Phase 5:

Project documentation



Name: Tharun VS

Reg.no:410621104110

College name: Dhaanish Ahmed College Of Engineering

## Problem statements in future sales prediction

The problem is to develop a predictive model that uses historical sales data to forecast future sales for a retail company. This project involves data preprocessing, feature engineering, model selection, training, and evaluation. This model will predict sales on a certain day after being provided with a certain set of inputs.

# Design Thinking process:

**Data Source**: Utilize a dataset containing historicalsales data. Here a csv file is converted to a DataFrame and the pandas object is used.

#### Data preprocessing:

* The describe() method returns description of the data in the DataFrame. If the DataFrame contains numerical data, the description containsthese information for each column such as count , mean , std , min , 25% , 50% , 75% , max .
* You can use the isnull() or isna() method of pandas. DataFrame and Series to check if each element is a missing value or not. isnull() is an alias for isna() , and both are used interchangeably.value.
* The fillna() method replaces the NULL values with a specified value. The fillna() method returns a new DataFrame object unless the inplace parameter is set to True , in that case the fillna() method does the replacing in the original DataFrame instead.
* The drop\_duplicates() method removes duplicate rows. Use the subset parameter if only some specified columns should be considered when looking for duplicates.
* The strategy is to convert each category into a column and assign it a 1 or 0 value. It is a process of creating dummy variables. We can see from the table above that all the unique

categories were assigned a new column. If a category is present, we have 1 in the column and 0 for others.

#### Feature engineering:

Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data.Here the Total spent is added as a feature using the datas of TV , Radio , Newspaper in the data set .

#### Model Selection:

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset. Here the ARIMA model is selected . An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends . Autoregressive Integrated Moving Average (ARIMA ) is a commonly-used local statistical algorithm for time-series forecasting.

#### Model Training:

Train/Test is a method to measure the accuracy of your model. It is called Train/Test because you split the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing. You train the model using the training set. You test the model using the testing set.

#### Evaluation:

Currently, the most popular metricsfor evaluating time series forecasting models are MAE, RMSE and AIC. To briefly summarize, both MAE and RMSE measures the magnitude of errors in a set of predictions. The major difference between MAE and RMSE is the impact of the large errors.

Dataset in future sales prediction

The ’test.csv’ data file that is a part of this dataset is only being used to predict values derived from the model with the lowest WMAE score. Because, the dataset contains no target variable, in our case ’Weekly Sales’, it cannot be used for testing for this analysis.Instead, the training dataset ’train.csv’ is being split into training and validation datasets for this study.

The main goal of this study is going to be to predict the department-wide weekly sales for each of these stores.

The dataset is already divided into separate training and testing data; the testing data is identical to the training dataset apart from the weekly sales information. The training dataset contains weekly sales information from 2010-02-05 to 2012-11- 01 about the stores and departments. It also contains a column that suggests whether a particular date falls on a holiday or not.

In total, there are 4,21,570 rows in the training dataset and 1,15,064 rows in the testing dataset.

The ‘inspectdf’ package in R does basically what the name suggests: it inspects crucial components of a dataframe under study. For this study, it was essential to get a sense of the several datasets before they were used to create complex models. This package inspects, compares, and visualizes the different components associated with the above-mentioned datasets. It gives a brief visualized column-wise sum

mary about missing or out of range values, distribution of values, types of columns,the correlation between numeric columns, etc. Starting with the initial explorations, it is imperative to understand the data types and ranges of the values in each column; the ‘inspecttypes()’ function from the package helps explore the data type for columns in

the dataset.

# Data Cleaning and Preprocessing

The data contains 421,570 rows, with some store-specific departments missing a few too many weeks of some columns in the features dataset contain missing values, however, after the features dataset is merged with the training dataset, the only missing values that exist are in the Markdown column.After the extensive EDA, it was determined that these five markdown files, with missing values, have barely any correlation to the weekly sales forWalmart, hence these five columns have been eliminated from the subsequent training and testing dataset.

Because the source already provides training and testing datasets, there is no need to create them for our study. Because the main focus of this study is to accurately predict weekly sales for different Walmart stores, the previously modified ‘Date’, ‘Month’, ‘Quarter’, and ‘Day’

columns have been dropped and only the ‘Week of Year’ column has been used in the upcoming models.



Data has been checked for inaccuracies, missing or out of range values using the ‘inspectdf’ package in R as part of the initial EDA. Columns with missing values have been dropped. The dataset contains information about weekly sales which was initially broken down to acquire information about monthly as well as quarterly sales.

#### Future Sales Prediction

The [**dataset**](https://raw.githubusercontent.com/amankharwal/Website-data/master/advertising.csv) given here contains the data about the sales of the product. The dataset is about the advertising cost incurred by the business on various advertising platforms. Below is the description of all the columns in the dataset:

1. **TV:** Advertising cost spent in dollars for advertising on TV;
2. **Radio:** Advertising cost spent in dollars for advertising on Radio;
3. **Newspaper:** Advertising cost spent in dollars for advertising on Newspaper;
4. **Sales:** Number of units sold;

So, in the above dataset, the sales of the product depend on the advertisement cost of the product. I hope you now have understood everything about this dataset. Now in the section below, I will take you through the task of future sales prediction with machine learning using Python.

Machine learning algorithm in future sales prediction

Forecasting sales is a common and essential use of machine learning (ML). Sales forecasts can be used to identify benchmarks and determine incremental impacts of new initiatives, plan resources in response to expected demand, and project future budgets.

The first step is to load the data and transform it into a structure that we will then use for each of our models. In its raw form, each row of data represents a single day of sales at one of ten stores. Our goal is to predict monthly sales, so we will first consolidate all stores and days into total monthly sales.

def load\_data():

url = ["""https://www.kaggle.com/](http://www.kaggle.com/c/demand-forecasting-kernels" /h)c/demand-[forecasting](http://www.kaggle.com/c/demand-forecasting-kernels" /h)-[kernels](http://www.kaggle.com/c/demand-forecasting-kernels" /h) only/download/ryQFx3IEtFjqjv3s0dXL%2Fversions%2FzjbSfpE39fdJl MotCpen%2Ffiles%2Ftrain.csv"""

return pd.read\_csv(url)def monthly\_sales(data):

data = data.copy() # Drop the day indicator from the date column

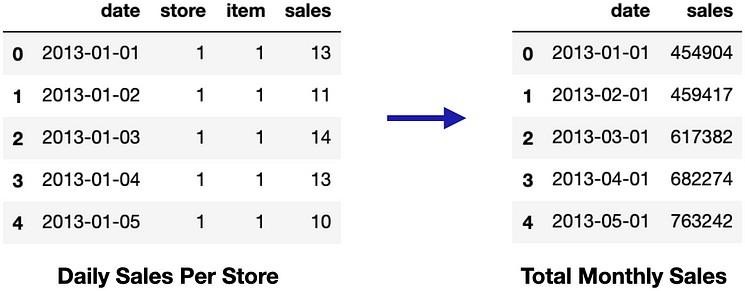
data.date = data.date.apply(lambda x: str(x)[:-3]) #

Sum sales per month

data = data.groupby('date')['sales'].sum().reset\_index() data.date = pd.to\_datetime(data.date) data.to\_csv('../data/monthly\_data.csv')

return datadata = load\_data() monthly\_data = monthly\_sales(data)

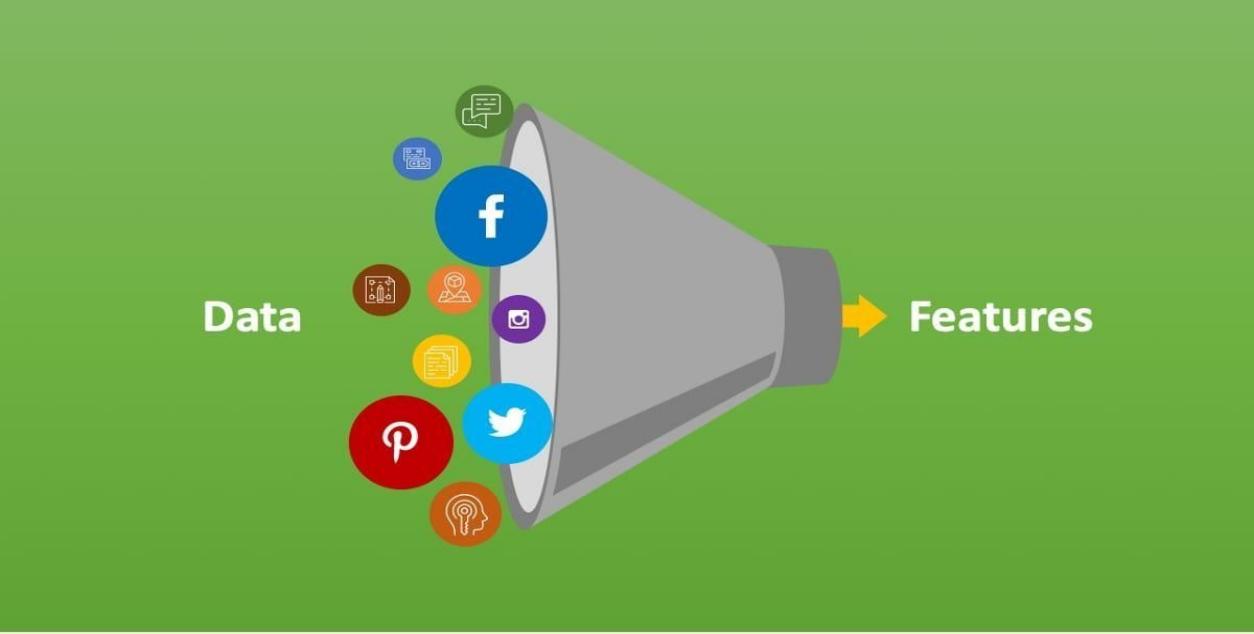
In our new data frame, each row now represents total sales in a given month across all stores.



If we plot the total monthly sales over time, we see that average monthly sales increase over time, which means that our data is not stationary. To make it stationary, we will calculate the difference between sales in each month and add this into our data frame as a new column.

One of the most common methods used to predict sales is **regression analysis.** This method involves using historical sales data to train a model that can predict future sales. The model can take into account factors such as **past sales, marketing campaigns, and economic indicators** to make its predictions.Another popular method for predicting sales is **time series analysis**. This method involves using historical sales data to

identify patterns and trends in sales over time. The model can then use these patterns to make predictions about future sales. This method is particularly useful for predicting sales in seasonal industries, such as retail and tourism.Another approach is using **decision tree-based algorithms** like **Random Forest, Gradient Boosting** etc. These algorithms are particularly useful when there are many factors that can influence sales, such as product features, customer demographics, and market conditions. The algorithm can help identify the most important factors and use them to make predictions.In addition to these methods, machine learning can also be used to predict sales through the use of **neural networks.** Neural networks are a type of machine learning algorithm that can learn to recognize patterns in data. They can be trained on large amounts of sales data and can make predictions about future sales.Machine learning can also be used to predict sales by using **clustering algorithms,** which can help identify groups of similar customers. This information can then be used to create targeted marketing campaigns and improve sales strategies.



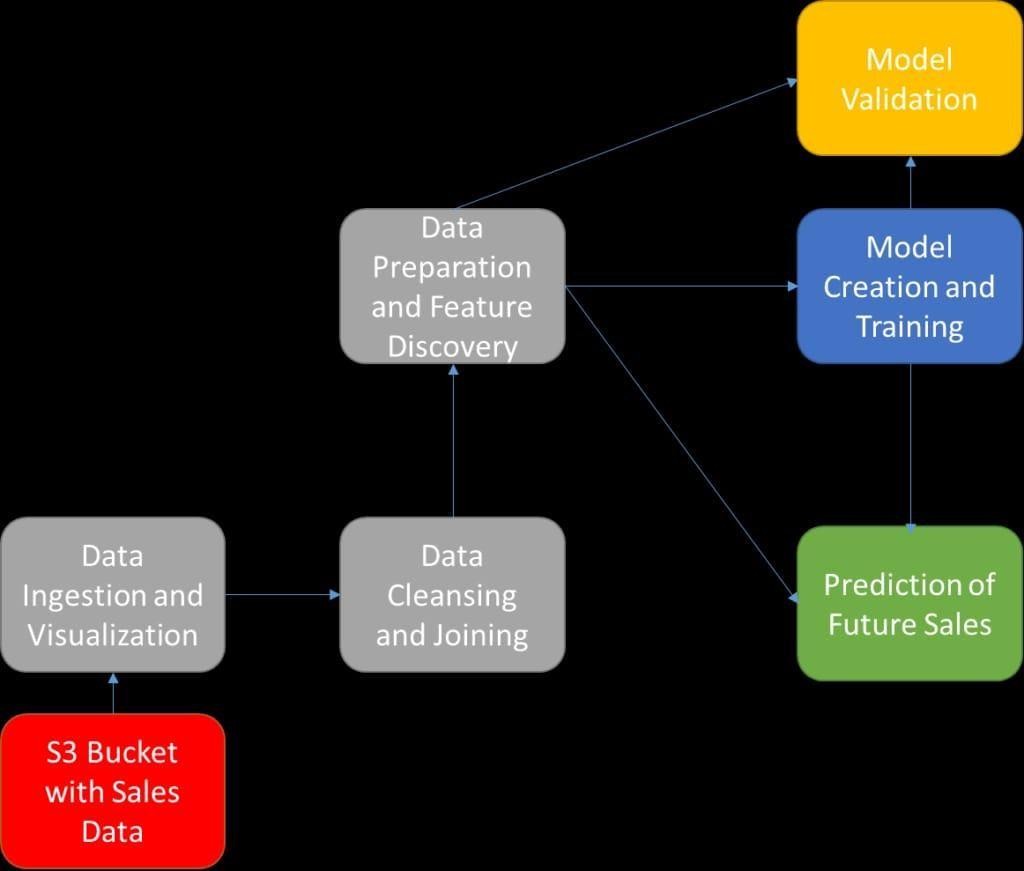
Evaluation metrics in future sales prediction

When forecasting sales with machine learning, the metrics you should track and measure vary depending on your business objectives, sales cycle, and industry. However, some of the most common metrics include revenue, quota attainment, conversion rate, pipeline velocity, and forecast accuracy. Revenue measures the amount of money generated by sales activities such as deals closed or renewals. Quota attainment is the percentage of revenue achieved by reps or teams compared to their targets. Conversion rate is the percentage of prospects or leads that move from one stage of the sales funnel to the next. Pipeline velocity is the speed at which prospects or leads progress through the funnel, while forecast accuracy is the degree of alignment between predicted and actual sales outcomes.



### Innovative techniques in future sales prediction

Sales forecasting with machine learning can draw from two categories of data sources: internal and external. Internal data sources are those collected and managed within the organization, such as CRM data, sales performance data, and sales feedback data. External data sources come from outside the organization and include market data, industry data, and social media data. These external sources can help to understand the size of the market, demand, trends, competitors, industry standards, regulations, best practices, benchmarks, brand awareness, reputation, trust, and relationships with prospects, customers, and influencers.



### Approaches using during the development in future sales prediction

In order to apply machine learning to your sales forecasting process, you must first define your forecasting objective, such as what you are trying to predict and why it will help you reach your sales goals. Additionally, you need to collect and prepare data from various sources, ensuring quality, relevance, and completeness. Furthermore, you must select and train the machine learning algorithm or technique, configuring and evaluating it for performance and accuracy. Finally, you must deploy and monitor the model in your sales system and workflow, updating and maintaining it as needed.

Sales forecasting with machine learning can provide many advantages, such as improved accuracy and efficiency, enhanced insights and recommendations, and better decision making and performance. However, there are also some challenges associated with this approach, including data quality and availability, model complexity and interpretability, and ethical and legal issues. Data sources need to be clean, relevant, and updated; models must be transparent, explainable, and trustworthy; and ethical, legal, and responsible best practices must be followed.

Date time features represent a useful way for data scientists to start their feature engineering work with time series data. In the next section, I introduce an additional approach to build input features for your dataset: lag and window features. In order to build these features, data scientists must leverage and extract the values of a series in previous or future periods.

Dataset in future sales prediction

Trying to find and implement the most effective model is the biggest challenge of this study. Selecting a model will depend solely on the kind of data available and the analysis that has to be performed on the data (UNSW, 2020). Several models have been studied as part of this study that were selected based on different aspects of our dataset; the main purpose of creating such models is to predict the weekly sales for different Walmart stores and departments, hence, based on the nature of models that should be created, the following four machine learning models

have been used:

* Linear Regression
* Lasso Regression
* Gradient Boosting Machine
* Random Forest

Each of these methods have been discussed briefly in the upcoming report. For each of the models, why they were chosen, their implementation and their success rate

(through WMAE) have been included.



These are the innovative techniques or approaches used during the development in future sales predictions.