Purchase Analysis with SciKit Learn

Jennifer Boyce

Summary

This study utilized a variety of knowledge discovery tools to better understand grocery marketing and customer behavior.

Specifically, analysis focused on four questions:
- Can customer demographics predict a customer's transaction amount?
- What items do customers frequently purchase together?
- How can customers be segmented by similar characteristics?
- Is it possible to predict which households will redeem coupons?

To answer these questions, regression analysis, agglomerative hierarchical clustering, association rules mining, classification, and logistic regression were employed. Utilizing model tuning, the agglomerative hierarchical clustering model was optimal using Ward's linkage with 5 clusters. Association rule mining performed well with a support criterion of 0.0001 and lift value greater than one. Among classification algorithms, Bernoulli naïve Bayes with an alpha value of 0.4 and refitting provided the greatest accuracy—60%.

The findings of the report included the following:
- Customer household income was the greatest predictor of the amount a customer spent in a single transaction, followed by the number of children in the household.
- Regular milk and chocolate milk are the items that customers most frequently purchase together.
- In segmenting customers by similar demographic characteristics for marketing purposes, customers are best characterized by their marital status and number of children, followed by income and age.
- Coupon redemption is most likely among higher-income customers ($125,000 - $175,000) and those who rent their homes.

Data Overview

Data for this study encompassed two years of sales and marketing data for a regional grocery chain. Among the factors included were such elements as customer demographics, item-level transactions, coupon offers and redemptions, and product descriptions. A full data schema is available in Appendix A.

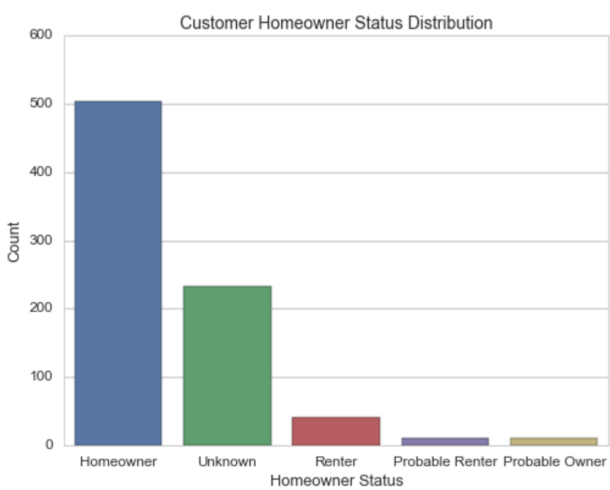
Transforming transaction data was the first step in data preparation. The stores represented in the data offered stand-alone gas stations on the store property. Gas transactions were not representative of customers' in-store shopping behaviors, and these transactions caused unnecessary noise in the models. Any line-item sales of gas were removed from the data before modeling. In the future, it might be interesting to analyze the full data set to understand the role and characteristics of gas shoppers.

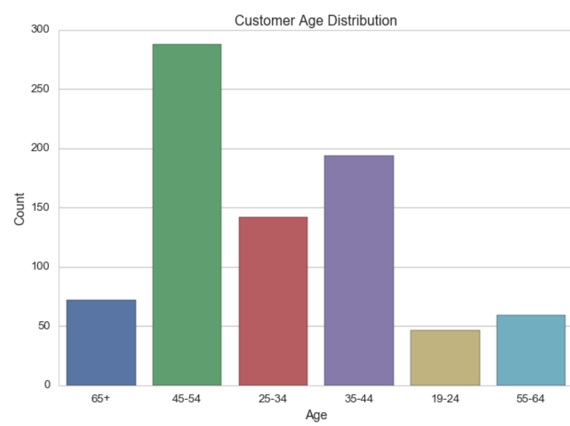
Demographic data were also coded into binary dummy variables, which were used for all techniques except market basket analysis. The original data provided pre-binned data for age and household income, so discretization was not needed.

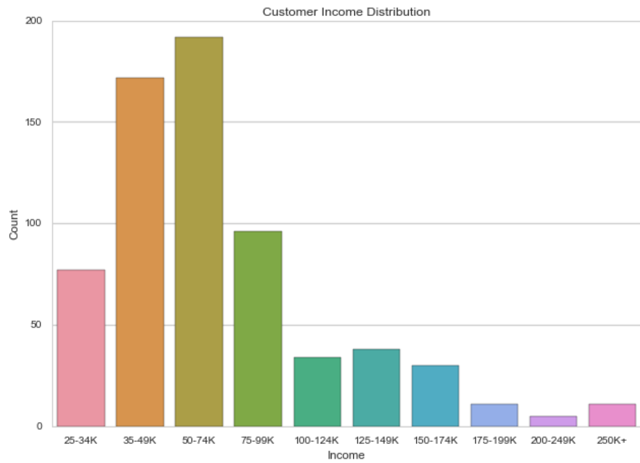
Customers

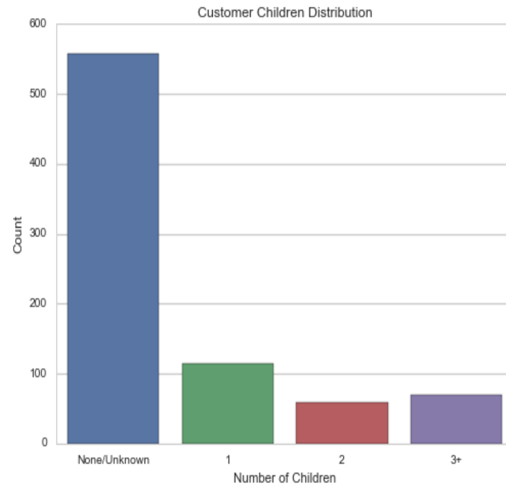
Although 2,500 households' shopping habits were included in the data set, demographic data was only available for 801 of these customers. Demographic variables included customer age, marital status, household income, homeowner status, number of adults/children in the home, and number of children.

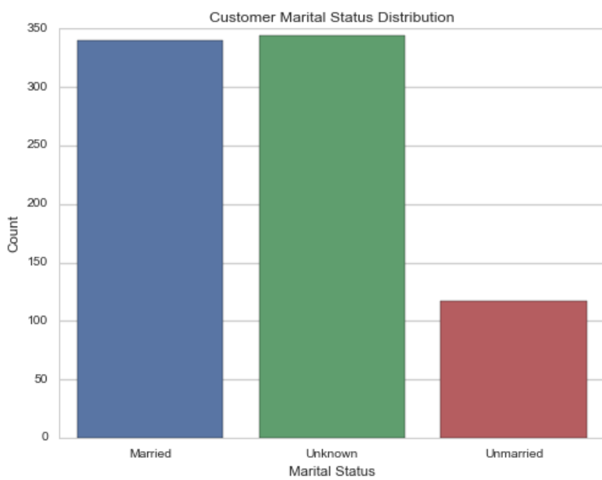
Customers were predominantly homeowners aged 45-54 with an income between $50,000 - $74,000 per year and an unknown number of children. Most customers were married or had an unknown marital status, and the majority of households contained two adults and two children.

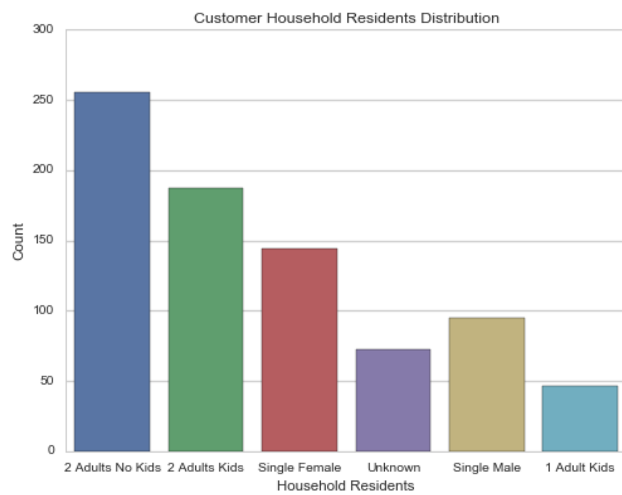






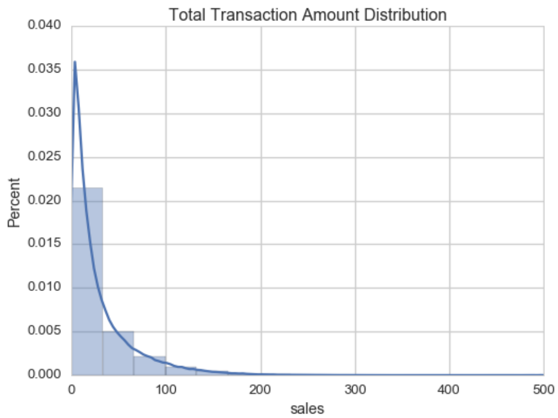


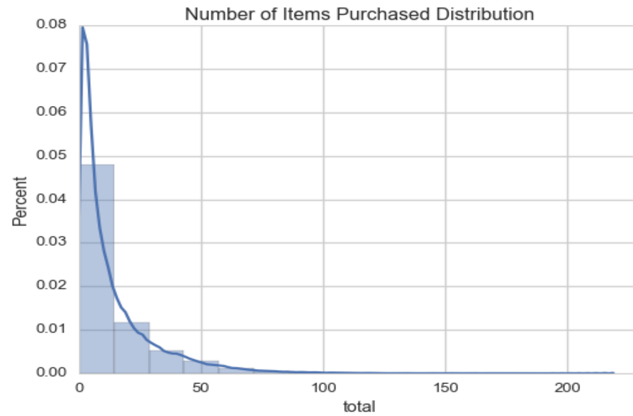




Transactions

Customers' transaction amounts varied substantially, ranging from $0 - $496.37, although the mean was $29.32. Among transactions, customers purchased between 1 and 215 unique items. The median number of items purchased per transaction was 17.





Coupons

In total, 124,811 coupons were available for 44,133 products. These coupons were mailed to 1,584 households. 2,318 coupons were redeemed by 434 households.

Demographics and Sales

The impact of individual customer demographics on total transaction amounts can be understood through linear regression. By creating a multivariate regression model, its possible to evaluate the overall effect of each factor.

While similar, various regression techniques provide slightly different results. To capitalize on this variability, several regression techniques were utilized: multiple regression, ridge regression, stochastic gradient descent (SGD), and lasso regression. Each was evaluated by root mean square error on a training/testing split, as well as in k-fold cross validation.

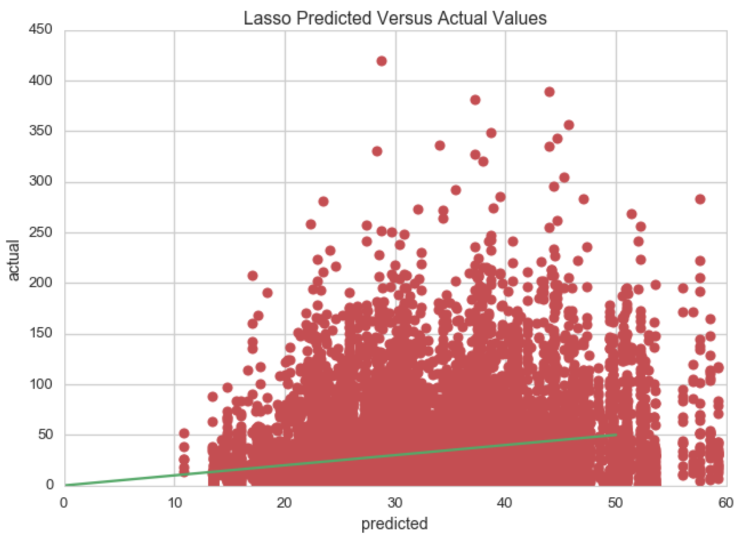
For ridge regression, SGD and lasso, GridSearchCV from Scikit Learn was used to identify the best set of input parameters. For ridge regression, the optimal alpha was 20. For lasso, the optimal alpha was 0.001. For SGD, the optimal parameters were an l1 penalty, with an alpha of 0.001.

Once the optimal parameters were identified, each algorithm was re-implemented using these best options. The outcomes are included in the table below.

	Train/Test RMSE	10-Fold Cross Validation RMSE
Multiple Regression	37.539	37.388
Ridge Regression	37.302	37.333

SGD	37.412	37.424
Lasso	37.328	37.369

Across the various algorithms, the RMSE value remained very consistent, and did not show significant improvement when the optimal parameters were identified through GridSearchCV. Although there was little variability among the models, the best-performing model—lasso—was used.



The full list of coefficients from the final model can be found in Appendix B, but the top five predictors are listed below.

Variable	Coefficient
INCOME_DESC_150-174K	15.08
INCOME_DESC_250K+	13.45
INCOME_DESC_200-249K	12.37
INCOME_DESC_175-199K	10.01
KID_CATEGORY_DESC_None/Unknown	-9.94

The most important variable is unsurprising—income. Household income plays a major role in transaction amount, and individuals with high incomes spend significantly more than those with low incomes. It's a bit more surprising, though, that as income increases, spending begins to decrease once income rises beyond $174,000.

The second-most important factor- number of children- is also understandable. It seems intuitive that those without children would be buying for a smaller number of people, and so grocery spending would be lower.

One outcome that was unexpected is that spending between married and unmarried individuals varied by only $0.10. It might be logical to predict that a married couple, as a household of two, would spend more than a single individual. The data show that it isn't the case with this group of customers. It's possible, however, that although an individual may be single, they may be living with a significant other or non-child family member, and thus spending a comparable amount to married couples.

These results indicate that retailers can maximize their customer transaction amounts by focusing marketing and sales efforts on targeting individuals in higher income brackets who have children.

Market Basket Analysis

Analyzing customer transactions with market basket analysis provides a better understanding of buying behaviors. Armed with this data, retailer can make more strategic decisions on in-store product placement, joint coupon offers, more effective mailers, and more.

Market basket analysis was conducted using the Apriori algorithm found in Machine Learning in Action. Due to the computational resources that would be needed to handle such a large dataset, a random sample of 50,000 baskets was used for analysis.

The following rule was generated using a minimum support value of 0.0001 and lift of 1:
CHOCOLATE MILK --> FLUID MILK WHITE ONLY   conf: 0.031 lift: 3.785
FLUID MILK WHITE ONLY --> CHOCOLATE MILK   conf: 0.014 lift: 3.785

Any support value larger than 0.0001 did not generate any rules. Unfortunately, processing 50,000 baskets took 12 hours of processing time, so significant iteration on parameter settings for the entire 50,000 baskets was not possible. A smaller set of 1,000 transactions was used, instead.

Experimentation with support values had a significant effect on the rules that were returned. In comparing rules from support values of 0.001 and 0.0001, there were no overlapping items represented in the rules. However, the choice of a confidence of 0.25 or greater versus lift value of 1 or greater did not alter the rules. Below are examples from each support level.

Support 0.001, Lift >=1
ORANGES NAVELS ALL --> MACARONI DRY   conf: 1.0 lift: 994.0
MACARONI DRY --> ORANGES NAVELS ALL   conf: 1.0 lift: 994.0
CONDENSED SOUP --> SAUERKRAUT   conf: 0.5 lift: 497.0
SAUERKRAUT --> CONDENSED SOUP   conf: 1.0 lift: 497.0
IWS SINGLE CHEESE --> ROMA TOMATOES (BULK/PKG) conf: 1.0 lift: 994.0

ROMA TOMATOES (BULK/PKG) --> IWS SINGLE CHEESE conf: 1.0 lift: 994.0
MUFFIN & CORN BREAD MIX --> DIPS (NON-REFRIGERATED) conf: 0.5 lift: 497.0
DIPS (NON-REFRIGERATED) --> MUFFIN & CORN BREAD MIX conf: 1.0 lift: 497.0

Support 0.0001, Lift >=1

PASTA: CANNED --> RTS SOUP: CHUNKY/HOMESTYLE ET conf: 1.0 lift: 989.0
RTS SOUP: CHUNKY/HOMESTYLE ET --> PASTA: CANNED conf: 1.0 lift: 989.0
DINNER SAUSAGE --> ALL FAMILY CEREAL conf: 1.0 lift: 989.0
ALL FAMILY CEREAL --> DINNER SAUSAGE conf: 1.0 lift: 989.0
BAKING BITS & MORSELS --> FRZN WHIPPED TOPPING conf: 1.0 lift: 989.0
FRZN WHIPPED TOPPING --> BAKING BITS & MORSELS conf: 1.0 lift: 989.0
SOFT DRINKS 6PK/4PK CAN CARB --> SNACK CAKE - MULTI PACK conf: 1.0 lift: 989.0
SNACK CAKE - MULTI PACK --> SOFT DRINKS 6PK/4PK CAN CARB conf: 1.0 lift: 989.0

When looking at the full 50,000 record set, it is interesting to see that customers see a need to buy both types of milk at once. It seems possible that an individual who was buying milk might simply buy chocolate syrup to make chocolate milk, but buying patterns show otherwise. This pattern suggests that shoppers see both types of milk as essential staple items. It might be an opportunity to shift retail displays to offer high-dollar options like packages of individual chocolate milk bottles directly next to the gallons of regular milk. Marketing promotions that offer discounts for buying the items together in higher volumes might also be worth exploring, as could offering a loss leader promotion on the combination of products.

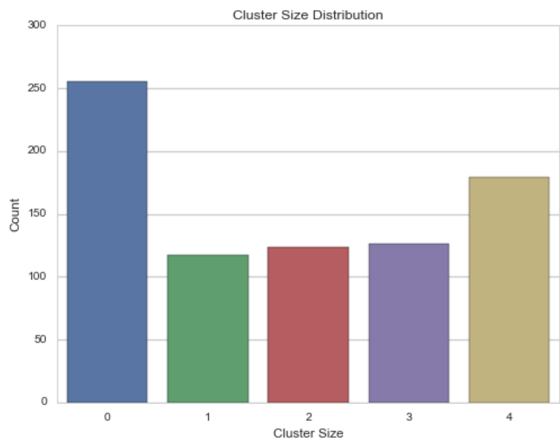
Customer Segmentation

An understanding of customers can boost marketing efforts. Clustering can be used to group similar customers and develop “profiles” for segments of customers.

Clustering was used to build demographic profiles of customers. Because the demographic data were exclusively binary-coded categorical data, Scikit Learn’s agglomerative hierarchical clustering was used. Clustering was performed using Ward’s linkage, complete linkage, and average linkage, with 2, 4, 5 and 7 clusters each.

The trial with 2 clusters was rejected, as it produced highly unbalanced clusters, such as in the 2-cluster trial with complete linkage, which produced one cluster with 702 of the 801 records. Five clusters were ideal; they grouped customers by demographic profiles that were simple and understandable, while still providing a degree of granularity.

Ward and complete linkage produced nearly identical cluster composition, although Ward produced the most balanced number of items in each cluster. Average linkage produced five clusters with highly imbalanced sizes, and was removed from consideration. Ward’s linkage with 5 clusters was selected as the optimal model.



The table below shows the traits that characterize each cluster. Cluster 0, the largest cluster, is composed of single females aged 45-54 with yearly household incomes of $50,000 - $75,000.

Cluster	Age	Household Residents	Homeowner Status	Income	Children	Marital Status
0	45-54	Single Female	Unknown	50-74K	None/Unknown	Unknown
1	35-44	2 Adults Kids	Homeowner	50-74K	3+	Married
2	45-54	2 Adults No Kids	Homeowner	35-49K	None/Unknown	Unknown
3	45-54	2 Adults Kids	Homeowner	35-49K	1	Married
4	45-54	2 Adults No Kids	Homeowner	50-74K	None/Unknown	Married

Cluster 1 represents married couples with larger families. They are aged 35-44, own a home, earn $50,000 to $75,000 per year, and have three or more children.

Clusters 2 and 4 differ only by income. Both contain individuals aged 45-54 who don't have children, own their home, and live in a household with another adult. Cluster 2's annual income, however, is $35,000 - $49,000, while cluster 4 is $50,000 to $74,000.

Cluster 3 are married couples aged 45-54, who own their home, have a single child, and earn $35,000 to $49,000 per year.

The composition of these clusters suggests that the presence of children in the home is the most straightforward means of segmenting customers for marketing purposes. Although clusters also differed in other respects, factors like age and household income fall within similar ranges of values, while marital status isn't a factor that can easily characterize customer behavior.

Coupon Redemption

Retailers could potentially save marketing dollars by only sending coupons to customers most likely to redeem them. To provide insight into redemption behavior, logistic regression and classification techniques were used with demographic characteristics to predict whether a customer would redeem any coupons. Logistic regression, naïve Bayes, k-nearest neighbors, and decision trees were chosen specifically because of their abilities in working with categorical variables with binary outcomes. The outcome from logistic regression was compared to the best-performing classification algorithm—decision trees.

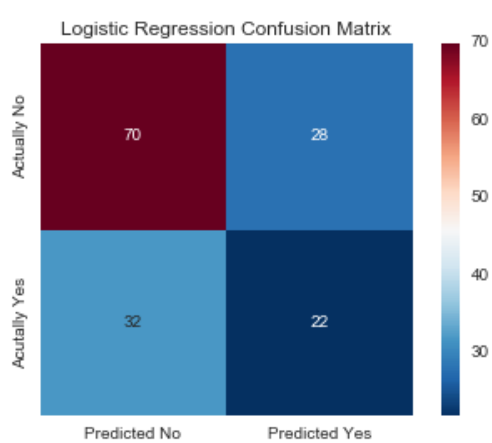
Logistic Regression

Logistic regression was conducted using Scikit Learn's LogisticRegression function, and evaluated by overall prediction accuracy rate.

	Accuracy Score
Training	0.638
Testing	0.605
10-Fold Cross Validation	0.586

Logistic Regression Classification Report

	Precision	Recall	F1-score	Support
Don't Redeem	0.69	0.71	0.70	98
Redeem	0.44	0.41	0.42	54
Average/Total	0.60	0.61	0.60	152



Classification Methods

To compare theoretical approaches, logistic regression was compared to Bernoulli naïve Bayes, k-nearest neighbors, and decision tree algorithms. Naïve Bayes offered the best results of these methods, with a 5-fold cross-validation accuracy rate of 60%.

Method	Accuracy Score
Naïve Bayes	0.600
K-Nearest Neighbors	0.582
Decision Trees	0.592

Naïve Bayes

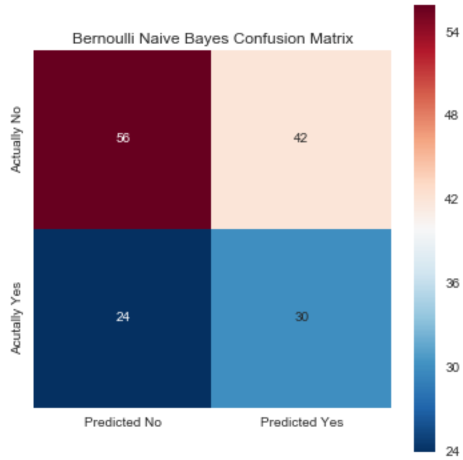
Naïve Bayes classification was implemented using Scikit Learn's BernoulliNB function. Bernoulli naïve Bayes was used, as it is designed specifically to work with binary variables. Overall accuracy was evaluated by prediction accuracy rate.

GridSearchCV from scikit learn was used to identify the best set of input parameters. The optimal alpha was 0.4, with fit_prior=True.

	Accuracy Score
Training	0.625
Testing	0.566
10-Fold Cross Validation	0.60

Bernoulli Naïve Bayes, 5-Fold Cross Validation Classification Report

	Precision	Recall	F1-score	Support
Don't Redeem	0.70	0.58	0.64	98
Redeem	0.42	0.56	0.48	54
Average/Total	0.60	0.57	0.58	152



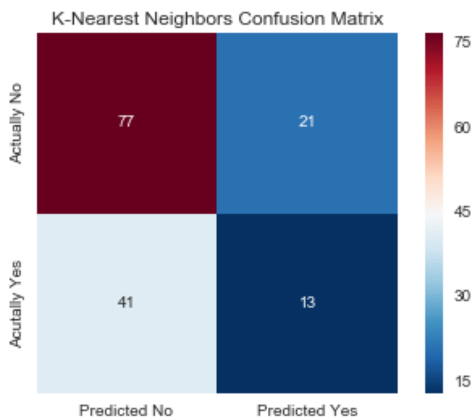
K-Nearest Neighbors

K-Nearest Neighbors classification was implemented using Scikit Learn's KNeighborsClassifier function. Iterations were conducted using several different k-values, as were trials with both distance and uniform weights. Overall accuracy was evaluated by prediction accuracy rate. A table of results is presented below. K=10 with uniform weighting produced the greatest accuracy rate, and showed the least indication of over-fitting.

K	Weight		Accuracy Rate
3	Distance	Training	0.533
		Testing	0.872
		5-fold Cross Validation	0.529
3	Uniform	Training	0.559
		Testing	0.753
		5-fold Cross Validation	0.539
5	Distance	Training	0.493
		Testing	0.872
		5-fold Cross Validation	0.551
5	Uniform	Training	0.474
		Testing	0.709
		5-fold Cross Validation	0.553
10	Distance	Training	0.553
		Testing	0.872
		5-fold Cross Validation	0.570
10	Uniform	Training	0.618
		Testing	0.658
		5-fold Cross Validation	**0.582**

K= 10, Uniform Weighting, 5-fold Cross Validation Classification Report

	Precision	Recall	F1-score	Support
Don't Redeem	0.67	0.82	0.73	98
Redeem	0.44	0.26	0.33	54
Average/Total	0.59	0.62	0.59	152



Decision Trees

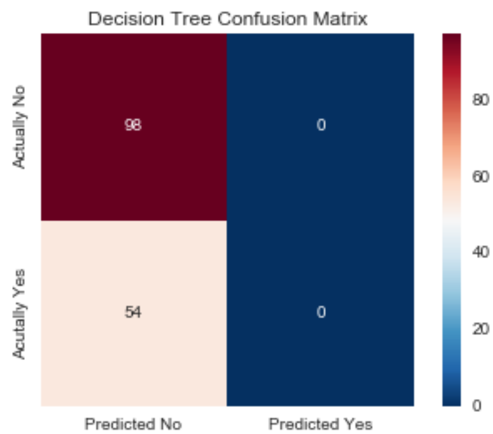
Decision tree classification was implemented using Scikit Learn's DecisionTreeClassifier function. Overall accuracy was evaluated by prediction accuracy rate. Tree parameters were adjusted individually. The best entropy tree was obtained with a minimum sample split of 3, maximum number of features of 5, and maximum tree depth of 2. The best gini tree was obtained with a minimum sample split of 5, maximum number of features of 6, and maximum tree depth of 2. Differences between the accuracy rates of the entropy versus gini trees were negligible.

Criterion		Accuracy Score
Entropy	Training	0.577
	Testing	0.645
	10-Fold Cross Validation	0.591
Gini	Training	0.582
	Testing	0.638
	10-Fold Cross Validation	**0.592**

Decision Tree (gini), 5-Fold Cross Validation Classification Report

	Precision	Recall	F1-score	Support

Don't Redeem	0.64	1.00	0.78	98
Redeem	0.00	0.00	0.00	54
Average/Total	0.42	0.64	0.51	152



Final Model- Logistic Regression

To more easily interpret the results of the logistic regression model, the log odds ratio for each variable was computed. An abbreviated table of the variables with the top 5 greatest odd values is included below. The full odds ratio table is available in Appendix C.

Variable	Odds
income_desc_150-174K	2.46
income_desc_125-149K	1.67
homeowner_desc_Renter	1.47
income_desc_50-74K	1.46
income_desc_175-199K	1.37

It's interesting to see that customers with relatively high household incomes are more likely to redeem coupons. This finding runs counter to the general belief that coupons are primarily used by individuals with limited financial resources. Renters are also likely to redeem coupons.

In looking at all variables with odds ratios of greater than one, the most influential variables are customer age, household income, and homeowner status.

Using the logistic regression model, it's obvious that the overall accuracy of the model is relatively low. Cross-validation on the testing set indicated that the overall accuracy of the classifier was 58.6%. When extending the model to the full dataset, the accuracy rate remains comparable. The model predicts 178 customer redemptions of coupons, when in reality, 311 customer redeemed coupons—an overall accuracy of 57.2%. It's interesting to note that in this case, the classifier's error rate resulted in a prediction that underestimated coupon redemption.

Final Model: Bernoulli Naïve Bayes

In cross-validation testing, the Bernoulli naïve Bayes classifier out-performed other classifiers, with an overall accuracy rate of 60%. When applied to the full dataset, the model predicts 353 redemptions, when in reality, 311 customers redeemed coupons. Here, the classifier has over-predicted the number of redeeming customers.

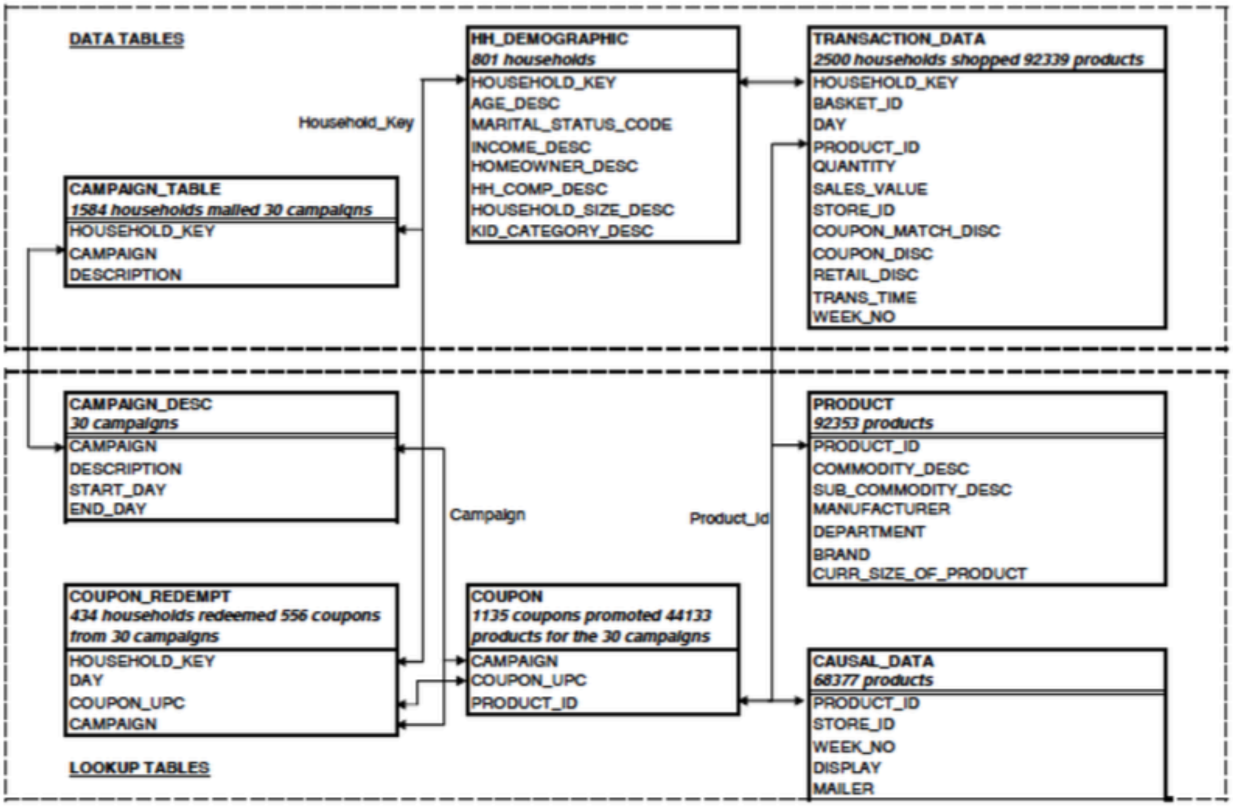
Model Comparison

While the naïve Bayes model out-performed the logistic regression model, both models display utility. If a marketing budget was tight, the logistic regression model may be a better choice, since the under-prediction will make it less likely that coupons are being sent to non-redeeming customers. When budget is less of a concern and over-estimation is acceptable, the improved accuracy rate of the naïve Bayes model may be better. Marketers may also find the logistic regression model's odds ratios to be helpful, since they can help in understanding the role of various predictors on overall redemption rates.

Conclusion

Knowledge discovery techniques such as regression, classification, clustering, and association rules offer tremendous benefits to marketers who seek to understand customer profiles, shopping behaviors, and promotion response rates.

Appendix A: Data Schema



Appendix B: Lasso Regression Coefficients

y-intercept: 34.76

Variable	Model Coefficient
AGE_DESC_19-24	-0.06
AGE_DESC_25-34	2.47
AGE_DESC_35-44	2.23
AGE_DESC_45-54	1.30
AGE_DESC_55-64	-1.53
AGE_DESC_65+	-3.82
HH_COMP_DESC_1 Adult Kids	-6.76
HH_COMP_DESC_2 Adults Kids	-8.58
HH_COMP_DESC_2 Adults No Kids	3.52
HH_COMP_DESC_Single Female	0.28
HH_COMP_DESC_Single Male	0.64
HH_COMP_DESC_Unknown	-3.88
HOMEOWNER_DESC_Homeowner	7.82
HOMEOWNER_DESC_Probable Owner	-7.38
HOMEOWNER_DESC_Probable Renter	0.00
HOMEOWNER_DESC_Renter	-0.64
HOMEOWNER_DESC_Unknown	2.88
HOUSEHOLD_SIZE_DESC_1	2.17
HOUSEHOLD_SIZE_DESC_2	0.00
HOUSEHOLD_SIZE_DESC_3	-6.93
HOUSEHOLD_SIZE_DESC_4	-1.97
HOUSEHOLD_SIZE_DESC_5+	3.04
INCOME_DESC_100-124K	-2.57
INCOME_DESC_125-149K	1.79
INCOME_DESC_15-24K	-7.31
INCOME_DESC_150-174K	15.08
INCOME_DESC_175-199K	10.01
INCOME_DESC_200-249K	12.37
INCOME_DESC_25-34K	-8.57
INCOME_DESC_250K+	13.45
INCOME_DESC_35-49K	-6.28
INCOME_DESC_50-74K	0.00
INCOME_DESC_75-99K	3.55
INCOME_DESC_Under 15K	-7.41
KID_CATEGORY_DESC_1	7.73
KID_CATEGORY_DESC_2	9.59
KID_CATEGORY_DESC_3+	4.15
KID_CATEGORY_DESC_None/Unknown	-9.94
MARITAL_STATUS_CODE_A	0.10
MARITAL_STATUS_CODE_B	0.00
MARITAL_STATUS_CODE_U	-1.42

Appendix C: Coupon Redemption Odds Ratios

variable	Odds Ratio
AGE_DESC_19-24	0.505141611
AGE_DESC_25-34	1.131521916
AGE_DESC_35-44	1.071539531
AGE_DESC_45-54	1.248935562
AGE_DESC_55-64	1.036310955
AGE_DESC_65+	0.994922018
marital_status_code_A	1.000980969
marital_status_code_B	0.784466214
marital_status_code_U	1.004392384
income_desc_100-124K	1.128469029
income_desc_125-149K	1.665770461
income_desc_15-24K	0.594212645
income_desc_150-174K	2.462421462
income_desc_175-199K	1.374585293
income_desc_200-249K	0.43306419
income_desc_25-34K	1.058283638
income_desc_250K+	0.301753434
income_desc_35-49K	1.07951613
income_desc_50-74K	1.461036826
income_desc_75-99K	0.841929444
income_desc_Under 15K	1.135926239
homeowner_desc_Homeowner	1.034583623
homeowner_desc_Probable Owner	1.296403106
homeowner_desc_Probable Renter	0.68939597
homeowner_desc_Renter	1.473571477
homeowner_desc_Unknown	0.578839109
HH_COMP_DESC_1 Adult Kids	0.988697779
HH_COMP_DESC_2 Adults Kids	0.962611608
HH_COMP_DESC_2 Adults No Kids	0.967062221
HH_COMP_DESC_Single Female	1.237546826
HH_COMP_DESC_Single Male	0.861767914
HH_COMP_DESC_Unknown	0.803493641
KID_CATEGORY_DESC_1	0.722531933
KID_CATEGORY_DESC_2	1.212535442
KID_CATEGORY_DESC_3+	1.135360225
KID_CATEGORY_DESC_None/Unknown	0.792899807