Springboard --- DSC

Capstone Project 3

Predicting Customer Spending & Coupon Use from Dunnhumby Dataset

By Annie Erbsen

December 2023

1. Introduction

The Dunnhumby Complete Journey dataset contains household level transactions over two years from a group of 2,500 households that shop at a popular retailer, along with data about marketing campaigns.

Each household in this study has been sent multiple direct marketing campaigns over the course of 2 years. Each campaign consists of coupons that were sent to specific households that were valid during certain time periods. The dataset includes data on which coupons were redeemed by a household, for what products coupons can be used on, the store location, and when. There are also data about what in-store displays were in specific stores over different time periods, as well as the locations of featured products in mail campaigns. In addition to this, we have all of the purchase histories from each household from this retailer over the two year time span, as well as demographic data from 32% of the households. For the purchases, we have the day, the week and the time of day the purchase took place, but the day and weeks are numbered, so we do not know the true dates or days of the week. We also do not know the locations of the retailer, the average income and cost of living demographics for store location zones, or details about the marketing campaigns aside from what customers used them for.

The primary objectives of this project are to understand how household demographics contribute towards customer spending and coupon use, to predict future customer spending and coupon use, and to make recommendations on how to develop more effective marketing campaigns based on my findings.

2. Approach

2.1 Data Acquisition, Wrangling & Storytelling

The data were acquired from Dunnhumby's Complete Journey dataset on Kaggle: https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey. It included 8 csv files that are summarized in the table below (Figure 1)

	df_name	number_of_rows	number_of_columns	column_names
0	campaign_desc	30	4	DESCRIPTION, CAMPAIGN, START_DAY, END_DAY
1	campaign_table	7208	3	DESCRIPTION, household_key, CAMPAIGN
2	causal_data	36786524	5	PRODUCT_ID, STORE_ID, WEEK_NO, display, mailer
3	coupon	124548	3	COUPON_UPC, PRODUCT_ID, CAMPAIGN
4	coupon_redempt	2318	4	household_key, DAY, COUPON_UPC, CAMPAIGN
5	hh_demographic	801	8	AGE_DESC, MARITAL_STATUS_CODE, INCOME_DESC, HOMEOWNER_DESC, HH_COMP_DESC, HOUSEHOLD_SIZE_DESC, KID_CATEGORY_DESC, household_key
6	product	92353	7	PRODUCT_ID, MANUFACTURER, DEPARTMENT, BRAND, COMMODITY_DESC, SUB_COMMODITY_DESC, CURR_SIZE_OF_PRODUCT
7	transaction_data	2595732	12	household_key, BASKET_ID, DAY, PRODUCT_ID, QUANTITY, SALES_VALUE, STORE_ID, RETAIL_DISC, TRANS_TIME, WEEK_NO, COUPON_DISC, COUPON_MATCH_DISC

Figure 1

The data can all be easily joined together, as shown in Figure 2.

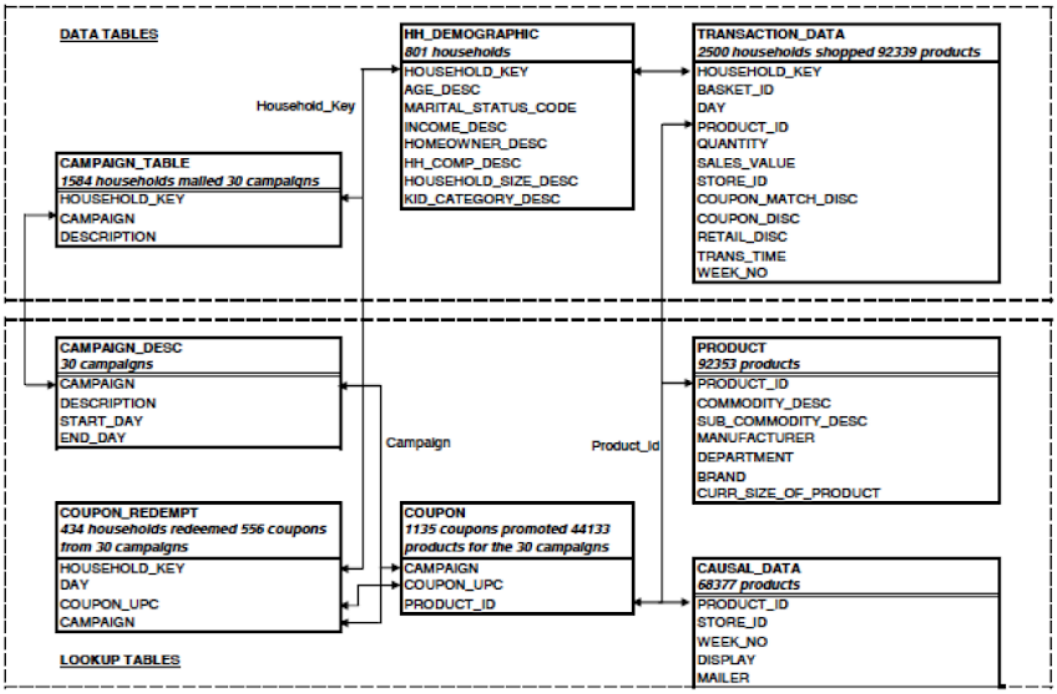
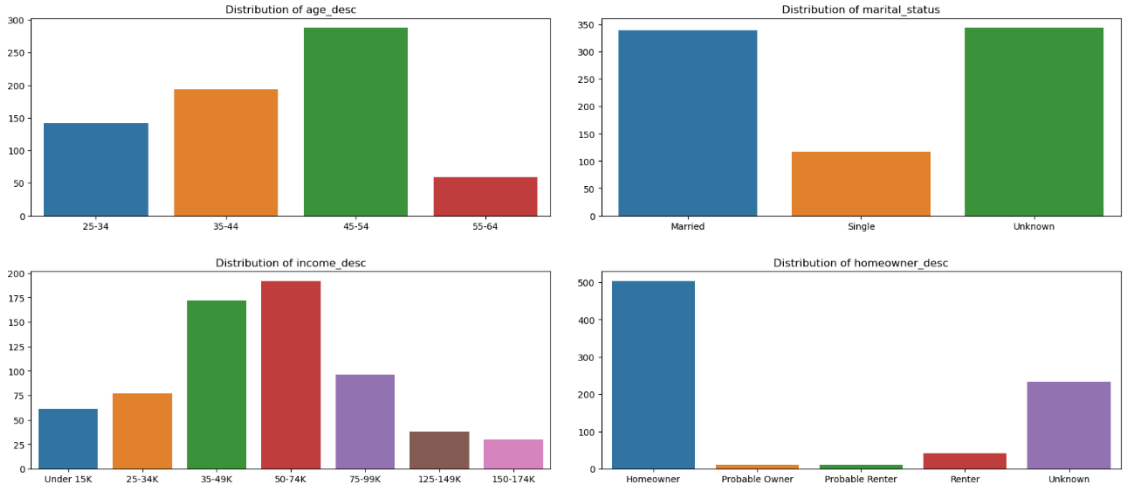


Figure 2

There were 30 campaigns, with a mean duration of 47 days. The campaigns started on day 224, ended on day 719, and the marketing campaigns sent out the most were numbered 8, 13 and 18. The 30 campaigns included 1135 unique coupons, most of which were valid for a number of different products and departments.

Of the 2500 households in the dataset, only 801 have demographics information available. Because this project focuses on how demographics contribute to sales and coupon use, I will only be using data for the households with demographics information. The distribution of the tracked demographics are shown below in Figure 3.



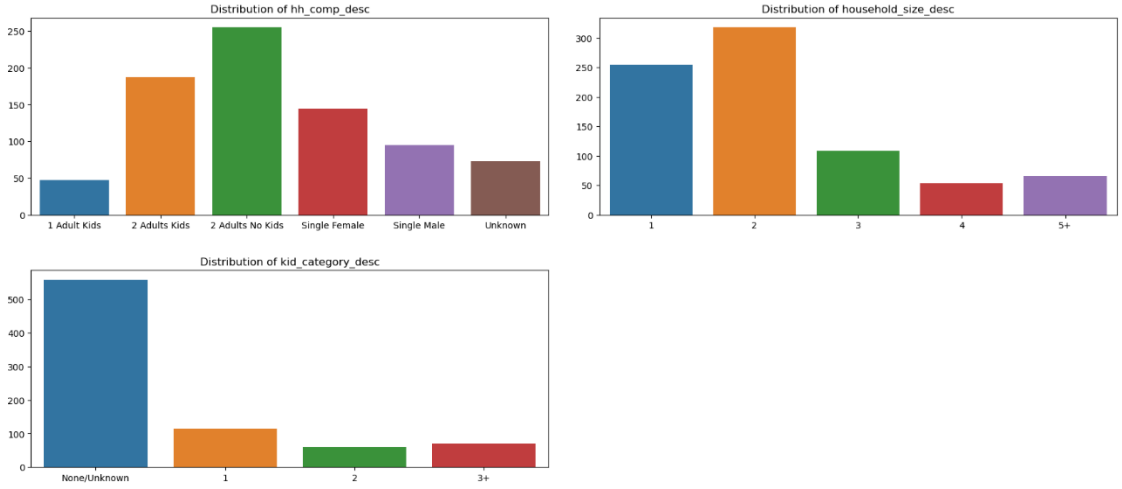


Figure 3

As shown above, there are certain demographics which are more common; in particular, ages 35-54, marital status of married and unknown, incomes 35-74K, homeowners, 2 adults no kids household makeup, 1-2 person households, and none/unknown number of children.

I next calculated the coupon redemption rates of the different demographic groups, shown in Figure 4. Here we can see that the households that have the highest redemption rates are ages 55-64, single, income over 125K, probable homeowners, 1 adult with kids households, households with 4 people, and households with 2-3 kids. It should be noted that most of the demographics with the highest coupon redemption rates were among the groups where the fewest coupons were sent.

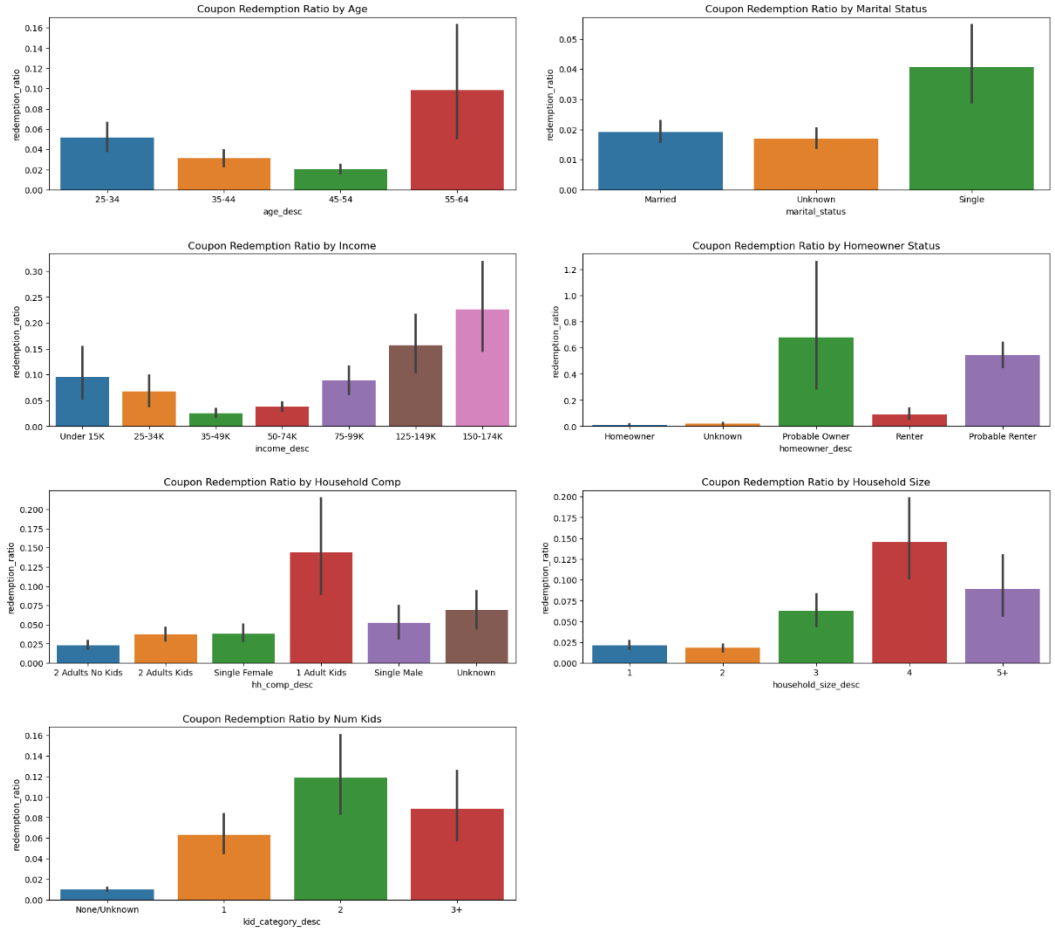


Figure 4

Looking next at the transaction data, the 2500 households were tracked over 102 weeks, and over that time there were 276484 basket IDs. Each item purchased has its own row, and includes the household key, basket id, day/time/week, product ID, sales value, store id, as well as columns for discounts. These discounts were from coupons, coupon matches from other stores, and loyalty program discounts. All of the customers are in the loyalty program and used this discount, and as we are interested in the marketing campaigns specifically for the retailer in this study, I only kept the coupon discount column and discarded the coupon match and loyalty discounts.

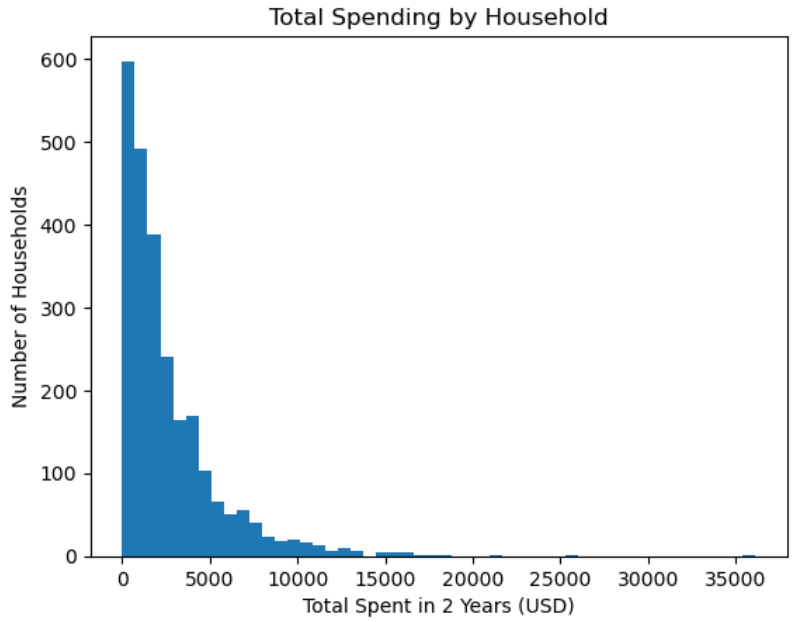


Figure 5

In Figure 5 we can see a histogram of the total spending by household. Most households spent significantly under $5000 a year, and there is a roughly exponential drop in spending as the values increase. However, because there is a long tail on how much households spend, I expect this to introduce higher error later on in the modeling phase.

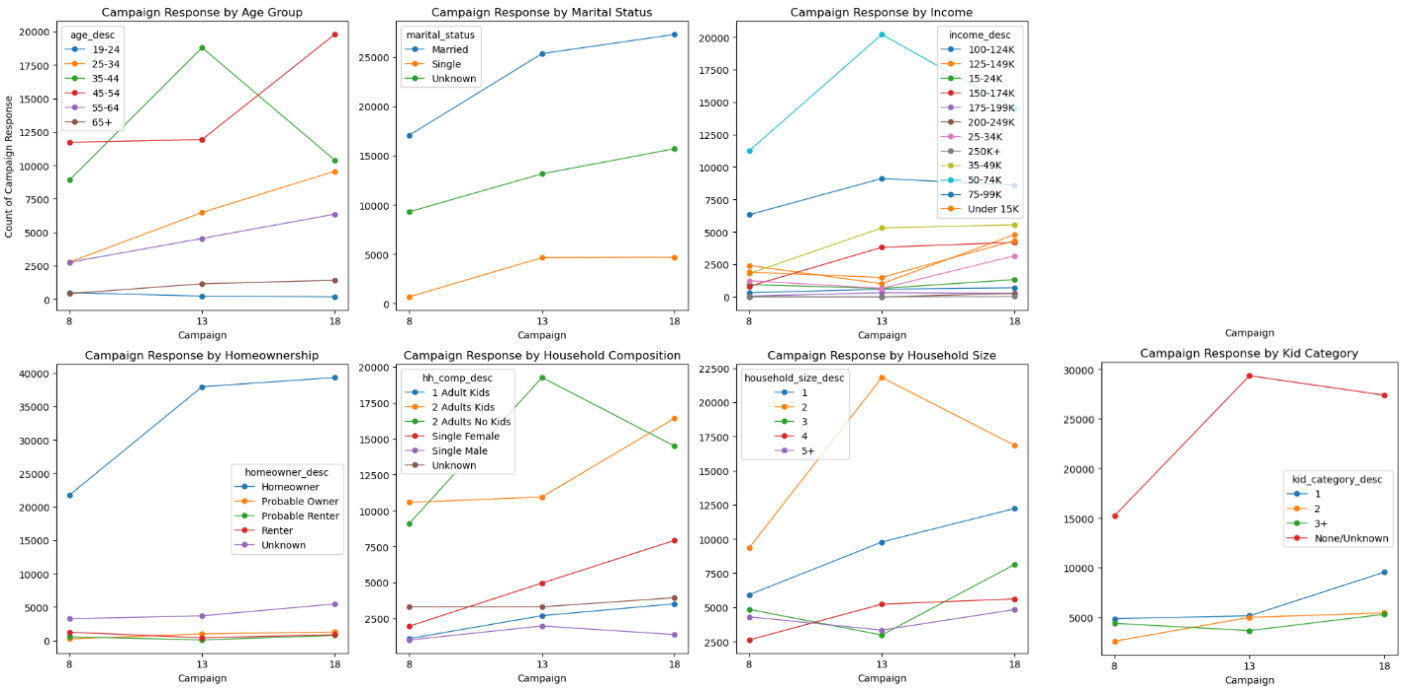


Figure 6

As shown in Figure 6, where I plotted the number of coupons redeemed from the 3 most widely mailed campaigns, we can see how the different demographics groups responded. The general trend that I see for campaign response is that overall there was the most response to 18, followed by 13 and 8. However, for any given demographic category there was usually one demographic group that responded most to 13. Generally, the groups that responded the most to campaigns were:

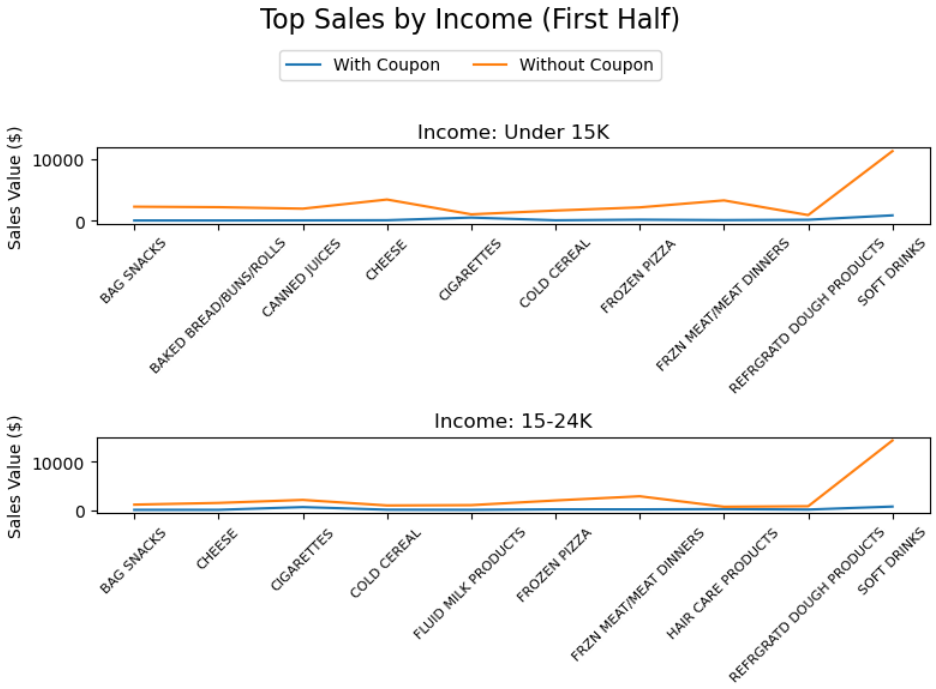
- Ages 35-54
- Married
- Earned 50-99K
- Homeowners
- 2 adult households (both with kids and no kids)
- 1-2 household size
- No/unknown number of kids

This mostly aligns with the distribution of how campaigns were sent out. However, I do see some notable differences:

- Married and 'Unknown' were sent similar numbers of campaigns, but the married households responded much more.
- Income group 35-49K were sent more campaigns than 75-99K, but the higher income group responded much more.
- Homeowners were sent about twice as many coupons as 'Unknown' owners, but redeemed 5-8x as many coupons.

The data suggest that focusing on the above demographics would yield higher campaign response rates.

Figure 7 shows product category sales by income, broken down by sales using a coupon and with no coupon. I am only including the income demographic group as an example; in this figure we can see that while the sales using no coupon vary quite a bit, the sales by category where coupons are used are quite flat. Because I saw the same trend across demographics groups, in proceeding I chose to look at sales overall, and not broken down by product or product categories.



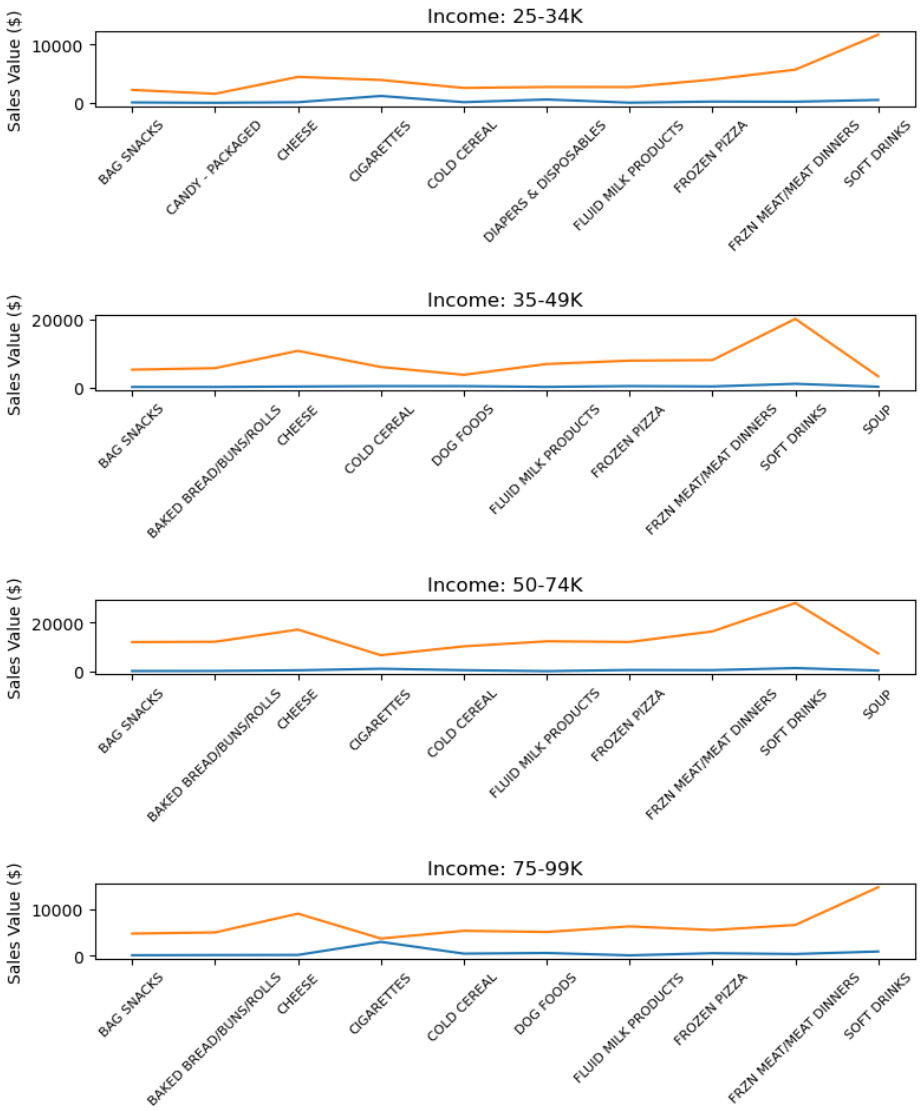


Figure 7

Moving on to the time component, I next looked at weekly sales over time (Figure 8) and active campaigns over time (Figure 9). The weekly sales gradually increased until around week 20, at which point they roughly stabilized. I suspect that data wasn't being gathered from all of the customers until this point, but as I don't have access to enough information about what was happening at this point, it is impossible to say. Because I am interested in sales values and coupon use, I decided to only keep the data during the times when there were active campaigns, which conveniently eliminated the tail at the beginning of Figure 9.

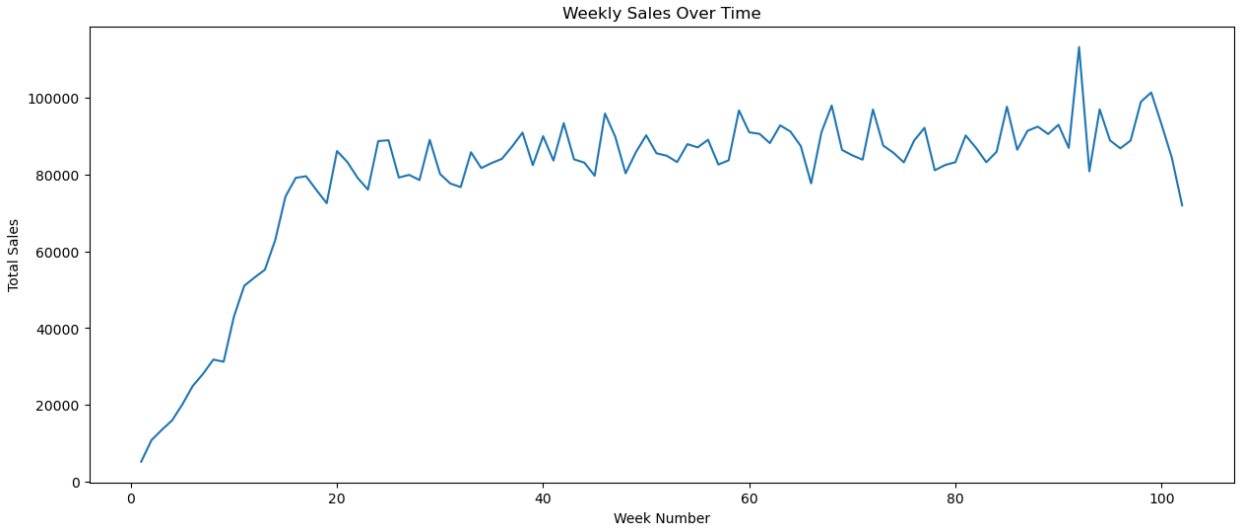


Figure 8

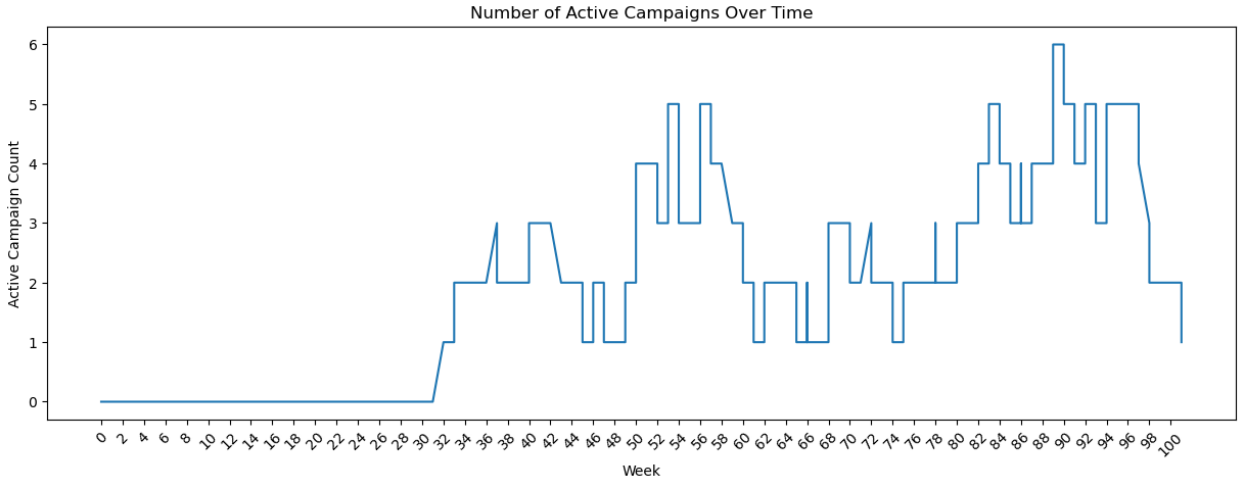


Figure 9

Based on the EDA done up until this point, I built a table for modeling that included the weekly sales totals for every customer with demographics data, the total coupon discounts, and the demographics data. The data were grouped by week and household key.

Using this newly created table, I did one last piece of EDA, which was to look at average sales volume per household vs average sales growth. In Figure 10, I segmented the households into quadrants. The high sales/high growth quadrant would be excellent targets for continued marketing campaigns, as I also found that customers on average spend more money when using coupons. However, all of these quadrants could benefit from targeted campaigns except for the low sales/low growth quadrant. These households are segmented into 4 different csv files in the project file.

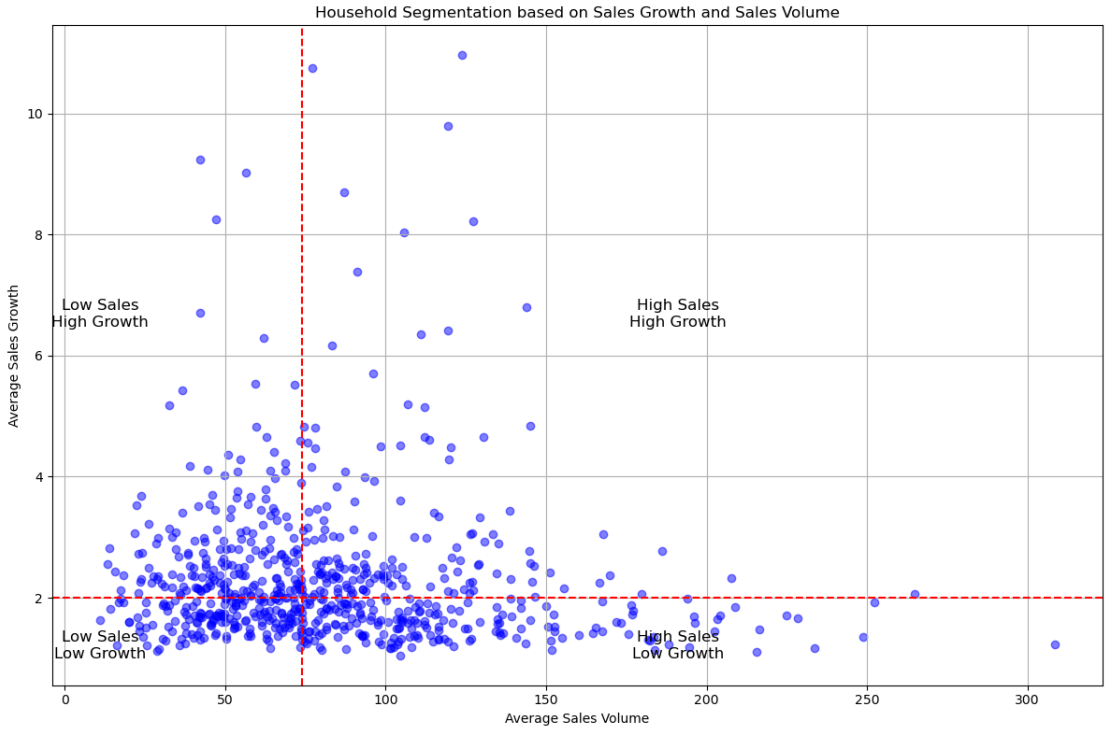


Figure 10

2.2 Baseline Modeling

As a first attempt at a model, I ran a K Nearest Neighbors (KNN) model with total sales value as the target. I did my train test split in such a way so that I did not break up any data mid-week; all data under 75 weeks went in the train group, and data 75 weeks and over went in the test group. The Mean Absolute Percentage Error (MAPE) was extraordinarily high at 2608569104823%, which told me that this was not a good model to start with.

I next tried an Ordinary Least Squares Regression model using the same train/test split from the KNN model. While a linear model was not likely to work well in this situation, it is easily interpretable and can provide meaningful information. The R-Squared was 0.25, which indicates a moderate level explanation. The results of the OLS model also showed that there were features with a P-value under 0.05, meaning that we can reject the null hypothesis (the features do not contribute towards sales values). These features with significant P-values were age, income, homeowner, household composition, and household size. While this model has little predictive power, it is important to note that we now know that most of the demographic features are statistically significant in their contributions towards the model results.

My final baseline models were XGBoost Regression and XGBoost Classification models. I added a coupon redemption rate feature; having the number of coupons redeemed as a feature introduced leakage in the regression model because the true number of coupons redeemed is collinear with sales values. The redemption rate is something that can be calculated for any customer where transactions are tracked, and is likely a contributor towards sales values without leakage. I added a ‘cluster’ feature, which was the result of K Means Clustering to see if that would improve the models, and I found 3 distinct clusters.

For the XGBoost Regression model where the mean sales value was $81.93, the baseline MAE was 42.43. I next split the data by clusters and ruan the model on the clusters separately, which did not improve the model at all, so I discarded this feature. I also tried to address outliers by ‘Windsorizing’ the data, which helped a little (MAE=39.07), and the IQR method, which was slightly better(MAE=36.8).

For the XGBoost Classification model for predicting coupon use, the recall score for the positive class was great (0.85), but the CV mean recall score was 0.46, indicating major overfitting. I determined this was likely due to imbalanced classes, and I used the scale_pos_weight feature of XGBoost as a first attempt at addressing the issue. I still had a similar level of overfitting, which I addressed in the build-out of the model.

2.3 XGBoost Regression for Sales Prediction Model

For the XGBoost Regression model, I dropped the week number and the household key, and the target feature was the total sales value. I did not stratify the train test split based on time for this model because I do not know if there is a seasonal component that could impact splitting by time. The mean weekly sales value was $80, with a MAE of 42.5, and the mean CV MAE was 55. This error is much better than our previous models, but is still quite high and there's some overfitting. This is more than likely because of outliers.

To create a better model, I performed an interquartile range (IQR) analysis, and selected the IQR with the fewest outliers, as seen in Figure 11 (IQR=0.25). The mean sales value for this subset of the data was $58, a full $12 lower than on the full dataset.

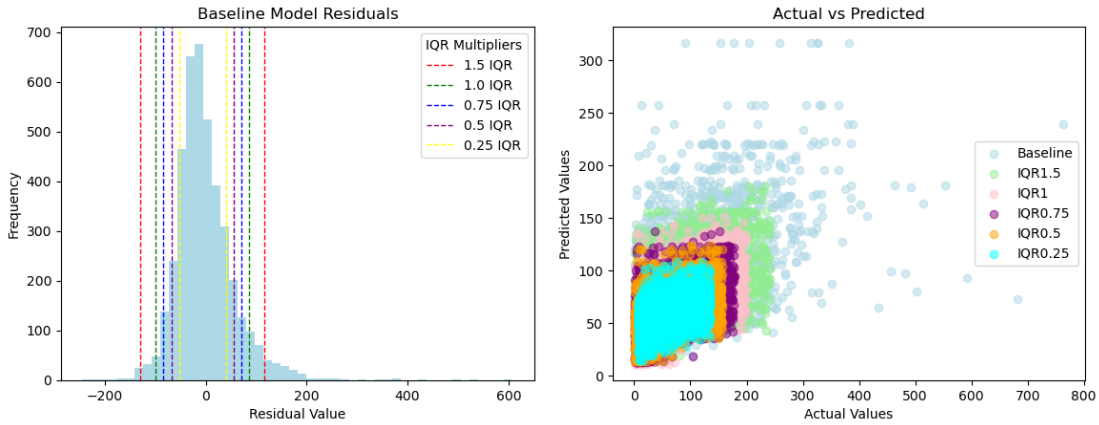


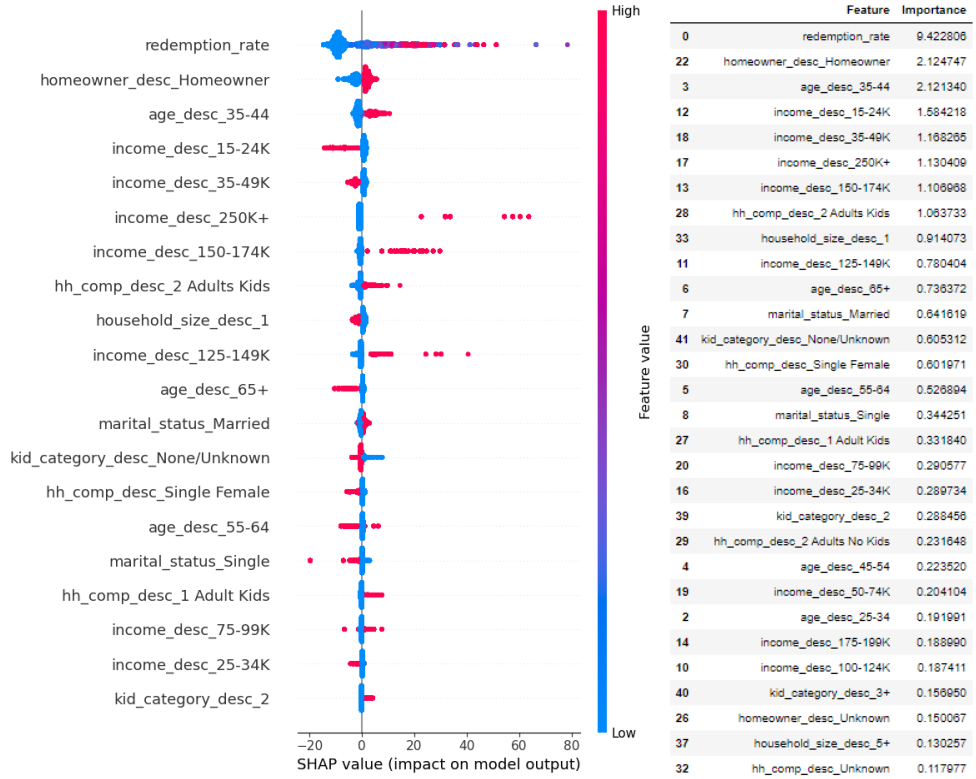
Figure 11

Running the XGBoost model on this smaller subset resulted in a much improved model; there was a MAE of 28.96, with a CV MAE of 29.07. This error is quite a bit lower (though still on the high side), and there is almost no overfitting now.

I next ran the model on the entire dataset to see how well it was able to generalize, and ended up with a MAE of 45.93 and CV MAE of 50.68. The MAE was slightly higher than the baseline XGBoost model, but as the CV MAE was lower and showed less overfitting, this model is a bit better than the baseline. None of these values are good enough to be useful as a predictive model, but they are good enough to show directionality, so SHAP analysis is an appropriate next step. This way we will be able to determine how features (or combinations of features) contribute towards sales.

2.3.1 SHAP Analysis for XGBoost Regression Model

Figure 12 shows the summary plot for the SHAP analysis and the top ranked feature importance values.



	Feature	Importance
0	redemption_rate	9.422806
22	homeowner_desc_Homeowner	2.124747
3	age_desc_35-44	2.121340
12	income_desc_15-24K	1.584218
18	income_desc_35-49K	1.168265
17	income_desc_250K+	1.130409
13	income_desc_150-174K	1.106968
28	hh_comp_desc_2 Adults Kids	1.063733
33	household_size_desc_1	0.914073
11	income_desc_125-149K	0.780404
6	age_desc_65+	0.736372
7	marital_status_Married	0.641619
41	kid_category_desc_None/Unknown	0.605312
30	hh_comp_desc_Single Female	0.601971
5	age_desc_55-64	0.526894
8	marital_status_Single	0.344251
27	hh_comp_desc_1 Adult Kids	0.331840
20	income_desc_75-99K	0.290577
16	income_desc_25-34K	0.289734
39	kid_category_desc_2	0.288456
29	hh_comp_desc_2 Adults No Kids	0.231648
4	age_desc_45-54	0.223520
19	income_desc_50-74K	0.204104
2	age_desc_25-34	0.191991
14	income_desc_175-199K	0.188990
10	income_desc_100-124K	0.187411
40	kid_category_desc_3+	0.156950
26	homeowner_desc_Unknown	0.150067
37	household_size_desc_5+	0.130257
32	hh_comp_desc_Unknown	0.117977

Figure 12

Detailed discussion of the findings will take place in section '3 Findings' below. Consistent with our EDA, coupon redemption rate is the most important contributor towards high SHAP values, followed by homeowners and age 35-44. It is surprising to see, however, that the higher income brackets overall have very high SHAP values, in particular those that earn 250K+. The feature interaction plots didn't reveal much, but I did see that the following feature interactions had higher SHAP values: income 35-49K & age 35-44, household size 1 & not a homeowner, single females & low redemption rate.

2.3.2 Findings for the Sales Prediction Model

In this section I picked a 25% IQR subset of the full dataset of weekly sales values and demographics for customers, and built an XGBoost regression model upon it to predict customer spending. Because of the variation in the data likely due to external factors not considered in the model, seasonality and multicollinearity, I was not able to build a model that could accurately predict spending, but I was able to use SHAP analysis to identify features and interactions of features which contribute towards customer spending.

Features that contribute towards higher sales:

- Higher redemption rates
- Homeowners earning over 24K a year
- 2 person households consisting of a married couple
- 35-44 year olds, especially if they earn 35-49K a year
- 75-99K a year
- 125-174K a year
- 250K+ a year
- 2 person households with kids who are not single
- Households composed of 1 adult and children, especially if they earn 35-49K a year
- 'Unknown' marital status with redemption rates over 1

Features that contribute towards lower sales:

- 15-24K income, especially if they are homeowners
- 35-49K income, especially if they are homeowners (except 35-44 year olds and single parents)
- 1 person households/single people, especially if they are homeowners and/or females
- Customers who are age 55+
- 'Unknown' homeowners

2.4 XGBoost Classification for Coupon Use Prediction Model

To build the coupon use prediction model, I first created a new binary column 'coupon_user' that was 1 If the redemption rate was greater than 0, and dropped the redemption rate column. The new column indicated if the customer ever used a coupon.

For the train test split, I made sure to split the households up so that they were kept together, thus avoiding leakage. In my baseline model I did not do this, which is why my recall score was so high. I also addressed the imbalanced class issue by running the model on multiple resampling methods, and compared them to a baseline model with no resampling: SMOTEN (oversampling), ADASYN (oversampling) RandomUnderSampler (undersampling), and combining SMOTEN & RandomUnderSampler (Figure 14). The ADASYN oversampled model was the best, so I proceeded with it.

	Strategy	Mean CV Recall	Recall	Precision	F1 Score	ROC AUC	PR AUC
0	Baseline Model	0.443911	0.663891	0.401373	0.500285	0.615077	0.390037
1	SMOTEN Oversampled Model	0.466664	0.753974	0.428019	0.546053	0.637263	0.408379
2	Undersampled Model	0.476857	0.704769	0.406195	0.515361	0.650510	0.422635
3	SMOTEN Oversampled+ Undersampled Model	0.466664	0.753974	0.428019	0.546053	0.637263	0.408379
4	ADASYN Oversampled Model	0.518990	0.753974	0.441881	0.557203	0.665492	0.443410

Figure 13

Hyperparameter tuning of the selected model resulted in improved CV scores, with much less overfitting: recall=0.66, mean CV recall=0.68. This is decent, but I wanted to see if I could further improve it, so I looked at feature importance, correlation heatmap, and permutation importance. The permutation importance plot was the most useful in terms of indicating a way of improving model importance, as seen in Figure 14. The negative feature values decreased model performance, so I dropped these features from the table.

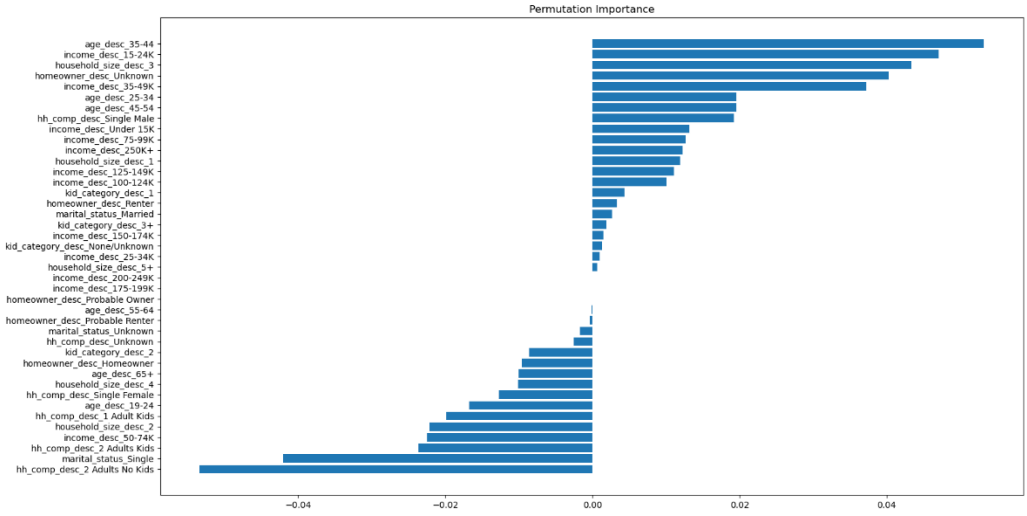


Figure 14

After performing hyperparameter tuning, my model performance greatly improved, with positive class recall of 0.72, and mean CV recall score of 0.75. There is now almost no overfitting, but there are a large number of false positives. Confusion Matrix, ROC curve, and PR Curve can be seen in Figure 15.

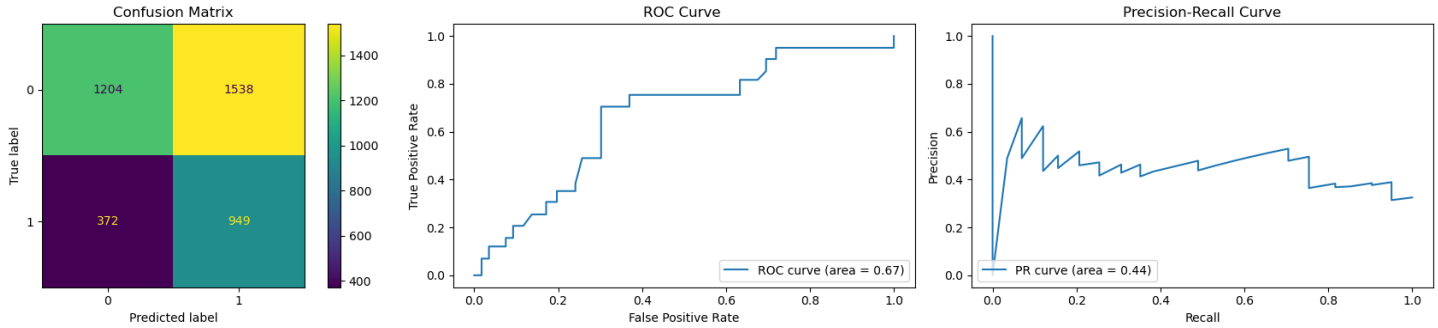
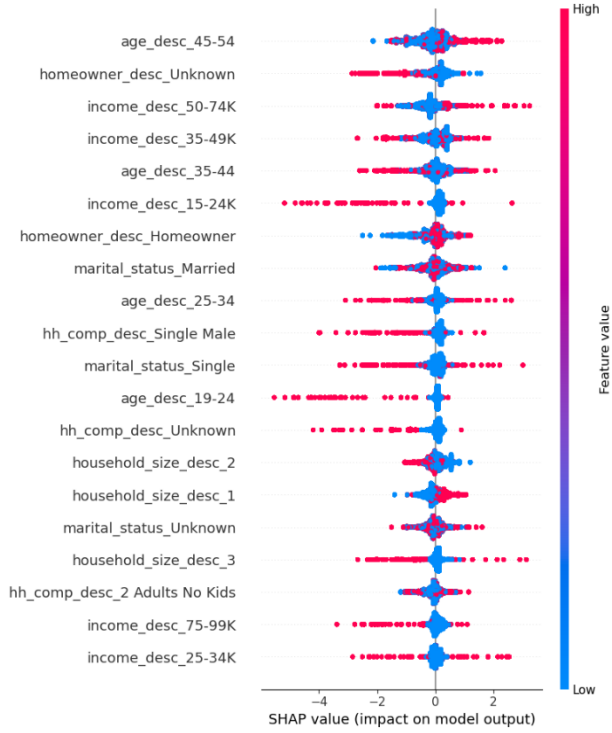


Figure 15

2.4.1 SHAP Analysis for XGBoost Classification Model

To perform SHAP analysis on the XGBoost classification model, I first ran the model on the entire dataset. To get metrics on model performance/cross validation, I calculated an estimated out of bag score of 0.143, which is a low error rate.

The SHAP summary plot and ranked feature importance values are in Figure 16. Unlike the previous sales prediction model, most of the features in this coupon prediction model have similar levels of importance, and in most cases, a single feature's value doesn't contribute much to the SHAP value.



	Feature	Importance
3	age_desc_45-54	0.442898
25	homeowner_desc_Unknown	0.416354
18	income_desc_50-74K	0.373590
17	income_desc_35-49K	0.367926
2	age_desc_35-44	0.345512
11	income_desc_15-24K	0.344020
21	homeowner_desc_Homeowner	0.328089
6	marital_status_Married	0.312629
1	age_desc_25-34	0.294148
30	hh_comp_desc_Single Male	0.290511
7	marital_status_Single	0.283317
0	age_desc_19-24	0.247352
31	hh_comp_desc_Unknown	0.233580
33	household_size_desc_2	0.228315
32	household_size_desc_1	0.224767
8	marital_status_Unknown	0.219150
34	household_size_desc_3	0.213640
28	hh_comp_desc_2 Adults No Kids	0.199943
19	income_desc_75-99K	0.193685
15	income_desc_25-34K	0.187499

Figure 16

However, the above SHAP summary plot in combination with the feature interaction plots did provide some useful information about how features and feature combinations contribute towards coupon use. As the model performance is quite good, the features clearly contribute towards model performance, but the interactions of the features must be more complicated than we can see on a 2 feature interaction plot. The results are detailed in the below '3. Findings' section.

2.4.2 Findings for the Coupon Use Prediction Model

The beeswarm summary plot of the features with the top SHAP values did not reveal many clear contributions towards coupon use. I did see that top contributors towards coupon use are ages 45-54, homeowners, and household size of 1. There were more features that contributed towards *not* using coupons: unknown homeowner status, income 15-24K, single males, 19-24 year olds, unknown household composition, and a household size of 2.

I also built feature interaction plots, so that I could see if there are combinations of features that jointly contribute towards coupon usage. Below I have listed the features and feature interactions that contribute towards and against coupon use.

Contribute towards coupon use:

- Income 50-74K who are renters
- Homeowners who make under 15K
- Ages 25-34 with 2 kids
- Household size of 1 who make under 15K
- Household size of 3 who make 15-24K

Contribute against coupon use:

- Ages 45-54 who make 35-49K a year
- Ages 35-44 with 2 kids
- Income 15-24K within the age bracket 45-54
- Married people with unknown homeowner status
- Single homeowners
- Income 15-34K who are age 35-44

3. Conclusions and Future Work

In this project, I conducted a thorough analysis of the 8 tables in the Dunnhumby Complete Journey dataset, identified the customer demographics with the highest coupon redemption rates, and segmented households into sales volume/sales growth quartiles. Upon looking at the weekly sales over time, I saw that while there is no visibly obvious seasonality, the autocorrelation plot indicated that there is a seasonal component. I did not further investigate this as of this time, but for future work I would like to go deeper into finding the seasonal trends, and make time series forecasts for weekly sales volume, both for all products and for the distinct product categories.

I created two models: An XGBoost Regression model for sales prediction, and an XGBoost Classification model for coupon use prediction.

I developed the sales prediction model on a 0.25 IQR subset of the data to address outliers that were introducing a lot of error. I was able to build a model with a mean sales value of $58, a MAE = 28.96 and a CV MAE = 29.07. When I ran the model on the entire dataset, the mean sales value was $80, a MAE = 45.93 and a CV MAE = 50.68. While the IQR = 0.25 model had very little overfitting, it did not generalize well to the entire dataset with outliers, and both model runs had relatively high error. Because of this, I would not use this model for predicting sales values, but it is useful for predicting *directionality* of sales, so I used it for SHAP analysis. From this analysis, I found that high coupon redemption rates and income over 125K a year were the strongest contributors towards customer spending. The complete list of features and feature combinations that contribute towards (or against) customer spending can be found in the 'Findings' section above.

In future work, I would like to find out what is different about the outliers, and to see if it makes sense to build separate models for customers who spend over 0.25 IQR. There are most likely external factors in common in different spending groups, and more accurate models could be built on segmented data.

The coupon use prediction model fortunately had better predictive results than the sales prediction model. After comparing multiple resampling methods to address class imbalance, I built an XGBoost Classification model using ADASYN oversampling,

and I removed the features with negative permutation importance values. I ended up with a model with 0.72 recall of the positive class (coupon users), with a CV recall of 0.75. There are quite a lot of false positives, which should be taken into account when basing any marketing campaign upon this model.

I also performed SHAP analysis on this model, but was only able to find features that clearly predict no coupon use. These include ‘unknown’ homeowner status, income 15-24K, age 19-24, and household size of 2. SHAP interaction plots revealed further relationships that can be seen in the above ‘Findings’ section, but overall the feature interactions must be more complicated than the level of analysis I performed.

4. Client Recommendations

I would recommend working with a marketing strategist to develop targeted marketing campaigns based on the identified customer demographics who are likely to spend more money and use coupons. I would also focus on the high growth/high sales volume customers in particular, as they are likely to respond well to marketing campaigns, although the high growth/low sales group could be a good area of focus as well.

The coupon use prediction model could be used to identify customers that are likely to use coupons, which would be the most effective customers to send coupons to. From there it would be possible to develop marketing campaigns targeted towards demographics that are likely to spend more money. This would decrease the cost of sending out marketing campaigns while increasing the response rates.