Personalized Ad- Analytics for small businesses

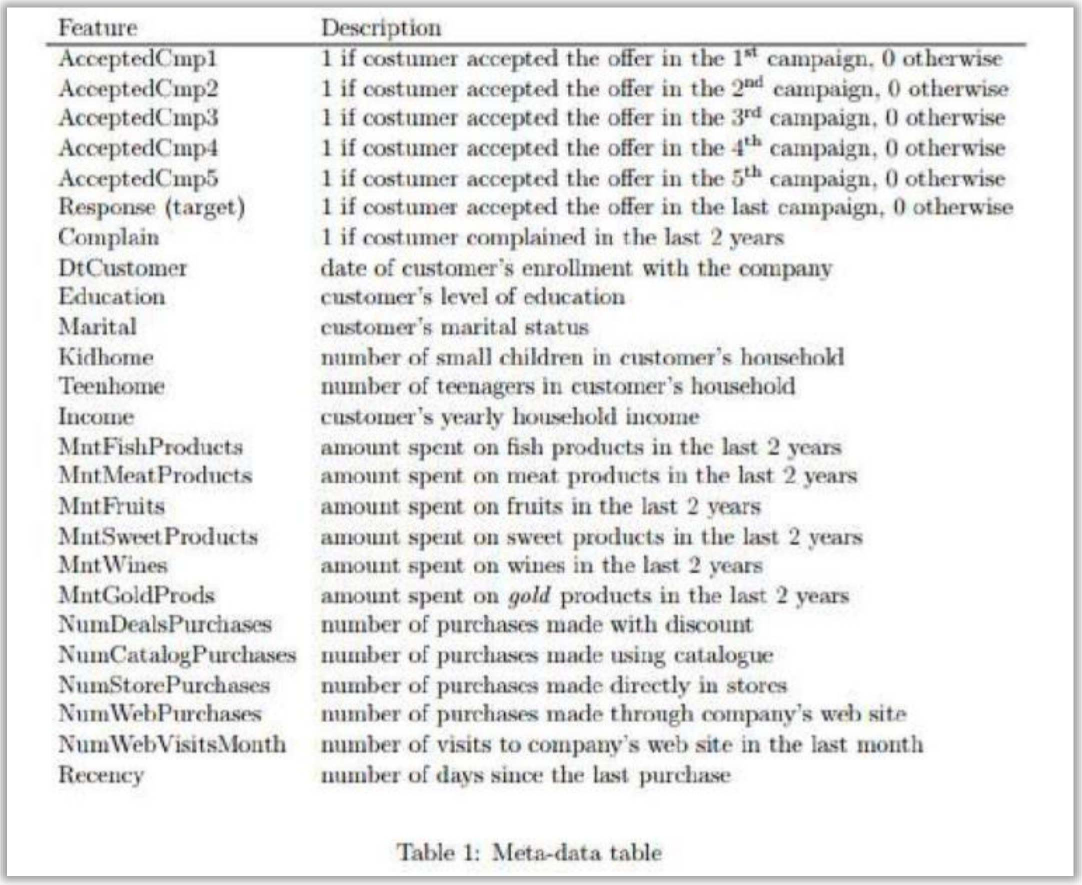
Team: - Data Explorers

A) Problem Statement: -

With the emergence of giant e-commerce firms such as Amazon or retail corporations such as COSTCO and Walmart, it is extremely crucial for new growing or small businesses to use various marketing methods to promote their business. Marketing campaigns cannot always be tailored to each individual client, but they do have the flexibility to send many adverts combined with imagery to a diverse range of individuals. Family members may be drawn to several types of advertisements, whereas single working professionals may be drawn to commercials that interest them. The issue for marketing teams arises here; they may know who they are attempting to reach, but they want to know what handful of advertising they should create to appeal to a broad range of individuals without splintering their campaign into thousands of tailored ads, which may be prohibitively expensive.

B) Data Source: -

We are using the Kaggle data set: - Kaggle | Marketing Analytics
In this dataset holds the information related to customer and their purchases over 3 years from 2012- 2014. There are 28 features in the dataset and about 2300 tuples as shown in the below image



Feature	Description
AcceptedCmp1	1 if costumer accepted the offer in the 1^{st} campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2^{nd} campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3^{rd} campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4^{th} campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5^{th} campaign, 0 otherwise
Response (target)	1 if costumer accepted the offer in the last campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on *gold* products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days since the last purchase

Table 1: Meta-data table

C) Methodology:

I. DATA PRE-PROCESSING:

Data available to us from Kaggle has some additional information which plan to clean, so that we will be able to proceed with analyzing the data and perform initial EDA on the same. For the first instance, the type of cleaning of data that we will need to perform are related to the following:

i. Income – This variable currently consists of string data type value, e.g., $150, here after importing the data we will need to perform string manipulation and conversion operations so that we are able to use this column for numerical analysis

ii. Education – This variable currently consists of string data type value, e.g., Masters or PhD, here after importing the data we will perform data manipulation operations to the column to convert this column to ordinal data, which will refer to these values.

iii. Marital Status- This variable is a categorical variable changed it to ordinal.

iv. Converted the Year of birth column to Age.

v. Converted the Customer joining date to number of customer days till today.

II. EDA: -

We currently have sufficient data of customers, their purchase history and past campaign data. Our initial task will be to try and understand the data through various EDA techniques, also identify if any outliers are available that may help us identify any issues that might help with modelling.

Below are few EDA findings on the dataset :-

i. Data for income variable required cleaning to be useful

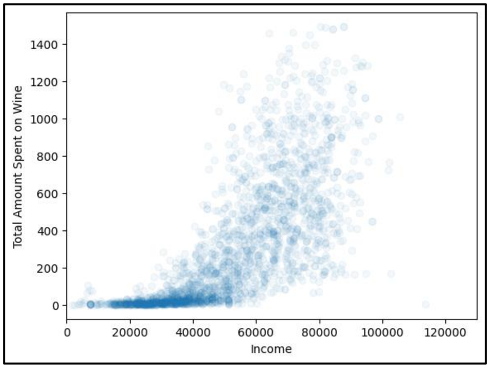
ii. Income data had few empty records which needed to be imputed

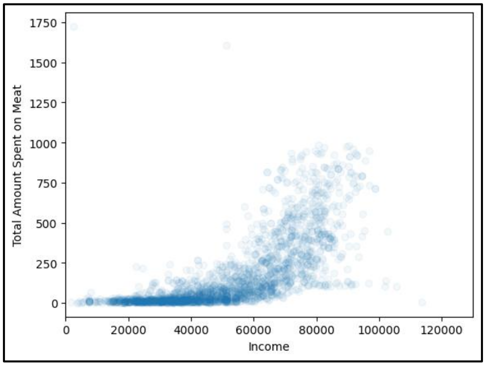
iii. Converted multiple columns from String to Ordinal values

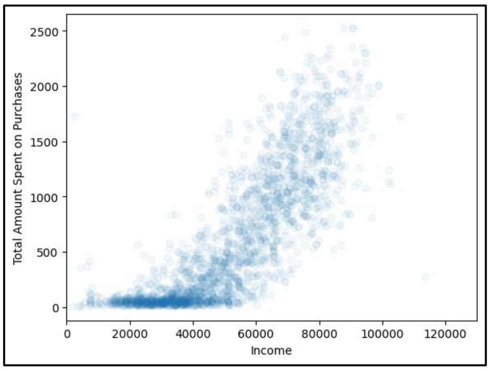
iv. From the heat map and the scatter plot high income we can observe a positive correlation between Total purchases and Total amount spent i.e., High Income individuals tend to spend more money and purchase more.

v. From the heatmap we can also understand that the Income is not correlated with Number of Discounted Purchases (numDealPurchases attribute), that means High income people tend to purchase fewer items with discount.

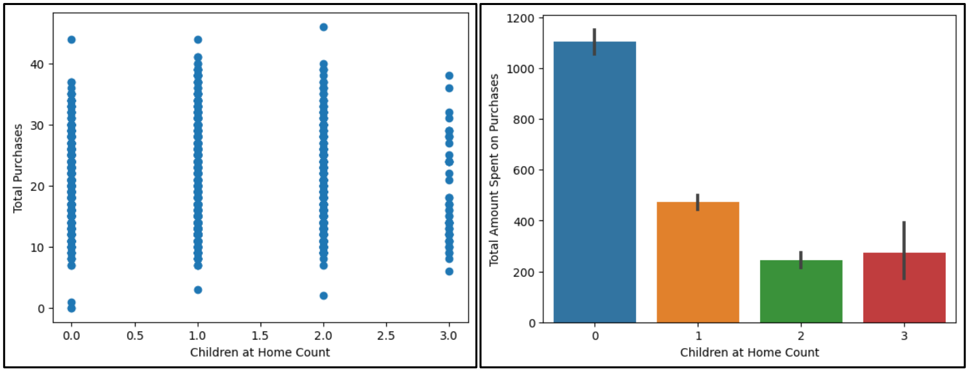
vi. From the scatter plot and heatmap we can observe that people with high income tend to buy wine & meat with higher price (quality) , while individuals with less income tend to buy wine and meat in more quantity but of cheaper price (quality)



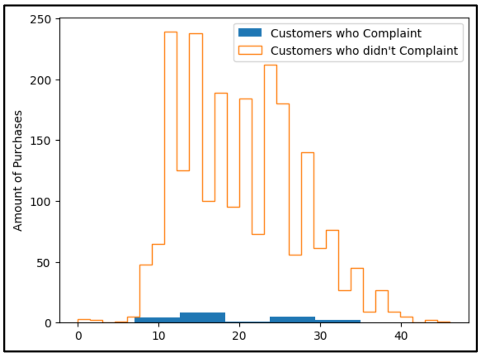




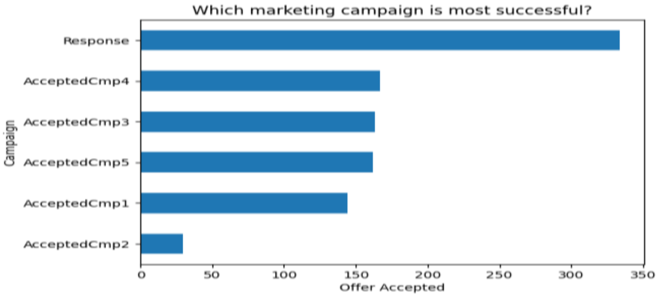
vii. There is no correlation between People having kids at home and Total number of purchases, however individuals with no children tend to spend more vs individuals with children. This might be due to more savings required for families with children.



viii. Here we can see drastic difference in customer purchases amongst people who have complained vs didn't complain, also helps us identify the importance of customer satisfaction for longer lasting business relations with our customers.



ix. Marketing Campaign Analysis- we found that the last campaign was successful as many people accepted the offer or deal in the last campaign.



III. Algorithms: -

i. We have customer data, relevant consumer purchase data, and past campaign information. We also have a response column that indicates whether the consumer accepted the prior campaign. We will attempt many techniques and compare the models to get the best fit for our dataset and forecast this response characteristic.

ii. One of the issue statement's main criteria is to discover items that would boost earnings for future campaigns. We shall use the following methods to forecast this. Classification modelling is used to determine the relationship and likelihood of a client purchasing specific items throughout a campaign.

iii. We will build a Decision Tree to determine whether the qualities can help us anticipate the proper values for campaign usage; however, determining which elements of the dataset can be used to correctly predict will be tough.

iv. The K-NN approach will assist us in identifying the sorts of clients who are most likely to respond depending on the types of items we have on campaign, hence increasing our earnings.

v. We may also use the Random Forest model to anticipate the number of store purchases, and then apply the feature significance method to rank the qualities that are highly related to the number of store sales.

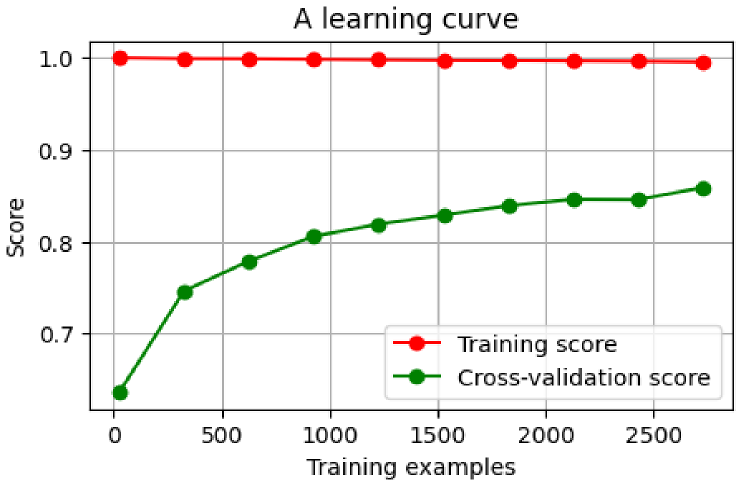
vi. Based on this, we will get our F-score from the prediction table for each algorithm to determine which method has the highest F-score and accuracy.

IV. Modelling : -

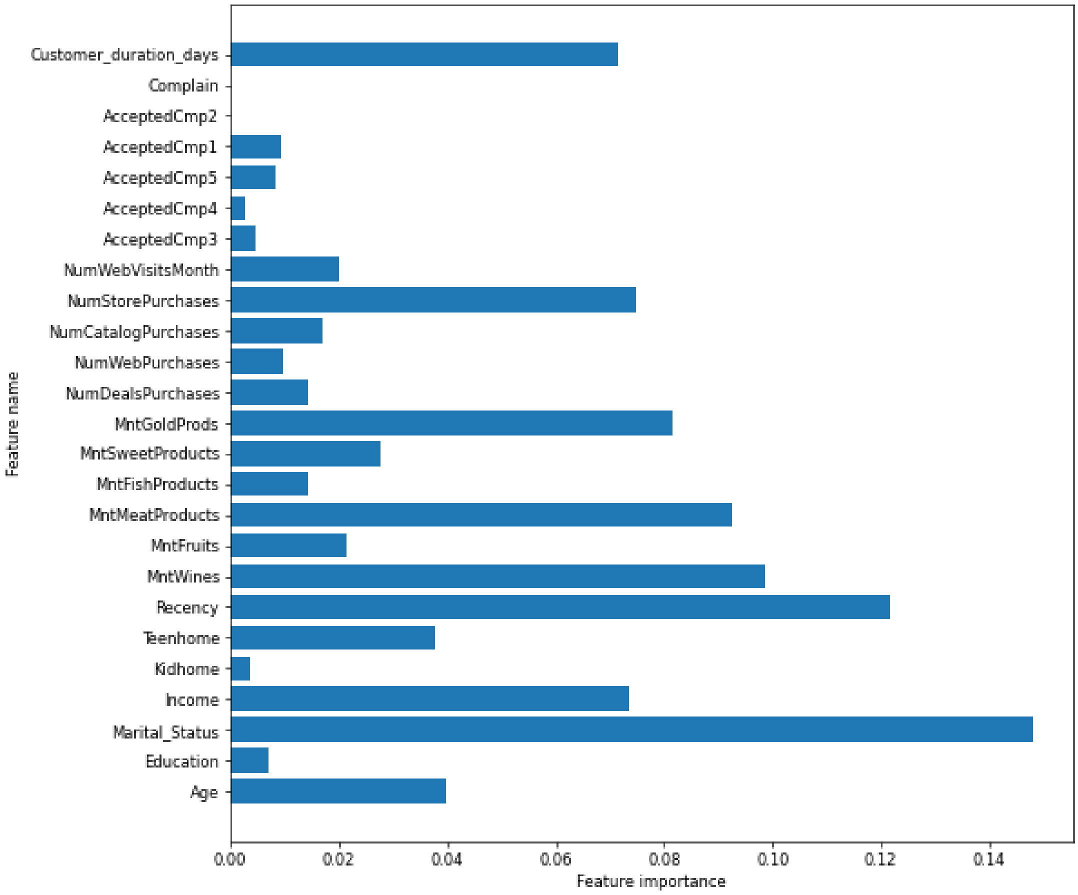
i. Various modelling methods are used to predict the campaign score by using supervised learning methods, along with methods to tackle overfitting the training data.

ii. Overfitting issue with Data was noticed due to a smaller number of records for Minority class which led to low precision and low recall while using Decision Tree classification. So, SMOTE(Synthetic Minority Oversampling Technique) with Random Under Sampling of majority was used to increase the recall rate for minority class.

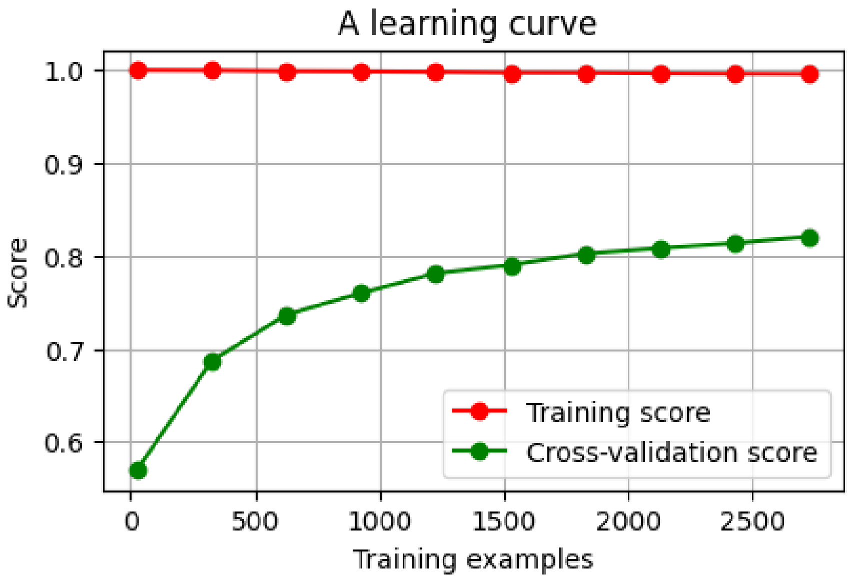
iii. Decision Tree with K Fold Model was used to identify how the model is learning and predicting the values.



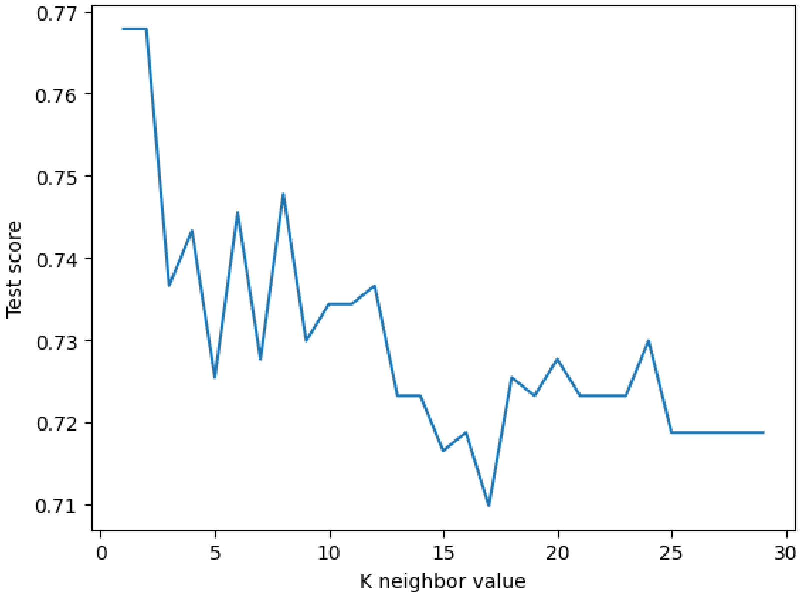
iv. Important Features were classified from the data using Tree classifier, which helped identify the most import features in the data.



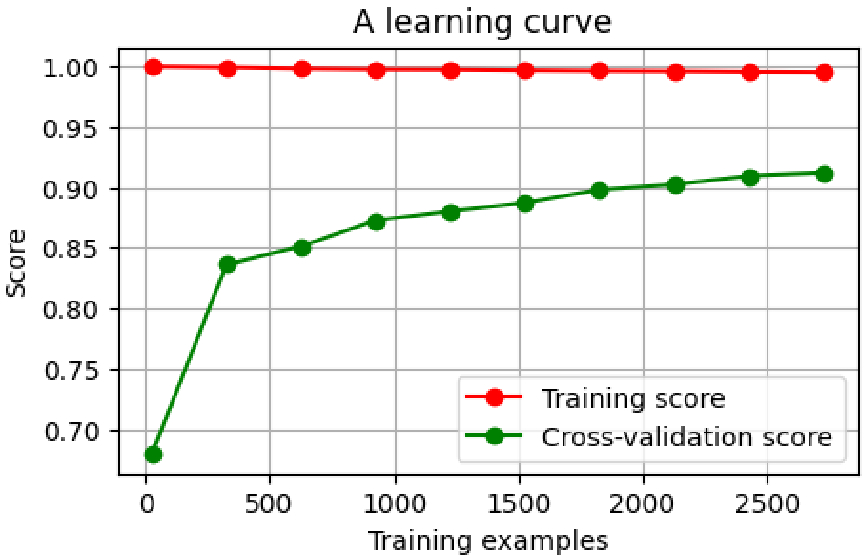
v. Next model used on the data to predict is K-Nearest Neighbor Model below is the learning curve for the same



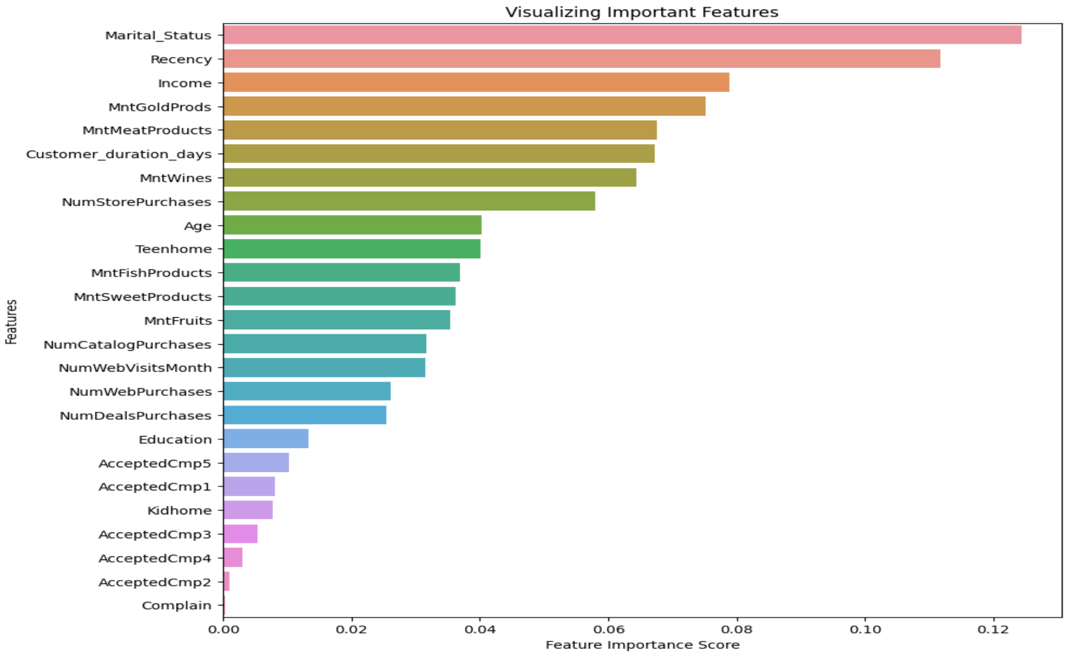
vi. Hyper Parameter tuning or K of KNN shows 3 to be the best value to predict



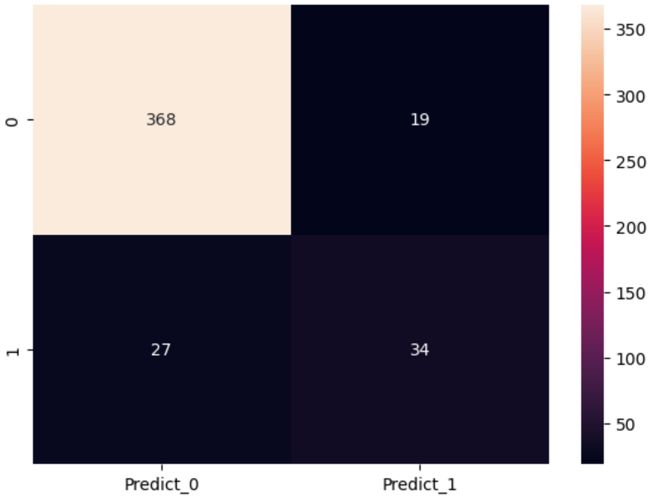
vii. Next model used on the data to predict is Random Forest Classification Model below is the learning curve for the same



viii. Grid Search with K-Fold were applied on to the Random Forest Classification model to identify best Parameters and most important features



ix. Confusion Matrix and classification report for Random Forest Classification :-



	precision	recall	f1-score	support
0	0.931646	0.950904	0.941176	387
1	0.641509	0.557377	0.596491	61
accuracy	0.897321	0.897321	0.897321	0.897321
macro avg	0.786578	0.754141	0.768834	448
weighted avg	0.89214	0.897321	0.894244	448

Model Analysis:

- KNN performance was very low with an accuracy of 72%
- Decision Tree Classifier gave an accuracy of 85%
- Random Forest Classifier performed well with an accuracy of 88% higher than other models

V. Risks: -

i. Although product and consumer purchase histories are supplied, basing research on them may be problematic. We must give analysis for customer and product relationships based on purchase history to see whether there is a pattern of growing or declining buy rates over time.

ii. Income categories are significant for analyzing client purchasing power and how it relates to various product kinds. However, there is always the risk that the information is incorrectly provided, as many surveys may contain incorrect or missing data.

iii. Future campaigns require analysis to improve sales and profit for the organization, and if the predictability of the model is low, the organization may incur heavy losses.

VI. Challenges: -

i. Customer survey information for campaign response does not include historical information older than two years, resulting in a smaller dataset and the inability to use data-hungry algorithms to understand correlations.

ii. The data does not provide information to determine if any currently available product should be priced higher or lower to boost profit margins or sales volume.

iii. Product bundling, based on the data provided, can we determine what kind of items offered for purchase are more likely to be purchased together?

VII. Completion Plan :-

i. EDA of dataset with initial report – 4th November 2022

ii. Train data Model with Classification and Decision Tree with different randomized datasets – 11th November 2022

iii. Train data Model with KNN and Random Forest models with different randomized datasets – 18th November 2022

iv. Compare results of various models and prepare learning curve for improvement – 20th November 2022

v. Improve the model according to the results and select the best model for predicting – 22nd November 2022

vi. Validation and Testing on Data by – 25th November 2022

VIII. Citations :-

i. SKLearn - Classification

ii. SKLearn - KNeighborsClassifier

iii. SKLearn - Decision Tree Classifier

iv. SKLearn - Random Forest Classifier

v. Kaggle - Dataset