Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2022-23

Analytics based Marketing

NAME: Shivam Manoj Musterya

SAP ID: 60009220082

Guided By

Dr. Kriti Srivastava,

Head of the Department, CSE-Data Science, D. J. Sanghvi College of Engineering

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

INDEX:

Python Notebooks:

https://colab.research.google.com/drive/1qTX_5AQwpice4QosBQOxibuEbVYc8fhd?usp=sharing

https://colab.research.google.com/drive/1lLj4OMiumQDCT0h8xYI9n8bL21iu5LtD?usp=sharing

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

CHAPTER 1: INTRODUCTION

In today's world, where competition is fierce, companies need to optimize their marketing campaigns to maximize their return on investment. Marketers have access to vast amounts of data, and they need to use this data to identify and target the most probable buyers from their target audience. Machine learning and data analytics have proven to be useful tools in this regard. This project aims to help marketers optimize their marketing campaigns by identifying and targeting the most probable buyers from their target audience using machine learning and data analytics.

The project focuses on a case study of ABC Supermarket, a major retail player in the UK, that launched a line of organic products and wants to penetrate the market quickly. The company has shared data for 10% of their loyalty program participants who received free sample kits and recorded their purchase decisions. The objective is to target the most probable buyers from the remaining 90% using demographic and loyalty program details to optimize profitability and focus on market penetration.

The proposed solution involves building a logistic regression model using a dataset of sample loyal customers to predict the probability of buying for the remaining loyal customers. The decile methodology will then be used to determine the portion of these loyal customers that ABC Supermarket should target to optimize profitability while focusing on market penetration.

Hence, this project presents a practical approach to optimize marketing campaigns by identifying and targeting the most probable buyers from the target audience using machine learning and data analytics. The case study of ABC Supermarket demonstrates how this approach can be applied to improve the effectiveness of marketing campaigns, reduce wasteful spending, and increase gross revenue within a limited marketing budget.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

CHAPTER 2: DATA DESCRIPTION

	A	B	C	D	E	F	G	H	I	J	K
1	ID	DemA	DemA	DemClusterGro	DemGen	DemReg	DemTVReg	LoyalCl	LoyalSpe	LoyalTir	TargetB
2	0017147654	5						Tin	0.01	5	0
3	0008415498	15			M			Gold	8000	5	1
4	0012107603				M	Midlands	East	Tin	0.01		1
5	0014400995	8	28		F			Tin	0.01		1
6	0028724674	14	67					Tin	0.01	7	0
7	0041251085		65		F			Silver	3000	3	0
8	0043266179	7	41					Tin	0.01	7	0
9	0000919551	13	50		F			Tin	0.01	5	1
10	0002510294	8	36		F			Tin	0.01	9	0
11	0003912957				M	Midlands	Ulster	Silver	2000	1	0
12	0007212720			B	F			Gold	6000	1	1
13	0008332525				F	Midlands	Ulster	Silver	5000	12	1
14	0009117733				U	South East	S & S East	Gold	18633.51	3	0
15	0009344462	11	59		U			Gold	12000	9	0
16	0010315063	13		B				Gold	6000	2	0
17	0016286049		57			North	Yorkshire	Tin	0.01	5	0
18	0017046734	5				Midlands	Ulster	Silver	0.02	5	0
19	0017126703		54			Midlands	Ulster	Silver	5000	4	0
20	0021076081		71	B				Gold	6000	26	0

Fig 1: Dataset 1:

	A	B	C	D	E	F	G	H	I	J
1	ID	DemAffl	DemAge	DemClusterGroup	DemGender	DemReg	DemTVReg	LoyalClass	LoyalSpend	LoyalTime
2	0000000140	10	76	C	U	Midlands	Wales & West	Gold	16000	4
3	0000000620	4	49	D	U	Midlands	Wales & West	Gold	6000	5
4	0000000868	5	70	D	F	Midlands	Wales & West	Silver	0.02	8
5	0000001120	10	65	F	M	Midlands	Midlands	Tin	0.01	7
6	0000002313	11	68	A	F	Midlands	Midlands	Tin	0.01	8
7	0000002771	9	72	D	U	North	N West	Platinum	20759.81	3
8	0000003131	11	74	A	F	Midlands	East	Tin	0.01	8
9	0000003328	13	62	D	M	North	N East	Tin	0.01	5
10	0000004529	10	62	F	M	Midlands	East	Silver	2038.76	3
11	0000005886	14	43	F	F			Gold	6000	1
12	0000007420	7	60	F	F	North	N East	Gold	11000	2
13	0000009814	5		C	M	South East	London	Silver	5000	1
14	0000010006	9	51	F	F	Midlands	Midlands	Silver	300	11
15	0000010219	6	64	C	F	South East	S & S East	Tin	0.01	9
16	0000010812	16	37	C	F	South East	London	Tin	0.01	4
17	0000011207	8	54	D	M	Midlands	Midlands	Silver	1420	1
18	0000011982	5	70	B	F	Midlands	Midlands	Gold	6104.66	8
19	0000014656		42	C	F	Midlands	East	Tin	0.01	5
20	0000015350	7		E	F	Scottish	C Scotland	Tin	0.01	5

Fig 2: Dataset 2

The data for this project consists of two datasets: Dataset 1 (a1_Dataset_10Percent) contains information on 10% of loyalty program participants who received free sample kits, while Dataset 2 (a2_Dataset_90Percent) contains information on the remaining 90% of loyalty program participants who did not receive free sample kits.

Both datasets have the same attributes, including ID, DemAffl, DemAge, TargetBuy, etc. The ID attribute is a unique identifier assigned to each customer in the dataset, while DemAffl represents the demographic affiliation of the customer, which can be used to group them based on their social status or income level. DemAge is the age of the customer, which can be used to group them based on their life-stage, and DemClusterGroup is a clustering algorithm-generated grouping of customers based on demographic variables such as age, income, and education level. DemGender is the gender of the customer, while DemReg is the region where the customer resides, and DemTVReg is the television region where the customer resides.

LoyalClass is a loyalty program-generated classification of customers based on their loyalty behavior, and LoyalSpend represents the total amount spent by the customer in the loyalty program. LoyalTime is the time duration for which the customer has been enrolled in the loyalty program. Finally, the TargetBuy attribute is the target variable in the dataset, which indicates whether the customer is likely to make a purchase or not.

Overall, this data provides valuable information about the behavior and characteristics of loyal customers in a supermarket's loyalty program, and can be used to build a logistic regression model to predict the probability of buying for the remaining 90% of loyalty program participants who did not receive free sample kits.

However, the dataset had some issues that made the data impure and potentially inaccurate for analysis and modeling. For instance, there were missing values in some of the records, which can create a bias in the data and lead to incorrect results. This can affect the analysis and modeling outcomes, and hence, it is important to identify and resolve these issues.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

CHAPTER 3: DATA ANALYSIS

Data analysis is the process of systematically examining and interpreting data using various statistical and analytical methods to derive meaningful insights and draw conclusions. It involves collecting, cleaning, transforming, and modeling data to identify patterns, trends, and relationships that can inform decision-making and drive business growth.

Here, we have performed two tasks:

1. Data Preprocessing
2. Exploratory Data Analysis (EDA)

Data Preprocessing:

Data preprocessing is the process of cleaning, transforming, and preparing raw data into a suitable format for analysis or modeling.

In the first dataset, several data preprocessing steps were performed. Firstly, the customer ID column was dropped from the dataset as it did not contain any valuable information for analysis or modeling. Secondly, missing values were identified and handled by filling them with either the mean or mode value of the column depending on whether the data was numerical or categorical. Thirdly, categorical variables were converted into numerical ones using label encoding. This involved mapping the categorical values to integers using a predefined encoding scheme. Finally, the dataset was visually inspected to ensure that there were no remaining issues and that the data was ready for further analysis or modeling.

In the second dataset, the same preprocessing steps were applied to ensure that both datasets had a consistent format and were suitable for further analysis or modeling. By performing consistent preprocessing steps on both datasets, any differences in data format or quality between the two datasets were minimized, enabling fair comparisons and accurate predictions.

	DemAffl	DemAge	DemClusterGroup	DemGender	DemReg	DemTVReg	LoyalClass	LoyalSpend	LoyalTime	TargetBuy
0	5.0	51.0	2	0	3	3	3	0.01	5.00000	0
1	15.0	51.0	2	1	3	3	0	8000.00	5.00000	1
2	8.0	51.0	2	1	0	2	3	0.01	6.56467	1
3	8.0	28.0	2	0	3	3	3	0.01	6.56467	1
4	14.0	67.0	2	0	3	3	3	0.01	7.00000	0
5	8.0	65.0	2	0	3	3	2	3000.00	3.00000	0
6	7.0	41.0	2	0	3	3	3	0.01	7.00000	0
7	13.0	50.0	2	0	3	3	3	0.01	5.00000	1
8	8.0	36.0	2	0	3	3	3	0.01	9.00000	0
9	8.0	51.0	2	1	0	10	2	2000.00	1.00000	0
10	8.0	51.0	1	0	3	3	0	6000.00	1.00000	1
11	8.0	51.0	2	0	0	10	2	5000.00	12.00000	1
12	8.0	51.0	2	2	3	8	0	18633.51	3.00000	0
13	11.0	59.0	2	2	3	3	0	12000.00	9.00000	0
14	13.0	51.0	1	0	3	3	0	6000.00	2.00000	0
15	8.0	57.0	2	0	1	12	3	0.01	5.00000	0
16	5.0	51.0	2	0	0	10	2	0.02	5.00000	0
17	8.0	54.0	2	0	0	10	2	5000.00	4.00000	0
18	8.0	71.0	1	0	3	3	0	6000.00	26.00000	0
19	8.0	47.0	4	0	3	3	3	0.01	8.00000	0

Fig 3: Dataset after preprocessing

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is the process of examining and visualizing a dataset to uncover patterns, relationships, and trends that can help inform decision-making. The goal of EDA is to understand the structure and nature of the data, identify potential issues or outliers, and generate insights that can be used to make data-driven decisions.

Here, various EDA techniques have been employed to investigate different aspects of customer behavior, such as loyalty classification, spending patterns, demographics, and purchase behavior.

Scatterplot and Distribution Plots: A scatterplot shows the relationship between two variables, while the distribution plots show the distribution of each variable. This technique has been used to identify patterns or correlations in the data, such as identifying any strong positive or negative correlations between different variables, or outliers or unusual patterns in the data.

Boxplots: Boxplots have been used to investigate the relationship between customer loyalty classification (LoyalClass) and loyalty spend (LoyalSpend). This technique helps in identifying which loyalty program tiers are the most profitable and can inform decisions about how to structure and incentivize loyalty programs.

Histograms: Histograms have been used to analyze the age and demographic affiliation of customers who made a purchase. This technique helps tailor marketing campaigns towards specific age and demographic groups that are more likely to make a purchase.

Stacked Bar Plot: Stacked bar plots have been used to investigate the relationship between DemAffl and TargetBuy. This technique helps in targeting marketing campaigns towards certain demographic groups that are more likely to make a purchase, allowing for more effective marketing strategies.

Combined Techniques: Various EDA techniques have been used together to investigate the relationship between LoyalTime, DemAge, DemAffl, LoyalClass, and LoyalSpend. This allows for a more comprehensive analysis and helps in developing targeted retention strategies to keep customers engaged and coming back for more.

Examples of EDA are shown below:

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

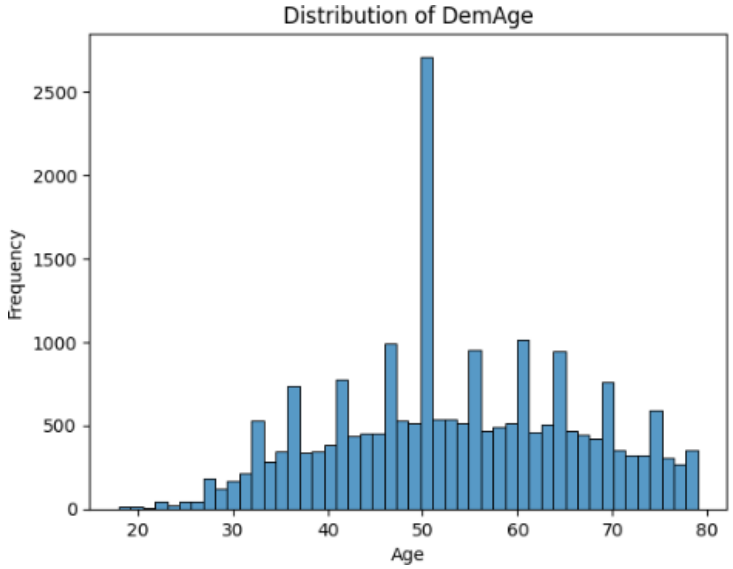

Fig 4: Histogram for Age Distribution

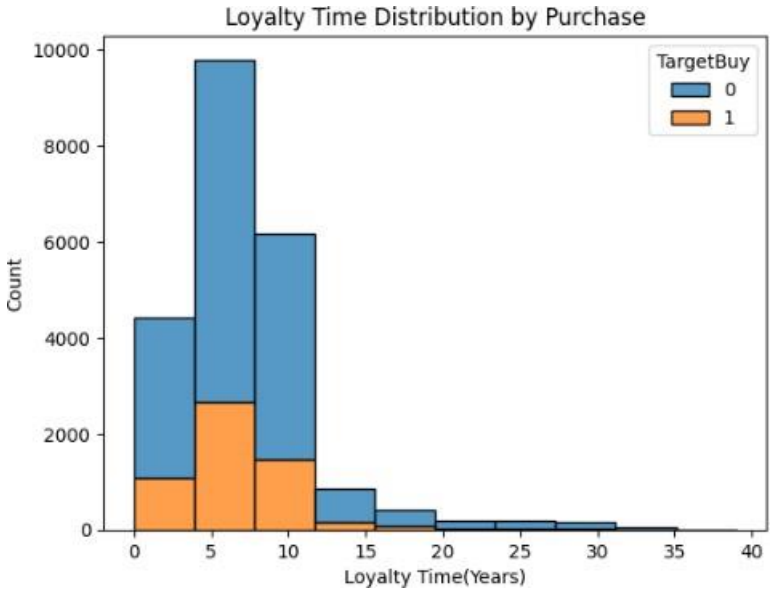

Fig 5: Histogram of LoyalTime distribution

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

4: DATA MODELLING

>> Reason to select ML model

Logistic regression is a popular algorithm for binary classification problems such as predicting customer behavior, in this case, buying or not buying. Logistic regression is a statistical method that estimates the probability of a binary outcome by examining the relationship between the input variables and the binary response variable. The algorithm learns from the given input features and then computes the probability of the binary outcome, making it useful in predicting customer behavior. Logistic regression is also computationally efficient and can handle large datasets, making it a good choice for marketing analytics tasks.

>> Logistic Regression Algorithm

Logistic regression is a statistical algorithm that is commonly used for binary classification problems, where the goal is to predict a binary outcome (e.g. yes or no, true or false) based on one or more input features.

The logistic regression algorithm models the relationship between the input features and the binary outcome by estimating the probability of the outcome. It does this by using a logistic function (also known as a sigmoid function) to map any real-valued input to a probability score between 0 and 1.

Once the logistic regression model is trained, it can be used to make predictions on new data by computing the probability of the positive class for each input, and then assigning the class with the highest probability as the predicted class.

Logistic regression is a classifier because it predicts the probability of the binary outcome, which can then be thresholded to obtain a binary classification decision.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

>> Building the model

The first step is to create an instance of the logistic regression classifier and fit it to the training data using the fit() method. The trained model is then used to predict the target outputs for a new dataset using the predict() method. The model is then saved using joblib.dump() function, which allows the classifier object to be exported to a file.

```
classifier =  LogisticRegression(max_iter=300)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
```

Fig 6: Applying the model

The confusion_matrix() and accuracy_score() functions are used to evaluate the performance of the trained classifier on the test data. The confusion matrix provides information on the number of true positives, true negatives, false positives, and false negatives. The accuracy score measures the proportion of correct predictions made by the model.

```
[[3191  176]
 [ 688  390]]
```

```
print(accuracy_score(y_test, y_pred))

0.8056242969628796
```

Fig 7: Confusion matrix and Accuracy of the model

The predict_proba() method is used to obtain the predicted probabilities of each sample belonging to class 0 and class 1. These probabilities are stored in a DataFrame along with the true binary classification labels and the input features for each sample in the test data. The resulting DataFrame is then sorted in descending order of predicted probabilities to apply the decile methodology.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

The decile methodology involves grouping the observations into ten equal parts based on their predicted probabilities. The resulting pivot table shows the total number of customers in each decile, the probability threshold above which a customer is predicted to buy the product, the number of customers who actually bought the product, the percentage of Good customers in each group, the number of customers who did not buy the product, the cumulative number of Good and Bad customers up to that decile, the percentage of cumulative Good and Bad customers up to that decile, and the percentage of Bad customers avoided by targeting that decile.

A1 | fx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		0	1	2	3	4	5	6	7	8	Actual Outcome	prob_0	prob_1	Decile
2	4134	29	29	2	0	3	3	0	13500	2	1	0.60%	99.40%	1
3	3531	30	38	4	0	2	1	2	1000	2	1	0.70%	99.30%	1
4	3041	28	33	5	0	0	11	3	0.01	8	1	0.86%	99.14%	1
5	317	31	35	5	1	1	7	2	300	8	1	0.96%	99.04%	1
6	3816	25	38	2	0	0	4	2	2000	4	1	2.39%	97.61%	1
7	3404	24	35	2	0	1	7	2	1000	5	1	2.64%	97.36%	1
8	4094	23	34	3	0	0	11	2	1000	2	1	3.14%	96.86%	1
9	3643	25	29	0	1	0	4	0	6000	2	1	3.17%	96.83%	1
10	985	24	46	5	0	0	4	2	2000	4	1	4.16%	95.84%	1
11	1508	20	30	2	0	1	7	2	600	4	1	5.06%	94.94%	1
12	1881	21	36	5	0	1	12	3	0.01	8	1	5.36%	94.64%	1
13	2406	22	41	3	0	0	11	0	6000	1	1	5.37%	94.63%	1
14	3392	20	34	5	0	3	3	2	50	5.564670495	1	5.51%	94.49%	1
15	954	21	37	4	0	3	3	3	0.01	10	1	5.57%	94.43%	1
16	1236	20	35	5	0	0	2	3	0.01	5	1	5.85%	94.15%	1
17	2424	20	34	4	0	0	4	3	0.01	5	1	5.86%	94.14%	1
18	1651	20	33	1	0	3	8	2	50	11	1	6.12%	93.88%	1
19	1613	19	31	5	0	3	3	3	0.01	5	1	6.35%	93.65%	1
20	427	24	51	3	0	3	8	0	11500	1	1	6.37%	93.63%	1

Fig 8: Result after applying Decile Methodology

To apply the model obtained from Dataset 1 to the new Dataset 2, the same set of input variables (X) was used for both datasets. The only difference is that the new dataset lacks purchase decisions, whereas the original dataset has purchase decisions.

First, the 90% dataset was imported into the notebook and the same nine input features were defined for this dataset as those used in building the model for Dataset 1. These input features could be the customer's demographic information, purchase history, and any other relevant information that can be used to predict the likelihood of buying.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Next, the classifier coefficients file that was exported to the Google Drive in the model building part using Dataset 1 was loaded. This file contains the trained model parameters that enable us to make predictions on new data.

```
] y_pred = classifier.predict(X_fresh)
  print(y_pred)
```

```
 [0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
  0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0
  1 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0
  1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1
  0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
  1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
  1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0 0 0
  0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
  0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0
  1 0 1 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0
  0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
  0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
  0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0
  0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
  0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0
  0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1
  0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1]
```

```
] predictions = classifier.predict_proba(X_fresh)
  predictions
```

```
 array([[0.97549312, 0.02450688],
        [0.96899159, 0.03100841],
        [0.94465062, 0.05534938],
        ...,
        [0.96593599, 0.03406401],
        [0.76761016, 0.23238984],
        [0.49754322, 0.50245678]])
```

Fig 9: New predictions made by the model

Using the predict_proba() function, the model's predicted probabilities for buying or not buying were computed for each customer in the new dataset. These predicted probabilities provide valuable information for businesses as they can identify customers with the highest likelihood of buying.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

CHAPTER 5: RESULT ANALYSIS

Firstly, we used the first dataset to build a predictive model that predicts the probability of a customer buying a product based on nine features. We then applied this model to the second dataset, which lacked purchase decisions, to predict the probabilities of buying or not buying for each customer. This allowed us to generate an output file that contains the model predicted buying probabilities for each customer.

Using this output file, the client can now target the most probable buyers based on their strategic objectives. For example, if the client wants to maximize profit, they can target the top 30% probable buyers. Alternatively, if the client wants to focus on market penetration with a slight compromise on profit, they can target the top 40% probable buyers as shown below in the pivot table.

	Strategic Option	Participants Covered (A)	% Cum. Good to Cum. Total (B)	% Total Buyers Reached	% Total Non Buyers Avoided	Probability Threshold	Profit Booked (in Mn INR)
No Model Scenario	All 100%	225,000	24%	100%	0%	0	- 176
Market Penetration	Top 40%	90,000	44%	72%	70%	24.4%	196
Profit Maximisation	Top 30%	67,500	51%	63%	80%	31.1%	214

Fig 10: Strategic marketing options for 90 percent Loyalty Base

Our approach had minimal operational cost to the business since we used Google's free computational resources. Based on the client's chosen strategy, they can gain up to 214 million rupees in profit. Overall, our analytics-enabled marketing strategy provides the client with a data-driven approach to optimize their marketing campaign and maximize their profits.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

CHAPTER 6: FUTURE SCOPE & CONCLUSION

We can explore opportunities to improve the predictive model for lead-to-sale conversion by trying out different feature engineering techniques, exploring more advanced machine learning algorithms, or even considering using deep learning models to see if we can enhance the model's accuracy. In addition, we can integrate the predictive model with Customer Relationship Management (CRM) software to automate the process of identifying the most promising leads and nurturing them through the sales funnel. To make quick decisions based on real-time data, we can work on developing a model that can make real-time predictions based on incoming data, as the current model uses historical data to make predictions.

In conclusion, this marketing analytics project using logistic regression and the decile methodology has enabled us to develop a predictive model for identifying the most profitable customers for our client's loyalty program. By analyzing a dataset of sample loyal customers and building a model to predict the probability of buying for the remaining customers, we have created a data-driven approach that can help our client optimize their loyalty program and increase profitability.

Our approach involves using the decile methodology to determine the portion of loyal customers that our client should target to optimize profitability while focusing on market penetration. By targeting the right audience, our client can increase profits through their loyalty program.