Project Report

Udacity Machine Learning Capstone Project

Starbucks Marketing Data

Aron Tharp

May, 2020

Problem definition

Overview of Project

Marketing data is a powerful resource for developing targeted marketing strategies to increase customer spending. We are provided with a simulation of data for users of the Starbucks mobile app and we will analyze this with the goal of determining which promotions are most and least effective. With many companies today having apps which support customer purchases, this is a large domain for businesses and provides significant potential for impacting business performance.

This project is one of the proposed capstone projects from Udacity, and all of the data used in this analysis was provided directly through Udacity. The customer information is simulated, but it resembles actual customer behavior. Some simplifications in the data exist. For example, we are only given information for purchasing amounts; we do not see any information related to product quantity or different product types. Data points for customer accounts are age, income, and date on which the account was created. Data points for customer behavior are purchases, receiving an offer (promotion), viewing an offer, and completing an offer.

In our project, we will focus on the transaction events and offer viewing events. We will assume that viewing a promotion is sufficient evidence of influence on the customers spending, and a further analysis of receiving or completing an offer is not considered here. The promotions vary in method of distribution, duration, and substance. A promotion may be BOGO, a discount, or informational.

Problem Statement

This is a classic project of evaluating marketing strategies. A simplification of the problem is that a marketing organization has applied promotions across a wide range of customers without knowledge of whether or not these promotions are successful and without targeting customers with specific promotions. The next logical step is to apply analytics to identify customer segments and determine effectiveness of promotions to develop a targeted marketing strategy to affect the company's bottom line. After clustering the customers into segments, linear regression will determine whether each promotional strategy had a positive, negative, or neutral effect on customer spending within a given segment.

It is anticipated that we will see different results of linear regression for different customer segments. Some promotions may be positive for certain segments and negative for others. Presumably, there will generally be a positive effect of the promotions on overall spending, however we may find a few negative impacts within our analysis. These positive and negative effects will directly guide improvements to the Starbucks' marketing strategy.

Metrics

The clustering algorithm will be measured based on Percentage of Variance Explained (PVE). An elbow graph will determine the appropriate number of clusters to use in our customer segmentation. For regression, each model will be compared to a null hypothesis to determine whether the variable of the promotion is statistically significant and we will look to disprove this hypothesis. Overall performance of each linear regression model will be measured with an adjusted R-squared value and Mean Absolute Percentage Error.

Analysis

Data Exploration

Inputs to this problem are 10 different marketing promotions provided in portfolio.json, basic demographic information of 17,000 customers in profile.json, and 300,000 interactions of the customers with the app in transcript.json. The 3 features of the different datasets are visualized in the 3 tables below.

Feature	Description
Age	18-118
Date Membership Began	Year-Month-Day for 5 years from 2013-07 to 2018-07
Gender	Female, Male, Other, or None
ID	Unique identifier of customer
Income	Dollar amount from $30,000 - $120,000

portfolio.json features

Feature	Description
Age	18-118
Date Membership Began	Year-Month-Day for 5 years from 2013-07 to 2018-07
Gender	Female, Male, Other, or None
ID	Unique identifier of customer
Income	Dollar amount from $30,000 - $120,000

profile.json features

Feature	Description
Event	Offer received, offer viewed, transaction, offer completed
Person	Unique identifer of customer - links to profile.json
Time	Hours since test started
Value	Unique identifer of offer - links to portfolio.json

transcript.json features

reward	channels	difficulty	duration	offer_type	id
10	[email, mobile, social]	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd

First row of portfolio.json

gender	age	id	became_member_on	income
None	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN

First row of profile.json

person	event	value	time
78afa995795e4d85b5d9ceeca43f5fef	offer received	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}	0
389bc3fa690240e798340f5a15918d5c	offer viewed	{'offer id': 'f19421c1d4aa40978ebb69ca19b0e20d'}	0
9fa9ae8f57894cc9a3b8a9bbe0fc1b2f	offer completed	{'offer_id': '2906b810c7d4411798c6938adc9daaa5', 'reward': 2}	0
02c083884c7d45b39cc68e1314fec56c	transaction	{'amount': 0.8300000000000001}	0

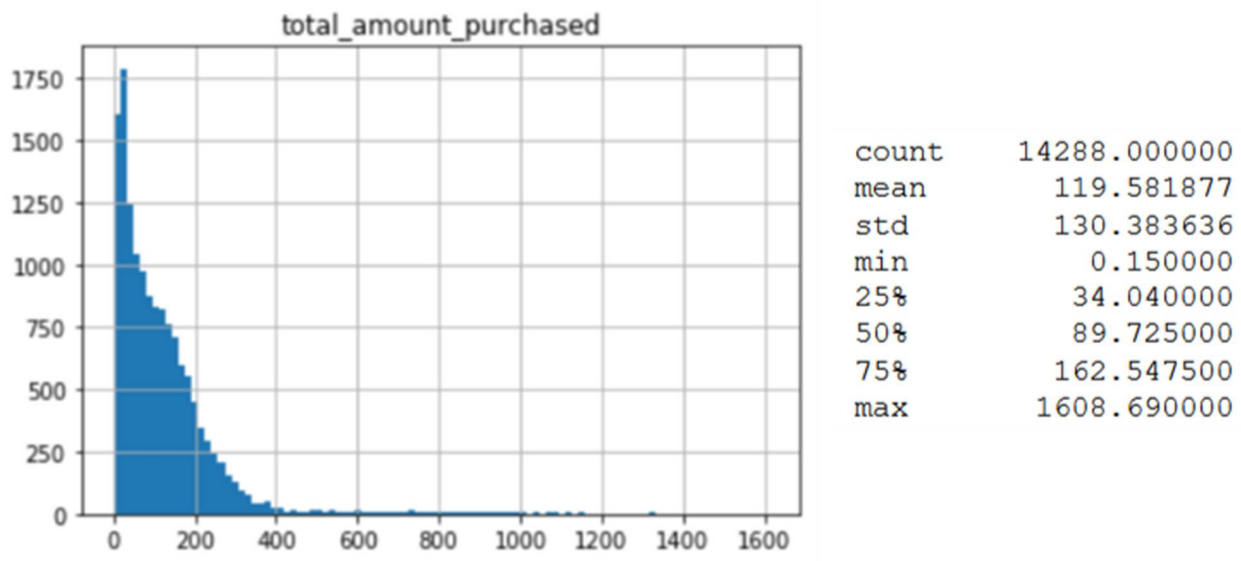
First row of each event type in transcript.json

The ID fields in portfolio.json are unique keys that will link promotions to the transcript.json file with its Value field. The ID field in profile.json will link the customers to transcript.json with its Person field. A few abnormal values exist in the profile dataset, namely in the age field. We will assume that our customers are not over 100 years old and will not consider data with values higher than 100.

After reviewing the files for missing data, this is only an issue in profile.json. There are accounts that appear to not have been populated by the user at the time of creation. Accounts that have missing data in any of the fields of age, gender, and income are missing in both the gender and income fields, and have an age of 118. Since customer segmentation will be a large part of our model, and the accounts with missing data only represent 13% of accounts, we will remove them from our data and they will not be considered in our results.

Transcript.json has unique formatting since each event type has different characteristics. Our model's predictor value will be the amount spent by customers from the transaction event type. For these rows, the value in the matrix is a dictionary with the value being the amount spent which must be extracted.

Exploratory Visualization



Histogram and descriptive statistics of total_amount_purchased by each customer

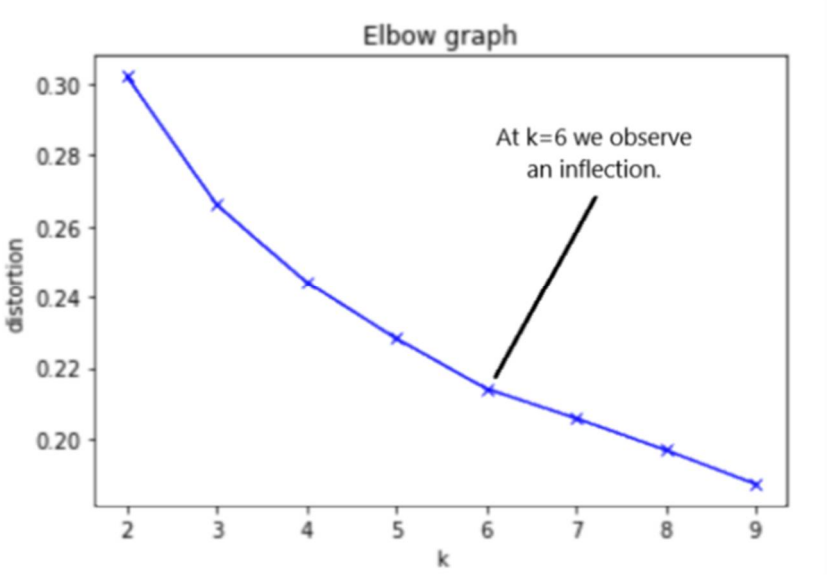
Our linear regression model will seek to determine the effects of the different features on the total amount purchased by each customer. A histogram and quantile analysis reveal that the majority of our customers purchased less than $100 during this timeframe. While there are some values much higher than 100, they appear to exist at a sufficient frequency to conclude that they are not incorrect data points. Due to the right skew of the data, an argument could be made to transform our dependent variable by taking its logarithm; however this would make the interpretability of the magnitude of the coefficients of regression less evident and will therefore not be used.

Algorithms and Techniques

Our process will be to perform a Principal Component Analysis (PCA) on the customer data and then feed the results into a K-mean algorithm to cluster our customers into distinct segments. Then, we will perform linear regression to analyze the effect of the promotions on the customer's spending according to the respective customer segment.

Our PCA will determine key relationships between features and rank them in order of importance. The PCA will serve as a preparatory step for our K-means clustering. We will follow the standard process of generating n-1 dimensions on our dataset. A benefit of PCA is a reduction in noise in variation. Another benefit is in speed of calculation due to a reduction in number of dimensions. In order to prepare our data for PCA, we will need to reduce our features to only numerical values. The gender feature does not support such a reduction, and we will resolve this by splitting the male and female data into separate datasets to analyze separately.

Our K-means algorithm will serve an important role in determining customer segmentation. The algorithm will search through the results of our PCA to identify clusters of customers. Since our PCA was performed separately for each gender, our K-means algorithm will also be performed separately for each gender. To identify the appropriate number of clusters, we will develop an elbow graph to identify an inflection point in distortion. In our data (graphed below), we find the ideal number of clusters to be six.



Comparison of different values for k to determine the optimal number

Now that we have distinct customer segments in our data, we can proceed to linear regression. Linear regression was chosen as a way to predict a numerical value – total amount purchased – based on the

linear relationship of our inputs. In order to capture the impact of each promotion, the promotions will be transformed into 10 distinct features with values equal to the number of times a customer viewed a specific promotion. Generally speaking, linear regression is a complex topic and the finer points are beyond the scope of this project and this course. Our implementation requires assumptions which are detailed later.

Benchmark Model

The benchmark for this model will be to naively analyze performance of promotional strategies through the same linear regression models as our complete model, without first doing customer segmentation analysis. This will provide validation of the usefulness of customer segmentation and should reveal that results of regression are more significant for the promotion variable when customer segmentation occurs. This benchmark will serve as a valuable reference point to see the effect of promotions on the overall dataset vs. individual customer segments.

Methodology

Data preprocessing (part 1)

Several steps of data processing were required to use our data in algorithms. The became_member_on field provided in the customer data is not in a clean format for our algorithms. The data is provided in string format as YYYYMMDD. Our algorithms require this to be in a numerical format. After converting the string into a "datetime" object, a new feature representing the number of months since the customer became a member was created called 'age of account' to replace our became_member_on feature.

We added two features based on spending behavior to increase the number of features for our clustering algorithm. These features were added since it is very likely that spending habits are indicative of a customer's responsiveness to a promotion, and their addition should improve overall model performance. We extracted data from the transactions on the number of customer purchases and the average amount of each purchase. The way that the data is formatted first suggests that number of purchases and total amount purchased would be logical features to add to the segmentation, however there is a high correlation between these features. By creating a third feature, amount per purchase, we eliminated the total amount purchased and resolved our correlation concern. These two features are added solely for the PCA and clustering algorithms and are removed before performing linear regression.

There are two sets of data points in our customer profiles which were disregarded in our implementation. The first, explained earlier, was in the case of customers who did not provide age, income, or gender upon account creation. The second is for customers who selected "other" under gender. While we feel that they are valuable customers whose spending should also be understood, the number of customers in our dataset who selected other was simply too small for us to make any meaningful analysis with the data we were given. Other represented barely 1% of the total data, only 200 rows, and therefore we chose to focus on the remaining data. Since we will generate many customer segments and use training and testing data slices, we cannot make meaningful analysis in this project. If more data were collected in the future from additional customers then analysis could be supported, but it would need to be at least a 500% or 1000% increase in data.

Data preprocessing (part 2)

Additional preprocessing was necessary between our k-means algorithm and linear regression. After adding the column for customer segment, we needed to link the customers with promotions which they viewed. First, the promotion IDs are long and incomprehensible strings. For our purposes, we renumbered them from 0 to 9 as shown in the below chart.

offer code	offer id
0	ae264e3637204a6fb9bb56bc8210ddfd
1	4d5c57ea9a6940dd891ad53e9dbe8da0
2	3f207df678b143eea3cee63160fa8bed
3	9b98b8c7a33c4b65b9aebfe6a799e6d9
4	0b1e1539f2cc45b7b9fa7c272da2e1d7
5	2298d6c36e964ae4a3e7e9706d1fb8c2
6	fafdcd668e3743c1bb461111dcafc2a4
7	5a8bc65990b245e5a138643cd4eb9837
8	f19421c1d4aa40978ebb69ca19b0e20d
9	2906b810c7d4411798c6938adc9daaa5

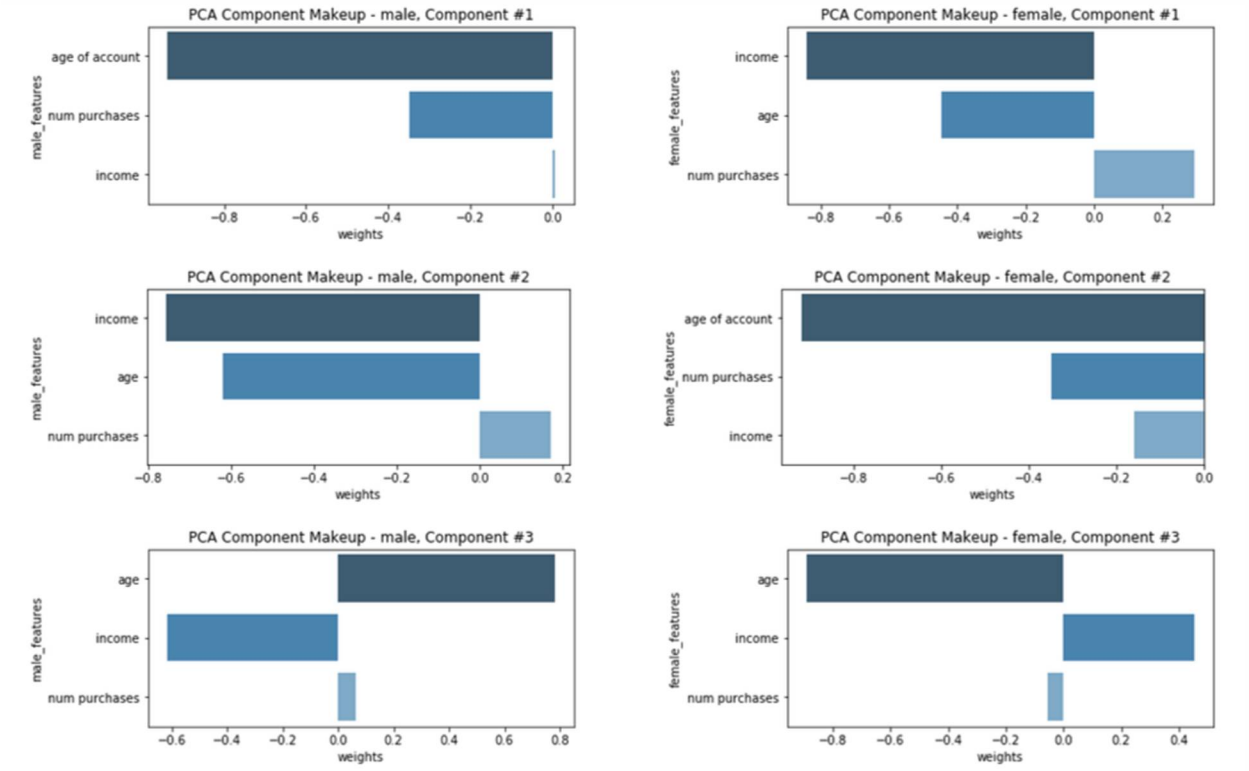
Key translating the promotions (offers) into a 0-9 coding

Then, we needed to count the views for each customer. 10 features were added to our data, one for each promotion, and the values of those features were the number of views per promotion. Within our data there were customers who did not view any promotion; this serves as an important base case, makes the coefficients in our linear regression more interpretable, and avoids overfitting in our regression models.

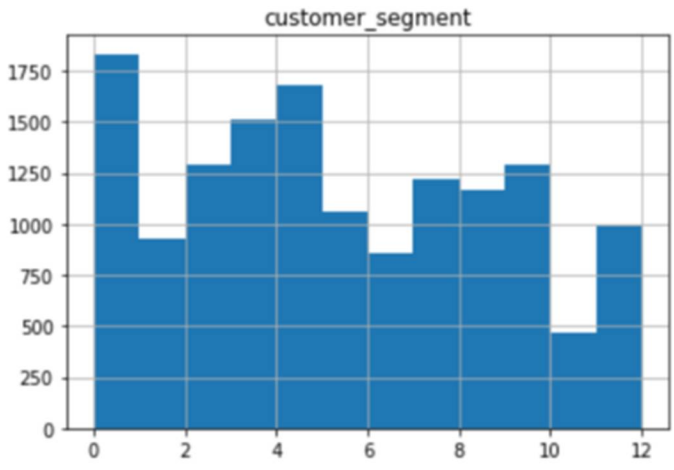
Implementation

The dataset input to our PCA was 5 features, and the output of our PCA was a dataset of 4 new features (or principal components) based on the key relationships between the 5 original features. In order to ensure that our features were equally weighted, each column was normalized to a scale from zero to one. A PVE analysis of our new features found that we could explain over 90% of the variance in both the male and female datasets with only 3 of the 4 principal components. 90% variance explained is sufficient for our purposes, and so the top 3 principal components were used for our K-means algorithm. A visualization of the top 3 components reveals some similarities across the genders in the relationships between the original features (visualization below). However, the data was unique enough and the volume of total data sufficient that we maintain the division between gender for further algorithms. A visualization of the top 3 components shows them ordered by PVE with male on the left and female on the right.

Our remaining 3 principal components are now ready to be input into our K-means algorithm. In order to determine the correct number of k clusters, we ran the algorithm with values for k ranging from 2 to 5. A visualization of the distortion did not reveal an ‘elbow’ in our data, therefore we ran it again with k values 6 to 9, finally revealing our optimal value for k as 6. In order to maintain our separation of male and female datasets, we numbered our 6 male customer segments from 0 to 5, and our 6 female segments from 6 to 11. A histogram of our customer segments reveal an relatively even distribution of customers between the segments, with the largest segments having approximately 1800 customers and the smallest having approximately 500 customers.



Top 3 principal components for each gender



Histogram of distribution of customers across customer segments

The last algorithm in our process is linear regression. In total, 13 regression models were created: 1 baseline model without customer segmentation and 1 model for each of the 12 customer segments. Due to the high complexity of linear regression being out of scope for this project and course, many assumptions were necessary for the purpose of this project. One assumption was that of linearity – we assume there is no polynomial relationship. Another assumption was that backward stepwise ordinary least squared (OLS) regression is the best method. This is one of many linear regression algorithms of which a comparison is out of the scope of this course. We also assume a significance level for p-values to be 95%.

This is a standard assumption that could be revisited in the future if the end-user was comfortable with a 90% or even 80% significance value. A significance value above 95% does not make sense in our scenario, as the consequences of inaccuracy are not as critical as other machine learning applications, such as those in healthcare.

Refinement

One way in which our models were refined was in the use of our elbow graph. When using a k-means model, too many clusters may reduce the usability and interpretability of a specific cluster, but too few clusters may have very different customers in the same cluster. This refinement determines an ideal balance between these two considerations.

Another refinement to our model was the addition of two features based on spending habits to our data before conducting the PCA. These features were further refined by reducing correlation between the quantity of purchases and the total amount spent. By using amount per purchase instead of total amount purchased in our PCA and k-means clustering, we were able to improve the accuracy of clustering.

	num purchases	total amount purchased	amount per purchase
num purchases	1	0.331584	-0.0940862
total amount purchased	0.331584	1	0.772835
amount per purchase	-0.0940862	0.772835	1

Correlation matrix showing the decrease in correlation by replacing the 2nd feature with the 3rd

Further refinement in our model could take place in a review our assumptions for linear regression. An argument against our linear assumption would be that income and disposable income often have a polynomial relationship and a higher disposable income could indicate higher spending. We could also explore forward stepwise regression or methods other than OLS.

Results

Model Evaluation and Validation

Our regression models were measured using adjusted R-squared on the training data and MAPE on the test data. These are two different ways to measure goodness of fit and they will be considered together. An important note on comparability, since the underlying data for each model is not the same, there is a lack of direct comparability of the results for each of these measures of fit across models. Still, these can be used for comparability if they are applied with a range of values. For our purposes, we will compensate for the difference in underlying data by making an assumption that the adjusted R-squared and MAPE must have a relative difference of at least 20% than the benchmark model in order for the model to be significantly better or worse. In our models, none of the individual models perform more than 20% worse on both adjusted R-squared and MAPE compared to the benchmark model. However, several models do perform more than 20% better, and the majority are within the 20% range. This is sufficient for us to consider the coefficients for the promotion variables in the individual models and, if they are considerably different than the coefficients for the benchmark model, then we will conclude that the individual models are more accurate and useful as our final conclusion.

	R2_adj	MAPE	Coeff_prom_0	Coeff_prom_1	Coeff_prom_2	Coeff_prom_3	Coeff_prom_4	Coeff_prom_5	Coeff_prom_6	Coeff_prom_7	Coeff_prom_8	Coeff_prom_9
baseline	0.54	217.1	1.4	1.0	10.8	8.5	5.8	15.1	26.4	NaN	10.2	16.3
seg_0	0.42	239.5	0.5	NaN	3.2	20.3	24.0	37.0	32.7	NaN	NaN	13.1
seg_1	0.59	76.9	42.6	4.1	NaN	54.3	NaN	45.1	40.0	32.5	60.0	36.2
seg_2	0.67	74.4	2.8	NaN	NaN	-2.8	NaN	24.9	14.2	16.7	20.7	13.4
seg_3	0.45	270.3	NaN	NaN	NaN	NaN	0.7	NaN	5.0	NaN	NaN	NaN
seg_4	0.48	211.8	0.8	3.7	13.8	9.6	23.3	NaN	20.1	NaN	23.0	25.1
seg_5	0.62	67.1	1.6	-1.6	20.5	19.3	31.9	19.5	31.6	24.8	30.7	9.2
seg_6	0.70	55.7	NaN	NaN	NaN	NaN	NaN	3.7	-2.8	NaN	36.1	NaN
seg_7	0.68	48.6	1.9	-2.1	NaN	17.6	32.8	25.6	32.4	NaN	21.5	46.5
seg_8	0.53	89.8	NaN	NaN	NaN	NaN	NaN	0.9	NaN	NaN	3.9	NaN
seg_9	0.52	250.8	1.5	NaN	3.1	NaN	16.0	NaN	NaN	NaN	NaN	19.6
seg_10	0.68	79.7	4.0	NaN	NaN	NaN	NaN	-3.9	NaN	NaN	20.5	NaN
seg_11	0.63	182.2	1.6	NaN	NaN	3.8	NaN	NaN	13.6	NaN	NaN	23.7

Matrix of adjusted R-squared, MAPE, and coefficients of each promotion for the baseline (benchmark) and segmented models

A quick glance shows that many of our models for individual customer segments are much more interesting than the baseline model. The amount of NaN coefficients, the magnitude of the numerical coefficients, and the existence of several negative coefficients strongly support this conclusion.

A NaN coefficient in our matrix implies than the promotion was considered to be insignificant. Or, more formally, we could not reject the null hypothesis that the given promotion had an impact on the y-variable (total amount purchased). For the Starbucks marketing department, this would imply that their promotion was a waste of time, effort, and money and should not be continued for the given customers. If we focus only on the benchmark model, we see that promotion 7 has the only NaN, which would suggest that promotion 7 should be canceled and the other kept (the others all had positive coefficients after all).

However, looking at the individual models provides much more interesting coefficients. We very quickly find that many segments are indifferent to many promotions. Reading the matrix column-wise reveals that every promotion is ineffective on at least 1 customer segment. We also see that promotion 7, which was insignificant when measured on the entire population, is now significant for 3 customer segments (with fairly large coefficients). Only one customer segment has no Nans - segment 5 – meaning they are responsive to all promotions. Many segments have many NaNs, with the most NaNs (row-wise) appearing in segment 8, for which only 2 promotions have impacts.

The interpretation of the magnitude of the coefficients is as follows - the larger the absolute value of the coefficient, the larger the impact of the promotion is on the total amount purchased by the consumer. If we remember from earlier, the base case is a customer who does not view any promotion. Focusing on the larger coefficients is the best strategy for increasing total amount purchased by consumers. The closer the coefficient is to 0, the less the overall impact. For those with a smaller magnitude, the marketing department may wish to evaluate the overall cost of the promotion (time, effort, and money) and may determine that those with the smallest coefficients are not worth the effort. In nine of our models we find coefficients greater than 20, however the coefficients for segment 3, 8 and 9 are low which may indicate a need to develop new promotions for those segments.

Another interesting observation is the difference in magnitude of coefficients for the same promotion across different models. For example, promotion 0 has a very small coefficient of 1.4 in our baseline model. However, in customer segment 1 the coefficient for the same promotion is very high at 42.6. While our baseline model may suggest that the promotion is barely worth the time and effort, our segmentation has indicated that this promotion is very impactful on customer segment 1.

Negative coefficients are especially interesting for the marketing department because they imply that the promotion caused a decrease in the total amount purchased by the consumer. This indicates that the promotion should immediately be canceled for a given customer segment as it is harming the company's revenue. We find 5 negative coefficients in our data. Fortunately, the negative coefficients' magnitudes are small indicating that their impact is also small, but they still should be immediately reconsidered.

Another evaluation of our model was the PVE metric for our PCA. As a reminder, we were able to explain over 90% of the variance by only using 3 of our 4 principal components, which allowed us to continue into the K-means clustering with only 3 components and still adequately explain the overall variance in the underlying data.

Justification

The improvement of our regression models by using customer segmentation is clear. We find that a promotion, which may appear useless without customer segmentation, is highly impactful for certain segments. We also find that a small amount of promotions are decreasing customer spending in some of our segments. We can comfortably develop a customer targeting strategy based on these findings and improve overall performance of future promotional activities.

If cost data were available for each promotion, a profitability analysis could further our models and dictate exactly which promotions are ideal for each customer segment. Since we are not provided with this data, we will make the assumption that coefficients greater than 20 are large enough to justify inclusion in a customer targeting promotion strategy. For promotions with NaN coefficients or positive coefficients from zero to 19.99, we suggest that the marketing department performs supplementary analysis to explore their usefulness and appropriateness. The immediate cessation of promotions with negative coefficients has already been discussed. A final targeting strategy would look as follows:

		Promotions to Target a Specific Segment
Segment	0	3,4,5,6
Segment	1	0,3,5,6,7,8,9
Segment	2	5,8
Segment	3	None
Segment	4	4,6,8,9
Segment	5	2,4,6,7,8
Segment	6	8
Segment	7	4,5,6,8,9
Segment	8	None
Segment	9	None
Segment	10	8
Segment	11	9

		Segments to receive a specific promotion
Promotion	0	1
Promotion	1	None
Promotion	2	5
Promotion	3	0,1
Promotion	4	0,4,5,7
Promotion	5	0,1,2,7
Promotion	6	0,1,4,5,7
Promotion	7	1,5
Promotion	8	1,2,4,5,6,7,10
Promotion	9	1,4,7,11

Targeting strategy showing best promotions for each segment