BBY Loyalty Customer Household Spending

By: Tom Herman
Bedding Bathing and Beyond (BBY)
December 15, 2024

Primary Objective: Predict household spending for prospective customers based on existing loyalty customer data and supplemental 3rd-party data
Secondary Objective: Provide profiles and insights around existing loyalty customers

Summary Findings: Used SEMMA workflow to build four machine learning models (both traditional and tree-based), of which XGBoost (an optimized implementation of gradient boosting machines) exhibited the best characteristics. The XGBoost model yielded predicted household spending for prospective customers, sorted by highest to lowest predicted spending.

Additionally, loyalty customer profiles and insights were provided, including customer locations and trends around spending patterns based on net worth, education, Experian personas and whether customers own computers and magazines in their homes.

YouTube Video Link: https://youtu.be/6R6QrdXbUFo

Sample:
Current in-house loyalty customer data was first enriched with supplemental 3rd party data, including consumer, donation, magazine and political information. The same enrichment was performed on the training and prospective customer data sets.

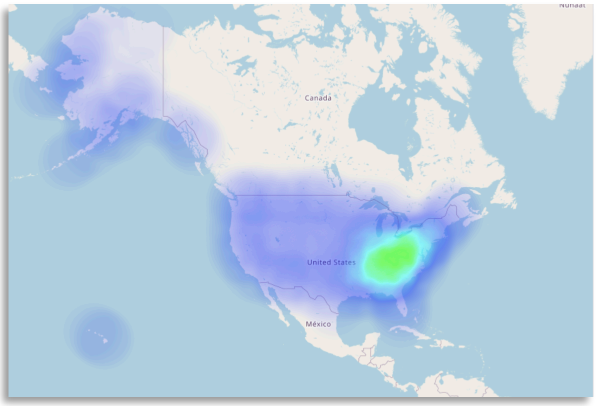
Next, sensitive and/or unethical attributes were removed, as well as columns that were more than 90%+ empty for customers. Additional cleanup was performed, including removing noise text from some columns, rounding ages to the nearest whole number, and rounding household spending amounts to dollars and cents.

Finally, some feature engineering was performed, including the creation of an age and education interaction (score of age times years education), and an overall "magazine enthusiast" rating based on number of magazine subscriptions in the home.

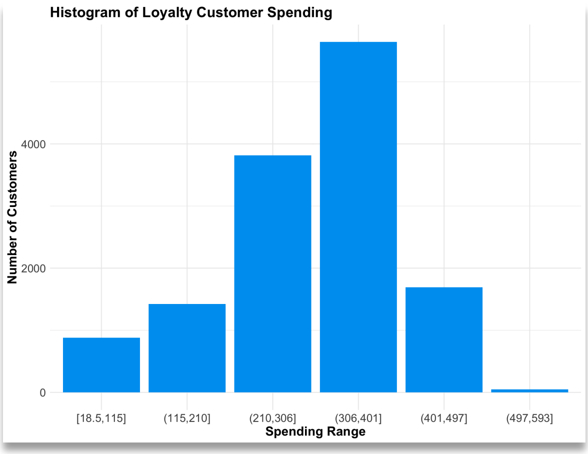
10% of the data (1,500 customers) was used for Vtreat rows, with the remaining 90% (13,500 customers) used for training data. Separate data sets were already provided for testing and prospective customers (5,000 and 5,992 customers, respectively).

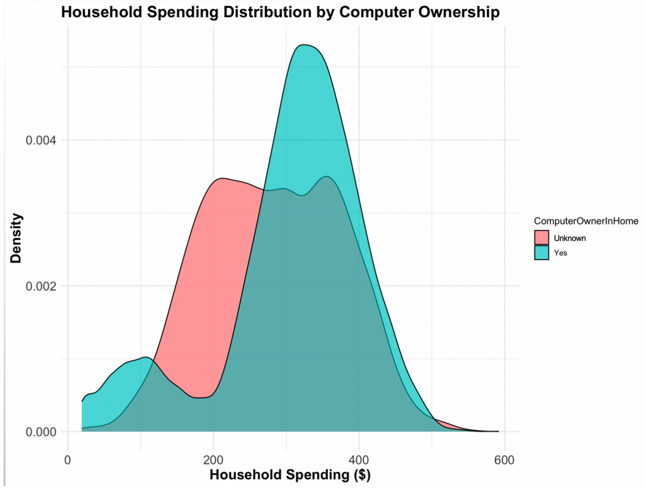
Explore:
The training data was explored for any interesting observations related to customer locations, net worth, education, Experian personas and whether customers own computers and magazines in their homes.

Current customers are dispersed across all 50 states, but there is a concentration of customers in the Ohio river valley area and eastern seaboard. There may be efficiencies for marketing in targeting these areas with higher customer concentrations.

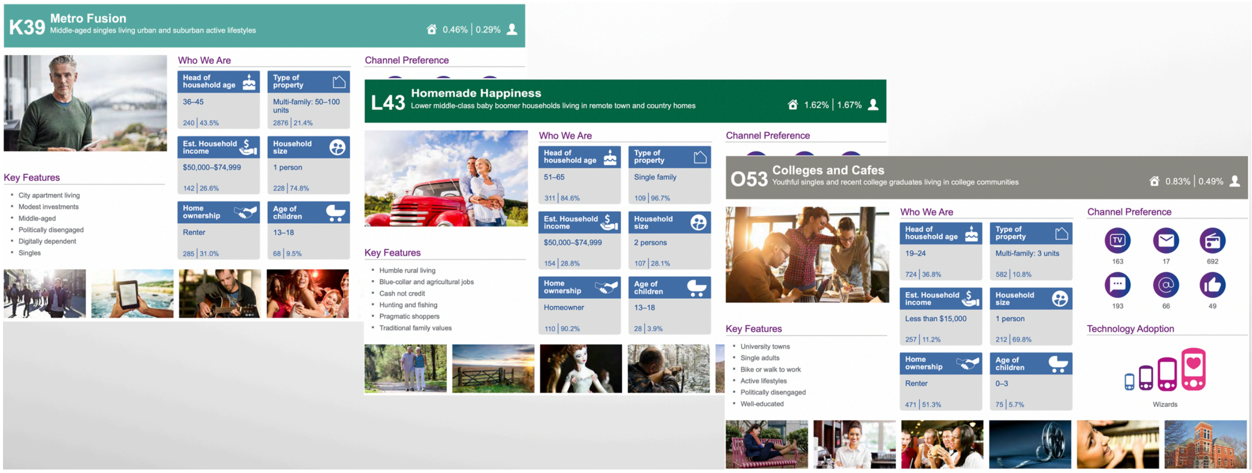


Most loyalty customers spend around $250 to $400, with a median of $300. Customer net worth and educational levels did not show meaningful differentiation, with similar median spending across all groups. However, customers with a computer in the home have a median spend of $41 more than customers whose computer status is unknown ($323 vs $282). Surprisingly, there was also a negative association (-.4 correlation) between household spending and the "magazine enthusiast" rating. In other words, customers with more magazines tended to spend less. This was also true specifically for customers with "Do It Yourself' magazines, which one would think would be associated with more spending on bed, bath and other home-related products. All exhibits are provided in slides 5-11 of the deck.





Finally, several Experian Mosaic personas displayed disproportionately higher household spending, namely "Metro Fusion," "Birkenstocks and Beemers," "Homemade Happiness," "Colleges and Cafes" and "Town Elders." Personas that still exist have been provided on slide 11 of the accompanying deck. Marketing can find these and all other personas here: [Experian Mosaic USA E-Handbook]



K39 Metro Fusion
Middle-aged singles living urban and suburban active lifestyles
0.46% | 0.29%

Who We Are

Head of household age	Type of property
36–45 240 | 43.5%	Multi-family: 50–100 units 2876 | 21.4%
Est. Household income	**Household size**
$50,000–$74,999 142 | 26.6%	1 person 228 | 74.8%
Home ownership	**Age of children**
Renter 285 | 31.0%	13–18 68 | 9.5%

Channel Preference

Key Features

- City apartment living
- Modest investments
- Middle-aged
- Politically disengaged
- Digitally dependent
- Singles

L43 Homemade Happiness
Lower middle-class baby boomer households living in remote town and country homes
1.62% | 1.67%

Who We Are

Head of household age	Type of property
51–65 311 | 84.6%	Single family 109 | 96.7%
Est. Household income	**Household size**
$50,000–$74,999 154 | 28.8%	2 persons 107 | 28.1%
Home ownership	**Age of children**
Homeowner 110 | 90.2%	13–18 28 | 3.9%

Channel Preference

Key Features

- Humble rural living
- Blue-collar and agricultural jobs
- Cash not credit
- Hunting and fishing
- Pragmatic shoppers
- Traditional family values

O53 Colleges and Cafes
Youthful singles and recent college graduates living in college communities
0.83% | 0.49%

Who We Are

Head of household age	Type of property
19–24 724 | 36.8%	Multi-family: 3 units 582 | 10.8%
Est. Household income	**Household size**
Less than $15,000 257 | 11.2%	1 person 212 | 69.8%
Home ownership	**Age of children**
Renter 471 | 51.3%	0–3 75 | 5.7%

Channel Preference

Technology Adoption

Key Features

- University towns
- Single adults
- Bike or walk to work
- Active lifestyles
- Politically disengaged
- Well-educated

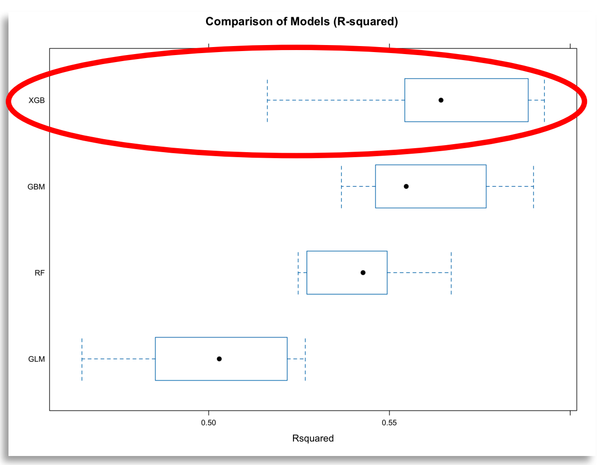
Modify:
The most informative features for the models were determined using an iterative process. First, a baseline Random Forest model was fitted on the training data, and then the VarImp() function was used to identify the least influential variables, which were then removed from the training data set. This was performed twice to remove noise and reduce the number of variables, resulting in a parsimonious model.

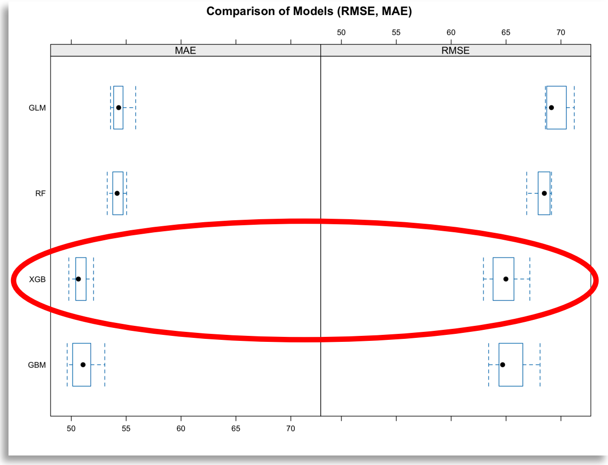
All three data sets were then treated using the Vtreat() function, where the outcome variable name was "HHSpend." These treated data sets were then used as a foundation for all four models in the following step.

Model:
First, a Random Forest model was fitted as a baseline model due to its versatility and ability to handle complex relationships while avoiding overfitting. A more traditional generalized linear model (GLM) was also created as an additional reference point and as an additional check for variable importance, which generally agreed with the variable importance as determined by the Random Forest model. Finally, two additional tree-based models were fitted, starting with a Gradient Boosting Machines (GBM) model for its ability to handle complex data sets with non-linear relationships. Since this model exhibited superior performance characteristics compared to the Random Forest and GLM models, a fourth model was constructed using XGBoost (XGB), which is an optimized implementation of GBM with enhancements that optimize speed and performance.

Note: In the interest of thoroughness, a K-Nearest Neighbors model was also fitted but not used because it showed a significantly higher RMSE value, which suggested it was not an optimal model for this data set. Also, an ensemble model was constructed using all four earlier models, but also not used because it showed similar – but slightly worse – performance metrics compared to the XGB alone.

The initial performance metrics (R-Squared, Root Mean Squared Error and Mean Absolute Error) on training data suggest that XGB was the most accurate model. In this case XGB exhibited the highest median R-Squared value, which means that approximately 55% of the variance in household spending can be explained by the features in the model. The model's RMSE of 65 means that the predicted household spending was off by $65 on average compared to the actual values. RMSE gives more weight to larger errors, unlike MAE, which simply measures the average absolute difference between the predicted and actual values. In this case, the MAE of 51 means that the predicted household spending was off by $51 on average compared to the actual values. With the highest R-Squared and the lowest RMSE and MAE, XGBoost appears to be the best model, but was further assessed in the next section.

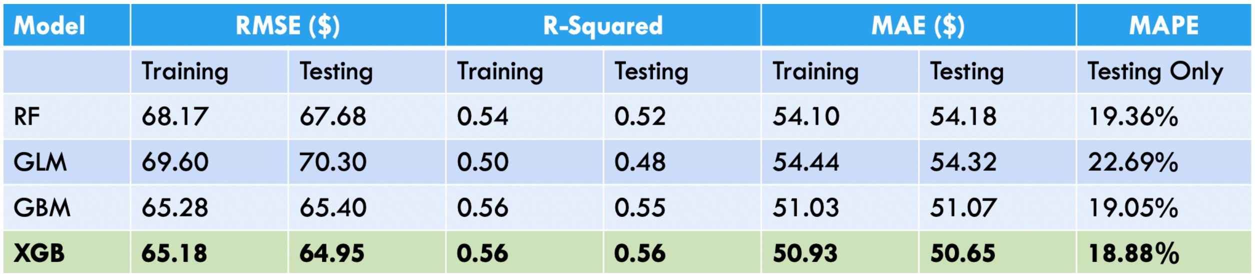




Assess:

On the following page we can see a comparison of all four models using four important metrics: R-Squared, RMSE, MAE and MAPE (the latter standing for "Mean Absolute Percentage Error"). As described previously, we want to maximize R-Squared and minimize RMSE and MAE. We also want to minimize MAPE, which expresses prediction errors as a percentage of the actual values.

As we can see, all models exhibit strong consistency between training and testing data, which suggests that none of the models are overfit or struggling with unseen data. However, there is a clear winner across all metrics: XGBoost.



Model	RMSE ($)		R-Squared		MAE ($)		MAPE
	Training	Testing	Training	Testing	Training	Testing	Testing Only
RF	68.17	67.68	0.54	0.52	54.10	54.18	19.36%
GLM	69.60	70.30	0.50	0.48	54.44	54.32	22.69%
GBM	65.28	65.40	0.56	0.55	51.03	51.07	19.05%
XGB	**65.18**	**64.95**	**0.56**	**0.56**	**50.93**	**50.65**	**18.88%**

Based on its consistency and strongest performance across the four key performance metrics, the XGBoost model was selected as the best model to apply to our prospective customers data and predict their expected future household spending.

Output:
The XGB model was applied to the prospective customer data, which yielded a column of predicted spending amounts. This information was joined with the original Prospects Data set (including all original in-house and third-party data in case useful for marketing), which was then reordered and to show the highest to lowest predicted spenders, ranging from $358.25 at most, to $76.21 at least. There are clear differentiation and targeting opportunities for marketing based on the predicted household spending and other customer insights provided in this analysis.

Finally, some additional exploration was performed on the final prospective customer data, which revealed that our prospects are largely concentrated in the same area (Ohio river valley and eastern seaboard). Some of the top spending Experian Mosaic personas remained the same as the current customers (Birkenstocks and Beemers, Colleges and Cafes), but the top three in the prospect set were: 1) Golf Carts and Gourmets, 2) Family Troopers, and 3) Boomers and Boomerangs. Finally, the proportion of prospects with computers remained approximately the same at 73%.

Conclusion:
Recap: The SEMMA data mining workflow was used to sample and explore the data from our current loyalty customers, modify the data for modeling, and then to assess the four best models that were created as a result.

The XGBoost model was used to predict the expected household spending for our 5,992 prospective customers. All of the data can be found in the accompanying CSV file titled “ProspectsPredictedSpending.csv,” with the prospective customers ranked from top to bottom according to their predicted spending amounts.

Key Takeaway: Marketing should consider how to segment and approach these different prospective customers based on their predicted household spending and other customer insights provided in this analysis.