')
customers = pd.read_csv('customers.csv')
order_items = pd.read_csv('order_items.csv')
orders = pd.read_csv('orders.csv')
payments = pd.read_csv('payments.csv')
products = pd.read_csv('products.csv')
sellers = pd.read_csv('sellers.csv')
shipping = pd.read_csv('shipping.csv')

# ----- DATA PREPARATION -----
print("Merging datasets...")

df = pd.merge(order_items, orders, on='order_id', how='left')
df = pd.merge(df, products, on='product_id', how='left')
df = pd.merge(df, category, on='category_id', how='left')
df = pd.merge(df, customers, left_on='customer_id', right_on='Customer ID', how='left')
df = pd.merge(df, sellers, on='seller_id', how='left')
df = pd.merge(df, payments, on='order_id', how='left')

# ----- FEATURE CREATION -----
df['total_sales'] = df['quantity'] * df['price_per_unit']
df['order_date'] = pd.to_datetime(df['order_date'])

# ----- TIME SERIES AGGREGATION -----

```

```
print("Aggregating monthly sales...")

monthly_sales = df.groupby(pd.Grouper(key='order_date', freq='M')).agg({
    'total_sales': 'sum',
    'quantity': 'sum',
    'order_id': 'nunique',
    'customer_id': 'nunique'
}).reset_index()

monthly_sales.columns = ['date', 'monthly_sales', 'total_quantity', 'total_orders', 'unique_customers']
monthly_sales = monthly_sales.sort_values('date')

# ----- FEATURE ENGINEERING -----
print("Creating time-based features...")

monthly_sales['year'] = monthly_sales['date'].dt.year
monthly_sales['month'] = monthly_sales['date'].dt.month
monthly_sales['quarter'] = monthly_sales['date'].dt.quarter

# Use fewer lags to prevent data loss
for lag in [1, 2, 3]:
    monthly_sales[f'sales_lag_{lag}'] = monthly_sales['monthly_sales'].shift(lag)
    monthly_sales[f'orders_lag_{lag}'] = monthly_sales['total_orders'].shift(lag)

# Rolling features
monthly_sales['sales_rolling_mean_3'] = monthly_sales['monthly_sales'].rolling(window=3).mean()
monthly_sales['sales_rolling_std_3'] = monthly_sales['monthly_sales'].rolling(window=3).std()

# Growth rates
monthly_sales['sales_growth_rate'] = monthly_sales['monthly_sales'].pct_change()
monthly_sales['orders_growth_rate'] = monthly_sales['total_orders'].pct_change()

# Seasonality flags
monthly_sales['is_holiday_season'] = monthly_sales['month'].isin([11, 12]).astype(int)
monthly_sales['is_year_start'] = monthly_sales['month'].isin([1, 2]).astype(int)

# Drop missing rows created by lag features
monthly_sales = monthly_sales.dropna().reset_index(drop=True)

print("✅ Data ready:", monthly_sales.shape)
```

```

# ----- PREPARE DATA FOR MODEL -----
features = [col for col in monthly_sales.columns if col not in ['date', 'monthly_sales']]
X = monthly_sales[features].select_dtypes(include=[np.number])
y = monthly_sales['monthly_sales']

split_index = int(len(monthly_sales) * 0.8)
X_train, X_test = X.iloc[:split_index], X.iloc[split_index:]
y_train, y_test = y.iloc[:split_index], y.iloc[split_index:]

print(f"Training samples: {X_train.shape[0]}, Testing samples: {X_test.shape[0]}")

# ----- MODEL TRAINING -----
print("\nTraining LightGBM model...")
ts_model = LGBMRegressor(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=6,
    random_state=42,
    subsample=0.8,
    colsample_bytree=0.8
)
ts_model.fit(X_train, y_train)

# ----- MODEL EVALUATION -----
y_pred = ts_model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("\nModel Performance:")
print(f"MAE: {mae:.2f}")
print(f"R²: {r2:.2f}")

# ----- FEATURE IMPORTANCE -----
print("\nGenerating feature importance plot...")

feature_imp = pd.DataFrame({
    'Feature': X.columns,
    'Importance': ts_model.feature_importances_
}).sort_values('Importance', ascending=False)

print("\nTop 10 Important Features:")

```

```

print(feature_imp.head(10))

plt.figure(figsize=(12, 6))
sns.barplot(data=feature_imp.head(10), x='Importance', y='Feature', color='skyblue')
plt.title('Top 10 Important Features for Sales Forecasting', fontsize=14, fontweight='bold')
plt.xlabel('Importance Score')
plt.ylabel('Feature')
plt.tight_layout()
plt.show()

# ----- FORECAST NEXT 6 MONTHS -----
def forecast_future_months(model, last_known_data, feature_cols, n_months=6):
    """
    Forecast next n months using recursive predictions.
    """
    predictions = []
    current_data = last_known_data.copy()

    for _ in range(n_months):
        next_pred = model.predict(current_data[feature_cols].values.reshape(1, -1))[0]
        predictions.append(next_pred)

        # Update lag features
        for lag in [3, 2, 1]:
            if f'sales_lag_{lag}' in feature_cols:
                current_data[f'sales_lag_{lag+1}'] = current_data[f'sales_lag_{lag}']
        current_data['sales_lag_1'] = next_pred

        # Update rolling mean
        if 'sales_rolling_mean_3' in feature_cols:
            current_data['sales_rolling_mean_3'] = (
                current_data['sales_lag_1'] + current_data['sales_lag_2'] + current_data['sales_lag_3']
            ) / 3

        # Update date-related fields
        current_month = current_data['month'].iloc[0]
        current_year = current_data['year'].iloc[0]

        next_month = current_month + 1
        next_year = current_year
        if next_month > 12:

```

```

        next_month = 1
        next_year += 1

    current_data['month'] = next_month
    current_data['year'] = next_year
    current_data['quarter'] = (next_month - 1) // 3 + 1
    current_data['is_holiday_season'] = 1 if next_month in [11, 12] else 0
    current_data['is_year_start'] = 1 if next_month in [1, 2] else 0

    return predictions

print("\nForecasting next 6 months...")
last_known_data = monthly_sales.iloc[-1:].copy()
future_sales = forecast_future_months(ts_model, last_known_data, X.columns, 6)

# ----- CREATE FUTURE DATES -----
last_date = monthly_sales['date'].iloc[-1]
future_dates = [last_date + timedelta(days=30*i) for i in range(1, 7)]

forecast_df = pd.DataFrame({
    'date': future_dates,
    'predicted_sales': future_sales
})

# ----- VISUALIZE FORECAST -----
plt.figure(figsize=(14, 8))
plt.plot(monthly_sales['date'], monthly_sales['monthly_sales'], label='Historical Sales', color='blue', marker='o')
plt.plot(forecast_df['date'], forecast_df['predicted_sales'], label='Forecasted Sales', color='red', linestyle='--', marker='s')
plt.fill_between(forecast_df['date'],
                 forecast_df['predicted_sales'] * 0.8,
                 forecast_df['predicted_sales'] * 1.2,
                 alpha=0.3, color='red', label='Confidence Interval')

plt.title('Amazon Sales Forecast - Next 6 Months', fontsize=16, fontweight='bold')
plt.xlabel('Date')
plt.ylabel('Sales Amount')
plt.legend()
plt.grid(alpha=0.3)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

```

```
# ----- DISPLAY FORECAST SUMMARY -----
print("\nSALES FORECAST FOR NEXT 6 MONTHS")
forecast_summary = pd.DataFrame({
    'Month': [d.strftime('%B %Y') for d in future_dates],
    'Predicted Sales': [f"${val:,.2f}" for val in future_sales],
    'Growth Trend': ['↑' if i == 0 or future_sales[i] > future_sales[i-1] else '↓' for i in range(len(future_sales))]
})
print(forecast_summary.to_string(index=False))

# ----- SAVE FORECAST -----
forecast_df.to_csv('amazon_sales_forecast_next_6_months.csv', index=False)
print("\n✅ Forecast saved to 'amazon_sales_forecast_next_6_months.csv'")
```

Merging datasets...

Aggregating monthly sales...

Creating time-based features...

✅ Data ready: (52, 20)

Training samples: 41, Testing samples: 11

Training LightGBM model...

C:\Users\admin\AppData\Local\Temp\ipykernel_7860\4092122585.py:42: FutureWarning: 'M' is deprecated and will be removed in a future version, please use 'ME' instead.

```
monthly_sales = df.groupby(pd.Grouper(key='order_date', freq='M')).agg({
```

```
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.000034 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
```

```
[LightGBM] [Info] Total Bins 185
```

```
[LightGBM] [Info] Number of data points in the train set: 41, number of used features: 14
```

```
[LightGBM] [Info] Start training from score 260674.410442
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

[illegible]

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

[illegible]

file:///C:/Users/admin/Downloads/Amazon lightGBM.html

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

Model Performance:

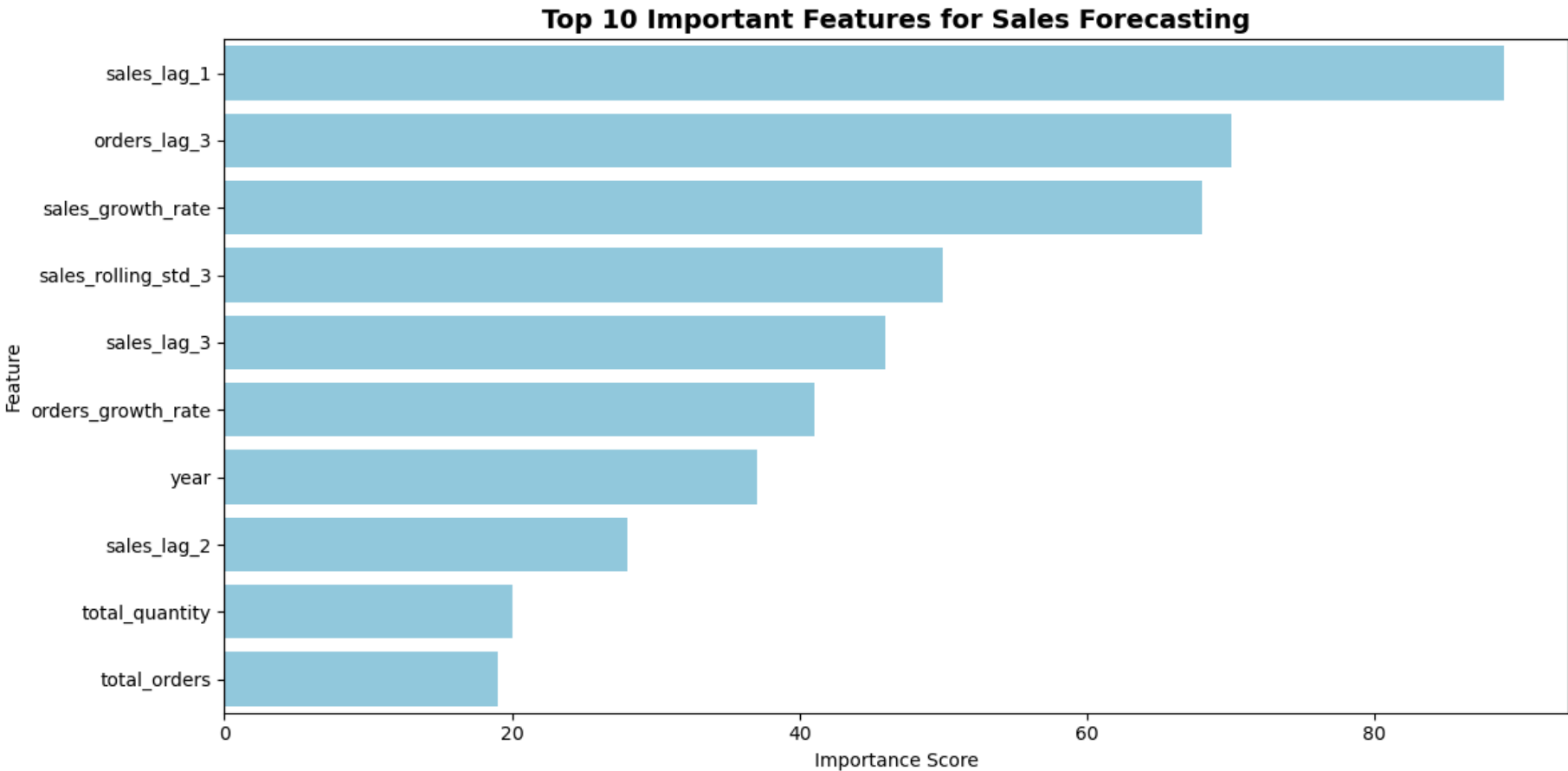
MAE: 100869.87

R²: -0.30

Generating feature importance plot...

Top 10 Important Features:

	Feature	Importance
6	sales_lag_1	89
11	orders_lag_3	70
14	sales_growth_rate	68
13	sales_rolling_std_3	50
10	sales_lag_3	46
15	orders_growth_rate	41
3	year	37
8	sales_lag_2	28
0	total_quantity	20
1	total_orders	19



Forecasting next 6 months...

```
C:\Users\admin\anaconda3\Lib\site-packages\sklearn\utils\validation.py:2739: UserWarning: X does not have valid feature names,
but LGBMRegressor was fitted with feature names
  warnings.warn(
C:\Users\admin\anaconda3\Lib\site-packages\sklearn\utils\validation.py:2739: UserWarning: X does not have valid feature names,
but LGBMRegressor was fitted with feature names
  warnings.warn(
C:\Users\admin\anaconda3\Lib\site-packages\sklearn\utils\validation.py:2739: UserWarning: X does not have valid feature names,
but LGBMRegressor was fitted with feature names
  warnings.warn(
C:\Users\admin\anaconda3\Lib\site-packages\sklearn\utils\validation.py:2739: UserWarning: X does not have valid feature names,
but LGBMRegressor was fitted with feature names
  warnings.warn(
C:\Users\admin\anaconda3\Lib\site-packages\sklearn\utils\validation.py:2739: UserWarning: X does not have valid feature names,
but LGBMRegressor was fitted with feature names
  warnings.warn(
C:\Users\admin\anaconda3\Lib\site-packages\sklearn\utils\validation.py:2739: UserWarning: X does not have valid feature names,
but LGBMRegressor was fitted with feature names
  warnings.warn(
```



SALES FORECAST FOR NEXT 6 MONTHS

Month	Predicted Sales	Growth Trend
August 2024	\$224,926.55	↑
September 2024	\$224,926.55	↓
October 2024	\$224,926.55	↓
November 2024	\$224,926.55	↓
December 2024	\$224,926.55	↓
January 2025	\$224,926.55	↓

✓ Forecast saved to 'amazon_sales_forecast_next_6_months.csv'

In []: