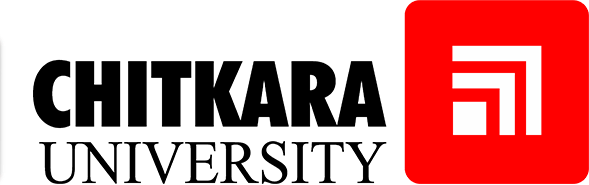
Artificial Intelligence and Machine Learning

Project Report Semester-IV (Batch 2022)

**BigMart Sales Prediction**



#### Supervised By: Submitted By:

Mr. Talib 2210992355: Shruti Bansal

2210992401: Soumya Makkar

2210992528: Vinay Kumar Mangal

2210992532: Vinit Kumar

**Department of Computer Science and Engineering Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab**

### **Abstract**

Retail industries are increasingly turning to advanced Artificial Intelligence and Machine Learning methodologies to optimize operations and enhance decision-making processes. In this project, we focus on predicting sales for BigMart retail stores, leveraging a comprehensive dataset comprising historical sales records alongside various product and store attributes. Through meticulous preprocessing techniques, including data cleaning and normalization, we prepare the dataset for analysis, ensuring the integrity and quality of the input data. Our approach involves extensive feature engineering, extracting relevant features such as item visibility, store size, and promotional activities, to enhance the predictive capabilities of our models.

We employ a diverse set of machine learning algorithms, including Linear Regression, Support Vector Regression, and Decision Tree to construct predictive models capable of forecasting sales with high accuracy. By systematically evaluating the performance of each model using metrics such as Mean Squared Error and R-squared value, we identify the most effective algorithm for the task at hand. Furthermore, we delve into the interpretability of the models, examining feature importance scores to discern the underlying factors driving sales trends. This analysis provides valuable insights for stakeholders, shedding light on the critical variables influencing sales performance within the BigMart retail environment.

Through rigorous experimentation and validation techniques, we demonstrate the efficacy of our predictive models in accurately forecasting sales figures for BigMart stores. Our findings not only showcase the potential of machine learning in the retail sector but also highlight its practical implications for inventory management, pricing strategies, and overall business optimization. By leveraging data-driven insights derived from our predictive models, retailers can make informed decisions, mitigate risks, and capitalize on emerging opportunities in an ever-evolving market landscape. This project underscores the significance of AI and ML techniques in revolutionizing traditional retail practices, paving the way for data-driven decision-making and sustainable business growth.

In conclusion, our project offers a comprehensive analysis of sales prediction in the context of BigMart retail stores, leveraging advanced AI/ML methodologies to derive actionable insights from raw data. Through meticulous data preprocessing, feature engineering, and model evaluation, we demonstrate the effectiveness of machine learning algorithms in forecasting sales trends with high accuracy. Our findings provide valuable guidance for retailers seeking to optimize their operations, enhance customer experiences, and drive business growth in a competitive market environment.

|  |  |  |
| --- | --- | --- |
| S No. | Topics | Page No. |
| 1 | Introduction | 4-5 |
| 2 | Scope | 6-7 |
| 3 | Problem Statement | 8 |
| 4 | Software and Hardware requirements | 9 |
| 5 | Proposed Solution | 10 |
| 6 | Results | 11 |
| 7 | References | 12 |

1. **Introduction**
   1. **Background:**

The retail industry is rapidly changing due to technology and shifting consumer behaviors. Traditional retailers face pressure to adapt as e-commerce grows. Predictive analytics, like sales prediction, is crucial for retailers to understand consumer preferences, manage inventory, and improve efficiency. However, accurately forecasting sales is complex due to factors like seasonality and promotions.

BigMart, a major retail chain, needs accurate sales predictions to succeed in a competitive market. They understand the value of data-driven insights for growth. By using advanced AI and ML, BigMart aims to improve sales predictions, optimize inventory, and boost revenue. Our project focuses on developing precise predictive models tailored to BigMart's needs, providing valuable insights for decision-makers throughout the organization**.**

**1.2 Objectives:**

The objective of this project is to predict sales for BigMart using AI/ML models to provide actionable insights and aid decision-making in sales and inventory management. The dataset consists of columns such as Item\_Identifier, Item\_Weight, Item\_Fat\_Content, Item\_Visibility, Item\_Type, Item\_MRP, Outlet\_Identifier, Outlet\_Establishment\_Year, Outlet\_Size, Outlet\_Location\_Type, Outlet\_Type, and Item\_Outlet\_Sales.

By focusing on sales prediction, the project offers BigMart a powerful tool to optimize operations, improve customer satisfaction, and increase revenue. The models and insights provided can support the company's decision-making process in various areas, including inventory management, marketing, and strategic planning.

**1.3 Significance:**

Accurate sales prediction holds immense significance for retailers like BigMart, enabling them to optimize various aspects of their operations and strategic decision-making processes. By leveraging advanced AI/ML techniques to forecast sales trends, retailers can effectively manage inventory levels, ensuring adequate stock availability while minimizing instances of overstocking or stockouts. This optimization of inventory management not only improves operational efficiency but also reduces storage costs and mitigates the risk of unsold inventory, thereby enhancing overall profitability.

Furthermore, the insights derived from sales prediction models empower retailers to devise targeted marketing strategies and pricing tactics tailored to specific product categories or customer segments. By understanding the underlying factors driving sales performance, retailers can optimize promotional activities, allocate resources more effectively, and enhance customer engagement. Moreover, accurate sales forecasts enable retailers to anticipate demand fluctuations, proactively adjust procurement strategies, and capitalize on emerging market trends. Ultimately, the implementation of AI/ML-driven sales prediction models equips retailers with a competitive edge in a dynamic and rapidly evolving retail landscape, fostering sustainable growth and profitability in the long term.

1. **Scope Of the Project:**
2. **Data Collection and Preparation:**

Gather a diverse dataset of electronic health records including as Item\_Identifier, Item\_Weight, Item\_Fat\_Content, Item\_Visibility, Item\_Type, Item\_MRP, Outlet\_Identifier, Outlet\_Establishment\_Year, Outlet\_Size, Outlet\_Location\_Type, Outlet\_Type, and Item\_Outlet\_Sales .

Clean the dataset, handle missing values, and perform necessary preprocessing steps such as data normalization and feature scaling.

1. **Feature Selection and Engineering:**

Identify the most relevant features for stroke prediction using feature selection techniques.

Engineer new features if necessary to improve the predictive performance of the model.

1. **Model Development:**

Implement and train multiple machine learning (ML) algorithms including linear regression, decision trees and support vector machines.

Optimize hyperparameters for each algorithm to maximize predictive performance.

1. **Model Evaluation:**

Assess the performance of the developed ML models using appropriate evaluation metrics such as accuracy, r2 score and root mean squared error.

Perform cross-validation to ensure the generalizability and robustness of the model.

1. **Comparison and Selection of Optimal Model:**

Compare the performance of different ML algorithms and select the most accurate and reliable model for stroke prediction.

1. **Interpretation and Validation:**

Interpret the results of the selected model, analyze feature importance, and validate the model using cross-validation techniques.

Validate the model on an independent dataset if available.

1. **Implementation and Deployment:**

Develop a user-friendly interface for the final ML model, allowing healthcare providers to input patient data and obtain personalized stroke risk assessments in real-time.

Deploy the model in a healthcare setting and assess its usability and effectiveness in a real-world environment.

**(viii)Documentation and Reporting:**

Document the entire process including data collection, preprocessing, feature selection, model development, evaluation, and deployment.

Prepare a comprehensive report detailing the methodology, results, and implications of the study.

1. **Future Work:**

Discuss potential areas for future research and improvement, including the integration of additional data sources, refinement of the model, and validation in larger and more diverse patient populations

1. **Problem Definition and Requirement:** 
   1. **Statement:**

BigMart, a leading retail company, faces challenges in accurately forecasting sales for its diverse range of products across multiple outlets. Inconsistent and unreliable sales predictions can lead to inefficient inventory management, resulting in either overstock or stockouts, which in turn affects profitability and customer satisfaction.

The project aims to develop a machine learning model that can accurately predict sales for each item-outlet combination based on historical sales data and other influencing factors such as item characteristics, pricing, outlet location, and other features. The model should be capable of providing actionable insights that can inform strategic decisions related to inventory planning, pricing, and promotions.

1. **REQUIREMENTS:**
   1. **Software Required:**

**Programming Language:**

Python (version 3.x)

Integrated Development Environment (IDE):

Jupyter Notebook or JupyterLab for interactive development and experimentation.

**Python Libraries:**

NumPy: For numerical computing and array operations.

Pandas: For data manipulation and analysis.

Scikit-learn: For machine learning algorithms and model evaluation.

Matplotlib: For data visualization.

Seaborn: For statistical data visualization.

* 1. **Hardware Requirement:**

The hardware requirements for the project include a personal computer or laptop with the following minimum specifications:

- Processor: Intel Core i5 or equivalent

- RAM: 8GB or higher

- Storage: 256GB SSD or higher

- Operating System: Windows 10, macOS, or Linux

- Stable internet connection

**PROPOSED SOLUTION:**

1. **Proposed Design and Methodology:**

**Data Collection and Preprocessing:**

Gather a diverse dataset of electronic health records including demographic information, medical history, lifestyle factors, and clinical biomarkers.

Clean the dataset, handle missing values, and perform necessary preprocessing steps such as data normalization and feature scaling.

**Feature Selection and Engineering:**

Identify the most relevant features for stroke prediction using feature selection techniques such as correlation analysis, feature importance, and domain knowledge.

Engineer new features if necessary to improve the predictive performance of the model.

**Model Development:**

Implement and train multiple machine learning (ML) algorithms including logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines using the scikit-learn library in Python.

Optimize hyperparameters for each algorithm to maximize predictive performance using techniques such as grid search or random search.

**Model Evaluation:**

Assess the performance of the developed ML models using appropriate evaluation metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve and precision-recall curve.

Perform cross-validation to ensure the generalizability and robustness of the model.

**Comparison and Selection of Optimal Model:**

Compare the performance of different ML algorithms and select the most accurate and reliable model for stroke prediction based on the evaluation metrics.

**Results on the Prediction Models: -**

1. **LINEAR REGRESSION:**

R2 score 0.48624380425175895

Root Mean Squared Error 1259.3664118274296

1. **SVM (SUPPORT VECTOR MACHINE):**

Average cross-validation score: 0.4657209479992114

Mean absolute error: 965.0922385659601

R-squared score: 0.4410506443960891

1. **DECISION TREE:**

Mean absolute error: 966.3037881138757

Mean squared error: 1783319.9713601293

R-squared: 0.4223269150841127

Median absolute error: 725.1378442519141

**REFERENCES:**

* Took the dataset from Kaggle
* Took help from sckitLearn documentation