DATA 240: Final Report
Customer Personality Analysis Using Data Mining

Team 11
Abinaya Seshadre, Jyoti Patel, Nirvisha Garara, Payal Padmanabhan, and Vraj Bharatkumar Mistry

1. Introduction

Given the boom in online shopping because of the pandemic, an immense amount of data has been churned through the observation of online buying patterns. Today we are generating so much data on our buying patterns whether it is online shopping on multiple ecommerce websites and apps or in-stores. These data that are generated are high dimensional datasets with complex connections suited for personality-based marketing. Customer Personality Analysis and Segmentation is a vital aspect that aids in decision making for customer-driven businesses. It is a comprehensive analysis and classification of the businesses' consummate customers. This would assist the organization with adapting their products and services to cater to their targeted demography based on personality traits apart from facets such as age, income, gender and buying history.

Examples

a. Personality analysis can be used to improve the need of usage and attractiveness of a product for example -any service, or a message which would result in increasing the usage, higher satisfaction, loyalty, and acceptance level. To say so, the Personality analysis of an online advertisement campaign could help in leading to more revenue and click–through rates (Boerman et al., 2017). The basic fact for the using the customer personality analysis in recommender systems is that every personality traits are known to be closely attached with individual differences in behavior (e.g. Harari et al., 2019; Jackson et al., 2010; Stachl et al., 2019) and preferences (Nave et al., 2018; Randler et al., 2017; Youyou et al., 2015). Hence, this user personality analysis is an intuitive way in increasing a system's attractiveness. Most importantly, personality–based targeting has been proved which has increased the effectiveness of marketing campaigns, increasing the sales for personality–congruent advertisements (Matz et al., 2017).

b. Another example can be seen as many ML approaches have been applied to customer personality data which helps in personalization of products and services through recommender systems. Personalization means by the help of information related to the users of a system it is useful to adapt the functionalities or characteristics related to the product or service helping in achieving a certain goal (e.g. Tkalcic et al., 2016, product recommendations on Amazon to facilitate purchase decisions). These adaptations are based on either the similarity of the user or objects to other users and objects (e.g. suggesting products based on similar products or based on purchases of users who also bought that product) or on predictive models (Aggarwal, 2016). The main motivation behind personality analysis is to reduce the amount of information which is associated with the user and it provides stimuli more suitable to the user's individual needs and interests (e.g. automatically rank movies by personal preference).

a. Why did you choose this topic?

Since the very beginning the companies have maintained the profile of their customers based on the demographics like age, gender, income, buying history. However, that kind of data can only help in extracting a fraction of relation and would still require companies to spend a hefty amount on marketing the new product. On the other hand, using personality models along with the demographic can be very fruitful. It has the capacity to even predict that youths with age range between 25-30 who are extroverts will show interest in the product advertisement email if it has bright colors and with an image of people etc. The other drawback in the initial days was that we lacked the computational power to model the complex highly dimensional multivariate dataset. With the advancement of processing power and machine learning algorithms like KNN and XGBoost, it became possible to study and model such data.

b. Why does this topic need classification methods?

Customer personality analysis is when the customers of a company are analyzed to find the different types of customers and their reactions to different product campaigns. It helps in understanding the buying behavior of customers in a better way. From this information the company can accordingly decide on the future product campaign and growth.

In personality analysis basically a company finds the target customers for a particular product. The last feature is the one that includes customers' response to this particular campaign. Using classification methods we are able to see for each customer depending on their income, family size, regular buying products that if that particular customer will be responding to the product campaign or not! The response column is going to be our targeted variable which we need to predict by customer's response. If a customer responds to the ad campaign then the response column will have value 1 and if they don't then it will have 0.

2. Literature Review

Laksono et al. (2019) discusses the use of Naïve Bayes algorithm for sentiment analysis of restaurant customer reviews in this paper. The accuracy from the Naïve Bayes algorithm and the TextBlob sentiment analysis is compared to better identify the best method to classify the reviews of customers and to in turn improve the customer service. Using the web crawling method using the WebHarvy tool, review data has been collected from TripAdvisor and then deployed on a WEKA (Waikato Environment for Knowledge Analysis) environment for further analysis. The dataset used consists of 337 data, of which, 269 data is used for training and 68 has been used for testing. After preprocessing the data, the models are trained and the results obtained from Naïve Bayes and TextBlob as analyzed using a confusion matrix and it was discovered that the accuracy of Naïve Bayes and TextBlob was 72.06% and 69.12% respectively (Laksono et al., 2019, p. 54). much the user depends on logical parameters. These types of users will buy products based on their experiences. The other personality trait described by authors is NFA (Need for Arousal). These types of customers buy products based on what feeling they are getting from those products. These users frequently buy items, switch brands, take risks in exploring new products and try to collect as much information about them. The other

personality trait which authors described was NFC (Need for Cognition). These types of users think rationally. They believe more in scientific facts. They don't get affected by the looks and feel of the product.

Iskandar et al. (2020) used the behavior and reaction shown on various tweets posted on Twitter as a way to analyze personality traits of a person. For the purpose of the experiment explained in this paper, around 50 to 500 tweets each from 312 users was used as the dataset and feature extraction, feature selection was performed on this dataset. The dataset was analysed on the basis of MBTI personality traits. The dataset consisted of 20 features, of which, "6 feature twitter and demography, 14 feature extraction from text and user demography." (Iskandar et al., 2020, p. 208). The evaluation metric considered here is accuracy and the experiment has been divided into three scenarios. As described by Iskandar et al. (2020, p. 208), " The first scenario is only implementing word frequency. The second scenario is implementing word frequency and feature extraction. The final and third scenario is implementing feature extraction and feature selection using Chi-Square." The proportion of training data and test data is constantly changed throughout the experiment. And the classification models that are used are Naïve Bayes and K-NN. Two things were concluded at the end of the paper. First, the third scenario of implementing feature extraction and feature selection using Chi-Square is the best scenario as compared to the other two scenarios and secondly, Naïve Bayes classifier provides a better accuracy than K-NN by an average of 10% higher.

Stachl C. et al. (2020) published a survey of using machine learning on personality psychology. They briefly talked about the challenges which researchers faced while applying machine learning models to the complex human behavior data. In their survey they mentioned many researchers used unsupervised learning methods to identify the psychological constructs in the data. They also talked about the recommender's system which were trained on personality data and were being used to suggest products to the user based on their past liking. Researchers used both supervised and unsupervised learning methods to recognize the psychological form in the data. The researchers explained how machine learning became very helpful in using complex datasets related to psychology and predicting personalities. Moreover, the insights gained from those models were used in applications in big organizations. The authors used the PhoneStudy mobile sending dataset and the Big 5 personality traits dataset.

Wang et al. (2020) used GCN techniques to predict personality traits of users from their Facebook status updates or essay information. They also used the Big 5 personality model as it related to the text features. Their GCN model learned from the personality graph per user which was created using the relationships like on user-document, document-word, and word co-occurrence. Their experiment on two public datasets showed that the general personality GCN is much better than the state-of-the-art methods for personality recognition when they don't use the word embeddings from outside. They also found that the personality GCN was more accurate if used on smaller datasets. At the embedding layer of the architecture of personality GCN the authors used word embedding, user embedding and document embedding. The authors used f1-score to evaluate the models. The authors showed from their results that a three-layer or at max four-layer personality GCN gave very good accuracy more than the state-of-the-arts.

Ringbeck et al. (2019), Stachl C. et al. (2020) and Wang et al. (2020) used the Big 5 traits dataset along with their own datasets for personality prediction. Stachl C. et al. (2020) used linear and nonlinear on personality psychology; Wang et al. (2020) used their personality GCN on the personality dataset while Ringbeck et al. (2019) used their Personality Trait Prediction Algorithm (PTPA) model for prediction personality of customers using their online shopping behavior.

3. Data Summary, EDA and Data Pre-Processing

Dataset and Features

A total of 29 columns and 2240 rows comprise the dataset. Different columns represent different aspects of the data, such as customer information, product information, promotions, and where the item was purchased. In addition, data preprocessing is required before any machine learning model can be used. Likewise, the data will be scaled and normalized as well. The table below includes all the features of the data.

ID	**unique identifier of customer**
Year_Birth	birth year
Education	education level
Marital_Status	marital status
Income	yearly income
Kidhome	Number of children in household
Teenhome	Number of teenagers in household
Dt_Customer	Date of enrollment with the company
Recency	Number of days since last purchase
Complain	if the customer complained in the last 2 years then 1 else 0
MntWines	Total money spent on wine in last 2 years
MntFruits	Total money spent on fruits in last 2 years
MntMeatProducts	Total money spent on meat in last 2 years
MntFishProducts	Total money spent on fish in last 2 years
MntSweetProducts	Total money spent on sweets in last 2 years
MntGoldProds	Total money spent on gold in last 2 years

NumDealsPurchases	Total number of purchases made with a discount
AcceptedCmp1	if customer accepted the offer in the 1st campaign then 1 else 0
AcceptedCmp2	if customer accepted the offer in the 2nd campaign then 1 else 0
AcceptedCmp3	if customer accepted the offer in the 3rd campaign then 1 else 0
AcceptedCmp4	if customer accepted the offer in the 4th campaign then 1 else 0
AcceptedCmp5	if customer accepted the offer in the 5th campaign then 1 else 0
Response	if customer accepted the offer in the last campaign then 1 else 0

Data Exploratory Analysis

We performed exploratory data analysis on our dataset to better understand customer buying behaviour and how various aspects affect the response to a marketing campaign. From **Fig. 1**, it can be seen that education has a significant impact on the no. of customers buying a product and responding to a marketing campaign. The reason behind this is that with better education, the chances of getting employed at a reputed firm and thereby earning an income above the average cost of living is more high. When the income is high, there is a high likelihood of buying more products as the buying capacity of the customer also increases. Hence why we see that 50.3% of the respondents are from the well-educated category.

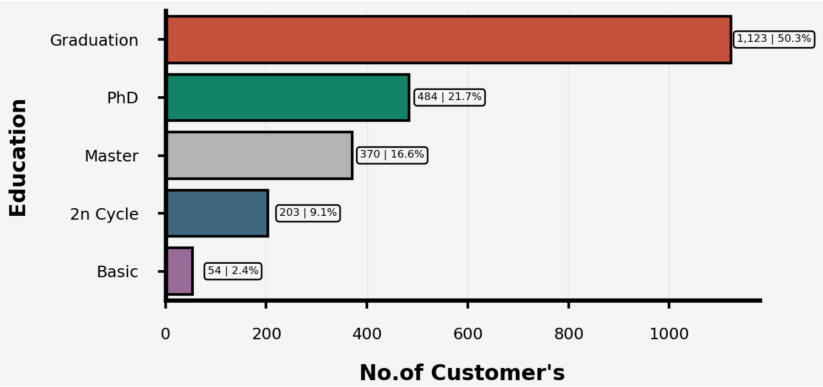

Fig. 1. An EDA of Education v/s No.of Customers

From **Fig. 2**, it can be seen that marital status also an an effect on the no. of customers buying a product and responding to a marketing campaign. The reason behind this can be attributed to the fact that those not in a relationship or bachelors, might not necessarily be cooking at home on an everyday basis. They might eat outside for half of the week, whereas someone who is in a relationship might be more wary of expenditures and tend to cook at home most of the time. Also, if it's a family, the needs of kids also takes higher priority and therefore the responses to marketing campaigns offering products at a competitive price. Therefore, why 64.5% of the respondents are in a relationship.

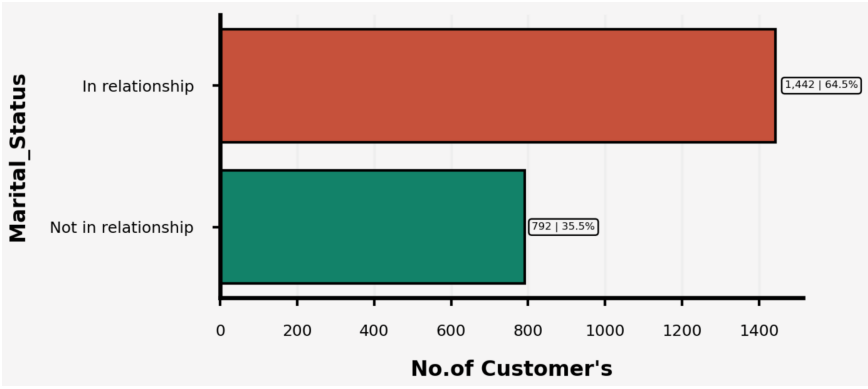

Fig. 2. An EDA of Marital Status v/s No.of Customers

From **Fig. 3**, it can be seen that those with a family size of 2 or 3 members are more. This is due to the fact that as a family of two, there is a possibility of two incomes coming into the household, therefore increasing the affordability of products. Whereas, in a family of three, needs of the kid are what drive for more purchases being made and as a way to reduce the expenses, parents are more willing to spend money when products are on promotions. This is why they are responsive to marketing campaigns.

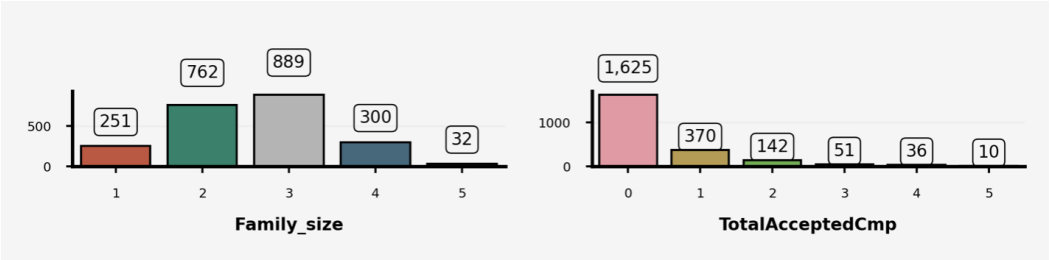

Fig. 3. An EDA of Family Size and Total Accepted Campaigns v/s No.of Customers

Data Pre-Processing

a. We are combining columns such as ‘Kidhome’ and ‘Teenhome’, as they are similar in context.
b. Using the ‘Year_Birth’ column, we are calculating the age of the customer as a separate column.
c. We are using the SDV library (Gaussian Copula method) to generate more data for our dataset.
d. We are performing data aggregation on certain columns to obtain extra information such as age of customer, family size of columns.
e. Outliers in Income where seven Customers’ ‘Income’ is near 150,000 and there is one Customer having ‘Income’ greater than 600000

4. Feature Selection Methods

a. Feature Selection Using Random Forest

As known, Random Forest is made up of multiple decision trees sometimes branching up to thousands of decision trees, where not all decision trees access all features or observations. This aspect of the Random Forest makes the model less susceptible to overfitting. As our dataset is comparatively large, Random Forest works well for our dataset. When coming to product development based on customer personality analysis it is quite necessary to understand various factors such as buying patterns, product interest etc. of the customer. Random Forest ranks the features as important or less important based on their level of purity. More the purity, the more important the feature.

b. Feature Selection Using Logistic Regression

Our dataset contains features that hold values of 0 and 1, for example, whether a campaign has been responded to or not has been indicated by a '0' for **'No'**, and by a '1' for **'Yes'**. Under such conditions, we can make use of Logistic Regression. Logistic Regression is quite useful for analyzing customer personalities.It helps in eliminating duplicating values from the dataset and introduces sparsity in the dataset.Logistic regression is less likely to get over-fit unless it highly dimensional datasets.It also help by shrinking the values of the coefficients of duplicate features to zero. We have also performed feature selection using Logistic Regression with a 95% confidence level, i.e., features that have a p-value less than 0.05 are taken into account as more important features. It usually helps in interpreting model coefficients as a benchmark of feature importance.

c. Selected Features After Feature Selection

A figure showing the features selected using Logistic Regression is given below. The number of features obtained after feature selection using Logistic Regression is **16 features**.

```
summary_as_html = summary.tables[1].as_html()
df_summary = pd.read_html(summary_as_html, header=0, index_col=0)[0]

df_summary.reset_index(level=0, inplace=True)

important_features = df_summary[(df_summary['P>|z|'] < 0.05)]['index'].tolist()

important_features
```

```
['Education',
 'Income',
 'Recency',
 'MntFruits',
 'MntMeatProducts',
 'NumDealsPurchases',
 'NumStorePurchases',
 'AcceptedCmp3',
 'AcceptedCmp4',
 'AcceptedCmp5',
 'AcceptedCmp1',
 'AcceptedCmp2',
 'Kids',
 'Family_size',
 'Age',
 'TotalAcceptedCmp']
```

Fig. 4. A snapshot of selected features using Logistic Regression

A figure showing the features selected using Random Forest is given below. The number of features obtained after feature selection using Random Forest is **10 features**.

```
from sklearn.feature_selection import SelectFromModel
from sklearn.ensemble import RandomForestClassifier

sel = SelectFromModel(RandomForestClassifier(n_estimators = 100))
sel.fit(X, y)
```

```
SelectFromModel(estimator=RandomForestClassifier())
```

```
sel.get_support()
```

```
array([False, False,  True,  True,  True, False,  True, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False,  True,  True, False,
        True, False,  True,  True])
```

```
selected_feat= X.columns[(sel.get_support())]
len(selected_feat)
```

```
9
```

```
print(selected_feat)
```

```
Index(['Income', 'Recency', 'MntWines', 'MntMeatProducts', 'Days_is_client',
       'MntTotal', 'AverageCheck', 'TotalAcceptedCmp', 'ActiveDays'],
      dtype='object')
```

Fig. 5. A snapshot of selected features using Random Forest

5. Comparison Between Different Methods

The baseline models, i.e. models without feature selection have been compared to the models after feature selection. We have implemented three different models, namely, XGBoost, Random Forest, and KNN. It can be seen that the XGBoost baseline model gives the highest accuracy as compared to other models.

Accuracy	XGBoost	Random Forest	KNN
Without Feature Selection	93.972	92.735	83.153
With Feature Selection with Random Forest	89.33	90.108	87.635
With Feature Selection with Logistic Regression	92.270	92.272	81.29

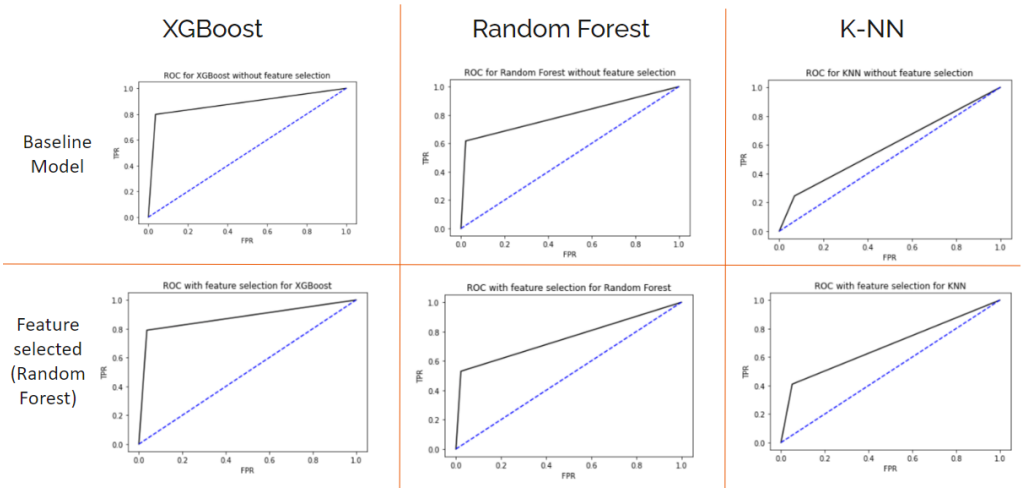

Fig. 6. A Snapshot of the ROC Curves of the Three Models

The above figure is the ROC curve that was obtained after performing all three methods on our data. The first row denotes the ROC curve for the baseline models and the second row shows the ROC curves for the methods obtained after random forest feature selection. From the ROC curves, it can be inferred that, similar to results obtained by using the accuracy metric, the XGBoost method shows the best ROC curve with the TPR (True Positive Rate) values being high and FPR (False Positive Rate) being low while K-NN's ROC curve is less ideal with TPR being low. In the case of Random Forest method, we can a slight decrease in the TPR value after performing feature selection, which could be due to high reduction in the input data features from 29 to 10 using the Random Forest feature selection method which reduces the features for the decision trees formed using the Random Forest method, in turn affecting the accuracy and TPR.

6. Discussions

a. Why is one method better than other methods in your data?

From the method comparison table, which compares the accuracy between the 3 methods XGBoost, KNN and Random Forest, we can see that XGBoost gives the highest accuracy in two cases - with feature selection using Logistic Regression and without feature selection. KNN gives the least accuracy with all cases - with and without feature selection.

KNN performs badly because this method is highly dependent on the number of input features as with many features it will be difficult to determine its nearest neighbors. Sometimes, even after feature reduction using feature selection methods such as using Logistic Regression does not improve the accuracy because of inadequacy of training data or because the model still ends up having more features to give high accuracy. Here, after Logistic Regression, the number of features used for the model is still 16, while the number of features used after using Random Forest for feature selection is 10. Correspondingly, we can see that there is a huge increase in accuracy using KNN when the number of input features become lesser and lesser.

For the Random Forest method, we can see that the feature selection using Logistic Regression works better compared to feature selection using Random Forest, when comparing the accuracy scores and gives an accuracy score comparable to the baseline model without feature selection. The baseline model has high accuracy because it uses several decision trees, selecting different features every time,to come up with the end result. So, more the features, more the variability with the trees and more the accuracy in prediction or classification. The baseline model and the feature selection with logistic regression have more features for model building than when feature selection with random forest is done on the data, resulting in only 10 features for the model. This results in accuracy being more for the baseline model and the model built using feature selection using logistic regression.

XGBoost gives higher accuracy in all three cases compared to KNN method and gives similar or more accuracy compared to Random Forest method because it uses

decision trees like Random Forest to generate results, but attaches a weight to iterations where predictions are wrong, thus learning from past mistakes. This boosts the accuracy of the results and, similar to Random Forest, will most probably work better when there are more features available to generate different decision trees with different features in various iterations.

b. Difference between all features and selected features. Why are selected features important?

The methods that are used for feature selection are random Forest and Logistic Regression using 95% confidence level. The feature selected by two methods is described in the following.

Feature Selection Method	Important Selected Features
Random Forest	Income, Recency, MntWines, MntMeatProducts, Days_is_client, MntTotal, AverageCheck, TotalAcceptedCmp, ActiveDays
Logistic Regression	Education, Income, Recency, MntFruits, MntMeatProducts, NumDealsPurchases, NumStorePurchases', AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, AcceptedCmp2, Kids, Family_size, Age, TotalAcceptedCmp

The features that are selected are described further. These features are very important for our project as we selected them using the random forest classifying features selected method which is one of the embedded methods of selecting the features. The way random forest works is by selecting the feature based on their impurity. Apart from that, the selected features have more importance on the basis of customer's personality analysis. Income is important because for each customer, it depends on their annual income on how much they are willing to spend on groceries and other products. Income is a very huge factor that affects a customer's buying decisions. As our project is based on a customer's decision to respond to an advertisement or not, Income factor is a very important feature.

Similarly, Education is also a very important feature which is related to income as well. These two features are highly correlated. High education most-likely suggests high income and vice versa. That is why these two features are equally very important. Kids, Family_size and Age are similarly important features as the customer's income and family size also influence how much products they are going to buy. Also, the age of the customer often also relates to their income. Recency also becomes a very important factor for the business because it helps keep track of when customers shop and if they decide to buy products after a certain amount of time has passed from their previous purchase. This feature infers the pattern of buying products for individual customers and that is the reason recency has been selected.

MntWines, MntFruits and MntMeatProducts are important features when it comes to customers responding to the next advertisement. These two features give information of each customer's very basic need of usage of wine, fruits and meat products. That

mostly depends on if a customer is going to respond for the next food advertisement or not! Days_is_client is the feature that gives information about the total amount of days a customer is a member. That is mostly important because old customers tend to respond to the ad campaign more than compared to the new ones. NumDealsPurchases and NumStorePurchases are the total number of purchases and total number of deals purchased by a customer in a certain amount of time. MntTotal and AverageCheck is essential features as they give information about total amount of products a customer buys in certain amount of days and average amount of check they receive in certain amount of days respectively. Finally, the TotalAcceptedCmp feature gives an overview of how responsive each customer is to the ad campaigns which will help decide the budget for such campaigns which inturn can increase the sales of the products and the business's profitability.

The correlation between the important features from random forest feature selection is shown in the figure below:

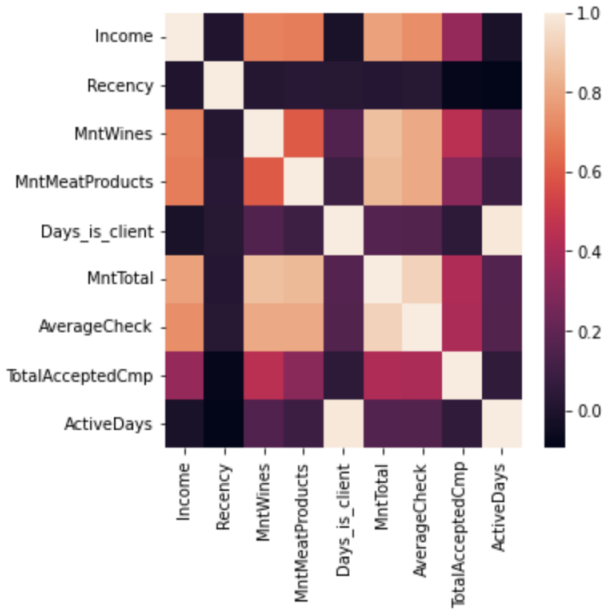

Fig. 6. Correlation Matrix for features selected using Random Forest

The correlation between the important features from logistic regression feature selection is shown in the figure below:

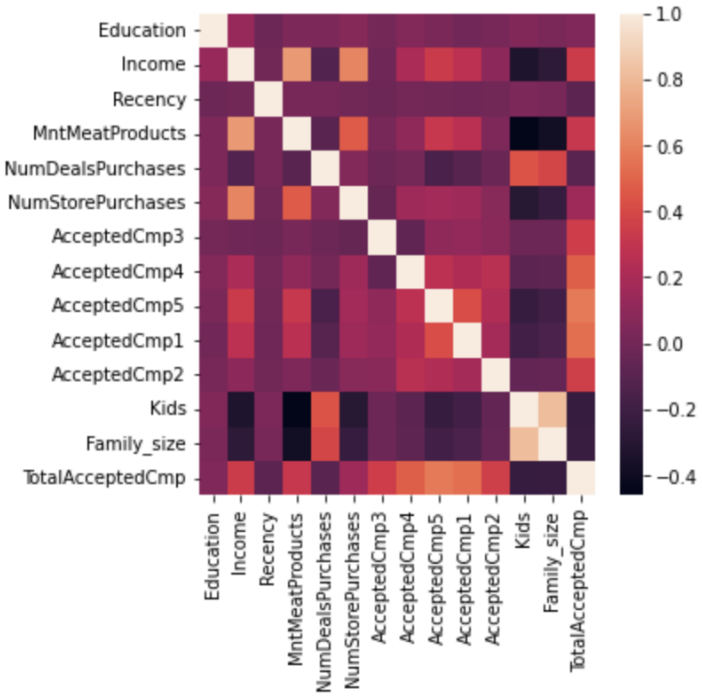

Fig. 6. Correlation Matrix for features selected using Logistic Regression

Some more features apart from the ones obtained by using the Random Forest method were selected in the case of feature selection using Logistic Regression. Features like AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, AcceptedCmp2 and ActiveDays throw lights on the total amount of online ad campaigns a customer has responded to and the total number of days a customer has been active online. All these have a direct impact on the sales of any business and their bottom-line production.

c. What is the meaning of your result? How to explain your result (interpretability) based on your domain knowledge and references.

Based on the domain knowledge, the importance of all the selected features are explained in the above section. Coming to the results obtained by three different algorithms and using two different feature selection methods, we can say that the best result obtained was by the XGBoost method although that was with the baseline model, meaning without using any feature selection methods. But the difference of accuracy is very minor that suggests that selected features are important but de-selected features are also equally important for the classification and it has the same effect on the target variable.

Similar to results discussed in one of the papers in the literature review, where, Iskandar et. al. (2020) observed that Naïve Bayes performed better than KNN. KNN performance dropped further after feature selection. In general, KNN is relatively less accurate and this can be attributed to various factors. One of them being that every characteristic of the method has the same result on calculating distance. KNN also

doesn't work well with large and high-dimensional data, both of which are characteristic to our dataset.

As conclusion, for our data, using our methods of XGBoost, Random Forest and KNN and using our feature selection methods of Random Forest and Logistic Regression, we can see that, when accuracy is the evaluation metric considered, sometimes more accuracy achieved by the baseline model can be achieved with reduced features as in the case of KNN, where accuracy increases almost 4 points when feature selection was performed using the Random Forest method. At the same time, KNN dropped in performance, with less accuracy when feature selection using logistic regression was performed. When coming to XGBoost and Random Forest methods, feature selection using Logistic Regression performed better and achieved accuracy comparable to using all features. This leads us to infer that not all methods are equal or suitable for every data or every process and proper, intelligible, selection of the best methods for the chosen metric would yield better results.

References

[1] Dubey, A. (2021, December 7). *Feature Selection Using Random forest - Towards Data Science*. Medium. https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

[2] Iskandar, A. F., Utami, E., & Prasetio, A. B. (2020). Impact of Feature Extraction and Feature Selection on Indonesian Personality Trait Classification. *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*. https://doi.org/10.1109/icoiact50329.2020.9332107

[3] Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019). Sentiment Analysis of RestaurantCustomer Reviews on TripAdvisor using Naïve Bayes. 2019 12th International Conference on Information & Communication Technology and System (ICTS). https://doi.org/10.1109/icts.2019.8850982

[4] K. (2021, October 8). *Customer Segmentation: Clustering* 🛍️🛒🛒. Kaggle. https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering

[5] Ringbeck, D., Seeberger, D., & Huchzermeier, A. (2019). Toward Personalized Online Shopping: Predicting Personality Traits Based on Online Shopping Behavior. *SSRN Electronic Journal*. Published. https://doi.org/10.2139/ssrn.3406297

[6] Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality Research and Assessment in the Era of Machine Learning. *European Journal of Personality*, *34*(5), 613–631. https://doi.org/10.1002/per.2257

[7] S. C. Boerman, S. Kruikemeier, and F. J. Zuiderveen Borgesius, "Online Behavioral Advertising: A Literature Review and Research Agenda," *Journal of Advertising*, vol. 46, no. 3, pp. 363–376, Jun. 2017, doi: 10.1080/00913367.2017.1339368.

[8] Syaliman, K. U., Nababan, E. B., & Sitompul, O. S. (2018). Improving the accuracy of k-nearest neighbor using local mean based and distance weight. *Journal of Physics: Conference Series*, *978*, 012047. https://doi.org/10.1088/1742-6596/978/1/012047

[9] M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, and A. Košir, Eds., *Emotions and Personality in Personalized Services*. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-31413-6.

[10] Manoj, S. (2021, April 23). *Feature Selection Methods | Feature Selection Techniques in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/discovering-the-shades-of-feature-selection-methods/

[11] Wang, Z., Wu, C. H., Li, Q. B., Yan, B., & Zheng, K. F. (2020). Encoding Text Information with Graph Convolutional Networks for Personality Recognition. *Applied Sciences*, *10*(12), 4081. https://doi.org/10.3390/app10124081