Bedding Bathing & Yonder (BBY)

A2 Assignment

HOUSEHOLD SPENDING TREND ANALYSIS AT BBY

INSIGHTS AND PREDICTIVE MODELING

Suraj Udasi
Lead Marketing Data Scientist - BBY

OBJECTIVES

Overview of BBY and the project

The importance of understanding customer spending patterns

BBY offers a membership and loyalty program designed to give customers easy access to great benefits and discounts through email, and physical direct mail coupons tailored to the household shopping habits

Maximizing our customer’s revenue potential by offering relevant promotions is important to the company’s success,

We want to build accurate predictive models to modeled household revenue from a predefined household data set of our loyalty customers.

DATA OVERVIEW

Description of the data sets used and preparation

Loyalty Customer Data sets

1. **Membership Data**
2. Consumer Purchasing Habits
3. Household Donation History
4. Household Magazine Subscription history
5. Household Political Leanings

Pre-processing Rationale

Handling Missing values

Relevance of variables

Outliers

15,000
TRAINING SET OBSERVATIONS

5,000
TESTING SET OBSERVATIONS

~6,000
TESTING SET OBSERVATIONS

80
VARIABLES

THE PROCESS

1. Exploratory Data Analysis (EDA)

2. Feature Selection

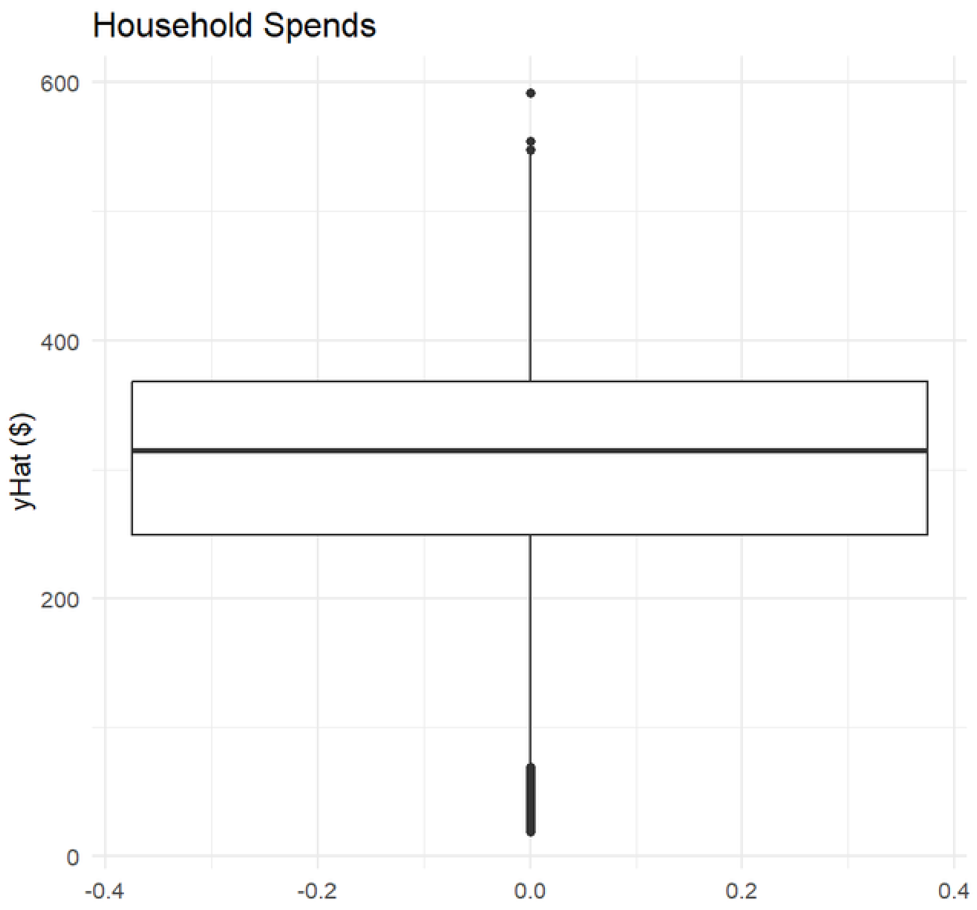
3. Model Selection

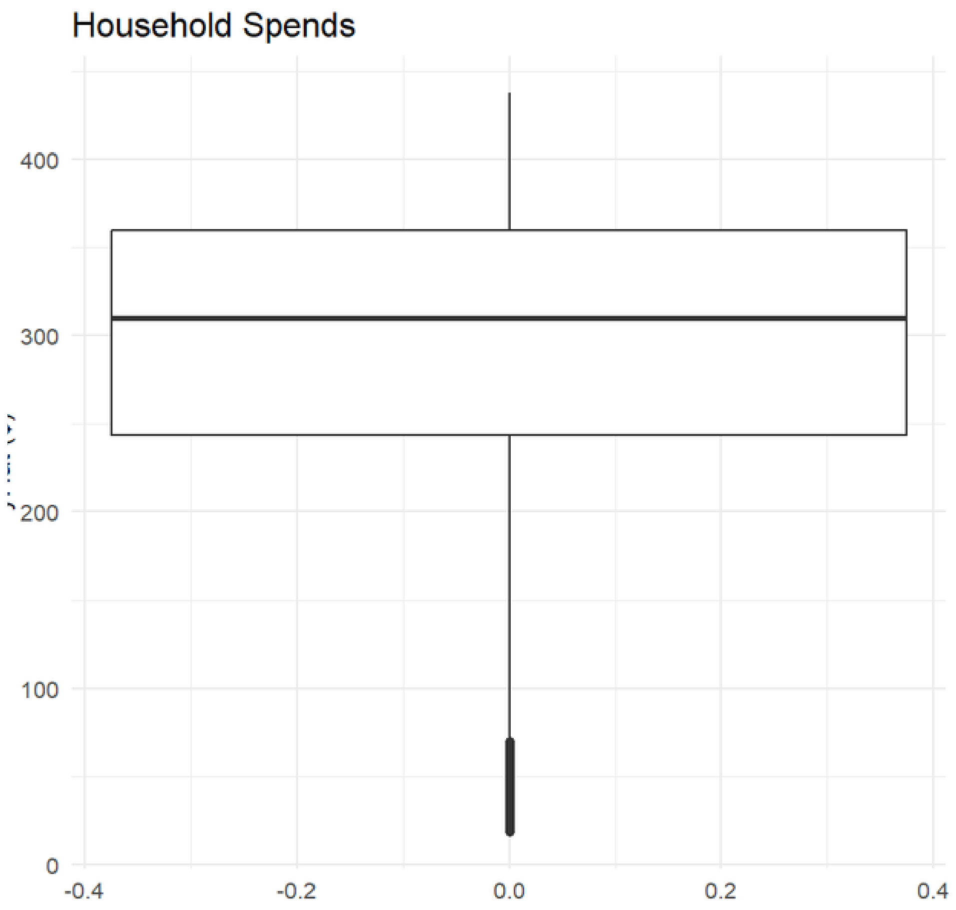
EXPLORATORY ANALYSIS (EDA)

Key highlights from EDA

Understanding the data

1. **Outlier Detection**



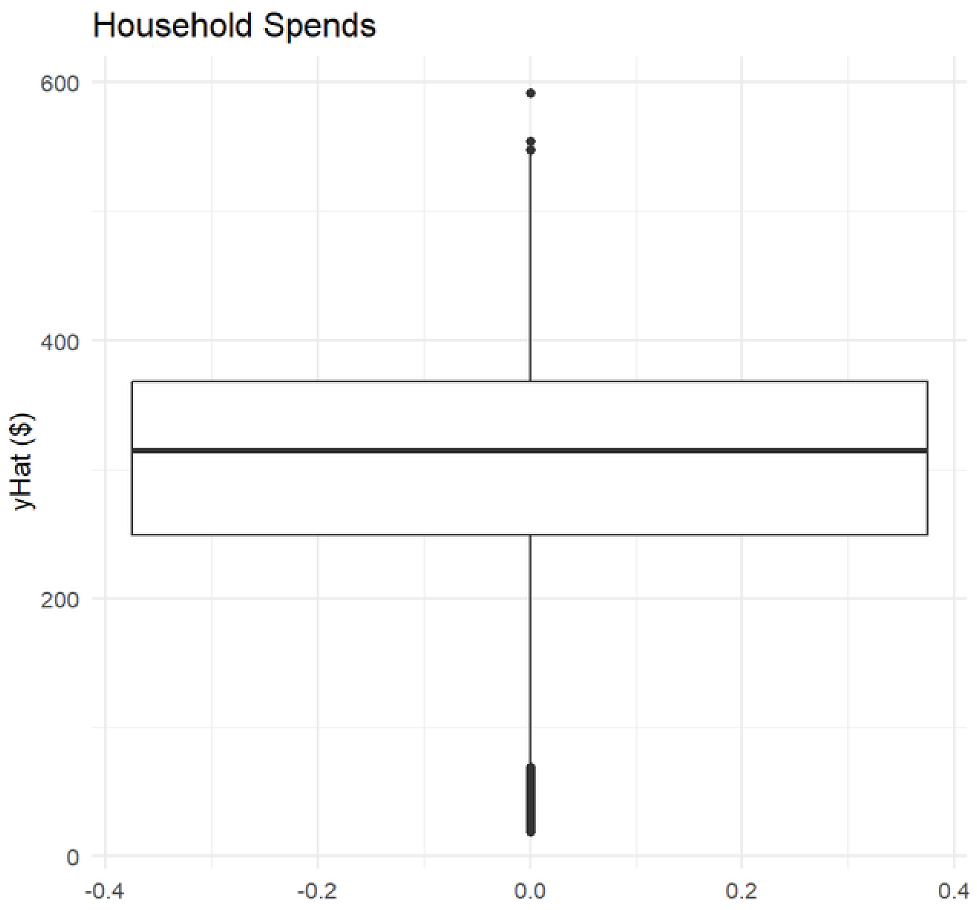


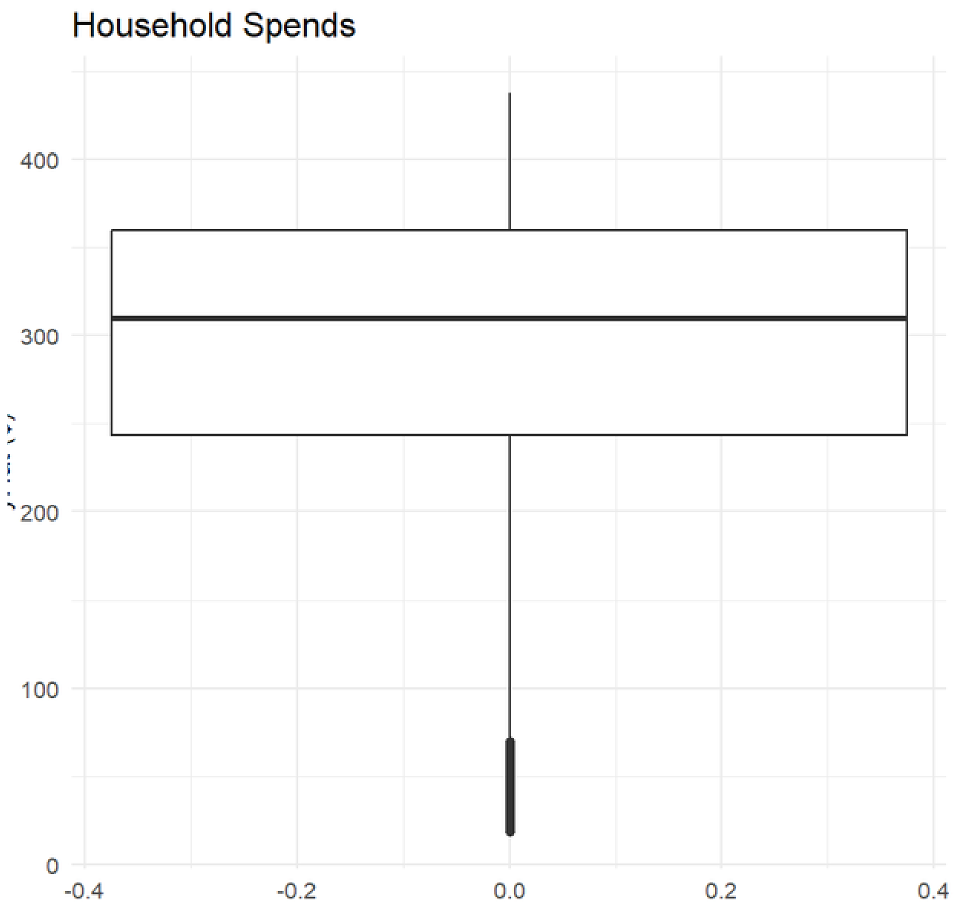
EXPLORATORY ANALYSIS (EDA)

Key highlights from EDA

Understanding the data

1. Outlier Detection





2. Numerical vs Categorical Variables

`EstHomeValue` & `LandValue`

$5,000

ORIGINAL CELL VALUE

▼
▼
▼

5,000

CORRECTED CELL VALUE

EXPLORATORY ANALYSIS (EDA)

Key highlights from EDA

Understanding the data

1. Outlier Detection
2. Numerical vs Categorical Variables
3. **Summarizing the Data**

11

NUMERICAL VARIABLES

69

CATEGORICAL VARIABLES

14,250

OBSERVATIONS

FEATURE SELECTION

Designing the data for predictive modeling

Addressing variables that were excluded from our analysis.

Set Criteria:

1. Atleast 50% of values must not be missing
2. Check for relevance of the features
    a. Uniqueness (`TmpID`)
    b. Value Add (`Veteran`)
    c. Ethical Concerns (`EthnicDescription`)

7

NUMERICAL VARIABLES

12

CATEGORICAL VARIABLES

14,250

OBSERVATIONS

MODEL SELECTION

Overview of the modeling methods employed

Linear regression, Decision tree, Random forest, XGBoost

Linear Regression

A fundamental statistical and machine learning technique where the goal is to model the relationship between a dependent variable (yHat) and explanatory variables independent variables.

We explore Linear Regression with all variables as well as the Parsimonious version

Decision Tree

A flowchart-like tree structure where an internal node represents a feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.

Prone to overfitting and often used as a baseline model.

Random Forest

Primarily used for classification and regression. It operates by constructing multiple decision trees during training and outputting the class that is the mean prediction of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set.

XG Boost

Extreme Gradient Boosting is an advanced implementation of gradient boosting algorithm. It's a scalable and accurate implementation of gradient boosting machines, designed to be highly efficient, flexible, and portable

MODEL PERFORMANCE

Based on KPIs

Fitting the first Linear Model

8.7%

R - squared

$89.4	**$88.5**
RMSE (train)	RMSE (valid.)
48.6%	**45.8%**
MAPE (train)	MAPE (valid.)

MODEL PERFORMANCE

Based on KPIs

Fitting the first Linear Model

8.7%

R - squared

$89.4	**$88.5**
RMSE (train)	RMSE (valid.)
48.6%	**45.8%**
MAPE (train)	MAPE (valid.)

Fitting the Parsimony Model

1.2%

R - squared

$93.2	**$91.9**
RMSE (train)	RMSE (valid.)
50.9%	**47.8%**
MAPE (train)	MAPE (valid.)

MODEL PERFORMANCE

Based on KPIs

Model Summary

Original Linear Model		Parsimony Model	
8.7% R - squared		**1.2%** R - squared	
$89.4 RMSE (train)	**$88.5** RMSE (valid.)	**$93.2** RMSE (train)	**$91.9** RMSE (valid.)
48.6% MAPE (train)	**45.8%** MAPE (valid.)	**50.9%** MAPE (train)	**47.8%** MAPE (valid.)

Broadening the parameters

- More lenient constraint
- No outlier detection
- Basic data cleaning only

MODEL PERFORMANCE

Based on KPIs

Original Model Summary

Refitting the Linear Model

Original Linear Model		Parsimony Model		New Linear Model with lower constraints	
8.7% R - squared		**1.2%** R - squared		**65.2%** R - squared	
$89.4 RMSE (train)	**$88.5** RMSE (valid.)	**$93.2** RMSE (train)	**$91.9** RMSE (valid.)	**$57.6** RMSE (train)	**$59.4** RMSE (valid.)
48.6% MAPE (train)	**45.8%** MAPE (valid.)	**50.9%** MAPE (train)	**47.8%** MAPE (valid.)	**22.6%** MAPE (train)	**23.0%** MAPE (valid.)

MODEL COMPARISON

Based on KPIs

Table comparing model performance metrics.

	Linear	Parsimony	Decision Tree	Random Forest	XG Boost
R - squared	65.2%	48.1%	71.0%	74.6%	74.3%
RMSE (train)	$57.6	$71.0	$50.6	$20.2	$39.0
RMSE (validation)	$59.4	$70.1	$53.4	$49.4	$49.3
RMSE (testing)	$80.1	$105.8	$88.9	$80.2	$85.8
MAPE (train)	22.6%	31.9%	18.0%	6.9%	12.6%
MAPE (validation)	23.0%	29.8%	18.4%	16.8%	16.4%
MAPE (testing)	32.6%	48.8%	34.2%	32.4%	35.4%

BEST PERFORMING MODEL

Deep dive

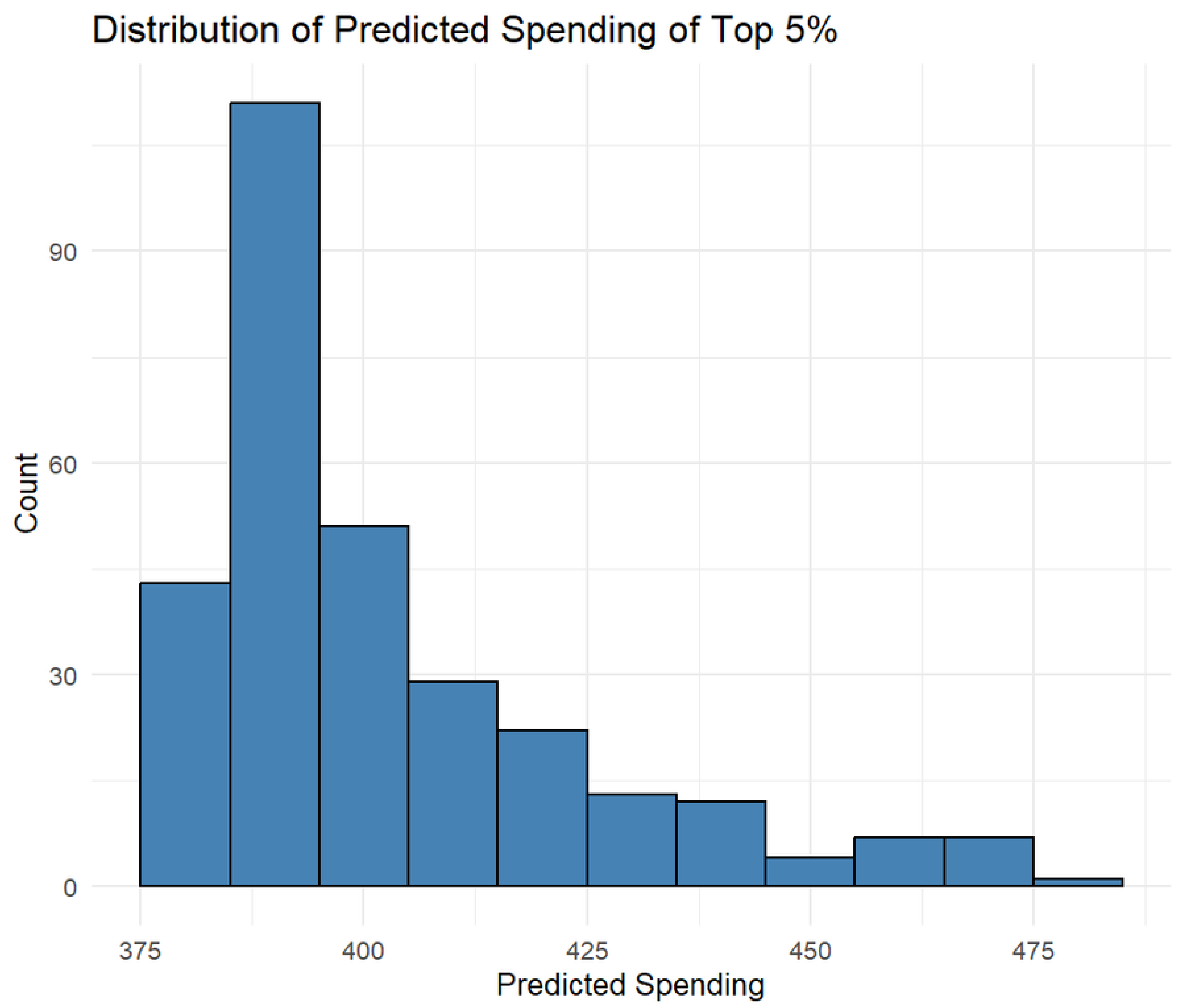
Interpretation of the model's key features and their influence

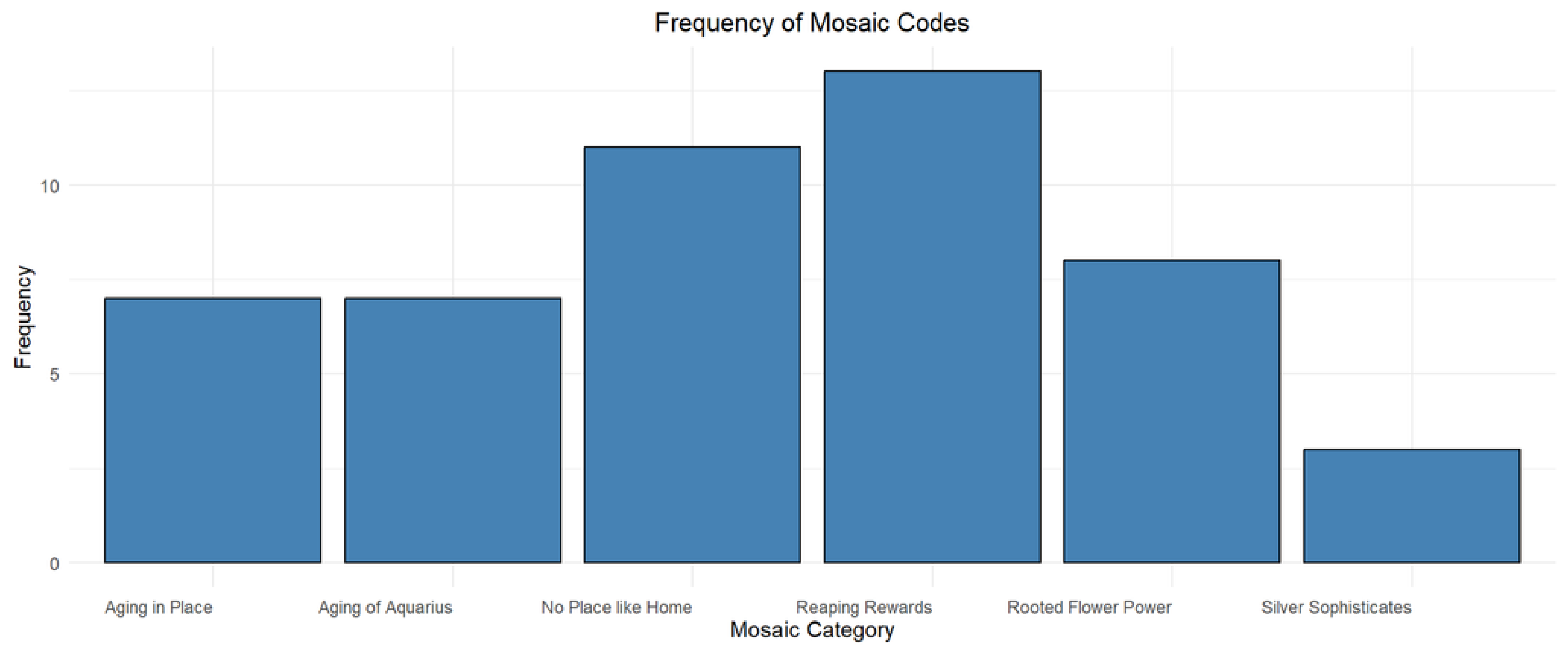
	Linear	Parsimony	Decision Tree	Random Forest	XG Boost
R - squared	65.2%	48.1%	71.0%	74.6%	74.3%
RMSE (train)	$57.6	$71.0	$50.6	$20.2	$39.0
RMSE (validation)	$59.4	$70.1	$53.4	$49.4	$49.3
RMSE (testing)	$80.1	$105.8	$88.9	$80.2	$85.8
MAPE (train)	22.6%	31.9%	18.0%	6.9%	12.6%
MAPE (validation)	23.0%	29.8%	18.4%	16.8%	16.4%
MAPE (testing)	32.6%	48.8%	34.2%	32.4%	35.4%

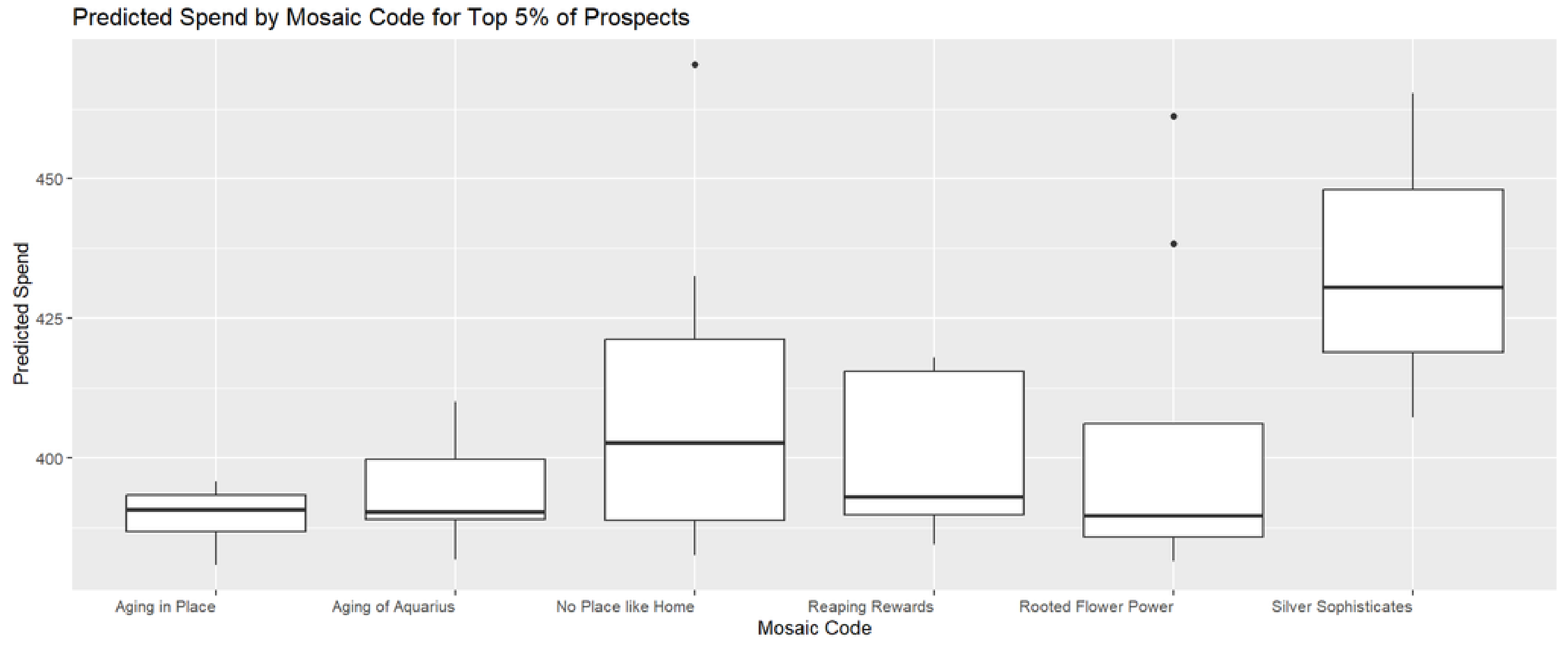
PROSPECTIVE ANALYSIS

Insights into prospective customer spending predictions.

Analysing the Top 5% of Prospects





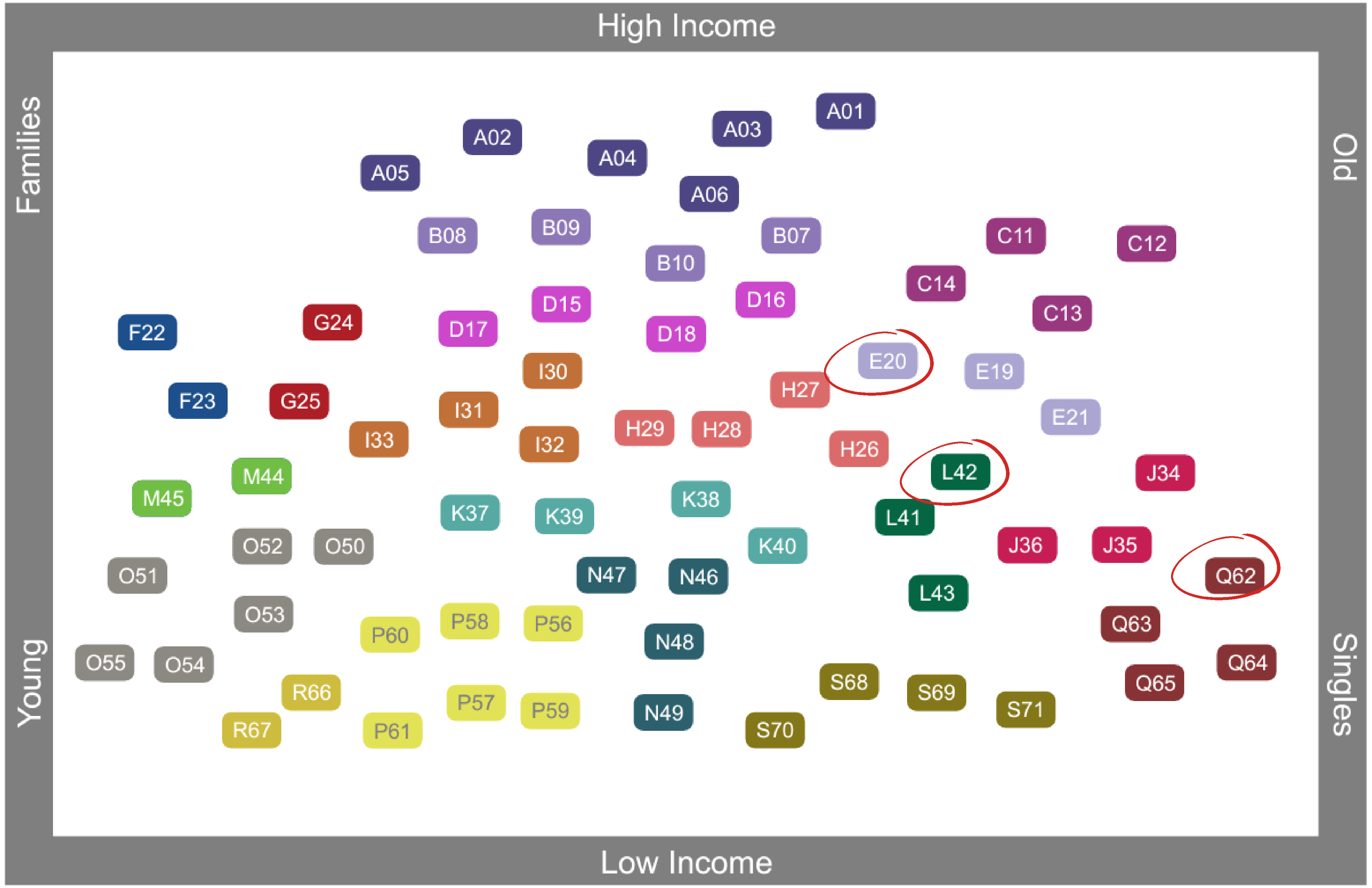


PROSPECTIVE ANALYSIS

Insights into prospective customer spending predictions.

Based on Mosaic Codes

The Mosaic USA family tree illustrates the major demographic and lifestyle polarities between the groups and types, and shows how the Mosaic types relate to each other.



1. Q62 - Repaing Rewards
2. E20 - No Place like Home
3. L42 - Rooted Flower Pot

PROSPECTIVE ANALYSIS

Insights into prospective customer spending predictions.

Q62 Enjoying Retirement

Relaxed, retired couples and individuals in suburban homes living quiet lives

1.78% | 1.37%

Who we are

Head of household age	Type of property
76+	Single family
629 | 61.1%	104 | 82.1%

Household income	Household size
$35,000–$49,999	2 persons
179 | 23.3%	138 | 34.4%

Home ownership	Age of children
Homeowner	7–9
127 | 81.4%	2 | 0.2%

Channel preference

TV	Mail	Radio
18	267	2

Text message	Email	Social media
17	109	11

Key features

- Retirees
- Established credit
- Cruise vacations
- Brand-loyal
- Traditional engagement
- Republican supporter

Technology adoption

Novices

THANK YOU

sudasi@student.hult.edu