A

PROJECT REPORT

ON

ADVERTISING SALES PREDICTION

A System-Based Data Science Project

By
AYESHA BANU

Under The Guidance Of
OASIS INFOBYTE
Data Science Internship Program

LinkedIn | GitHub | ayesha24banu@gmail.com

Completion Date: Oct, 2025

DECLARATION

I, Ayesha Banu, hereby declare that the project work entitled “Advertising Sales Prediction” has been carried out by me during my internship under Oasis Infobyte Internship Program. This project is a part of the academic and professional training to enhance my skills in Data Science and Machine Learning.

I further declare that:

- The work presented in this project is my original contribution, carried out under the guidance and mentorship provided during the internship.
- The dataset used in this project was sourced for research and learning purposes and has been processed, analyzed, and modeled by me.
- No part of this project has been copied or reproduced from any other work without proper acknowledgment.
- The project demonstrates my ability to apply data preprocessing, feature engineering, exploratory data analysis, machine learning modeling, and deployment techniques in a real-world scenario.

This project has been completed with sincere effort, dedication, and adherence to ethical practices in research and implementation.

I submit this project as part of my internship evaluation with Oasis Infobyte, and I am fully responsible for its authenticity and content.

Name: Ayesha Banu
Internship Program: Data Science Internship
Organization: Oasis Infobyte
Date: Oct, 2025

Signature:
Ayesha

ACKNOWLEDGMENT

I take this opportunity to express my deepest gratitude to Oasis Infobyte for offering me the opportunity to work on this project as part of the Data Science Internship Program. This internship has been an invaluable learning experience that allowed me to apply my theoretical knowledge of data science and machine learning to a real-world business problem.

I sincerely thank the entire Oasis Infobyte team for their continuous support, guidance, and encouragement throughout the duration of this project. Their structured internship program enabled me to strengthen my understanding of data preprocessing, exploratory data analysis, feature engineering, machine learning modeling, and deployment.

I would also like to thank my mentors and peers who provided constructive feedback and motivation, helping me to successfully complete this project.

Finally, I express my gratitude to my family and friends for their constant encouragement and moral support during the course of this project.

This project, *"Advertising Sales Prediction"*, has been a significant step in my journey to becoming a proficient Data Scientist, and I am grateful to Oasis Infobyte for giving me this platform to showcase my skills.

ABSTRACT

Advertising is one of the most significant investments for product-based companies, yet determining the effectiveness of each advertising channel remains a major challenge. This project, Advertising Sales Prediction, leverages machine learning and data science techniques to predict product sales based on advertising expenditures across Television, Radio, and Newspaper platforms.

The dataset used in this project consists of 200 records, each capturing advertising spend and corresponding product sales. After rigorous data preprocessing (handling duplicates, missing values, and irrelevant columns), feature engineering was applied to introduce interaction features such as *TV* × *Radio* and *TV* × *Newspaper*, along with a cumulative spend feature (*Total_Ads*). These enhancements capture the synergistic effects between channels and improve predictive performance.

Exploratory Data Analysis (EDA) revealed critical insights:

- TV advertising has the strongest positive influence on sales.
- Radio advertising shows a moderate impact, while Newspaper advertising contributes the least.
- Interaction features demonstrate that multi-channel campaigns outperform isolated spends.

Three regression models — Linear Regression, Random Forest, and Gradient Boosting — were trained and compared using RMSE and R^2 metrics. Among them, the Gradient Boosting Regressor emerged as the best-performing model with an R^2 of 0.989 and RMSE of 0.599, indicating high accuracy in predicting sales.

To ensure practical usability, the model was deployed using Streamlit, providing a simple, interactive web application where business users can input advertising budgets and receive real-time sales predictions. This deployment bridges the gap between technical development and business decision-making, making the solution accessible to non-technical stakeholders.

The project demonstrates the power of data-driven decision-making in optimizing advertising strategies. By accurately forecasting sales outcomes, companies can allocate budgets more effectively, maximize ROI, and reduce wasteful spending. Future enhancements may include integration of digital channels (social media, online ads), real-time data pipelines, and advanced ensemble models for even greater predictive power.

This work highlights how machine learning solutions can transform traditional marketing strategies into intelligent, optimized, and scalable systems, driving both efficiency and profitability

TABLE OF CONTENTS

CHAPTER - 1
Introduction

1. Introduction

1.1 Company Profile

The company is a consumer goods and retail organization that invests heavily in marketing campaigns to promote its products. Advertising is conducted across multiple channels, including Television (TV), Radio, and Newspapers, with significant budget allocation. Despite these investments, the company has faced challenges in optimizing advertising spend and maximizing sales returns.

To stay competitive in a data-driven marketplace, the company seeks to adopt advanced analytics and predictive modeling to understand how different channels contribute to sales. By leveraging machine learning, the organization can allocate its advertising budget more effectively, improve customer reach, and boost overall revenue.

The company emphasizes evidence-based decision making, where data science solutions play a critical role in designing and refining marketing strategies. This project is designed as a showcase of how data science can directly align with business growth objectives.

This project was completed as part of the Oasis Infobyte Internship Program, where the objective was to apply data science methodologies to solve real-world business challenges.

1.2 Problem Statement

In a competitive consumer market, companies spend millions of dollars on advertising without fully understanding the return on investment (ROI) from each channel. Traditional decision-making methods, such as intuition or isolated campaign analysis, often lead to inefficient budget allocation.

The company faces challenges such as:

- Identifying which advertising channel (TV, Radio, Newspaper) contributes most to sales.
- Determining the interaction effect of combining multiple channels.
- Predicting future sales given a new advertising budget allocation.
- Reducing waste in marketing expenditure while maintaining or increasing revenue.

Without an intelligent prediction system, the company risks overspending on low-impact channels and underutilizing high-impact ones. This leads to suboptimal sales performance and lost market opportunities.

1.3 Objective

The primary objective of this project is to develop a machine learning-based sales prediction model that can accurately forecast product sales based on advertising expenditure.

Specific objectives include:

- To analyze the impact of TV, Radio, and Newspaper spending on sales.
- To apply exploratory data analysis (EDA) to identify patterns, correlations, & outliers.
- To design feature engineering techniques (e.g., interaction terms like *TV × Radio*).
- To evaluate and compare multiple regression models (Linear Regression, Random Forest, Gradient Boosting).
- To select and deploy the best-performing model for real-time prediction through a Streamlit app.
- To provide actionable insights to marketing and business teams for decision-making.

1.4 Business Use Case

This project represents a real-world business case for organizations investing in advertising campaigns. The predictive model will help the company answer critical business questions, such as:

- *"If we increase TV spend by $10,000, how much will sales increase?"*
- *"Does a combination of TV and Radio advertising generate better results than TV and Newspaper?"*
- *"Which channel provides the highest ROI, and which can be minimized?"*

Use Case Benefits:

- **Marketing Optimization** → Identify the most effective channel and allocate resources strategically.
- **Cost Savings** → Reduce unnecessary expenditure on low-impact channels.
- **Revenue Growth** → Predict and maximize future sales outcomes.
- **Decision Support System** → Provide company executives with data-driven insights for campaign planning.

In essence, this project showcases how machine learning can transform advertising strategy into a scientifically optimized process, ensuring maximum efficiency and profitability for the company.

CHAPTER - 2

Dataset Description

2. Dataset Description

2.1 Dataset Overview

The dataset used in this project is the Advertising dataset, which contains details of marketing spend across three channels — TV, Radio, and Newspaper — and the corresponding product sales. This dataset provides the foundation for analyzing the relationship between advertising expenditure and sales outcomes.

It consists of 200 records (rows) and 4 primary attributes (columns). Each row represents one advertising campaign with its associated spending on different channels and the resulting sales.

2.2 Attributes / Features

Feature	Description
TV	Advertising expenditure on Television campaigns (in thousands of dollars).
Radio	Advertising expenditure on Radio campaigns (in thousands of dollars).
Newspaper	Advertising expenditure on Newspaper campaigns (in thousands of dollars)
Sales	Product sales generated from the advertising spend (target variable).

2.3 Feature Engineering

To improve the predictive capability of the models, additional features were created:

- **Total_Ads** → Sum of TV + Radio + Newspaper (total advertising budget).
- **TV_Radio** → Interaction term between TV and Radio spend.
- **TV_Newspaper** → Interaction term between TV and Newspaper spend.
- **Radio_Newspaper** → Interaction term between Radio and Newspaper spend.

These engineered features help capture **synergistic effects** between advertising channels.

2.4 Data Quality & Preprocessing

- **Missing Values** → None found in the dataset.
- **Duplicates** → Removed during preprocessing.
- **Unnecessary Columns** → Dropped (`Unnamed: 0` index column).
- **Scaling** → StandardScaler applied to normalize features.
- **Outliers** → Detected using IQR method; not removed but kept for model robustness.

CHAPTER - 3

Tools & Technologies

3. Tools & Technologies

3.1 Programming Language

- **Python 3.12** → Primary programming language used for all data science tasks.

3.2 Libraries & Frameworks

- **pandas, numpy** → Data manipulation and numerical computations.
- **matplotlib, seaborn** → Exploratory Data Analysis (visualizations, histograms, heatmaps, pairplots).
- **scikit-learn** → Machine learning models (Linear Regression, Random Forest, Gradient Boosting), feature scaling, model evaluation.
- **joblib** → Model and scaler persistence.
- **Streamlit** → Deployment of user-friendly web application for real-time prediction.

3.3 Tools Used

- **Jupyter Notebook** → For analysis, EDA, and stepwise model building.
- **VS Code / PyCharm** → For project development and modular code organization.
- **Git & GitHub** → Version control and repository hosting.
- **ReportLab** → Used to generate project documentation in PDF format.

CHAPTER - 4

Project Workflow

4. Project Workflow

The project follows a systematic end-to-end Data Science pipeline, ensuring accuracy, reproducibility, and business relevance. Each stage is carefully structured to deliver a deployable machine learning solution.

4.1 Workflow Stages

Step 1: Data Collection

- The dataset (`Advertising.csv`) was sourced as the primary data input.
- Data was ingested into the pipeline using Python (`pandas`) for further processing.

Step 2: Data Preprocessing

- Removal of duplicates and missing values.
- Dropped unnecessary index column (`Unnamed: 0`).
- Ensured data consistency and correctness.
- Saved processed dataset as `processed_Advertising.csv`.

Step 3: Exploratory Data Analysis (EDA)

- Generated statistical summaries and distribution plots.
- Created histograms, boxplots, and pairplots to study feature distributions.
- Built a correlation heatmap to analyze relationships between advertising channels and sales.
- Conducted outlier detection using the Interquartile Range (IQR) method.

Step 4: Feature Engineering

- Added new engineered features to capture interaction effects:
    - Total_Ads (aggregate spend)
    - TV_Radio, TV_Newspaper, Radio_Newspaper (interaction terms)
- Improved model robustness by incorporating synergistic effects.

Step 5: Model Training & Evaluation

- Data split into training (80%) and testing (20%) sets.
- Features scaled using StandardScaler.
- Trained and compared three regression models:
    - Linear Regression
    - Random Forest Regressor
    - Gradient Boosting Regressor

- Evaluation metrics:
    - Root Mean Squared Error (RMSE)
    - Coefficient of Determination (R^2 Score)

Step 6: Model Selection

- Gradient Boosting Regressor achieved the best performance:
    - RMSE: 0.599
    - R^2 Score: 0.989
- Selected as the final model for deployment.
- Best model and scaler saved as `best_model.pkl` in the `models/` directory.

Step 7: Deployment

- Built a Streamlit-based web application (`app.py`).
- Users input advertising spend for TV, Radio, Newspaper.
- The app automatically calculates interaction features and predicts sales.
- Results include:
    - Predicted sales value
    - Display of entered input data

4.2 Project Workflow

A simplified workflow of the project:

```
Raw Data → Preprocessing → EDA → Feature Engineering → Model Training → Model Evaluation → Deployment (Streamlit App)
```

CHAPTER - 5

Modeling & Results

5.Modeling & Results

The core objective of this project was to build predictive models that accurately estimate product sales based on advertising spend across various channels. A comparative analysis of different algorithms was conducted to identify the most suitable model for deployment.

5.1 Data Splitting & Scaling

- Dataset divided into training (80%) and testing (20%) sets.
- Features standardized using StandardScaler to ensure consistent scale across variables.
- Both original and engineered features were included in training.

5.2 Models Implemented

Three regression models were implemented and evaluated:

1. Linear Regression
    - Simple, interpretable baseline model.
    - Captures linear relationships between advertising channels and sales.
2. Random Forest Regressor
    - An ensemble method using multiple decision trees.
    - Captures nonlinear interactions and reduces variance through averaging.
3. Gradient Boosting Regressor
    - Boosting-based ensemble technique.
    - Sequentially improves performance by focusing on errors made by previous models.
    - Provides high predictive accuracy for complex datasets.

5.3 Evaluation Metrics

The following metrics were used for model comparison:

- Root Mean Squared Error (RMSE) → Measures prediction error magnitude.
- R^2 Score (Coefficient of Determination) → Measures how well features explain variability in sales.

5.4 Results

Model	RMSE	R^2 Score
Linear Regression	0.887	0.975
Random Forest Regressor	0.616	0.988
Gradient Boosting	0.599	0.989

5.5 Model Selection

- The Gradient Boosting Regressor was selected as the final model due to its superior accuracy (lowest RMSE, highest R^2).
- Both the trained model and the scaler were saved as `best_model.pkl` for deployment.

5.6 Key Insights

- TV advertising spend had the strongest positive impact on sales.
- Radio advertising showed moderate influence.
- Newspaper spend contributed the least, but interactions (e.g., TV × Newspaper) showed significant effects.
- Combining multiple channels yielded better performance than using a single channel alone.

CHAPTER - 6

Deployment

6. Deployment

The final stage of this project was to deploy the best-performing machine learning model into a user-friendly web application. Deployment ensures that business teams, marketing analysts, and executives can easily use the predictive model without requiring technical expertise.

6.1 Deployment Framework

- **Tool Used:** Streamlit
- **Reason for Selection:**
    - Lightweight and simple to set up.
    - Interactive, real-time prediction interface.
    - Ideal for showcasing machine learning projects in a professional setting.

6.2 Application Workflow

1. User opens the Streamlit web app (`app.py`).
2. Application displays input fields for advertising spends:
    - TV (in thousands of dollars)
    - Radio (in thousands of dollars)
    - Newspaper (in thousands of dollars)
3. The app automatically generates engineered features:
    - `Total_Ads`
    - `TV_Radio`
    - `TV_Newspaper`
    - `Radio_Newspaper`
4. Input data is scaled using the saved StandardScaler.
5. The Gradient Boosting Regressor model (`best_model.pkl`) predicts sales.
6. The app displays:
    - Predicted Sales Value
    - Entered Input Data (JSON format) for transparency.

6.3 Example Application Output

Input (User Entry):

```
TV: 100
Radio: 30
Newspaper: 20
```

Engineered Features (Auto-calculated):

```
Total_Ads: 150
TV_Radio: 3000
TV_Newspaper: 2000
Radio_Newspaper: 600
```

Predicted Sales:

```
Predicted Sales = 13.77
```

6.4 Application Interface

- **Title Page:** *Advertising Sales Prediction App*
- **Sections:**
  - Input fields (sliders/numeric input) for TV, Radio, Newspaper.
  - Prediction result with highlighted sales output.
  - Summary of entered data.
- **Footer:** Developed by *Ayesha Banu* | *Data Scientist*.

6.5 Benefits of Deployment

- **Business Readiness** → Non-technical users can directly interact with the predictive system.
- **Decision Support** → Provides real-time insights for campaign planning.
- **Scalability** → The app can be extended to include more features (e.g., Social Media spend).
- **Portability** → Deployed via Streamlit, making it shareable with stakeholders over web.

6.6 Streamlit web app:

Advertising Sales Prediction

Enter the advertising spend details below to predict **Sales**. The model uses historical data with engineered features for accurate prediction.

Enter Advertising Spend

TV Advertising Spend

100.00 − +

Radio Advertising Spend

30.00 − +

Newspaper Advertising Spend

20.00 − +

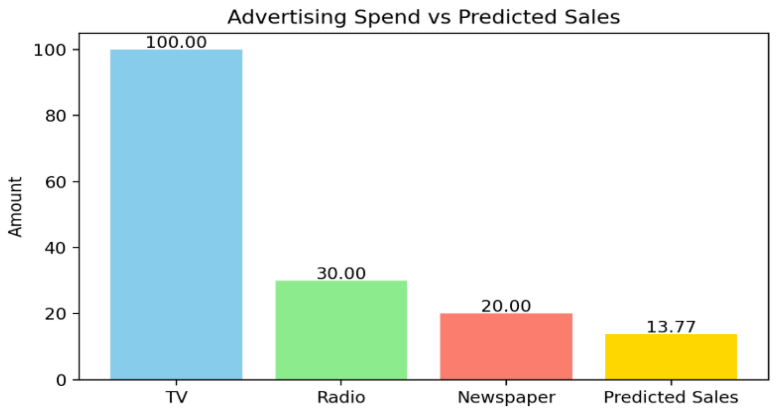
Predict Sales

Predicted Sales: 13.77

Entered & Derived Features


```
{
  "TV" : 100
  "Radio" : 30
  "Newspaper" : 20
  "Total_Ads" : 150
  "TV_Radio" : 3000
  "TV_Newspaper" : 2000
  "Radio_Newspaper" : 600
}
```


Visualization



Developed by: **Ayesha Banu** | Data Science Project

CHAPTER - 7

Insights, Findings & Conclusion

7. Insights, Findings & Conclusion

7.1 Key Insights from the Analysis

1. **Impact of Advertising Channels:**
    - TV advertising showed the strongest correlation with sales, making it the most effective medium for driving revenue.
    - Radio advertising had a moderate positive effect on sales.
    - Newspaper advertising had the least influence compared to TV and Radio, though interaction effects with other channels added some value.
2. **Interaction Effects:**
    - Features such as TV × Radio and TV × Newspaper highlighted synergistic effects, meaning combined spending across channels boosted sales more effectively than isolated spending.
3. **Outlier Analysis:**
    - A few extreme values were detected (particularly in Newspaper spend), but they were retained to maintain model robustness and real-world variability.
4. **Model Performance:**
    - Linear Regression performed well but struggled with nonlinear relationships.
    - Random Forest significantly improved predictions.
    - Gradient Boosting Regressor delivered the best performance, with $R^2 = 0.989$ and RMSE = 0.599, indicating excellent predictive accuracy.

7.2 Business Findings

- Allocating higher budgets to TV and Radio will yield the greatest increase in sales.
- Newspaper advertisements alone do not significantly contribute to sales but can complement other channels.
- A balanced multi-channel advertising strategy is more effective than focusing on a single channel.
- Machine learning models can serve as a decision support system for optimizing marketing budgets.

7.3 Conclusion

This project successfully demonstrated how machine learning can be applied to optimize advertising strategy by predicting sales based on ad spend.

- The project followed a complete data science pipeline:
  Data Collection → Preprocessing → EDA → Feature Engineering → Modeling → Deployment.
- The deployed Streamlit app enables business teams to input advertising budgets and receive real-time sales predictions, ensuring practical usability.
- The best-performing model (Gradient Boosting Regressor) provided highly accurate predictions, giving the company a reliable tool for budget optimization and revenue maximization.

CHAPTER - 8

Future Enhancements

8. Future Enhancements

While the current project provides a strong foundation for sales prediction using advertising data, there are several opportunities to expand and enhance the solution in future iterations.

8.1 Data Enhancements

- **Incorporate Additional Channels**: Extend the dataset to include modern advertising mediums such as Social Media (Facebook, Instagram, YouTube, TikTok), Search Engine Marketing, and Email Campaigns.
- **Real-Time Data Integration**: Connect the model to live marketing campaign data streams (via APIs) for dynamic and up-to-date predictions.
- **Customer Demographics**: Add demographic attributes (age, income level, region) to improve personalization of predictions.

8.2 Modeling Enhancements

- **Hyperparameter Tuning**: Apply advanced techniques such as Grid Search or Bayesian Optimization to fine-tune model parameters.
- **Cross-Validation**: Implement K-Fold cross-validation to ensure model generalizability.
- **Ensemble Techniques**: Explore model stacking, blending, or advanced boosting algorithms (e.g., XGBoost, LightGBM, CatBoost) for improved performance.
- **Time-Series Forecasting**: Extend the project to predict future sales trends over time using ARIMA, Prophet, or LSTM models.

8.3 Deployment Enhancements

- **Cloud Deployment**: Host the Streamlit app on cloud platforms (AWS, Azure, or Google Cloud) for global access.
- **REST API Development**: Wrap the prediction model in a REST API to integrate with enterprise systems, CRM tools, or mobile applications.
- **Interactive Dashboards**: Enhance the app with dashboards (using Plotly Dash or Power BI) to visualize ad spend vs. sales relationships.
- **Multi-User Authentication**: Add login systems to support different user roles (Marketing Analyst, Manager, Executive).

8.4 Business Impact Enhancements

- **ROI Analysis Module**: Extend predictions to include return on investment (ROI) for each channel.
- **What-If Scenario Testing**: Allow users to simulate scenarios like "What if we increase TV spend by 20%?" and instantly observe sales impact.
- **Budget Optimization Tool**: Use optimization algorithms to suggest the best allocation of ad budgets for maximum sales.

8.5 Long-Term Vision

- Transition the solution from a predictive model to a decision intelligence platform that integrates with the company's entire marketing ecosystem.
- Enable AI-driven campaign recommendations for automated budget planning and execution.

CHAPTER - 9

Appendix

9. Appendix

9.1 Data Dictionary

Feature	Type	Description
TV	Float	Advertising spend on TV (in thousands of dollars).
Radio	Float	Advertising spend on Radio (in thousands of dollars).
Newspaper	Float	Advertising spend on Newspaper (in thousands of dollars).
Sales	Float	Product sales (target variable).
Total_Ads	Float	Engineered feature: sum of TV, Radio, and Newspaper spends.
TV_Radio	Float	Engineered feature: interaction between TV and Radio spends.
TV_Newspaper	Float	Engineered feature: interaction between TV and Newspaper spends
Radio_Newspaper	Float	Engineered feature: interaction between Radio and Newspaper spends.

9.2 Algorithms Summary

Algorithm	Description	Description
Linear Regression	Statistical method to model linear relationship between input features and sales.	Simple, interpretable baseline model.
Random Forest Regressor	Ensemble of decision trees using bagging to reduce variance and improve accuracy.	Handles nonlinear data, reduces overfitting.
Gradient Boosting Regressor	Sequential ensemble method improving errors of previous trees.	High accuracy, best-perform ing model in this project.

9.3 Module Summary

Module	Purpose
data_processing.py	Loads and preprocesses raw dataset (removes duplicates, handles missing values).
feature_engineering.py	Adds engineered features like Total_Ads and interaction terms.
eda (notebook)	Performs exploratory data analysis, generates visualizations, detects outliers.
model.py	Trains multiple regression models, compares results, saves best model & scaler.
deploy.py	Loads saved model & scaler, provides prediction function for new inputs.
app.py (Streamlit)	User interface for real-time predictions with interactive input fields.
reports/	Stores visualizations such as histograms, boxplots, heatmaps, pairplots.

9.4 References

1. **Advertising.csv Dataset** – Widely used dataset in machine learning regression problems.

2. **Scikit-learn Documentation** – https://scikit-learn.org/stable/

3. **Streamlit Documentation** – https://docs.streamlit.io/

4. **Ensemble Learning Methods** – Research articles on boosting and bagging techniques.