Coupon Campaign Analysis

Team 61

Gene Fountain, Orlando Gonzalez,
Tad Kabage, Tyler Uher & Jake Vance

Coupon Campaign Analysis

Contents

Coupon Campaign Analysis

Problem Statement

Consumer-packaged goods companies (CPGs) and the grocery retailer would like to evaluate whether the coupon campaign is worth a CPG's investment and determine how to optimize future coupon campaign design.

Literature Surveys

The following source titled "The Business of Coupons-Do coupons lead to repeat purchases?" seeks to answer the question of whether a coupon is an effective tool in increasing the repeat purchases of a product. In order to answer this question, a sample of 2,500 households (grocery store patrons) and their transactions over a two-year period were analyzed. In their approach the author chose to focus on 12/15 pack canned carbonated drinks. Below are some key insights derived from this paper.

- Household ages between 45 to 54 had the most transactions using coupons, yet a larger percentage of households between 55 to 64 used coupons (86.17% vs 75.20%).
- Household ages between 45 to 54 made up most of the sample. Most non-coupon transactions come from this age group as well.
- Households making less than $15k and between $100k to $124k yearly used coupons in over 99% of their transactions.
- Over 99% of all transactions using coupons were from homeowners.
- Households tend to stock up on these products when prices are lowered due to the coupon.
- There is no sustained after-effect of a coupon campaign since quantity sold returns to a similar amount as before the coupon was issued.

The following source titled "An Exploratory Investigation into How Socioeconomic Attributes Influence Coupons Redeeming Intentions" seeks to analyze how coupon redemption rates can be influenced by socioeconomic attributes. The sample data is taken from a survey of 2250 respondents. The grocery products were the focal product category chosen for this study.

- The perceived income effect (the perception of saving money) is a significant factor in increased coupon redemption rates.
- The savings from the coupons can be used to incentivize the purchase of other products.
- The opportunity cost of time and effort spent researching and planning towards the use of coupons is a significant factor in decreasing the coupon redemption rate.

The following source titled "Measuring and Managing Returns from Retailer-Customized Coupon Campaigns" investigates the effect of retailer-customized coupon campaigns on trip incidence and revenues. The study looked at purchase histories of 2,500 households that were members of a group of regional grocery chains' shopper card programs for a two-year period. A total of 40 customized coupon campaigns were mailed.

- An analysis of variance test showed that the percentage increase in weekly trip revenue per customer in the test group (24%) was significantly higher than the percentage increase in the weekly trip revenue per customer in the control group (11%).

Coupon Campaign Analysis

Background Information

This dataset, which comes from the "completejourney" R package, contains household-level transactions over the span of one year from a group of 2,469 households who are frequent shoppers at a grocery store.

Consumer-packaged goods companies (CPGs), such as Coca-Cola, Kraft, PepsiCo, etc., have the goal of increasing their "share of stomach" at the grocery store, and one way to acquire/retain new/existing households to their brand is by launching a coupon campaign at the grocery store. The grocery store has historical data such as purchase data, historical coupon campaign data, demographic data, and other promotional data.

Anticipated Conclusions/Hypothesis

We hypothesize that the coupon campaign will cause an incremental sales uplift for the test households versus the control households. With our distance minimizing algorithm, we matched test households to their nearest control counterparts to then build a difference-in-difference model to understand whether the coupon campaign indeed supports our hypothesis.

We hypothesize that having a family, being more price-sensitive (lower income), or being both young and having a low-income could increase one's likelihood of redeeming a coupon. We also hypothesize our most loyal customers will be more likely to redeem a coupon. Our logistic regression approach will allow us to isolate each of these variables to understand how much each affects the probability of redeeming a coupon or not.

We also hypothesize that the random forest model will yield the highest classification accuracy in predicting whether the household will redeem a coupon or not due to its ability to capture interaction effects and capture nonlinear patterns in the data. Our confusion matrix output from the logistic regression model, support vector machine, and random forest will help us determine what the best model is for management to use to target households for future coupon campaigns.

Approach

Figure 1 below summarizes the approach we took analyzing 17 coupon campaigns (campaign IDs 2-18) that provided sufficient pre-period and post-period data for analysis.

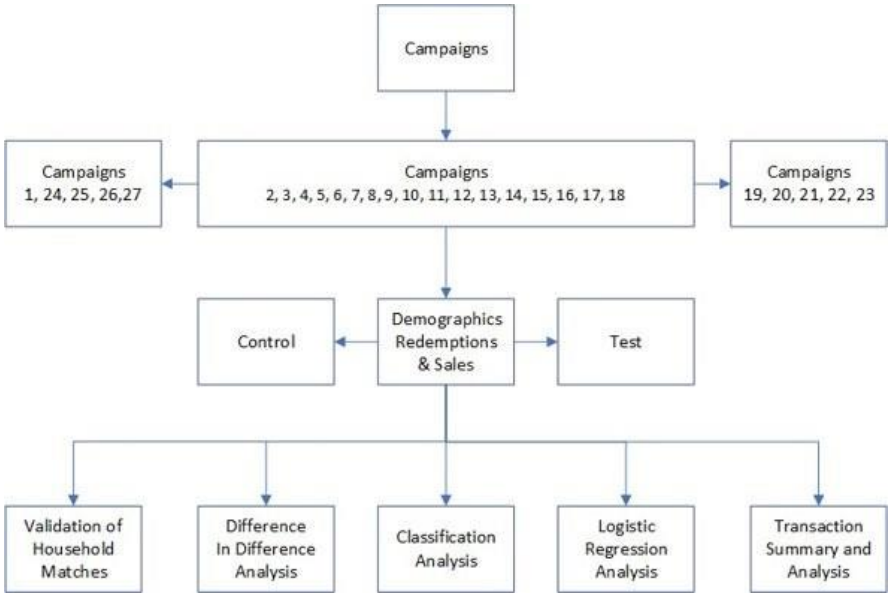

Figure 1: The Approach

We cleaned and engineered features about each household in the 7 weeks *before* it received the coupon campaign, *during* the coupon campaign window, and the 7 weeks *after* the coupon campaign ended. Features included the household's net sales (gross sales – coupon discounts) and total number of visits at the grocery store, the coupon's product categories (e.g., "APPLES"), and the coupon's products (e.g., "APPLES GOLD DELICIOUS (BULK & BAG)"), demographic information such as the household's estimated age range, income, homeownership status, marital status, household size, household composition, and number of kids in the house, and whether that household has redeemed a coupon before.

Then, using a simple matching algorithm, we matched "test" households to "control" households. This algorithm calculated the test household's Euclidean distance, after the pre-period numeric data has been scaled, from all the potential control households and selected the nearest potential control household to be matched to the test household. Of note, we only used numeric features due to matching households via minimizing Euclidean distance, and we also believe that using the pre-period sales, grocery store visits, and previous coupon campaign redemption information alone would allow us to identify reasonable matches for our test households. After each test household was matched to its nearest control household, we removed the control household from the eligible control household dataset to prevent control households from being overused. Also, to validate household matches, we built a logistic regression model to see if using pre-period numeric features and demographic features whether we could predict whether a household was a test or control household.

After matching test households to control households, we built 3 difference-in-difference models to measure if the coupon campaign caused the household to purchase more on the coupon campaign's products, categories, and/or at the total grocery store. First, we calculate net sales in the pre-period, which is 7 weeks before the coupon campaign for both test and control households. Then, we measured sales the test and control households for the campaign duration, plus 7 weeks to capture longer-term effects. The goal is to find that the coupon campaign caused an "uplift" in sales from the test vs. control households, as this would indicate to the CPG and grocery store that this coupon campaign design is

favorable to future investment. Moreover, we built a built a paired t-test on these pre-period and post-period differences between the test and control groups for each campaign to understand if there was a statistically significant difference in sales between the test and control households at the 95% confidence level.

Additionally, we built a series of logistic regression models with demographics and the pre-period engineered numeric features to predict whether that test household redeemed the coupon or not. Output from this logistic regression model could help inform a CPG/advertising agency with some specific characteristics upon which to target households for future campaigns. With this output, we can determine if there are certain characteristics that make a test household more likely to redeem a coupon or not.

In addition to the logistic regression model, we explored building other classification models, such as a support-vector machine and random forest model to predict whether a test household will redeem a coupon or not. First, we split our test households into an 80 % training dataset and 20 % validation dataset. Of note, we used a random stratified split to make sure that the proportion of redeeming households was similar in the training and validation data, which was important given that coupon redemptions were low across each campaign. To compare our models, we produced confusion matrices that showed how effective our models were at predicting whether a validation household redeemed a coupon or not. We also focused on the sensitivity metric, as we believe that correctly finding all redeeming households is very important from a business perspective. High classification accuracy can be misleading when coupon redemption is already very low, so sensitivity stood out as the metric to focus on when comparing models. With our "best" model, management will be able to use this model to target the best set of households for future coupon campaigns.

Campaign Size

Our coupon redemption data was collected from point-of-sale transaction data collected at each store location. A total of 2,469 households participated in the study and household demographic data was collected for each household. Table 1 shows the number of households included in each campaign.

Campaign ID	Targeted Households
campaign-02	48
campaign-03	12
campaign-04	81
campaign-05	166
campaign-06	65
campaign-07	197
campaign-08	1,075
campaign-09	176
campaign-10	123
campaign-11	214
campaign-12	169
campaign-13	1,075

campaign-14	224
campaign-15	17
campaign-16	188
campaign-17	202
campaign-18	1,125

Table 1: Targeted Households by Campaign

We used a stratified random sampling split with 80% in train and 20% in test. Ultimately, the small campaign size negatively impacted model performance, which will be explained further in the classification models. Our recommendation for future campaigns is to set the minimum campaign size to 1000 households before actionable business insights can be made.

Household Matching Validation with Logistic Regression

Intuitively, if we have effectively matched test households with control households for each campaign, then it should be hard to distinguish between the test and control households. Therefore, for each campaign, we built a logistic regression model to predict whether a household was a test household or control household using pre-period numeric features and demographic data. If the household matching were effective, we should find no statistically significant variables that make a household more likely to be a test household. With household matching, the goal is to make sure that our test households and control households look identical, except for the test household receiving the treatment (being targeted with the coupon). Table 2 below summarizes the results of household matching by campaign.

Campaign ID	Bad	Okay	Good
campaign-02			X
campaign-03			X
campaign-04			X
campaign-05		X	
campaign-06		X	
campaign-07			X
campaign-08	X		
campaign-09		X	
campaign-10	X		
campaign-11		X	
campaign-12			X
campaign-13		X	
campaign-14		X	
campaign-15			X
campaign-16	X		
campaign-17		X	
campaign-18		X	

Table 2: Household Matching Validation Results

We defined a “bad” match as having 4+ statistically significant predictors, which happened to 3/17 campaigns, an “okay” match as having 1-3 statistically significant predictors, which happened to 8/17 campaigns, and a “good” match as having 0 statistically significant predictors, which happened to 6

campaigns. Given that most household matches were "okay" or "good," we would grade ourselves as "decently" matching test and control household. For future work, we should consider reusing control households (rather than only using each control household once) to find better matches for each test household. Additionally, we should explore incorporating demographic features and more granular engineering for each household when matching, such as a household's sales with specific products and/or brands.

Difference-in-Difference Model

With the use of the difference-in-difference model, following test and control household matching, we were able to conclude that there does seem to exist an uplift in total net sales in the total grocery store after the issuance of the coupon. This uplift is equivalent to $115 on average across all campaigns. However, we were not able to see an uplift for total coupon category sales (-$0.45) nor for total coupon product sales (-$0.05).

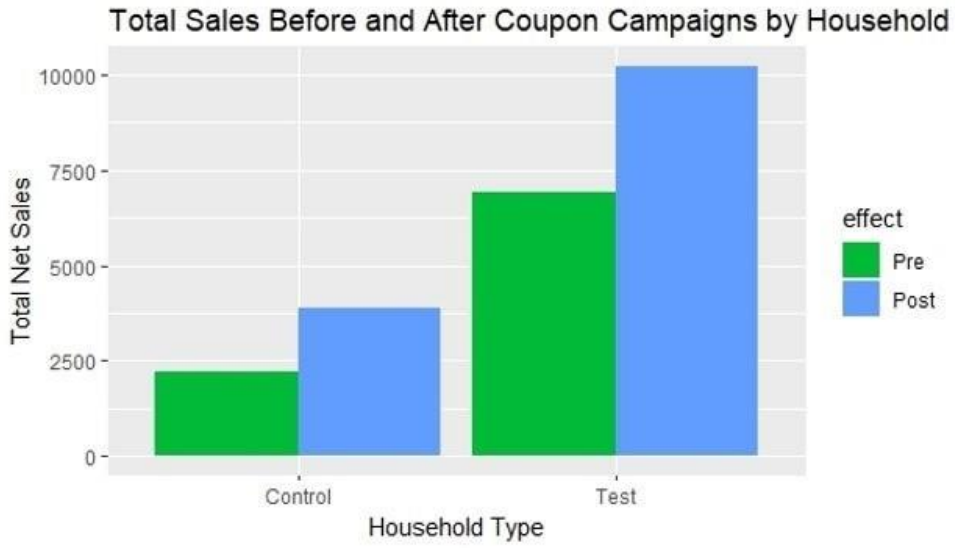

Figure 2: Total Net Sales at the Grocery Store Pre- and Post-Coupon Campaign

Overall, the coupon campaign does not appear to drive sales for specific coupons and products, but it does appear to drive sales across the grocery store. Therefore, based on this difference-in-difference model, the grocery store should be more confident that any potential losses on the coupon's products itself can be more than offset by increased purchases across the grocery store.

Moreover, in order to validate the difference-in-difference model, we conducted a paired t-test and concluded that there was a statistically significant difference in total net sales between test and control households for 3 out of the 17 campaigns. There was a statistically significant difference in sales between the test and control households from pre-to-post period only for campaigns 8 (+$183.14 per household), 13 (+$204.25 per household), and 18 (+$31.07 per household) in the difference-in-difference model for average net sales. For the coupon category and the coupon product net sales models, we did not find any statistically significant campaigns based on the paired t-test.

Logistic Regression

For logistic regression, we analyzed a total of twelve predictors of coupon redemption. Of the twelve predictors we identified three significant features that increased a household's likelihood of coupon redemption:

- Higher pre-campaign period sales
- Higher previous coupon usage
- Homeownership (best predictor)

We ran 3 series of logistic regression analyses and fitted a total of 19 logistic regression models.

- Series 1 – Fitted one model per campaign (17) using all predictors
- Series 2 – Fitted a single model using all predictors
- Series 3 – Fitted a single model using the significant (3) predictors

We summarized model performance using a confusion matrix and highlighted the distribution of sensitivity in Figure 2 below. Sensitivity, in this model referred to the rate at which the household would redeem the coupon given that the model predicted they would.

As a reminder, due to low coupon redemption rates, understanding how performant the model was at identifying all households that would eventually redeem the coupon was most important from a business perspective.

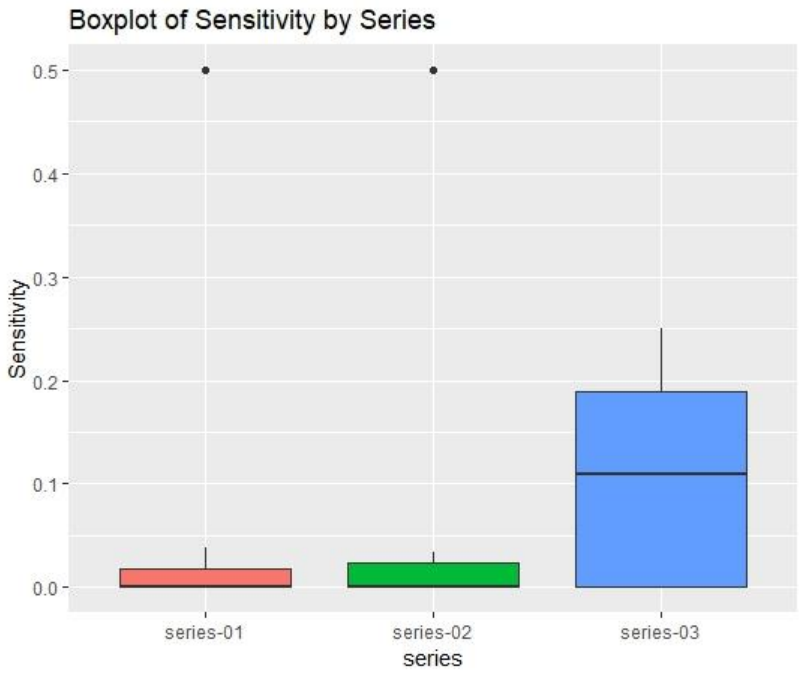

Figure 3: Boxplot of Sensitivity by Logistic Regression Model

As result of this analysis, we identified the following key takeaways:

- Collect more household data both pre and post campaign
- The best sensitivity was achieved with the larger sample sizes
- Minimum campaign size should be 1000 households to achieve a reasonable model fit

Other Classification Methods: Support Vector Machine and Random Forest

Using the features that were determined to be significant by our logistic regression models, we ran both a random forest and a support vector machine model. These features included: pre-period net sales, pre-period visits to the grocery store, marital status, household size, household composition, and the number of kids in the household. We were consistent with our datasets using the same training and validation datasets for each campaign's model, across all three methods. We believe that the random forest model is the best model for the grocery store to use for future coupon campaign targeting because it is very versatile, quick and is an ensemble method, so it generally produces a higher accuracy and minimizes variance in predictions. The SVM model was chosen because the data had a very high dimensionality especially considering the number of rows. There was also very little noise within each campaign, which is great for the distance calculations used within a support vector machine model. For both methods we created a model for all 17 campaigns and calculated accuracy and sensitivity. The downfall of these two models is that they aren't easy to interpret so we were stuck looking at the end performance only. This is acceptable though because overall performance is what we cared about most: how well the model could predict whether a household would redeem a coupon so the grocery store and CPG would know whom to target coupon with.

We had similar issues with these models when it came to the limited row counts of a few campaigns causing a lot of overfitting. In general, this resulted in a very high accuracy for every campaign, but many sensitivity calculations were undefined as the model would simply predict all the coupons to not be redeemed. This is because a lot of the campaigns we used for modeling consisted of a very little number of redemptions. We did pursue some hyperparameter tuning to eliminate some of the model biases that I just explained above, but it was impossible to eliminate the bias out of some campaigns as their composition consisted of an extremely small redemption rate.

Comparing Classification Methods

We compared our models by looking at both the aggregate accuracy and sensitivity of all the models combined. We looked at the aggregates because the campaigns had a huge variance in redemptions leading to a unique model for each campaign. We wanted to know what method would perform best on any new campaign that we would, in the future, put into action. Therefore, an aggregate of accuracy and sensitivity would be best. We not only looked at the aggregates but also the distribution of both the accuracy and sensitivity which can be seen below in our boxplots.

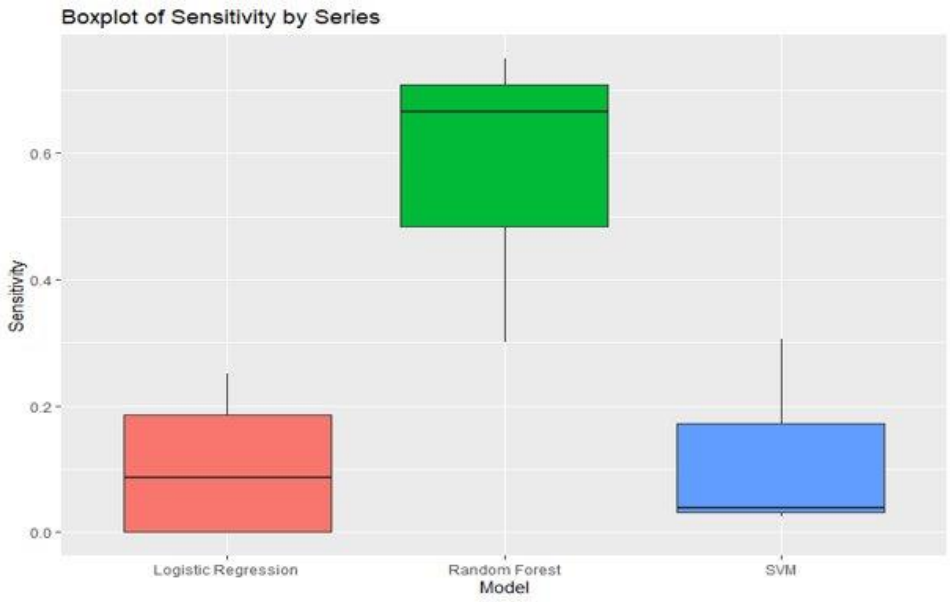

Figure 4: Sensitivity Distribution by Model Type

As you can see from the boxplots, all the models produced a very high accuracy for almost every campaign. This is in large part due to overfitting as many of the campaigns had very few rows. The campaign size can be seen earlier in the Campaign Size section of the report. Although accuracy supports our decision for the most effective model, it's not the key statistic for this analysis. Sensitivity, clarified in the previous Other Classification Models section, is the most important statistic for comparing these models, and the random forest model produced a higher sensitivity for every campaign. Therefore, we determined that moving forward the random forest method would be best for determining which households to target coupons with, as the grocery store and CPG are much better equipped with the random forest model to predict whether will redeem a coupon or not.

Conclusions

First, we initially hypothesized that the coupon campaign would cause an uplift in sales for the test households (those targeted with the coupon) versus the control households, and our analysis showed evidence in support of this hypothesis. Specifically, although our difference-in-difference model did not support an uplift in sales on the coupon's products and coupon category, our analysis did find that there was an increase in sales across the grocery store, which indicates that perhaps the coupon campaign's losses are more than offset by increased purchasing across the store. In other words, it might be the case that the issuance of coupons makes one more likely to visit the grocery store, thereby driving up total grocery store sales for the test households than the control households.

Second, we initially hypothesized that having a family, having a lower income, and being younger would increase one's likelihood to redeem a coupon. However, our logistic regression classification models suggest that higher pre-period net sales, higher previous coupon redemption, and homeownership most increase one's likelihood of coupon redemption. Therefore, when targeting households for future coupon campaigns, perhaps the grocery store and/or CPG should target their coupons toward households that tend to purchase more at the grocery store (higher loyalty), have a history of using coupons, and are estimated to be homeowners.

Third, we initially hypothesized that a random forest model would yield the highest classification accuracy in predicting coupon redemption, and our classification modeling results support this hypothesis. Specifically, the random forest far out-performed the logistic regression and SVM models in terms of sensitivity, or given that a household redeemed a coupon, what percent of coupon-redeeming households could the model identify? Given that coupon redemption was typically very low with the campaigns, such as less than 10% for most campaigns, it is very important for a grocery store and/or CPG to find those households that will redeem a coupon. Once a household tries a product for the first time, that household increases its likelihood of becoming loyal to the grocery store and/or brand. Therefore, a grocery store and/or CPG wants to make sure it prioritizes targeting households that appear likely to redeem a coupon and try the coupon's products.

Findings, Recommendations, and Future Work

After completion of our difference in difference analysis, our classification analysis, and our logistic regression analysis, we offer the following findings, recommendations, and guidance for future work:

Findings

- Campaigns with more than 1000 households increase model stability
- There is inconclusive evidence over whether coupons drive uplift for its redemption products and categories, but the difference-in-difference model suggests that total grocery store sales increase following coupon redemption.
- Homeownership is the best demographic predictor at increasing one's likelihood to redeem coupons. Higher pre-period sales and previous coupon usage also increase one's likelihood to redeem coupons.
- The best model for identifying coupon redemption is the random forest.

Recommendations

- Understand that issuing coupons can lead to statistically significant total grocery store sales uplift, even if there is not statistically significant sales uplift on the coupon's specific products.
- If possible, collect more pre and post campaign sales and redemption data across more households to allow for better household matching.
- Increase the size of each campaign – the minimum campaign size is 1000 households before reasonable models can be fit for more actionable insights.
- Use the random forest model to target households for coupon redemption.

Guidance for Future Work

- Include additional feature data to support modeling
    - Product placement, in-store promotions, and geographic data
- Emphasize feature engineering supported by additional data to improve
    - Evidence for sales uplift resulting from coupon campaigns
    - Household matching
- The face value of the coupon might have a significant effect on coupon redemption
    - Explore the relationship between coupon redemption and the coupon value
- Explore the effect of personalized vs. non-personalized coupons

Appendix

Works Cited

Barat, Somjit, et al. "An Exploratory Investigation into How Socioeconomic Attributes Influence Coupons Redeeming Intentions." *Journal of Retailing and Consumer Services*, Pergamon, 30 Jan. 2013, https://www.sciencedirect.com/science/article/pii/S0969698913000052.

Ross, Margaret P. "The Business of Coupons-Do Coupons Lead to Repeat Purchases?" *TRACE*, The University of Tennessee Libraries, June 2014, https://trace.tennessee.edu/pursuit/vol5/iss1/14/.

Venkatesan, R., & Farris, P. W. "Measuring and Managing Returns from Retailer-Customized Coupon Campaigns." *Journal of Marketing*, 76(1), 76–94, 1 Jan. 2012 https://doi.org/10.1509/jm.10.0162.