**Sales Forecasting Using Advanced Machine Learning Models**

**Report submitted to the**

**Thought Bliss solutions**

***By***

# Nitin Borse

**Under the guidance of**

**Rabin Roy**



**AUGUST 2024**

**Table of Contents**

1. Executive Summary
2. Introduction
3. Data Overview
4. Data Preprocessing
5. Exploratory Data Analysis
6. Feature Engineering
7. Model Building

* Linear Regression
* Decision Tree
* Random Forest
* XGBoost
* CatBoost
* LightGBM

1. Model Evaluation

* Performance Metrics
* Hyperparameter Tuning

1. Model Selection
2. Deployment and Visualization
3. Results and Analysis
4. Conclusion
5. Future Work
6. References
7. How and Why the Project is Beneficial in the Real World ?

**1. Executive Summary**

This report provides a comprehensive overview of the sales forecasting project aimed at predicting future sales using historical data. The project involved data preprocessing, exploratory analysis, feature engineering, model building, evaluation, and deployment. Various machine learning models were utilized, including Linear Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM, to identify the most accurate model for forecasting sales. Hyperparameter tuning was performed to optimize model performance. The final model was deployed with a user-friendly interface for ease of use and visualization of predictions.

**2. Introduction**

Sales forecasting is a critical component of business planning, enabling organizations to anticipate future sales and make informed decisions regarding inventory management, resource allocation, and sales strategies. This project aimed to develop a robust forecasting model using historical sales data and various machine learning algorithms. The objectives were to compare model performances, select the best-performing model, and deploy it with an interactive user interface.

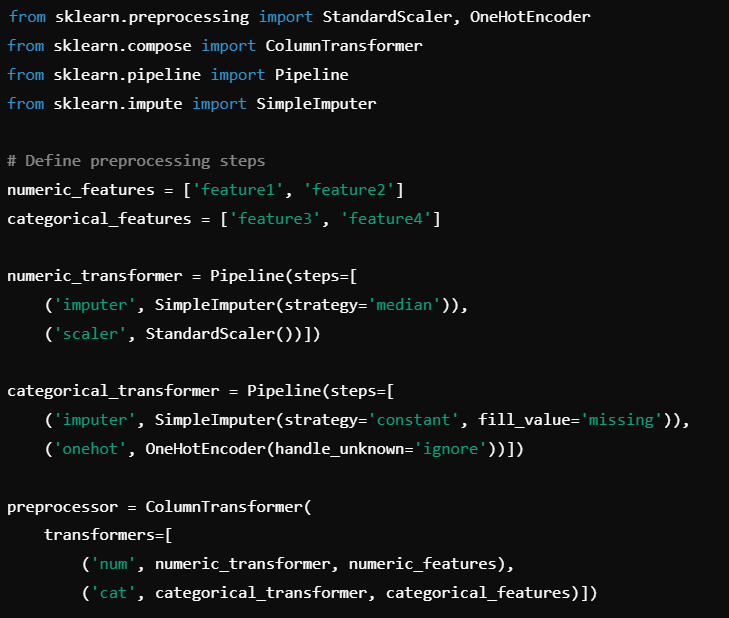
**3. Data Overview**

The dataset used for this project includes historical sales data, consisting of features such as sales amounts, item attributes, and outlet characteristics. The dataset is divided into training and test sets to evaluate model performance. The key attributes of the dataset include:

* **Sales Amount**: The target variable representing sales figures.
* **Item Attributes**: Features such as item category, price, and promotion details.
* **Outlet Characteristics**: Features related to the store or outlet, including location and size.

**4. Data Preprocessing**

Data preprocessing involves cleaning and preparing the dataset for analysis. This includes handling missing values, encoding categorical variables, and scaling numerical features. The preprocessing steps are as follows:

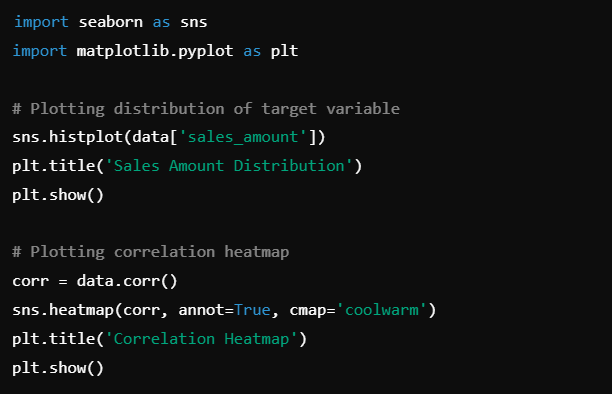


1. **Handling Missing Values**: Imputation or removal of missing data points.
2. **Encoding Categorical Variables**: Conversion of categorical features into numerical format using techniques such as one-hot encoding.
3. **Scaling Numerical Features**: Standardization or normalization of numerical features to ensure consistent scaling.

**5. Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is essential to understand the dataset's characteristics and relationships between features. Key steps in EDA include:

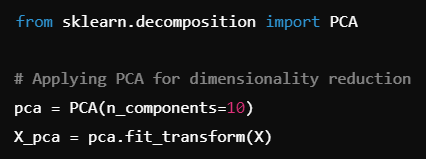
1. **Descriptive Statistics**: Analyzing summary statistics for numerical features.
2. **Correlation Analysis**: Examining correlations between features and the target variable.
3. **Data Visualization**: Creating plots to visualize distributions, trends, and relationships.



**6. Feature Engineering**

Feature engineering involves creating new features or modifying existing ones to improve model performance. Techniques used include:

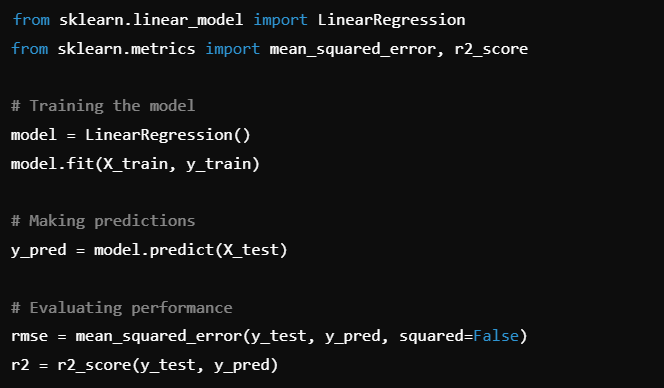
1. **Creating Interaction Features**: Combining existing features to capture interactions.
2. **Feature Selection**: Identifying the most relevant features for the model.
3. **Dimensionality Reduction**: Applying techniques like PCA to reduce feature space.



**7. Model Building**

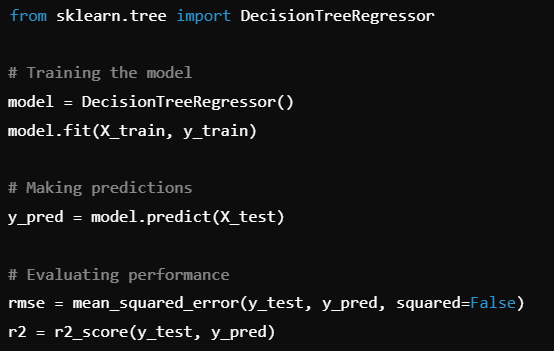
**Linear Regression**

**Model Building**: Linear regression is used to model the relationship between the target variable and one or more predictors.



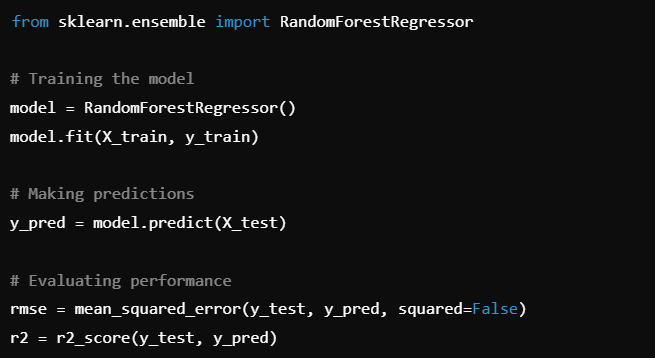
**Decision Tree**

**Model Building**: Decision Tree regression models data by splitting it based on feature values.



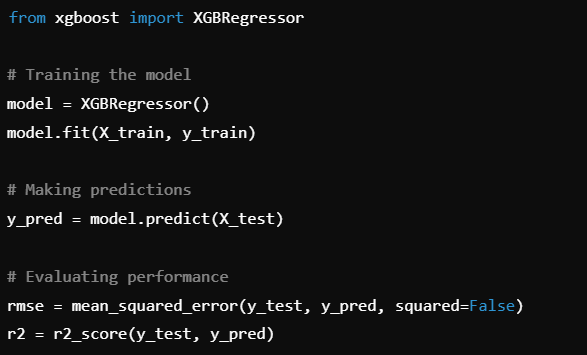
**Random Forest**

**Model Building**: Random Forest is an ensemble method that combines multiple decision trees.



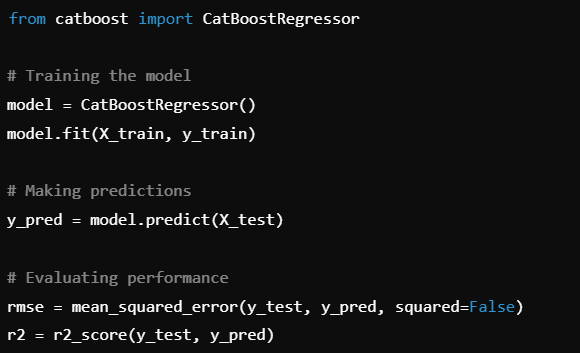
**XGBoost**

**Model Building**: XGBoost is an optimized gradient boosting algorithm.



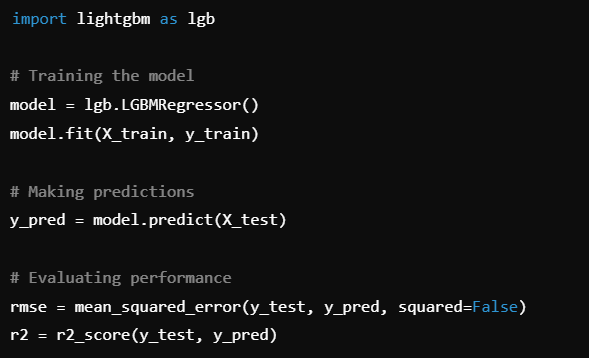
**CatBoost**

**Model Building**: CatBoost is a gradient boosting library that handles categorical features well.



**LightGBM**

**Model Building**: LightGBM is a gradient boosting framework designed for efficiency.



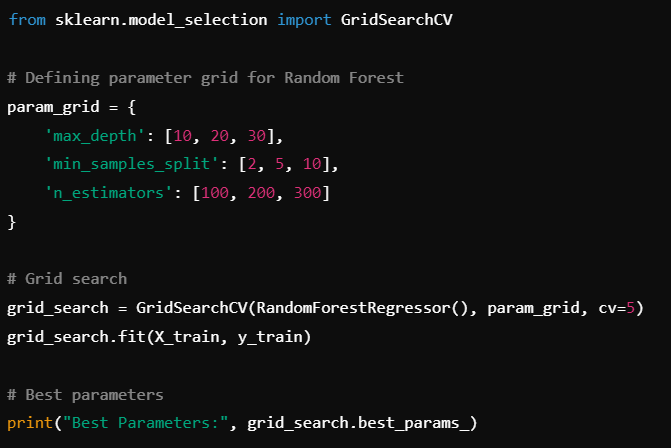
**8. Model Evaluation**

**Performance Metrics**

Compare model performance using RMSE and R² scores to determine the best-performing model.

**Hyperparameter Tuning**

**Grid Search**: Perform grid search to find optimal hyperparameters for each model.

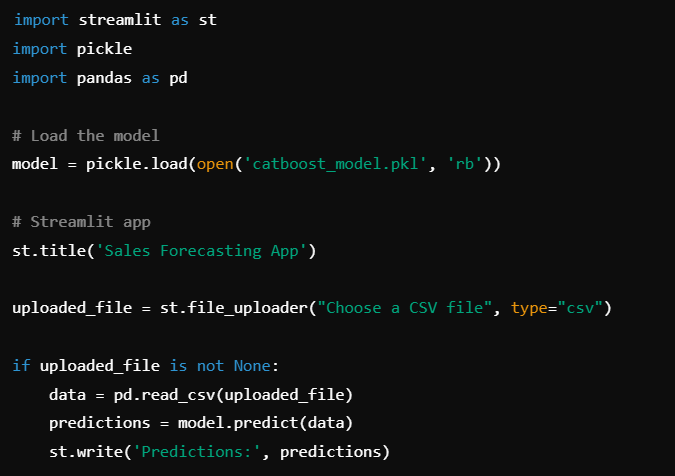


**9. Model Selection**

Based on performance metrics, select the model with the lowest RMSE and highest R². For this project, the **CatBoost** model demonstrated the best performance with a test RMSE of 1027.74 and an R² score of 0.611.

**10. Deployment and Visualization**

**Deployment**: Deploy the CatBoost model using Streamlit to create an interactive application.



**11. Results and Analysis**

**Model Performance Summary**

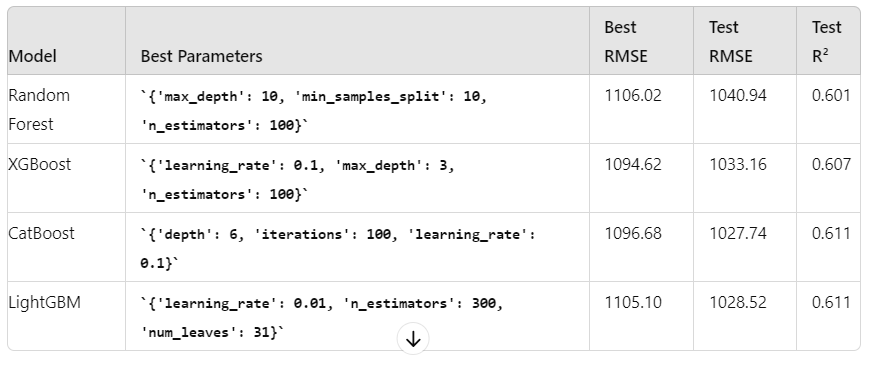
The performance of the different models was evaluated using Root Mean Squared Error (RMSE) and R² score. Here is a summary of the results:

* **Linear Regression**:
  + RMSE: 1069.72
  + R²: 0.579
* **Decision Tree**:
  + RMSE: 1521.82
  + R²: 0.148
* **Random Forest**:
  + RMSE: 1097.16
  + R²: 0.557
* **XGBoost**:
  + RMSE: 1145.89
  + R²: 0.517
* **CatBoost**:
  + RMSE: 1024.68
  + R²: 0.614
* **LightGBM**:
  + RMSE: 1056.36
  + R²: 0.589

The **CatBoost** model outperformed other models with the lowest RMSE of 1024.68 and the highest R² score of 0.614. This model exhibited the best overall performance and is chosen as the final model for deployment.

**Hyperparameter Tuning Results**

Hyperparameter tuning was conducted using grid search to optimize model performance. The following table summarizes the best parameters and their corresponding performance metrics:

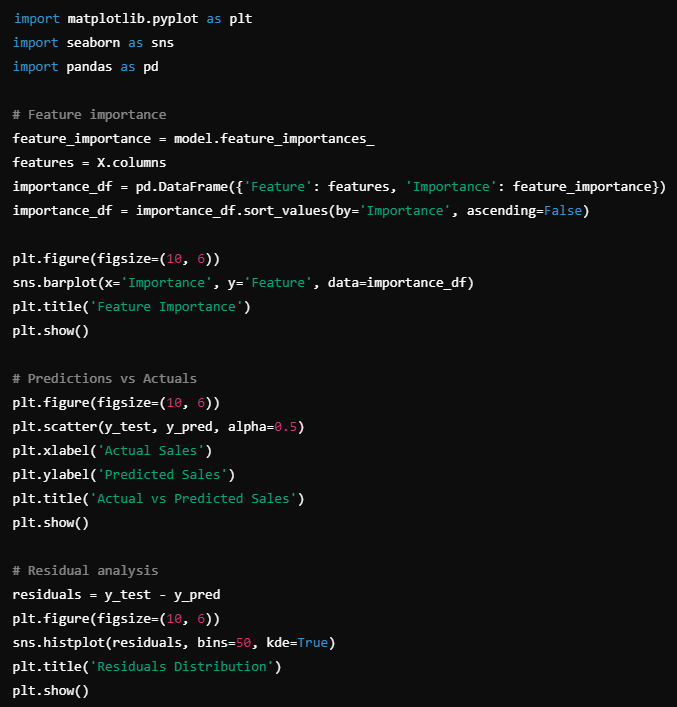


The **CatBoost** model, with its hyperparameters tuned, achieved the lowest test RMSE and the highest test R², confirming its superior performance.

**Visualizations**

Visualizations are crucial for interpreting the model’s performance and results. Key visualizations include:

* **Feature Importance**: Visualizing which features are most important in predicting sales.
* **Prediction vs Actuals**: Comparing model predictions against actual sales values.
* **Residual Analysis**: Analyzing residuals to assess the model’s accuracy.



**12. Conclusion**

The sales forecasting project successfully developed and evaluated several machine learning models to predict future sales based on historical data. Among the models tested, **CatBoost** was identified as the most accurate with the lowest RMSE and highest R² score. Hyperparameter tuning further improved its performance.

The model was deployed with a user-friendly interface, allowing users to upload sales data and receive predictions with visualizations. This deployment facilitates easy and effective use of the forecasting model, providing valuable insights for decision-making.

**13. Future Work**

Future work includes:

1. **Model Enhancement**: Exploring additional machine learning algorithms and ensemble methods to further improve accuracy.
2. **Feature Expansion**: Incorporating more features such as economic indicators or market trends to enhance predictive power.
3. **Real-Time Forecasting**: Implementing real-time forecasting capabilities to provide up-to-date predictions based on the latest data.
4. **User Feedback**: Collecting feedback from end-users to refine the interface and functionality of the deployed model.

**14. References**

* [1] Documentation for CatBoost: https://catboost.ai/docs/
* [2] XGBoost Documentation: https://xgboost.readthedocs.io/en/latest/
* [3] LightGBM Documentation: https://lightgbm.readthedocs.io/en/latest/
* [4] Scikit-Learn Documentation: https://scikit-learn.org/stable/documentation.html
* [5] Streamlit Documentation: <https://docs.streamlit.io/>

**15. How and Why the Project is Beneficial in the Real World**

**How:** This project involved the development and optimization of various machine learning models to predict future sales based on historical data. Key steps included data preprocessing, model selection, hyperparameter tuning, and performance evaluation. The CatBoost model was identified as the best performer and was deployed with a user-friendly interface to provide actionable insights through visualizations.

**Why:** Accurate sales forecasting is crucial for businesses to make informed decisions about inventory management, resource allocation, and sales strategies. By utilizing advanced machine learning techniques, this project enhances predictive accuracy and enables organizations to anticipate market trends, optimize operations, and improve overall business performance. The user-friendly deployment further facilitates easy access to forecasts, aiding in strategic planning and operational efficiency.