UNIVERSIDADE NOVA DE LISBOA

NOVA INFORMATION MANAGEMENT SCHOOL

Group Project: Spice Alley – Predictive

DATA SCIENCE AND MACHINE LEARNING

POST GRADUATION IN ENTERPRISE DATA SCIENCE AND ANALYTICS

PROFESSORS:

Carina Albuquerque

Ricardo Santos

GROUP DSML 202223 21

André Oliveira nº. 20222156

Diogo Fernandes nº. 20220507

Gonçalo Eloy nº. 20222162

Gonçalo Matos nº. 20221194

Rafael Chamusca nº. 20222174

2023

Index

1. Introduction & Methodology

This project aims to predict responders. Customers that will respond to Spice Alley's new product campaign and make a purchase, based on the analysis of the customers value, behavior, channel preference, and demographics. The team will employ supervised machine learning models, such as Neural Network, Support Vector Machines and Logistic Regression. These models will be trained using the historical data obtained from Spice Alley, followed by the utilization of new customer data to identify potential respondents. Predictive response models have been widely used in machine learning with the purpose of selecting those customers that will be most interested in a particular offer. In building such a model the goal is to enable Spicey Alley to strategically allocate its resources towards the target audience that has the highest probability of a positive response to the campaign, that is, to purchase the new product.

2. Business Understanding

Spice Alley is a restaurant that has recently gained notoriety and offers a range of options including meat, fish, and vegetarian, cooked by the restaurant's chefs. Due to the recent increase in the use of data science in business contexts, Spice Alley seeks to apply the potential that this resource offers, to obtain greater knowledge about the behaviour of its customers. The restaurant's objective is for our data scientist team to build a model that allows maximizing the profit generated in a future marketing campaign related to a new food segment. For this model, a sample of 2500 customers were contacted by email to inquire about purchasing the product. Those who purchased in the following three months were classified as 1 and those who did not respond as 0. With the data obtained in this process we will infer about the behavior of the remaining population.

3. Understanding & Exploring Data

In this project, data from two datasets were used, one of them with information on whether the customer joined the marketing campaign or not, this dataset will be used to build the forecast model. There is also another dataset that does not have the response variable, and with this data we will test the predictive capacity of the model that we created earlier. The variable description is presented in detail in **Table 1**.

4. Data Preparation

Before applying any model into our data, it is crucial to perform some manipulation in the datasets provided by Spice Ally, in the further subpoints we will describe the transformations applied.

4.1. Data Cleaning

To determine the outliers, we used the z-scores method and the Inter-quartile range method. The Z-score method is a statistical technique used to identify outliers in a dataset by measuring how many standard deviations a data point is away from the mean. We calculated the Z-scores for each data point by defining a threshold equal to 3. The IQR (Interquartile Range) method is another statistical technique commonly used to detect outliers in a dataset. It calculates the range between the first quartile (Q1) and the third quartile (Q3) and defines a threshold to identify values that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. Initially we applied these methods to non-standardized data and obtained a large number of outliers for both methods (26568 for IQR and 372 for Z-score). Then, we applied the same methods after using the Yeo-Johnson transformation method (see 4.3.) and the values obtained were much smaller (79 outliers for the Z-score method). Finally, we end up removing those outliers, 29 from Income, 28 from NumAppPurchases and 22 from NumAppVisitsMonth.

4.2. Missing Values (MV)

The variables Education, Recency and MntDrinks, have, respectively, 33, 48 and 21 missing values. We applied the mode to fill the values in “Education”, the median to fill the values in “Recency” and KNN to fill the values in “MntDrinks” (three highly correlated features were used) the detailed analysis of missing values is in **Table 2**.

4.3. Data Transformation and Feature Engineering

In this step we applied some feature engineering that we think that can give us different perspectives of approaching our data and give us meaningful insights. The detailed description is present in **Table 3**. Since there are variables that are quite unbalanced in terms of distribution, we tried to find ways to mitigate this situation. So, we decided to apply the Yeo-Johnson transformation. It is an extension of the Box-Cox transformation and allows for transformations of data that may contain negative values. To scale our data, we utilized 3 standardization techniques: MinMaxScaler, Standard and Robust Scaling. MinMaxScaler rescales the data set so that all feature values are within the range [0, 1] or [-1, 1] if there are negative values, it also preserves the original distribution of the data and is less affected by outliers when compared to StandardScaler. StandardScaler transforms each feature in the data set to have a mean of 0 and a standard deviation of 1. Resulting in a standard normal distribution, it allows for fair comparisons between different features, and it is useful for algorithms that assume data to be normally distributed, such as linear or logistic regression. RobustScaler scales the features using statistics that are robust to outliers, specifically by subtracting the median and dividing by the interquartile range (IQR), it is less affected by outliers than MinMax or Standard Scaling and preserves the shape of the original distribution while standardizing features to a similar scale.

5. Feature Selection

During the feature selection process, we focused on identifying the most pertinent features that align with the business problem and are influential for the machine learning models employed.

To evaluate the independence of categorical variables, we utilized the Chi-Square method. To ensure consistent representation of variables across diverse data subsets, we employed Stratified k-fold.

Categorical variables were analyzed for their independence with the Chi-Square method, and to ensure consistent variable representation across the different data susbets a Stratified k-fold was applied.

Through this approach we determined that Marital Status is the only categorical variable to be retained in the final set.

Regarding numerical features, the analysis began with variance assessment, to understand if our numerical data and ordinal data is constant or quasi-constant. Zero variance means no information which in turn means the variable is not important. Then, with the help of a Spearman's correlation heatmap that showed the correlation among features, we identified 2 groups of highly correlated features:

- **Group 1** Income, MntMeat&fish, MntVegan&Vegetarian, NumTakeAwayPurchases, NumStorePurchases, Mnt_Total,
- **Group 2** Freq., MntMeat&fish, MntVegan&Vegetarian, NumAppPurchases, NumTakeAwayPurchases, NumStorePurchases, Mnt_Total

Following this analysis, and while employing Stratified K-fold cross-validation, we utilized the Decision Tree Classifier, Recursive Feature Elimination (RFE), and Lasso techniques to assess the significance of the independent variables and determine their feature importance (**Figure 19-233 and Table 4**. Validating all the findings and above information we identified the most influential variables, that we should retain, the variables to remove and other potential variables to consider, summarized in **Table 5** and **Table 6.** After Experimentation and model performance assessment we came to the conclusion that the variables to be applied are the the following: Marital_Status; Income; Recency; MntVegan&Vegetarian; NumOfferPurchases; NumAppPurchases; NumAppVisitsMonth; Campaign_ordinal.

Other methods were used to assess which variables were most relevant, including PCA and Random Forest, however the results were unsatisfactory and thus dropped.

6. Model Assessment

At this stage we tried to analyze which were the most promising models to help us predict which customers will buy the product that Spice Alley will launch. For this, we tested several models with adjustments in the hyperparameters of each one, to extract the best performance from them. We used "GridSearch" to test the best possible combinations of hyperparameters together with Stratified-K-fold to ensure that we had a proportional representation in the validation and test sets. We used F1-score metric to evaluate the performance of the model. This metric combines precision and recall in a single metric, Very low values in

one of them will result in a low score. It is particularly useful when both false positives and false negatives are important.

6.1. Comparing models with default hyper parameters

Initially we tested our datasets with Logistic Regression, KNeighbors, DecisionTree, Neural Networks and Support Vector Machines. All of them with the default parameters and with the application of Stratified K Fold with 5 splits. The results obtained can be seen in **Table 7**. Analyzing the obtained results, we conclude that the most promising models are **Logistic Regression, Neural Networks and Support Vector Machines**.

6.2. Comparing multiple combinations of different hyperparameters

We focused on the top performing models, Logistic Regression, Neural Networks and Support Vector Machines, and aimed to enhance the results through exploration of the multiple combinations of their different hyperparameters. We have gained a better understanding on hyperparameter techniques that can be used to achieve better model performance and reduce overfitting. Some that are used and that were initially considered were reducing model complexity through layer number reducing, reducing the number of iterations to mitigate potential overfitting that may occur, and using smaller learning rates and variables. Nonetheless, considering the number of features and parameters available the team applied GridSearch for Logistic Regression and SVM and also RandomSearch to NN MLP. Our main goal was to increase the performance of our models given the scores obtained in the Kaggle competition and also to reduce overfitting which was a major issue particularly in our MLP model.

MLPClassifier – Neural Network

To improve the model's performance and address overfitting concerns, we explored parameter adjustments as part of our optimization process.

An analysis of the Neural Network Model Performance for Different Hidden Layer Sizes was also conducted to best identify the number o layers, and we can see that increasing the number of layers of neurons to 150 will lead to overfitting. Here the best model was achieved with the following hyperparameters:

- activation = 'relu – the active function Rectified Linear Unit, a simple and effective actication function
- hidden_layer_sizes = (90) Which represent the number and size of hidden layers in the neural network
- learning_rate = 'adaptive', which is sets a decreasing learning rate, and helps finding a more optimal set of weights
- learning_rate_init = 0.005 - which sets the initial learning rate used, a higher learning rate may result in faster convergence but may also lead to less stability

- solver = 'sgd' - Stochastic Gradient Descent, which is the weight optimization algorithm used and popularly used in neural networks

This process also included calculation and plotting of the Precision-Recall Curve where the best value ide ntified was 0.413526 ***Figure* 266**

Logistic Regression

For this model we mainly wanted to increase performance, and the changes made on the parameters were the following:

- solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'], which specifies the algorithm used fot eh optimization, in this case the best was the Large Linear Classification. This is aligns with the dataset which involved a large number of features, since this algorithm is commonly used in datasets with high dimensionality.
- penalty: ['none', 'l1', 'l2', 'elasticnet'], Also controls overfiting and applies the penalty to mitigate it. In this case it was chosen L1, Lasso).
- C: [100, 10, 1.0, 0.1, 0.01] : which controls the regularization strength inverse, meaning smaller values result in stronger results. The model reduces the impact of individual features to mitigate overfitting. Here the value was given by GridSearch was 1, which aligns with our results considering we didn't have overfitting problems.

SVM

For this model we mainly wanted to increase performance, and the changes made on the parameters were the following:

- kernel': ['linear', 'poly', 'rbf', 'sigmoid']: The kernel is chosen based on the data distribution, and in this case GridSearch identifies that Radial Basis Function is the best parameter. This is a very versatile function that can be used in different data distributions considering its flexibility
- 'C': [100, 10, 1.0, 0.1, 0.001]: Similarly, to the C paramater in Logistic Regression it allows for regularization strength control, and misclassication penalty. In this case it was set as 10, a moderate strength which considering the absence of overfitting issues in our model, this choice aligns with the overall decision-making process.

Despite our extensive efforts to optimize the models by exploring multiple combinations of hyperparameters and conducting numerous iterations, we encountered challenges in achieving significant performance improvements. Although we tested numerous parameter configurations and excluded those exhibiting overfitting, and bad results, the overall performance did not increase as expected. It became apparent that the primary factor influencing performance enhancement would be the feature selection process conducted

prior to model training. We believe that refining the feature selection methodology would have a more significant impact on improving model performance.

7. Results Analysis

Upon analyzing the final models, we obtained the results summarized in **Table 7**. Among the models evaluated Neural Networks emerged as the best performing, with an F-Score of 0.826 for the training set and 0.708 for the validation set. Although there is a noticeable difference of 0.118 between the two sets, indicating some level of overfitting, this model obtained the best score in our Kaggle performance, hence validating its selection as the final model. Our analysis also included the utilization of a ROC curve which assesses the performance of classification models by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold values. In **Figure 255** we were able to identify that the Neural Network demonstrated the best predictive capabilities.

8. Conclusions

The goal of this project was to predict which customers will buy the new product to be launched by Spice Alley through the application of various forecasting models. To achieve this goal, it was necessary to analyze different combinations of features with more relevance to the models and apply them to these features, adjusting the parameters to obtain a better performance from the model. After carrying out these tasks, we concluded that the model with the best performance was the Neural Network, having presented an F1 Score of 0.36. In our point of view the results were not satisfactory, however there are several reasons for this result. The fact that the variable we intend to predict is very unbalanced (315 buyers vs. 2185 non-buyers) causes the models' predictive capacity to be greatly affected. We believe that applying other feature selection techniques and improving the hyperparameters of the explored models could have led to better performance.

9. Annex I – Tables & Figures

9.1. Tables

Table 1. Variable Description, Data Types and Distribution Observations

Data Types	Variables	Distribution Observations
Categorical	Name, Date_Adherece, Marital_Status	
Categorical	Education	
Categorical	Kid_younger6, Children_to18, Complain	
Numerical	Income	Positive skewness of 0.827, indicating a right tail and A kurtosis of 2.550. 31 observations have a value greater than 200000, those are considered outliers. It also has a standard deviation of 35505.417 and a mean of 77557.227.
Numerical	Recency	Positive skewness of 0.028, and A kurtosis of -1.150. Does not show outliers or extreme values. It also has a standard deviation of 28.636 and a mean of 48.983.
Numerical	MntMeat&Fish	Positive skewness of 1.135, and A kurtosis of 0.477. Does not show outliers or extreme values. It also has a standard deviation of 3376.433 and a mean of 3071.254.
Numerical	MntEntries	Positive skewness of 2.091, indicating a right tail and A kurtosis of 4.292 suggests a moderate peak and tails. Does not show outliers or extreme values. It also has a standard deviation of 761.351 and a mean of 526.582.
Numerical	MntVegan&Vegetarian	Positive skewness of 2.479, indicating a right tail and A kurtosis of 8.099 suggests an extreme peak and tails. Some extreme values (greater than 20000). It also has a standard deviation of 3875.425 and a mean of 2748.278.
Numerical	MntDrinks	Positive skewness of 2.016, indicating a right tail and A kurtosis of 3.742 suggests a moderate peak and tails. Does not show outliers or extreme values. It also has a standard deviation of 793.028 and a mean of 545.916.
Numerical	MntDesserts	Positive skewness of 2.071, indicating a right tail and A kurtosis of 4.084 suggests a moderate peak and tails. Does not show outliers or extreme values. It also has a standard deviation of 763.868 and a mean of 524.163.
Numerical	MntAdditionalRequests	Positive skewness of 1.849, indicating a right tail and A kurtosis of 3.246 suggests a moderate peak and tails. Does

		not show outliers or extreme values. It also has a standard deviation of 49.576 and a mean of 42.555.
Numerical Discrete	NumOfferPurchases	Positive skewness of 2.875, indicating a right tail and A kurtosis of 11.293 suggests an extreme peak and tails. Does show some extreme values since the mean is 2.454 and max value is 16. It also has a standard deviation of 2.300.
Numerical Discrete	NumAppPurchases	Shows a skewness of 0.535, indicating a normal distribution and a kurtosis of -0.215 suggests a moderate peak and tails. Does not show outliers or extreme values. It also has a standard deviation of 2.757 and a mean of 5.996.
Numerical Discrete	NumTakeAwayPurchases	Shows a skewness of 2.390, indicating a right tail and a kurtosis of 9.146 suggests an extreme peak and tails. Some extreme values (greater than 20). It also has a standard deviation of 3.425 and a mean of 3.852.
Numerical Discrete	NumStorePurchases	Shows a skewness of 0.611, indicating a normal distribution and A kurtosis of -0.724 Does not show outliers or extreme values. It also has a standard deviation of 3.339 and a mean of 5.828.
Numerical Discrete	NumAppVisitsMonth	Skewness of 0.970, indicating a distribution a bit skewed to the right and A kurtosis of 4.989 suggests a moderate peak and tails. Some extreme values (22 greater than 15). It also has a standard deviation of 2.712 and a mean of 5.292.
Numerical Discrete	Responses (1,2,3,4,5)	Low variability.
Numerical Discrete	CostContact	Is a constant feature, so does not give relevant insights.
Numerical Discrete	DepVar	Target variable, highly unbalanced

Table 2. Missing Values and Applied Processes

Variable	Applied Processes
Education	33 missing completely at random values (MCAR) since those seem be independent from the rest of “Education’s” values and other values.We decided to fill them with mode because it’s a categorical variable and the number of mv is low compared to the total number of observations.
Recency	48 missing completely at random values (MCAR) since those seem be independent from the rest of “Recency’s” values and other values.

	We decided to fill them with media because it's a numerical variable and the number of mv is low compared to the total number of observations.The Knn option was discarded since this variable does not correlate with any of other variables.
Mntdrinks	21 missing completely at random values (MCAR) since those seem be independent from the rest of "MntDrinks's" values.We decided to fill them with Knn imputer with "MntDrinks", "MntEntries", "MntVegan&Vegetarian", "MntDesserts" since those have shown a strong correlation (0.7).

Table 3. Feature Engineering - New Variables

Variable	Applied Transformations
Gender	All values start with Ms. or Mr. prefix were used to create a gender Male or Female.
Antiquity	Date_Adherence had 16 invalid observations (2/29/2022), we replace them to '3/01/2022' and transformed this variable into another called Antiquity that has the date of adherence converted into the number of days
Age	Age variable created by transforming birthyear
Freq	Variable created by summing the NumAppPurchases, NumTakeAwayPurchases, NumStorePurchases, NumOfferPurchases.
Mnt_Total	Variable created by summing the MntMeat&Fish,MntEntries, MntVegan&Vegetarian, MntDrinks , MntDesserts, MntAdditionalRequests
Education_bins_2	Grouping Education Values with the following binning: 'Phd': 'High' | 'Master': 'High' | 'Graduation': 'Medium' | 'Basic' : 'Low' |'Highschool' : 'Medium
Education_bins_3	Grouping Education Values with the following binning: 'Phd': 'High' | 'Master': 'High' | 'Graduation': 'Low' | 'Basic' : 'Low' |'Highschool' : 'Low
Campaign_ordinal	Aggegrated sum of the Customer response to pervious Campaigns.
Marital_Status_2	Grouping Marital_Status values with the following binning: 'Married' : 'Together' | 'Single' : 'Single' | 'Divorced' : 'Single' | 'Widow' : 'Single
Marital_Status_4	Grouping Marital_Status values with the following binning: 'Married' : 'Together' | 'Single' : 'Single' | 'Divorced' : 'Divorced'| 'Widow' : Widow'
Have_kids	Variable created by summing Kid_Younger6 and Children_6to18 counts

Table 4. RFE Feature Selection for the Numerical Variables

PREDICTOR	RFE Split 1	RFE Split 2	RFE Split 3	RFE Split 4	RFE Split 5
Income	False	False	False	False	False
Recency	True	True	True	True	True
Mntmeat&Fish	False	False	False	False	False
MntEntries	False	False	False	False	False
Mntvegan&Vegetarian	True	True	True	True	True
MntDrinks	False	False	False	False	False
MntDesserts	False	False	False	False	False
MntAdditionalRequests	False	False	False	False	False
NumOfferPurchases	False	False	False	False	False
NumAppPurchases	False	False	False	False	False
NumTakeawayPurchases	True	True	True	True	True
NumstorePurchases	False	False	False	False	True
NumAppvisitsMonth	True	True	True	True	False
Antiquity	False	False	False	False	False
Age	False	False	False	False	False
Freq	False	False	False	False	False
Mnt_Total	True	True	True	True	True

Table 5. Categorical Variables Feature Selection Analysis

Predictor	Chi-Square	Decision
Education	0 Yes / 5 No	Remove
Education_Bins_2	0 Yes / 5 No	Remove
Education_Bins_3	5 Yes / 0 No	Keep
Kid_Younger6	0 Yes / 5 No	Remove
Children_6to18	5 Yes / 0 No	Consider Keep/Switch
Have_Kids	5 Yes / 0 No	Keep
Response_Cmp1	5 Yes / 0 No	Consider Keep/Switch
Response_Cmp2	5 Yes / 0 No	Consider Keep/Switch
Response_Cmp3	5 Yes / 0 No	Consider Keep/Switch
Response_Cmp4	5 Yes / 0 No	Consider Keep/Switch
Response_Cmp5	5 Yes / 0 No	Consider Keep/Switch
Campaign_Ordinal	5 Yes / 0 No	Consider Keep/Switch
Complain	0 Yes / 5 No	Remove
Gender	0 Yes / 5 No	Remove
Marital_Status	5 Yes / 0 No	Consider Keep/Switch
Marital_Status_2	5 Yes / 0 No	Consider Keep/Switch
Marital_Status_4	5 Yes / 0 No	Consider Keep/Switch

Table 6. Numerical Variables Feature Selection Analysis

Predictor	RFE	Lasso	DT	Decision
Recency	5 Yes	5 Yes	5 Yes	Keep
Income	0 Yes	5 Yes	5 Yes	Consider
Mntmeat&Fish	0 Yes	5 Yes	5 Yes	Consider
MntEntries	5 Yes	0 Yes	0 Yes	Consider
MntVegan&Vegetarian	5 Yes	5 Yes	5 Yes	Keep
MntDrinks	0 Yes	0 Yes	0 Yes	Remove
MntDesserts	0 Yes	0 Yes	0 Yes	Remove
MntAdditionalRequests	0 Yes	0 Yes	0 Yes	Remove
NumOfferPurchases	0 Yes	0 Yes	5 Yes	Consider
NumAppPurchases	0 Yes	5 Yes	0 Yes	Consider
NumTakeAwayPurchases	5 Yes	4 Yes	1 Yes	Consider
NumStorePurchases	1 Yes	4 Yes	1 Yes	Consider
NumAppvisitsMonth	4 Yes	5 Yes	5 Yes	Keep
Antiquity	0 Yes	0 Yes	2 Yes	Remove
Age	0 Yes	0 Yes	2 Yes	Remove
Freq.	0 Yes	4 Yes	0 Yes	Consider
Mnt_Total	5 Yes	0 Yes	1 Yes	Consider

Table 7. Model Performance Comparison: F1 Scores on Training and Validation Sets (A – Initial Model Run, B – Final Model Run

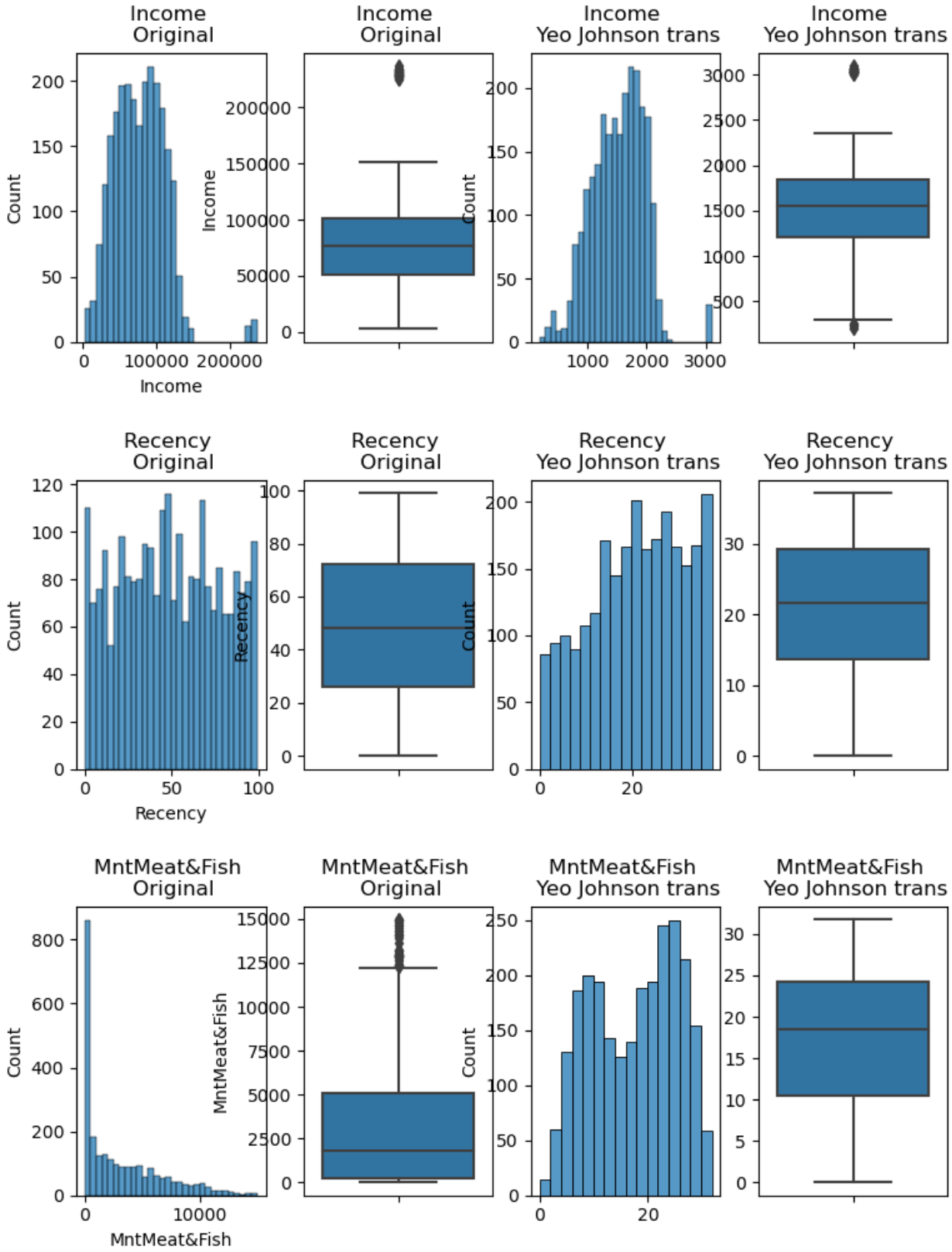
A

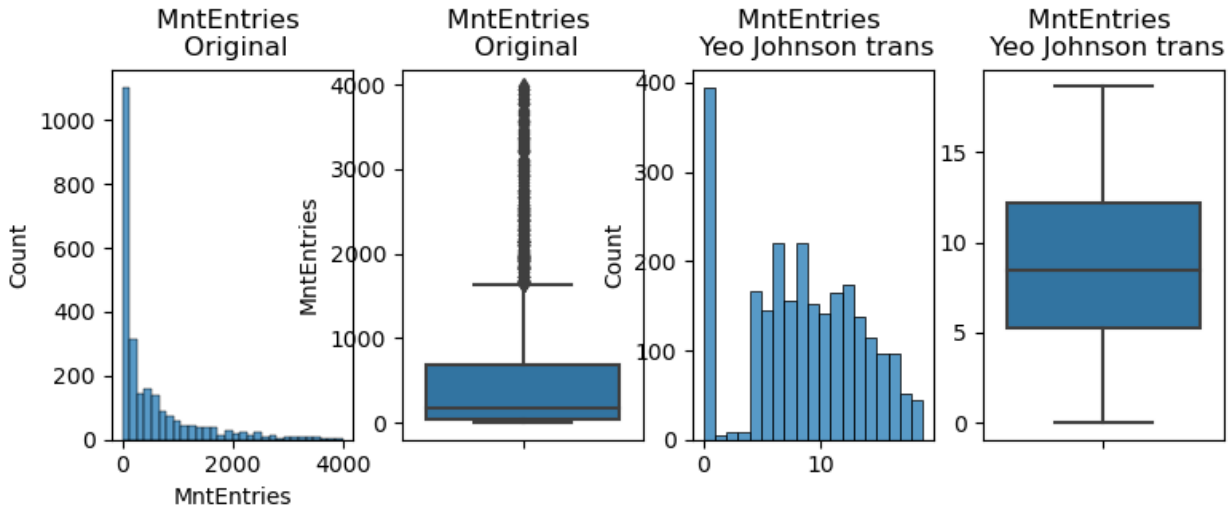
	Train	Validation	Diff
Logistic Regression	0.666+/-0.01	0.656+/-0.03	0.01
KNN	0.722+/-0.02	0.588+/-0.09	0.134
DT	0.532+/-0.05	0.505+/-0.06	0.027
NN	0.826+/-0.02	0.708+/-0.04	0.118
SVC	0.765+/-0.01	0.713+/-0.05	0.052

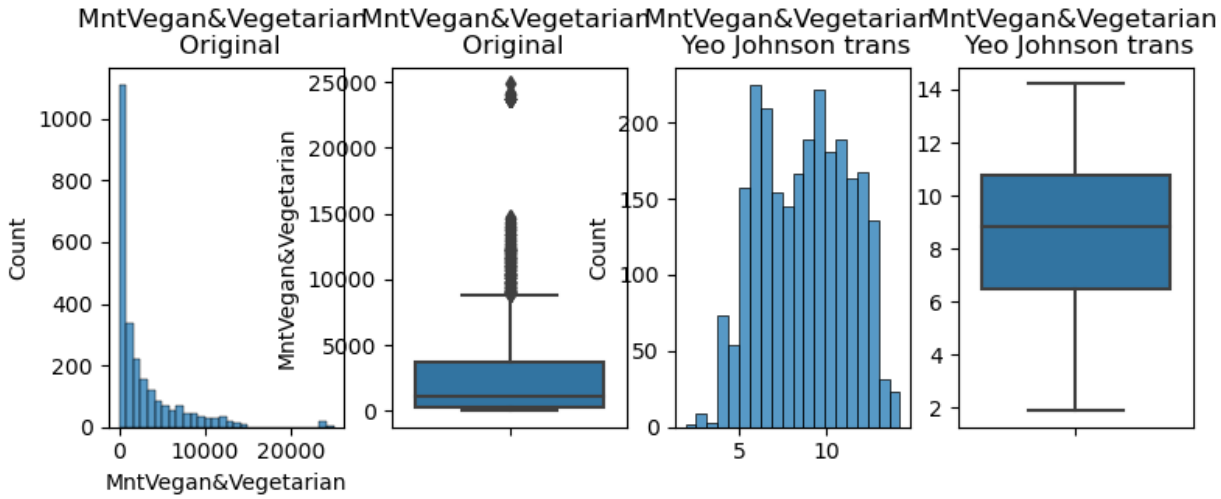
B

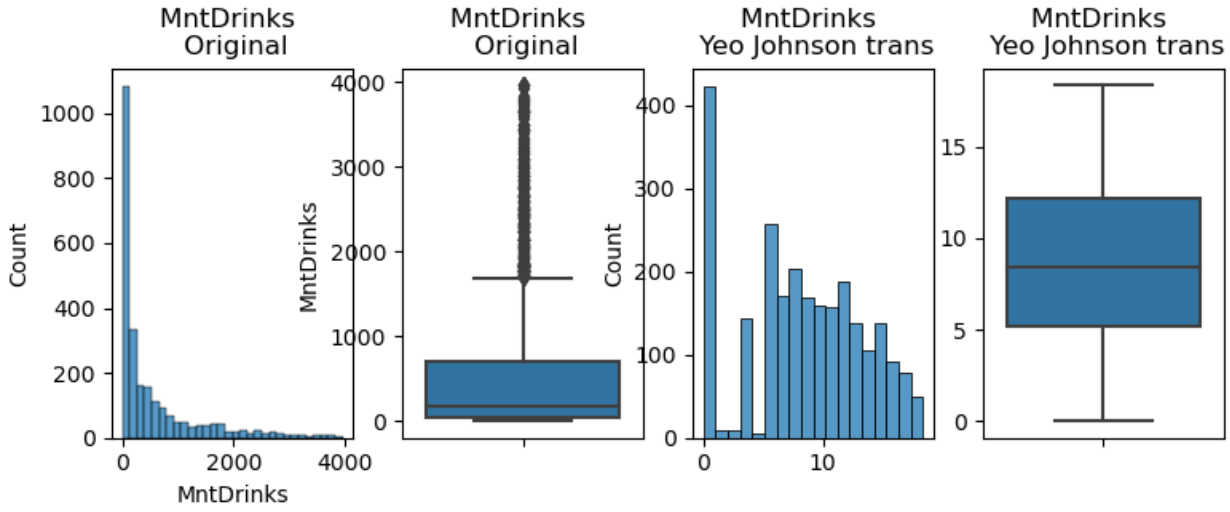
	Train	Validation	Diff
Logistic Regression	0.665+/-0.01	0.651+/-0.03	0.014
NN	0.78+/-0.01	0.736+/-0.03	0.044
SVC	0.843+/-0.01	0.704+/-0.04	0.139

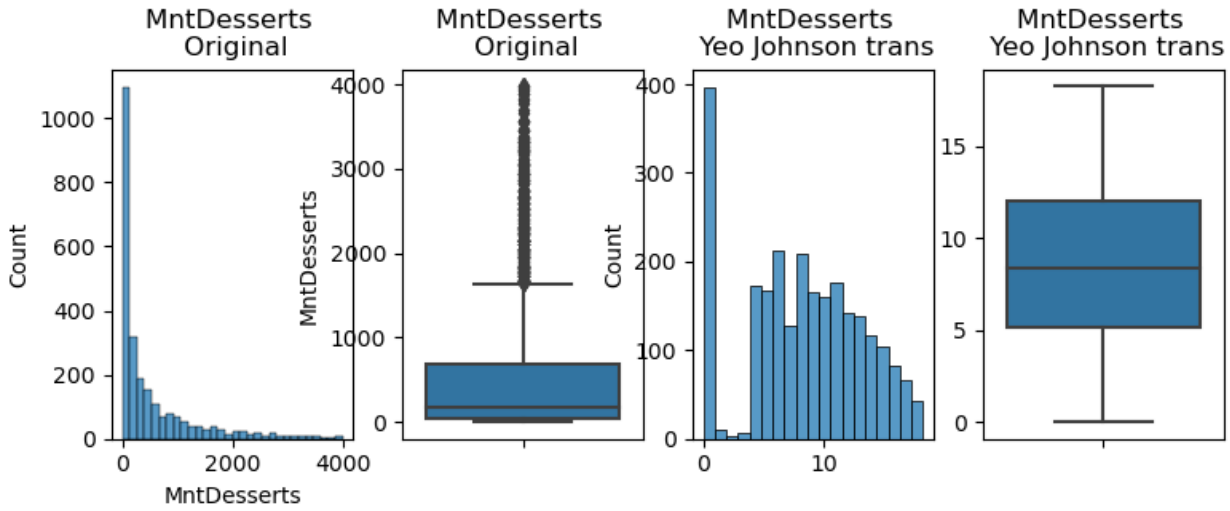
9.2. Figures

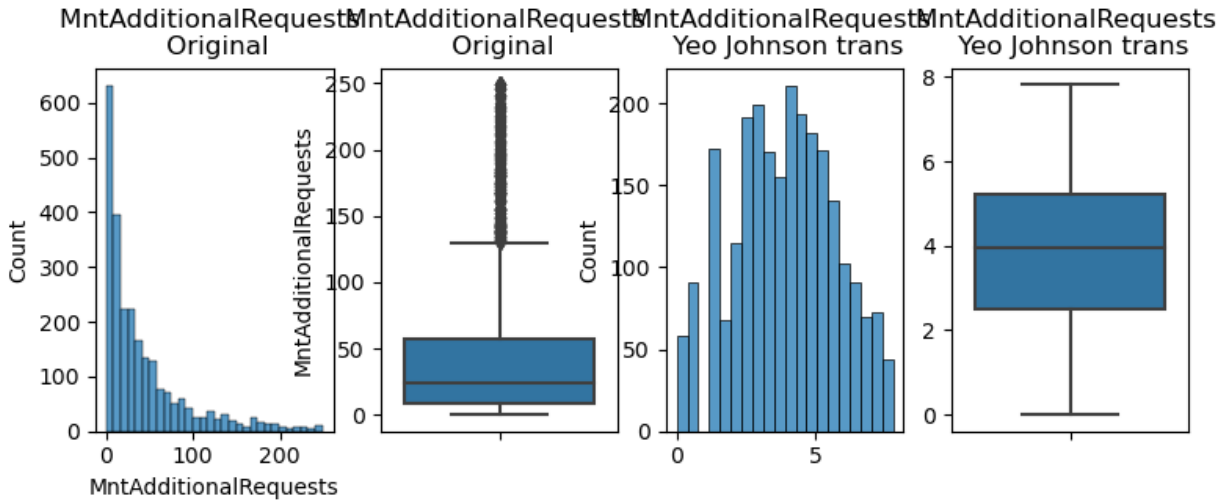

Income
Original
Count
Income
Income
Original
Income
Income
Yeo Johnson trans
Income
Yeo Johnson trans
Recency
Original
Count
Recency
Recency
Original
Count
Recency
Yeo Johnson trans
Recency
Yeo Johnson trans
MntMeat&Fish
Original
Count
MntMeat&Fish
MntMeat&Fish
Original
Count
MntMeat&Fish
Yeo Johnson trans
MntMeat&Fish
Yeo Johnson trans

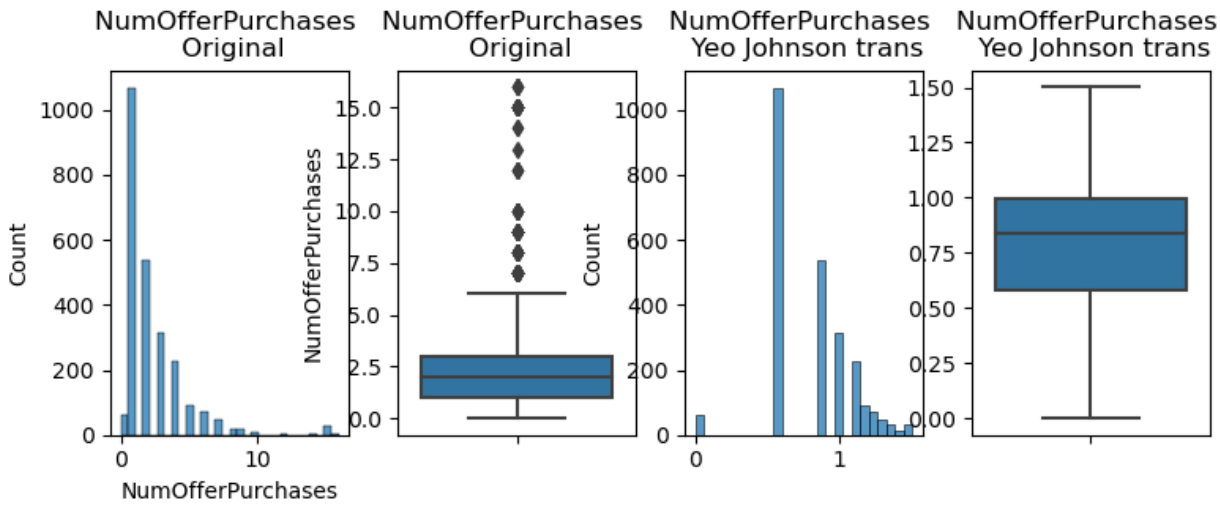

MntEntries
Original
MntEntries
Original
MntEntries
Yeo Johnson trans
MntEntries
Yeo Johnson trans
Count
MntEntries
Count

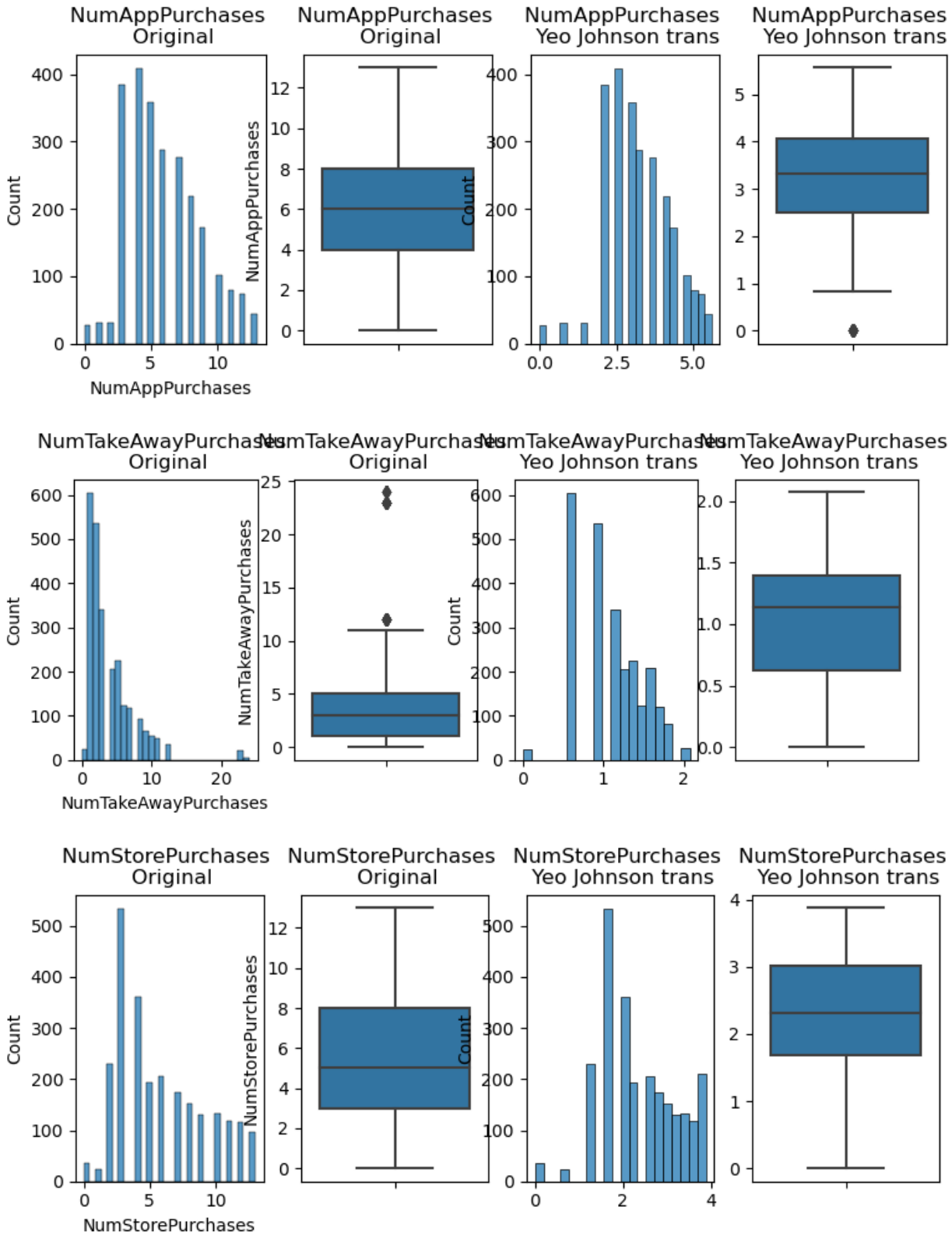

MntVegan&Vegetarian
Original
MntVegan&Vegetarian
Original
MntVegan&Vegetarian
Yeo Johnson trans
MntVegan&Vegetarian
Yeo Johnson trans
Count
MntVegan&Vegetarian
Count

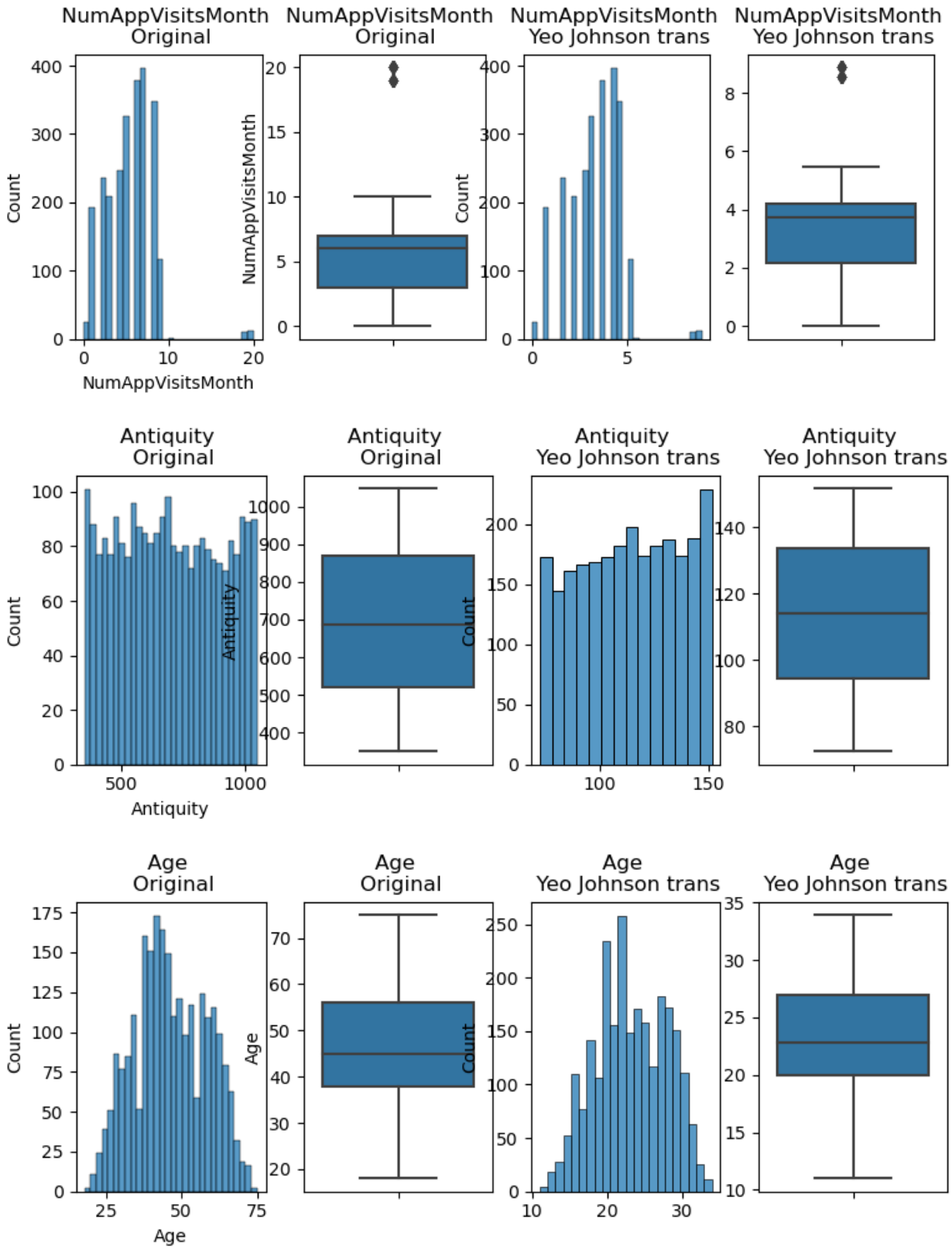

MntDrinks
Original
MntDrinks
Original
MntDrinks
Yeo Johnson trans
MntDrinks
Yeo Johnson trans
Count
MntDrinks
Count


MntDesserts
Original
MntDesserts
Original
MntDesserts
Yeo Johnson trans
MntDesserts
Yeo Johnson trans
Count
MntDesserts
Count
MntDesserts


MntAdditionalRequests
Original
MntAdditionalRequests
Original
MntAdditionalRequests
Yeo Johnson trans
MntAdditionalRequests
Yeo Johnson trans
Count
MntAdditionalRequests
Count
MntAdditionalRequests


NumOfferPurchases
Original
NumOfferPurchases
Original
NumOfferPurchases
Yeo Johnson trans
NumOfferPurchases
Yeo Johnson trans
Count
NumOfferPurchases
Count
NumOfferPurchases


NumAppPurchases
Original
NumAppPurchases
Original
NumAppPurchases
Yeo Johnson trans
NumAppPurchases
Yeo Johnson trans
Count
NumAppPurchases
NumTakeAwayPurchases
Original
NumTakeAwayPurchases
Original
NumTakeAwayPurchases
Yeo Johnson trans
NumTakeAwayPurchases
Yeo Johnson trans
Count
NumTakeAwayPurchases
NumStorePurchases
Original
NumStorePurchases
Original
NumStorePurchases
Yeo Johnson trans
NumStorePurchases
Yeo Johnson trans
Count
NumStorePurchases


NumAppVisitsMonth
Original
NumAppVisitsMonth
Original
NumAppVisitsMonth
Yeo Johnson trans
NumAppVisitsMonth
Yeo Johnson trans
Count
NumAppVisitsMonth
Antiquity
Original
Antiquity
Original
Antiquity
Yeo Johnson trans
Antiquity
Yeo Johnson trans
Count
Antiquity
Age
Original
Age
Original
Age
Yeo Johnson trans
Age
Yeo Johnson trans
Count
Age

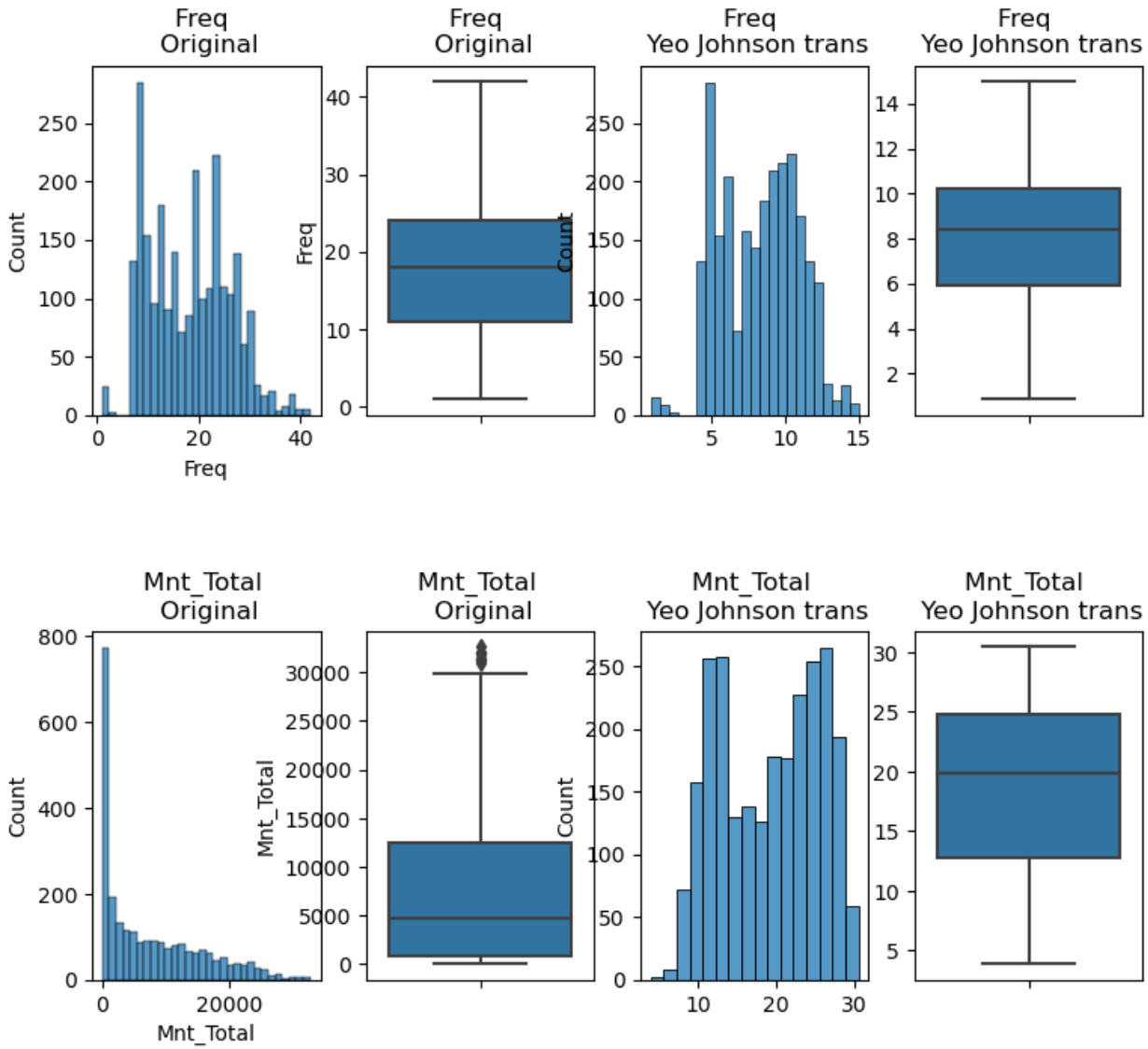


Figure 1-17 Histogram and boxplots of new and transformed variables.

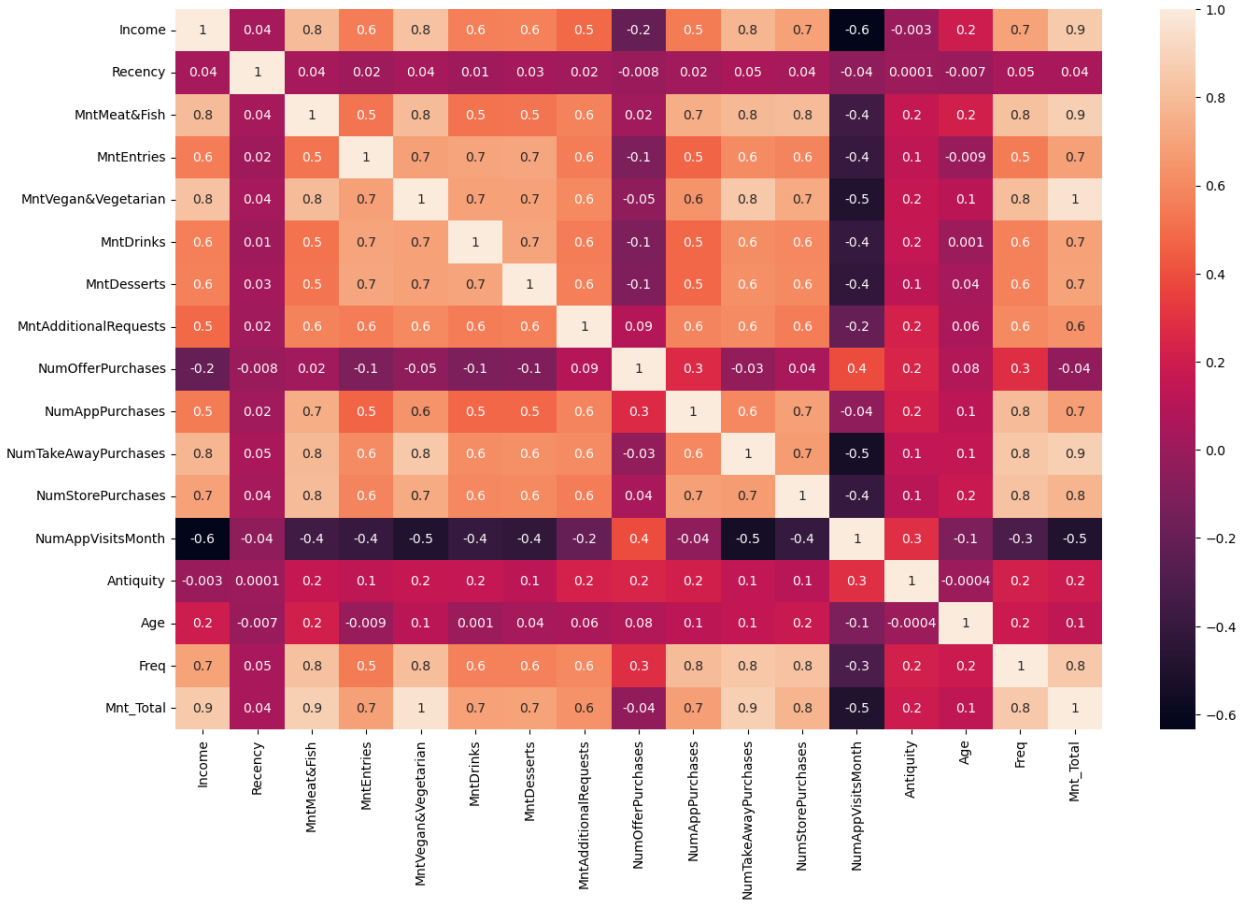
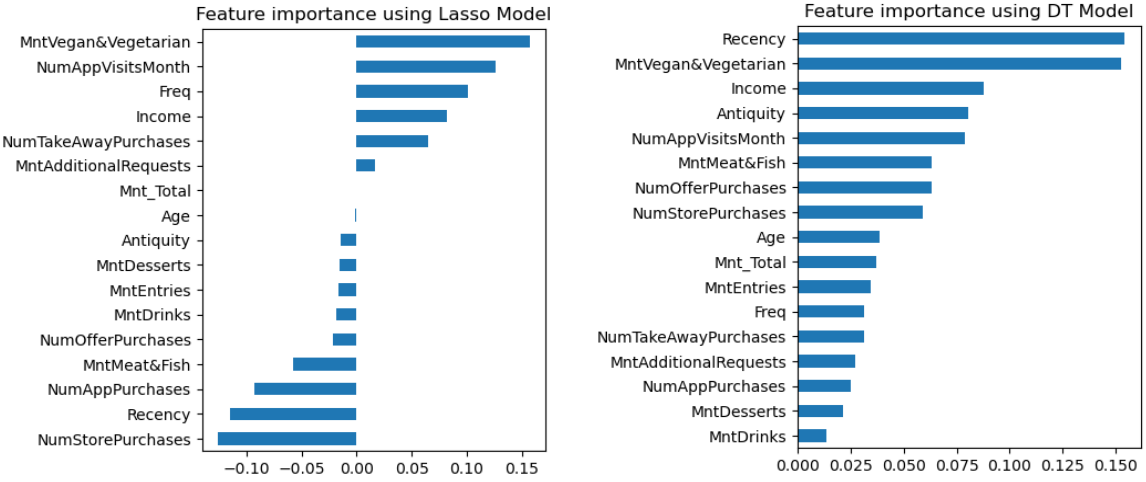
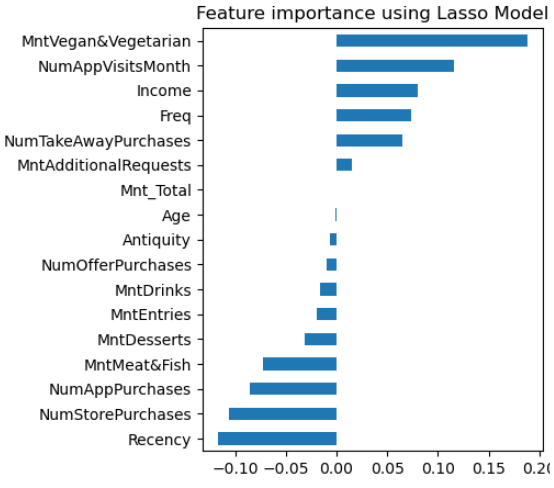
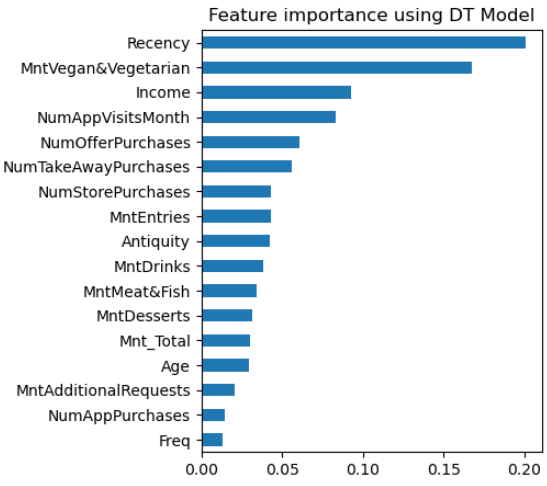
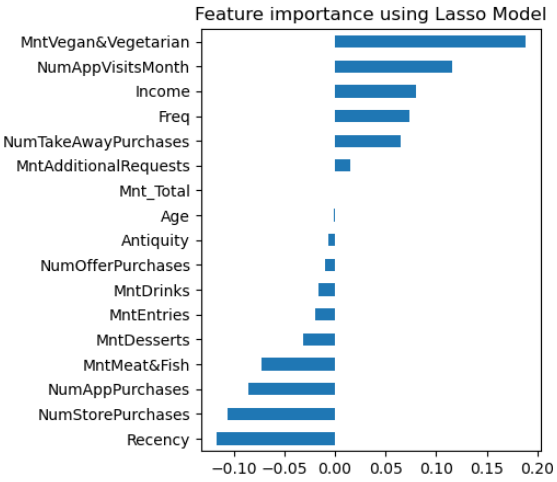


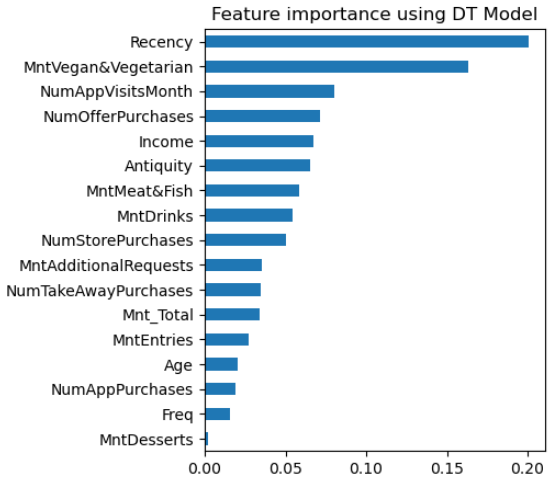
Figure 18. Spearman Correlation Heatmap.

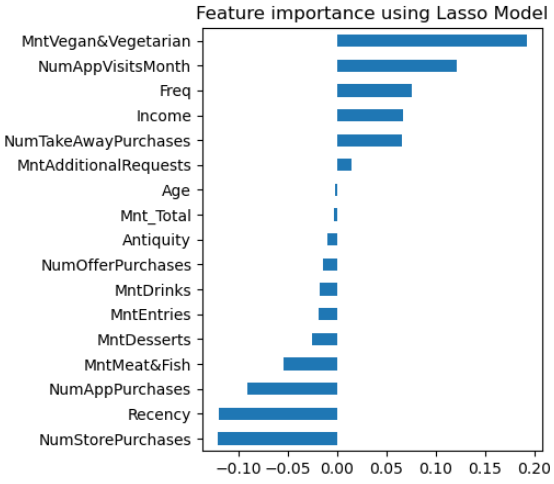


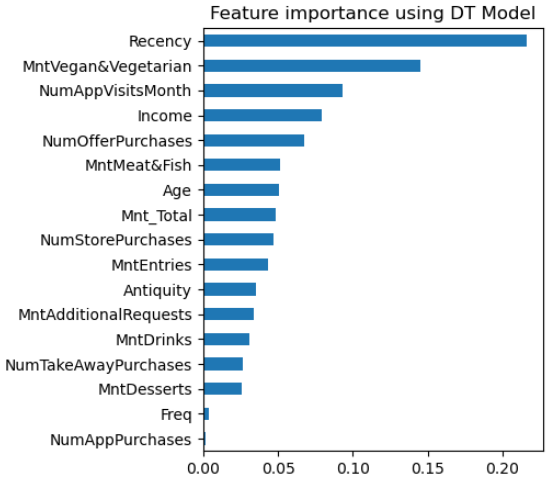

Feature importance using Lasso Model
MntVegan&Vegetarian
NumAppVisitsMonth
Income
Freq
NumTakeAwayPurchases
MntAdditionalRequests
Mnt_Total
Age
Antiquity
NumOfferPurchases
MntDrinks
MntEntries
MntDesserts
MntMeat&Fish
NumAppPurchases
NumStorePurchases
Recency
−0.10 −0.05 0.00 0.05 0.10 0.15 0.20


Feature importance using DT Model
Recency
MntVegan&Vegetarian
Income
NumAppVisitsMonth
NumOfferPurchases
NumTakeAwayPurchases
NumStorePurchases
MntEntries
Antiquity
MntDrinks
MntMeat&Fish
MntDesserts
Mnt_Total
Age
MntAdditionalRequests
NumAppPurchases
Freq
0.00 0.05 0.10 0.15 0.20


Feature importance using Lasso Model
MntVegan&Vegetarian
NumAppVisitsMonth
Income
Freq
NumTakeAwayPurchases
MntAdditionalRequests
Mnt_Total
Age
Antiquity
NumOfferPurchases
MntDrinks
MntEntries
MntDesserts
MntMeat&Fish
NumAppPurchases
NumStorePurchases
Recency
−0.10 −0.05 0.00 0.05 0.10 0.15 0.20


Feature importance using DT Model
Recency
MntVegan&Vegetarian
NumAppVisitsMonth
NumOfferPurchases
Income
Antiquity
MntMeat&Fish
MntDrinks
NumStorePurchases
MntAdditionalRequests
NumTakeAwayPurchases
Mnt_Total
MntEntries
Age
NumAppPurchases
Freq
MntDesserts
0.00 0.05 0.10 0.15 0.20


Feature importance using Lasso Model
MntVegan&Vegetarian
NumAppVisitsMonth
Freq
Income
NumTakeAwayPurchases
MntAdditionalRequests
Age
Mnt_Total
Antiquity
NumOfferPurchases
MntDrinks
MntEntries
MntDesserts
MntMeat&Fish
NumAppPurchases
Recency
NumStorePurchases
−0.10 −0.05 0.00 0.05 0.10 0.15 0.20


Feature importance using DT Model
Recency
MntVegan&Vegetarian
NumAppVisitsMonth
Income
NumOfferPurchases
MntMeat&Fish
Age
Mnt_Total
NumStorePurchases
MntEntries
Antiquity
MntAdditionalRequests
MntDrinks
NumTakeAwayPurchases
MntDesserts
Freq
NumAppPurchases
0.00 0.05 0.10 0.15 0.20

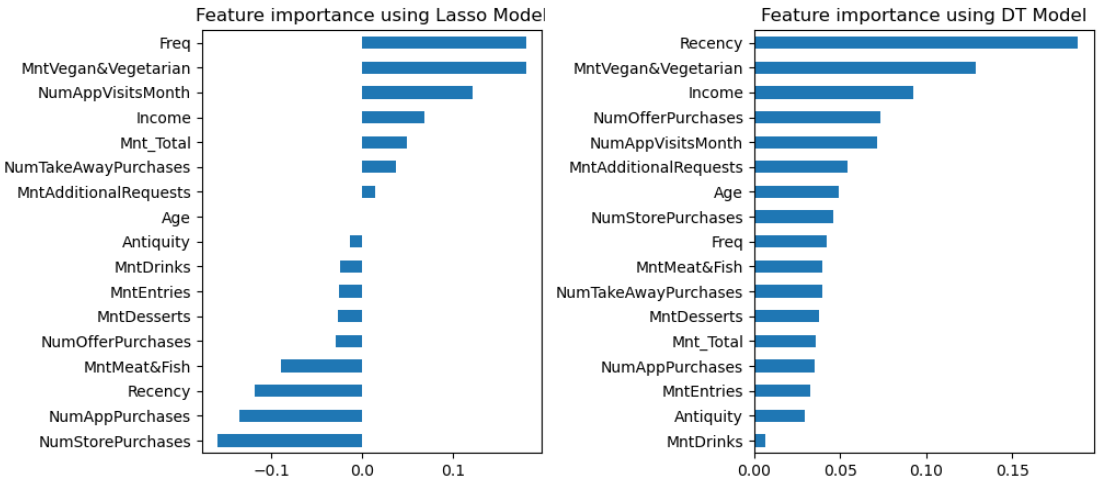


Figure 19-23. *DT and Lasso feature importance (Splits 1-5)*

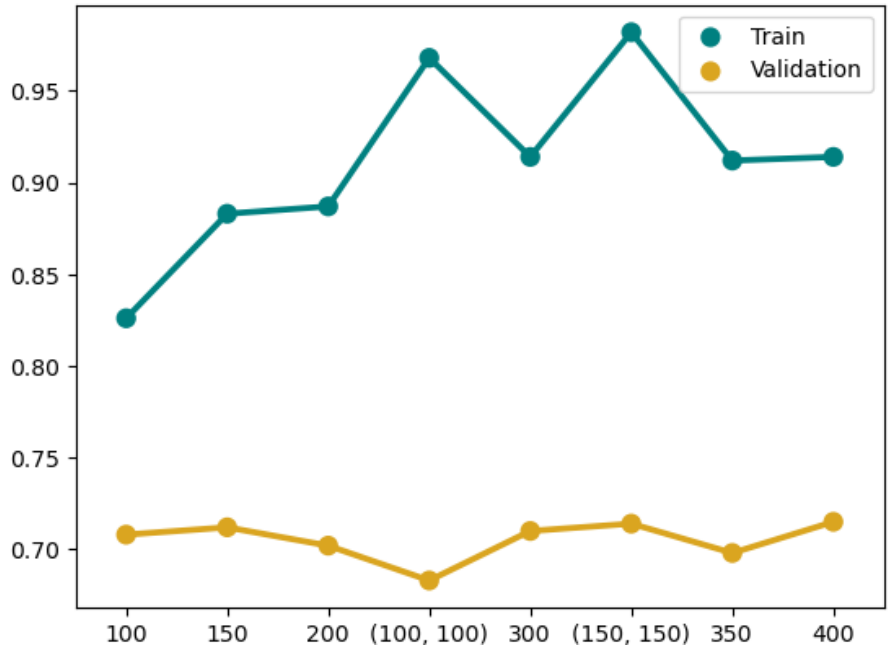


Figure 24. Neural Network Model Performance for Different Hidden Layer Sizes

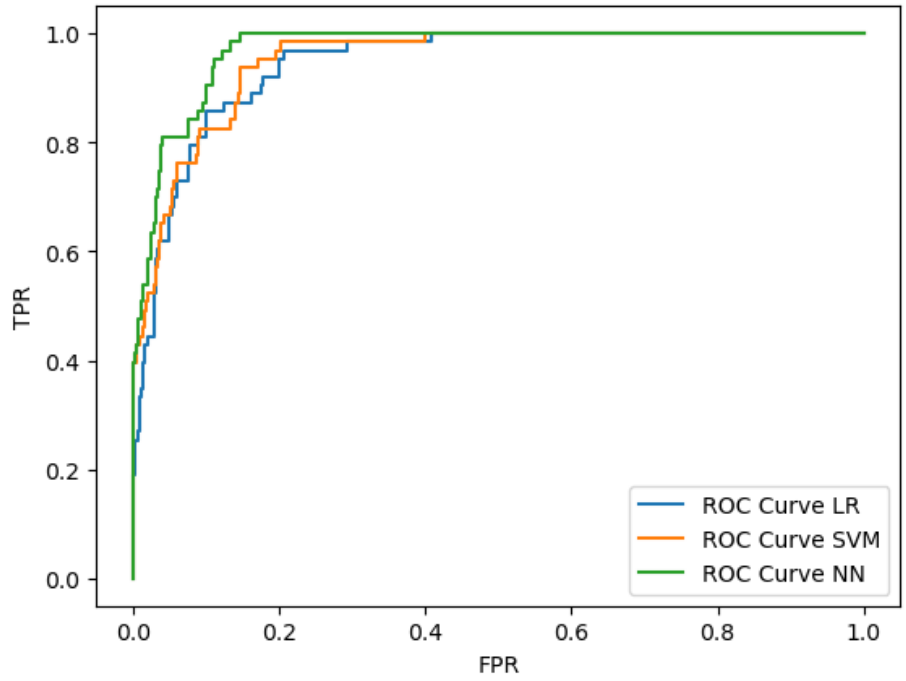


Figure 25. Receiver Operating Characteristic (ROC) Curves for Logistic Regression, SVM and NN models

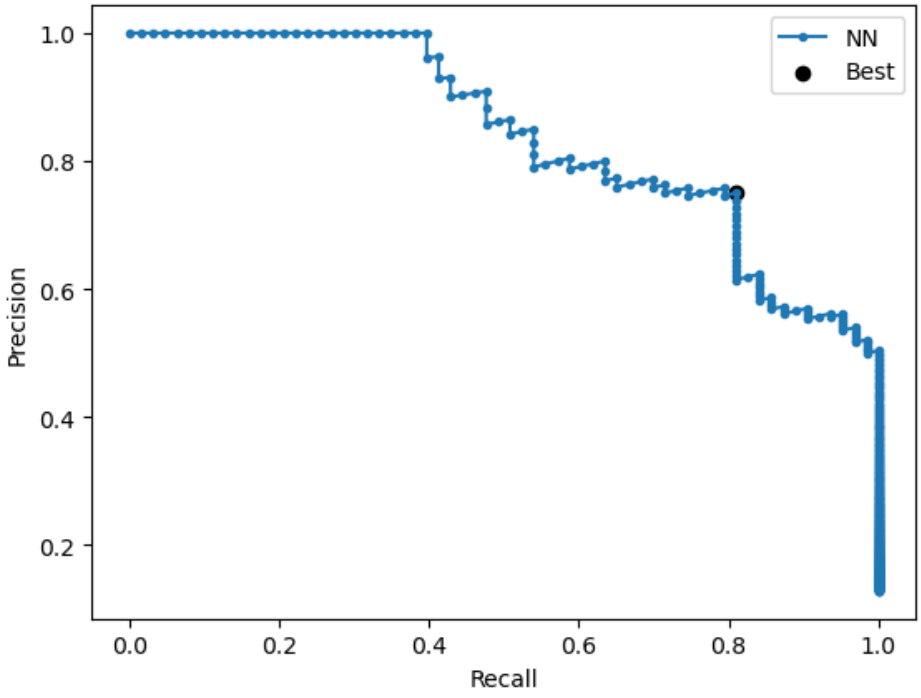


Figure 26. Precision-Recall Curve for Neural Network Model.

10. Annex II – Theory

10.1. Z-score

Z-score is a concept in statistics that helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, the Z score tells how many standard deviations away an observation is from the mean. It's important to note that the z-score method assumes that the data follows a normal distribution. If the data is not normally distributed or contains outliers, alternative normalization methods may be more appropriate. Assuming a normal distribution, we know that 68% of the observations lie between +/- 1 standard deviation, 95% between +/- 2 standard deviation and 99.7% between +/- 3 standard deviation. If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier. For our problem, we considered the threshold of 3 to be considered as an outlier.

10.2. Yeo-Johnson method

The Yeo-Johnson transformation is a method used for data normalization or for stabilizing variance in statistical analysis. It is an extension of the Box-Cox transformation that allows for both positive and negative values in the data. This is one of the main advantages of the model when compared with Box-Cox.

10.3. Support Vector Machines (SVM)

The goal of SVM is to find a hyperplane in an N-dimensional space(N - nº of features) that distinctly classifies the datapoints. The hyperplane should have the maximum margin (the maximum distance between datapoints of both classes), because maximizing the margin distance provides reinforcement so that future datapoints can be classified with more confidence. When there is only 2 features the hyperplane is just a line. With 3 features is a two-dimensional plane. From 4 features on is a hyperplane.

The hyperplane is a decision boundary that helps to classify the datapoints. The support vectors are datapoints closer to the hyperplane that influence his position and orientation. These datapoints help to build the SVM, and deleting them will change the hyperplane position. So, if the output of the hinge loss

function is > 1 it belongs to one class. If the output is > -1 it belongs to another class. This threshold range of values (1, -1) will act as margin.

The hinge loss function helps the model to maximize the margin between the datapoints and the hyperplane. If the actual value and the predicted value are of the same class, the cost is 0. If they're not, the function calculates the loss value and add a regularization parameter that will balance the margin maximization and loss. The function uses partial derivatives respecting the weights to find the gradients, that help updating the weights. When the model correctly predicts the class of the datapoint, the gradient from the regularization parameter is updated. When the model makes a wrong prediction, it includes the loss along with the regularization parameter to perform the gradient update.

In summary, the main ideas of the SVM are:

1. Start with data in relatively low dimension,
2. Move data into higher dimension and
3. Find a support vector classifier that supports the higher dimension data into 2 groups.
4. The kernel function systematically finds support vector classifiers in higher dimensions. For instance, when the polynomial kernel is used and d = 1, it compares the relationship between each pair of observations in 1-dimension. If d = 2, it compares the relationship between each pair of observations in 2-dimensions.

11. References

Hoaglin, D.C. (2013). Volume 16: How to Detect and Handle Outliers.

Javaheri, S. H. (2008). Response Modeling in Direct Marketing. In *Elsevier eBooks* (pp. 153–180). https://doi.org/10.1016/b978-0-12-411511-8.00006-2

Z-Score
Z score for Outlier Detection - Python - GeeksforGeeks

Yeo-Johnson method:
Johnson Transformation In Python (Full Code) » EML (enjoymachinelearning.com)