1. Introduction:

In the realm of data science, data mining is all about finding patterns and useful information from large amounts of data. One important task in data mining is predictive modeling, which helps us make predictions about the future based on the past data. This predictive model plays a crucial role in various industries such as retails, where predicting sales accurately can significantly lead to better and enhanced decision making and operational efficiency. Sales prediction is an important part of modern business intelligence [1,2,3]. It can be a complex problem, especially in the case of lack of data, missing data, and the presence of outliers. Sales can be considered as a time series. In this project we will use data mining techniques to predict future sales of different product types in stores. The dataset we are working with comes from KAGGLE and is named Store Sales - Time Series Forecasting dataset. This dataset includes important details like store IDs, product families, sales numbers, holidays, oil prices and store transactions. All of this information can significantly help us build a robust predictive model.To address this problem, we will use a machine learning model called Random Forest Regressor. This model is great for handling structured data and can also find complex patterns in the data. By looking at past sales information, the model will help predict future sales, enabling better inventory management and resource allocation for each store. The objective of this project is to provide useful insights that will help stores make better decisions about how much stock to keep, when to run promotions, and how to improve their supply chain. As a result, stores can minimize stock shortages, reduce overstocking, and ultimately increase their profitability by ensuring the right products are available at the right time.

2. Data Preprocessing:

2.1. Data Cleaning:

The dataset was thoroughly scanned to fish out missing values and contradictions. Fortunately, there were no missing values in the dataset, which simplified things. However, during the analysis of the sales column, we identified a huge range in sales values, with some sales going up to 124,000 units. This is an indication that there may be potential outliers, which would be detrimental to model performance. Further investigation revealed that the distribution of sales was right-skewed. In the following steps, either capping methods or data transformation will be used to reduce the effect of any one outlier on the model.

2.2. Data Transformation:

The following key transformations were performed at this stage in order to prepare the dataset for modeling. One-hot encoding for the family column was also done, representing product categories, into a numerical format so that it could be understood by machine learning models. We have further created time-based features by extracting month, day, and day of the week from the date column to get seasonal trends in sales. Besides, this is going to introduce a lag feature that will represent the sales of the previous day to take possible temporal dependencies into account. Then the date column was dropped as it had no further use.

2.3. Exploratory Data Analysis:

A comprehensive EDA was conducted to examine the structure and distribution of the data. The sales distribution was visualized in the histogram of quantity, which indicated a highly skewed distribution: most sales are at the lower end and a few are at the extreme. Grouping with store, family, and date allowed us to make some aggregates on sales and observe some trends. Time series plots were made for the detection of any seasonal patterns or anomalies. The EDA clearly showed that sales trends are very different across stores and product categories, thus feature engineering, especially in terms of categorical encoding and lag features, is warranted.

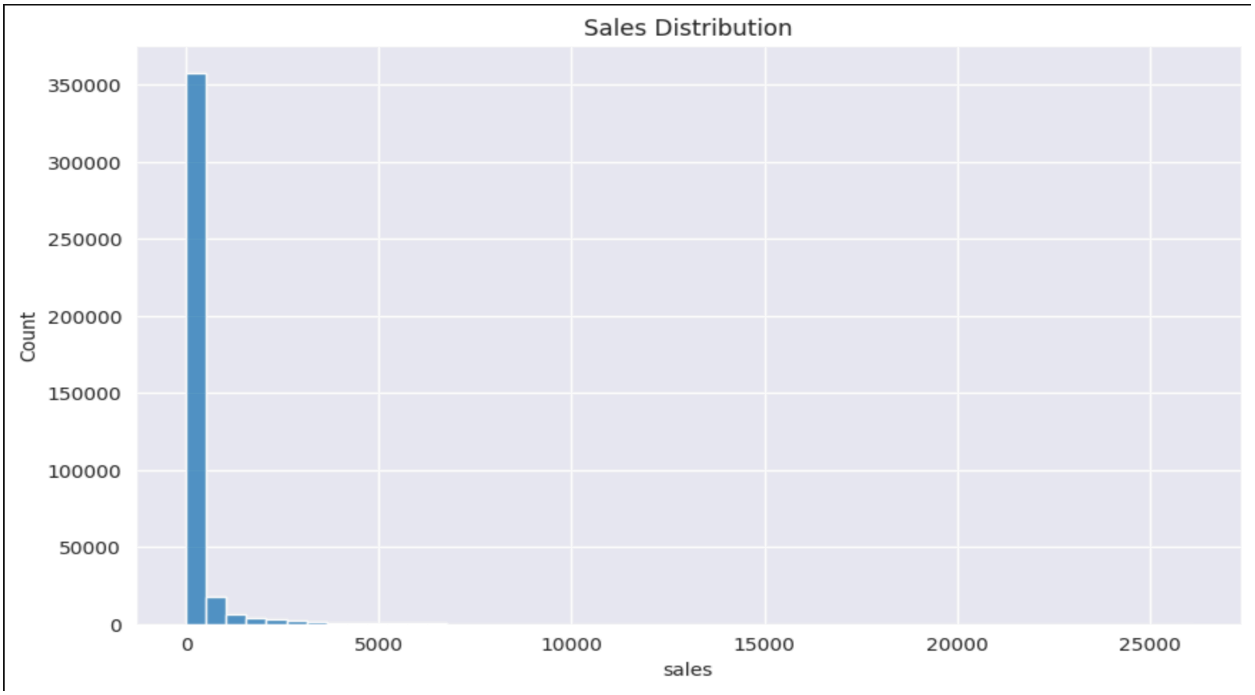

Figure 1: the figure shows the distribution of sales

1. Model Building:

The Random Forest Regressor was selected for this project to predict future sales for each product family across all stores. Random Forest is an ML technique that particularly performs very well in making predictions with structured data. It makes predictions based on the multiple decision trees built during the training phase and helps subdue overfitting and improves Performance.

3.1. Train-Test Split

We performed a train-test split such that 80% of the data remains for training, while 20% remains for testing. It is done in a time order so that the temporal nature of data can be preserved, and the model has information only from the past to predict. These methodologies go a long way in simulating actual forecasting, where future sales are estimated based on historical data.

3.2. Model Implementation

The implementation of Random Forest Regressor used the number of trees (estimators) at 100, which is standard to balance performance and computational cost. We paid attention to the use of mean absolute error (MAE) as the principal evaluation metric, since it clearly shows the model's predictive accuracy through the quantification of the average prediction error in terms of sale values.

Metric	**Value**
Mean Absolute Error (MAE):	104.39686634987822

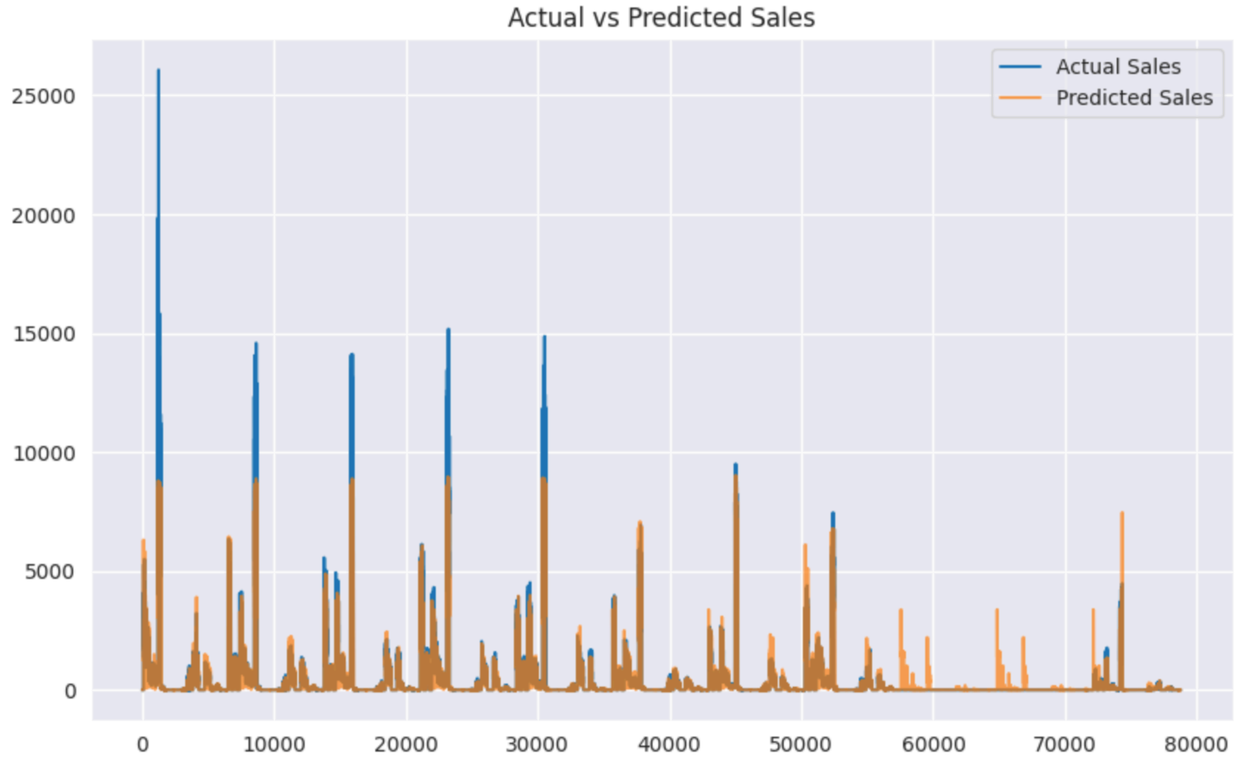

Figure 2: the figure shows distribution of actual sales and predicted sales

2. **Model Evaluation:**

After the model was built using the Random Forest Regressor technique, it was crucial to evaluate its performance with the appropriate evaluation metrics. The evaluation of the model was done using Mean Absolute Error (MAE), which quantifies the average difference between the predicted and actual sales values. We chose MAE because it provides an easily interpretable metric that directly reflects the magnitude of errors in our predictions. Additionally, it penalizes all errors equally, making it suitable for our sales forecasting task, where under-predicting and over-predicting by the same amount have similar consequences.

Metric for Evaluation: Mean Absolute Error (MAE)
The Mean Absolute Error on the test data was 104.47. This means that, on average, the model's predictions deviated from actual sales by about 104.47 units. Given the large range of sales in the dataset, this margin of error is reasonable but indicates that further tuning is needed to improve accuracy.

Why MAE?
We chose MAE for its simplicity and clear interpretability. Unlike metrics like Mean Squared Error (MSE), which exaggerate larger errors due to squaring, MAE provides a fairer evaluation across all predictions, ensuring that a few large errors do not disproportionately affect the overall model evaluation.

Insights from the Assessment
The model effectively captured key sales patterns, such as seasonality and store-level differences. However, challenges like a skewed sales distribution and outliers likely contributed to the overall error. The right-skewed distribution means most sales were low, but a few stores or product families had very high sales, which might have inflated the MAE. Techniques like log transformation could help reduce the impact of outliers and improve the model's performance.

The model's performance on test data also suggests that temporal features, such as month, day, and lagged sales, positively contributed to the prediction, allowing it to capture time-based trends and cyclic sales fluctuations. However, further improvements could be achieved through hyperparameter tuning, such as adjusting the number of decision trees or exploring alternative models like Gradient Boosting or XGBoost.

Visual Inspection
In addition to MAE, we performed a visual inspection of actual vs predicted sales. A graph comparing real sales figures with the model's predictions showed that the model captured the overall trend but missed some high and low points in the sales data. These deviations could be due to factors not included in the dataset, such as promotional events, sudden demand changes, or other market dynamics.

Insight from model results:
The Random Forest Regression provided valuable insights into store sales patterns across different product families. The model's ability to predict future sales based on historical data has significant real world implications, particularly in retail industries, where accurate forecasting plays a crucial role in optimizing operations and improving profitability.