International Journal of Innovative Research in Engineering
Volume 7, Issue 2 (March - April 2026), PP: 280-285.
https://www.doi.org/10.59256/ijire.20260702034
www.theijire.com
ISSN No: 2582-8746

Sales Forecast Prediction Using Machine Learning

Md.Asad Meraj [1], Dr.Hazique Aetesam[2]
[1,2]Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Patna Campus, Bihar, India.

How to cite this paper:
Md.Asad Meraj[1], Dr.Hazique Aetesam[2] "Sales Forecast Prediction Using Machine Learning", IJIRE-V7I2-280-285.

Abstract: *Sales forecasting is a vital business activity because it assists businesses in planning, managing inventory as well as decision-making. Proper forecasting of sales helps companies to minimize losses, prevent overstocking or understocking and enhance customer satisfaction. The primary goal of this project is to create a machine learning-based system that is capable of forecasting future sales based on the previous data and give meaningful information to use in business planning. The data in this project is historical sales records that have details of various stores, items and their daily sales per store over time. The raw data is not usable to model train and therefore a data preprocessing step is done. In this step, the dataset is cleaned, and the date column is converted into a proper datetime format. Based on this date column, year, month and day are extracted as new features that would assist the model to learn about time based trends in sales. Such characteristics are significant as sales are usually determined by seasonal and periodic patterns. In order to enhance the performance of the model further, the feature engineering techniques are implemented. New features are developed like lag values (past sales information) and rolling mean (mean sales in a period). These attributes assist the model to learn the effect of the past sales on the future sales which is quite significant in predicting the problems. Once feature engineering has occurred, missing values that are created in the process are properly dealt with to maintain data quality. To ease the prediction process, the continuous values of sales are transformed into a categorical variable i.e. low, medium and high. This is possible due to this transformation that enables the application of machine learning algorithms based on classification to predict. In this project, four different models are implemented, including Random Forest, XGBoost, K- Nearest Neighbors (KNN), and Logistic Regression. To compare various methods and to realize which model works better in sales prediction, these models are chosen. The data is separated into the training and testing sets in such a way that the models can be trained on one half of the data and performance is analyzed on the unknown data. The various metrics used to measure the performance of each model include accuracy, confusion matrix and ROC curve. Such evaluation methods assist in the determination of the performance of the models and the extent to which they can categorize sales in various groups. The given project shows the effectiveness of the application of machine learning methods to analyze previous sales records and create a forecasting system. These systems have the potential to guide businesses towards making superior decisions, better inventory management and future strategies. Overall, this study highlights the importance of data-driven approaches in solving real-world business problems.*

Key Words: *K-Nearest Neighbors, Machine Learning, Predictive Analytics, Random Forest, Sales Forecasting, Time Series Analysis, XG Boost.*

I.INTRODUCTION

Sales forecasting is a significant aspect in business analysis that assists organizations to forecast their future sales and make more informed decisions. It plays a key role in inventory management, demand planning, and improving overall business Performance. Proper forecasting enables the firms to minimize losses, prevent overstocking or understocking and satisfy customer demand effectively.

Historically, sales forecasting has been performed by traditional statistical techniques which were not well suited to large and complex data. Machine learning is a potent tool that can evaluate large volumes of data and reveal the hidden patterns and trends with the development of technology. Such models have the ability to acquire past information and make more precise and valid predictions.

The main aim of this project is to develop a machine learning-based system for sales forecast prediction using historical data. The dataset utilized includes the sales of the stores and items over periods. With the help of this data, one can extract meaningful patterns to comprehend the change of sales over time.

Several machine learning models are applied in this project to predict the categories of sales: these include, but are not limited to, Random Forest, XGBoost, K-Nearest Neighbors, and Logistic Regression. These models are then compared in terms of performance to come up with the best performing model.

This project shows how machine learning methods can be used in a real-life business to enhance the accuracy of the forecasts and contribute to better decision-making.

II.MATERIAL AND METHODS

1. Dataset Description

The data utilized in this project is a retail sales data that is sourced online. It has past sales history of various products in various stores over time. Each record is the sales of that day in a specific store.

The data set has the following attributes:
Date Store ID Item ID Sales
The data is huge and includes thousands of records, which is why it can be used to implement machine learning methods.

2. Data Preprocessing

Another step that is significant in preparing the dataset to be analyzed and trained on the model is the preprocessing of data. The following preprocessing steps were carried out in this project:

The date column was changed to the format of datetimes. The date column was used to derive new features like year, month and day. After extraction of features, the original date column was dropped. Missing values that were a result of feature engineering were treated by dropping row nulls.
These processes aid in transforming raw data into a structured format that could be used in machine learning models.

3. Feature Engineering

There was feature engineering to enhance the performance of the models by extracting trends in the data. The following features were developed:

Lag Features: Lag features like lag 1, lag 2, lag 7, were developed to show the previous sales. Rolling Mean: A moving average of sales over a predetermined window was computed to reflect the trends.
These characteristics assist the model to know the effect of previous sales on the future sales.

4. Data Transformation

The continuous sales values were transformed into categorical classes to make the problem of prediction easier. The sales were categorized into three:
Low Sales Medium Sales High Sales
The conversion enables the classification algorithms to be used in prediction.
Model Implementation

This project implemented and compared four machine learning models:
Random Forest XGBoost K-Nearest Neighbors (KNN) Logistic Regression

These models were chosen to make comparisons between various types of algorithms, such as ensemble methods and conventional classification models.

5. Model training and testing

The train-test split method was used to divide the dataset into training and testing sets. Training data was used to train the models Performance of models was tested using data. This strategy helps to make sure that the model is tested on unknown data.

6. Evaluation Metrics

The models performance was measured in terms of the following metrics:

- **Accuracy:** Percentage of correct predictions.
- **R2 Score:** The measure of the model which explains the data.
- **Mean Absolute Error (MAE):** Indicates the mean error in prediction.
- **Confusion Matrix:** Presents the right and wrong predictions.
- **ROC Curve:** Tests the performance of a model at varying thresholds.

III.RESULT

1. Model Performance Comparison

	Model	Accuracy	R2 Score	MAE
0	Random Forest	0.870717	0.806158	0.129283
1	XGBoost	0.875826	0.813819	0.124174
2	KNN	0.834484	0.751561	0.165576
3	Logistic	0.869534	0.804089	0.130532

Model Comparison Table

Accuracy, R 2 score and Mean Absolute Error (MAE) were metrics used to determine the performance of the various machine learning models. The table below summarises the results:

Based on the findings, we note that XGBoost had the highest accuracy and minimum error compared to all the models and is thus the best performing model. Random Forest worked well as well, and KNN was relatively poorly-performing.

2.Confusion Matrix Analysis

Confusion matrices were used to evaluate the classification performance of each model. The diagonal values signify correct prediction and off-diagonal ones signify incorrect prediction.

The XGBoost and the Random Forest models demonstrate a high degree of diagonals, which implies high accuracy. KNN and Logistic Regression have a minor misclassification than other models.

Random Forest:

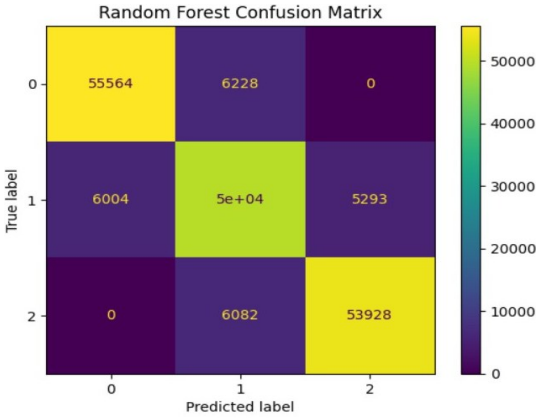

Fig 1: Random Forest confusion matrix

XG Boost :

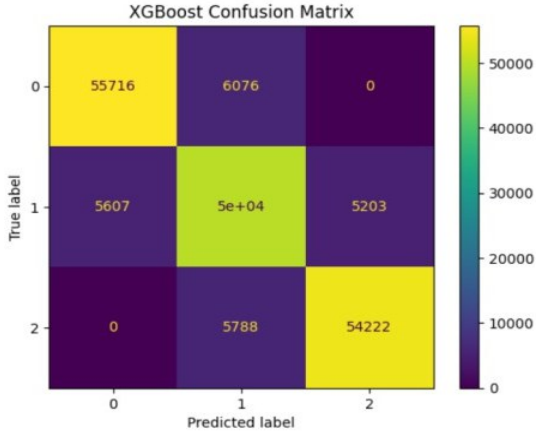

Fig2: XG Boost Confusion Matrix

KNN

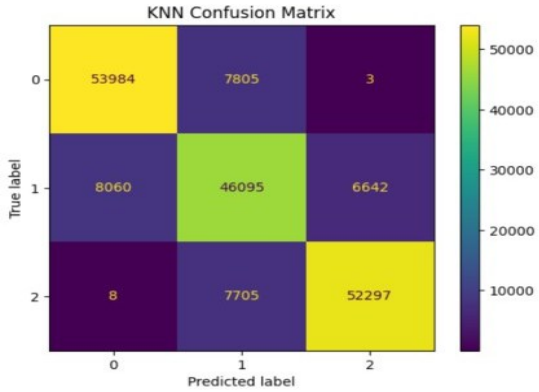

Fig3: KNN Confusion Matrix

Logistic Regression:

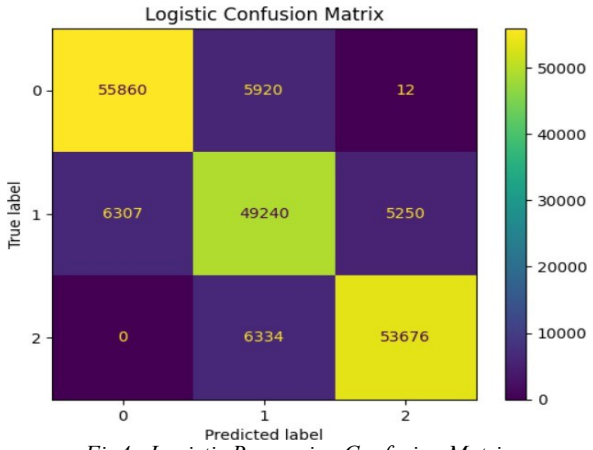

Fig4: Logistic Regression Confusion Matrix

3. ROC Curve Analysis

To assess the performance of the models, ROC (Receiver Operating Characteristic) curves were plotted. The curves that are nearer to the upper-left side are more indicative of better performance.

The Area Under Curve (AUC) of the models was high, which means that the models have a strong capability of classification.

XG Boost had the highest ROC performance of all the models

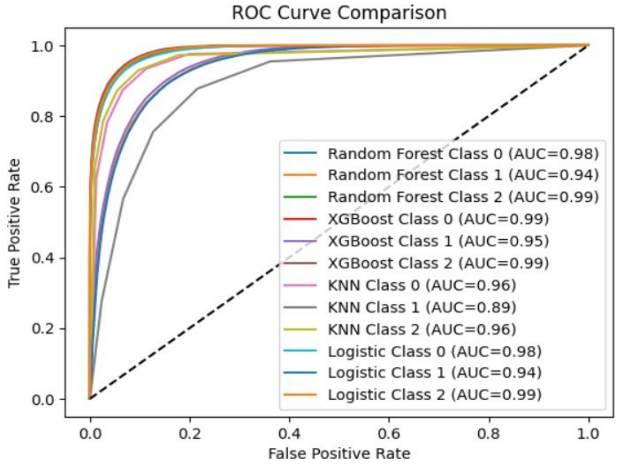

Fig5: ROC Curve

4. Sales Trend Analysis

The sales trend graph indicates the variation in sales. One can see that sales have an upward trend with fluctuations, which means that there are seasonal changes.

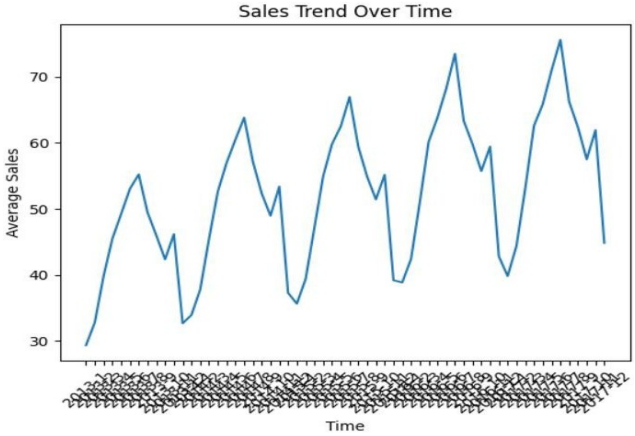

Fig 5: predicted Sales

5. Future Sales Prediction

The predictive model was trained to make future sales predictions, given the input features. The result as predicted indicates that the sales will be in a middle range of demand.
This shows that one can make future projections and business plans using the model.

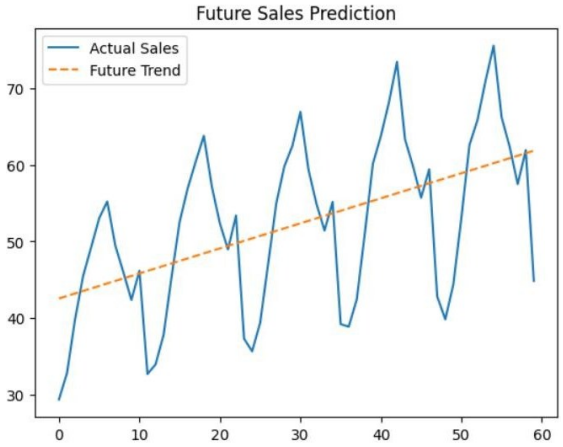

Fig 7: Future Prediction Graph

IV.DISCUSSION

Various machine learning models were used in this project to forecast sales categories using past data. The results indicate that all models tended to perform fairly well with evident differences in their performance. XGBoost had the greatest accuracy and minimum error compared to all other models. The reason is that XGBoost is a more sophisticated form of ensemble learning which uses various decision trees and can better capture important patterns in the data. Random Forest also works quite well as it minimizes overfitting and can work with the large data sets. K-Nearest Neighbors (KNN) performed relatively worse, on the other hand. This is primarily due to KNN being data scaling sensitive and it might not perform well when dealing with a large and complex dataset. Logistic Regression also fared slightly worse than ensemble models because it is based on the assumption that there is a linear relationship between features and output, which is not necessarily applicable to such data. The use of feature engineering was very significant in enhancing the performance of the models. Lag features and rolling mean were used to enable the models to know the past sale pattern and trend. These characteristics enabled the models to model time based relationships as needed in sales forecasting. The obtained results of the confusion matrices indicated that most of the predictions were correctly identified, particularly in the case of XGBoost and Random Forest models. It was also confirmed by the ROC curve analysis that these models can be highly classified in terms of high AUC values. The sales trend analysis indicated that the trend followed by sales shows some fluctuations over time, which shows that there are seasonal changes. This justifies the use of time-based features in the model. On the whole, the findings indicate that machine learning models, in particular, ensemble models are efficient in addressing sales forecasting issues and can be used to make credible predictions in the context of the real-life business environment.

V.CONCLUSION

This was a project aimed at creating a machine learning based system of sales forecast prediction with the help of past sales records. The quality of the data and important patterns were enhanced with the use of different preprocessing and feature engineering methods. Various machine learning models such as Random Forest, XGBoost, K-Nearest Neighbors and Logistic regression were applied and compared. XGBoost was the most appropriate model to use in this problem, as it was the most accurate and the lowest error model of all models. As the findings of this paper indicate, machine learning methods can be used to efficiently analyze previous sales data and make accurate predictions. Time-based features like lag values and rolling averages were important because they greatly enhanced the performance of the model. The system can also be helpful to business in making improved decisions regarding the inventory management, demand planning, and general business efficiency. It exhibits how machine learning can be applied in real-life issues.

References

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
2 . Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD Conference.
3 . Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13(1), 21–27.
4. Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. Wiley.
5 . James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
6. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly.
7 . Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
8. Kaggle. (2018). Demand Forecasting Dataset. Retrieved from https://www.kaggle.com
9. Brownlee, J. (2017). Machine Learning Mastery with Python. Machine Learning Mastery.
10. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice.
11. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

12. Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
13. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
14. Han, J., Kamber, M., & PeiMitchell, J. (2011). Data Mining: Concepts and Techniques. Elsevier.
15. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
16. Scikit-learn Documentation. (2023). https://scikit-learn.org
17. XGBoost Documentation. (2023). https://xgboost.readthedocs.io