### Predicting H&M Customer Value

*Natania Christopher, Roopam Chakrabarty, Helen Cunningham, Janine Mis*

## Background and Business Justification

A number of trends in recent years have contributed to the increasing value of customer loyalty programs. Firstly, digital advertising has gained market share at the expense of more traditional channels (McKinsey), and direct-to-consumer advertising has grown in popularity with the rise of digital and corresponding low barriers to entry into the market (McKinsey). This means more brands have crowded the market, and, without a loyalty program, switching costs for consumers are low.

Brands must be strategic in structuring a loyalty program, as it is a constant balance of cost to the business versus value to the consumer (McKinsey). A common problem for businesses is high participation in their loyalty program, but low purchase volume for each consumer (McKinsey). This is why it is important to use predictive analytics to determine the factors that truly drive purchase volume, and to identify consumers likely to purchase more in the future, so that they can be incentivized accordingly.

The first step in predicting customer loyalty is to dig into the factors that drive current customer purchasing behavior (Forbes). It is essential to make sure your loyalty program is relevant to your consumer so that the business does not take on unnecessary costs, and so that the customer derives value from the program as well (Forbes). A win-win loyalty program for customer and business can look like early access to sales or new products, which costs the business very little, but drives high value for customers (yotpo). Over half of customers indicate willingness to spend more with a brand they are loyal to, rather than a cheaper alternative, which highlights the importance of identifying and engaging with loyal customers (yotpo). Furthermore, the cost to drive a purchase from a loyalty program participant is typically lower than that to drive a purchase from a new customer.

Team 72 has chosen to focus on one retailer, H&M, to explore customer loyalty and purchasing behavior in the retail sector. Understanding each customer’s estimated value to the company would allow H&M to optimize many aspects of the business toward profit and sales volume. Customer Lifetime Value enables intelligent targeted marketing to optimize ROI, offering priority customer service to the highest potential value customers, value informed product decisions, and more. Tailoring efforts toward driving and retaining profitable, high-value business could reduce inefficient advertising spend, customer service expenses, and other waste.

Many other business decisions can be informed by customer value also. Marketers could target marketing to optimize ROI. Knowing a customer’s predicted future purchase behavior and value, H&M marketers can tailor spend to different groups. The firm can choose to exclude low-value customers from advertising (in channels with that capability) and seek out those that will bring in the most profit. When determining their product mix, H&M could use their understanding of the types of products being purchased by repeat purchasers, and high-value customers, to prioritize future product offerings that appeal to this group. Predicted sales and value could also be used to inform loyalty programs and customer retention efforts. The retailer may choose to prioritize higher-value customers in customer service queues since keeping them happy should have a larger, positive impact on sales. Another incentive would be to offer free shipping.

## Problem Statement and Objective

Retailers spend a great deal of time and effort in understanding how to get maximum value out of their customers. From loyalty programs to diverse product lines and more, there are many factors that affect how much customers spend and how often they shop at stores like H&M. The primary purpose of this analysis is to understand the future value of a H&M customer (as measured in $s spent OR quantity purchased with H&M). Our objective is to predict how much (in $s OR quantity) a current customer will purchase with H&M in the future based on their current attributes, such as demographic information and current purchasing behavior. Our secondary research questions are determining if “loyalty” can be used to identify high spending customers, and examining which customer attributes are most highly correlated with spend.

## Data Description

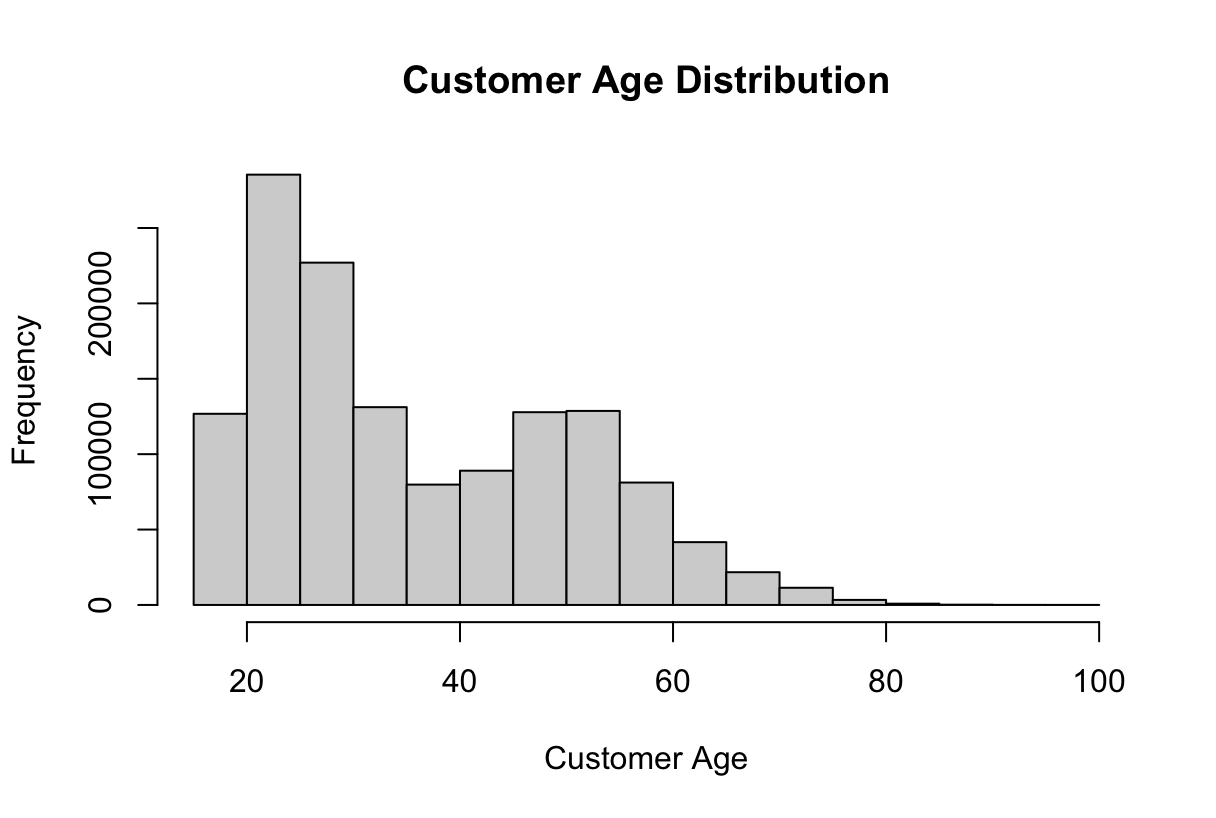
The H&M data, sourced from Kaggle, consists of multiple tables relating to H&M customers. There is a “customers” data table with demographic information, an “ articles” table detailing which articles of clothing are purchased by each customer, and a “transactions” table with information about customers’ purchases. The team merged these tables for analysis.

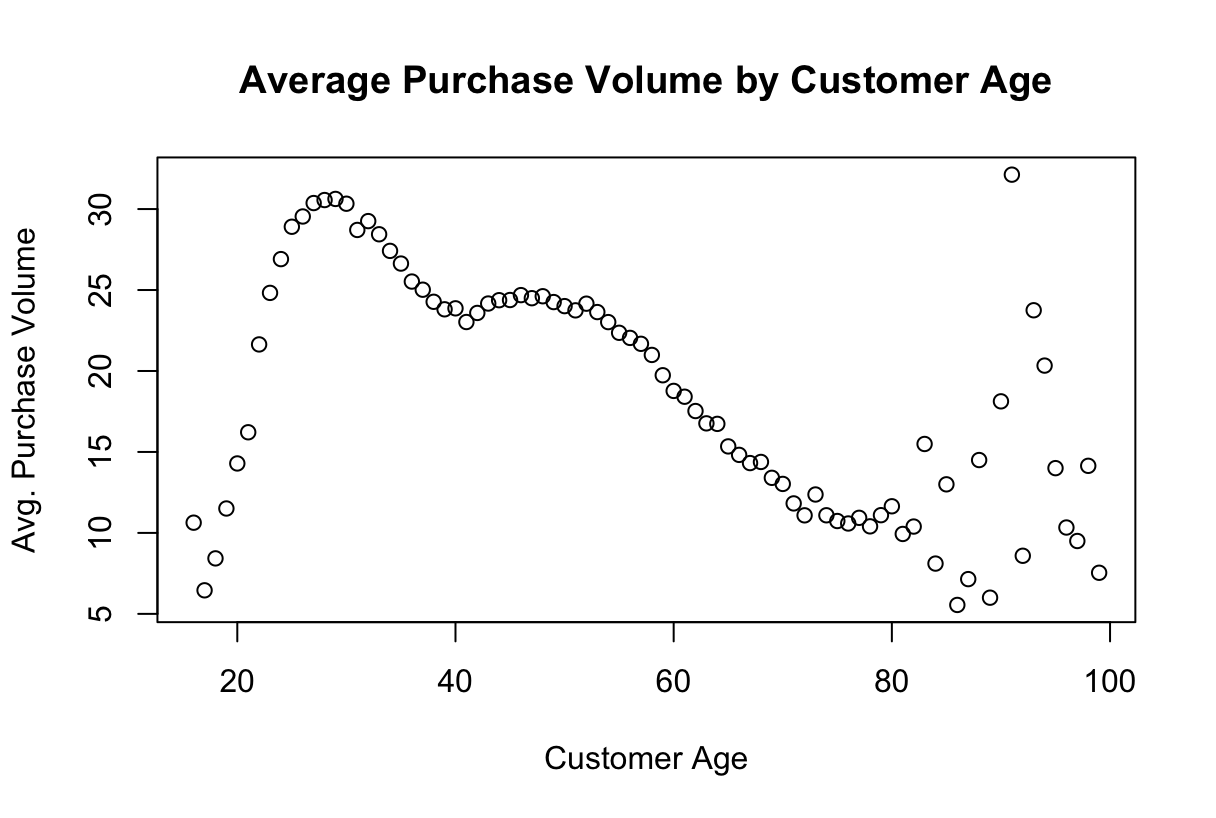
Key variables the team cleansed and transformed to be predictors in this analysis are: receiver of newsletter (Y/N), newsletter frequency, club member status (Y/N), club member stage, age, postal code, clothing category, # items purchased, # transactions per month, $ amount of purchases Sep 2018 - Feb 2019, and online vs. in-store medium. The variable used as a response in this analysis is $ amount of purchases or quantity of items purchased from Sep 2019 - Feb 2020.

The team chose to segment the data into two time periods of equal length, Sep 2018 - Feb 2019 (t1) and Sep 2019 - Feb 2020 (t2), for the purposes of prediction. Care was taken not to include any time periods that could be affected by COVID-19, as effects from global shutdowns and macroeconomic factors would make the true effects of predictors more difficult to determine.

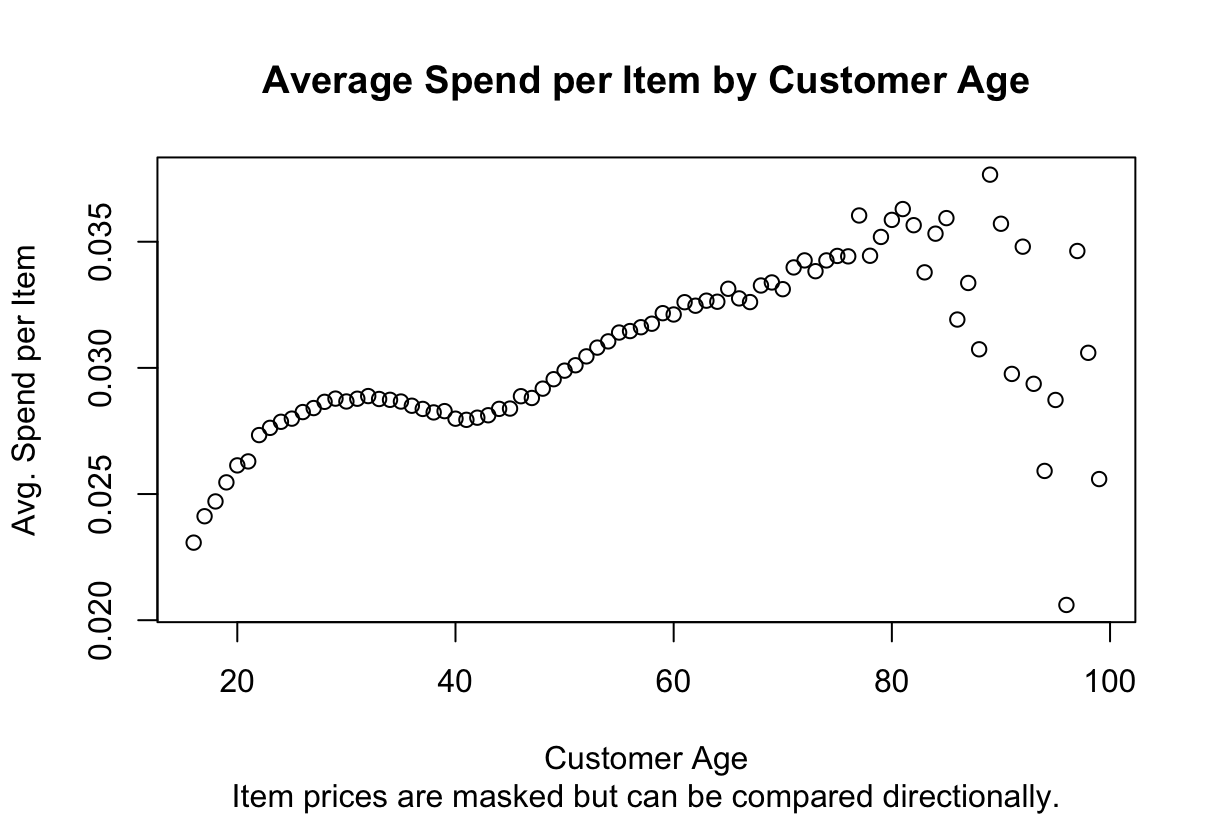
The hypothesis is that previous purchasing behavior will be a strong predictor of future purchasing behavior. It is also expected that club members will be more committed to the brand, and therefore would be likely to have a higher $ amount of purchases.

## Exploratory Data Analysis and Preliminary Models

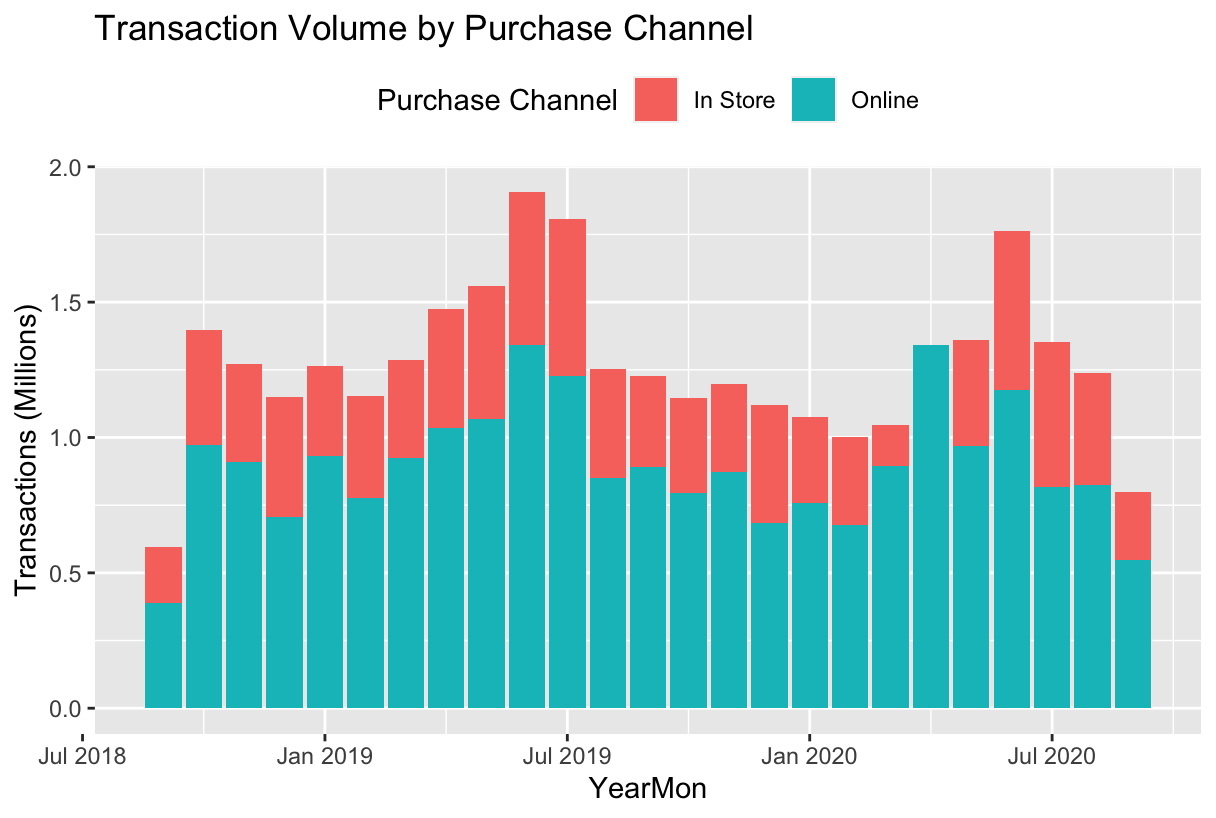
The team’s initial approach to exploratory data analysis was to understand the attributes of the customers first. A histogram was created to view the age distribution of the customers in the dataset. The two age groups most highly represented in the data are customers in their late teens through 20s, as well as customers in their 50s. H&M is primarily known as a brand that caters to young people, so the high volume of purchasers under 30 is unsurprising. The high frequency of purchases among those in their 50s could be due to customers in their 50s purchasing for younger children or relatives, or could be because customers in their 50s are looking to buy more “youthful” clothes as they age.



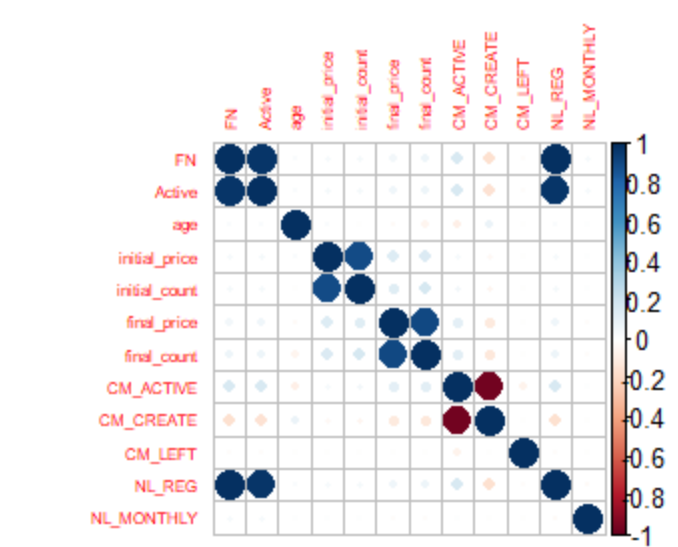
Beyond the simple age distribution of customers in the dataset, the team sought to understand how purchasing volume (as measured in number of items purchased) was correlated with age. From a scatterplot of purchasing volume vs. age, it can be seen that customers under 20 years old purchase an increasing volume as they age to 20. This trend is consistent with the understanding that in general, purchasers have more disposable income as they age. Customers between 20 and 40 purchase the highest volume compared to customers of other ages. This is the target demographic of H&M, so this observed trend makes sense. Customers between 40 and 60 tend to purchase a consistent volume, and customers over 60 purchase a decreasing value as they age out of the clothes, or as their children and relatives age out as well. Interestingly, the team observed messy data for customers over the age of 80. There is no trend to be observed, and the presence of multiple data points for ages close to 100 leads the team to be suspicious of data entered mistakenly or intentionally in the case of a customer unwilling to provide their real birthday. For this reason, the team decided to remove outliers from the age variable in addition to removing NA values during data cleaning.



Similar noise can be seen when plotting average spend per item vs. customer age. While a trend of increasing average spend with increasing customer age is seen up until age 80, customers over 80 do not show any consistent average spend behavior. Most intriguing to the team is that the relationship between age and purchase volume does not look to be linear. From the values of the spend per item, it is also clear that the data is masked in some way. Values cannot therefore be interpreted as literal prices, and instead must be compared only to each other to get a sense of directionality.



Sales data is known to be seasonal, and to test this assumption, the team plotted month vs. transaction volume. Color is used to indicate the channel where items were purchased. As a general trend, it is clear H&M sells a greater volume of items online than in store. Because early 2020 is included in the dataset, it can be seen that in-store sales were minimal in March 2020 and 0 in April 2020, which is consistent with the team’s historical knowledge of store closures due to the global COVID-19 pandemic. Interestingly, it seems that by June and July of 2020, in-store sales had increased back to pre-pandemic levels as customers wished to shop outside their homes when allowed again. Overall, it seems that sales are highest in the summer months: May - Jul. We do not see the spike we expected around the holiday season. One potential explanation is that H&M could be more popular for its summer clothing rather than winter clothing.

The transaction data frame was filtered for transactions in the initial period (Sep 2018 - Feb 2019) and the final period (Sep 2019 - Feb 2020). In these 2 periods, transaction price was summed and the number of transactions was counted for each customer and merged with the customer data frame on customer id. Any customers with age missing were removed; customers with “NA” values in FN and Active columns were replaced with 0. Outliers were observed in age (lower and upper values), prices (upper values only) in the initial period and final period, and the outliers were removed using the IQR method. Indicator variables were created for club\_member\_status and fashion\_news\_frequency. 

A correlation matrix showed strong correlation between FN and Active. Strong correlations were also observed with FN, Active and subscribing to fashion news regularly. Correlation was also observed between final price and final count, as well as between initial price and initial count. Negative correlation was observed between Club Member Status Pre-create and Active.

The newsletter and club member data could be more valuable for assessing H&M customer value if it included dates when statuses changed since the current data only exists for a point-in-time while the transaction data spans multiple years. If the data could be examined over time, a time series model could be appropriate to analyze trends in club member status signups, as well as correlation of newsletter subscriptions, specific newsletter releases, and sales. It would be interesting to explore how spending behavior changed for customers who were once not club members, and later became active club members.

The team believes geographical location and socioeconomic factors likely impact customer value, so an attempt was made to merge IRS salary data (based on postal code information found in the customer data frames) with the customer table, but unfortunately this was not possible. Ten different hashing algorithms were used to attempt to hash the IRS zip codes in the same way that the H&M customer data table zip codes are hashed. None returned a match, and therefore the team was unable to merge the IRS datasets with the H&M customer data table. In future analysis, if the team were able to determine the hashing algorithm used in the H&M data or if un-hashed zip codes were available, all of the metadata related to total annual wages, filing status, and income distribution by zip code could be incorporated into the analysis.

## Modeling Approach and Methodology

The primary approach was to build a linear regression model and then build a supplementary logistic regression to predict high-volume (“loyal”) purchaser/non-high-volume purchaser based on a threshold T. The loyalty threshold is determined by using the 75th percentile of spend for time period t, defined as Sep 2018 - Feb 2019. After merging the customer, transaction, and article tables of data, sums, aggregate fields, and indicator variables were created for interpretation. Indicator variables created as predictors for the linear regression model include: active club member status (1 if active, 0 if not), pre-create club member status (1 if “pre-create”, 0 if not), terminated club member status (1 if customer has left the club, 0 if not), regular newsletter readers (1 if receive newsletter “regularly”, 0 if not), and monthly newsletter readers (1 if receive newsletter “monthly”, 0 if not). Two aggregate variables used in the linear regression model as well were fields to aggregate spend per customer in two periods: Sep 2018 - Feb 2019 (t1) and Sep 2019 - Feb 2020 (t2).

The regression model used spend per customer in t2 as the response with the following predictors: spend per customer in t1, age, and all 5 indicator variables discussed above. The predictors significant at a .05 level are: customer spend in t1, age, active club member status, regular newsletter reader, and monthly newsletter reader. Predictors not significant at the .05 level are: pre-create club member status and club members who have left the club. The coefficient for age is -0.00016. This means an increase of 10 years in age only drives a -$0.0016 decrease in purchasing volume in t2. Even considering that prices are masked in this data, and the true magnitude is unknown, the coefficient for age is still more than 100x smaller than the other coefficients in the model. Therefore, we can discard this variable from the model.

A second linear regression model was created with only the significant predictors from the first model. The coefficient for customer spend in t1 is 0.6, which is the largest value of all coefficients returned by the model, meaning any change in customer spend in t2 has a greater effect than the same amount of change in the other predictors, all else held equal. A $1 increase in customer spend in t1 leads to $0.60 increase in customer spend in t2, all else held constant. All other coefficients of significant predictors are in the hundredths place. The model has an R^2 value of .3697.

Following the creation of two basic linear regression models, the team pursued a more complex method to attempt to increase the R^2 value. The merged data frame containing transaction and customer based on customer id was further divided into train (60%), test (20%) and validation (20%). Seven different linear regression models were created based on the significant variables. New polynomial variables were defined and products of variables were also introduced. The final model (Model#7) that had the square of price and the product of age and price showed 33% improvement compared to the initial model that had most of the variables based on the initial data analysis. The top 3 trained models with the highest value of adjusted R squared were used to predict the final price for the test data set, and it was observed that the final model (Model#7) had the least RMSE and MSE values. Lastly, a Random Forest model was created with the significant variables from the Linear Regression models. The MSE, RMSE and MAE for the Random Forest Model were lower than those of the highest performing Linear Regression Model (Model#7), but the MAPE was observed to be lower for the Linear Regression model (Model#7) than that of the Random Forest Model. It is important to note that MAPE is a relative measure of error that depends on the scale of the data and may not be directly comparable between models.

Additional linear regression models were created using additional variables created from merged customer, transactions, and articles datasets (joined on customer\_id & article\_id. Step-wise regression was used to select the factors for the model from the following possible factors (for each customer) from time t1 to predict customer spend in t2: total spend, number of purchases, # of days on which purchases were made, time between first and last purchase, days since last purchase, age, most purchased category, most purchased department, % of purchases that occurred online (vs. in-store), newsletter status & frequency, and club member status. The selected model included total spend, # of days on which purchases were made, days since last purchase, age, most purchased department, % of purchases that occurred online (vs. in-store), regular newsletter recipient (binary), and active club member (binary). The R-squared was 0.3993 and the Adj. R-squared was 0.3991, meaning ~40% of the variance in spend in t2 can be explained by the model. Running the model on the test set (30% of the data), MSE was 0.25, the RMSE was 0.5, the MAE was 0.29, and the MAPE was 5.82. The model shows that the most valuable customers are those that are younger, subscribe to the newsletter, are club members, have high pre-period spend, made purchases on numerous days, made purchases most recently, purchase online, and frequently buy ladies’ accessories.

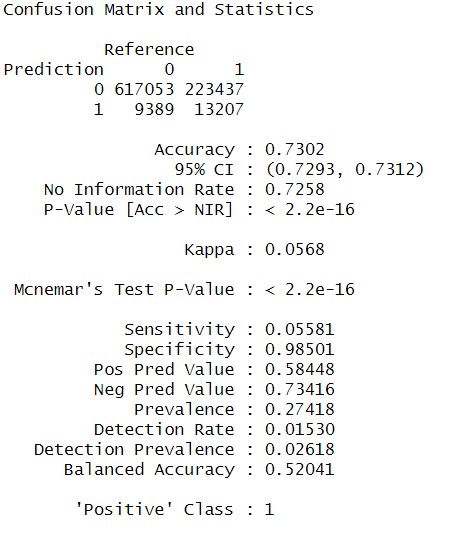
Adjusted R2 and R2 values of the linear regression models based on the training data:

| **Model #** | **Model Formula** | **Adjusted R2** | **R2** |
| --- | --- | --- | --- |
| 1 | lm(final\_price ~ initial\_price + initial\_count + age + FN + Active + CM\_ACTIVE + CM\_CREATE + CM\_LEFT + NL\_REG + NL\_MONTHLY, data=train) | 0.03817 | 0.03824 |
| 2 | lm(final\_price ~ FN + Active + CM\_ACTIVE + CM\_LEFT +NL\_MONTHLY, data = train) | 0.01783 | 0.01786 |
| 3 | lm(final\_price ~ initial\_price + age + FN + Active + CM\_ACTIVE +CM\_LEFT + NL\_MONTHLY, data=train) | 0.03802 | 0.03807 |
| 4 | lm(final\_price ~ Sq\_i\_price + Sq\_age + FN + Active + CM\_ACTIVE +CM\_LEFT + NL\_MONTHLY, data=train) | 0.05 | 0.05005 |
| 5 | lm(final\_price ~ Sq\_i\_price + cube\_age + FN + Active + CM\_ACTIVE + CM\_LEFT + NL\_MONTHLY, data=train) | 0.05004 | 0.05009 |
| 6 | lm(final\_price ~ Sq\_i\_price + cube\_age + FN + Active + CM\_ACTIVE + NL\_MONTHLY + age\_price, data=train) | 0.05078 | 0.05083 |
| 7 | lm(final\_price ~ Sq\_i\_price + age + FN + Active + CM\_ACTIVE + NL\_MONTHLY + age\_price, data=train) | 0.05094 | 0.05099 |

MSE, RMSE, MAE and MAPE values for the best 2 Linear Regression models and Random Forest model on the Test data frame:

| **Model** | **MSE** | **RMSE** | **MAE** | **MAPE** |
| --- | --- | --- | --- | --- |
| Linear Regression Model 6 | 0.0178 | 0.1334 | 0.1028 | 0.9558 |
| Linear Regression Model 7 | 0.0178 | 0.1334 | 0.1028 | 0.9555 |
| Random Forest | 0.0166 | 0.1290 | 0.0978 | 0.9696 |

To pursue the secondary research question regarding customer loyalty, the team first created a variable “Loyal” defined as 1 for customers who purchased 6 or more items in t2, and defined as 0 for customers who purchased fewer than 6 items. The threshold of 6 items was determined by reviewing the summary of the count of purchases per customer in t2 and determining the 75th percentile of items purchased per customer. “Loyal” therefore represents the top quartile of purchasers in t2.

A logistic regression was run with Loyal as the response variable, and age, club member status, newsletter reader status, $s spent in t1 and quantity purchased in t1 as predictors. After analyzing various cutoff values for the logistic regression, balancing the tradeoff of specificity and sensitivity, it was determined that the best logistic regression model was that with a 0.5 cutoff value. In practice, this means that when the model predicts a Loyal response of 0.5 or above, the team categorizes the customer as “Loyal”, and when the model predicts Loyal to be less than 0.5, the team categorizes the customer as not Loyal. 

As seen in the confusion matrix, the model the team chose yields an overall accuracy of 73%, specificity of 98.5%, and sensitivity of 5.6%. When the customer truly is not loyal, the model correctly classifies these customers nearly 100% of the time. Any attempt to improve sensitivity resulted in significant decreases of both accuracy and specificity. Furthermore, the impressive specificity value means that H&M could predict with high accuracy which customers are likely to not be loyal in the future based on current purchasing behavior, club member status, and newsletter reading habits.

## Conclusions & Results

As expected, club member & newsletter status are strong predictors of future sales along with past purchase behavior, age, and purchase channel. Customers with an active club member status generally are loyal to the brand as well as aware of the latest promotions, and are expected to make higher future purchases. Current sales ($ amount of purchases in t1) is also a strong predictor, as expected, as customers who are currently purchasing more are expected to have higher future sales.

The results of the logistic regression model could help H&M’s marketing team strategically allocate its budget. Since the model created has a near perfect specificity rate, the team knows with near certainty which customers are unlikely to purchase frequently in the next time period. In order to maximize revenue per ad dollar spent, the marketing team should avoid sending marketing to the customers classified as not loyal by the logistic regression model. This could be accomplished by creating suppression audiences consisting of customers not expected to be loyal. H&M should focus their advertising on customers who are already loyal, and could be convinced to add another item to their cart, driving up each transaction price, and correspondingly, H&M revenue.

The linear regression model could help H&M find high-value customers to market to as well as prioritize high-value customers in customer service queues since losing them due to poor service would have the highest cost. The insights gained from the model could also steer H&M’s product development - they may want to produce new offerings that align with the purchase behavior of high spenders (more accessories, young adult clothing, etc.).

Together the logistic regression and linear models can be used to decide where and where not to focus efforts and investments in order to optimize ROI.

### 

### Works Cited

Huang, Jess, and Thomas O'Toole. “Customer Loyalty: The New Generation.” *McKinsey & Company*, McKinsey & Company, 14 Aug. 2020, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/customer-loyalty-the-new-generation.

Katz, Elie Y. “Council Post: How to Build Customer Loyalty.” *Forbes*, Forbes Magazine, 10 June 2022, https://www.forbes.com/sites/forbesbusinesscouncil/2022/06/09/how-to-build-customer-loyalty/?sh=78bee284b0ae.

Polkes, Aliza. “The State of Brand Loyalty 2021: Global Consumer Survey.” *Yotpo*, 28 Dec. 2022, https://www.yotpo.com/blog/the-state-of-brand-loyalty-2021-global-consumer-survey/.

### 