ROBERT H. SMITH
SCHOOL OF BUSINESS

Project Report

Team 6 Members: Angela Chavez, Srushti Suresh, Danfei Li,
Manesht Al Rubaye, Sai Srujana Vepa

University of Maryland, College Park

BUMK746 - Data Science for Customer Analytics

We pledge in our honor that we have not given or received any unauthorized assistance on this Assignment.

Executive Summary

H&M is a global fashion retailer that sells a wide variety of clothing and accessories. For this project, the goal was to help the company predict how much a customer is likely to spend on their next purchase based on the types of products they have previously purchased.

To explore this topic, we used data from H&M with customer transaction history, product categories, and customer details. After cleaning and simplifying the data, we selected key product categories and calculated each customer's average order value (AOV). We then used stepwise regression, k-nearest neighbors, decision trees, and random forest regression, to predict the AOV based on customer's purchase histories and profiles.

The models provided valuable insights, with the random forest performing the best since it had the highest R^2 and lowest prediction error. It showed that product categories like trousers, t-shirts, and dresses, impacted node purity the most, and were important predictors of future spending. In all, these insights can significantly help H&M better target their marketing efforts, stock the right products, and focus their marketing efforts on high-value clients.

1. Introduction

In the trend-driven fashion industry, timely and accurate predictions of consumer preferences are essential for staying competitive. For a major retailer like H&M, identifying not only what customers buy but also how much they are likely to spend can offer powerful insights for personalization and strategy. Accurately predicting the AOV of a customer's next purchase can guide marketing efforts, enhance discount strategies, and focus resources on high-value customer segments. This project addresses the question: *What is the expected average order value of a customer's next purchase, based on their previous product category purchases?*

Solving this problem requires understanding both the categories of products customers purchase and the spending patterns that emerge over time. With thousands of unique transactions and detailed customer and product data, this project presents a real-world application of machine learning in revenue prediction.

2. Dataset Description

Our H&M dataset offers a detailed view of customer purchasing behavior on H&M's online platform, including historical transaction data, product metadata, and customer demographics. With millions of rows and a diverse mix of temporal, categorical, and demographic variables, it provides a valuable base to predict the order value of a customer's next purchase.

The original files contained over 10 million observations, making it much too large of a dataset to process and analyze efficiently. Therefore, after a preliminary analysis of the dataset, we modified the dataset by choosing to work with transactions from the latest 1 year of data and to

include a subset of 50,000 observations through random sampling. We then proceeded to choose the variables most relevant to our research question.

To improve interpretability and analytical clarity, we restructured the product categories based on existing variables in the dataset. We designed more intuitive and enhanced groupings that aligned with H&M's online site groupings, leading to analyzed results that could be more easily comprehended and implemented by management. We created 12 clear product categories that render more meaningful insights into customer purchasing behavior.

1. Accessories
2. Blouses, Tops & Shirts
3. Dresses, Jumpsuits & Sets
4. Hoodies & Outerwear
5. Knitwear
6. Nightwear
7. Shoes & Socks
8. Skirts, Shorts & Tights
9. Swimwear
10. T-shirts
11. Trousers
12. Underwear

Our independent variable utilized the computation of the average order value, calculated through finding the total number of purchases a unique customer had and the price of each item. Our model's key function was to predict a given customer's next purchase amount through analyzing their previous product category purchases.

3. Exploratory Data Analysis

Our exploratory analysis focused on understanding customer behavior patterns and product preferences that might influence AOV. We examined customer demographics, engagement metrics, purchase frequency, and spending patterns across different product categories.

The dataset includes customers across a wide age range (16-99 years), with a median age of 30 and a bimodal distribution showing peaks around ages 20-25 and 45-55 (Figure 1). This suggests H&M's customer base spans different generational groups, potentially with different spending behaviors and product preferences. Over 97% of members are active club members, and 40% are subscribed to regular fashion news updates (Table 1). These engagement metrics suggest varying levels of brand interaction that could correlate with spending patterns.

The purchase frequency distribution (Figure 2) is heavily right-skewed, with 75% of customers making fewer than 5 purchases. This purchase frequency distribution indicates a small segment of repeat customers accounts for a disproportionate share of transactions.

When examining the relationship between AOV and purchase frequency (Figure 4), we observe an inverse relationship where customers with fewer purchases tend to have more variable (and sometimes higher) AOVs, while frequent shoppers show more consistent, moderate AOVs. This may suggest that occasional shoppers make larger single-purchases, while regular customers make smaller, more frequent purchases. Our analysis of spending across product categories reveals distinct purchasing patterns: Categories like T-shirts and Trousers show the highest

purchase frequencies. Premium categories like Dresses, Jumpsuits & Sets demonstrate higher per-item spending.

All these findings directly informed our modeling approach, suggesting that:

1. Category-level features will be important predictors
2. Models capturing complex relationships would be more effective

These findings guided our feature engineering and modeling strategies, leading us to test both simpler methods like linear regression and more complex approaches like tree-based models that can capture the intricate patterns observed in the data.

4. Modeling Approach and Evaluation

To predict the AOV of a customer's next purchase, we tested several supervised regression models: Forward Stepwise Linear Regression, *k*-NN, Decision Tree and Random Forests. Each model was selected to balance interpretability, flexibility, and predictive accuracy.

4.1. Forward Stepwise Linear Regression

The forward stepwise linear regression was set as a baseline due to its simplicity, interpretability, and utility in feature selection. It incrementally builds a linear model by adding predictors that improve model fit, using Mallow's Cp to balance accuracy and complexity. Our initial goal was to identify a parsimonious set of relevant predictors while avoiding overfitting.

In our case, the model selected all 16 candidate predictors, indicating that each variable provided incremental explanatory value for predicting AOV. This suggests that the full set of behavioral, demographic, and category-level variables meaningfully influences the outcome. While this model achieved lower predictive accuracy compared to other models (Adjusted R^2 = 15.19%, MAE = 0.00811), it offers clear interpretability and supports feature relevance validation for downstream modeling.

4.2. *k*-Nearest Neighbors

The *k*-Nearest Neighbors algorithm provides a non-parametric approach to estimating the AOV. The algorithm captures local structure in the data and predicts the outcome based on the average of *k* closest data points. To support regression in our context, we used the knn.reg function from the FNN library. The MAE was minimized when 28 neighboring data points were used in estimating the average order value, with an R^2 of 21.89% (Table 3).

4.3. Decision Tree

Decision Trees are useful in creating interpretable "if-else" rules based on the variables. Not only do they capture potential non-linearities and variable interactions, but also provide a rather

simple visual that is easy to explain. To support regression in our context, we used the "anova" method within the rpart function.

An arbitrary tree was initially built and then pruned by minimizing the X-error to prevent overfitting. This resulted in an optimized complexity parameter value of 0.00071458 and a tree with 28 splits (Figures 4 and 5). The tree was able to capture around 24.38% of the variation in the AOV (R^2), with an MAE of 0.00736 on the test data.

4.4. Random Forest

This method was selected for its ability to robustly handle interactions, automatically rank variable importance, and resist overfitting, making it well-suited for datasets with a mix of behavioral, demographic, and categorical features.

As part of our modeling process, we tested different settings for the number of trees (specifically 60 and 100) in the forest to evaluate both performance and efficiency. The results showed that both configurations produced similar predictive accuracy, but using 60 trees offered a computational advantage with no meaningful drop in performance. We chose to move forward with the 60-tree model, which achieved an MAE of 0.00726 and an R^2 of 24.63% on the test set.

4.5. Evaluation and Selection

The Random Forest model proved to be a clear winner, outperforming the baseline regression model by 9.44 percentage points (Table 4). While all models had relatively low R^2 values and similar MAEs, the tree-based models outperformed classic regression and k-NN algorithms, highlighting the predictive advantages that machine learning algorithms can have over classic models that may be more interpretable with well defined functional forms.

A deeper dive into the performance of the Random Forest (Figure 6) reveals that the following categories of purchase contribute significantly to node purity, a measure of variable importance: Trousers, T-shirts, Dresses, Jumpsuits & Sets, and Accessories. Along with these categories, the average age of the customers also has a significant contribution.

5. Managerial Implications

The analysis we performed shows that predictive modeling, especially the Random Forest, could help H&M move from broad marketing and inventory tactics to targeted, data-backed actions. The model revealed that product categories like trousers, T-shirts, and dresses, along with customer age and purchase frequency, are strong indicators of future spend.

With these insights, H&M can personalize marketing by sending exclusive and customized offers (refer to Figure 7 for an example) or early access to high-value customers based on their preferred categories, while encouraging lower-AOV customers to increase their basket size with targeted bundles or discounts. Additionally, inventory decisions can be optimized: stores and

regions with higher predicted AOV for certain categories can be prioritized for new stock, reducing overstock and missed sales.

Moreover, predictive segmentation allows H&M to focus loyalty efforts and retention campaigns on customers most likely to drive future revenue, rather than just those who spent the most in the past. By integrating these models into planning and CRM systems, H&M can improve sales forecasting, campaign targeting, and overall operational efficiency. These data-backed strategies enable H&M to deliver more relevant experiences, optimize resource allocation, and strengthen customer loyalty in a competitive retail market.

Appendix

Table 1. Frequency Distribution of Customer Club Member Status

ACTIVE	LEFT CLUB	PRE-CREATE
97.998%	0.008%	1.994%

Table 2. Frequency Distribution of Customer Fashion News Frequency

Monthly	NONE	Regularly
0.030%	59.594%	40.376%

Table 3. Choosing Optimal *k* for *k*-NN

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MAE	0.010248	0.008979	0.008365	0.008125	0.008043	0.007884	0.007806	0.007735	0.007699	0.007651
R^2	0.066158	0.099082	0.125035	0.144282	0.149778	0.162741	0.170212	0.181531	0.185230	0.191474
k	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
MAE	0.007618	0.007599	0.007579	0.007549	0.007532	0.007536	0.007525	0.007518	0.007519	0.007516
R^2	0.195137	0.199198	0.202352	0.206967	0.208238	0.208192	0.209258	0.210127	0.210940	0.211549
k	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
MAE	0.007503	0.007513	0.007504	0.007502	0.007488	0.007492	0.007484	0.007479	0.007485	0.007483
R^2	0.213349	0.213847	0.214155	0.215474	0.218145	0.217394	0.218102	0.218924	0.218666	0.219069

Table 4. Summary of Model Performance

Note: adjusted R^2 considered for regression

Algorithm	**Mean Absolute Error (MAE)**	**R-Square**	**Difference from Baseline R^2**
Forward Stepwise Regression	0.00811	15.19%	0
k-Nearest Neighbors (*k*=28)	0.00748	21.89%	6.70
Decision Tree	0.00736	24.37%	9.18
Random Forest (ntree=60)	0.00726	24.63%	9.44

Figure 1. Distribution of Customers' Age

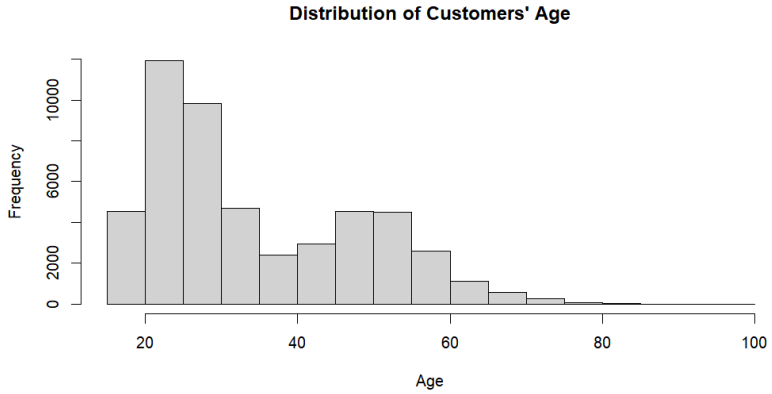

Figure 2. Distribution of Purchase Frequency

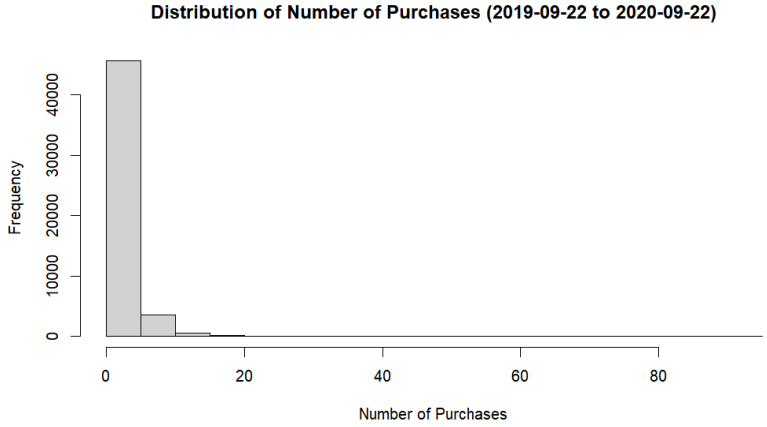

Figure 3. Correlation of AOV and Purchase Frequency

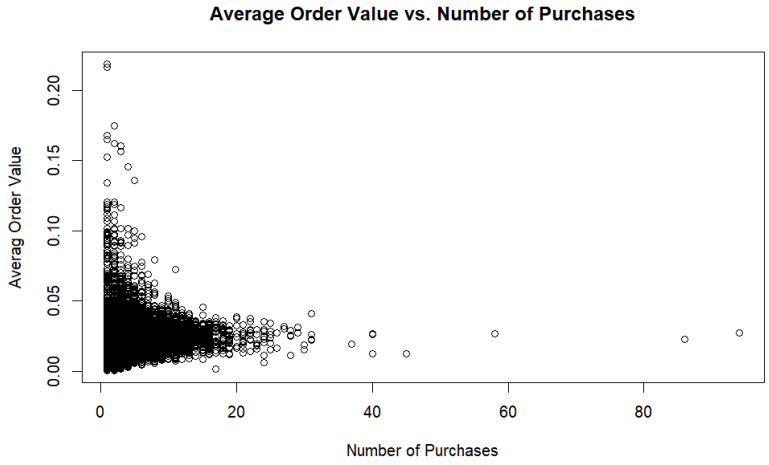

Figure 4. Choice of Complexity Parameter for Tree Pruning

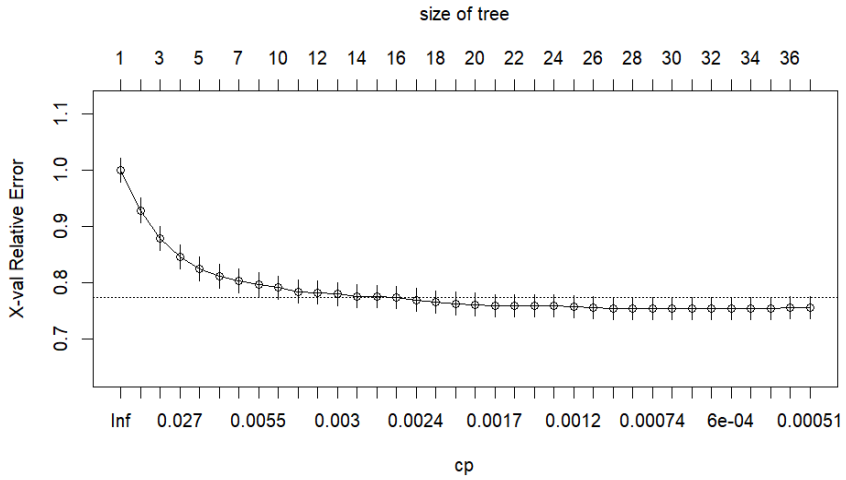

Figure 5. Pruned Decision Tree Structure

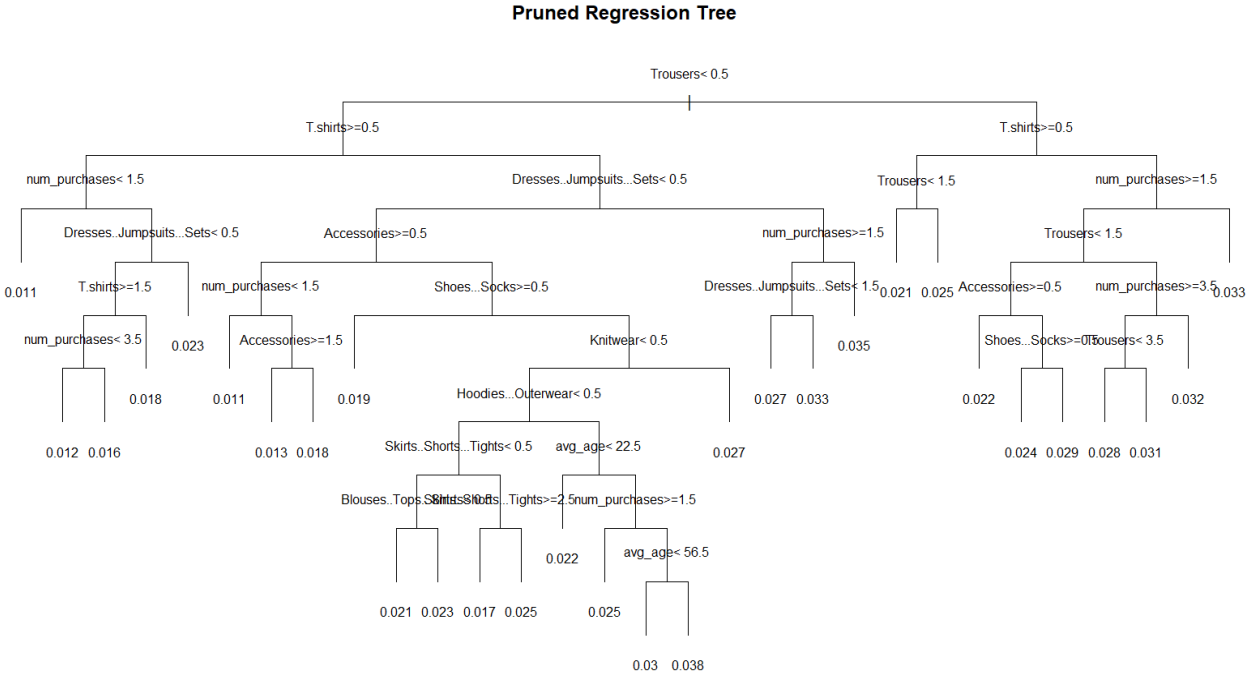

Figure 6. Variable Importance for Random Forest with 60 Trees

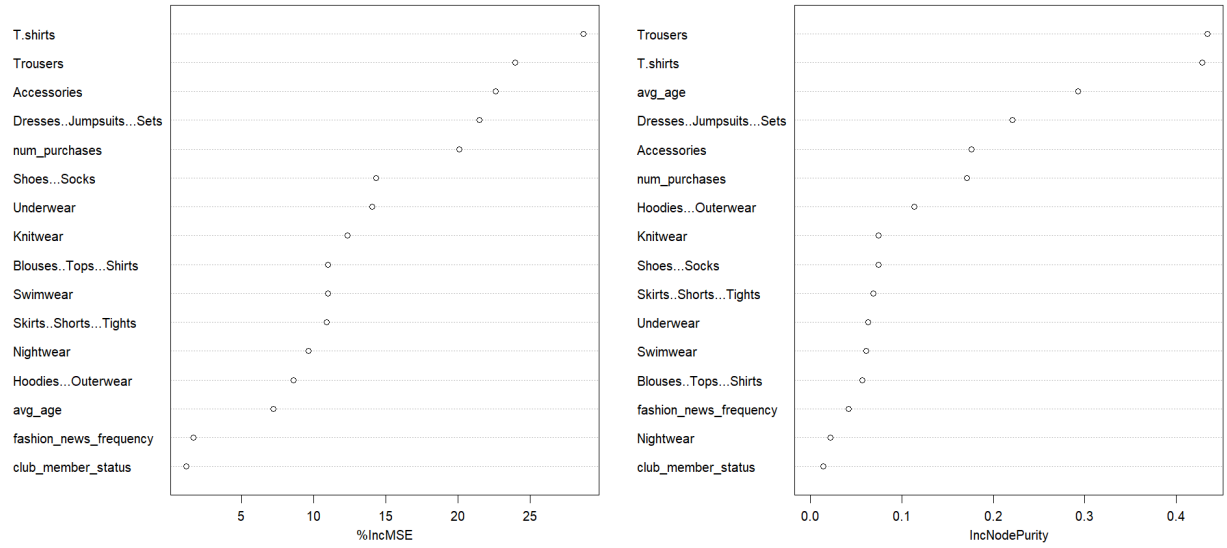

Figure 7. Personalised Marketing Communication Sample