 **Mini Project Report on**

**Sale prediction analysis by using ensemble learning over Walmart**

**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name – Daman Rakwal University Roll No. 2018772**

***Under the Mentorship of***

**Dr. Ankit Tomar**

**Assistant Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**December-2023**



**CANDIDATE’S DECLARATION**

I hereby certify that the work presented in the project report entitled " Sale prediction analysis by using ensemble learning over Walmart " in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University , Dehradun) has been carried out by me under the mentorship of Dr. Ankit Tomar, Assistant Professor, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name – Daman Rakwal University Roll no.- 2018772

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 1-2 |
| Chapter 2 | Literature Survey | 3-4 |
| Chapter 3 | Methodology | 5-7 |
| Chapter 4 | Result and Discussion | 8-9 |
| Chapter 5 | Conclusion and Future Work | 10 |
|  | References | 11 |

**Chapter 1**

**Introduction**

**1.1 Introduction:**

Ensemble learning, a powerful meta-approach in machine learning, enhances predictive performance by aggregating predictions from multiple models. In the context of retail, where accurate sales prediction is paramount, this project focuses on "Sale Prediction and Analysis using Ensemble Learning over Walmart." By leveraging advanced techniques such as XGBoost and Random Forest, the project aims to amalgamate predictions and gain valuable insights from Walmart's historical sales data. Through meticulous data preprocessing, feature engineering, and hyperparameter tuning, the project endeavors to deliver a robust model facilitating informed decision-making and strategic planning within the retail giant.

**1.2 Problem Statement**:

Traditional sales prediction approaches often fall short due to complex data analysis, limited model accuracy, and an inability to adapt to dynamic market conditions. Manual forecasting methods may result in inefficiencies in inventory management and missed business opportunities. The mini-project addresses these challenges by developing a "Sales Production" application. This application is designed to offer an advanced, accurate, and scalable sales prediction system, allowing businesses to predict sales trends, optimize resource allocation, and make data-driven decisions.

**1.3 Scope of Work:**

The project's scope involves the design and implementation of a comprehensive sales prediction system using ensemble learning techniques. The application allows users to input historical sales data, automates data preprocessing, and employs ensemble models, such as

XGBoost and Random Forest, for accurate predictions. Users can visualize sales trends, gain insights into influential factors, and make informed decisions based on predictive analytics. The system incorporates features like hyperparameter tuning, model evaluation, and exportable

reports to enhance usability. The overarching goal is to create a versatile platform empowering businesses with accurate sales forecasts, contributing to improved strategic planning and operational efficiency.

**Chapter 2**

**Literature Survey**

**2.1 Introduction**

This chapter conducts a comprehensive literature survey to explore existing research and technologies in the domain of sales prediction and ensemble learning. The review aims to identify strengths and weaknesses in various approaches and technologies, providing insights to guide the development of the project.

**2.2 Existing Sales Prediction Applications**

Several applications and techniques are available for sales prediction, each offering unique features and methodologies. Some notable examples include:

**2.2.1 Salesforce:**

Salesforce is a widely-used customer relationship management (CRM) platform that incorporates predictive analytics for sales forecasting. It leverages machine learning algorithms to analyze historical data and predict future sales trends.

**2.2.2 IBM Watson Analytics:**

IBM Watson Analytics is a cognitive computing platform that includes predictive analytics for sales. It enables businesses to uncover patterns and make data-driven decisions by analyzing historical sales data.

**2.2.3 RapidMiner:**

RapidMiner is a data science platform that provides tools for predictive analytics, including sales forecasting. It allows users to build, evaluate, and deploy predictive models to enhance decision-making.

**2.2.4 Ensemble Learning in Sales Prediction:**

Research in ensemble learning for sales prediction has gained traction. Studies explore the effectiveness of combining multiple models, such as Random Forests, XGBoost, and neural networks, to improve the accuracy of sales forecasts.

**2.3 Technologies and Features**

The literature survey identifies various technologies and features employed in sales prediction applications:

**1. Python**- The entire code is written in Python, a versatile and widely-used programming language for data science and machine learning.

**2. Libraries and Frameworks**:

**NumPy and Pandas**:-Used for numerical and data manipulation, respectively.

**Matplotlib and Seaborn**:- Utilized for data visualization.

**Scikit-learn**:-Provides tools for machine learning tasks, such as data preprocessing, model selection,

and evaluation.

**XGBoost**:- An implementation of gradient boosting, used for regression in this project.

**RandomForestRegressor** :- Part of the scikit-learn library, used for ensemble learning.

**2.4 Comparison and Analysis**

Comparative analyses of various sales prediction approaches highlight their strengths and limitations:

**2.4.1 Ensemble Learning Effectiveness:**

Ensemble learning techniques, particularly when combining diverse models, demonstrate superior performance in handling the complexity of sales data and improving prediction accuracy.

**2.4.2 Feature Importance:**

Studies emphasize the significance of identifying and utilizing the most relevant features in sales prediction models, underscoring the impact of feature selection techniques.

**2.4.3 Time Series vs. Machine Learning Models:**

Comparisons between traditional time series models and machine learning models showcase the advantages of machine learning approaches, especially in capturing nonlinear relationships and complex patterns.

**Chapter 3**

**Methodology**

**1. Problem Definition and Business Understanding:**

**1.1 Problem Statement**

The project focuses on predicting Walmart weekly sales to facilitate better business planning and decision-making.

**1.2 Business Context**

Accurate weekly sales predictions are crucial for optimizing inventory, workforce planning, and overall operational efficiency.

**2. Data Acquisition and Exploration:**

**2.1 Data Source**

The dataset is sourced from Kaggle, containing information on weekly sales and relevant features.

**2.2 Exploratory Data Analysis (EDA)**

- Explore dataset characteristics, structure, and size.

- Identify key features and their distributions.

- Check for missing values and outliers.

**3. Data Preprocessing:**

**3.1 Date Handling**

- Convert the 'Date' column to datetime format.

- Extract temporal features such as 'Year,' 'Month,' and 'Day.'

**3.2 Categorical Encoding**

- Use Label Encoding to convert categorical features ('Store' and 'Holiday\_Flag') to numerical format.

**4. Data Splitting:**

**4.1 Train-Test Split**

- Divide the dataset into training and testing sets for model evaluation.

- Use an 80-20 split, ensuring a random seed for reproducibility.

**5. Technology Stack:**

**5.1 Programming Language**

- Python: Utilized for its extensive libraries and frameworks for machine learning.

**5.2 Libraries and Frameworks**

- NumPy: For numerical computations.

- Pandas: For data manipulation and analysis.

- Matplotlib and Seaborn: For data visualization.

- Scikit-learn: For machine learning models, preprocessing, and evaluation.

- XGBoost: For the XGBoost regression model.

- RandomForestRegressor: For the Random Forest regression model.

**6. Model Selection and Training:**

**6.1 XGBoost**

- XGBoost (Extreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting.

- Key Features:

- Regularization for controlling overfitting.

- Parallel computing for faster training.

- Handles missing data.

- Flexible and robust.

**6.2 Random Forest**

- Random Forest is an ensemble learning method based on constructing a multitude of decision trees.

- Key Features:

- Diversity in tree construction through bootstrap sampling.

- Reduction of overfitting through averaging.

- Suitable for regression tasks.

- Handles missing values.

**7. Model Evaluation:**

**7.1 R-squared Metric**

- Utilize the R-squared metric to assess model performance.

- Evaluate on both training and test datasets to identify overfitting or underfitting.

**8. Ensemble Learning:**

**8.1 Averaging Predictions**

- Combine predictions from XGBoost and RandomForestRegressor through simple averaging.

- Evaluate the ensemble model's performance using the R-squared metric.

**Chapter 4**

**Result and Discussion**

**1. Model Performance:**

**1.1 XGBoost Model Performance**

The XGBoost model demonstrated strong predictive capabilities for Walmart weekly sales. The R-squared value for the training data was found to be 0.99675(as per Kaggle dataset) , indicating a high level of explained variance. However, the model's performance on the test data yielded an R-squared value of 0.97792(as per Kaggle dataset). The discrepancy between training and test performance suggests a potential for overfitting.

**1.2 Random Forest Model Performance**

The Random Forest model, another powerful regression algorithm, exhibited promising results. The R-squared value for the training data was 0.99368(as per Kaggle dataset), demonstrating effective learning on the training set. On the test data, the R-squared value was 0.95496(as per Kaggle dataset). While the model generally generalizes well, further exploration is needed to optimize its parameters.

**2. Ensemble Learning Results:**

**2.1 Averaging Predictions**

The ensemble model, created by averaging predictions from XGBoost and Random Forest, aimed to leverage the strengths of both algorithms. The resulting R-squared value for the test data was 0.97414 (as per Kaggle dataset), showcasing a potentially improved overall performance compared to individual models.

**3. Comparison and Insights**

**3.1 Model Comparison**

Comparing the individual models and the ensemble, it's evident that the ensemble model outperformed the standalone models on the test set. The ensemble approach leveraged the diversity in predictions from XGBoost and Random Forest, resulting in a more robust and accurate model.

**3.2 Feature Importance**

Insights into feature importance reveal that key features played a crucial role in predicting weekly sales. Further analysis is needed to understand the dynamics of these features and their impact on sales fluctuations.

**4. Future Directions:**

**4.1 Feature Engineering**

Exploration of additional features and advanced feature engineering techniques could contribute to improved model accuracy. Considering external factors such as economic indicators might further enhance the predictive power of the models.

**4.2 Advanced Models**

Investigation into more advanced regression models or ensemble methods could be a focus for future work. Deep learning models may capture complex patterns that traditional regression models might miss.

**Chapter 5**

**Conclusion**

In conclusion, the project successfully developed and evaluated machine learning models for predicting Walmart sales using ensemble learning. The ensemble model, in particular, demonstrated superior performance on the test set, showcasing the potential of combining diverse algorithms for better predictive accuracy. Ongoing efforts in feature engineering and model optimization will contribute to refining the model's predictive capabilities.

In future I will try to add some more algorithms to this project so that the prediction rate of this project improves and certainly gives more accurate results and also I will try to make a front-end for this project .

**References**

1. Walmart Dataset - <https://www.kaggle.com/datasets/yasserh/walmart-datasets>.

2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754.

3. Breiman, L. (2001). Random Forests. Machine learning, 45(1), 5-32.

4. Wes McKinney. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. Python for High Performance and Scientific Computing, 14.