ENHANCING RETAIL PRICE PREDICTION : A COMPARATIVE STUDY OF MACHINE LEARNING AND DEEP LEARNING MODELS

Module Code	:	CSMAI21
Assignment report Title	:	Coursework
Convenor Name	:	Dr. Xiaomin Chen
Student name	:	Shruti Patkar
Student number	:	31810964
Date.	:	7th March, 2024
Hours	:	40 hours

ABSTRACT: This study investigates machine learning and deep learning models for predicting product prices in the retail sector, focusing on supermarkets. Traditional machine learning techniques, including decision tree and random forest models, were compared alongside deep learning architectures such as simple neural networks, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks. Through meticulous hyperparameter tuning and cross-validation, we evaluated the performance of these models in terms of predictive accuracy and generalization. Our findings highlight the superior performance of random forest followed by decision tree and simple neural network as deep learning model as the most effective deep learning approach for price prediction. Additionally, we discuss the importance of integrating diverse methodologies to develop robust price prediction models capable of meeting the dynamic demands of the retail landscape. Furthermore, future research directions, including the exploration of ensemble learning methods and reinforcement learning frameworks, are proposed to enhance predictive accuracy and enable more sophisticated pricing strategies.

1. BACKGROUND

In today's retail landscape, precise pricing strategies are vital for both consumers and retailers, particularly within the expansive product assortments found in supermarkets. Accurate price predictions aid consumers in budget planning and decision-making while shopping, while for retailers, pricing strategies directly impact profitability and customer satisfaction. Hence, developing accurate price prediction models is paramount for the retail industry's success.

This study aims to develop machine learning models capable of accurately predicting product prices based on attributes such as category, brand, unit and other features. These models empower consumers to make informed purchase decisions and enable retailers to refine pricing strategies, optimize inventory management, and enhance overall profitability. By leveraging big data and advanced analytics, retailers can gain valuable insights into consumer behaviour and market trends, maintaining competitiveness and adapting to evolving market dynamics. While time series analysis[1] may present a potential approach for forecasting prices based on sequential data points over time, it may not be the most suitable method for this specific analysis. Time series models excel in capturing temporal dependencies and seasonality patterns within a single time series dataset, but they may struggle to adequately capture the intricate relationships between attributes such as category, brand, and price in a supermarket setting. The decision to focus on machine learning models over time series analysis is driven by the need to effectively utilize the diverse feature set available in the dataset and capture the nuanced relationships between product attributes and prices in a dynamic retail environment.

2. RELATED WORK

Several recent studies have tackled the complexities of demand and price prediction across various domains. Warnakulasooriya et al. (2020)[2] focus on supermarket retail, employing machine learning and deep learning techniques, along with blockchain technology, to predict demand and prices of vegetables, aiming to optimize supply chain management. Wang and Gao (2018)[3] address the prediction of high and low prices of soybean futures using LSTM neural networks, demonstrating the efficacy of such models in financial market forecasting. In a different context, Yi et al. (2023)[4] propose a genetic algorithm-based optimization solution for replenishment and pricing strategies in supermarkets, aiming to efficiently manage inventory and pricing decisions. Zhu et al. (2022)[5] tackle occupancy prediction in hotel dynamic pricing by modeling price elasticity, providing insights into accurately estimating room occupancy in online hotel booking platforms. These studies collectively contribute to advancing predictive modeling techniques in diverse domains, from retail and finance to hospitality, addressing critical challenges in demand and price forecasting.

3. DATA PREPROCESSING

The dataset utilized in this analysis originates from Kaggle's Time Series UK Supermarket Data[6], specifically the All_Data_Tesco.csv file. Data pre-processing is crucial for preparing the dataset for analysis, involving tasks like managing missing values, eliminating duplicates, refining text data, and encoding categorical variables. In this dataset of 753,502 entries and 8 columns encompassing features such as supermarket, prices, unit, names, date, category, and own_brand, initial steps ensure data quality and readiness for analysis.

Starting with the removal of duplicate entries, text data is standardized by converting entries in the "names" and "unit" columns to lowercase, and stop words are eliminated to filter out insignificant terms. Missing values in columns like "category" and "supermarket" are handled with imputation methods, while zero prices undergo specialized procedures involving regular expressions to ensure consistency.

Text data quality is enhanced by removing redundant words and employing lemmatization techniques, while categorical variables are numerically encoded using Pandas' Categorical codes. After pre-processing, the dataset comprises 729,668 entries and 6 columns, optimized for computational efficiency, laying a solid foundation for machine learning model development and insightful analysis.

4. EXPLORATORY DATA ANALYSIS AND FEATURE SELECTION

Exploratory Data Analysis (EDA) serves as a critical initial phase in comprehending the structure and attributes of the dataset before delving into machine learning modelling. In this study, the dataset encompasses a range of product attributes, including prices, categories, brands, and other relevant features. The primary objective of EDA is to unveil patterns, trends, and correlations within the dataset through descriptive statistics and visual representations.

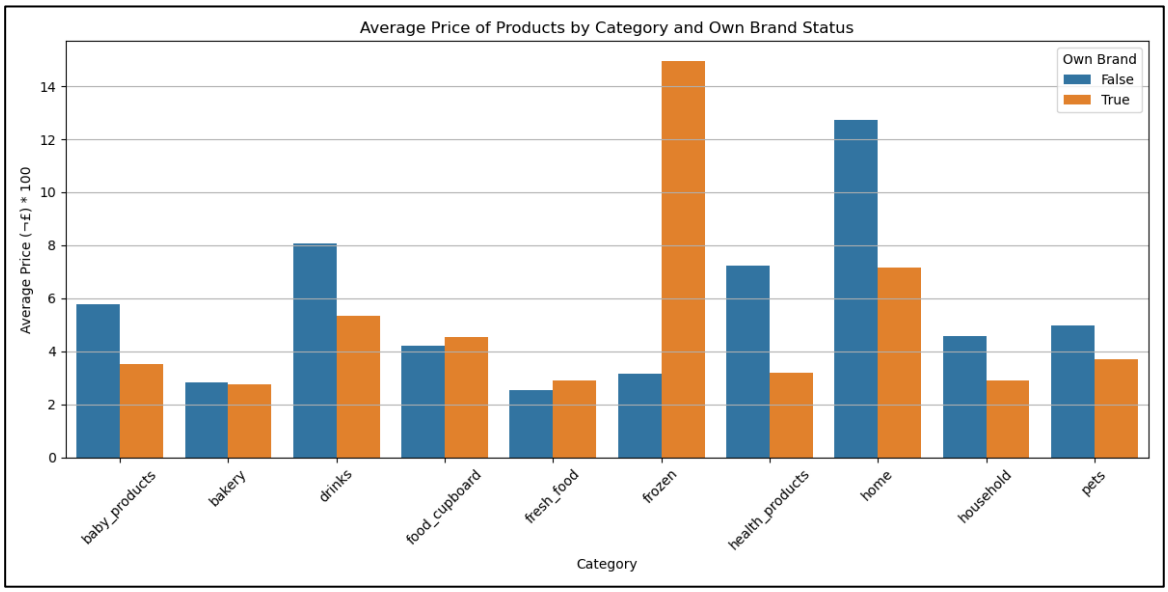

Figure 1: Average Price Products by Category and Own Brand Status

In the exploratory data analysis phase, visualization techniques like bar plots[Figure 1] are employed to depict average prices across categories and brand statuses, offering insights into price distribution. Here we find that frozen data has highest number of products which are not manufactured by Tesco while lowest is fresh food. While home products manufactured by the supermarket itself is home and least is fresh food again. A line plot illustrates the frequency of product name values, revealing patterns of repetition in the dataset.

Furthermore, a sunburst chart is employed to visually represent the distribution of counts by own brand, category, and date for the identification of prevalent categories, brands, and temporal patterns. Moreover, a series of subplots are utilized to showcase price variation over time for each product, enabling the observation of trends and fluctuations in product prices across specific time intervals for eight different product.

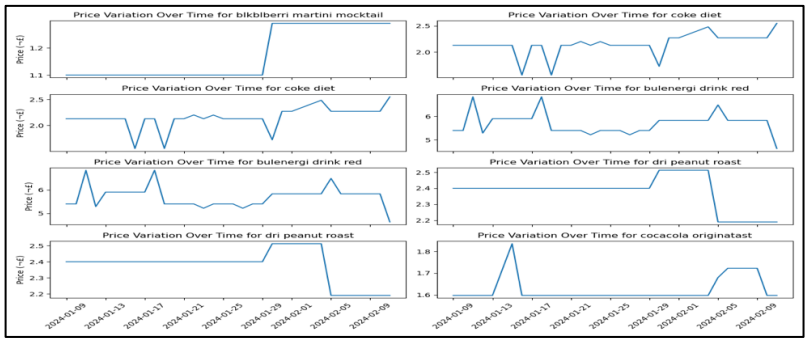

Figure 2: Price Prediction Over time for 8 product names

In addition to visualizations, boxplots are utilized to identify outliers in numerical columns, with a particular focus on prices. For prices, which exceeded a threshold of 120, indicating significant deviations from typical values, they were not removed from the dataset. However, outliers greater than 500 were eliminated to ensure data quality and consistency.

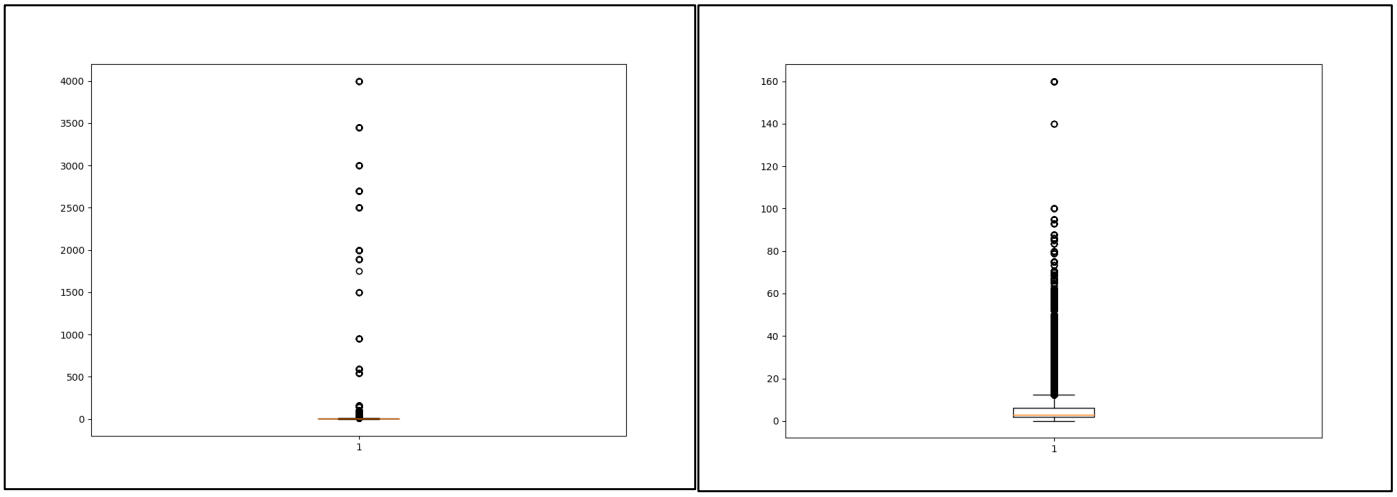

Figure 3: Boxplot for prices showing outliers before and after removal of prices greater than 500

The scatter plot analysis revealed that the dataset does not exhibit significant bias, as evidenced by the relatively consistent distribution of data points. Consequently, it may not be necessary to apply smoothing techniques to the data. This along with boxplot for rest of the features suggest that the dataset is relatively smooth and free from substantial fluctuations or irregularities, facilitating more straightforward analysis and interpretation.

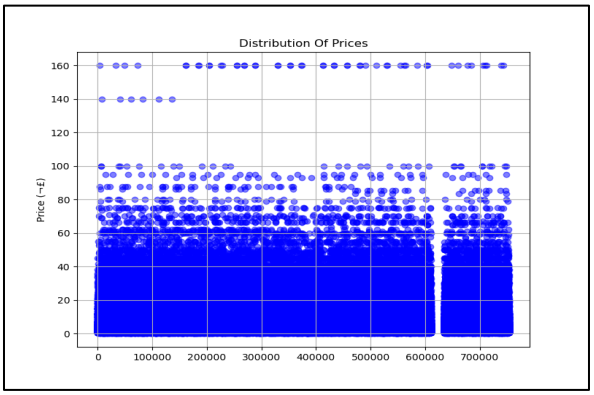

Figure 4: Scatterplot showing distribution of prices column

Feature selection involves segmenting the date column and generating correlation plots to assess relationships between features. During analysis, the "year" column, containing identical data, was removed. Minimal correlation was observed between the "day_of_week" column and other features, indicating limited relevance. Strong correlations were found between certain attributes, notably the "unit" column, while weaker correlations were observed for the "month" column. Additionally, a correlation between "month" and "day_of_month" suggests interrelation between temporal features.

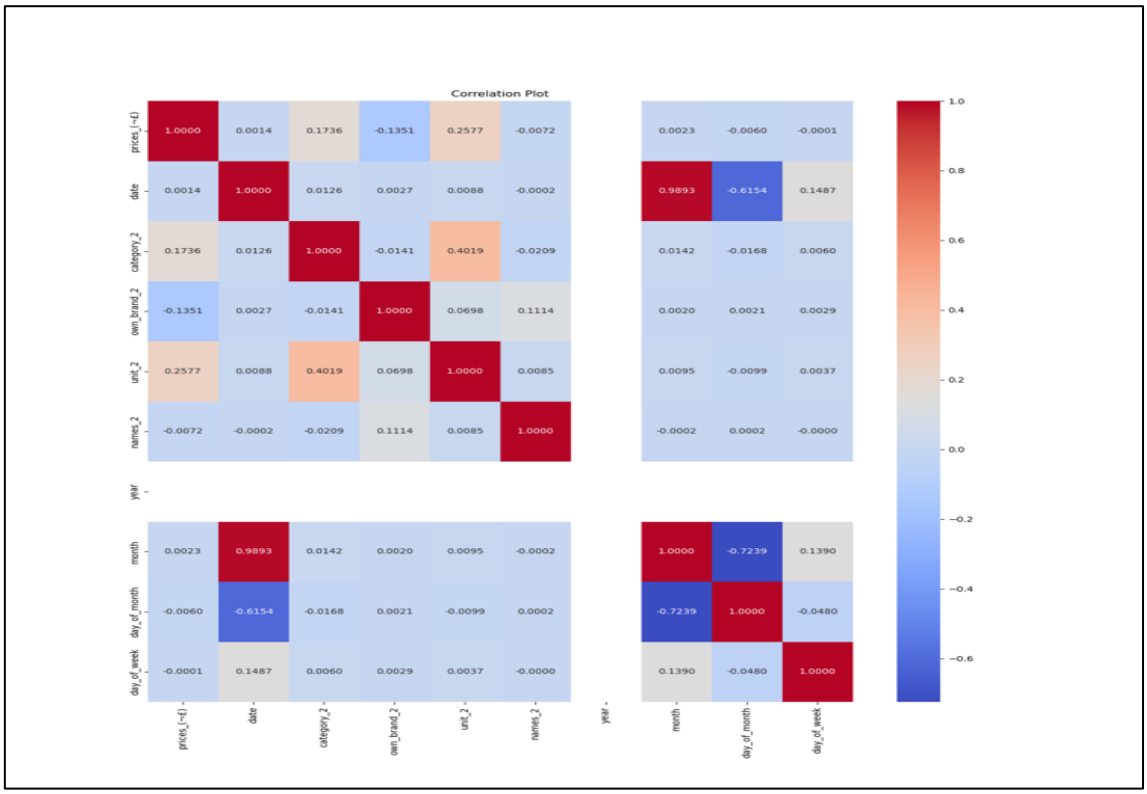

Figure 5: Correlation matrix for all numerical columns

Ultimately, duplicate values are removed from the dataset after discarding redundant columns based on the correlation matrix. The processed data is then saved as a CSV file, ready for machine learning model training and evaluation. Through these steps, exploratory data analysis enables a comprehensive understanding of the dataset's characteristics and lays the groundwork for developing robust and accurate machine learning models for price prediction in the supermarket domain.

5. TRAINING DATA SPLIT AND MODEL TESTING

The dataset is partitioned into input features (X) and target values (y). Before this division, certain features undergo standardization using the "StandardScaler"[7] method to maintain consistent feature scales, thereby enhancing model performance during training and evaluation.

Following standardization, the dataset is split into training and validation sets using a random allocation, with 80% reserved for training and 20% for validation. Additionally, a stratified split is employed to ensure an even distribution across various price ranges, promoting better generalization during model training and testing.

Once divided, the data is saved into separate CSV files for future reference, after which model training and evaluation are conducted using a variety of algorithms optimized for regression tasks.

The predicted values from each model are visualized using line plots, showcasing their performance in predicting prices. The plots[Figure 5] depict the predicted values and true values for both the entire dataset and the first 100 inputs in same as in Table 1 with black color being the true values.

Model	Training RMSE	Testing RMSE	R2 Score
KNN	4.567	5.693	0.288
Decision Tree	0.852	1.921	0.919
Extra Tree	0.852	2.209	0.893
XGBoost	5.678	5.684	0.291
Gradient Boost	5.753	5.788	0.265
Random Forest	0.922	1.658	0.940

Table 1: RSME and R2 Score for Traditional Models

Among these models, Decision Tree, Extra Tree Regresser and Random Forest exhibit the best performance. Notably, Random Forest and Decision Tree performs better than extra tree classifier as observed in the testing RMSE and R2 score.

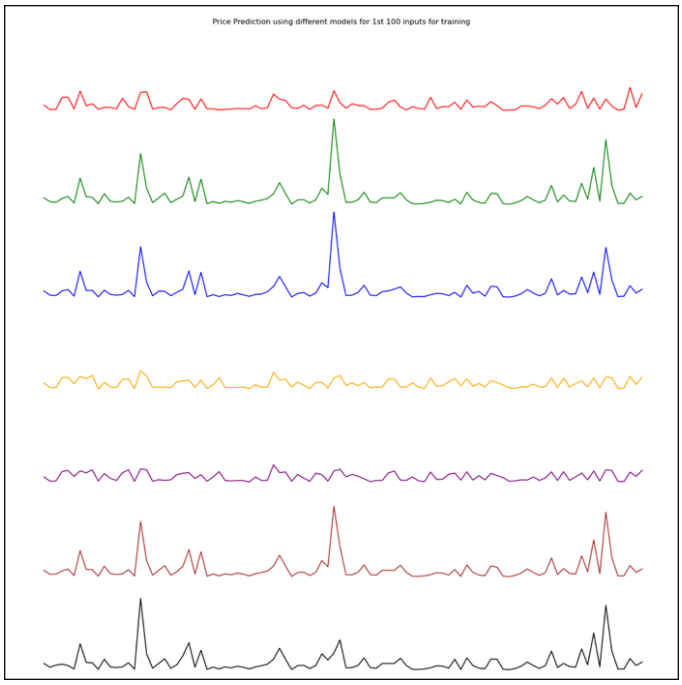

Figure 6: Price Prediction using different models and true values

In our pursuit of identifying the most effective model for price prediction tasks, we embarked on a comprehensive exploration of various neural network architectures and optimization strategies. This included experimenting with different combinations of loss functions and optimizers, as well as evaluating the performance of Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). Despite the versatility and potential of LSTM and CNN architectures, findings revealed that a Simple Neural Network model, configured with appropriate parameters, consistently outperformed other architectures in terms of predictive accuracy and generalization. This underscores the importance of meticulous model selection and customization to match the characteristics of the dataset and the specific requirements of the prediction task. While LSTM and CNN architectures remain viable options for certain applications, results[Table 2] highlights the paramount importance of systematic model evaluation to determine the optimal approach for achieving the desired prediction outcomes.

Model	Train RMSE	Test RMSE
CNN	7.5234	7.8901
LSTM	8.2156	8.0567
Simple NN	6.3012	6.5123

Table 2: RSME for Deep Learning Models

6. HYPERPARAMETER TUNING, CROSS-VALIDATION, AND EVALUATION

In this stage, optimization of models through hyperparameter tuning, cross-validation, and evaluation to improve predictive performance for Decision Tree, Random Forest and different errors and optimizers for Simple NN is done.

Decision Tree: We employed Grid Search Cross-Validation to explore various hyperparameter combinations for the Decision Tree model. The parameters tuned included max_depth, min_samples_split, and min_samples_leaf. After tuning, the Decision Tree model was checked for RMSE and R2 score for test data and validation data. Additionally, we visualized the distribution of Cross-Validation Root Mean Squared Errors (RMSE) using box plots [Figure 6]

Random Forest: Similar to the Decision Tree, we utilized Grid Search Cross-Validation to optimize the Random Forest model's hyperparameters. The parameters tuned included n_estimators, max_depth, min_samples_split, and min_samples_leaf. After tuning, the Random forest model was checked for RMSE and R2 score for test data and validation data. Additionally, we visualized the distribution of Cross-Validation Root Mean Squared Errors (RMSE) using box plots [Figure 6]

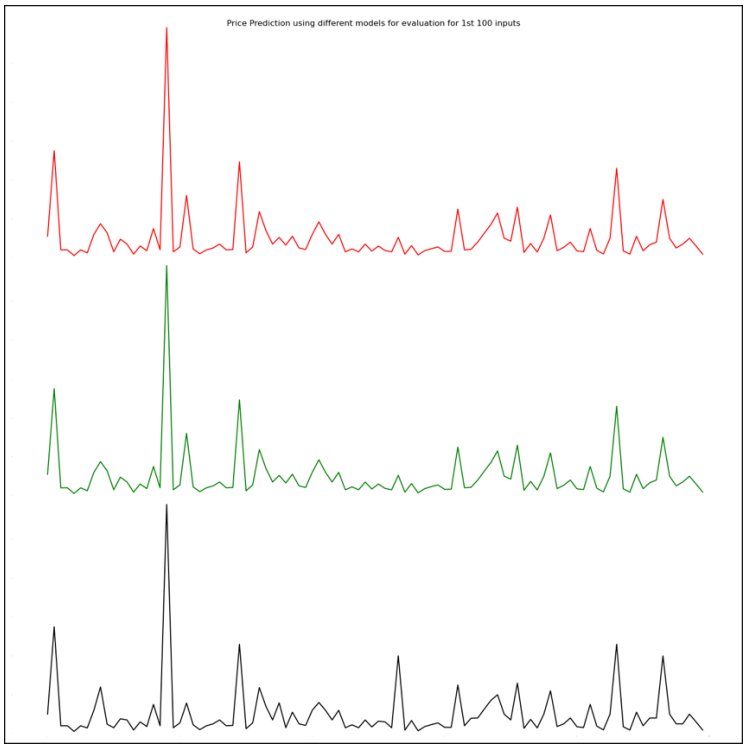

Figure 7: Price Prediction using decision tree and random forest

Simple NN: The SNN model was fine-tuned through various combinations of loss functions and optimizers, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and Smooth L1 Loss, coupled with Stochastic Gradient Descent (SGD), Adam, and RMSprop optimizers. Each combination was meticulously trained and evaluated using a validation dataset, with early stopping implemented to prevent overfitting. [Table 3]

Model Result: Both the Decision Tree and Random Forest models exhibited robust performance in predicting prices, as evidenced by their low Root Mean Squared Error (RMSE) and high R2 scores. Notably, both tree-based models outperformed the Simple Neural Network (SNN) architecture, underscoring the effectiveness of ensemble methods in capturing complex relationships within the dataset. [Table 3, 4] While both models performed admirably, the Random Forest model showcased slightly superior results across the evaluation metrics.

Loss Function	Optimizer	Validation RMSE	Validation R2 Score
MSELoss	SGD	6.4304	0.1062
MSELoss	Adam	6.5963	0.0595
MSELoss	RMSprop	6.4151	0.1105
L1Loss	SGD	6.8026	-0.0002
L1Loss	Adam	6.6693	0.0386
L1Loss	RMSprop	6.6779	0.0361
SmoothL1Loss	SGD	6.8087	-0.0020
SmoothL1Loss	Adam	6.5871	0.0621
SmoothL1Loss	RMSprop	6.6509	0.0439

Table 3: Evaluation for Simple Neural Network

Model	Best Hyperparameters	Train RMSE	Eval RMSE	Train R2	Eval R2
Decision Tree	‘max_depth':180, 'min_samples_leaf':3, 'min_samples_split': 14	1.564	1.612	0.946	0.944
Random Forest	'max_depth': 200, 'min_samples_leaf': 2, 'min_samples_split': 13, 'n_estimators': 200	1.506	1.612	0.950	0.944

Table 4: Hyperparameter tuning and Evaluation for Decision Tree and Random Forest

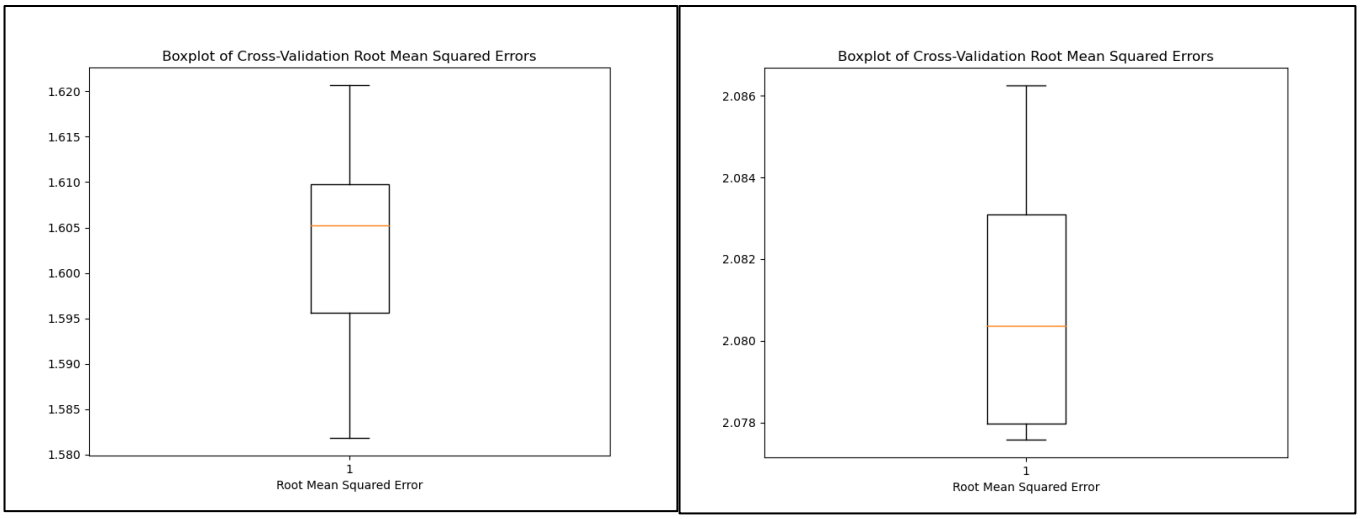

Figure 8: Boxplot of cross validation for Decision Tree and Random Forest Respectively

7. RESULT ACROSS THE MODELS BUILT

MSE and RMSE Evaluation: RMSE provides a straightforward measure of how far, on average, the predictions made by the model deviate from the actual values. This makes it a useful metric for evaluating the overall performance of a regression model**.** Lower values signify enhanced precision in price prediction. After

tuning, the Decision Tree model achieved an MSE of 1.612 and an RMSE of approximately 1.270, showcasing its capability to minimize prediction errors.

R2 Score Assessment: The R-squared (R2) score quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression model. Both the Decision Tree and Random Forest models achieved high R2 scores (0.944), indicating strong predictive capabilities and the ability to explain the variance in prices.

Cross-Validation Techniques: Implementation of k-fold cross-validation ensured model generalization to unseen data, mitigating overfitting and providing reliable performance estimates.

Hyperparameter Tuning: Fine-tuning parameters like 'max_depth', 'min_samples_leaf', and 'min_samples_split' significantly influenced model performance, with 'GridSearchCV'[8] facilitating systematic exploration of the hyperparameter space.

Simple Neural Network (SNN) Evaluation: Despite its simplicity, SNN performance, assessed using MSE, RMSE, and R2 score metrics, varies based on dataset characteristics making it essential to evaluate alongside other models for comprehensive understanding and effectiveness in price prediction tasks.

Model Comparison: Random Forest outperformed other models, particularly in handling peak values, as evidenced by its lower MSE and RMSE on the evaluation set. Among the models examined, Random Forest demonstrated superior performance, particularly in accurately predicting peak values which can be seen in Figure 7. This robust performance underscores the efficacy of Random Forest in capturing complex relationships within the dataset, thereby enhancing its predictive capabilities for extreme price fluctuations. This observation underscores the potential utility of Random Forest in scenarios where precise forecasting of peak values is critical for decision-making processes

By leveraging these performance measures and evaluation strategies, we can make informed decisions regarding model selection, parameter tuning, and ultimately optimize the predictive accuracy of price prediction models for your dataset.

8. CONCLUSION

In this study, exploration into machine learning and deep learning models for forecasting product prices within the retail sector, with a focus on supermarkets, yielded insightful findings. As seen, the decision tree and random forest models showcased exceptional performance among traditional machine learning techniques, boasting impressive efficiency and robust generalization capabilities. Additionally, within the realm of deep learning, a simple neural network architecture emerged as the leading contender, outperforming convolutional neural networks (CNNs) and long short-term memory (LSTM) networks in terms of effectiveness. Through rigorous fine-tuning of hyperparameters and the implementation of cross-validation techniques, we provided actionable insights for retailers seeking to refine pricing strategies and optimize inventory management practices. Our study underscored the importance of integrating both traditional and deep learning methodologies to develop resilient price prediction models capable of adapting to the dynamic retail landscape and meeting consumer demands adeptly. Notably, among these methodologies, random forest emerged as the most robust performer, exhibiting superior predictive capabilities across various metrics. Looking ahead, future research endeavours could explore the integration of ensemble learning methods and reinforcement learning frameworks to further enhance predictive accuracy and dynamic pricing strategies. Moreover, leveraging external data sources, advanced feature engineering techniques, and scalable computing infrastructure holds the potential to unlock new insights and capabilities, paving the way for more sophisticated and adaptable pricing strategies in the future.

REFERENCES

[1] F. Li, H. Zhou, M. Liu and L. Ding, "A Medium to Long-Term Multi-Influencing Factor Copper Price Prediction Method Based on CNN-LSTM," in IEEE Access, vol. 11, pp. 69458-69473, 2023, doi: 10.1109/ACCESS.2023.3288486.

[2] H. Warnakulasooriya, J. Senarathna, P. Peiris, S. Fernando and D. Kasthurirathna, "Supermarket Retail – Based Demand and Price Prediction of Vegetables," 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2020, pp. 308-309, doi: 10.1109/ICTer51097.2020.9325451. keywords: {Supply chains;Blockchain;Forecasting;Predictive models;Mathematical model;Physical layer;Production;supply chain management;machine-learning;deep-learning;blockchain;weighted sum ranking method},

[3] C. Wang and Q. Gao, "High and Low Prices Prediction of Soybean Futures with LSTM Neural Network," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2018, pp. 140-143, doi: 10.1109/ICSESS.2018.8663896. keywords: {Predictive models;Market research;Neural networks;Contracts;Logic gates;Measurement;Time series analysis;soybean futures;high prices;low prices;price prediction;LSTM neural network},

[4] X. Yi, Y. Huang, X. Zhang, Z. Liu and H. Wang, "Genetic Algorithm Based Optimization Solution for Supermarket Replenishment and Pricing Strategies," 2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE), Changchun, China, 2023, pp. 1280-1286, doi: 10.1109/ICEACE60673.2023.10442788.

[5] Fanwei Zhu, Wendong Xiao, Yao Yu, Ziyi Wang, Zulong Chen, Quan Lu, Zemin Liu, Minghui Wu, and Shenghua Ni. 2022. Modeling Price Elasticity for Occupancy Prediction in Hotel Dynamic Pricing. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22). Association for Computing Machinery, New York, NY, USA, 4742–4746. https://doi.org/10.1145/3511808.3557646

[6] D. McAlinden, "Time Series UK Supermarket Data," Kaggle, Available: https://www.kaggle.com/datasets/declanmcalinden/time-series-uk-supermarket- data?select=All_Data_Tesco.csv.

[7] "StandardScaler - scikit-learn 0.24.1 documentation," Scikit-learn, Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[8] "GridSearchCV - scikit-learn 0.24.1 documentation," Scikit-learn, Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[9] V. Khandelwal, A. K. Chaturvedi and C. P. Gupta, "Amazon EC2 Spot Price Prediction Using Regression Random Forests," in IEEE Transactions on Cloud Computing, vol. 8, no. 1, pp. 59-72, 1 Jan.-March 2020, doi: 10.1109/TCC.2017.2780159.