Universidade NOVA de Lisboa

Information Management School

# Spice Alley

*Data Science & Machine Learning*

*Author:*
Celso Cassamá
Student ID: 20222138

*Author:*
Guilherme Silva
Student ID: 20221209

*Author:*
Murilo Oliveira
Student ID: 20222137

*Author:*
Nuno Marques
Student ID: 20222145

2023

# Contents

# List of Tables



# List of Figures

# 1 Introduction & Methodology

## 1.1 Introduction

This project's aim is to develop a predictive model that is able to maximize Spice Alley's profit in the upcoming marketing campaign. This campaign is geared towards promoting a new frozen food product to the customer database. In order to achieve this goal, it is crucial that the model is able to accurately select the customers more prone to purchase the campaigned product, disregarding those less likely to act.

## 1.2 Methodology

The CRoss Industry Standard Process for Data Mining (CRISP-DM) methodology served as the basis for this project, dividing it into the following sections: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

# 2 Business Understanding

Typically, businesses when executing a promotional campaign pay a fixed amount per contacted or reached customer. Therefore, in order to have a positive ROI (Return-on-Investment) it is important that a high percentage of contacted customers buy the promoted product, otherwise the money spent in the campaign will not generate increased future sales. In Space Alley's case, there is an estimated cost of $3 per contacted customer, while each customer that accepts the offer contributes $16 in revenue. Therefore, to be a successful campaign, it must have at least 19% of acceptance rate to break-even. It is our aim to develop a model, capable of providing this level of acceptance rate by the existing customer database. Interpreting the historical performance of previous campaigns it is possible to identify there are 544 observations that responded at least to one of the 5 campaigns, amounting to 21,7% of the sample. Further analysis to this dataset suggests a significant overlap between those who accept the offer (DepVar = 1) and those that responded positively to at least one of the campaigns (Response Campaign 1/2/3/4/5 = 1). While in the pilot campaign only 12.5% of customers accepted the offer, 60% of those who responded positively to one of the campaigns are also in the group that accepted the offer. This fact suggests that Campaign Responses have a high business value while trying to predict customer response to the acquisition of the new product.

# 3 Data Understanding

The Data Understanding stage focuses on getting to know the data given on the project resources, identify data quality and get some initial insights. To do so, several steps were performed, from loading and merging data, looking for duplicates and understanding the data types of each variable. Moreover, a first glance at the data, simple statistical analyses (summary statistics, kurtosis, and skewness), variables distribution and the correlation between variables, also enabled the gathering of valuable information on Spice Alley's available dataset described below.

## 3.1 Basic Exploration

The dataset initially had 18 duplicated rows which were immediately removed. The resulting dataframe had 5000 rows and 29 columns. It had different data types, namely, 2 float64, 4 object and the rest int64. Furthermore it had missing values on Education (32), Recency (48), and MntDrinks (21). Regarding summary statistics, these are the main insights we gathered: In the numerical variables, the Income had a minimum of (2678,00), a maximum of (237117,00), the mean (77456,27), median (76579,00), and standard deviation of (355151,56), such values suggest the presence of outliers with a right skew, due to very distant maximum value either from mean and median values, as well as, a large standard deviation.

Comparable results were found in other variables too, such as on all the monetary ones (MntMeat & Fish, MntEntries, MntVegan & Vegetarian, MntDrinks, MntDesserts, and MntAdditionalRequest), NumOfferPurchases, and NumTakeAwayPurchases (see Annex - Statistics). Skewness wise, we saw moderate skew (between —0.5— and —1.0—) in the Income, Kid_Younger6, NumAppPurchases, NumStorePurchases, NumAppVisitsMonth variables, and high skewness (higher than —1.0—) in the MntMeat & Fish, MntEntries, MntVegan & Vegetarian, MntDrinks, MntDesserts, MntAdditionalRequests, NumOfferPurchases, NumTakeAwayPurchases, Complain, Response_Cmp1, Response_Cmp2, Response_Cmp3, Response_Cmp4, Response_Cmp5, which indicate that the distributions were not Gaussian distributions that could impact the performance of some algorithms. At

last, and yet regarding numerical variables, the kurtosis has higher than 3 (kurtosis from Gaussian distribution) on the following variables: MntEntries, MntVegan & Vegetarian, MntDrinks, MntDesserts, NumOfferPurchases, NumAppVisitsMonth, which again indicate the presence of outliers.

Regarding the categorical variables: we understood that the Name variable has the title (Mr., Miss, and Mrs.) before the actual name, Education has 9 levels (Graduation, PhD, Master, HighSchool, Basic, master, graduation, phd, highschool) although we can perceive that there should only be 5 levels, the Marital_Status has 10 levels (Married, Together, Single, Divorced, Widow, married, together, single, divorced, widow) but only 5 of them are different according to human standards, and Date_Adherence has the following format "2020-09-19 00:00:00" but is of the object type, meaning that we will need to transform it to perform calculations based on dates.

# 4 Visual Exploration and Data Preparation

After an initial exploration, it is necessary to prepare data to be successfully introduced in the model. This step comprises the removal of outliers, duplicates and missing values. Additionally, incoherencies were also dealt with and new features were created in order to maximize the value extracted from the original dataset. A brief explanation of each step is provided below.

## 4.1 Data Cleaning

### 4.1.1 Duplicates

The initial dataset had 2518 rows, from which 18 duplicate rows were removed. We then proceed to the next preparatory steps with 2500 rows.

### 4.1.2 Outliers

There was further data treatment applied to this dataset, regarding the removal of outliers. Through visual exploration (images 5, 4, 2, 1, 3), it was possible to detect outliers in Income (31), MntVegan&Vegetarian (25), NumTakeAwayPurchases (26), NumAppVisitsMonth (22), and NumOfferPurchases (31). As some entries had outliers in more than one category, the combined value of outliers is 62, which is below the 3% recommended threshold. Therefore, all of these outliers were removed, and the dataset was kept with 2438 rows.

### 4.1.3 Missing Values

It was identified the existence of missing values in the following columns: Education (33), Recency (48) and MntDrinks (21). Education and Recency were filled with mode and mean, respectively. For MntDrinks the decision was to fill it using the KNNImputer method.

### 4.1.4 Encoding

Gender was encoded (1 - Male, 0 - Female), after first extracting, after splitting the Title in the "Name" column. This column was then dropped as it didn't add value to the model. The encoding process is required as the used model allows only numeric values as inputs.

### 4.1.5 Dates and Standardization

The columns "Date Adherence" and "Birthday" also required cleaning and preparation before the introduction of its values on the model. Regarding "Date Adherence" it had an incorrect date (29/2/2022) that was corrected. Following this correction, it was modified to "Days as Client". Applying this same procedure, the column "Birthday" was also modified to "Age" thus representing the actual valuable variable from a business perspective.

"Marital Status" also had inconsistencies on capitalization and description. A procedure to standardize the column values was made to transform the values into one of the following ("Married, Single, Widow and Divorced).

## 4.2 Data Transformation

### 4.2.1 New Variables

To gather further insights, the creation of more meaningful variables is a common practice. It seemed suitable to create two additional variables: *Response Campaigns*, which is the sum of *Response Campaigns 1/2/3/4/5*, and *Total Kids*, which sums *Kid_Younger6* and *Children_6to18*. The goal behind the creation of these variables was to consolidate the information that was previously spread out among multiple variables into a single variable.

#### 4.2.2 Incoherencies

There was an intention to check if real world incoherencies were present in the dataset:

- Customers with monetary spending but without purchases;
- Customers with Purchases using offers but without any purchase on the available channels (App, TakeAway or Store); After the removal of outliers, no incoherencies were found.

#### 4.2.3 Skewness Correction

From the visual exploration have seen non-normal distributions, on the variables: MntMeat& Fish, MntVegan& Vegetarian, MntEntries, MntDrinks, MntDesserts, MntAdditionalRequests, NumOfferPurchases, NumTakeAwayPurchases. To try to make it more normal-like distributions we've applied logarithm base 10+1 on them.

### 4.3 Feature Selection

In this step there was the application of various feature selection techniques, including Variance, Spearman Correlation, Recursive Feature Elimination (RFE), Lasso Regression, and Decision Trees. To address the sensitivity of regression models to correlated features, it is advisable to remove such features before utilizing techniques like Lasso Regression and Recursive Feature Elimination.

The feature selection techniques should be applied in the following sequence: (1) Variance, to identify constant variables; (2) Spearman Correlation, to detect correlated features; (3) Decision Trees, to retain one variable from a group of correlated features; (4) RFE, to iteratively select features by considering different subsets of features; and (5) Lasso, to identify and select important features in the dataset.

Additionally, it is important to determine the data type of each variable before applying feature selection techniques, as different techniques are specific to different data types. Finally, we applied a Stratified K-fold technique, using 5 splits in order to correct for imbalances on the dataset.

#### 4.3.1 Variance Analysis

The analysis revealed that the features "CostContact" and "Revenue" exhibit a constant value across the dataset. Therefore, these variables were removed as they do not provide additional information.

Additionally, the features "Complain" has very low variance, indicating limited variation within its value. However, considering its potential significance from a business perspective, it was retained acknowledging its potential usefulness in explaining the business context.

#### 4.3.2 Categorical Variables Analysis

In order to decide on maintaining or excluding certain categorical variables we used the Chi-Squared test for categorical variables. The results are shown in table 1. As such we decided to exclude 'Education' and maintained 'Marital_ Status'.

#### 4.3.3 Spearman Correlation

The use of this method aims at identifying features with high correlation ($|corr| > 0.8$) between them and excluding one variable of the pair, as highly correlated variables tend to not provide additional value when modeled together. In this case, the highly correlated pairs (12 between the 4 features) were:

- Income
- MntMeat & Fish
- MntVegan & Vegetarian
- NumTakeAwayPurchases

Therefore we kept only MntVegan& Vegetarian as an explainer.

#### 4.3.4 Decision Tree Classifier Analysis

Feature importance was also evaluated using Decision Trees. Through the use of this method as shown in image 6 the top 5 features were:

- Recency
- Response_ Campaigns
- MntVegan & Vegetarian
- NumOfferPurchases
- NumStorePurchases

#### 4.3.5 RFE

Recursive Feature Elimination helps identify the subset of features that contribute the most to the model's predictive performance while discarding less informative or redundant features. The obtained results regarding the most relevant features were:

- Recency
- Response_ Campaigns
- MntVegan& Vegetarian

- NumOfferPurchases
- NumAppVisitsMonth

#### 4.3.6 Lasso Regression

Finally, a Lasso Regression was also used to further extend the evaluation process of the features to include in the model. The most relevant features according to this technique were:

- Response_ Campaigns
- MntVegan& Vegetarian
- Recency
- NumOfferPurchases
- NumAppVisitsMonth

## 5 Modeling

The modeling phase starts with two datasets, the keep_ data and all_ data with the features that come out from the feature selection phase. A pipeline to understand which dataset would give us better results was used. The LogisticRegression, KNeighborsClassifier, DecisionTreeClassifier (max_ depth = 3), and the MLPClassifier (max_ iter = 2000) with the default parameters, except the ones mentioned before, were chosen to build the final model. The models were applied to both datasets and the results were according to Table 3, Table 4 for all_ data and keep_ data datasets, respectively. The StratifiedKFold split (with n_ splits = 5) was used and the missing values were filled only after the split between the train and the validation dataset, with the models being fitted to the training set.

The data suggest to use the all_ data but the keep_ data were consistently getting better results in kaggle, so the keep_ data set was chosen for the Model Optimization.

### 5.1 Model Optimization

First, the Manual optimization approach for the K-Nearest Neighbor (KNN), DecisionTreeClassifier (DT), and MLPClassifier was tried as it requires less processing power.

For KNN, figure 7 suggests that n=3 is the better parameter since it's the higher value of the f-1 score for the validation dataset. Regarding, DT model figure 8 suggests that max_ depth = 3 is the better choice. Lastly, for the MLPClassifier the figure 9 suggest that the hidden_ layer_ sizes should be around 300 or 350, considering all the other parameters the default ones, those are the values tried [(100),(150),(200),(100,100),(300),(150,150),(350),(400)]. Due to the better result shown using the MLPClassifier the RandomizedSearchCV and GridSearchCV were used to fine-tune the parameters of the model, and returned the best parameter and best F1-Score for which of them. Best RandomizedSearchCV - Best Hyperparameters: 'solver': 'adam', 'learning_ rate_ init': 0.1, 'learning_ rate': 'invscaling', 'hidden_ layer_ sizes': 300, 'activation': 'logistic' Best Score: 0.6516093466748323

Best GridSearchCV - Best Hyperparameters: 'activation': 'relu', 'alpha': 0.0001, 'hidden_ layer_ sizes': 350, 'learning_ rate': 'adaptive', 'learning_ rate_ init': 0.01, 'solver': 'adam' Best Score: 0.6784304764249183

With the given result, the GridSearchCV output parameters were used.

The table 5, shows the Final Models Comparison F1-Score for Train and Validation using the final version of the 3 mentioned models after optimization.

After model optimization the final versions of the models were fitted in the split dataset, 80% training 20% validation ration, the prediction probabilities were calculated and used to draw a ROC Curve, figure 10.

In order to further optimize our results we decided to change the threshold for classification (default = 0.5). This is shown in figure 11. The best values were Threshold=0.092266 and F-Score=0.716.

## 6 Evaluation

Finally, with the model trained, and the threshold defined, the dataset ('predict.xlsx') was transformed and scaled to match all the variables existing in the training dataset. Then, the final model predicted the value of the target variable in the test dataset ('predict.xlsx').

The predictors used in the final submission were:

- 'Recency',
- 'MntVegan& Vegetarian'
- 'NumAppVisitsMonth'
- 'NumOfferPurchases'
- 'Marital_ Status'
- 'Response_ Campaigns'
- 'NumAppPurchases'
- 'Days_ As_ Client'

The predictions were then submitted to the kaggle competition resulting in a score of 0.38283.

# 7 Conclusion

Taking into account the obtained F1-score of 71, it is necessary to determine the probability of success of Spice Alley's campaign and the expected profitability. Assuming a cost per contact of €3 and an expected revenue of €16 it is possible to calculate the profitability in the following manner:

Conversion rate = F1-score / 100 Conversion rate = 71 / 100 Conversion rate = 0.71

Profitability = (Conversion rate * Revenue per successful sale) - Cost per contact

Profitability = (0.71 * 16€) - 3€

Profitability = 11.36€ - 3€

Profitability per customer = 8.36€

All in all, for a pool of 2500 clients there is an expected profit of €20.900 making this classification model valuable from a business perspective.

If we use the obtained F1-score obtained in Kaggle (38) the profitability per customer would be as such:

Conversion rate = F1-score / 100

Conversion rate = 38 / 100

Conversion rate = 0.38

Profitability = (Conversion rate * Revenue per successful sale) - Cost per contact

Profitability = (0.38 * 16€) - 3€

Profitability = 6.08€ - 3€

Profitability per customer = 3.08€

In this case, for a pool of 2500 clients there is an expected profit of €7.700 making this classification model also valuable from a business perspective.

# 8 Annexes

## 8.1 Tables

Table 1: Chi-Square filtering

| Predictor | Chi-Square | Remove |
|---|---|---|
| Education | 1 Yes 4 No | Yes |
| Marital_Status | 5 Yes | No |

Table 2: Summary of feature selection techniques

| Top 5/ 6 for Lasso | RFE | LASSO | DT | Remove |
|---|---|---|---|---|
| Recency | 5 | 5 | 5 | KEEP |
| MntVeganVegetarian | 5 | 5 | 5 | KEEP |
| NumOfferPurchases | 5 | 5 | 2 | KEEP |
| NumAppVisitsMonth | 5 | 5 | 5 | KEEP |
| Response_Campaingns | 5 | 5 | 5 | KEEP |
| NumAppPurchases | 0 | 5 (always top 6) | 0 | Try with and without |
| Days_As_Client | 0 | 0 | 3 | Try with and without |
| Marital_Status | | | | KEEP - chi2 |
| Education | | | | REMOVE - chi2 |
| Kid_Younger6 | | | | REMOVE |
| Children_6to18 | | | | REMOVE |
| MntEntries | | | | REMOVE |
| MntDrinks | | | | REMOVE |
| MntDesserts | | | | REMOVE |
| MntAdditionalRequests | | | | REMOVE |
| Response_Cmp1 | | | | REMOVE |
| Response_Cmp2 | | | | REMOVE |
| Response_Cmp3 | | | | REMOVE |
| Response_Cmp4 | | | | REMOVE |
| Response_Cmp5 | | | | REMOVE |
| Complain | | | | REMOVE |
| Gender | | | | REMOVE |
| Age | | | | REMOVE |
| Total Kids | | | | REMOVE |
| CostContact | | | | REMOVE - Constant |
| Revenue | | | | REMOVE - Constant |
| Income | | | | REMOVE - Redundant |
| MntMeat&Fish | | | | REMOVE - Redundant |
| NumTakeAwayPurchases | | | | REMOVE - Redundant |
| NumStorePurchases | | | | REMOVE - Redundant |

Table 3: all_data Results

| | Train | Validation |
|---|---|---|
| Logistic Regression | 0.622 +/- 0.01 | 0.609 +/- 0.03 |
| KNN | 0.681 +/- 0.02 | 0.542 +/- 0.08 |
| DT | 0.493 +/- 0.03 | 0.484 +/- 0.11 |
| NN | 0.725 +/- 0.01 | 0.694 +/- 0.02 |

Table 4: keep_data Results

| | Train | Validation |
|---|---|---|
| Logistic Regression | 0.585 +/- 0.01 | 0.579 +/- 0.06 |
| KNN | 0.732 +/- 0.01 | 0.595 +/- 0.05 |
| DT | 0.493 +/- 0.03 | 0.484 +/- 0.11 |
| NN | 0.67 +/- 0.02 | 0.646 +/- 0.04 |

Table 5: keep_data Results after optimization

| | Train | Validation |
|---|---|---|
| Best KNN | 0.789 +/- 0.01 | 0.627 +/- 0.05 |
| Best DT | 0.493 +/- 0.03 | 0.484 +/- 0.11 |
| Best NN | 0.753 +/- 0.03 | 0.646 +/- 0.04 |

## 8.2 Figures

Figure 1: NumAppVisitsMonth - Histplot

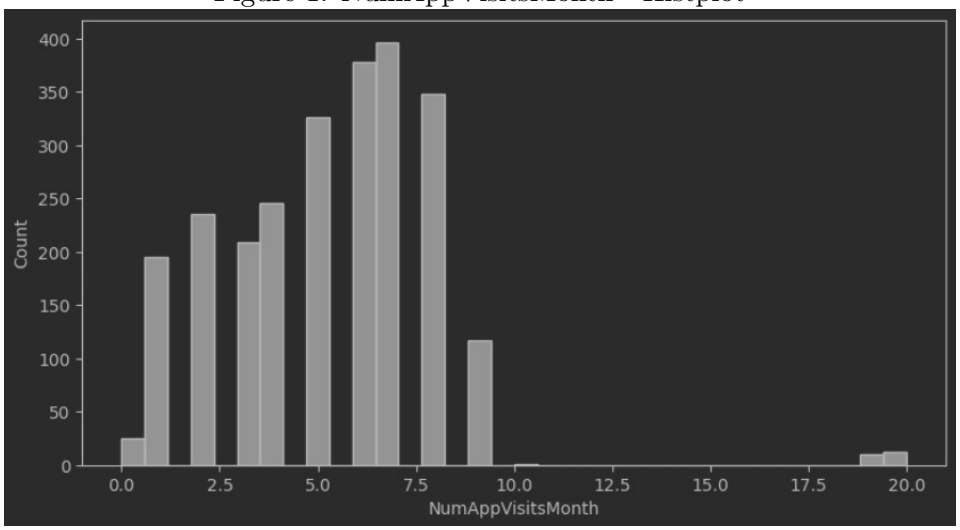


Figure 2: NumTakeAwayPurchases - Histplot

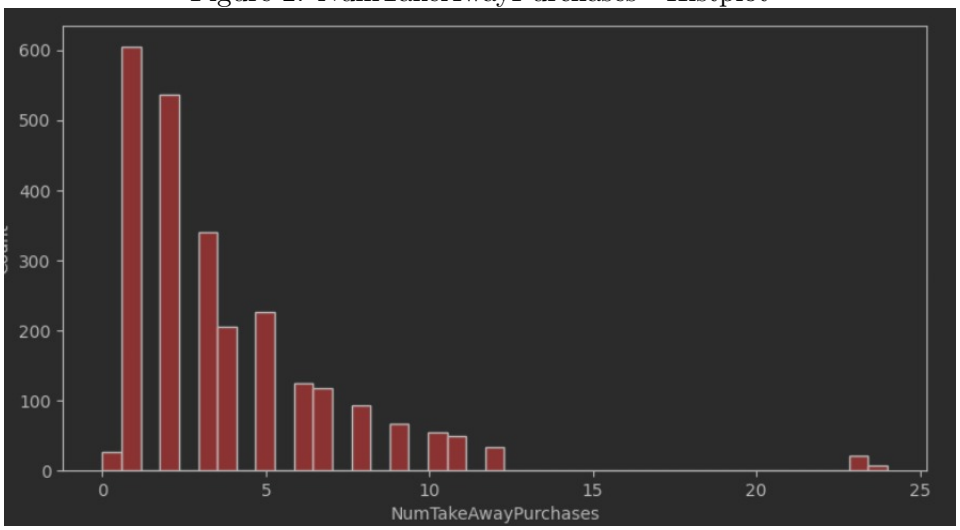

Figure 3: NumOfferPurchases - Histplot

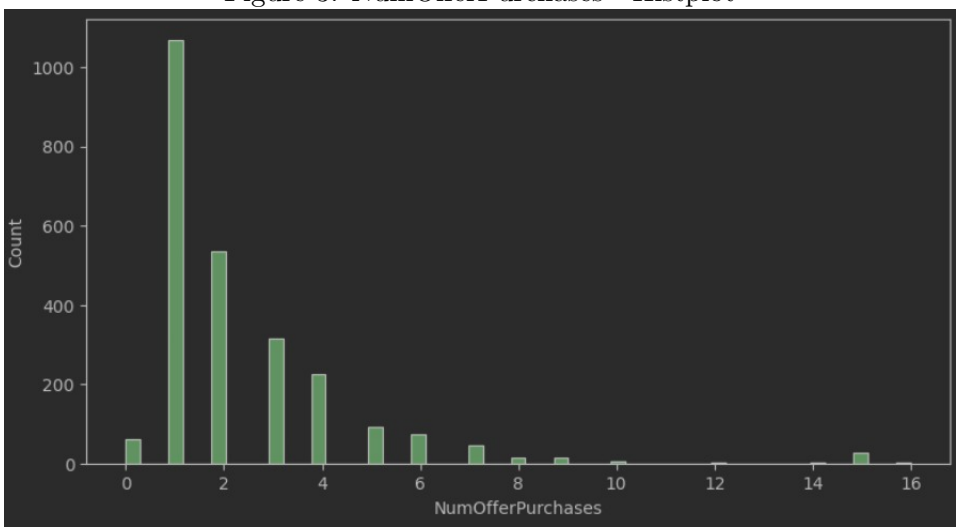


Figure 4: MntVegan and Vegetarian - Histplot

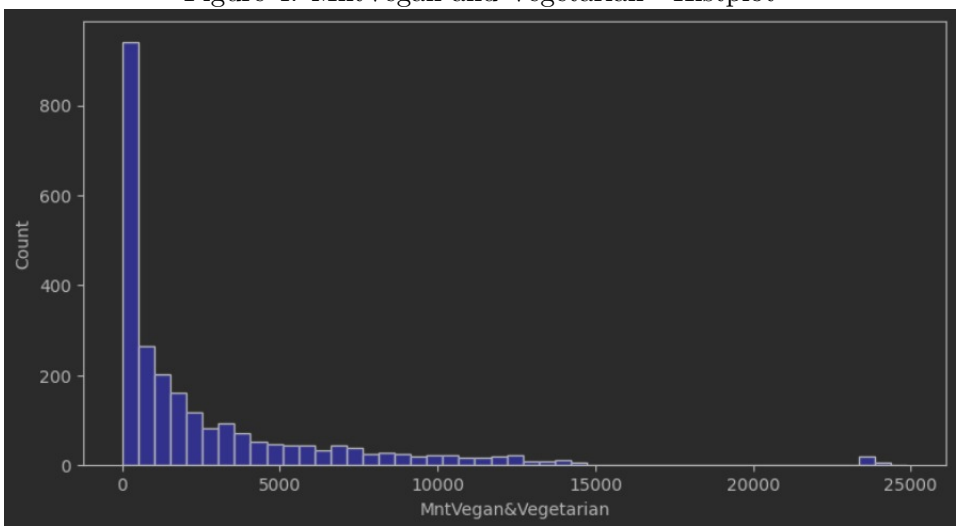


Figure 5: Income - Histplot

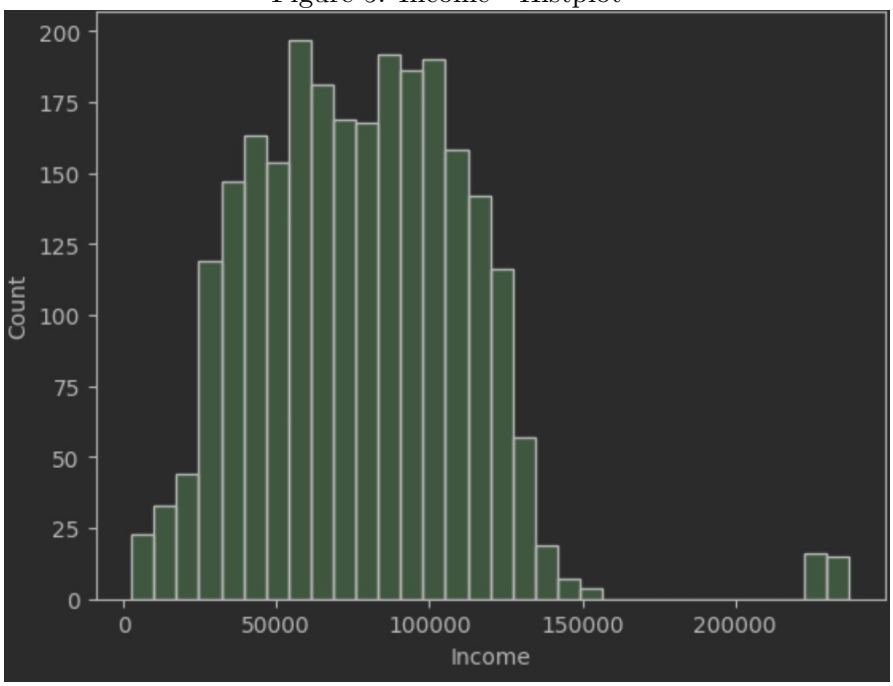

Figure 6: Feature Importance using DT Model

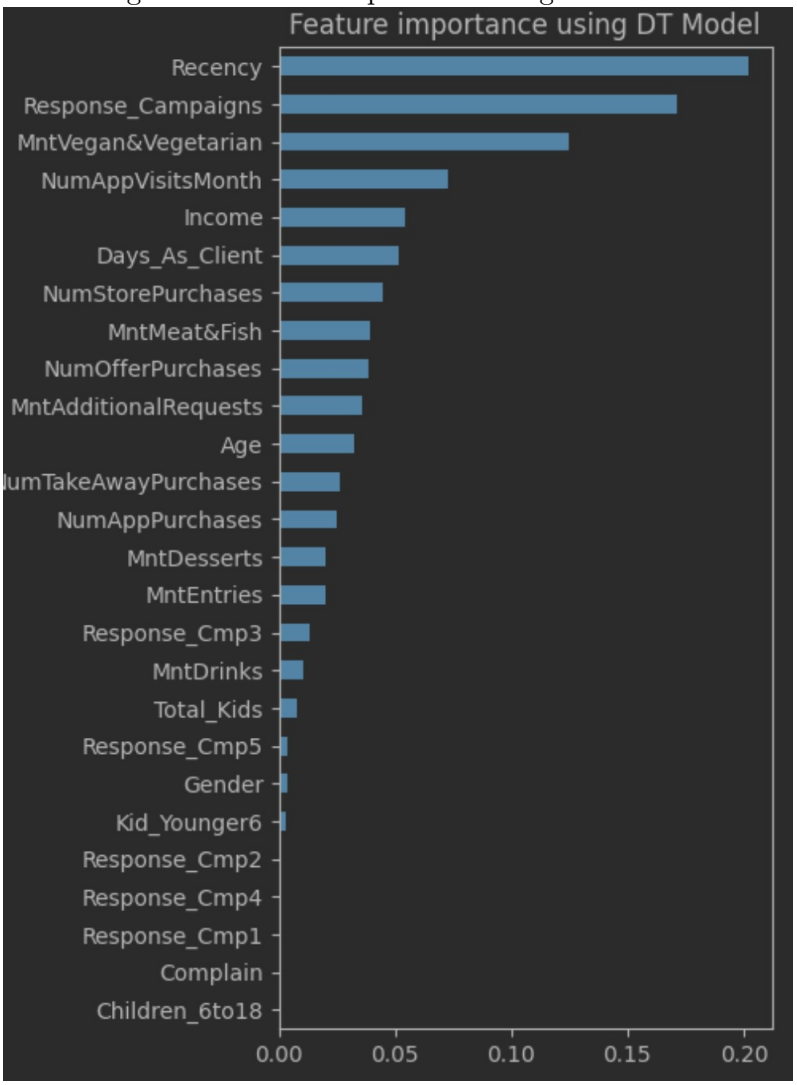


Figure 7: KNN manual optimization result

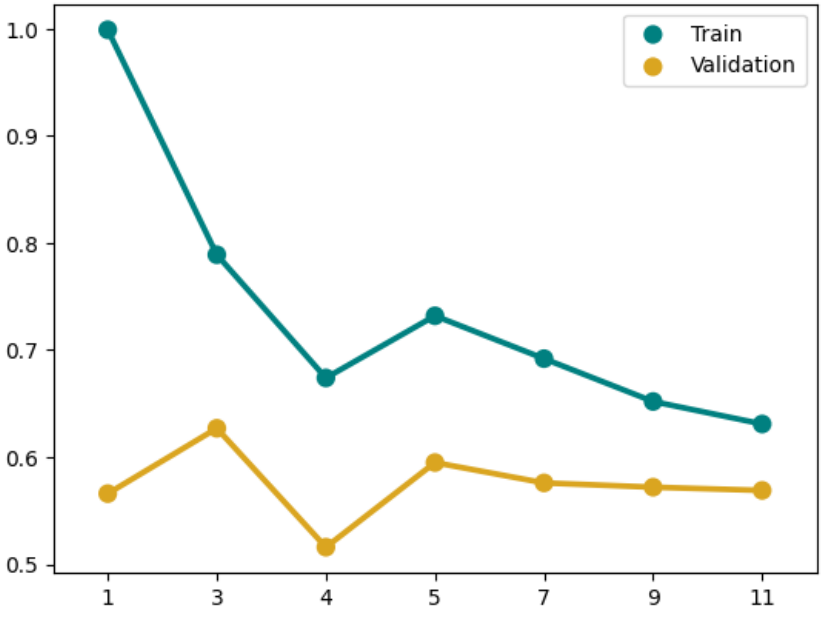

Figure 8: Decision Tree manual optimization result

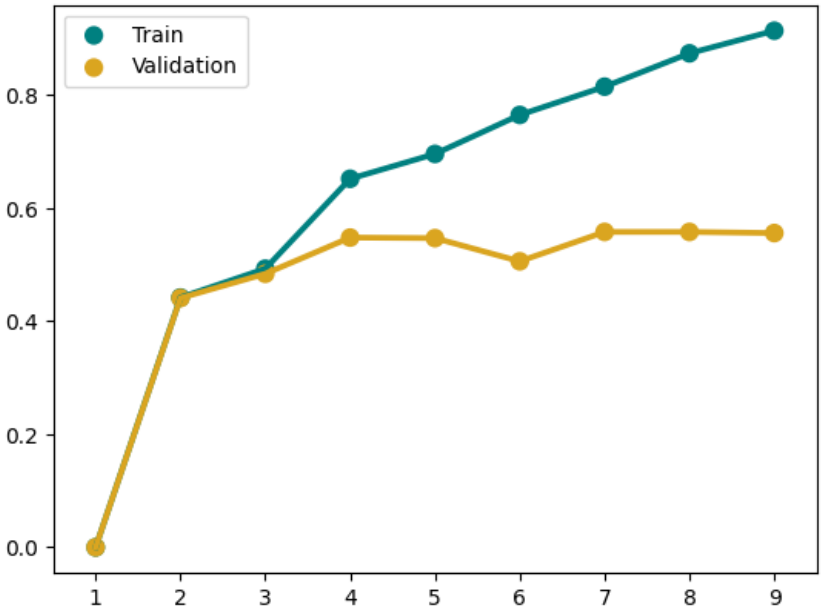


Figure 9: MLP Classifier manual optimization result

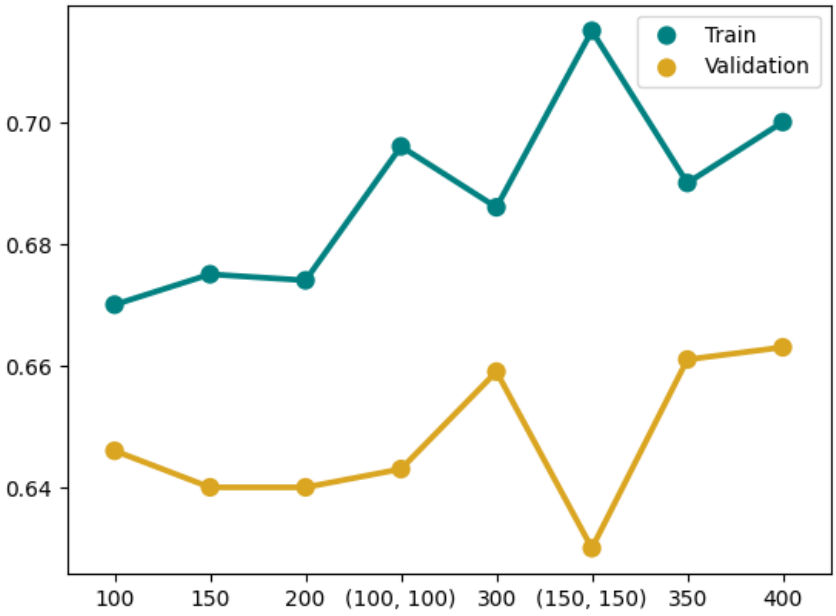

Figure 10: ROC Curve

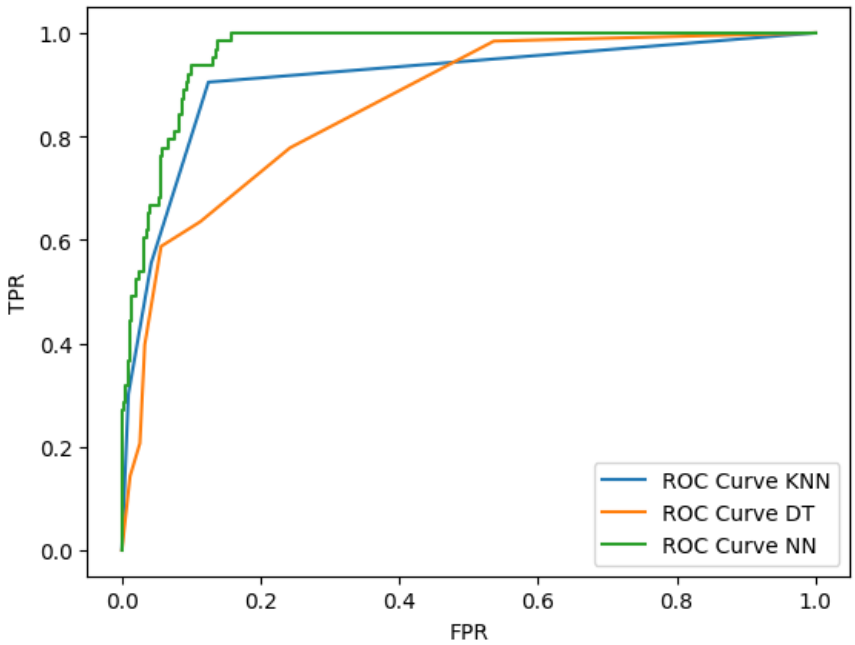


Figure 11: Precision Recall Curve Best Threshold=0.092266, F-Score=0.716