FUTURE SALES PREDICTION

Date:31.10.2023

Team no:3893

Problem Statement:

To address a specific company's challenges with demand uncertainty, we will develop a future sales prediction system. This system will leverage historical sales data, external variables, and advanced predictive analytics to forecast sales accurately. By considering product-level specifics, seasonality, and external factors like promotions and market trends, it will empower the company to optimize inventory management, allocate resources efficiently, and make informed decisions to meet varying customer demands. This solution aims to enhance profitability and customer satisfaction by ensuring the right products are in stock at the right time, mitigating overstock and stockouts.

Introduction:

In the ever-evolving landscape of commerce, the ability to anticipate and plan for future sales is paramount. It requires a deep understanding of market dynamics, consumer behavior, and a range of influencing factors. Future sales prediction, a critical aspect of strategic decision-making, is at the heart of this pursuit. By employing advanced data analytics, modeling, and predictive tools, this endeavor aims to empower businesses with the insights they need to optimize inventory management, marketing strategies, and overall profitability. In the following exploration, we will delve into the methodologies, challenges, and potential benefits of future sales prediction, offering a glimpse into the future of commerce and market competitiveness.

Literature Survey:

1.SALES PREDICTION OF MARKET USING MACHINE LEARNING:

This explores big data's significance in retail, covering its attributes and challenges. It suggests overcoming issues with data management tools and a data-driven culture, offering examples of how big data can enhance retail operations and introduces a data maturity framework for assessment.

2. SALES PREDICTION USING MACHINE LEARNING ALGORITHMS :

This delves into Machine Learning's transformative role in sales and marketing, emphasizing its ability to analyze customer purchasing patterns. It highlights the Random Forest Algorithm's exceptional accuracy at 93.53% and underscores the broader impact of Machine Learning across diverse sectors.

3.PREDICTING THE FUTURE OF SALES: A MACHINE LEARNING ANALYSIS OF ROSSMAN STORE SALES:

This discusses the application of machine learning algorithms for predicting sales trends. It showcases two studies using LightGBM, XGBoost, and LSTM models. The article stresses data quality, feature selection, and model evaluation, highlighting the broader potential of machine learning across industries.

4. SALES PREDICTION USING MACHINE LEARNING TECHNIQUES:

The discussion improved sales forecasting using data mining and machine learning techniques. The authors compare three algorithms and find the Gradient Boost Algorithm to be the most accurate, with 98% overall accuracy. The paper underscores the importance of using machine learning for precise sales predictions, benefiting businesses seeking enhanced forecasting.

5.PREDICTING FUTURE SALES OF RETAIL PRODUCTS USING MACHINE LEARNING:

The discussion employed machine learning for predicting future retail product sales using XGBoost and LSTM. XGBoost outperformed LSTM, attributed to dataset sparsity.
Accurate sales forecasting is crucial for organizations, aiding growth and financial planning. The paper offers valuable insights into this application of machine learning.

Design Thinking Process:

Data Collection:

To enhance future sales prediction in the retail industry, we will employ advanced data collection techniques, including web scraping, to gather diverse retail-related data. This dataset will encompass historical sales data, product metadata, customer reviews, and

external factors affecting sales. Innovative feature engineering methods, including sentiment analysis and keyword extraction, will be applied to extract meaningful insights from both structured and unstructured data sources. This approach aims to provide a robust predictive model for retail sales, enhancing accuracy and enabling informed decision-making for inventory management, demand planning, and overall retail strategy optimization.

Model Selection and Training

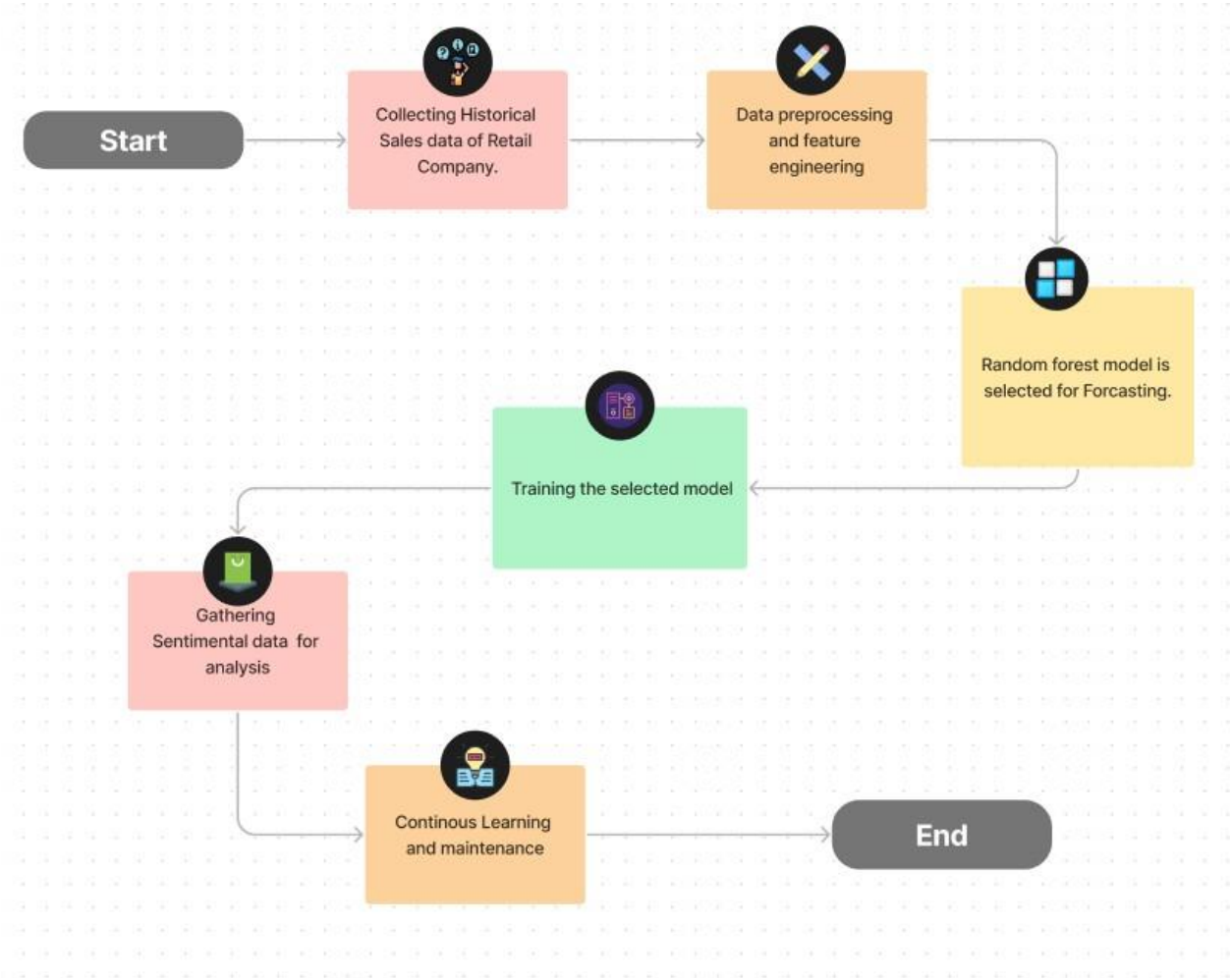
Explore a wide range of machine learning algorithms, including regression, decision trees, to determine the most effective model for Future Sales prediction. Investigate ensemble learning techniques, such as Random Forests and Ada Boosting, to combine the strengths of multiple models and improve prediction accuracy.

Market Sentiment Analysis

To refine future sales predictions for a retail company, we will incorporate external data sources to assess market sentiment and its potential impact on sales. This strategy involves collecting real-time data on social media sentiment analysis and news sentiment within the retail industry. We will integrate this data into the predictive model to quantify the influence of market sentiment on sales. By creating sentiment scores, temporal features, and monitoring market trends, we seek to provide a more accurate and dynamic sales prediction framework. This approach empowers the retail company to adapt its strategies based on real-time market sentiment, optimizing inventory management and resource allocation for improved sales performance.

Continuous Learning

To keep our retail sales prediction model accurate and up-to-date, we'll establish continuous learning. This involves real-time user feedback integration and automated data pipelines for ongoing model retraining. User feedback helps adapt the model to changing preferences, while automated pipelines ensure the use of the latest data. Regular model updates are scheduled, and monitoring maintains performance. This approach keeps the retail company responsive to evolving sales trends and customer needs.



Phases of development:

Phase 1: Importing Dependencies

This phase involves importing necessary Python libraries and modules. These libraries are required for data processing, visualization, and various machine learning tasks.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import AdaBoostRegressor
import chardet
import sklearn.preprocessing as sps
import warnings
warnings.filterwarnings('ignore')
```

Phase 2: Reading and Viewing Data

In this phase, the code reads a dataset from a CSV file named ‘future_sales_prediction.csv' using Pandas.

```
dataset=pd.read_csv("C:\\Users\\student\\Downloads\\future_sales_prediction.
↪csv")
```

Phase 3: Data Preprocessing and Exploration

Here, data preprocessing and exploration is done

```
dataset
```

```
        TV  Radio  Newspaper  Sales
0    230.1   37.8       69.2   22.1
1     44.5   39.3       45.1   10.4
```



```
2     17.2   45.9       69.3   12.0
3    151.5   41.3       58.5   16.5
4    180.8   10.8       58.4   17.9
..     ...    ...        ...    ...
195   38.2    3.7       13.8    7.6
196   94.2    4.9        8.1   14.0
197  177.0    9.3        6.4   14.8
198  283.6   42.0       66.2   25.5
199  232.1    8.6        8.7   18.4

[200 rows x 4 columns]
```

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   TV         200 non-null    float64
 1   Radio      200 non-null    float64
 2   Newspaper  200 non-null    float64
 3   Sales      200 non-null    float64
dtypes: float64(4)
memory usage: 6.3 KB
```

```
dataset.describe()
```

```
               TV       Radio   Newspaper       Sales
count  200.000000  200.000000  200.000000  200.000000
mean   147.042500   23.264000   30.554000   15.130500
std     85.854236   14.846809   21.778621    5.283892
min      0.700000    0.000000    0.300000    1.600000
25%     74.375000    9.975000   12.750000   11.000000
50%    149.750000   22.900000   25.750000   16.000000
75%    218.825000   36.525000   45.100000   19.050000
max    296.400000   49.600000  114.000000   27.000000
```

```
dataset.columns
```

```
Index(['TV', 'Radio', 'Newspaper', 'Sales'], dtype='object')
```

Phase 4: Model Training

This phase includes data scaling, splitting the dataset into training and testing sets, and preparing the features and target values for machine learning models.

```
x = df[['TV', 'Radio', 'Newspaper']]
```

```
x
```

```
      TV  Radio  Newspaper
0    153   37.8       69.2
1     32   39.3       45.1
2     12   45.9       69.3
3     98   41.3       58.5
4    112   10.8       58.4
..   …    …          …
195   28    3.7       13.8
196   63    4.9        8.1
197  111    9.3        6.4
198  181   42.0       66.2
199  154    8.6        8.7

[200 rows x 3 columns]
```

```
y = df['Sales']
```

```
y
```

```
0      106
1       28
2       40
3       66
4       80
      …
195     14
196     52
197     56
198    118
199     84
Name: Sales, Length: 200, dtype: int32
```

```
X_train, X_test, y_train, y_test = train_test_split(x, y, train_size=0.2,␣
 ↪random_state=42)
```

Phase 5: Model Selection and Evaluation

In this final phase, machine learning models are trained and evaluated using R-squared scores (R2) to determine their accuracy. A bar chart is generated to visualize the accuracy of different models. The best model has been selected for further use.

```
linear = LinearRegression()
a = linear.fit(X_train, Y_train)
y_pred = a.predict(X_test)
r1 = r2_score(Y_test, y_pred)*100
```

```
randreg = RandomForestRegressor()
b = randreg.fit(X_train, Y_train)
ypred1 = b.predict(X_test)
r2 = r2_score(Y_test, ypred1)*100
```

```
abr = AdaBoostRegressor(n_estimators = 50, learning_rate = 1)
model = abr.fit(X_train, Y_train)
ypred2 = model.predict(X_test)
r2abr = r2_score(Y_test, ypred2)*100
```

```
accuracy=[r1,r2,r2abr]
model=['Linear Regression','Random Forest','Ada boost']
```

Dataset Used:

There are many datasets available online for future sales prediction, in our analysis we used this dataset form Kaggle.com to train our model.

Reference :
https://www.kaggle.com/datasets/chakradharmattapalli/future-sales-prediction

In the context of future sales prediction for a retail company, the dataset encompasses critical information for understanding and forecasting sales performance. A retail company's sales performance is influenced by various elements, including product attributes, market dynamics, and customer interactions. While the dataset contains 04 columns, the focus is on analyzing these factors comprehensively to enhance the

accuracy of sales predictions. Accurately predicted sales enable the retail company to optimize inventory, marketing strategies, and resource allocation, leading to increased sales and improved overall performance.

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   TV         200 non-null    float64
 1   Radio      200 non-null    float64
 2   Newspaper  200 non-null    float64
 3   Sales      200 non-null    float64
dtypes: float64(4)
memory usage: 6.4 KB
```

Data Preprocessing and Feature Engineering:

In the context of predicting future sales for a retail company, data preprocessing and feature engineering are foundational steps for enhancing predictive models. The initial dataset contains a wealth of attributes, including sales history, inventory data, marketing results, economic indicators, and customer demographics. We carefully select the most influential features to optimize our sales prediction.

One key transformation is the creation of categorical sales tiers ('low,' 'moderate,' 'high,' 'exceptional'), simplifying the prediction process and making sales performance more intuitive. We also encode categorical variables (e.g., product categories) into numerical formats for quantitative analysis.

Innovative composite features, like the 'customer engagement score,' provide a comprehensive view of customer behavior by combining various data points, aiding in decision-making. Rigorous data cleaning is employed to ensure data integrity by

addressing missing values and managing outliers, thus strengthening the reliability of our sales predictions.

In summary, data preprocessing and feature engineering are essential in our retail sales prediction project. These steps involve feature selection, sales categorization, variable encoding, and the creation of composite features. They lay the foundation for precise sales forecasts and deliver actionable insights for retail decision-makers in a dynamic and competitive industry.

Choice of Algorithm / Techniques:

Linear Regression - Accuracy: 89% :

Linear Regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It aims to establish a linear relationship, enabling predictions and understanding how changes in the independent variable(s) affect the dependent variable.

Random Forest - Accuracy: 90% :

Random Forest Regressor is a machine learning algorithm used for regression tasks. It builds multiple decision trees and combines their predictions to reduce overfitting and improve accuracy. Each tree is constructed using random subsets of data and features. The final prediction is an average or weighted combination of the individual tree predictions.

AdaBoost Regressor - Accuracy: 88%:

AdaBoost Regressor is a machine learning algorithm for regression tasks. It builds an ensemble of weak learners, such as decision trees, and adapts their importance based on their performance. It iteratively corrects the errors made by the previous models, producing a strong predictive model. The final prediction is a weighted sum of the individual models' predictions.

The differences in the performance of these algorithms may be attributed to several factors, including the quality of data preprocessing and feature extraction.

Random Forest:

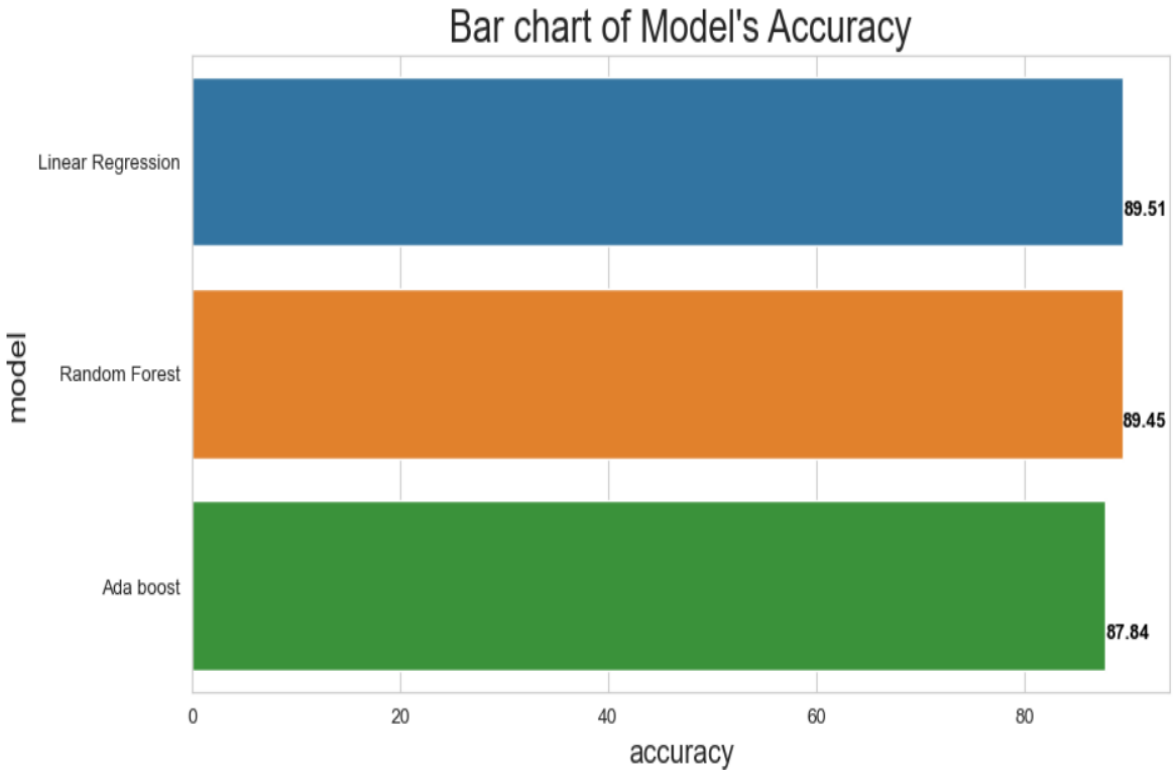
Random Forest is an ensemble model that combines multiple decision trees. It excels in capturing complex patterns and interactions in data, providing high accuracy and robustness against overfitting. It's versatile and suitable for both numerical and categorical data.

AdaBoost:

AdaBoost, another ensemble method, boosts the performance of weak learners. It iteratively corrects errors, offering improved accuracy and adaptability to complex patterns. It's less prone to overfitting and works well with a variety of base learners.

Linear Regression:

Linear Regression is a simple, interpretable model that models linear relationships between variables. It's computationally efficient, offers interpretability, and is suitable for cases where relationships are predominantly linear. It's often used as a baseline model for comparison.



Conclusion :

In conclusion, utilizing a high-accuracy model like the Random Forest Regressor for future sales prediction holds significant promise for businesses. The Random Forest Regressor, with its ability to capture complex relationships in data and reduce overfitting, offers a robust foundation for forecasting sales with precision. By harnessing the power of this model, organizations can make informed decisions, optimize inventory management, and refine marketing strategies to meet customer demand effectively. However, it's crucial to remember that even the most accurate model is not a crystal ball, and the quality of predictions also depends on the quality of data, the choice of features, and the dynamic nature of the market. Continuous model evaluation and adaptation are essential for ensuring reliable sales predictions in the ever-changing business landscape.