The determinants of coupon usage: the key customer and coupon characteristics that determine the proneness to redeem coupons

MSc Economics & Business

Data Science & Marketing Analytics

Academic year: 2021-2022

Name: Célina Maria Morad

Student number: 501226

Supervisor: Erjen van Nierop

Second assessor: Kathrin Gruber

Abstract

The use of coupons is prevalent in the consumer goods industry and in many other markets. Advertisements and promotion deals are among the most crucial marketing strategies for those businesses. However, a significant challenge the consumer goods industry faces regarding coupon campaigns is the low redemption rates. Therefore, to efficiently and effectively utilize the coupon campaigns identification of coupon-prone customers is needed. With the Logistic regression and the Random Forest classification algorithm the key customer and coupon characteristics are identified that determine coupon redemption. According to the results obtained in this paper, the key customer characteristics are income and age. Moreover, the marital status of the customer, their accommodation status and their family size show all to be statistically significant. Regarding the coupon characteristics, the category of the couponed product shows to be of importance for coupon redemption. The results also suggest that the likelihood of coupon redemption is higher for discounted products from established brands than for products from local store brands. With this identification managers and marketers know what type of customers to target with their available coupons and what type of coupons to issue in order to reach higher coupon redemption rates.

Table of contents

1. Introduction

In the year 1887, the Coca-Cola Company used an innovative advertising technique where they offered consumers hand-written tickets that they could redeem for a complimentary glass of Coca-Cola, which is now known as couponing (Tuttle, 2010). Since then, couponing has become a popular promotion tool and is defined as a certificate allowing consumers to get reduced prices at the time of purchase (Schultz et al., 1998).

Providing coupons to customers has been popular for many years and still is in the consumer goods industry. Advertisements and promotion deals are among the most crucial marketing strategies for those businesses. However, a significant challenge the consumer goods industry faces regarding coupon campaigns is the low redemption rates. Previous studies have shown that approximately 320 billion USD worth of coupons were issued to customers in the consumer goods industry in 2015, of which only 2.5 per cent got redeemed (Inmar, 2016; Lalwani & Wang, 2019).

Many benefits can be identified from promotion deals, such as brand exposure, boosting sales, and increasing customer loyalty. However, with the low redemption rate, the efficiency and effectiveness of the coupon campaigns can be questioned. Managers and marketers must discover the optimal utilisation of promotional campaigns to achieve a satisfactory effect. The ideal coupon campaign would add value to the customers while at the same time does not add any additional costs for the company (Goh & Bockstedt, 2013).

Therefore, identifying consumers who may be more or less likely to redeem a coupon can play a valuable role in the consumer goods industry. With this identification, the company can apply a better targeting strategy to achieve the desired goal. Moreover, the popularity of online shopping and ordering leads to the increasing usage of behavioural targeting. For example, targeting customers based on their personal life, online behaviour, geographic location, and purchase history could increase the coupon redemption rate. Moreover, given the possible 50 million USD budget for advertisement and coupon campaigns for one product alone in the consumer goods industry, a highly effective and efficient rate is crucial (Reibstein & Traver, 1982).

From a managerial perspective, having more knowledge and insights regarding consumers who redeem the coupon provides multiple advantages. Companies could fine-tune their segments to whom the coupons are distributed to achieve redemption and minimise the negative effect on profit. Additionally, recognising coupon characteristics that could influence the coupon redemption rate, such as the type of product or brand, may provide valuable insights to customize the coupon.

To evaluate the promotional campaigns, marketers need a measurement or identification of the customers' coupon proneness and the prediction of their redemption behaviour. Therefore, with the fictional data of the company ABC, this paper aims to shed light on the key customer and coupon characteristics that are of importance for coupon redemption. Logistic regression and Random Forest will be performed during the analysis to illustrate the key customer and coupon features and their relationship to coupon redemption.

Accordingly, the main research question entails:

> *What is the relation of customer and coupon characteristics with the redemption rate of coupons?*

To answer this research question more precisely and in the context of the data set, several sub-questions are obtained and analysed in this study:

1. *What is the relationship between income and coupon redemption?*
2. *What is the relationship between age and coupon redemption?*
3. *What is the relationship between household size and the redemption of coupons?*
4. *Is the coupon redemption rate different for different product categories, such as food, no-food and services?*
5. *Does the coupon redemption rate differ between established store brands and local store brands?*

After the provision of an introduction where the research topic and relevance of this thesis are highlighted, the rest of this study will be structured as follows. Section 2 presents the literature review and the established hypotheses. Section 3 elaborates on the methodology and the data used in this paper. Section 4 offers the hypotheses testing and research results. Section 5 provides the discussion, and Section 6 discusses the limitations of this study and suggests a path for further research. Lastly, section 7 features the conclusion.

2. Literature review and hypotheses development

Previous studies and associated literature related to coupon redemption can be divided into two streams. One stream focuses on the customer characteristics and examines socio-economic, demographic, and psychological drivers of coupon usage by individuals and households. In contrast, the other stream emphasises the coupon characteristics and other non-demographic characteristics to identify the significant features of coupon redemption. Therefore, the following sub-chapters include the literature review divided into those two streams and illustrate the developed hypotheses.

2.1 Customer characteristics

Coupons continue to be an essential promotional tool in many consumer goods industries. However, due to the low redemption rates, managers are seeking a solution or, better said, an enhancement in the effectiveness of their promotional campaigns by identifying coupon-prone customers. The socio-economic and demographic determinants most often studied in research to analyse coupon-prone customers include income, age, education, household size, presence of children and marital status (Cronovich et al., 1997).

2.1.1 Income

According to Cronovich et al. (1997), households with high incomes use coupons less often than households with lower incomes. The study by Lee and Brown (1985) contravenes this outcome since their research states that high-income households are significantly more likely to use coupons than low-income households. Bawa and Shoemaker (1987) examine the coupon usage for different products and find a statistically insignificant relationship between income and coupon usage. The insignificant income effects are also proven by Narasimhan (1984), Goodwin (1992), and by Chiou-Wei and Inman (2008).

From a theoretical viewpoint, microeconomic theories, including opportunity cost and marginal utility theory, suggest that consumers with lower incomes have more excellent utility and lower opportunity costs associated with coupon redemption than consumers with high incomes. Similarly, the higher-income groups incur higher opportunity costs related to coupon search, handling, and redemption than lower-income groups and should, therefore, exhibit relatively

lower coupon redemption rates. However, the available research does not strongly support these theories (Noble et al., 2017).

2.1.2 Age and education

The demographic characteristics of consumers that may also influence the redemption of coupons are age and education. Goodwin (1992) and Ward and Davis (1978) included age categories in their models to account for the differences in coupon usage related to age. Both studies found that middle-aged customers are more likely to redeem a coupon. Yet, Goodwin (1992) states that age does not strongly influence the propensity of coupon usage. Similar results are found by Cronovich et al. (1997). The study done by Cronovich et al. (1997) controls for the household's life cycle, as it experiences changes in the employment status of its adults and changes in the presence and age of its children. Their findings indicate that households aged 45-54 and adults over 65 years are slightly more likely to use coupons. The obtained results for the other included age categories show an insignificant relationship between age and coupon usage.

Analysing the relationship between education and coupon redemption, Cronovich et al. (1997) suggest no systematic relationship between education level and coupon propensity. The findings of other studies did indicate a significant positive relationship between education and coupon redemption and discussed this outcome by explaining that households with more excellent education want more diversity, have lower substitution costs, and more efficient management of time to collect the offered coupons (Narasimhan, 1984; Bawa & Shoemaker, 1987).

2.1.3 Household size

The studied effect of the socio-economic variables household size and the presence of children on the coupon redemption rates shows again conflicting results in the literature. Some studies report that the presence of children decreases the likelihood of using coupons (Lee & Brown, 1985; Narasimhan, 1984). On the contrary, Cronovich et al. (1997) show that the household size and number of children in the households positively and significantly affect coupon redemption rates.

Considering the opportunity costs theory, it would predict that households with children may have less time available for managing coupons. Therefore, this should lead to lower coupon redemption rates for households with children than households without children. Yet, marginal utility theory predicts, holding income constant, that the presence of children holds less disposable income in the household, which makes the economic benefit of coupon usage more useful (Noble et al., 2017; Schiller, 1997).

To summarise, the previous studies have produced conflicting results on the various socio-economic and demographic determinants of coupon usage. These inconsistencies are probably due to the differences across studies. This could be explained by the types of grocery products analysed, the type of data analysed, and the set of control variables included. These differences inhibit and confound the comparison of the relative importance of the many socio-economic and demographic determinants.

The emerging hypotheses that will be tested in this paper are the following:

H_1: Ceteris paribus, customers with lower incomes are more likely to redeem coupons than customers with higher incomes.

H_2: The customer's propensity for coupon usage increases with age, ceteris paribus.

H_3: Ceteris paribus, bigger households are more likely to redeem coupons than smaller households.

2.2 Coupon characteristics

Much research has been done on traditional couponing, while new forms of coupons have emerged over the years, like online couponing. Due to this change in couponing, the customer's attitude towards couponing can be affected differently. Nielsen (1965) identified five coupon characteristics influencing coupon redemption rates. These factors are 'method of distribution', 'size of product class', 'rate of discount', 'the face value of coupon' and 'brand distribution'. Prior research also shows that coupon redemption behaviour varies between product categories. Therefore, managers must gain insights into the coupon proneness at the product category level

before formulating a promotional strategy (Bawa and Shoemaker 1987; Blattberg and Neslin 1990; Webster 1965).

2.2.1 Face value and product category

Reibstein and Traver (1982) studied the impact of coupon characteristics on redemption rates and found that higher face value coupons are associated with higher redemption rates. These authors mean by coupon face value, the value of the coupon in cents which are discounted from the product related to the coupon. This significant positive relation between the coupon face value and the coupon redemption rate is also proven in the study done by Leone and Srinivasan (1996).

Swaminathan and Bawa (2005) estimated the coupon proneness at the product category level and analysed the categories of coffee, detergent, beauty salon, and oil change. Their findings suggest that consumers vary in their coupon usage across the product categories. Coupon-prone customers could disregard coupons in a specific product category since they are brand loyal, and at the same time, no coupons being available for their favoured brand. Danaher et al. (2015) covered the product categories of snack food, menswear, shoes, and others and found the highest coupon redemption rate for the snack food product category.

Moreover, Swaminathan and Bawa (2005) found that the propensity to redeem coupons correlates across categories. The extent of the correlation depends on the nature of the class. Their findings suggest that a joint couponing campaign for detergent and coffee would be more effective than promoting either product with a service like an oil change.

2.2.2 Established brands vs local store brands

Chiou-Wei and Inman (2008) differentiate between the effect of private brands and national brands on coupon redemption. Private brands or so-called local store brands are often regarded as having a lower price and lower quality, which implies being inferior in quality. National brands or established brands use nationwide advertising to reaffirm their quality and, therefore, are less likely to be considered low in quality and more prone to coupon use (Sawyer & Dickson, 1984, Garretsona et al., 2002; Chiou-Wei & Inman, 2008). According to Cronovich et al. (1997), households that buy local brands instead of established brands are significantly less likely to use coupons than consumers who only occasionally purchase local store brands.

Cronovich et al. (1997) explain this finding by the fact that most coupons distributed in the USA are for established brands.

2.2.3 Distribution

Due to the increasing penetration of Internet access, the traditional form of couponing has evolved into online couponing distributed online through, for example, direct mail. Many previous studies (Ward & Davis, 1978; Reibstein & Traver, 1982; Henderson, 1985) have highlighted that more easily available coupons tend to have higher redemption rates. Jung and Lee (2010) found a higher average redemption rate of online coupons than that of offline coupons. One of the main reasons for the high redemption rate of online coupons would be the 'selective' nature of the distribution. Moreover, they obtained that the age of customers influences the online coupon redemption rate since mainly online users are between 20 and 30 years old (Jung & Lee, 2010).

The emerging hypotheses that will be tested in this paper are the following:

H_4: Ceteris paribus, coupon redemption within the product category Food is higher than in the product categories Non-food and Services.

H_5: Ceteris paribus, coupons for established brands induce higher coupon redemption rates than coupons for local store brands.

3. Methodology

The objective of this paper is to identify the customer and coupon characteristics that are key determinants of the consumers' likelihood of redeeming coupons. Managers and marketers can target the coupon campaigns more effectively to reach their desired goals by analysing these determinants. Therefore, they know who and how to target to increase coupon redemption rates with this knowledge.

3.1 The data

To analyse the essential determinants of coupon redemption and the relation of these key determinants to the redemption rate, the data provided by Kaggle will be used. Kaggle is an online public data platform where users can find published data sets, experiment with them, and share their skills with other users. The so-called 'Predicting Coupon Redemption' data set is used for this research.

The retrieved data set contains information on a fictional company called ABC. ABC is an established Brick & Mortar retailer that regularly organises coupon marketing campaigns for its diverse product range. The promotions given out by ABC are shared across various channels, including email and notifications (Kanojia, 2020).

The data set is divided and available through 6 different data sets containing information on the campaigns, the coupon, user demographics, the products and information on previous transactions. Therefore, to perform the analyses, the relevant data sets are merged. The data sets are joined by the unique item, coupon and customer identification codes.

After merging the data sets into one complete data set, including all the relevant attributes, it is checked for missing values. No missing values were present, however, the data set included empty cells. For the attribute covering the number of children in the household, it is straightforward that an empty cell stands for no children in the household. This is also proven by the variable measuring the family size since it shows a size of 1 or 2 for the observations with empty cells. Therefore, the empty cells are converted to cells having the value "0" for this attribute measuring the number of children. Empty cells were also detected for the attribute covering the customer's marital status. These observations were removed due to the missing

information. This reduced the number of observations in the data set from 3,807,856 to 2,219,752.

The variable measuring the income of the customer is given by 12 different brackets of income. Meaning that a higher income corresponds to a higher income bracket. However, to make the results more useful and to reduce the imbalance between the number of observations in the income brackets, the brackets are merged. After merging, the income brackets 1 till 4 correspond to low income, 5 till 7 correspond to middle income and 8 till 12 indicate the high-income bracket.

Lastly, the categorical variable measuring the product category included 19 different levels. To simplify this categorical variable, the levels are merged into the levels Food, Non-food and Services. The Food category covers Alcohol, Grocery, Natural Products, Prepared Food, Seafood, Vegetables (cut), Bakery, Meat, Packaged Meat, Dairy Juices & Snacks and Salads. The Non-food category covers Flowers & Plants, Fuel, Skin & Haircare, Garden, Miscellaneous and Pharmaceutical. And the Services category includes Travel and Restaurant. In Table 1, an overview is given of all the variables in the data set.

Variable name	Variable type	Measure
redemption_status	Factor	Binary variable measuring the redemption of a coupon. 0 - Coupon not redeemed 1 - Coupon redeemed
brand_type	Factor	Assigning the valid brand type of the coupon. Whether it is for an established brand or a local store brand. 0 – Established brand 1 – Local store brand
category	Factor	The product category of the couponed product. The three categories are: Food, Non-food and Services

age_range	Factor	The age range of the customer in years. Age ranges are: 18-25, 26-35, 36-45, 46-55, 56-70 and 70+.
marital_status	Factor	The marital status of the customer, which is assigned married or single. 0 – Married 1 - Single
rented	Factor	Binary variable measuring if the accommodation of the customer is rented or not. 0 – not rented accommodation 1 - rented accommodation
family_size	Integer	The number of family members. Given by 1, 2, 3, 4, and 5.
no_of_children	Integer	The number of children in the family. Given by 0, 1, 2, 3 and 4
income_bracket	Factor	Label Encoded Income Bracket. 1 till 4 = low income 5 till 7 = middle income 8 till 12 = high income

Table 1: Description of all the included variables in the data set

3.2 Classification methods

Consumers have an unobserved tendency to use coupons. This coupon proneness depends on a combination of the attractiveness of the coupon and the features of the customer. Two classification algorithms will be used to identify these key customer and coupon characteristics for coupon redemption. The applied algorithms to this classification problem are the Logistic regression and the Random Forest model. The programming language R processes the data and performs the research.

3.2.1 Logistic regression

Logistic regression is used to identify the relationships between the dependent variable and the included explanatory variables in the model. The dependent variable is a binary variable

indicating only two outcomes with the binary Logistic regression. In this analysis, the dependent binary variable, measuring the coupon redemption, is coded as 0 being equal to "No" and 1 being equivalent to "Yes" (Sperandei, 2014; Pusztova & Babic, 2020; Menard, 2002).

The Logistic regression predicts the likelihood of a customer redeeming the coupon or not. If the possibility of redemption is greater than 0.5, the assumption is that the customer redeemed the coupon. No coupon redemption is assumed if the likelihood is less than 0.5. Therefore, using components of linear regression transformed in the logit scale, Logistic regression identifies the most robust linear combination of variables with the most significant probability of detecting the observed outcome, coupon redemption or no coupon redemption (Stoltzfus, 2011; Ren et al., 2021).

Some assumptions are checked before training the Logistic regression on the data set. The first assumption states that the observations in the data set are independent. This assumption is met since the data set shows observations per individual customer together with their original identification number, and no repeating is detected. Second, the assumption indicating the absence of multicollinearity among the explanatory variables is checked (Stoltzfus, 2011). This assumption is analysed according to the Generalized Variance Inflation Factors (GVIF) due to categorical variables in the data set. The chosen threshold is 5 for the GVIF, indicating no presence of multicollinearity. See Table 2 for the output of the GVIF analysis.

Explanatory variable	GVIF	df	GVIF^(1/(2*df))
brand_type	1.0252	1	1.0125
category	1.1289	2	1.0308
age_range	2.3393	5	1.0887
marital_status	4.3397	1	2.0832
rented	1.3630	1	1.1675
family_size	42659.2163	4	3.7910
no_of_children	18412.4134	3	5.1387
income_bracket	1.8867	11	1.0293

Table 2: GVIF multicollinearity analysis results

The GVIF provides a combined measure of collinearity for each group of explanatory variables. To make the outcomes of GVIF comparable across dimensions, Fox and Monette (1992)

introduced the GVIF^(1/(2*df)), where df is equal to the number of coefficients in the subset—specifying the proportional change of the standard error of confidence interval of the coefficients due to the level of collinearity. Therefore, it indicates the reduction in the precision of the estimated coefficients due to collinearity (Fox & Monette, 1992). Table 2 shows the highest GVIF and GVIF^(1/(2*df)) for the independent variables measuring the size of the family and the number of children in the household. This presence of multicollinearity is rational since the number of children will also be counted in the total size of the family and vice versa. According to the threshold of 5 and only referring to GVIF^(1/(2*df)), the variable measuring the number of children shows high multicollinearity.

The independent variables in the constructed Logistic regression model are selected based on the researched literature. However, the occurrence of overfitting in the full model, including all the explanatory variables, should be acknowledged. Moreover, due to the presence of multicollinearity, a Logistic regression model without the variable measuring the number of children in the household is obtained. Therefore, the following Logistic regressions are used in this research.

<u>The full Logistic regression model:</u>

$$\text{logit}\,(P_i) redemption\_status = \log\left(\frac{P_i}{(1-P_i)}\right)$$

$$= \beta_0 + \beta_1 brand\_type + \beta_2 category + \beta_3 age\_range + \beta_4 marital\_status + \beta_5 rented + \beta_6 family\_size + \beta_7 no\_of\_children + \beta_8 income\_bracket + \varepsilon_i$$

<u>The Logistic regression model accounted for multicollinearity:</u>

$$\text{logit}\,(P_i) redemption\_status = \log\left(\frac{P_i}{(1-P_i)}\right)$$

$$= \beta_0 + \beta_1 brand\_type + \beta_2 category + \beta_3 age\_range + \beta_4 marital\_status + \beta_5 rented + \beta_6 family\_size + \beta_7 income\_bracket + \varepsilon_i$$

3.2.2 Random Forest

Decision trees are a simple and easily interpretable classifier tool. Classification decision trees can handle high-dimensional data and generally show good accuracy. However, a fully-grown decision tree will overfit the training data, and the resulting model might not be performant for predicting the outcome of the test data (Safavian & Landgrebe, 1991; Song & Ying, 2015). The technique of pruning can be used to control this overfitting problem, which results in a more straightforward tree with fewer splits and better interpretation at the cost of a bit of bias. Nonetheless, the Random Forest (RF) model goes beyond the decision tree method.

RF is a machine learning method that uses a non-parametric algorithm to predict an outcome and select essential determinants. The RF model consists of many individual uncorrelated trees that operate as an ensemble. Each decision tree in the forest is constructed from a randomly selected bootstrap sample. Therefore, during the construction of the decision trees, each time a split in the tree is considered, a random sample of m predictors is chosen as split candidates from the complete set of p predictors. This method ensures that the strongest predictor is not in the top split every time, so other predictors will have a higher chance of being chosen. Accordingly, the fundamental concept behind the RF model is that the combined uncorrelated trees outperform a single tree, resulting in a more accurate model in the end without pruning that has high variance and low bias (Belgiu & Drăguţ, 2016; Breiman, 2001; James et al., 2013; Pal, 2005; Pusztova & Babic, 2020).

With the RF classifier, the relative importance of the different characteristics in the data set can be identified. It will yield a variable importance plot that ranks the accounted variables in terms of their predictivity regarding the target variable. The variable importance analysis shows the ordering of the variables by their mean decrease in accuracy estimated by the RF model or by using the mean decrease in Gini (Belgiu & Drăguţ, 2016; Breiman, 2001; Hamdiui et al., 2018; Pal, 2005). The mean decrease in accuracy is defined as the difference between the out-of-bag (OOB) error resulting from a random subset of the data and the OOB error from the original data set. Moreover, to select the important variables, the ordering of mean decrease in accuracy is used. The mean decrease in Gini is more prone to different kinds of biases, such as selection bias and bias due to the correlation between predictors since the Gini variable importance increase when a predictor appears more often in the trees (Nicodemus et al., 2010; Boulesteix et al., 2012).

However, this method does not define how a feature contributes to the model's predictions. To identify the relationship between the target variable and the features, multiple partial dependence (PD) plots will be used. A PD plot shows the marginal effect that a feature has individually on the predicted output, in this case, coupon redemption. The used feature is modified, holding all else constant, and the changes in the mean prediction are observed. For classification, the PD plot displays the probability for a particular class given the different values of the feature (Friedman, 2001; Cutler et al., 2007; Greenwell, 2017).

Another point needing awareness is that those classification problems are mainly imbalanced, which means that one of the classes is underrepresented in the data set. A high imbalance in the data set can badly affect the Logistic regression and RF model since they are constructed to minimise the overall error rate. Therefore, the models will focus more on the prediction accuracy of the majority class, which results in poor accuracy for the minority class (Chen et al., 2004; Maalouf, 2011). This imbalance problem, giving biased classification and overfitting of the models, can be handled by several methods. (Batista et al., 2004; Chen et al., 2004; Xie et al., 2009). In this paper, the over-sampling technique will be used.

The re-sampling technique can be divided into under-sampling and over-sampling. With under-sampling, the majority class decreases, while with over-sampling the minority class increases. Therefore, with the under-sampling method, the majority class's observation gets deleted to match the number of the minority class. This will lead to a loss of information. With over-sampling, synthetic data is produced randomly based on the minority class. The most used methods for over-sampling are SMOTE (Synthetic Minority Over-sampling Technique) and ROSE (Random Over Sampling Examples). SMOTE works by taking minority class samples and its k nearest neighbours generating synthetic examples (Chawla et al., 2002). The ROSE over-sampling method also generates new artificial data from minority classes but is based on a smoothed bootstrap approach (Menardi & Torelli, 2014). The sub-section 'Exploratory analyses' will cover both methods and show the resulting balance in the data sets. Moreover, the Logistic regression models and the RF model will be applied to the imbalanced and balanced data sets to identify the performance differences.

Lastly, the RF model is trained and fine-tuned on the train data, both balanced and unbalanced, and evaluated on the validation set. Accordingly, the best mtry, corresponding to the number

of predictors considered for each split in the tree, in the RF function is set to be 4 providing the lowest OOB error. Regarding the number of decision trees (Ntree), research has shown that Ntree has a low impact on classification accuracy. Many studies identified a Ntree of 500 to be the threshold since the errors stabilise, and no further improvement is identified when increasing the number of trees beyond 500 (Belgiu & Drăguţ, 2016; Du et al., 2015; Ghosh et al., 2014; Kulkarni & Sinha, 2012; Lawrence et al., 2006). This also holds for this research and, therefore, the number of trees in the RF model is set to 500.

3.2.3 Evaluation

The Logistic regressions and the RF models are trained on the imbalanced, ROSE balanced, and SMOTE balanced train data sets. Therefore, the established models are evaluated and compared according to their performance measures in order to obtain the best models. The performance measures used are the accuracy metric, the Area Under the Curve (AUC) value and the confusion matrix analysing sensitivity and specificity. The accuracy metric indicates the percentage of correct classification and is defined as:

$$Accuracy = TP + TN / (TP + FP + FN + TN)$$

Where,

TP = True Positive, the number of positive classes correctly predicted
TN = True Negative, the number of negative classes correctly predicted
FP = False Positive, the number of positive classes wrongly predicted
FN = False Negative, the number of negative classes wrongly predicted.

The other performance measure is the AUC value given by the ROC curve. The ROC curve is a graph with the False Positive rate on the x-axis from 0 to 1 and the True Positive rate on the y-axis from 0 to 1. The ROC curve captures the classification rate when varying the classification thresholds. After that, the AUC value is constructed, indicating the overall performance ability. If the AUC value is closer to 1, the classifier shows good accuracy. However, the classifier offers no good accuracy if the AUC value is more comparable to 0.5 (Mandrekar, 2010).

Lastly, the sensitivity and specificity measures are obtained from the confusion matrix. Sensitivity, the True Positive Rate, measures the proportion of actual positives that are correctly identified. Specificity, the False Positive Rate, measures the proportion of real negatives correctly identified. Therefore, the sensitivity measure shows the rate of accurately estimated redeemed coupons. In contrast, the specificity measure shows the rate of correctly estimated non-redeemed coupons. Sensitivity and specificity are defined as:

$$\text{Sensitivity} = TP / (TP + FN),$$

$$\text{Specificity} = TN / (TN + FP).$$

With the obtained performance measures, the optimal models are identified for analysing the key customer and coupon characteristics and their relationship with coupon redemption (Kirasich & Sadler, 2018; Pusztova & Babic, 2020).

3.3 Exploratory analysis

After cleaning the data set and making a few adjustments, the data set is analysed. For the RF model, the existing data set is too large and not processable. Therefore, a smaller data set is randomly sampled from the original data set and used for the research. The sample contains 739.916 observations, which is approximately 1/3 of the actual size of the data set. See appendix A for the descriptive statistics.

First, the target variable is checked for balance. The target variable measuring coupon redemption shows that 91.50 per cent of the coupons included in the data set did not get redeemed, and 8.50 per cent of the coupons did get redeemed. This shows that data is imbalanced, and the main interest lies in the minority class of redeemed coupons. As described in the previous section, both the Logistic regression and RF model are sensitive to imbalanced data. Therefore, before training the models the data used for training needs to be balanced in order to reduce the risk of biased outputs and to obtain robust models. In order to achieve balance in the data set, the over-sampling methods of SMOTE and ROSE are applied to the train data only.

The data set is split into a train data set that covers 80 per cent of the data and into a validation and test data set that both cover 10 per cent of the data. After balancing the train data set with the SMOTE methods, it shows a perfect balance of 50 per cent of redeemed coupons and 50 per cent of not redeemed coupons. Balancing the train data with the ROSE method shows a balance of 50.04 per cent of not redeemed coupons and 49.96 per cent of redeemed coupons. The time to perform the balance in the train data set was more significant for the SMOTE method than for the ROSE method. The Logistic regression and the RF model are applied to the unbalanced data set and to the obtained balanced data sets computed by SMOTE and ROSE.

Moreover, the explanatory variables in the data set are analysed concerning the target variable, which measures the coupon redemption status; 0 is equal to “No”, and 1 is equivalent to “Yes”. The outcome of these analyses is given in Figure 1 by enumeration and in Figure 2 by percentages.

The redemption of coupons is the outcome of interest. According to Figure 1 and Figure 2, coupon redemption is higher for established brands than for local store brands. However, the

distribution of coupons for established brands is higher than for local store brands, partially explaining this difference. The redemption rate is shown to be less than 10 per cent for established brands but even lower for local store brands, which indicates to be lower than 5 per cent. Due to these low values, it is not clearly visible in the figures. Therefore, Table 3 is added to provide the division more precisely in percentages. Moreover, Table 3 provides the ratio of coupon redemption per included variable and the ratio of no coupon redemption in percentages.

Coupon redemption shows to be higher for married customers and for customers who do not have rented accommodation. Both show to be between 5 and 10 per cent. This is interesting since married customers are more likely to have a double income to cover the expenses and are, therefore, in less need to cut costs than customers with a single payment. As well as for customers who do not have a rented accommodation are expected to earn enough money to possess a house and have less need to cut expenses by using coupons on their store purchases. However, another explanation for this difference in redemption between having a rented accommodation and not having a rented accommodation can be homeowners' mortgage costs and property taxes. These extra costs can cause homeowners to need to cut expenses and use coupons more than customers with rented accommodation.

As mentioned, the variable measuring the different product categories is modified from 19 product categories to only three, divided by Food, Non-food and Services. The product category Food shows the highest coupon redemption rate compared to the other categories. It shows that coupon redemption is less than 10 per cent in the Food category and less than 5 per cent in the Non-food category. Regarding the category Service, the number of redeemed coupons in this category is too low to show the redemption status visually. According to Table 3, coupon redemption in the category Services is approximately 0.002 per cent. Moreover, this figure of product category and coupon redemption shows the problem of redemption rates well since the number of not redeemed coupons outweighs the number of redeemed coupons.

The plots accounting for the variables measuring the size of the family and the number of children in the household show presence of coupon redemption in all classes. The family size of 2 shows the highest coupon redemption rate compared to the other classes and is almost 5 per cent. This could indicate a married customer or a single customer with one child. This can also be drawn from the figure accounting for the number of children in the customer’s household since having no child at all and one child shows the highest rate of coupon

redemption, see Table 3 for the exact percentages. The next highest coupon redemption rate is detected by a family including four individuals, which indicates a higher number of children in the family. The 2 and 3 children classes show coupon redemption of approximately 1 per cent. However, these insights are surprising since larger households have higher grocery expenses, and coupons can help reduce these expenses. Regardless, this is not shown in the figures since it shows that a customer with a small household makes the most use of the coupons. This could be explained by the fewer leisure time customers with large households have due to the more children to take care of and, therefore, less time to manage the available coupons.

Customers between 36 and 45 years old show the most use of coupons of almost 2.5 per cent, which could be explained by the presence of children and or having a house indicating high costs. Customers within the 26-35 and 46-55 age range show the second highest rates of coupon redemption, being approximately 2 per cent. Moreover, customers that earn an income within the middle class show the highest rate of coupon redemption. However, a decrease in coupon use can be detected for customers with high incomes since the coupon redemption rate decreases from 6.48 per cent precisely to 0.27 per cent.

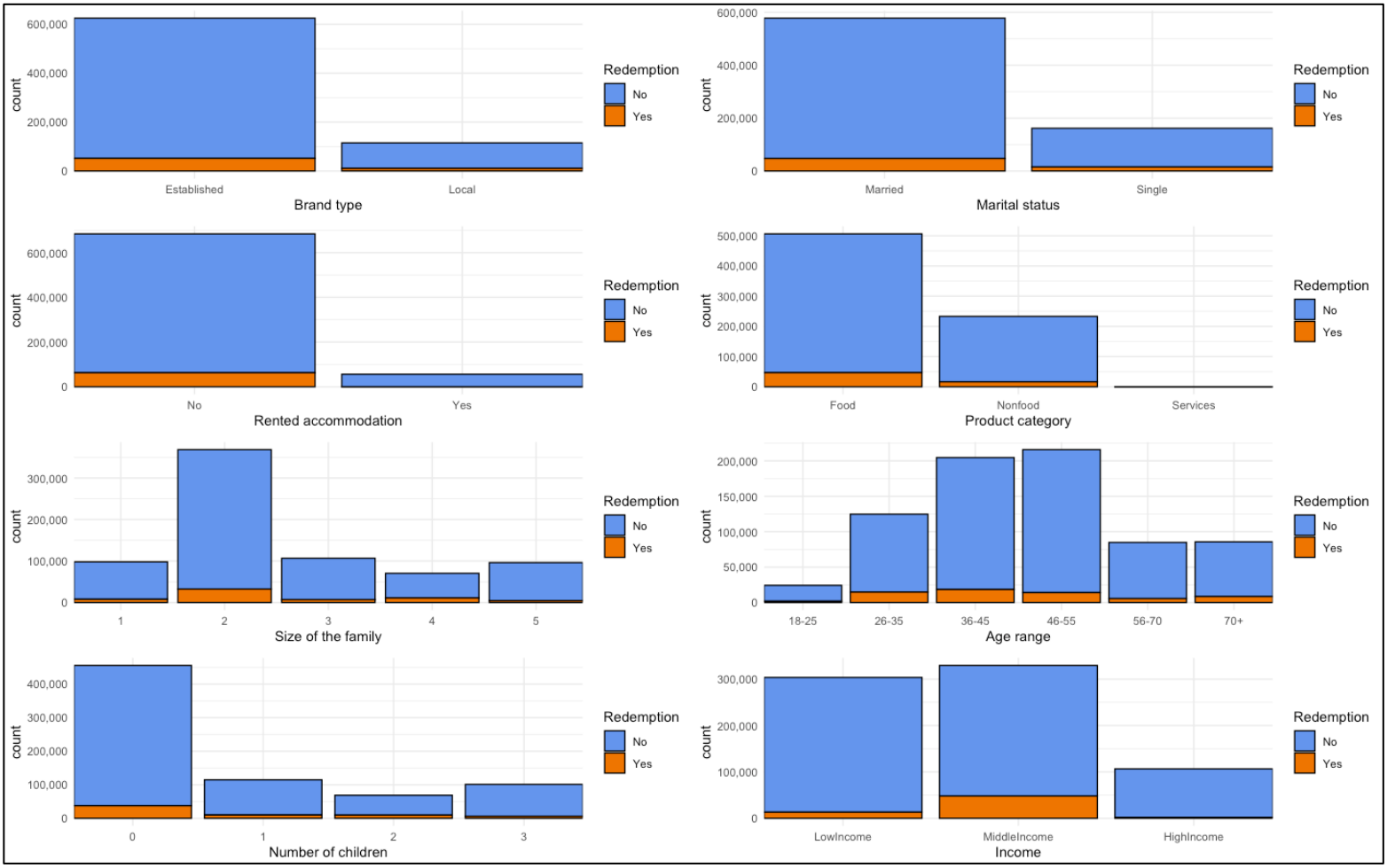

Figure 1: All Explanatory variables concerning the target variable in enumeration

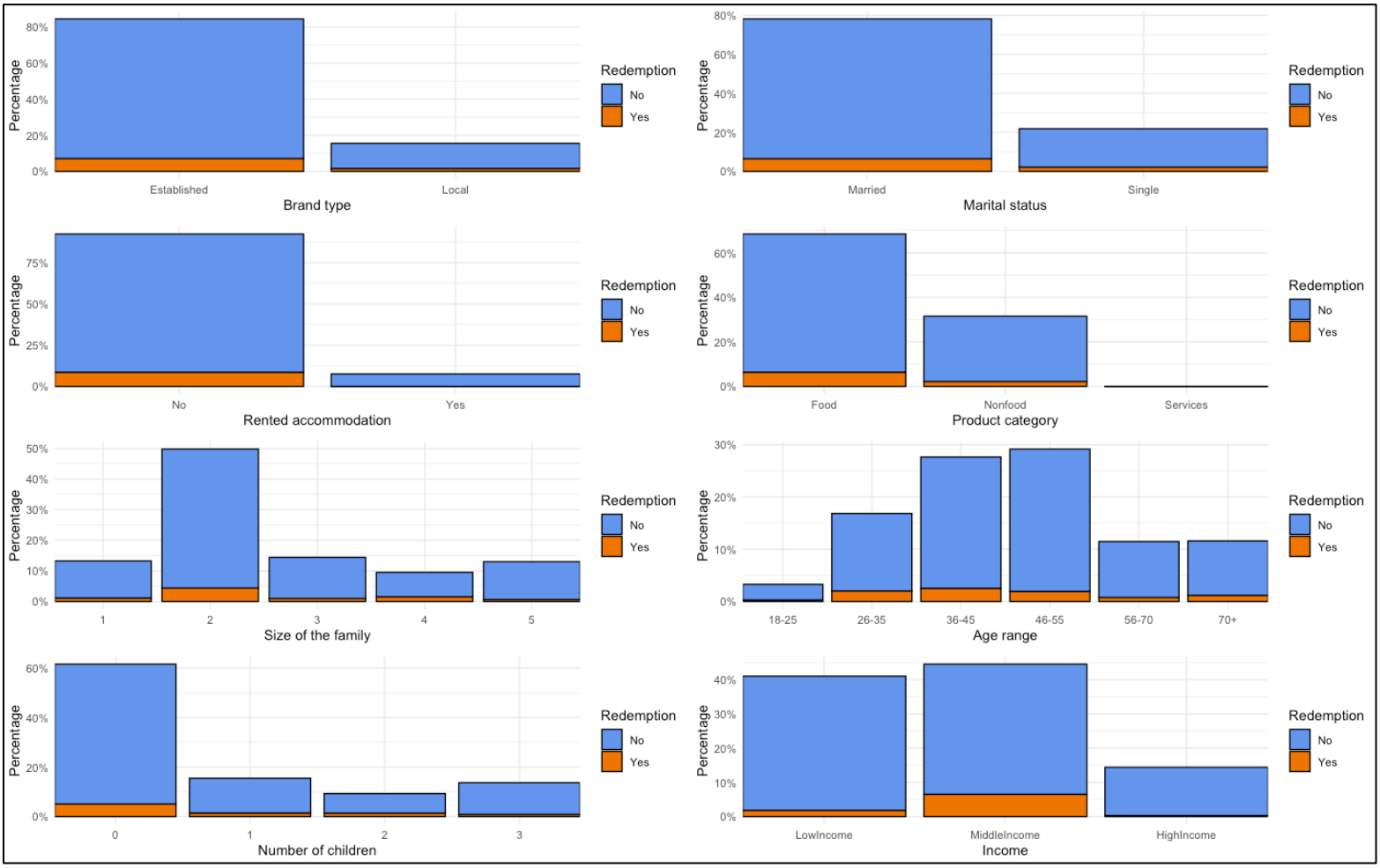

Figure 2: All Explanatory variables concerning the target variable in percentages

Variable	Coupon redemption - No	Coupon redemption - Yes	Ratio – No coupon redemption	Ratio – Coupon redemption
Brand type – Local	14.04%	1.49%	90.4%	9.6%
Brand type – Established	77.48%	6.99%	91.7%	8.3%
Marital status – Married	71.82%	6.40%	91.8%	8.2%
Marital status – Single	19.70%	2.08%	90.5%	9.5%
Rented accommodation – No	83.97%	8.44%	90.9%	9.1%
Rented accommodation – Yes	7.55%	0.04%	99.5%	0.5%
Product category – Food	62.08%	6.28%	90.8%	9.2%
Product category – Non-food	29.40%	2.19%	93.1%	6.9%
Product category – Services	0.04%	0.002%	96.2%	3.8%
Family size – 1	12.13%	1.05%	92%	8%
Family size – 2	45.47%	4.40%	91.2%	8.8%
Family size – 3	13.53%	0.93%	93.6%	6.4%
Family size – 4	8.06%	1.49%	84.4%	15.6%
Family size – 5	12.33%	0.61%	95.3%	4.7%
Number of children – 0	56.53%	5.05%	91.8%	8.2%
Number of children – 1	14.18%	1.32%	91.5%	8.5%
Number of children – 2	8.00%	1.29%	86.1%	13.9%
Number of children – 3	12.81%	0.81%	94%	6%
Age range – 18 till 25	3.05%	0.24%	92.8%	7.2%
Age range – 26 till 35	14.86%	1.98%	88.3%	11.7%
Age range – 36 till 45	25.15%	2.49%	91%	9%
Age range – 46 till 55	27.31%	1.89%	93.5%	6.5%
Age range – 56 till 70	10.68%	0.75%	93.4%	6.6%

Age range – 70+	10.48%	1.13%	90.3%	9.7%
Low income	39.27%	1.77%	95.7%	4.3%
Middle income	38.07%	6.48%	87.5%	12.5%
High income	14.14%	0.27%	96.6%	3.4%

Table 3: All Explanatory variables concerning the target variable in percentages and the ratio's in percentages

3.4 Model selection

Before testing the hypotheses and analysing the results, the best Logistic regression and Random Forest model need to be chosen. The obtained models are compared and evaluated according to the aforementioned performance measures explained in sub-chapter 3.2.3. The performance measures are the accuracy metric, AUC value and the confusion matrix analysing sensitivity and specificity. First, the Logistic regressions not accounting for multicollinearity are compared, and an overview of their performance measures is given in Table 4.

Logistic regression	Accuracy	AUC	Sensitivity	Specificity
Unbalanced	0.9152	0.5000	0.0421	0.9910
Balanced with ROSE	0.6403	0.6972	0.7659	0.6286
Balanced with SMOTE	0.6432	0.7106	0.7917	0.6295

Table 4: Performance measures accuracy, AUC, sensitivity and specificity for the Logistic regression models not accounting for multicollinearity

The accuracy measure shows to be the highest for the Logistic regression model trained on the unbalanced data set. The model shows to be 91.52% accurate. However, the accuracy metric provides deceiving results since the model is trained on an unbalanced data set. Here the unbalanced Logistic regression model is biased towards the majority class, which is no coupon redemption, and the minority class, coupon redemption, hold a minimum effect on the overall accuracy.

Referring to the Logistic regression models trained on the balanced ROSE and SMOTE data set no significant difference can be obtained between the performance measures. The accuracy of the Logistic regression model shows a difference of 0.29 percentage points between the ROSE and SMOTE methods in Table 4.

The AUC for the Logistic regression model trained on the unbalanced data set shows to be the lowest, indicating that the model has a low discriminatory ability. The Logistic regression models not accounting for multicollinearity and trained on balanced data show similar AUC values. The highest value is given by the SMOTE model, which shows a 0.7106 probability of correctly distinguishing between coupon redemption and no coupon redemption.

The full Logistic regression model trained on the unbalanced data set gives the highest specificity rate shown in Table 4. However, the sensitivity rate is very low, implying that the model correctly estimates only 4.21% of redeemed coupons. Therefore, in evaluating a model based on sensitivity and specificity, both values need to be as high as possible for the model to be a good fit. Comparing the Logistic regressions of ROSE and SMOTE balanced data sets, again, little difference in measurements is obtained. However, the Logistic regression trained on the balanced SMOTE data set shows the best sensitivity and specificity measures.

Next, the performance measures of the Logistic regression models accounting for multicollinearity are discussed and an overview is given in Table 5.

Logistic regression	Accuracy	AUC	Sensitivity	Specificity
Unbalanced	0.9152	0.5000	0.0415	0.9942
Balanced with ROSE	0.6465	0.7013	0.7695	0.6352
Balanced with SMOTE	0.6458	0.7009	0.7673	0.6345

Table 5: Performance measures accuracy, AUC, sensitivity and specificity for the Logistic regression models accounting for multicollinearity

Again, the highest accuracy metric is given by the Logistic regression model trained on the unbalanced data set. However, the possibility of biased outcomes from this model needs to be recognized. The Logistic regression trained on the ROSE balanced data set shows a higher accuracy compared to the Logistic regression model trained on the SMOTE balanced data set. However, the difference is just 0.07 percentage points between their accuracy measures.

Regarding the AUC value, the lowest measure is given by the Logistic regression model trained on the unbalanced data set. The AUC measures of the balanced Logistic regression models are again similar. The ROSE model shows a probability of 0.7013 while the SMOTE model shows a probability of 0.7009 correctly distinguishing between coupon redemption and no coupon redemption.

The sensitivity rate and the specificity rate of the Logistic regression model trained on the unbalanced data set indicate a not good fit of the model. The Logistic regression models trained on the balanced data set show high values for both the sensitivity and specificity measures. However, again little difference between the two models is identified. The sensitivity rate of

the ROSE model is the highest and shows to be 0.7695 which implies that the model correctly identifies the redeemed coupons by approximately 77%. The specificity rate is also higher for the ROSE model by 0.07 percentage points compared to the SMOTE model.

To select the best Logistic regression model, the models not accounting for multicollinearity are dropped. Since they violate one of the Logistic regression assumptions and include the problem of overfitting the train data. Moreover, the unbalanced Logistic regression model that accounts for multicollinearity is also dropped due to its bias towards the majority class no-coupon redemption. Therefore, the best Logistic regression for testing the hypotheses in this paper is the Logistic regression trained on the ROSE balanced data set, which also accounts for multicollinearity. The performance measures of the ROSE model are shown to be higher compared to the measures of the Logistic regression model trained on the SMOTE balanced data set.

Next, the RF models are analysed with their performance measures given in Table 6.

Random Forest	Accuracy	AUC	Sensitivity	Specificity
Unbalanced	0.9326	0.6288	0.2629	0.9947
Balanced with ROSE	0.8524	0.8859	0.9264	0.8455
Balanced with SMOTE	0.8295	0.8892	0.9611	0.8173

Table 6: Performance measures accuracy, AUC, sensitivity and specificity for the Random Forest models

The accuracy measure is the highest for the RF model trained on the unbalanced data set, implying an accuracy rate of 93.26%. The RF models trained on the balanced data set, with the ROSE and SMOTE method, also provide high accuracy measures of 85.24% for the ROSE RF model and 82.95% for the SMOTE RF model.

According to the AUC value, the best performance is the balanced RF model trained on the SMOTE data set. However, a little difference of 0.33 percentage points is shown between the RF model trained on the balanced ROSE data set and the model trained on the SMOTE data set. The lowest AUC value is given by the unbalanced RF model.

The low AUC value of the unbalanced RF model is also proven by the sensitivity and specificity rates. The sensitivity rate measuring the True Positive Rate is equal to 0.2629. This implies that the model correctly estimates only 26.29% of redeemed coupons. However, the specificity rate is shown to be the highest compared to the other RF models. Evaluating both rates shows the consequence of training the model on an unbalanced data set, indicating a bias towards the majority class of not-redeemed coupons.

The RF models trained on the balanced data set do not show this bias and perform better than the unbalanced RF model. For the balanced RF models, both sensitivity and specificity are high and little difference between the models can be obtained. However, the best RF model based on the performance measures is the RF model trained on the SMOTE balanced data set. Therefore, the SMOTE RF model shows the highest ability of 96.11% to correctly classify the class of interest, which is redeemed coupons.

To conclude, the models chosen to identify the key customer and coupon characteristics and their relationship to redemption rate are 1) the Logistic regression model accounting for multicollinearity and trained on the balanced ROSE data set and 2) the RF model trained on the balanced SMOTE data set.

4. Results

With the purpose of testing the hypotheses and providing an overview of the key customer and coupon characteristics for coupon redemption and their relationship, the best Logistic regression and Random Forest model are used. The used Logistic regression is accounted for multicollinearity and trained on the ROSE balanced data set. The best obtained Random Forest model is trained on the balanced SMOTE data set. First, the results of the Logistic regression are presented, and the hypotheses are tested. Thereafter, the results from the RF model are provided and analysed.

4.1 Logistic regression results and hypotheses testing

With the purpose of testing the hypotheses and providing an overview of the relation between the customer and coupon characteristics with redemption, the Logistic regression is used, and the results are given in Table 7.

	Coupon redemption = Yes	
Variable	**Estimate Log(odds)**	**Odds-ratio**
Middle-income class	1.6378***	5.1435
High-income class	-0.6716***	0.5109
Age range from 26 till 35	0.9219***	2.5141
Age range from 36 till 45	0.6125***	1.8450
Age range from 46 till 55	0.0395***	1.0402
Age range from 56 till 70	0.1435***	1.1543
Age range from 70 and above	0.7486***	2.1140
The size of the family	0.0785***	1.0817
Non-food product category	-0.5281***	0.5897
Services product category	-1.0098***	0.3643

Local store brand	-0.0981***	0.9066
Marital status being single	0.6189***	1.8569
Rented accommodation	-3.1536***	0.0427
Intercept	-1.2203***	0.2951

Table 7: Logistic regression results and Odds-ratio with significant codes
**** p<0.0001; ** p<0.01; * p<0.05; . p<0.1*

4.1.1 H_1: Customers with lower incomes are more likely to redeem coupons than customers with higher incomes, ceteris paribus.

Examining the results from the Logistic regression in Table 7, the income classes middle-income and high-income are both statistically significant at the 0.0001 p-level. The reference group is the income bracket measuring low incomes.

Referring to the middle-income bracket, the odds of redeeming a coupon increase compared to the odds of the low-income bracket, holding all else constant. For example, the odds of a customer within the middle-income bracket redeeming a coupon is 5.1435 times that of a customer redeeming a coupon with low income, ceteris paribus.

For customers with income within the high-income bracket, the odds of redeeming a coupon is 0.5109 times less than that of customers with low-income, holding all else constant.

Therefore, according to the Logistic regression results, hypothesis 1 can only be partially rejected. The hypothesis cannot be rejected for the comparison between the low- and high-income customers. This also holds for the comparison between the middle- and high-income customers since customers with middle-income are more likely to redeem coupons than customers that earn more and belong to the high-income class. However, the comparison between the low and middle income indicates higher redemption rates for customers with middle income. Moreover, the estimated coefficients indicate a non-linear relationship.

4.1.2 H_2: The customer's propensity for coupon usage increases with age, ceteris paribus.

The Logistic regression results given in Table 7 show that the categories for the variable measuring age are all highly significant at a significance level of 0.01 per cent. The reference group is the age range 18 till 25 years old.

The results for customers within the age range of 26 to 35 years old show that the odds of them redeeming coupons is 2.5141 times that of customers redeeming coupons within the 18-25 age range, holding all else constant.

However, after the 26-35 age range, the odds show a decrease. For customers within the 36-45 age range the odds of them redeeming a coupon is 1.8450 times that of customers in the 18-25 age range, holding all else constant. For customers within the 46 till 55 age range, the odds of them redeeming coupons is 1.0402 times that of customers in the 18-25 age range, ceteris paribus.

After the 46-55 age range, the odds increase again. The odds of redeeming a coupon in the 56-70 age range is 1.1543 times that of redeeming a coupon within the 18-25 age range. Moreover, for the age range of 70+, the odds of redeeming a coupon is 2.1140 times that of redeeming a coupon within the 18-25 age range. Holding all else constant.

Therefore, hypothesis 2 can be rejected according to the results of the Logistic regression.

4.1.3 H_3: Bigger households are more likely to redeem coupons than smaller households, ceteris paribus.

According to the Logistic regression results in Table 7, the variable measuring the number of people in the household of the customer is statistically significant at the 0.0001 p-level.

For one unit increase in the size of the customer's family, the odds of coupon redemption increase by a factor of 1.0817, ceteris paribus. Therefore, hypothesis 3 cannot be rejected.

4.1.4 H_4: Coupon redemption within the product category food is higher than in the product categories non-food and services, ceteris paribus.

The Logistic regression results in Table 7 show that the variable indicating the product categories are statistically significant at the 0.0001 p-level for both the category measuring non-food products and for the category measuring the services.

The odds of coupon redemption in the Non-food product category is 0.5897 times that compared to coupon redemption in the reference group Food, ceteris paribus. The odds for coupon redemption in the product category Services is lower. Here it indicates that the odds of coupon redemption in the Service category is 0.3643 times that of coupon redemption in the product category Food, holding all else constant.

Therefore, hypothesis 4 cannot be rejected.

4.1.5 H_5: Coupons for established brands induce higher coupon redemption rates than coupons for local store brands, ceteris paribus.

According to Table 7, providing the results of the used Logistic regression model, the variable indicating the type of brand shows to be statistically significant at the 0.01 per cent level. The reference group is the established brands.

Holding all else constant, the odds of coupon redemption for local store brands is 0.9066 times less than that of coupon redemption for established brands. Therefore, hypothesis 5 cannot be rejected.

Then lastly, the remaining estimated coefficients of the variables indicating the marital status of the customer and their type of accommodation are discussed. Both variables are given statistical significance at the 0.01 per cent level. The reference group is ‘married’ for the variable indicating the marital status and ‘no rented accommodation’ for the variable measuring the accommodation. Therefore, the odds of coupon redemption for single customers is 1.8569 times that of married customers. The odds of coupon redemption for customers with a rented accommodation is 0.0417 times less than that of coupon redemption for customers with no rented accommodation. Holding all else constant.

4.2 Random Forest results

The RF model trained on the balanced SMOTE data set obtained the following variable importance plot shown in Figure 3. The variable importance analysis shows the ordering of the predictors by their mean decrease in accuracy. See appendix section B for the variable importance plot based on the mean decrease in Gini.

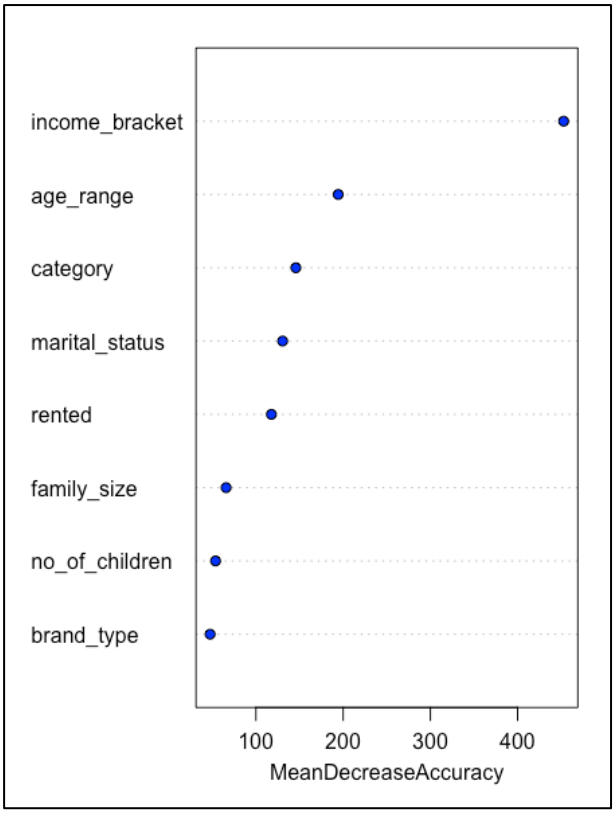

Figure 3: Variable importance plot Random Forest model

According to Figure 3, the predictors measuring income and age have the most impact on the target variable measuring coupon redemption. Moreover, the predictor 'category' is given high importance by the RF model, implying high predictivity regarding the target variable. Therefore, the top three variables have a high contribution to the model's prediction power and show the importance of specific customer and coupon characteristics for coupon redemption.

Next, the PD plots are given to identify the relationship between the key predictor variables and the target variable coupon redemption, holding all else constant. The predictors of interest are shown in Figures 4a till 4c. See appendix section C for the PD plots of the remaining predictors.

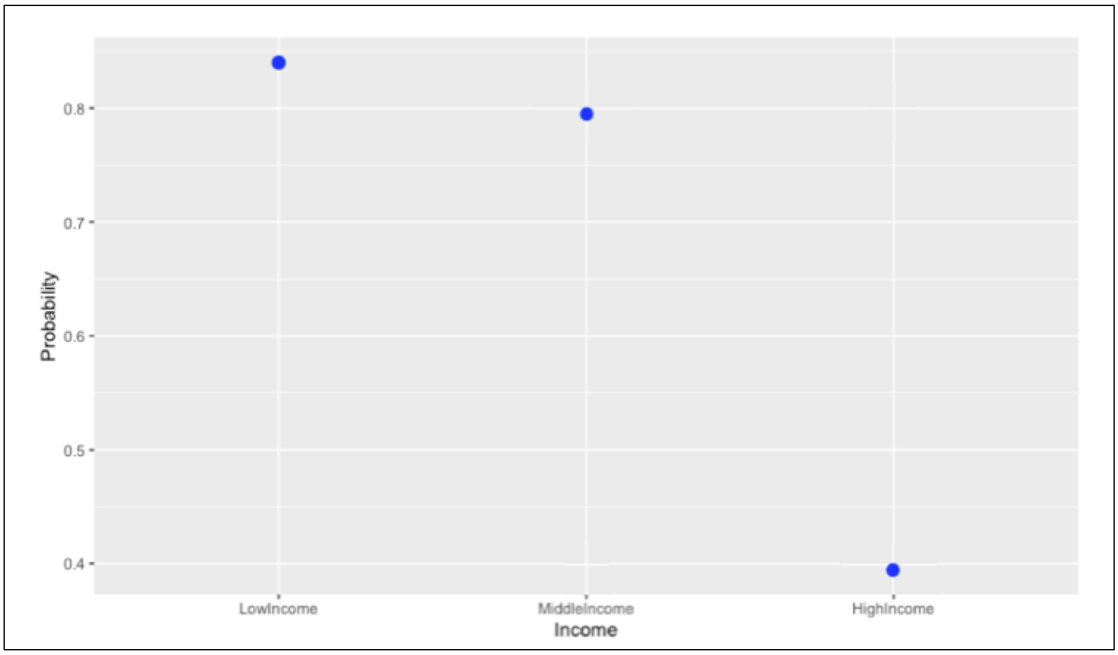

Figure 4a: Partial dependence plot on the predictor income_bracket

The marginal effect of the predictor measuring the income bracket of the customer on coupon redemption shows the highest probability of coupon redemption for customers with low income. A lower chance of coupon redemption is shown for customers within the middle-income group and the lowest likelihood is given for customers within the high-income bracket.

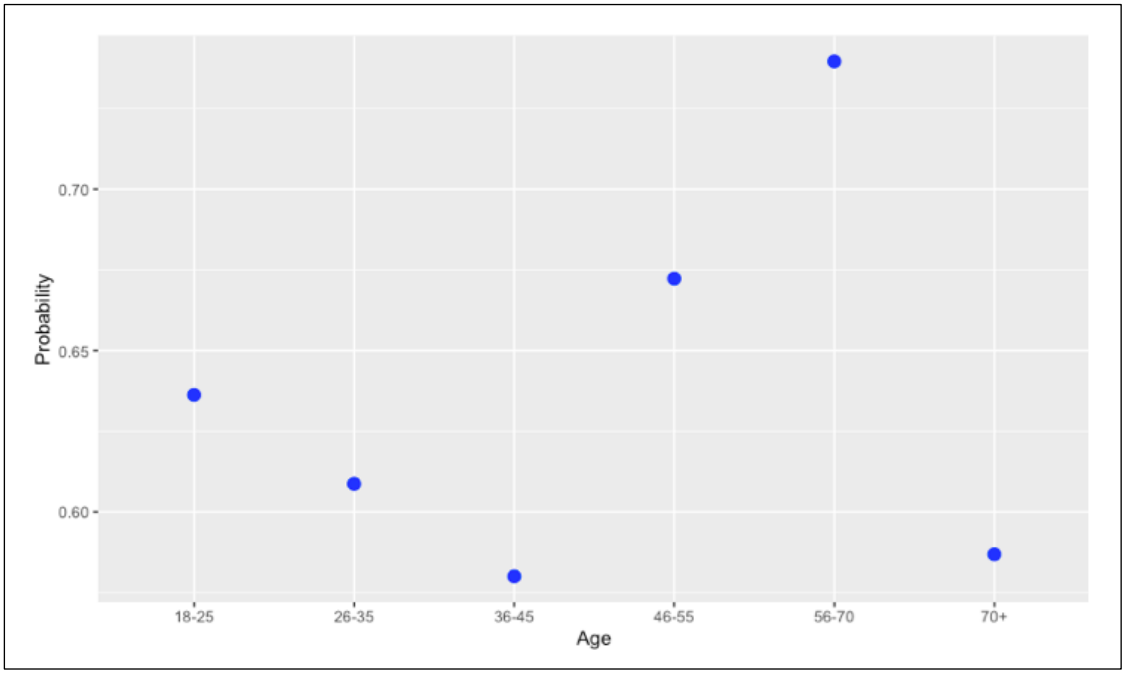

Figure 4b: Partial dependence plot on the predictor age_range

The marginal effect of age on coupon redemption shows the highest likelihood of coupon redemption for customers aged 56 to 70 years old. Moreover, a relatively high probability is displayed for customers in the age range of 18 to 25 and 46 to 55 years old. The lowest likelihood of coupon redemption is given by customers in the age range of 36 to 45 years old.

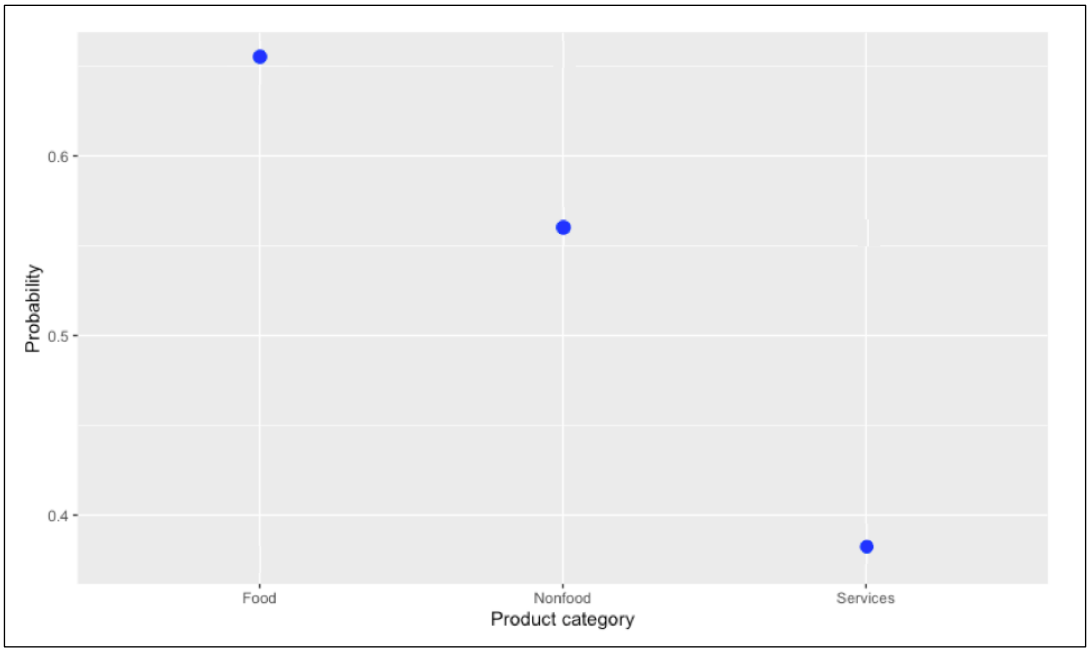

Figure 4c: Partial dependence plot on the predictor category

The marginal effect of the variable measuring the different product categories on coupon redemption is given in Figure 4c. Here the likelihood of coupon redemption is the highest for coupons valid within the Food category. The probability of coupon redemption shows to be the second-highest for coupons valid within the Non-food category. Lastly, the lowest probability of coupon redemption is shown for the category Services.

5. Discussion

The discussion seeks to illustrate the key customer and coupon characteristics and the relationship of the included variables with coupon redemption, given the results in the previous section.

First, discussing the results regarding the customer characteristics, the income of the customer shows to be a significant determinant regarding coupon redemption. This significant relationship between income and coupon redemption is also proven by Cronovich et al. (1997) and by Lee and Brown (1985). Cronovich et al. (1997) stated that customers with high incomes are less likely to use coupons than customers with low incomes. The results from this paper also prove this relation but only between the low-income class and the high-income class and between the middle-income class and high-income class. The obtained results are also supported by the microeconomic theory regarding opportunity costs and utility since customers with lower incomes have higher utility and lower opportunity costs associated with coupon redemption than customers with higher incomes. However, the obtained insights also support the findings of Lee and Brown (1985) since customers with middle income are more likely to redeem coupons compared to customers with low income. Overall the relationship between income and coupon redemption is given to be non-linear. According to the results of the RF model, customers with low incomes have the highest probability of coupon redemption.

The obtained results regarding the age of customers show a significant relation with coupon redemption. However, according to the found results coupon redemption does not increase with age. The results indicate higher redemption rates within the 26-35 age range and above 70 years old. The PD plot indicates high probabilities of coupon redemption for middle-aged customers. This is also proven by Cronovich et al. (1997), Goodwin (1992) and by Ward and Davis (1978).

The relationship between household size and coupon redemption is statistically significant. According to the obtained result, the likelihood of coupon usage increases with the customer's family size, ceteris paribus. Arguably, bigger households are prone to having higher expenses and, therefore, receive higher utility from coupon redemption. However, only Cronovich et al. (1997) found similar results while other studies suggest that bigger households decrease the likelihood of coupon usage due to the time constraint of childrearing. The PD plot of family

size in Appendix C supports the increase in the likelihood of coupon redemption when the size of the family increases.

The results regarding the variable measuring whether the customer is married or single, indicate a higher probability of coupon redemption for customers that are single. See the results in Table 7 in Chapter 4 and Appendix C for the PD plot. This is interesting since smaller households are expected to be less likely to use coupons than bigger households (Goodwin, 1992). The reason behind this result could be that single people have more time to collect and manage the available coupons than people who are married and maybe even have children.

Moreover, according to the obtained results of the Logistic regression, customers with rented accommodation are less likely to use coupons compared to customers with non-rented accommodation. This could be explained by the extra costs homeowners have compared to people renting a residence. However, according to the PD plot obtained with the RF model, customers with rented accommodation have a higher probability of coupon redemption than customers with no rented accommodation. The location of the residence indicates the distance from the redemption location, the ABC Brick & Mortar store, and is an important factor for redemption behaviour. Customers that are located further away from the redemption location have concomitantly more cost in time. Therefore, this greater time constraint and that rented accommodation are more present in cities instead of the suburbs, could explain the higher likelihood of coupon redemption for customers with rented accommodation (Chiou-Wei, 2004; Chiou-Wei & Inman, 2008; Rhee & Bell, 2002)

The obtained variable importance plot by the RF model identifies income and age as key customer characteristics for coupon redemption. The size of the family is given low predictivity regarding coupon redemption. The variables measuring marital status and accommodation also show low importance but higher predictivity power than the variable measuring family size.

Next, the coupon characteristic ‘category’ shows statistical significance. Therefore, coupon redemption lies within different categories of available products, which is also proven by Swaminathan and Bawa (2005). The obtained results indicate higher coupon redemption for the category Food. This is also proven by Danaher et al. (2015) since they found high redemption rates within their researched food category. The probability of coupon redemption is the lowest for the product category Services.

The obtained results for the coupon characteristic ‘brand type’ indicate higher coupon redemption for products of established brands. Products of established brands are regarded as high-quality products and are sold at a higher price than local store brands. Therefore, when a coupon is available for those inferior products more customers will be interested in using the coupon due to receiving higher quality for a lower price. This is also supported by the provided literature in Chapter 2.2. See Appendix C for the PD plot regarding the variable measuring the type of product brand.

According to the variable importance plot of the RF model, the category of the product is an important determinant for coupon redemption. However, the brand type of the product is given the lowest importance of all included variables.

6. Limitations

Coupons are not new to the consumer goods industry. They are being extended to other markets and developed into more up-to-date digital coupons. In this paper, the research is done on data from a fictional company in the consumer goods industry that shares the coupons through different channels such as email and flyers. However, no information is given in the data set about which type of coupon is used by the customer during the transaction. Researching the effectiveness of the individual channels could provide insights into the best distribution channel to reach coupon-prone customers and the type of coupon mostly redeemed.

Moreover, information regarding the basket size of the customer could further develop the research. Basket size is the total number of products purchased by the customer in a single shopping visit. Including this in the study could improve identifying the different relations of coupon redemption for customers with low and high incomes. Noble et al. (2017) state that customers with lower incomes purchase fewer products during a shopping trip which disguises their likelihood of coupon redemption. The data used only included the number of products bought with coupons but not the total items bought during the shopping trip.

Therefore, further research can address the mentioned limitation and provide more insights regarding coupon redemption behaviour to make the research more complete. Furthermore, due to the increasing use of technology and the rise in online coupons, research only focusing on E-coupons could provide different insights regarding the key customer and coupon characteristics for coupon redemption. As well as establishing this research on real consumption data.

By conducting similar research on digital coupons, different insights can be obtained and applied to online coupon campaigns. Therefore, increasing the obtained insights for different types of coupons and markets will help to develop more accurate and effective managerial actions to increase coupon redemption since customers behave differently in various markets.

7. Conclusion

The use of coupons is still prevalent and used in many markets. In the consumer goods industry, coupons are used to promote the products for which the coupon is valid. However, despite the billions worth of issued coupons, the number of redeemed coupons is relatively low, approximately 2.5 per cent in 2015 in the consumer goods industry. Therefore, to enact efficient and effective coupon redemption campaigns, managers and marketers need to know the key determinants of coupon redemption to identify their coupon-prone customers.

This coupon proneness depends on a combination of the attractiveness of the coupon and the features of the customer. Therefore, the objective of this paper is to identify the key customer and coupon characteristics that determine coupon redemption and to analyse the relationship of these characteristics with coupon redemption. In order to establish this research, fictional consumption data is used, and the applied classification algorithms are the Logistic regression and the Random Forest method.

According to the findings in this paper, managers and marketers should focus the coupon distribution on customers with low and middle incomes. Moreover, customers that are middle-aged show to be more prone to the use of coupons as well as customers that have big households including children.

Regarding the coupon characteristics, coupons for products that are included in the Food category have a higher possibility of being redeemed than coupons that are valid for products in other categories. Moreover, managers and marketers should focus their coupon campaigns on products of established brands to make the campaign more effective since the likelihood of coupon redemption is higher for products from established brands than from local store brands.

To make the research usable for more markets, conducting similar research with respect to digital coupons could provide more and different insights regarding online coupon campaigns. For the reason that customers behave differently in different markets as well as their behaviour online. Moreover, including a variable accounting for the number of items bought during the shopping trip could help identify coupon prone customers.

References

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29. https://doi.org/10.1145/1007730.1007735

Bawa, K., & Shoemaker, R. W. (1987). The coupon-prone consumer: some findings based on purchase behavior across product classes. *Journal of marketing*, *51*(4), 99-110. https://doi.org/10.1177/002224298705100409

Bawa, K., & Shoemaker, R. W. (1987). The effects of a direct mail coupon on brand choice behavior. *Journal of Marketing Research*, *24*(4), 370-376. https://doi.org/10.1177/002224378702400404

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Blattberg, R. C., & Neslin, S. A. (1990). *Sales Promotion: Concepts, Methods and Strategies*. Prentice Hall. https://doi.org/10.1016/S0927-0507(05)80035-0

Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 493-507. https://doi-org.eur.idm.oclc.org/10.1002/widm.1072

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
https://doi.org/10.1613/jair.953

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*(1-12), 24.
https://statistics.berkeley.edu/tech-reports/666

Chiou-Wei, S. Z. (2004). The determinants of direct mail coupon usage revisited: Evidence from count panel data models. *Agribusiness: An International Journal*, *20*(2), 189-200.
https://doi-org.eur.idm.oclc.org/10.1002/agr.20007

Chiou-Wei, S. Z., & Inman, J. J. (2008). Do shoppers like electronic coupons? A panel data analysis. *Journal of Retailing*, *84*(3), 297-307.
https://doi.org/10.1016/j.jretai.2008.07.003

Cronovich, R., Daneshvary, R., & Schwer, R. K. (1997). The determinants of coupon usage. *Applied Economics*, *29*(12), 1631-1641.
https://doi-org.eur.idm.oclc.org/10.1080/00036849700000039

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology, 88*(11), 2783–2792.
https://doi.org/10.1890/07-0539.1

Danaher, P. J., Smith, M. S., Ranasinghe, K., & Danaher, T. S. (2015). Where, when, and how long: Factors that influence the redemption of mobile phone coupons. *Journal of Marketing Research*, *52*(5), 710-725.
https://doi-org.eur.idm.oclc.org/10.1509/jmr.13.0341

Du, P., Samat, A., Waske, B., Liu, S., & Li, Z. (2015). Random forest and rotation forest for

fully polarized SAR image classification using polarimetric and spatial features. ISPRS Journal of Photogrammetry and Remote Sensing, 105, 38-53. https://doi.org/10.1016/j.isprsjprs.2015.03.002

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178-183. https://doi.org/10.2307/2290467

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232. https://doi.org/10.1214/aos/1013203451

Ghosh, A., Fassnacht, F. E., Joshi, P., & Koch, B. (2014). A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation, 26*, 49–63. https://doi.org/10.1016/j.jag.2013.05.017

Goh, K. H., & Bockstedt, J. C. (2013). The Framing Effects of Multipart Pricing on Consumer Purchasing Behavior of Customized Information Good Bundles. *Information Systems Research*, *24*(2), 334–351. https://doi.org/10.1287/isre.1120.0428

Goodwin, B. K. (1992). An Analysis of Factors Associated with Consumers' Use of Grocery Coupons. *Journal of Agricultural and Resource Economics*, *17*(1), 110–120. http://www.jstor.org/stable/40986744

Greenwell, B. M. (2017). pdp: an R Package for constructing partial dependence plots. *R J.*, *9*(1), 421. https://doi.org/10.32614/RJ-2017-016

Hamdiui, N., Stein, M. L., Timen, A., Timmermans, D., Wong, A., van den Muijsenbergh, M.

E., & van Steenbergen, J. E. (2018). Hepatitis B in Moroccan-Dutch: a quantitative study into determinants of screening participation. *BMC medicine*, *16*(1), 1-11 https://doi.org/10.1186/s12916-018-1034-6

Henderson, C. M. (1985). Modeling the coupon redemption decision. *ACR North American Advances*. https://www.acrwebsite.org/volumes/6374

Inmar (2016), "Promotion Industry Trends: A Year in Review," http://go.inmar.com/rs/134-NXN-082/images/Inmar-Promotion-Industry-Trends-A-Year-In Review.pdf.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer. https://doi.org/10.1007/978-1-0716-1418-1

Jung, K., & Lee, B. Y. (2010). Online vs. offline coupon redemption behaviors. *International Business & Economics Research Journal (IBER)*, *9*(12). https://doi.org/10.19030/iber.v9i12.345

Kanojia, M (2020). Coupon redemption data set. *Kaggle datasets download -d meghakanojia/predicting-coupon-redemption.* Available at: https://www.kaggle.com/datasets/meghakanojia/predicting-coupon-redemption

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review, 1(3), 9. https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Kulkarni, V. Y., & Sinha, P. K. (2012). Pruning of random forest classifiers: A survey and future directions. In 2012 International Conference on Data Science & Engineering (ICDSE) (pp. 64-68). IEEE. https://doi.org/10.1109/icdse.2012.6282329

Lalwani, A. K., & Wang, J. J. (2019). How do consumers' cultural backgrounds and values

influence their coupon proneness? A multimethod investigation. *Journal of Consumer Research*, *45*(5), 1037-1050. https://doi.org/10.1093/jcr/ucy033

Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). *Remote Sensing of Environment*, *100*(3), 356-362. https://doi.org/10.1016/j.rse.2005.10.014

Lee, J. Y., & Brown, M. G. (1985). Coupon redemption and the demand for frozen concentrated orange juice: A switching regression analysis. *American Journal of Agricultural Economics*, *67*(3), 647-653. https://doi-org.eur.idm.oclc.org/10.2307/1241088a

Leone, R. P., & Srinivasan, S. S. (1996). Coupon face value: Its impact on coupon redemptions, brand sales, and brand profitability. *Journal of retailing*, *72*(3), 273-289. https://doi.org/10.1016/S0022-4359(96)90030-5

Maalouf, M. (2011). Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281-299 https://doi.org/10.1504/IJDATS.2011.041335

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315-1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage. Available at: https://books.google.it/books?id=EAI1QmUUsbUC&lpg=PP7&ots=4VGMM-mQHN&dq=logistic%20regression&lr&hl=nl&pg=PP9#v=onepage&q=logistic%20regression&f=false

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced

data. *Data mining and knowledge discovery*, *28*(1), 92-122. https://doi.org/10.1007/s10618-012-0295-5

Narasimhan, C. (1984). A price discrimination theory of coupons. *Marketing Science*, *3*(2), 128-147. https://doi-org.eur.idm.oclc.org/10.1287/mksc.3.2.128

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics*, *11*(1), 1-13. https://doi-org/10.1186/1471-2105-11-110

Nielsen Jr, A. C. (1965). The impact of retail coupons. *Journal of Marketing*, *29*(4), 11-15. https://doi.org/10.1177/002224296502900403

Noble, S. M., Lee, K. B., Zaretzki, R., & Autry, C. (2017). Coupon clipping by impoverished consumers: Linking demographics, basket size, and coupon redemption rates. *International Journal of Research in Marketing*, *34*(2), 553-571. https://doi.org/10.1016/j.ijresmar.2016.08.010

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, *26*(1), 217-222. https://doi.org/10.1080/01431160412331269698

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*(364), 699-705 https://doi.org/10.1080/01621459.1978.10480080

PUSZTOVÁ, Ľ., & Babic, F. (2020). Performance Assessment of Different Classification Methods for Coupon Marketing in E-Commerce. *Acta Electrotechnica et Informatica*, *20*(3), 11-16. https://doi.org/10.15546/aeei-2020-0014

Reibstein, D. J., & Traver, P. A. (1982). Factors affecting coupon redemption rates. *Journal*

of Marketing, *46*(4), 102-113. https://doi.org/10.1177/002224298204600411

Ren, X., Cao, J., & Xu, X. (2021). A two-stage model for forecasting consumers' intention to purchase with e-coupons. Journal of Retailing and Consumer Services, 59, 102289. https://doi.org/10.1016/j.jretconser.2020.102289

Rhee, H., & Bell, D. R. (2002). The inter-store mobility of supermarket shoppers. *Journal of Retailing*, *78*(4), 225-237. https://doi.org/10.1016/S0022-4359(02)00099-4

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674 https://doi-org.eur.idm.oclc.org/10.1109/21.97458

Schiller, B. R. (1997). *The Economics of Poverty and Discrimination* (7th edition). Pearson Education (Us). Available at: https://agris.fao.org/agrissearch/search.do?recordID=US201300491113

Schultz, D. (1998). *Sales Promotion Essentials: The 10 Basic Sales Promotion Techniques… and How to Use Them* (3rd edition). McGraw-Hill. Available at: https://www.worldcat.org/title/sales-promotion-essentials-the-10-basic-sales-promotion-techniques-and-how-to-use-them/oclc/318281527?referer=di&ht=edition

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic Emergency Medicine*, *18*(10), 1099-1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130. https://dx-doi-org.eur.idm.oclc.org/10.11919%2Fj.issn.1002-0829.215044

Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1),

12-18.
http://dx.doi.org/10.11613/BM.2014.003

Swaminathan, S., & Bawa, K. (2005). Category-specific coupon proneness: The impact of individual characteristics and category-specific variables. *Journal of Retailing*, *81*(3), 205-214
https://doi.org/10.1016/j.jretai.2005.07.004

Tuttle, B. (2010). The History of Coupons. Available at:
http://business.time.com/2010/04/06/the-history-of-coupons/

Ward, R. W., & Davis, J. E. (1978). A pooled cross-section time series model of coupon promotions. *American Journal of Agricultural Economics*, *60*(3), 393-401.
https://doi-org.eur.idm.oclc.org/10.2307/1239936

Webster Jr, F. E. (1965). The "deal-prone" consumer. *Journal of Marketing Research*, *2*(2), 186-189.
https://doi.org/10.1177/002224376500200209

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445-5449.
https://doi.org/10.1016/j.eswa.2008.06.121

Appendices

Appendix A: Descriptive statistics

Variable	Observations
redemption_status:	
No	676,835
Yes	63,081
brand_type:	
Established	624,811
Local	115,105
category:	
Food	506,361
Nonfood	33,205
Services	350
age_range:	
18-25	24,241
26-35	124,693
36-45	204,584
46-55	215,791
56-70	84,874
70+	85,733
marital_status	
Marries	578,152
Single	161,764
rented	
No	683,844
Yes	56,072
family_size	
1	97,951
2	368,425
3	106,855
4	70,520
5	96,165
no_of_children	
0	455,518
1	114,603
2	68,592
3	101,203

income_bracket	
low-class	303,666
middle-class	329,635
high-class	106,615

Appendix B: Variable importance plot - Gini

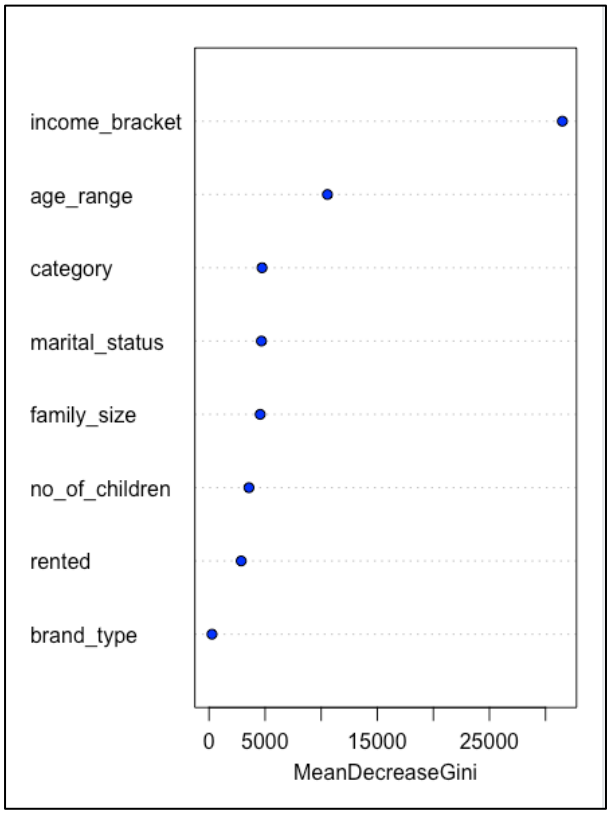

Figure: Variable importance by mean decrease in Gini

Appendix C: Partial dependence plots

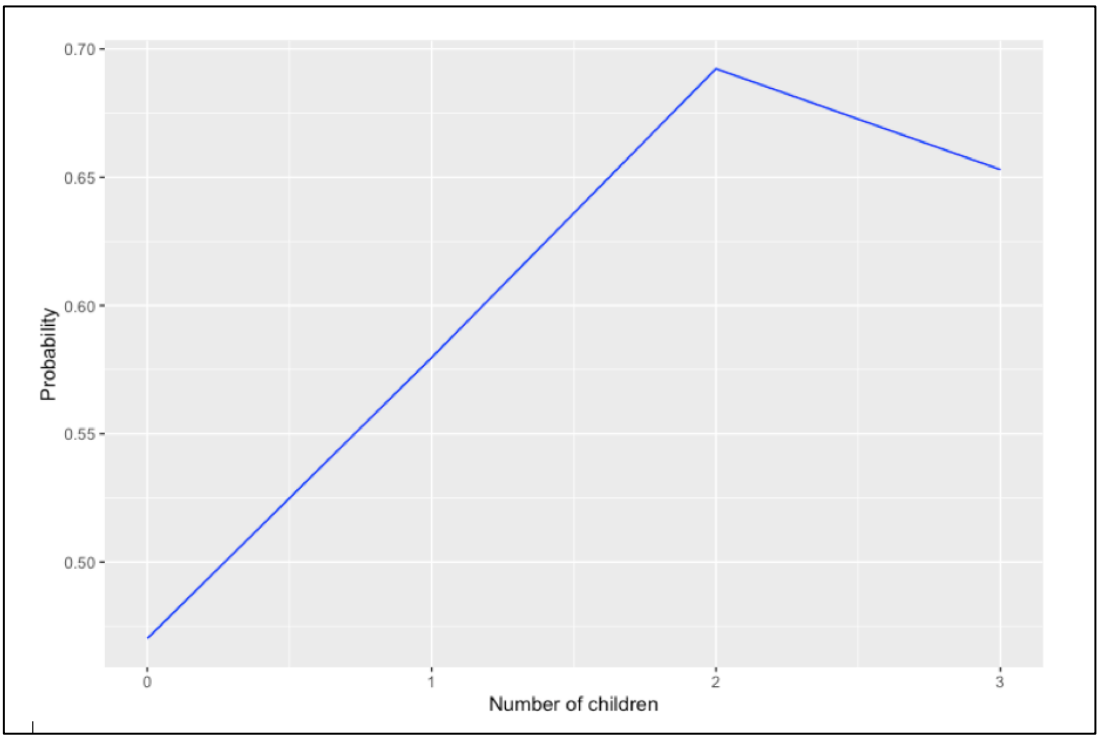

Figure: Partial dependence plot on the predictor no_of_children

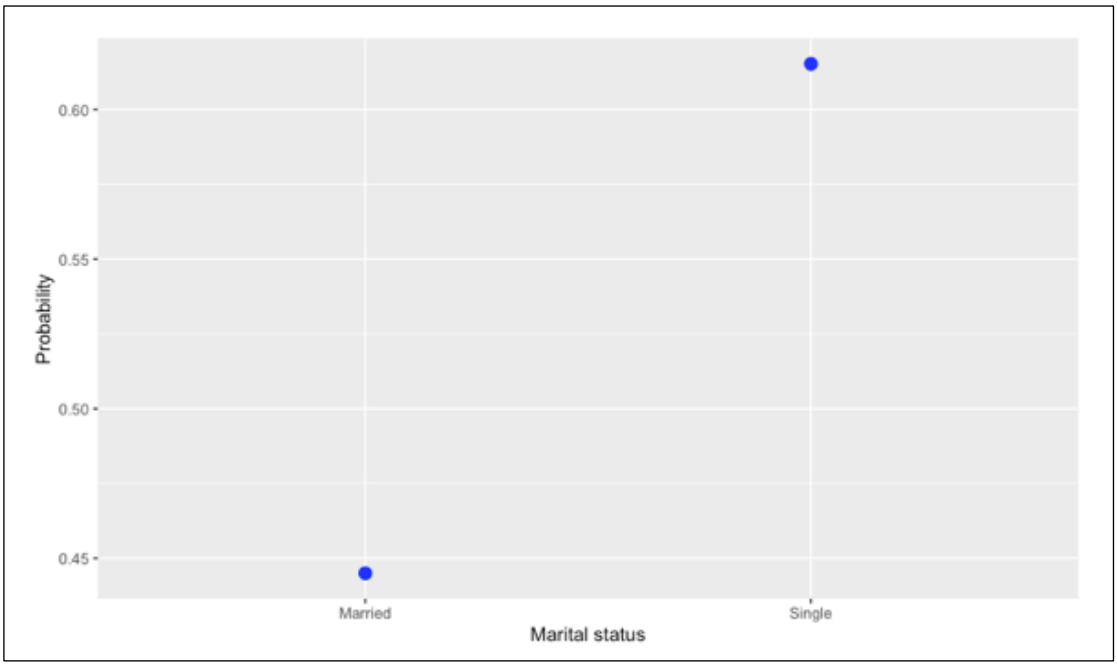

Figure: Partial dependence plot on the predictor marital_status

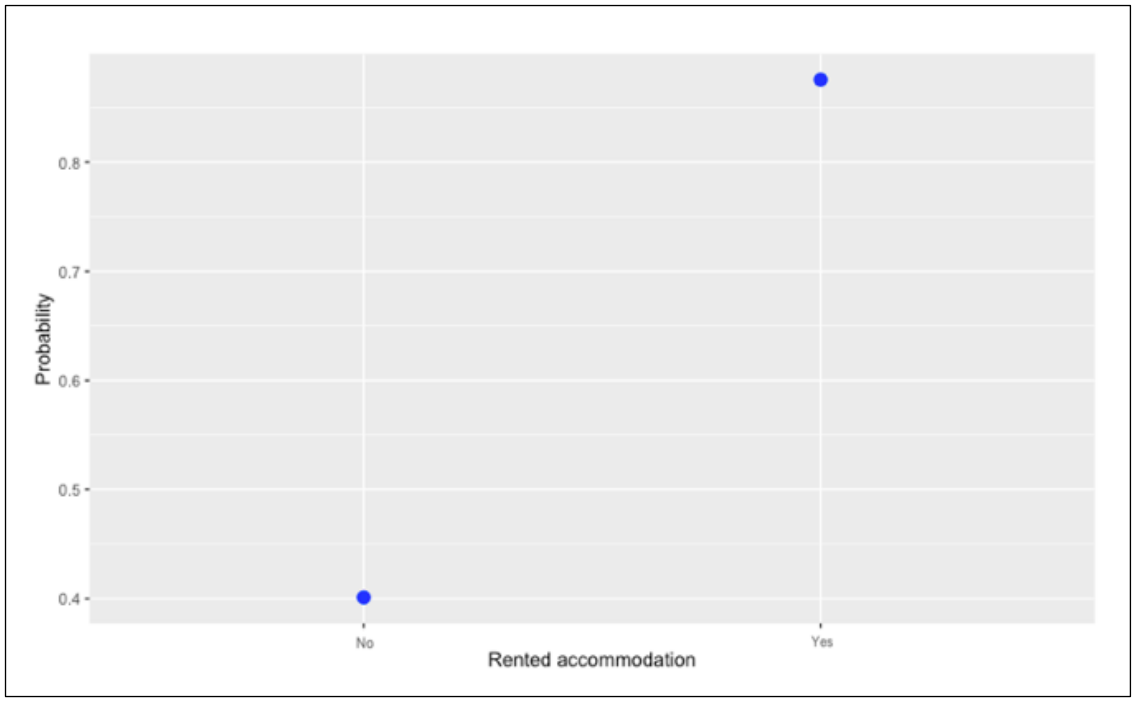

Figure: Partial dependence plot on the predictor rented

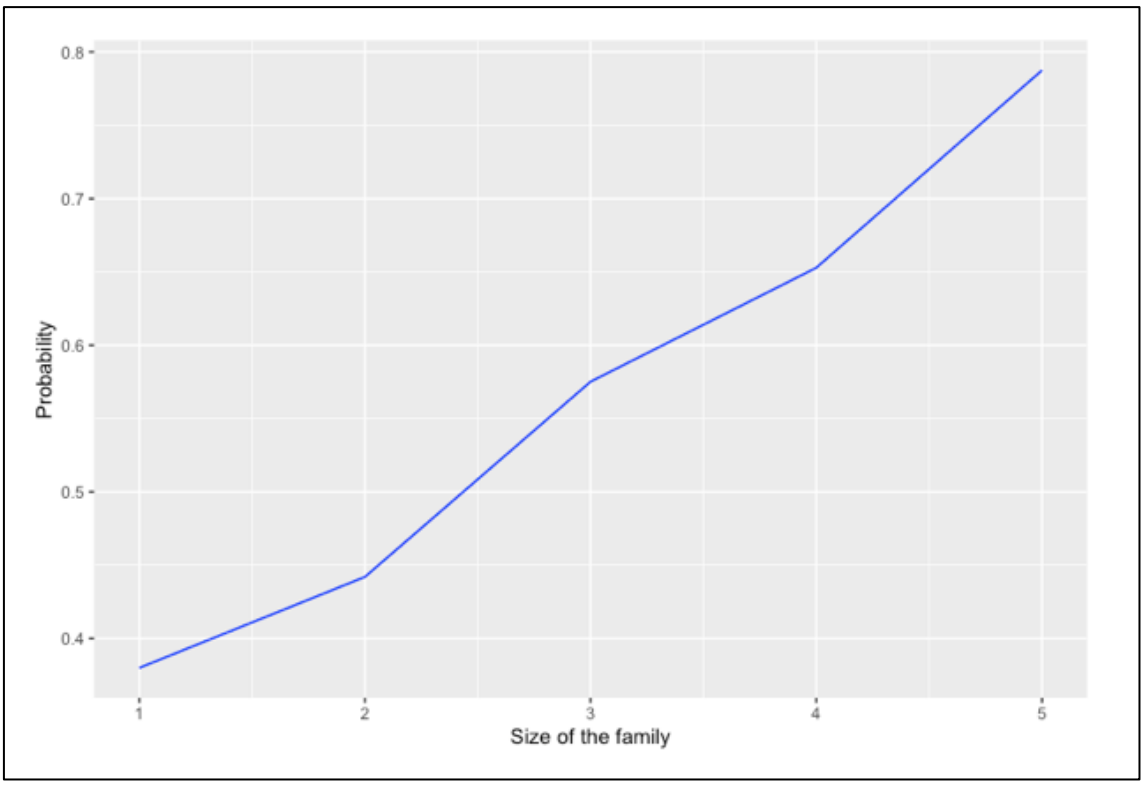

Figure: Partial dependence plot on the predictor family_size

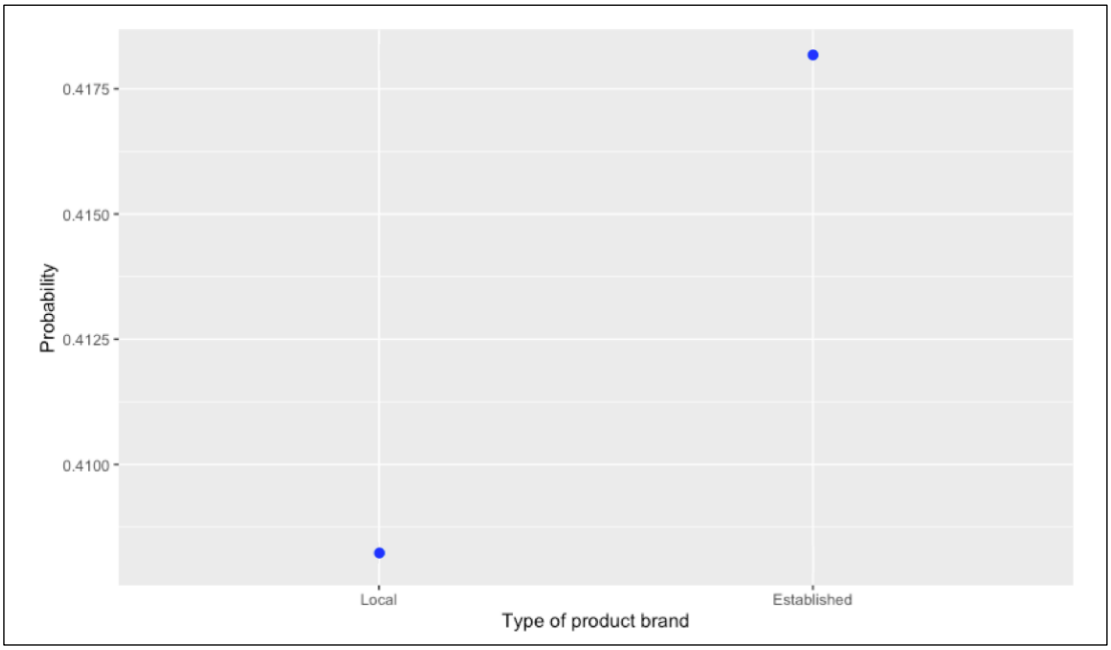

Figure: Partial dependence plot on the predictor brand_type