**Product Sales Analysis with Machine Learning - Project Design**

College code: 0001

**Phase 2: Project Definition and Design Thinking**

**Team Members:**

Diya Arshiya S (202115033) [diya.arshiya@gmail.com](mailto:(diya.arshiya@gmail.com))

Dhivyadharshini S K(2021115030) [dhivyadharshini0907@gmail.com](mailto:dhivyadharshini0907@gmail.com)

Mukesh Raja K(2021115065) [mukeshrajatmr2021@gmail.com](mailto:mukeshrajatmr2021@gmail.com)

Mukilarasan V (2021115066) mukilarasan.v@gmail.com

Karthik V (2021115321) karthiksk9360@gmail.com

**Table of Contents**

1. Project Overview

2. Objectives

3. Data Overview

4. Data Preprocessing

5. Machine Learning Models

- Supervised Learning

- Unsupervised Learning

6. Conclusions and Future Steps

**1. Project Overview**

This section will set the foundation for your focus on machine learning. Mention that the project aims to leverage machine learning algorithms for a data-driven approach to analyzing product sales and predicting future trends of the company.

**2. Objectives**

The primary objectives is to apply machine learning models to predict sales trends, identify customer behaviors .

**3. Data Overview**

Data for this Project is taken from the below link https://www.kaggle.com/datasets/ksabishek/product-sales-data

**4. Data Preprocessing**

**i. Data Collection:**

Gather all relevant data sources, such as sales records, customer information, product details, and any other relevant data.

**ii. Data Exploration:**

Begin by exploring the data to understand its structure, the types of features, and the overall quality. This involves checking for the presence of missing values, outliers, and data distributions.

**iii. Handling Missing Values:**

* Identify and handle missing values in your dataset. Common strategies include:
* Removing rows with missing values if they are a small percentage of the dataset.
* Imputing missing values using techniques like mean, median, mode, or more advanced imputation methods.
* Using domain knowledge or predictive models to estimate missing values.

**iv. Data Cleaning:**

Address any data quality issues, such as correcting inconsistent or erroneous values, removing duplicates, and standardizing data formats.

**5. Machine Learning Models**

This will be the meat of your document. Break this section down into the types of machine learning models you'll use:

**i)Supervised Learning:**

**Random Forest**

**Overview:**

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during the training phase. The more trees in the 'forest', the more robust the model. For predictions, the output of individual trees is averaged (for regression) or voted upon (for classification) to produce the final result.

**Key Features:**

* Bagging Technique: Random Forest uses bootstrap aggregating (bagging) to create different subsets of the original dataset, training each tree on a different subset.
* Feature Randomness: In addition to data randomness, Random Forest selects a random subset of features for splitting nodes, adding an extra layer of randomness that makes the model more robust.
* Overfitting Resistance: The ensemble nature of Random Forest makes it resistant to overfitting, especially when using many trees.
* Hyperparameters: Important hyperparameters include the number of trees (`n\_estimators`), the maximum depth of the tree (`max\_depth`), and the minimum samples required to make a split (`min\_samples\_split`).

**Use-Cases:**

- Sales prediction

- Customer segmentation

- Anomaly detection

**ii) XGBoost (eXtreme Gradient Boosting)**

**Overview:**

XGBoost is an open-source software library that provides a gradient boosting framework. It is renowned for its performance and computational speed, and it's designed to be highly efficient, flexible, and portable.

**Key Features:**

* Boosting Technique: Unlike Random Forest, which builds each tree independently, XGBoost builds trees sequentially, each one correcting the errors of its predecessor.
* Handling Missing Values: XGBoost has an in-built routine to handle missing values, which can be especially useful for real-world data.
* Regularization: Includes L1 (Lasso regression) and L2 (Ridge regression) regularization terms in its cost function, reducing the likelihood of overfitting.
* Hyperparameters: Important hyperparameters include learning rate (`eta`), number of boosting rounds (`n\_estimators`), and tree complexity (`max\_depth`).

**Use-Cases:**

- Price prediction

- Customer churn prediction

- Natural Language Processing tasks

**Unsupervised Learning:** For customer segmentation or finding hidden patterns, clustering algorithms like K-means could be used.

**6. Conclusions and Future Steps**

Sum up the expected outcomes of the project and what steps could be taken next. This could include deploying the model into a production environment or integrating it with other systems for automated insights.