**Team Group - Machine Learning 6**

**Topic** - **Product Personalization**

**Project Report**

**Description:** Predict which offers will be most attractive to each individual customer, resulting in more targeted marketing campaigns and higher brand value.

**Team Members**

Saloni Jain(PGI18CS035)

Prashant Jain(PGI18CS030)

Anmol Agarwal(PIET18CS017)

Aachal Kala(PIET18CS001)

Mohan Gayan(PGI18CS024)

Shivam Kushwah(PGI18CS040)

**Mentor**

Dr. Geeta Chhabra Gandhi

Associate Professor

Poornima University

**Table of Contents**

**(1) Introduction**

1.1 Problem statement

1.2 About the data-set

1.3 Methodology

1.4 Software tools & Hardware requirements

**(2) Data collection & Loading**

2.1 Importing essential libraries

2.2 Reading The Data-set

**(3) Data Pre-processing**

3.1 Data Cleaning & Inspection

3.2 Identifying and handling the missing values

**(4) Exploratory Data Analysis**

4.1 Univariate Analysis

4.2 Bivariate Analysis

**(5) Feature Engineering**

5.1 Handling imbalanced data.

5.2 Handling outliers.

5.3 Encoding.

5.4 Feature Scaling

**(6)Feature Selection**

6.1 Selection important features

**(7) Model Creation**

7.1 Implementation of various ML models

**(8) Model Evaluation**

8.1 Evaluation ML models

**(9) Model Deployment**

9.1 Deployment of ML model using python and Django

**(10) References**

**Abstract**

In this project, we have predicted which offers will be most attractive to each individual customer, resulting in more targeted marketing campaigns and higher brand value.

This report consists of complete life cycle of our machine learning project. We have described each and every phase in a short description that we have implemented.

**(1) Introduction**

Product Customization or Product Personalization is a process of delivering customized goods and services to the customers as per their needs and desire. Customers can either approach a merchant to make certain customization in a product or personalize the products themselves, exactly the way they want.

Product customization has made a stand in the

E-commerce market and has caught the attention of the e-retailers. Let’s try to understand why the Product Configuration service has gained popularity among the E-Commerce merchants.

**1.1 Problem Statement**

Predict which offers will be most attractive to each individual customer, resulting in more targeted marketing campaigns and higher brand value.

**1.2 About The Data-set**



**1.3 Methodology**

* **Step 1- Data collection:** This will involve scrapping of structured data from the source.
* **Step 2- Data Preprocessing:** In this phase, the data is prepared for the analysis purpose which contains information. **Pre-processing** and **data cleaning** are some of the most important tasks that must be done before data-set can be used for machine learning.
* **Step 3- Feature Extraction Set/Training :** Feature set or training data can be prepared from the cleaned data by using any of the available techniques. The feature sets and training set that has obtained by using any method will be used for the implementation of machine learning algorithms.
* **Step 4- Implementation of Machine Learning Algorithm:** Here after extracting the important features from the data-set we have to train our model with these features . Here we implement different machine learning algorithms on our data-set according to our given problem
* **Step 5- Testing of Data:** Testing of data is done based on training model which is classified using supervised learning algorithm.

**1.4 Software tools & Hardware**

**Requirements:**

* Software - Google colab, Power BI
* Libraries - NumPy, Pandas, Matplotlib and Seaborn, SKlearn
* Languages - Python , Html , CSS
* Framework - Django

**(2) Data Collection & Loading**

Suppose if you want to start a ML project then what is the first and most important thing you would require? It is the data that we need to load for starting any of the ML project. With respect to data, the most common format of data for ML projects is CSV (comma-separated values).

Basically, CSV is a simple file format which is used to store tabular data (number and text) such as a spreadsheet in plain text.

**2.1 Importing Neccessary Libraries**

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

import seaborn as sns

import Sklearn

**2.2 Reading the Dataset**

from google.colab import files

uploaded = files.upload()

df = pd.read\_csv("File Path")

df.head()

**(3) Data Preprocessing**

* Data Preprocessing is a Data Mining technique that involves transforming raw data into an understandable format. Why Do We Need Data Preprocessing?
* Since real-world data tends to be incomplete, noisy or inconsistent, we need this process in order to eliminate such to avoid poor quality models we'll build later.
* Data Preprocessing provides operations which can organise the data into a proper form for better understanding in data mining process.

**3.1 Data Cleaning & Inspection**

In Data Science, data cleaning can be described in many ways, one of them being: the process of fixing and/or removing incorrect, corrupted, wrongly-formatted, incomplete or duplicate data within a dataset.

Quite mouthful and yet comprehensive. We want to ensure that our dataset has no duplicate data or does not contain any corrupted entries that will otherwise lead to wrong/less useful models.

As we make use of multiple data sources in data analysis, the chances of duplicating data are very high, and so are those of mislabeling the data. It goes without saying that incorrect data leads to unreliable algorithms and predictions or outcomes. While there are no sure steps of going about data cleaning due to the nature of different datasets, it is however vital to build some sort of a framework or template for data cleaning process for future references so as to at least be close to doing it right each time.

**3.2 Handling Missing Values**

One of the important stages of data mining is pre-processing, where we prepare the data for mining. Real-world data tends to be incomplete, noisy, and inconsistent and an important task when pre-processing the data is to fill in missing values, smooth out noise and correct inconsistencies.

If we specifically look at dealing with missing data, there are several techniques that can be used. Choosing the right technique is a choice that depends on the problem domain

So after collecting and reading the dataset we have cleaned our dataset using appropriate approach for our dataset

**(4) Exploratory Data Analysis**

Exploratory data analysis is simply the process of performing initial investigations on data to discover pattern, anomalies with the aid of graphical representations and summary statistics

It is worth noting that EDA is not really a formal process with rigid set of rules or path to follow, it is what one makes it. The aim is to uncover whatever may be hidden in the data, so one should feel free to investigate whatever idea that comes to mind. Some ideas will yield some positive outcomes, others not so much. To successfully perform data cleaning, we'll need to deploy EDA tools such as visualization, transformation and modelling.

4.1 **Univariate Analysis**

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn't deal with causes or relationships (unlike regression ) and it's major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. So we have taken different features and performed univariate analysis on them

4.2 **Bivariate Analysis**

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y. We have taken some of the features from our dataset and performed bivariate analysis to analyze which feature is depending on other feature

**(5) Feature Engineering**

This is the process of selecting features which are most relevant in predicting the output variable.

* It helps reduce data dimensionality and
* Ensures that models' accuracy can be trusted when those features are out.

**5.1 Handling imbalanced Data**

Most datasets used for hands-on practicals are cleaned up or have their imbalanced nature ignored in a bid to focus on learning specific algorithms or techniques without getting distracted by other issues but we have checked whether our data is balanced or not ,In our case the dataset is completely balanced

**5.2 Handling Outliers**

In machine learning, an approach to tackling the problem of outlier detection is one-class classification.One-Class Classification, or OCC for short, involves fitting a model on the “normal” data and predicting whether new data is normal or an outlier/anomaly.A simple approach to identifying outliers is to locate those examples that are far from the other examples in the feature space. We can also use boxplot for checking outliers in a feature. We can remove them with the help of IQR or Z-Score method. We have also checked for outliers in our dataset.

**5.3 Encoding**

Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called feature encoding.

Data frame analytic automatically performs feature encoding. The input data is pre-processed with the following encoding techniques:

one-hot encoding: Assigns vectors to each category. The vector represent whether the corresponding feature is present (1) or not (0).

target-mean encoding: Replaces categorical values with the mean value of the target variable.

frequency encoding: Takes into account how many times a given categorical value is present in relation with a feature.

**5.4 Feature Scaling**

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. We have scale down our all the features .

**(6) Feature Selection**

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.There are various features in our dataset and we have taken only the best features from the dataset

(7) **Model Creation**

This is the one of the major phase in every machine learning project life cycle. Here we have to train our model with the help of the dataset so that the machine can learn

We have used multiple classification algorithms in order to train our model such as KNN , Logistic regression ,

Support Vector Machine. After implementing all the algo's we have checked the accuracy of each model in our last phase.

**(8) Model Evaluation**

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data. Methods for evaluating a model's performance are divided into 2 categories: namely, holdout and Cross-validation. Both methods use a test set (i.e data not seen by the model) to evaluate model performance.Model Evaluation is the subsidiary part of the model development process. It is the phase that is decided whether the model performs better. After evaluating the dataset on different ML algo's we get the maximum accuracy on KNN.

**(9) References**

* Geeks for geeks <https://www.geeksforgeeks.org/machine-learning/>
* Great Learning blog <https://www.mygreatlearning.com/blog/machine-learning-tutorial/>
* Kaggle <https://www.kaggle.com/>