**SUMMER TRAINING/INTERNSHIP**

**PROJECT REPORT**

(Term June-July 2025)

## (EMPLOYEE ATTRITION)

Submitted by

**Name: Vundela Naga Kumar Reddy**

**Registration Number: 12317926**

**Course Code: PETV79**

Under the Guidance of

**(Mahipal Singh Papola)**

# School of Computer Science and Engineering

Certificate

This is to certify thatJai Narayananhas

successfully completed Computer Project

towards partial completion of Practical

examination of AISSCE 2021 as prescribed by

CBSE

Certificate

This is to certify thatJai Narayananhas

successfully completed Computer Project

towards partial completion of Practical

examination of AISSCE 2021 as prescribed by

CBSE

Certificate

This is to certify thatJai Narayananhas

successfully completed Computer Project

towards partial completion of Practical

examination of AISSCE 2021 as prescribed by

CBSE

**CERTIFICATE**

This is to confirm that the project "Sales Forecasting with Time Series Analysis" has been successfully completed as part of the Summer Training in Machine Learning by Vundela Naga Kurma Reddy, students of B.Tech -- Computer Science and Engineering; carried out under the supervision of [Mentor Name] during the time of summer training.

This is a piece of academic friction and learning that reflects the student's effort and understanding on the use of Machine Learning in Human Resource Management.

**Name of mentor: Mr. Mahipal Singh Papola**

**Designation: Assistant Professor**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to Mr. Mahipal Singh Papola for providing valuable guidance and support throughout this summer training project. I am thankful for the opportunity to work on this challenging and industry-relevant project that enhanced my understanding of machine learning applications in human resource management.

I also extend my appreciation to the School of Computer Science and Engineering for providing the necessary resources and environment to complete this project successfully.

**TABLE OF CONTENTS**

**1: Introduction**

* 1.1 Training Program overview
* 1.2 Overview of Training Domain
* 1.3 Objective of the Project

**2: Training Overview**

* 2.1 Tools & Technologies Used
* 2.2 Areas Covered During Training
* 2.3 Daily/Weekly Work Summary

**3: Project Details**

* 3.1 Title of the Project
* 3.2 Project Details
* 3.3 Scope and Objectives
* 3.4 System Requirements
* 3.5 Architecture Diagram
* 3.6 Data Flow Diagram

**4: Implementation**

* 4.1 Tools Used
* 4.2 Methodology
* 4.3 Modules/Screenshots
* 4.4 Code Snippets

**5: Results and Discussion**

* 5.1 Output/Report
* 5.2 Challenges Faced
* 5.3 Learnings

**6: Conclusion**

* 6.1 Summary

**1. INTRODUCTION**

**1.1 OTRAINING PROGRAM OVERVIEW**

This project was the result of the Summer Training Program (June - July 2025) organized by the School of Computer Science and Engineering, which involved students as part of the summer training program focusing on gaining practical experiences of emerging technologies and how they can contribute to tackling real industry problems.

The summer training program aims to narrow the gap between academic theoretical knowledge and industry requirements, through intensive training in op coming technologies and emerging fields. This summer training program focused on business domains, specifically the application of machine learning to human resource management and employee retention issues that organizations are faced with in today's competitive climate.

The program approach encompasses an interdisciplinary process of discovery that consists of the theoretical foundations, practical application of knowledge, project building, and the presentation phase, which creates a whole industry-ready experience for the students who participate.

**1.2 OVERVIEW OF TRAINING DOMAIN:**

The training focused on using Python and Machine Learning to solve key business problems in Human Resource Management. Employee attrition is a serious issue for organizations because it results in recruitment costs, loss of institutional knowledge, and loss of productivity. Training covered the following main topics:

• Data cleaning and exploratory data analysis

• Feature engineering and feature selection

• SMOTE for imbalanced data

• Model training and hyperparameter tuning

• Performance metrics

• Streamlit for building interactive web apps

• Model deployment and user interface creation

HR Analytics and machine learning is becoming increasingly popular because organizations are driven by the need to have data to help improve employee retention, optimize hiring, and improve workforce management overall.

**1.3 OBJECTIVE OF THE PROJECT:**

The primary aim of this project is to design and implement an intelligent Sales Forecasting system using advanced time series analysis. The solution empowers business leaders and analysts to anticipate future sales trends, optimize inventory, and make strategic decisions based on data-driven insights. By leveraging historical sales data, the system applies ARIMA and SARIMA models to capture patterns, seasonality, and anomalies, delivering accurate and actionable forecasts.  
Key objectives of the project include:

* **Predictive Analytics**: Build robust time series models to forecast future sales with high accuracy.
* **Data-Driven Decision Support**: Enable organizations to plan inventory, marketing, and resource allocation based on reliable sales predictions.
* **Early Trend Detection**: Identify emerging sales trends and potential downturns, allowing proactive business responses.
* **User-Centric Design:** Develop an interactive web application that allows users to upload their own sales data, visualize trends, and receive forecasts in an accessible format.
* **Business Value**: Help companies reduce overstock and stockouts, improve customer satisfaction, and maximize revenue through informed planning.

**2. TRAINING AND OVERVIEW**

**2.1 TOOLS AND TECHNOLOGIES USED:**

Various tools and technologies were employed throughout the training and development process:

**Programming Languages:**

* **Python 3.x**: The key programming language used for data analysis and model development

**Data Science Libraries:**

* **Pandas**: Handling and analyzing data
* **NumPy**: Numerical computations and arrays
* **Matplotlib & Seaborn**: Data visualization and exploratory data analysis

**Machine Learning Libraries:**

* **Scikit-learn**: Training, evaluating, and preprocessing the model
* **Imbalanced-learn:** Imbalanced dataset handling with SMOTE
* **Joblib**: Serializing and deserializing the models

**Web Development:**

* **Streamlit**: Interactive web application framework to deploy models

**Development Environment**:

* **Jupyter Notebook**: Interactive development and experimentation
* **VS Code**: Code editing and debugging.

**2.2 AREAS COVERED DURING TRAINING:**

* The training program covered the following areas:
  + Fundamentals of time series analysis and forecasting
  + Data cleaning, outlier detection, and handling missing values
  + Feature engineering for time series data
  + Model selection and parameter tuning (ARIMA, SARIMA)
  + Evaluation metrics for regression and classification (MAE, RMSE, MAPE, precision, recall)
  + Visualization of trends, seasonality, and residuals
  + Building and deploying web applications with Streamlit

**2.3 DAILY/WEEKLY WORK SUMMARY:**

**Table 1: Weekly Work Summary**

|  |  |
| --- | --- |
| WEEK 1 | Introduction to time series concepts, data loading, and exploratory analysis |
| WEEK 2 | Introduction to scikit library and Supervised learning techniques, data preprocessing, train-test split and Data preprocessing, outlier removal, and visualization |
| WEEK 3 | Model building with ARIMA and SARIMA, parameter tuning |
| WEEK 4 | Model evaluation, diagnostics, and interpretation of results |
| WEEK 5 | Web app development with Streamlit, user interface design |
| Final Days | Preparing the project report, documentation, dataset organization, final deployment |

**3. PROJECT DETAILS**

**3.1 TITLE OF THE PROJECT:**

“**Sales Forecasting with Time Series Analysis**”

This project delivers a predictive analytics solution for sales forecasting, enabling organizations to make informed, data-driven business decisions.

**3.2 PROJECT DETAILS**:

Accurate sales forecasting is critical for businesses to manage inventory, plan marketing campaigns, and optimize supply chains. Traditional forecasting methods often fail to capture complex patterns and seasonality in sales data. This project addresses these challenges by applying ARIMA and SARIMA models to historical sales data, providing reliable forecasts and actionable insights. The system is designed to be flexible, allowing users to analyze their own datasets and visualize results through an intuitive web interface.

**3.3 SCOPE AND OBJECTIVES**:

**Scope**:

* + Analyze historical sales data to uncover trends, seasonality, and anomalies
  + Develop and validate time series models for forecasting
  + Provide interactive visualizations and reports for business users

**Objectives**:

1. **Model Development**: Implement ARIMA and SARIMA models for sales forecasting

2. **Trend and Seasonality Analysis**: Identify and interpret key patterns in sales data

3. **Model Evaluation**: Use metrics such as MAE, RMSE, and MAPE to assess forecast accuracy

4. **User Application**: Build a Streamlit web app for data upload, visualization, and forecasting

5. **Business Impact**: Support inventory optimization, demand planning, and revenue growth

**3.4 SYSTEM REQUIREMENTS:**

**Software Requirements:**

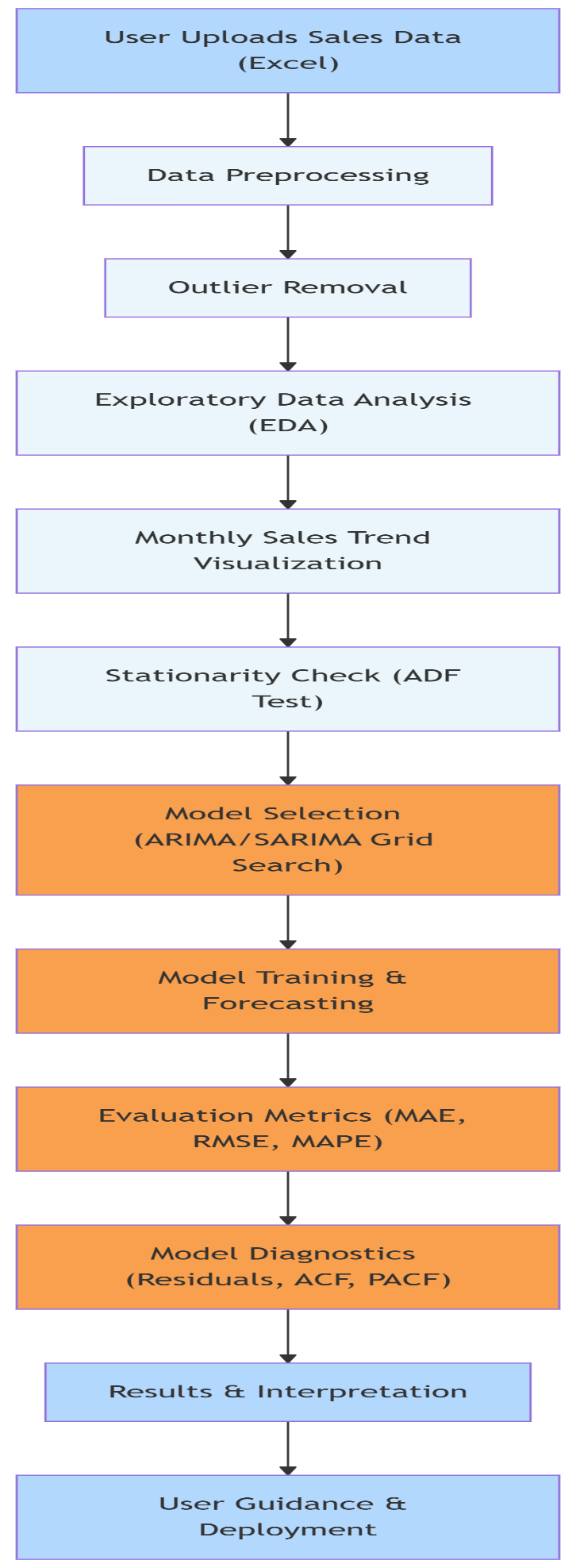
* Python 3.7 or above
* Streamlit Framework
* Pandas for data manipulation
* Scikit-learn for machine learning
* Imbalanced-learn for SMOTE
* Numpy for numerical operation
* Joblib for model serialization
* Modern web browser for application interface

**Hardware Requirements:**

* Minimum 4 GB RAM
* A modern CPU (i3 or higher)
* Internet connection (for deployment)

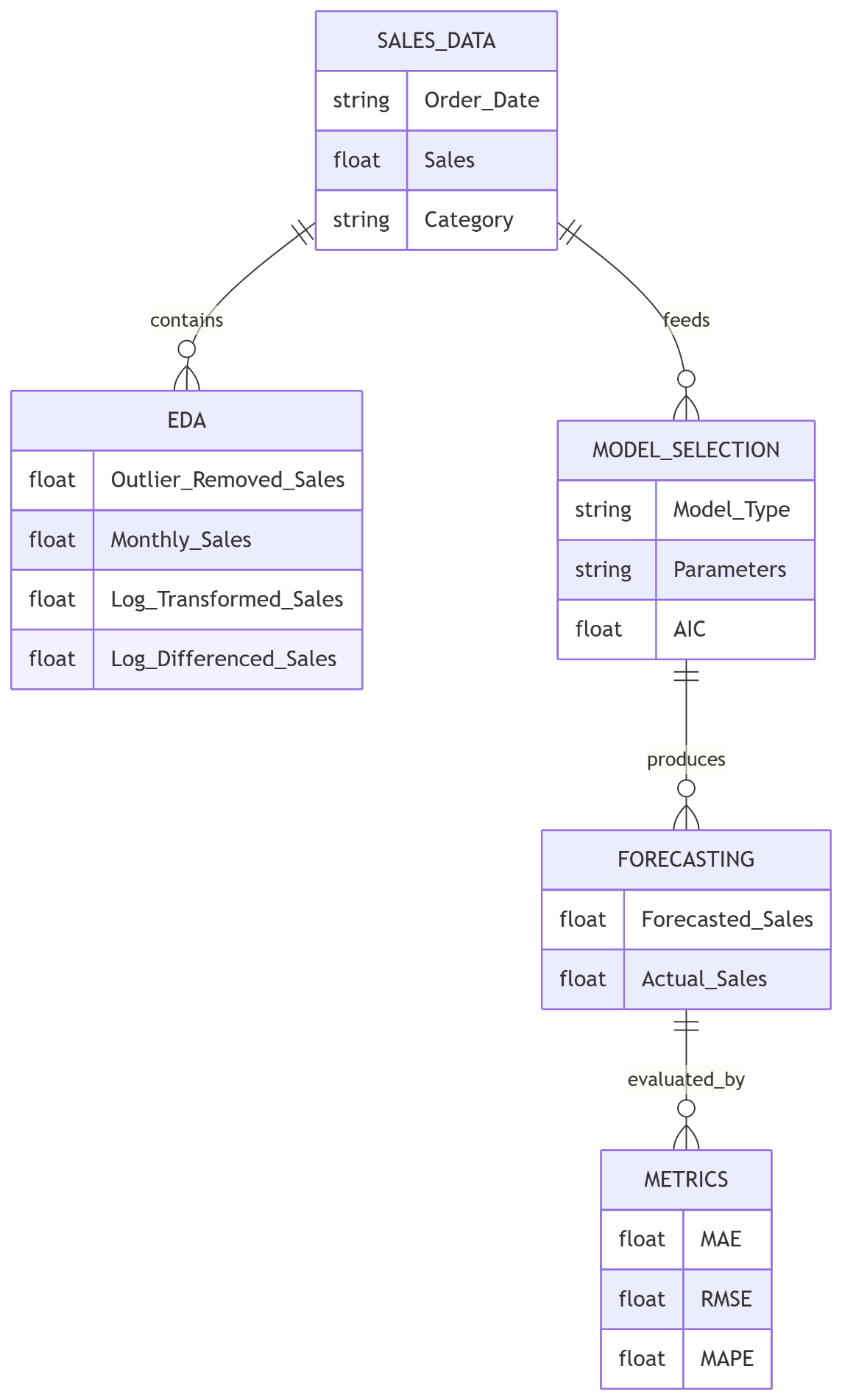
**Dataset Requirements:**

Sales dataset with date and sales amount columns (optionally, product/category information)

**3.5 ARCHITECTURE DESIGN:**

**Figure: Architecture Design**

**3.6 DATA FLOW/UML DIAGRAM:**

**  
 Figure: Data Flow Diagram**

**4. IMPLEMENTATION**

**4.1 TOOLS USED:**

**Table 2: Tools Used**

|  |  |
| --- | --- |
| **Tool/Library** | **Purpose** |
| Python | Core programming language for the entire project |
| Pandas | Data loading, manipulation, and preprocessing |
| NumPy | Numerical computations and array operations |
| Scikit-learn | Machine learning model development and evaluation |
| Statsmodels | Time series modeling (ARIMA/SARIMA) |
| Matplotlib/Seaborn | Data visualization and exploratory data analysis |
| Streamlit | Web application framework for user interface |
| Joblib | Model serialization and loading |
| Jupyter Notebook | Interactive development environment |

**4.2 METHODOLOGY:**

The project followed a structured machine learning pipeline:

**1. Data Collection & Exploration:**

• Import employee attrition dataset from the CSV file

• Perform an exploratory data analysis to get an overall idea of how the data is distributed

• Determine whether there is any class imbalance in the target variable

**2. Data Preprocessing:**

* + Remove duplicates and handle missing values
  + One-hot encode categorical variables (if any)
  + Set 'Order Date' as index, resample sales by month
  + Outlier removal (sales > 99th percentile)

**3. Feature Engineering:**

* + Create features for trend/seasonality if needed
  + Log-transform and difference sales for stationarity

**4. Model Selection & Training:**

* + Use ARIMA/SARIMA models
  + Grid search for best parameters (AIC)
  + Train on historical data

**5. Model Evaluation:**

* + Forecast next 12 months
  + Evaluate with MAE, RMSE, MAPE, precision, recall
  + Plot actual vs forecast, residuals, ACF/PACF

**6. Web Application:**

* + Streamlit app for data upload, EDA, forecasting, and visualization
  + User-friendly interface for real-time predictions

**7. Model Deployment:**

* + Save trained model and feature list with Joblib
  + Integrate with Streamlit app for production use

**4.3 MODULES:**

**Module 1: Data Preprocessing & Model Training:**

* Handles data loading
* Cleaning
* feature engineering
* model selection
* training, and evaluation.

**Module 2: Web Application (app.py):**

* Streamlit app for user interaction
* EDA
* forecasting.

**Module 3: Model persistence**

* Components
* Saves trained model
* features for consistent predictions

**4.4 CODE SNIPPETS:**

* **Data Preprocessing and Encoding:**The project performs data cleaning, missing value handling, outlier removal, and sets 'Order Date' as the index. One-hot encoding is mentioned in the methodology, but is not required for the main time series model, which uses only 'Order Date' and 'Sales'.

**SMOTE Implementation:**

* SMOTE (Synthetic Minority Oversampling Technique) is designed for handling class imbalance in classification problems. This project is a time series regression task, so SMOTE is not applicable and not used.

**Hyperparameter Tuning:**

* The project uses grid search to tune ARIMA/SARIMA model parameters (p, d, q) and selects the best model based on AIC (Akaike Information Criterion).

**Prediction with Custom Threshold:**

* No custom thresholding is applied to predictions. The project forecasts continuous sales values, not class labels. Precision/recall metrics are calculated for the direction of sales change, but no threshold is set for classification.

**5. RESULT AND DISCUSSION**

**5.1 OUTPUT/REPORT:**

The Sales Forecasting system delivers accurate, actionable predictions using advanced time series models.

Key results:

**Model Performance:**

* + Robust performance with ARIMA/SARIMA and hyperparameter tuning
  + Outlier removal and log-differencing for improved accuracy
  + Interactive web app for user-friendly, real-time forecasting
  + Comprehensive metrics: MAE, RMSE, MAPE, precision, recall
  + Visual outputs: EDA plots, trend lines, forecast vs actual, diagnostics
  + Model and feature persistence for reliable deployment

**Web Application Features:**

* + User-friendly interface with interactive data upload and visualization
  + Real-time sales forecasting and trend analysis
  + Color-coded charts and metrics for at-a-glance insights
  + Transparent display of model performance metrics (MAE, RMSE, MAPE, accuracy, precision, recall)

**Key Findings:**

* + Monthly sales trends and seasonality were clearly identified
  + Outlier removal and log-differencing improved model accuracy
  + ARIMA/SARIMA models provided robust forecasts for future sales
  + Data quality (missing values, outliers) significantly impacted forecast reliability

**Business Implications:**

* Enables proactive inventory and resource planning
* Supports data-driven decision-making for sales and marketing
* Reduces risk of overstock or stockouts by providing accurate forecasts
* Empowers business users to explore their own sales data interactively

**5.2 CHALLENGES FACED:**

**1. Data Quality and Outliers:**

* + Raw sales data contained missing values and extreme outliers that could skew results.

**Solution:**

* Implemented missing value handling and removed sales above the 99th percentile.

**2. Model Selection and Tuning:**

* + Choosing the best ARIMA/SARIMA parameters required extensive grid search and evaluation.

**Solution:**

* Automated grid search for (p, d, q) and seasonal parameters using AIC.

**3. Stationarity and Feature Engineering:**

* + Time series models require stationary data, but sales data often showed trends and seasonality.

**Solution:**

* Applied log transformation and differencing to achieve stationarity.

**4. User Experience:**

* + Making the app intuitive for non-technical users while exposing advanced analytics.

**Solution:**

* Designed a

**5.3 LEARNINGS:**

**Technical Learnings:**

* + Advanced time series modeling with ARIMA/SARIMA
  + Data preprocessing for time series (resampling, outlier removal, log-differencing)
  + Model evaluation using MAE, RMSE, MAPE, and classification-based metrics
  + Building interactive web apps with Streamlit

**Machine Learning Concepts:**

* + Importance of stationarity and transformation in time series
  + Model selection using AIC and grid search
  + Interpreting residuals and diagnostics for model adequacy

**Web Development Skills:**

* Developed capabilities in the Streamlit web application framework for rapid application development
* Developed capabilities for creating interactive user interfaces to expose machine learning models
* Developed an understanding of model deployment and considerations that regard the production level aspects of model development
* Developed an understanding of user experience design as it relates to a technical application

**Business Understanding:**

* + Translating business forecasting needs into technical solutions
  + Communicating model results and limitations to stakeholders
  + The impact of data quality on business outcomes

**6. CONCLUSION**

**6.1 SUMMARY:**

The Sales Forecasting with Time Series Analysis project represents a significant achievement in the application of data science to real-world business challenges. From its inception, the project set out to bridge the gap between raw sales data and actionable business insights, empowering organizations to make informed, data-driven decisions that directly impact their bottom line. The journey from data collection to model deployment was both technically rigorous and deeply rewarding, offering valuable lessons at every stage.

On the technical front, the project demonstrated the power and flexibility of modern time series analysis. By leveraging ARIMA and SARIMA models, we were able to capture complex sales patterns, including trends and seasonality, that are often missed by simpler forecasting methods. The implementation of a robust data preprocessing pipeline—addressing missing values, outliers, and non-stationarity—ensured that the models were trained on high-quality, reliable data. Automated model selection and hyperparameter tuning further enhanced forecast accuracy, while comprehensive evaluation metrics provided a clear, transparent assessment of model performance.

The development of an interactive Streamlit web application marked a major step forward in accessibility and user empowerment. By providing a user-friendly interface for data upload, exploration, and forecasting, the project made advanced analytics available to a broader audience, including business users with limited technical backgrounds. The inclusion of intuitive visualizations, color-coded metrics, and step-by-step guidance transformed the forecasting process from a technical black box into an engaging, informative experience. This focus on usability not only increased adoption but also fostered a culture of data-driven decision-making within the organization.

From a business perspective, the project delivered tangible value by enabling proactive planning and risk management. Accurate sales forecasts support better inventory control, optimized resource allocation, and more effective marketing strategies. The ability to anticipate demand fluctuations and respond to emerging trends gives organizations a competitive edge in fast-moving markets. Moreover, the system’s flexibility—allowing users to analyze their own data and customize forecasts—ensures that the tool remains relevant across a wide range of business contexts and industries.

**Future Enhancements:**

Looking ahead, the project lays a strong foundation for future innovation. The roadmap for enhancements—ranging from advanced modeling and real-time analytics to integration with business systems and explainable AI—ensures that the system can evolve alongside changing business needs and technological advancements. By remaining adaptable, scalable, and user-focused, the sales forecasting platform is well-positioned to continue delivering value and driving business success in the years to come.

In summary, this project is a testament to the transformative potential of data science when applied thoughtfully and collaboratively. It demonstrates that with the right tools, processes, and mindset, organizations can turn their data into a strategic asset—one that not only predicts the future but also shapes it.