High Level Design
Black Friday Sales Prediction

Written By	**Aditya Gautam**
Document Version	**1.0**
Last Revised Date	**05-Dec-2022**

Document Version Control

Date Issued	Version	Description	Author
05-Dec-2022	1.0	First Version of Complete HLD	Aditya Gautam

Contents

1 Introduction

1.1 Why this High-Level Design Document?
The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

(1) Present all of the design aspects and define them in detail
(2) Describe the user interface being implemented
(3) Describe the hardware and software interfaces
(4) Describe the performance requirements
(5) Include design features and the architecture of the project
(6) List and describe the non-functional attributes like:
    (a) Security
    (b) Reliability
    (c) Maintainability
    (d) Portability
    (e) Reusability
    (f) Application compatibility
    (g) Resource utilization
    (h) Serviceability

1.2 Scope
The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

2 General Description

2.1 Product Perspective & Problem Statement

Retail is the sale of goods and services from individuals or businesses to the end-user. The retail industry provides consumers with goods and services for their everyday needs. In retail one of crucial part is to understand the consumer behaviour and make various arrangements for the sales of the company.

A retail company "ABC Private Limited" wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month.

The data set also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

The Objective of this project is to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.
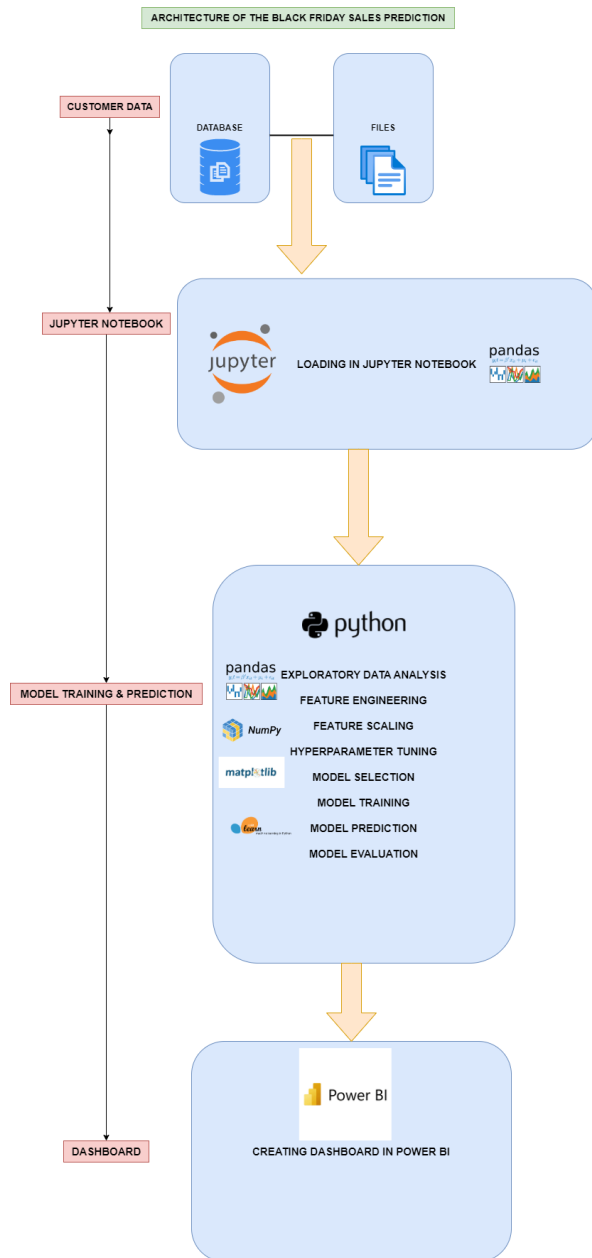
2.2 Tools Used

Jupyter Notebook, Python Libraries such as Pandas, Numpy, Matplotlib, seaborn, Sklearn and
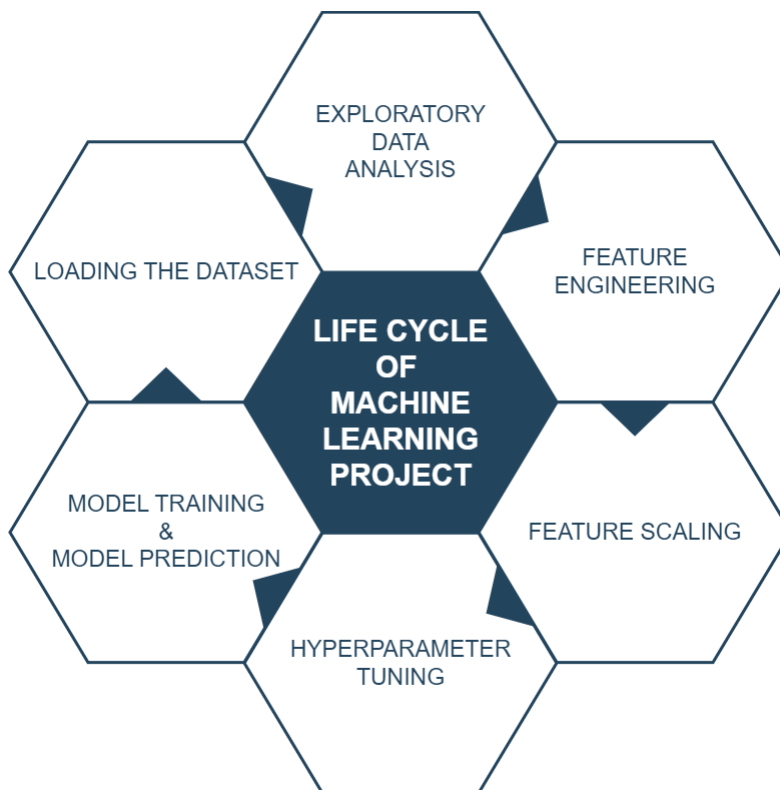
Business Intelligence such as Power Bi.

3 Design Details

3.1 Architecture of the Black Friday Sales Prediction

ARCHITECTURE OF THE BLACK FRIDAY SALES PREDICTION

CUSTOMER DATA

DATABASE

FILES

JUPYTER NOTEBOOK

LOADING IN JUPYTER NOTEBOOK

pandas

MODEL TRAINING & PREDICTION

python

pandas — EXPLORATORY DATA ANALYSIS

FEATURE ENGINEERING

NumPy — FEATURE SCALING

HYPERPARAMETER TUNING

matplotlib — MODEL SELECTION

MODEL TRAINING

MODEL PREDICTION

MODEL EVALUATION
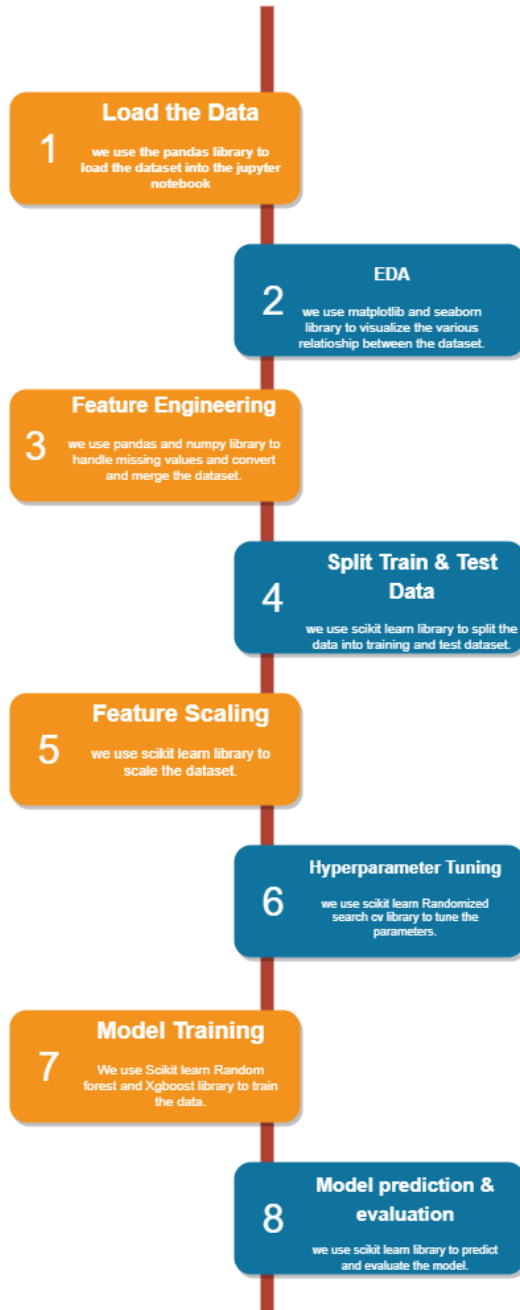
DASHBOARD

Power BI

CREATING DASHBOARD IN POWER BI

The detailed architecture of the Black Friday Sales Analysis has been discussed in the above architecture diagram which gives a overview of the step by step process of the project which gives an idea about flow of the data from original sources to database, then exporting the data from database to importing the data into jupyter notebook by using pandas library for data cleaning process, then for visualize the data, visualization library such Matplotlib and seaborn is used for the purpose and pandas library is used for Feature engineering. Then scikit learn library is used for feature selection , model training, hyperparameter tuning and model evaluation of the data. And finally, deploying the trained data into Power Bi for creating an interactive dashboard.

3.2 **Life Cycle of Machine Learning Project**

3.3 **Detailed Architecture of Black Friday Analysis**

Load the Data

1 we use the pandas library to load the dataset into the jupyter notebook

EDA

2 we use matplotlib and seaborn library to visualize the various relatioship between the dataset.

Feature Engineering

3 we use pandas and numpy library to handle missing values and convert and merge the dataset.

Split Train & Test Data

4 we use scikit learn library to split the data into training and test dataset.

Feature Scaling

5 we use scikit learn library to scale the dataset.

Hyperparameter Tuning

6 we use scikit learn Randomized search cv library to tune the parameters.

Model Training

7 We use Scikit learn Random forest and Xgboost library to train the data.

Model prediction & evaluation

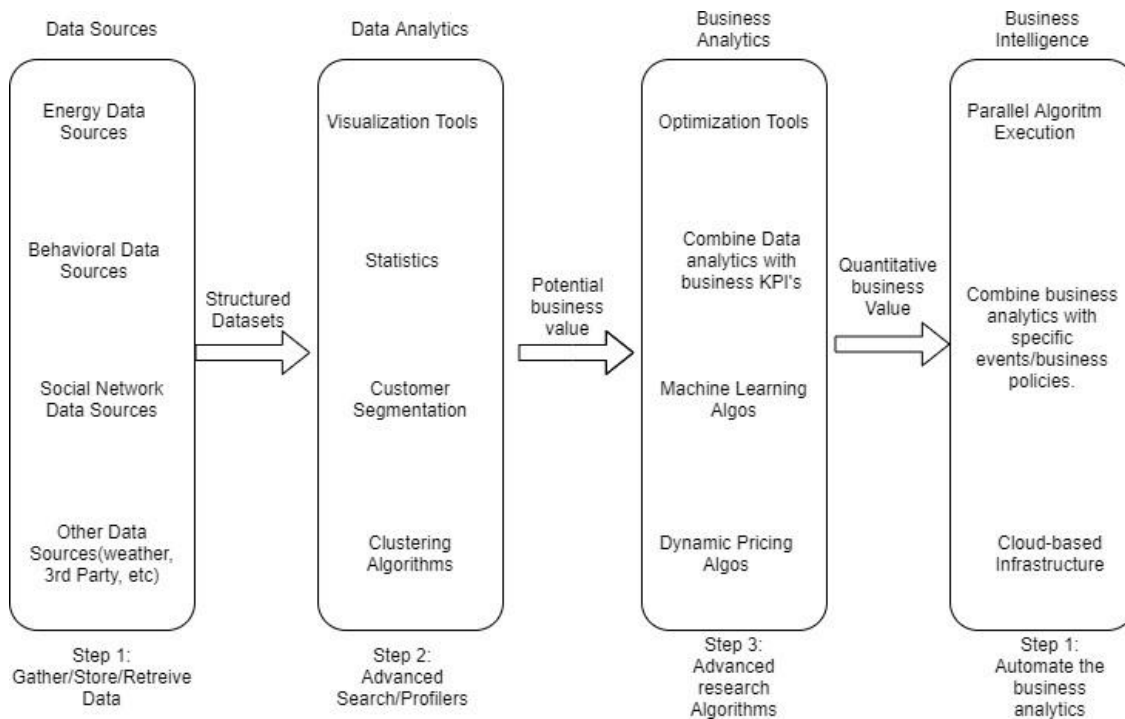**8** we use scikit learn library to predict and evaluate the model.

3.4 **Function Design**

Figure 1: Functional Architecture of Business Intelligence

3.5 **Optimization**

(1) Handling the missing values

  For optimizing the model, missing values can be replace with zero or with the average value.

(2) Handling the Categorical values

  For optimizing the model, categorical values can be replace with dummy variables or can be mapped with the required numerical values.

(3) Handling the Multi-co-linearity between the variables

  For optimizing the model, handling the multi-co-linearity between variables is a vital step Which is can be check by variance inflating factor and if VIF is above 4 or tolerance is below 0.25 indicates that multi-co-linearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1 there is significant multi-co-linearity that needs to be corrected. Which can be handled by dropping some redundant variables.

(4) Selecting the Important feature

For optimizing the model, from Sklearn library import ExtraTreesRegressor , which helps to select the vital features for the model.

(5) Hyper parameter Tuning

For optimizing the model, from sklearn library import RandomizedSearchCV, which helps to fine tune the parameter of the model before training, which helps to attain an optimized model.

(6) Select the optimized model for training the dataset.

To attain the optimized model, model have to be evaluated using R2 score and RMSE, if the model is performing low in the R2 score and RMSE then change the model and select another model and follow this step until desired level of accuracy is attained.

Repeat the Following steps until desired level of accuracy is obtained.

4 **Key Performance Indicator**

Key Performance Indicator provides vital insights about the predicted model

4.1 **Insight of the Black Friday Sales Prediction**

(1) Sum of Actual Purchase and Sum of Predicted purchase using Random Forest Regressor in Train Dataset, it is clear from the graph that the product id - P00025442 has the highest Actual Purchase with $10,013 k and has the highest Predicted Purchase with $ 5,347k.

(2) Sum of Actual Purchase and Sum of Predicted purchase using XGBoost Regressor in Train Dataset, it is clear from the graph that the product id - P00025442 has the highest Actual Purchase with $10,013 k and has the highest Predicted Purchase with $ 5,315k.

(3) Sum of Predicted Purchase Using Random Forest Regressor and XGBoost Regressor in Test Dataset, it clear from the graph that the product id - P00112142 has the highest Predicted Purchase with $10,243 k.

(4) Sum of Predicted Purchase Using Random Forest in Train data, it clear from the graph that the age between 26-35 has the highest Predicted purchase with $5,16,096K for Male and $1,53,589 K for female, where for city category B the predicted purchase is $5,31,360 K for male andfor female $1,77,477 K and for Overall predicted Purchase for male is $12,69,751 K and for female is $4,13,162 K.

(5) Sum of Predicted Purchase Using XGBoost Regressor in Train data, it clear from the graph that the age between 26-35 has the highest Predicted purchase with $5,12,842K for Male and $1,52,567 K for female, where for city category B the predicted purchase is $5,27,806 K for male and for female $1,76,253 K and for Overall predicted Purchase for male is $12,61,389 K and for female is $4,10,426 K.

(6) Sum of Predicted Purchase Using Random Forest in Test data, it clear from the graph that the age between 26-35 has the highest Predicted purchase with $6,82,712 K for Male and $1,89,600 K for female, where for city category B the predicted purchase is $6,95,655 K for male and

for female $2,11,451 K and for Overall predicted Purchase for male is $16,75,002 K and for female is $5,08,443 K.

(7) Sum of Predicted Purchase Using XGBoost in Test data, it clear from the graph that the age between 26-35 has the highest Predicted purchase with $6,78,448 K for Male and $1,88,209 K for female, where for city category B the predicted purchase is $6,91,135 K for male and for female $2,09,962 K and for Overall predicted Purchase for male is $16,64,475 K and for female is $5,05,006 K.

5. **Deployment**

5.1   **Load the Dataset in Power BI**

Now, Load the Dataset into Power Bi for creating the dashboard.
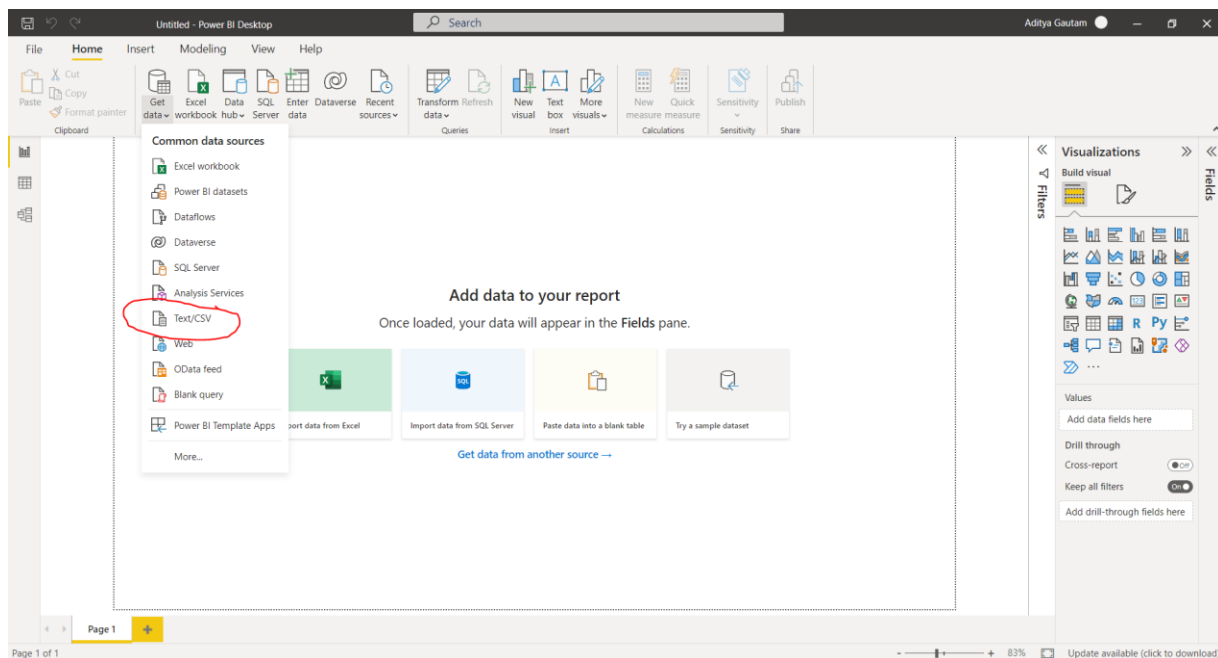
5.2 Create the interactive dashboard

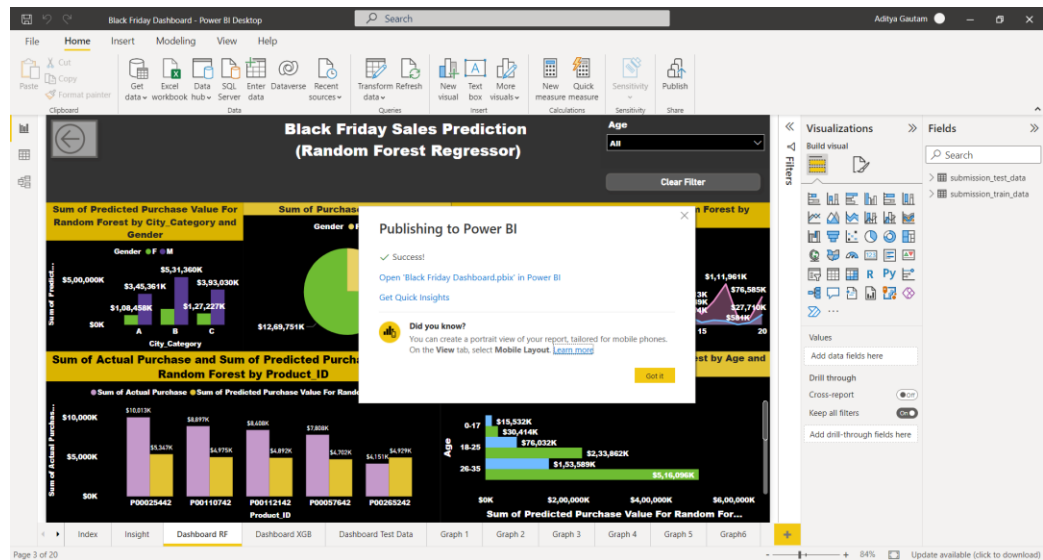### Random Forest Regressor
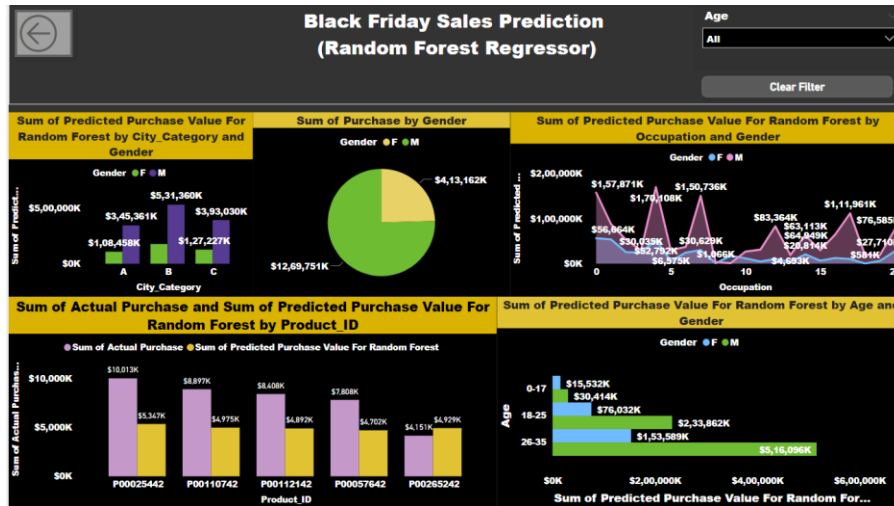
XGBoost Regressor
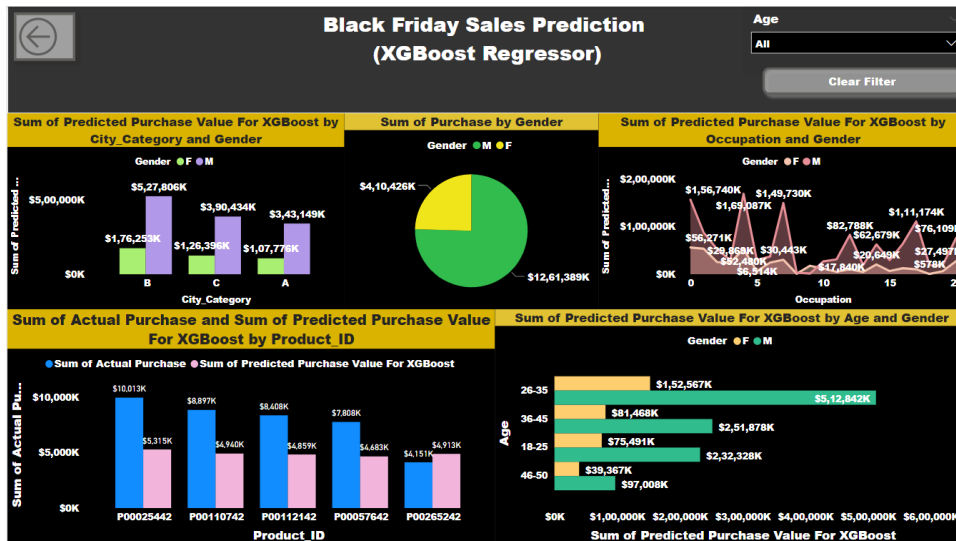
5.3 **Publish to Power Bi account**

(1)
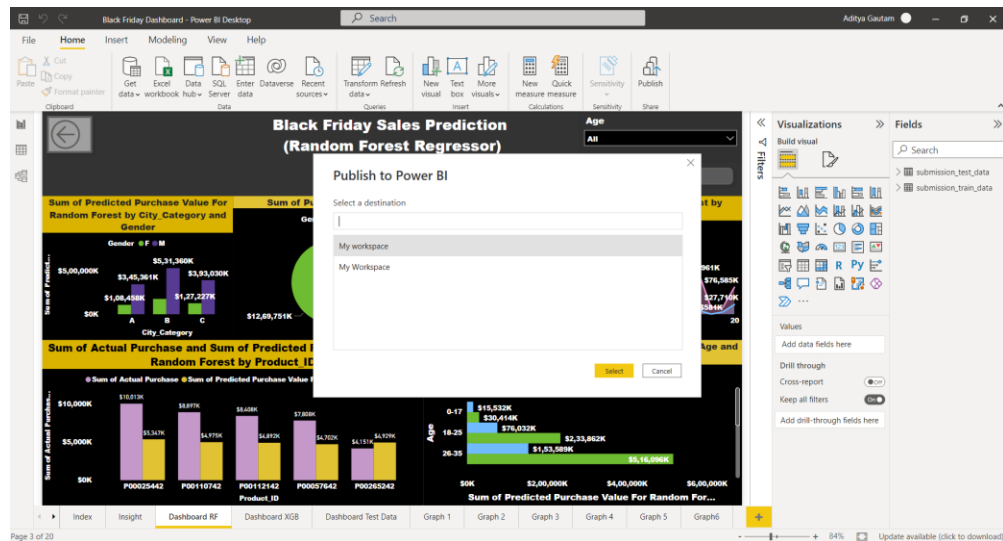
(2)

5.4 **Publish to Web**

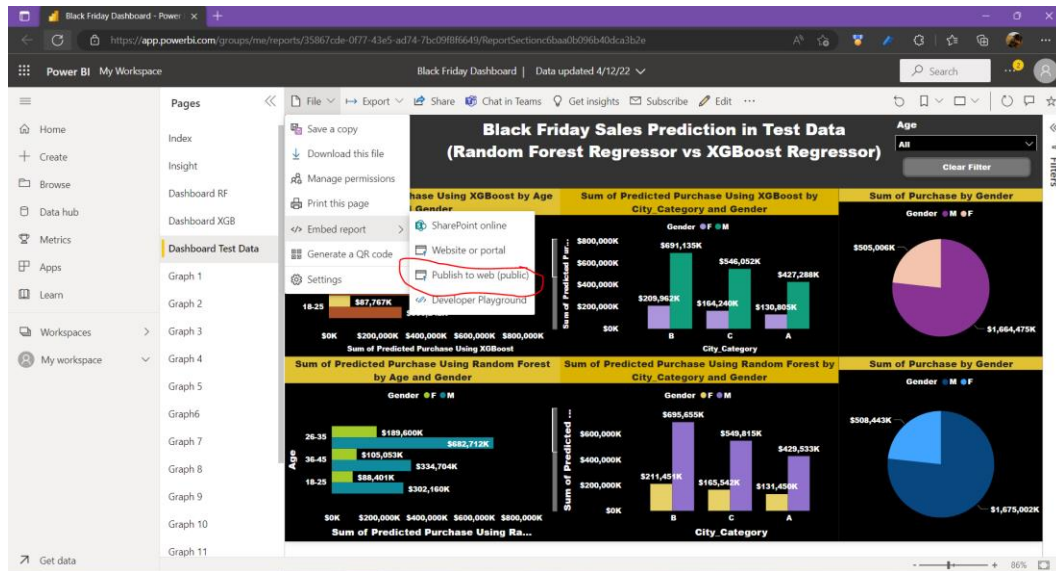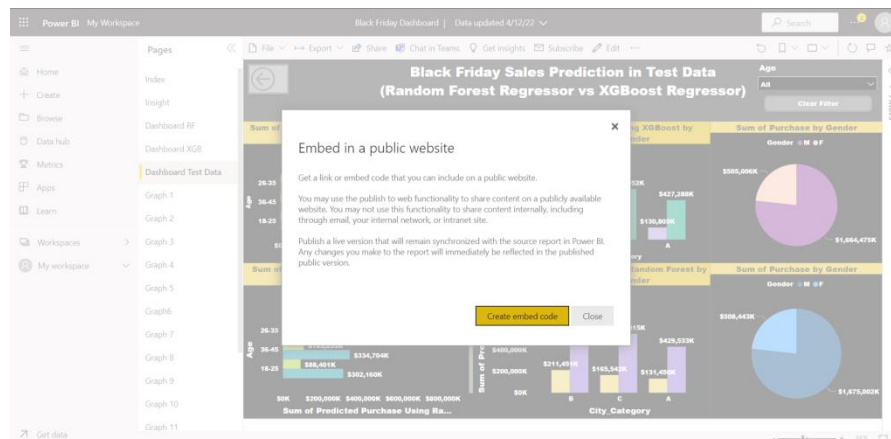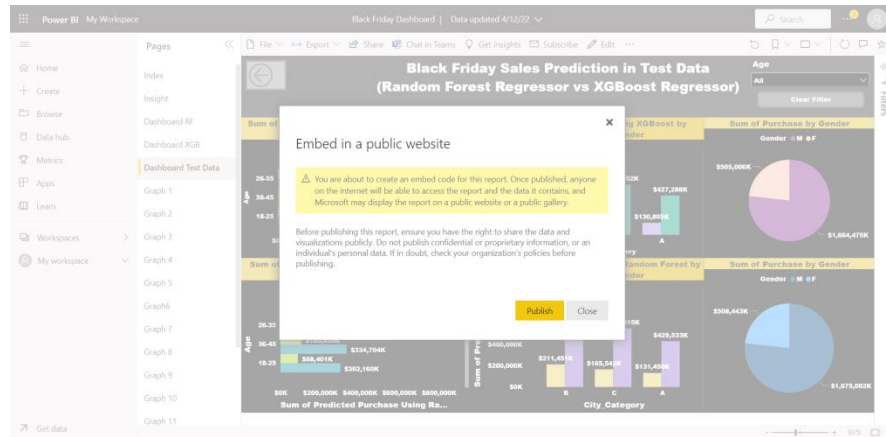## 5.5 **Create the embedded code**

(1)

(2)

5.6 **Share the Public Link to Client**