SALES FORECASTING FOR WALMART DATASET

BERKELEY AI ML - FINAL CAPSTONE PROJECT

BY LALITYA SAWANT

CONTENTS

EXECUTIVE SUMMARY

Objective

• Utilize AI/ML models to predict sales forecasts for Walmart.

Rationale

• Sales forecasting is crucial for revenue optimization and profit maximization.

• Here are some key reasons why sales forecasting is essential:

- Strategic Planning
- Financial Management
- Inventory Management
- Production Planning
- Marketing Strategy
- Customer Service
- Risk Management
- Performance Evaluation
- Investor Confidence
- Adaptation to Market Changes

BUSINESS BENIFITS

- Understanding sales trends enables organizations to strategically order the necessary quantities of goods across various departments and locations.

Leveraging AI/ML for Sales Forecasting

- Optimized Inventory Management
- Improved Supply Chain Efficiency
- Enhanced Financial Planning
- Maximized Revenue Generation
- Customer Satisfaction
- Data-Driven Decision Making
- Competitive Edge

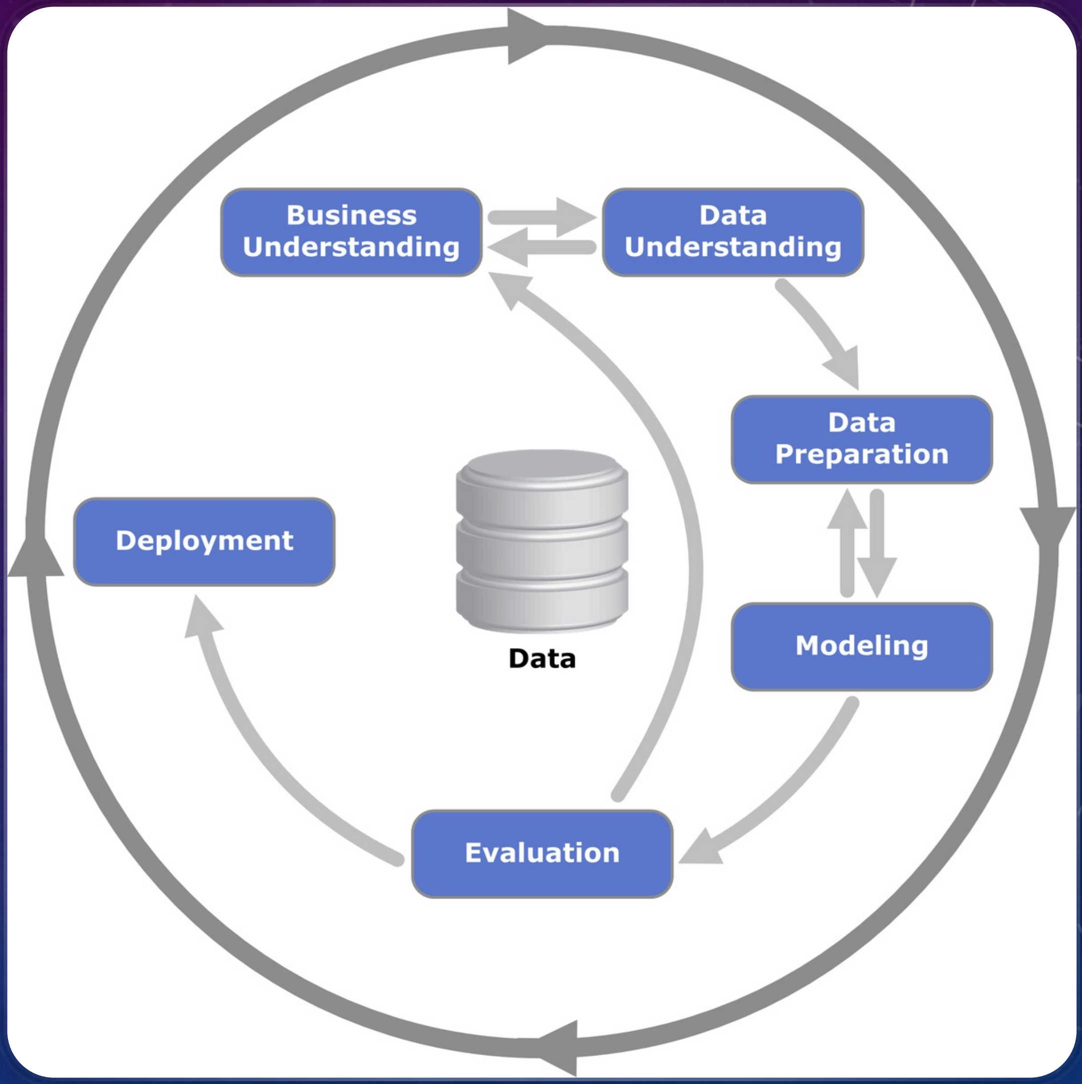
DATA SOURCE

- I picked up the Walmart sales dataset from Kaggle

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 421570 entries, 0 to 421569
Data columns (total 16 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Store         421570 non-null  int64
 1   Dept          421570 non-null  int64
 2   Date          421570 non-null  object
 3   Weekly_Sales  421570 non-null  float64
 4   IsHoliday     421570 non-null  bool
 5   Temperature   421570 non-null  float64
 6   Fuel_Price    421570 non-null  float64
 7   MarkDown1     150681 non-null  float64
 8   MarkDown2     111248 non-null  float64
 9   MarkDown3     137091 non-null  float64
 10  MarkDown4     134967 non-null  float64
 11  MarkDown5     151432 non-null  float64
 12  CPI           421570 non-null  float64
 13  Unemployment  421570 non-null  float64
 14  Type          421570 non-null  object
 15  Size          421570 non-null  int64
dtypes: bool(1), float64(10), int64(3), obje
memory usage: 51.9+ MB
```

METHODOLOGY

- I used the CRISP framework for this analysis and modeling
- Steps involved were as below:
  - Data Cleaning
  - Outlier Detection
  - Bias Assessment
  - Data Transformation
  - Data Distribution
  - Application of Algorithms



ANALYSIS KEY FINDINGS

Data Compilation:

The data was initially provided in 4 separate CSV files.

I merged the store, features, and train CSVs to create a comprehensive dataset.

Data Quality Enhancement:

Identified and addressed null values in markdown columns by removing those columns.

Ensured better data quality for subsequent analysis.

Sales Data Anomalies:

Detected and addressed rows with negative sales values, likely data anomalies.

Removed such instances, maintaining the integrity of the dataset.

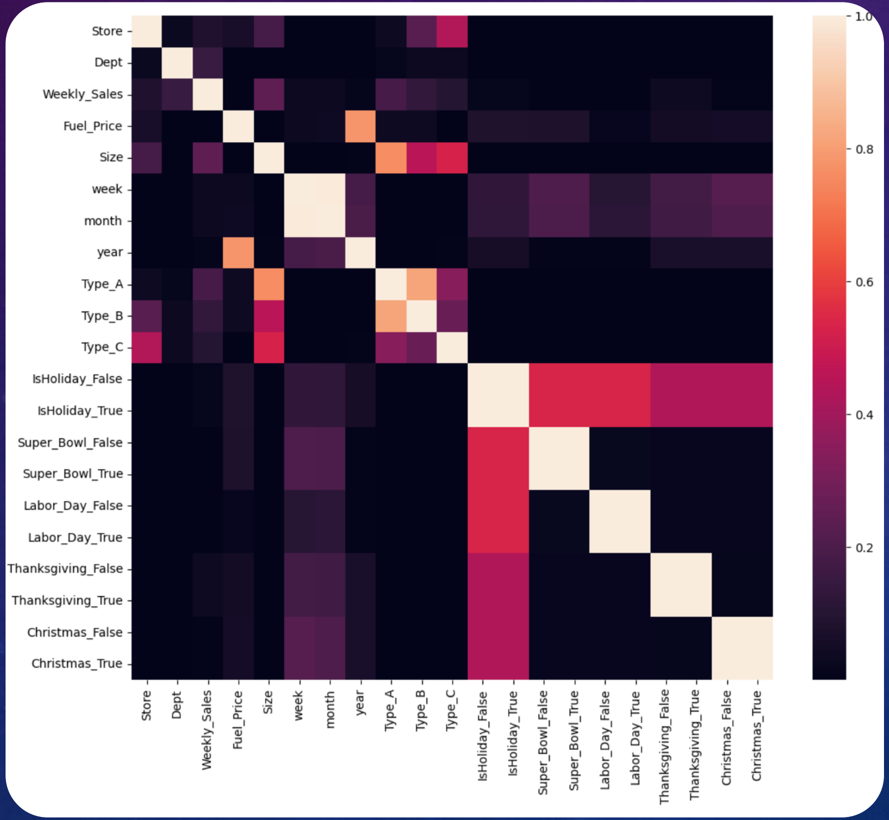
Key Attributes Impacting Sales:

Explored attributes like holidays, fuel price, unemployment, and temperature.

- **Holiday Analysis:**
  - Categorized holidays into four types: Labor Day, Super Bowl, Thanksgiving, and Christmas.
  - Thanksgiving showed a strong positive impact on sales, while Super Bowl had a moderate impact.
  - Labor Day and Christmas did not exhibit a significant positive impact on sales.
- **Other Sales Influencers:**
  - Explored factors beyond holidays, finding no clear positive or negative impact on sales.
- **Yearly Sales Trend:**
  - Observed a consistent pattern of increased sales at the end of each year.

FEATURE SELECTION

- Upon executing Ridge Regression for feature selection, I obtained the following correlation coefficient data



#	Features	Coefs
3	Size	6111.455355
1	Dept	3272.028832
9	Type_C	1379.949679
5	Month	1168.625192
2	Fuel_Price	701.886808
17	Thanksgiving_True	341.183575
18	Christmas_False	206.217420
14	Labor_Day_False	94.529108
13	Super_Bowl_True	80.195029
11	IsHoliday_True	62.867514
10	IsHoliday_False	-62.867514
12	Super_Bowl_False	-80.195029
15	Labor_Day_True	-94.529108
19	Christmas_True	-206.217420
16	Thanksgiving_False	-341.183575
7	Type_A	-410.515532
8	Type_B	-427.978958
4	Week	-430.756689
6	Year	-663.120183
0	Store	-1681.637899

TIME SERIES ANALYSIS - ARIMA

Time Series Analysis and Modeling

After performing time series decomposition and the augmented Dickey-Fuller test, we concluded that the data is nonstationary. Subsequent decomposition at weekly and monthly intervals revealed a repetitive pattern in the data.

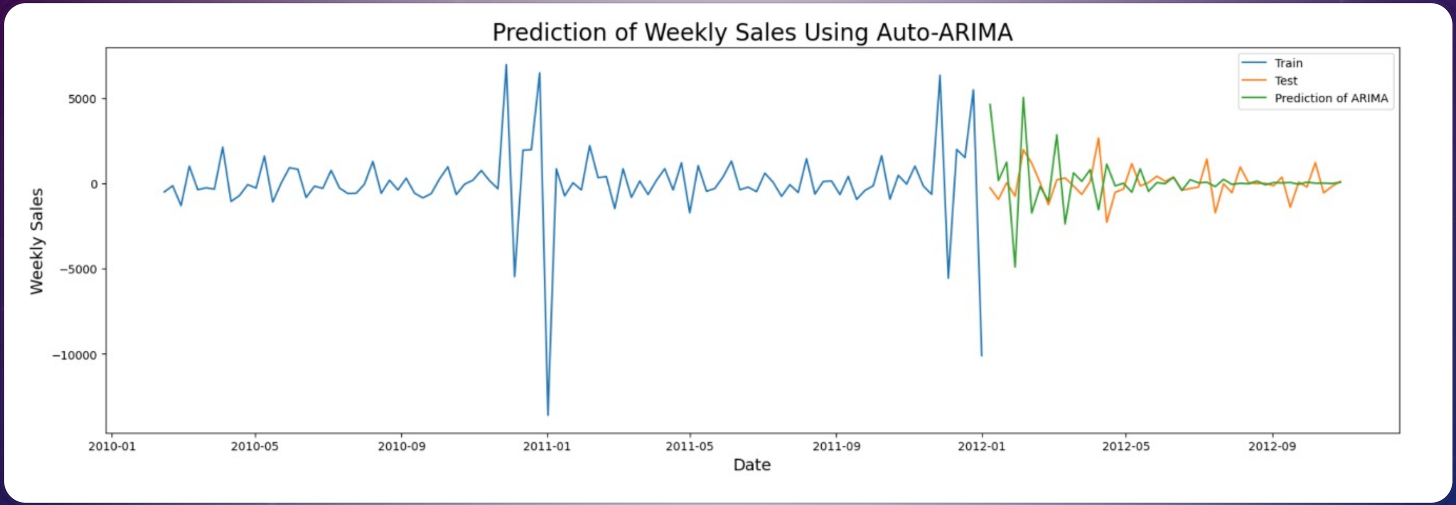
To address nonstationarity, I applied difference, shift, and log algorithms. The differential data emerged as the most effective in achieving stationarity.

For the final time series model, we utilized the auto_arima algorithm, identifying the following as the optimal model for predictions:

Best model: ARIMA(3,0,2)(0,0,0)[1] intercept

Total fit time: 10.236 seconds

AUTO-ARIMA PREDICTION



- **Next steps**

The predictions from the above model exhibit a slightly lower trend than the test data. Further tuning or exploring alternative algorithms may help achieve a closer alignment between the predictions and the test data.

MODEL EXPLORATION AND TUNING

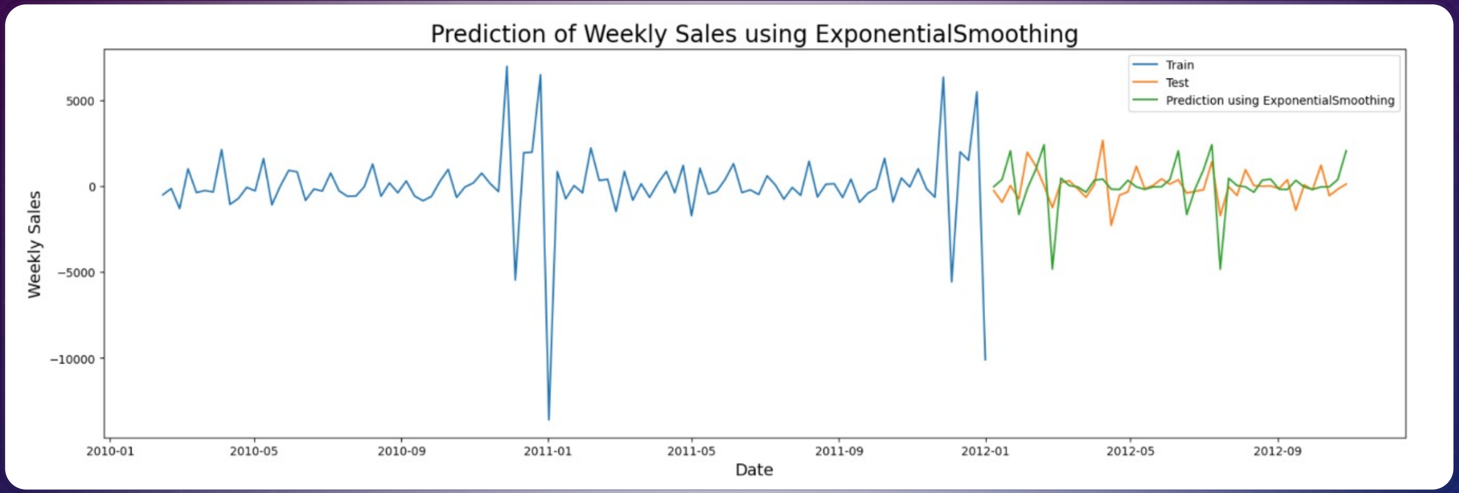
The previous iteration of the sales forecasting model has shown a slightly lower trend than the test data. To enhance the model's performance and achieve a closer alignment with the test data, I have undertaken further steps in model tuning and exploration of alternative algorithms.

New models created to forecast the data:

Exponential Smoothing

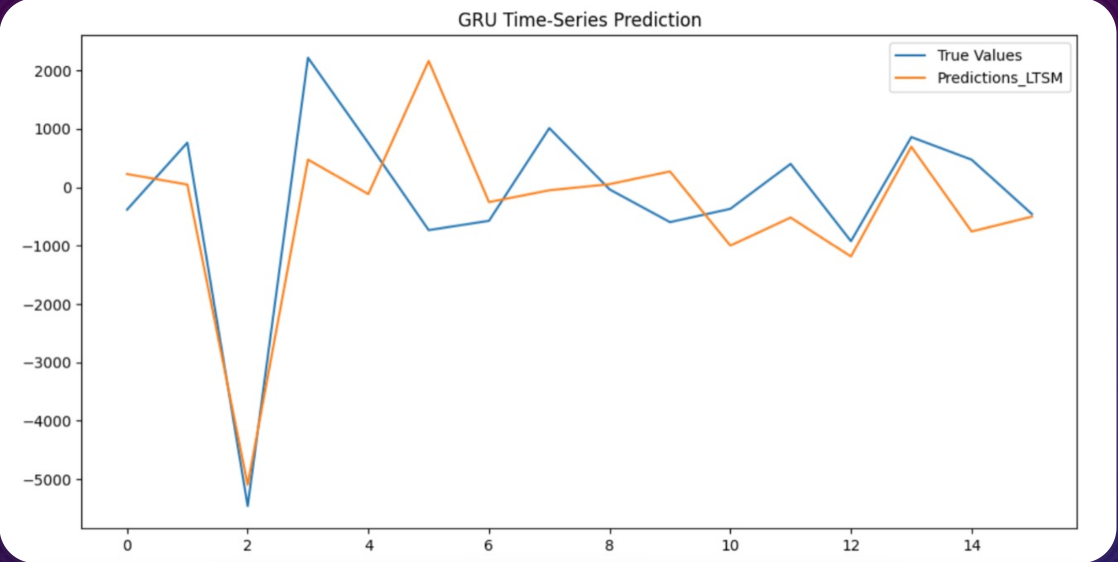
CNN LSTM and GRU Models (Required a data transformation step to fit a CNN model)

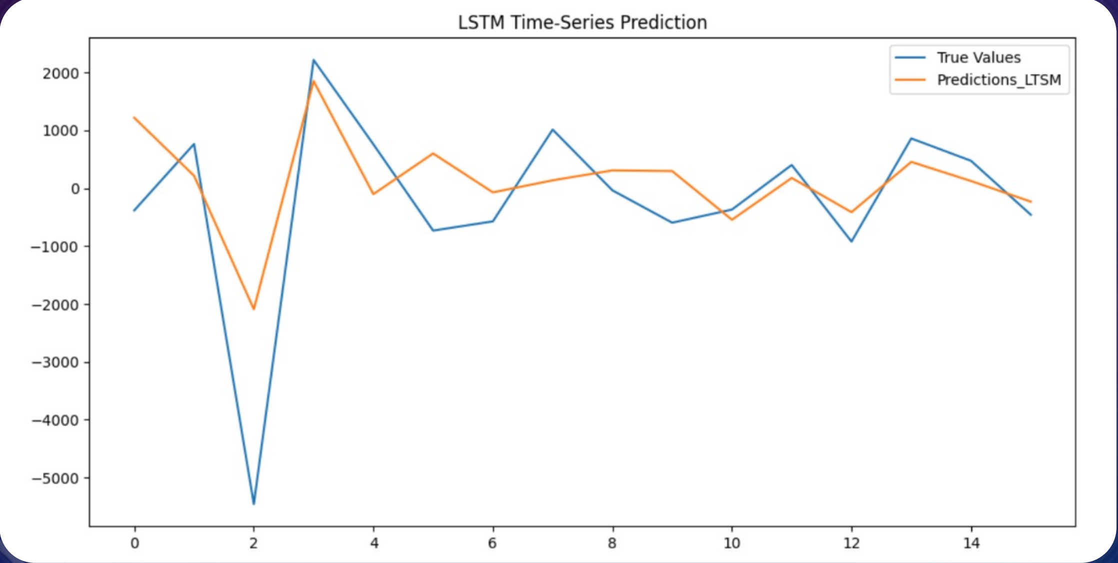
EXPONENTIAL SMOOTHING - PREDICTION



- The Exponential Smoothing model was applied to the dataset, revealing promising results in terms of prediction accuracy.
- This method leverages a weighted average of past observations, assigning exponentially decreasing weights to older data points.
- The adaptability of Exponential Smoothing makes it effective in capturing trends and seasonality in time-series data.

LSTM AND GRU - PREDICTION





- Two deep learning models, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), were implemented to further explore the dataset.
- Both LSTM and GRU exhibited notable improvements in prediction accuracy. Long Short-Term Memory Networks are known for their ability to capture long-term dependencies in sequential data, while Gated Recurrent Units, a more efficient variant of LSTM, also demonstrated competitive performance.

RECOMMENDATIONS / LESSONS LEARNED FOR CURRENT ITERATION:

FINE-TUNING PARAMETERS: ONGOING EFFORTS HAVE BEEN MADE TO FINE-TUNE THE PARAMETERS OF THE CURRENT MODELS, ESPECIALLY FOCUSING ON HYPERPARAMETER TUNING FOR LSTM AND GRU.

FEATURE ENGINEERING: SOME INITIAL ATTEMPTS AT FEATURE ENGINEERING HAVE BEEN MADE TO ENHANCE THE REPRESENTATION OF UNDERLYING PATTERNS IN THE DATA. ADDITIONAL FEATURES ARE BEING CONSIDERED FOR FUTURE ITERATIONS.

DATA AUGMENTATION: DATA AUGMENTATION TECHNIQUES HAVE BEEN EXPERIMENTED WITH TO ARTIFICIALLY EXPAND THE DATASET, PROVIDING THE MODELS WITH MORE DIVERSE EXAMPLES FOR IMPROVED GENERALIZATION.

EVALUATION METRICS: EVALUATION METRICS HAVE BEEN REASSESSED AND FINE-TUNED TO ENSURE ALIGNMENT WITH THE SPECIFIC GOALS OF THE PROJECT. ITERATIVE TESTING WITH VARIOUS COMBINATIONS OF FEATURES, ALGORITHMS, AND HYPERPARAMETERS IS ONGOING.

FURTHER STEPS AND CONCLUSION

- **Further Steps:**

Ensemble Modeling: Exploration of ensemble modeling can be conducted, combining the strengths of multiple models to enhance predictive performance.

- **In Conclusion:**

The current iteration has seen progress in the exploration, tuning, and experimentation of models. Exponential smoothing and LSTM models performed better than the previously used ARIMA model on the Walmart sales forecasting data.