SMU Data Science Review

Volume 6
Number 1 *Spring 2022*

Article 6

2022

A Machine Learning Approach to Revenue Generation within the Professional Hair Care Industry

Alexander K. Sepenu
Southern Methodist University, asepenu@smu.edu

Linda Eliasen
Southern Methodist University, l.eliasen7@comcast.net

A Machine Learning Approach to Revenue Generation within the Professional Hair Care Industry

Alexander Kwaku Sepenu[1], Linda Eliasen[1], Advisor: Gordon E. Berry
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA

[1]{asepenu, leliasen}@smu.edu

Abstract. – The cosmetic and beauty industry continues to grow and evolve to satisfy its patrons. In the United States, the industry is heavily science-driven, innovative, and fast-paced, suggesting that to remain productive and profitable, companies must seek smart alternatives to their current modus operandi or risk losing out on this multi-billion-dollar industry to fierce competition. In this paper, the authors seek to utilize machine learning models such as clustering and regression to improve the efficiency of current sales and customer segmentation models to help HairCo (pseudonym for confidentiality), a professional hair products manufacturer, strategize their marketing and sales efforts for revenue growth. The present challenge facing HairCo is the lack of models that learn from aggregated data centered on the buying behavior, demographic, and other publicly available data of end consumers tied to historical sales data of their customers, i.e., salons and stylists. The proposed clustering and regression models achieved notably improved results using the aggregated data in comparison to models solely using internal company-provided data. Recommendations on which features are most important from both models that improve customer profiling and predicting sales were presented. With these results, HairCo can increase its revenue and expand its market share.

1 Introduction

The ancient use of cosmetics and beauty products originates from the early Egyptian, Roman and Greek civilizations. Archaeological findings also show that Neanderthals painted their faces with brown, red, and yellow arsenic, clay, and mud as an early form of beautification Kumar [1]. Other studies also suggest that hair was rolled on bones to curl, and the use of tattoos, makeup, and adornments were used to convey status and class. As described by [1] an ancient Greek physician, called Galen, invented cold cream whereas Romans used oil-based perfumes and fragrances in their baths, fountains, and on their bodies to help them relax and de-stress. As trade and migration became more commonplace during the 13th-century, merchants brought perfumes back from the Far East to Europe giving a rebirth to this important and valuable industry [1].

The beauty and cosmetics industry contributes significantly to the regional and national economies across the world. Purchases, patronage of services, and the payment of wages

and taxes generated contribute significant wealth and jobs that benefit citizens and countries as a whole.

The importance of this industry cannot be overlooked as the production of products and its ancillary services affect a myriad of other industries. A top-down analysis of the supply chain processes touches on raw materials sourced from mostly the agricultural and the chemical industries, the manufacturing industry, the oil and gas industry, distribution through the transportation industry, sales through retail, wholesale, and e-commerce industries, the financial industry, the media industry and by and large the beauty and salon industry that serves as the end-users.

However, despite its importance and reliance on various industries and with tremendous potential for brands to remain competitive, there are razor-thin margins for error. To improve the odds of success, new entrants, and established companies often either outsource their products to vendors through franchising agreements or through robust advertisement campaigns to realize new sales. Though there are in some cases successes, many of these expensive marketing campaigns often yield low sales and disappointments, driving some of these nascent brands and products to the brink of survival [2].

According to 2019 Mordor-Intelligence, the salon hair care market in the United States is expected to reach 3.1 billion dollars by 2024, primarily driven by consumer interest in specialized hair care products and increased salon visits. The report also suggests that consumers rely on their stylist's knowledge of product compatibility and expertise in the field for pre-purchase advice driving demand for professional products through salon channels [3].

Touching on challenges in the industry are fluctuations in hairstyle trends, increased disposable income, population growth, and a rise in air pollution coupled with a plethora of competing products consumers can choose from. Available products range from colorants, straightening and perming products, shampoos, conditioners, styling products, sprays, and treatments. These products are developed with a wide range of formulations, designed to address the differing hair needs of the consumer; hence requiring innovation to drive customer satisfaction and loyalty. Professional hair care products are differentiated by the high unit cost and the limited distribution networks. Kumar [1] points out that to remain profitable, industry leaders embark on common strategies such as remarketing of core brands and select product relaunches as well as market expansions in various regional markets across the world to sustain and increase growth.

HairCo manufactures and sells professional hair care products primarily through the salon and stylist sales channels. This presents challenges as HairCo does not have direct access to data about the end consumer of their products, hence limited advertising campaigns directed to this audience. HairCo's influence is rather on the salons and stylists, who, in turn, use or recommend products to their clients. Inferences about the end consumer can be drawn from the salon or stylist sales data, however it does not provide direct knowledge about the end consumer as this data is protected by the salons and stylists due to professional code of conduct standards.

The company has an array of brands targeting differing consumer markets in their product line dedicated to these channels. These brands provide a range of professional hair

color, treatments, styling, and home care products to meet consumer needs and preferences. There are three strategic brands of interest that will be the main focus of this research.

In a salon setting, retail product sales play a key role in its profitability. HairCo estimates these consumer purchases represent 7-15% of total salon sales, with profit margins ranging from 42% to 48%. Based on their market analysis, there are approximately 200,000 salons and 1.6 million stylists in the United States.

In HairCo's experience, salon owners typically carry products from only a couple of manufacturers, or are devoted to one brand, with each stylist representing an annual profit on average of $5,000 to the product manufacturer. As such, the industry is highly competitive with numerous product manufacturers competing for market share and dominance.

HairCo faces fierce competition from rival brands, new entrants into the landscape, the rise in e-commerce, and product innovation across the industry. As a result of the highly competitive landscape, HairCo is experiencing a drop in sales volume for the three strategic brands among longer-tenured customers. Prior twelve-month sales indicate accounts are peaking at the two-year mark and declining after that point as customer tenure increases. This is causing HairCo to lose market share to competitors. Current sales efforts rely on hosted educational events, customer buying histories, and salesforce experience, incorporating data mining and marketing insights into the mix.

To build on these traditional methods, HairCo is seeking to leverage machine learning modeling to improve revenue and to prevent customer churning. Two areas of focus to address HairCo's desire to use machine learning to accomplish these goals will be in clustering and regression. Customer clustering, or segmentation, divides a company's customers into distinct groups that reflect similarity to each other based on identified variables and metrics of importance. The data points in the cluster share common features but are distinct when compared to data points in other clusters. Machine learning methods allow advanced algorithms to surface insights and groupings that companies may have difficulty discovering on their own and form the groundwork to establish customer profiles to systematically identify high revenue-generating customers.

This will enable HairCo to determine the best way to engage with customers in each segment. From a marketing perspective, they can develop specific customer profiles for each segment in order to create a more tailored marketing campaign to suit each of the customer segments based on their individual needs and characteristics as well as persuasive advertisements which spurs products sales.

The results of the sales prediction model will identify factors that influence revenue the most. The knowledge gained can be incorporated into strategic planning in order to better focus on resource allocation and marketing efforts. It also allows the sales organization to pursue high-value opportunities; focusing time and effort on the more lucrative customers to obtain higher quality revenues. These tailored strategies are expected to maximize the value of each customer to the business and drive loyalty, promoting the stability of the market base.

The current problem facing HairCo's channel sales and largely the entire professional hair care industry is the lack of models that learns from a wide pool of data such as location

demographics, end-consumer behavior, sales campaign data as well as internal records to learn the buying behavior of customers to improve the efficiency of segmentation models as well as sales prediction models. This is important as a data-centric approach is ideal to strategize the way the company sell its products to salons and hairstylists that meets the needs of end consumers to maximize profits.

This research aims to solve this dilemma by presenting a novel approach that aggregates data by incorporating location based demographic census and social context data with internal company data to feed machine learning algorithms such as clustering and regression analysis to predict sales as well as to segment customers in order to create more robust customer profiles. The belief is that aggregating these data for use in machine learning models will result in improved models versus those that the company currently uses. The insights realized from these two models will guide HairCo in formulating various strategies to best target current and future customers to increase market share and drive revenue growth in an increasingly crowded industry.

2 Related Work

Historians have it that the quest for beauty and body enhancements to combat the perception of attractiveness and aging to affirm one's identity in society or a group have come a long way since the earliest human civilization Ehlinger-Martin *et al* [4]. The authors suggest that these ancient humans employed strategies that built self-esteem, hierarchical roles, and a higher status in society. They indicate that over the centuries the attitude of women towards appearance, aging, and in retrospect "beauty" has undergone tremendous changes. They attribute this assertion to women's longer life expectancy that transcends their reproductive years. This gives birth to the beauty industry if one can spare the definition and what it entails.

Evolving aristocracies over the centuries according to Hornsey [5] introduced a cult of a class system in which women in the twentieth century forged new identities built on ostentatious appearances reflective of their social class. Black [6]in their research affirms the assertion of [5] that throughout history men and women altered their bodies with beauty products that conform to their defined aesthetic standards. The authors [5], [6], go further to point out that even though the contemporary beauty industry has popularity, it has its antecedent in the nineteenth century owing to commercials or ads that target women. This, however, did not gain traction until the mid-'20s to late '30s during the post-war era when the beauty industry exploded and became fully recognizable, requiring formal training, licensing, and certifications. It was during this period that ethical standards and a professional code of conduct were established for the industry.

According to Black [6, pp. 15-21], the beauty market for elixirs began with women spending their disposable income at the wholesale or drugstores where these products were sold. Jones [7, p. 5] also refers to this trend as an evolution of the beauty industry over time, suggesting that instead of the industry being "capital-intensive, mass marketing and mass production industries" rather, it is a "large number of small and medium-sized

entrepreneurial firms" that have become multi-billion-dollar global conglomerates. The successes of these enterprises solely depended on the distribution networks, knowledge of products, and salon franchises [6].

According to Winship [8], the development of the industry can be attributed to the emergence of departmental stores from the nineteenth to the twenty-first century. This is a result of the rise in consumer culture as more women gained economic independence by entering public spaces and gaining paid employment. These among many others created the demand for beauty products and the birth of an industry Peiss [9].

2.1 Conventional Marketing, Sales, and Use

Hair and beauty salons according to Black [6] become the new retail hubs for many beauty products as workers within them mixed up preparations for use on the client's hair and body as well as distributing products by mail and door-to-door, a technique pioneered by Madam C. J. Walker and Annie Turnbo Malone [6]. As the industry expanded and became more professionalized, the market, as well as product offerings, expanded according to Willett [10].

Although salons, according to Linnan & Ferguson [11] are an innovative and "unconventional way" to reach a defined population. Over the decades or perhaps centuries, sales of beauty products have witnessed dramatic shifts attributed to various economic, social, and cultural transformations as consumer values and tastes changed Lopaciuk and Looda [12]. These changes according to Jones, [13] required new marketing techniques.

These marketing techniques are hinged on the level of customer satisfaction, perceived quality of products and quality of service from service providers or product manufacturers Oliver [14], [15]. According to Zeithaml, Berry & Parasuraman [16] these consumer behaviors are contributory factors to post-purchase attitude and intended repeated purchase for a particular product.

However, in the beauty and hair care industry, hair care professionals according to Auh, Salisbury and Johnson [17] make sales based on product loyalty and satisfaction of customers. Gimlin [16], [17], [18] also bring another dimension to the discussion by arguing that, due to cultural exchanges during the process of provider-client interaction, hair care service providers or hairstylists introduce clients to products. These sales and marketing technique relies on the stylist's depth of knowledge to cater to the customer's individualized hair care needs as well as a bond of trust between them often and ultimately leads product purchase and customer satisfaction Jacob-Huey [19].

According to Oliver [21, p. 34] brand loyalty can be defined as "a deeply held commitment to rebuy or patronize a preferred product or service consistently in the future." This drives repetitive purchasing, regardless of situational influences and marketing efforts that have the potential cause "switching behavior" or sway these brand loyalists. This affirms that customers buy products that provide them the level of satisfaction and self-fulfillment they desire. Rooted in this assertion, Chandahuri and Holbrook [22] points out that the product brand sale performance is conditioned on brand trust, hedonic value,

loyalty, relative price, market share, and both rapport and the symbiotic relationship between the customer, stylist, and/or the hair care provider.

Thus, Spake *et al* [23] further suggest that the provider-client relation leads to customer comfort that includes confidence, brand disclosure, reduced risk, and improved satisfaction which Brown and Beale [20] agrees with. They further argue that marketing techniques in the hair and beauty care industry are dependent on marketing strategies that link product mix, quality of service, and profits with superior services that correlate with the desire to grow the market share of a product [20]. Features such as scheduling appointment times, timely services, and convenience among others are influential determinants to a salon or a stylist's ability to engender or win a customer's loyalty and satisfaction for a particular type of service or product [20], [24].

Another point by [25] posits the idea from traditional marketing in the hair and beauty care industry that employs strategic tools to attract standard customers. This uses the strategy of transactional marketing to profile potential customers and the process of completing sales transactions with them. They further cite Robinson [26] to detail four ps' (Product, Price, Place, and Promotion) of marketing as the precepts of transactional marketing to address customer segmentation, targeting, and positioning to optimize product sales and customer satisfaction.

However, due to lack of unanimity and differing opinion on which marketing strategy gives the optimum outcome for product sales and customer satisfaction Amoakoh and Naong [25], [27] argues that no marketing strategy can yield the best results.

2.2 The Role of Machine Learning

The approaches discussed so far require a holistic and scientific approach that uses data to study the various relationships or features that reflects a customer's propensity to patronize a product or hair salon service in a given marketing strategy.

Pearl [28] makes a compelling case for using machine learning to analyze the causal structure between features that influence an outcome. This method is used by Horita and Yamashita [29] to highlight marketing effect and strategy on hair salons growth and their customer base. To further narrow it down, several researchers cited by [29] used the clustering technique of machine learning to divide customers into clusters based on probabilistic features to study their relationships on their expected outcomes. Weerasinghe and Yidanagama [30] also use convolutional neural networks in machine learning to build a recommendation system that promotes customer preference for a particular hairstyle.

That said, to define machine learning, Müller and Guido [31, p.1] define this as "extracting knowledge from data". To further expand, they define it as "predictive analytics or statistical learning" that intersect computer science, statistics, and artificial intelligence.

These predictive analytics focus on the large and high-dimensional dataset for predictive accuracy compared to hypothesis-driven inference. It is further classified into unsupervised, supervised, semi-supervised, and reinforcement learning Bi *et al* [32]. Complex decisions were made using hand-coded rules that often-required domain

knowledge [31] but with the advent of machine learning methods and its evolution, the application has become ubiquitous both in personalized and commercial applications automate decision making processes [31].

Conversely, [16] also define machine learning as algorithms that learn from data by sifting through vast amounts of predictors to look for covariates and statistical functions that predict outcomes. A look at the approaches or techniques of machine learning, Russel and Norvig [31, p. 653] discusses supervised learning and how it works. The authors define supervised learning as "the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples".

Kotsiantis [34] citing many authors, discusses supervised learning algorithms as most widely used techniques to include Support-vector machines, Linear regression, Logistic regression, Naive Bayes, K-NN algorithm, Linear discriminant analysis, Neural networks (Multilayer perceptron), Decision trees, and Similarity learning.

On unsupervised learning, Hinton and Sejnowski [35] define it as "a type of machine learning in which the algorithm is not provided with any pre-assigned labels or scores for the training data".

With Semi-supervised learning, Chapelle *et al* [36] define it as "halfway between supervised and unsupervised learning" this puts it in context as combining labeled and unlabeled data to use this machine learn approach.

Reinforcement Learning according to Hu *et al* [34] is a type of machine learning that defines how software agents behave in a given environment to maximize output. For instance, in clinical studies, Taylor *et al* [37] in their publication used this machine learning approach to predict in-hospital mortality rate of patients with sepsis in the emergency department using random forest methods. The research showed outstanding results and outperformed existing clinical decision rules and traditional analytic techniques. This outcome they suggest can be automated and applied to other clinical studies of interest and deployed locally on all healthcare servers to enable real time clinical predictions.

Zhu *et al* [38] applied machine learning approaches to forecast credit risk of small and medium-sized enterprises in supply chain finance as financial institutions current traditional credit risk forecasting models cannot meet the forecasting complexities in this modern times. Indeed, supply chain finance has become the most critical issue in financial decision-making to ascertain the creditworthiness of small medium-sized enterprises in reducing non-performing loan ratios hence removing bottlenecks that stifle small medium-sized enterprise development and growth.

Another use case of machine learning is by Nguyen *et al* [39] to predict finance risk by companies to estimate their carbon footprints in climate analyses. In this research, [39] looked at the limitations of existing methods of estimating carbon footprint of companies that did not guarantee high prediction accuracy, such as regression models, which often times resulted in lower predictive power on the assumption of linearity in parameters with some unbiased restriction in estimates. Compared to machine learning approaches, these do not rely on statistical assumptions giving flexibilities of use allowing non-linear, high order

interactions in the dataset to minimizes prediction errors hence making this novel in addressing the research problem.

To recap why using machine learning for this research is important in the professional hair care industry, several machine learning approaches have been presented by the various authors to improve revenue and maximized production outputs. In Tarallo *et al* [40], machine learning is leveraged to predict demand for fast-moving consumer goods. In this research [40] presents using machine learning techniques to achieve high model accuracy that surpassed the accuracy level of traditional statistical methods, which are often fraught with statistical error, to improve inventory balancing throughout the value chain. In turn, reducing points of sale re-stocking rates, improving product availability to consumers, and increasing revenue.

In Wu and Zheng [41], the researchers presented a model for sales forecasting using machine learning methods that outperformed traditional statistical techniques in accuracy for fast-fashion retail products that have high demand volatility and short life cycles.

Chen and Ou [42] in their research used machine learning to forecast sales and weather data which improved production line decision-making processes as well as increased profitability and reduced production losses. The model presented in [42] also proved to improve customer satisfaction at the point of sale for products as they were available for purchase. [41], [42], [43] all used machine learning techniques to improve revenue, the business process and forecast with significant accuracy.

Citing the advantages of machine learning algorithms by Yu and Liu [44, p. 8], this research will employ these approaches and their advantages to handle the high dimensionality of the data to test the proposed hypothesis. Just as Lakoju *et al* [47], unsupervised clustering technique was used in the study to identify features with similar attributes and an evaluation of the best performing model was measured. Also in Suwalka and Agrawal [48], unsupervised clustering technique like SOM and K-means was used to segment data for the detection of alzheimer's disease.

Given the background study discussed above, this research will test the hypothesis that aggregating external data with internal data into the models does not improve customer profiles and sales forecasting of HairCo's products given a salon or stylist.

3 Methods

The data provided by HairCo spans nearly seven years, from January 2015 through November 2021, for salons and stylists in the state of Florida. According to the company, the Florida market is unique in that the varying demographics are reflective of the product markets that HairCo sells products into across the United States. The dataset was pared down to approximately 3,000 customers (Salons and Stylists) who purchased the three brands of interest within a defined revenue range over a twelve-month period. There were several features with a large proportion of missing data that could not be imputed with a high degree of accuracy and were dropped from the dataset. Features with a correlation greater than 95% were identified and removed.

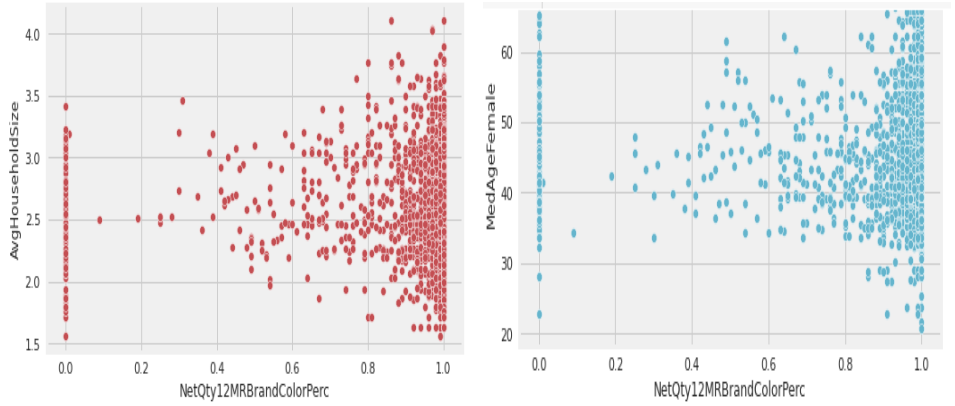
In addition to the company-provided data, publicly available demographic data based on zip code from the American Community Survey (ACS) [58] published by the U.S. Census Bureau and social context data at the county level from StatsAmerica [50] was added to the dataset. The final dataset includes 76 features pertaining to net sales, product mix (color versus all hair products), transactional data, and other demographic and social context characteristics. Table 1 below shows the final features used for the analysis.

Table 1. Table of Features used for Analysis

Company Provided	StatsAmerica	American Community Survey
Identifier	Agreeableness	PerCapitaIncomeR12
Group	BeliefInScience	HousingCostsMonthlyMedian
SoldToYears	Collectivism	Households
NetSlsLifeAvgPerYr	ConflictAwareness	AvgHouseholdSize
NetSlsLifeBrandAvgPerYr	Conscientiousness	MedAgeFemale
NetSlsLifeBrandPerc	Empathy	FemPop
NetSlsLifeColorAvgPerYr	EmploymentRate	FemBelowHS
NetSlsLifeBrandColorPerc	Entrepreneurship	FemHS
NetSls12MRBrand	Extraversion	FemSomeCol
NetSls12MRBrandColor	GenderEquality	FemBach
NetSls12MRBrandPerc	Hopefulness	FemMast
NetSls12MRBrandColorPerc	IncomeMobility	FemProfSch
NetQty12MRBrand	IncomePerCapita	FemDoctorate
NetQty12MRBrandPerc	Neuroticism	WhitePop
NetQty12MRBrandColor	Openness	AfrAmerPop
NetQty12MRBrandColorPerc	Religiosity	AmerIndPop
EducationClasses12MR	RiskTaking	AsianPop
EducationClasses24MR	Selflessness	PacIslanderPop
EducationColorClasses12MR	Tolerance	OtherPop
EducationColorClasses24MR	WorkEthic	TwoplusPop
TransactionCount12MR		FemPopCount
TransactionCount12MRColor		FemPop10_19
TransactionCount12MRBrand		FemPop20_29
TransactionCount12MRBrandColor		FemPop30_44
LevelTier		FemPop45_54
LoyaltyMember		FemPop55_59
SummitMember		
LifeisSuite		
RiskRatingDesc		
CustomerGroup6Description		
AccountAssignmentGroup		

3.1 Exploratory Data Analysis

The exploratory data analysis reveals that a majority of the data is not normally distributed among the features and have a high degree of skewness seen in the Figures 1 to 3 below. In order to keep as much information as possible to inform the models, outliers were kept in the data due to lack of domain knowledge as most features have high degree of outliers as seen in the sampled boxplot in Figure 2.



(a) (b)

Fig. 1. Scatter plot relationship between Net Quantity 12 MR Brand Color Percenatge aganst Average Household Size (a) and Net Quantity 12 MR Brand Color Percenatge against Medium Age Female (b)

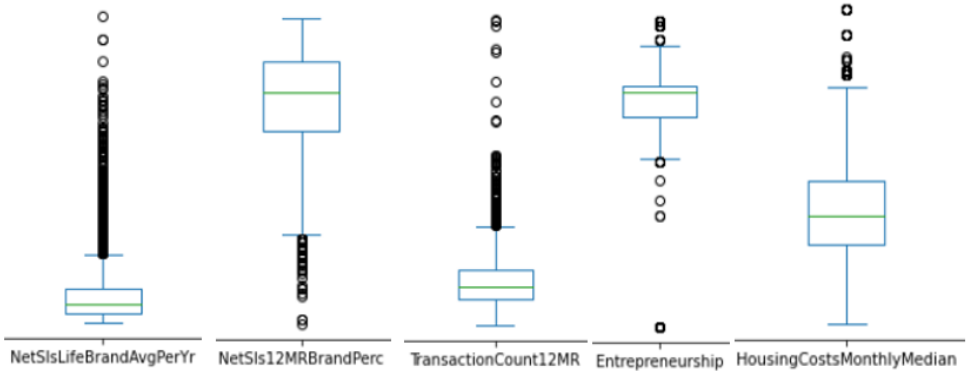


Fig. 2. Boxplot showing a few features with outliers in the dataset

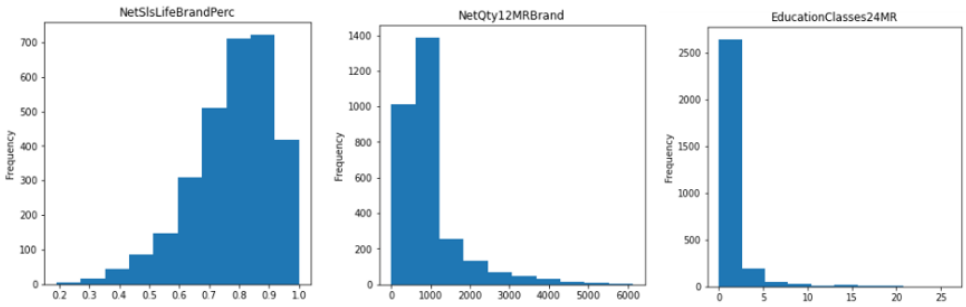


Fig. 3. A histogram plot of a few features not normally distributed.

Within the dataset salons make up 40%, and stylists the remaining 60%, of which 84.5% are independent stylists and 15.5% are commission based. The majority of customers i.e., salons and stylists, 70%, have been customers for five years or more, and this group represents 60% of the brand net sales figure for the rolling 12-month period. In terms of product mix, 66% of net sales come from professional hair color sales.

Examination of customer average sales from the company data for the same period indicates that sales begin dropping off after a customer reaches the two-year mark and declines steadily from that point as seen in Figure 4 below.

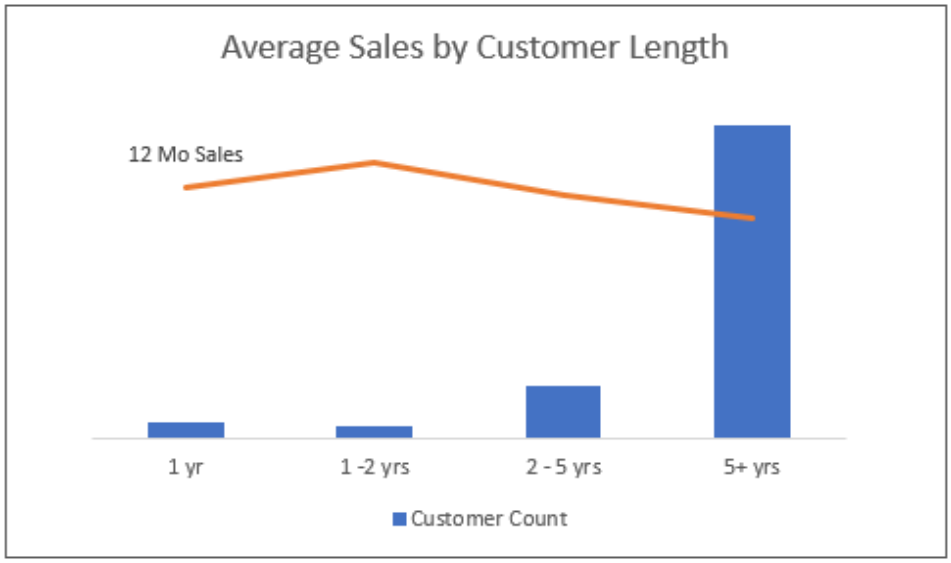


Fig. 4. Average Sales by Customer Length

HairCo sponsored education events are geared towards increasing sales and engendering brand loyalty. These events range from large-scale shows and events to in-salon product education seminars. HairCo also has its own academies where product education and artistry events are offered to stylists. Interestingly, in reviewing the data, it does appear that there is a weak positive correlation between these educational events and sales as seen the scatter plot below in Figure 5 as the data points are heavily concentrated in the third quadrant. Salons and stylists who participated in these events during the prior two years overall did not generate notably higher levels of brand net sales versus non-participants for the 12-month period. It will be interesting to see what, if anything, subsequent modeling uncovers about this revelation. This will be key to understand why there is not a stronger relationship between brand net sales and these educational events.

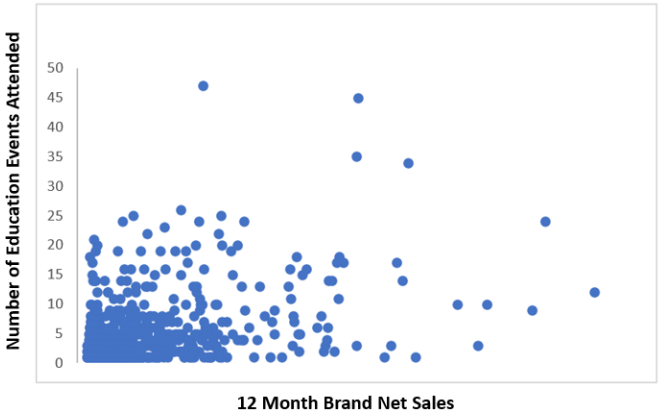


Fig. 5. Education events attended vs. 12 Month Brand Net Sales [NDA restricts displaying dollar values]

The next sectionslook at the application of machine learning methodologies, specifically clustering techniques, and regression algorithms, to gain more insight into the aggregated data and then drill down on details that will help answer the hypothesis and solve the research problem. Cluster analysis was performed to find the number of clusters that best suits the data. The resulting cluster assignment for each customer was then incorporated into the original dataset. Subsequently, classification modeling was conducted to identify the most influential features in determining a customer's cluster assignment. These characteristics were then used to develop a profile for each customer group, or segment. Separately, regression modeling and analysis was performed for sales prediction. This analysis considers competing models and their comparative metrics to choose which performs better on the dataset. The best regression model is then used to find feature characteristics that are believed to drive sales to increase revenue when integrated into corporate information systems or enterprise resource planners.

3.2 Customer Clustering

Clustering analysis is a technique for discovering interesting patterns in a dataset based on their characteristics. Though there are many clustering algorithms, no one clustering algorithm is suitable for use with all datasets. In that, it is good practice to explore a variety of algorithms on this particular dataset. It is against this backdrop that this research employs this technique to achieve its goals. The data has a high degree of skewness and is heavily laden with outliers as shown in Figure 2. In this regard, the data required a transformation to normalize it for use in hierarchical and distance-based clustering algorithms. This ensures all dimensions are treated equally and contribute the same impact on the distance.

The robust scaler [51] algorithm was used to normalize the dataset.This scaling method is prefered because it minimizes the influence the outliers have on the sample mean by removing the median and scaling the data according to the interquartile range and centers the data independently feature by feature. Scaling also transforms the data to a common scale, with an approximate range of 0 to 1. Categorical features were converted to numerical using one-hot encoding.

To reduce feature dimensionality of the dataset and improve efficiency prior to clustering, Principal Component Analysis (PCA) [51] [52] was conducted. The dimensionality reduction from the PCA entails zeroing out the smallest components thereby lowering dimensionality of the data vis-a-vis preserving the maximum explainable variance and also helps remove multicollinearity. This helps analyze the data more accurately in a low dimensional space for regression and classification type problems rather than in a high dimensional space.

To identify the number of principal components that preserves the most information needed, the optimal threshold for number of components was set to 85% after successive tries in the range of 80 – 90%. This implies that 85% of the total explainable variability in the dataset that is preserved lies in the first 18 principal components. These 18 new features were then used for clustering.

The clustering techniques considered include Mini-Batch K-Means, BIRCH, K-Means, Agglomerative Clustering and Gaussian Mixture Models for which resulting metrics were compared and the best one chosen. K-Means randomly assigns centroids by spreading out the initial centroids based on maximum calculated Euclidean distance between the centroids, then optimizes the iterations for the clusters to reduce the variance within each cluster [55]. Mini-Batch K-Means uses small and random fixed size batches of data held in memory for the iterative process that is used to update the cluster [57]. BIRCH uses the natural closeness of data points for clustering decisions without scanning all observations [56]. Agglomerative Clustering utilizes a hierarchical clustering approach that uses nested clusters from the bottom up where each observation starts off in its own cluster and then is merged with other singular clusters using a linkage criterion [51]. Finally, Gaussian Mixture Model is distribution based, assuming that each distribution represents a cluster [51].

To estimate the number of clusters prior to clustering, two methods were evaluated to find the optimum number. One was using the Elbow method or Sum of Squares Error (SSE) method and the other is the Silhouette Score method. The elbow approach is used to find the elbow point where the within-cluster sum of square error curve declines and starts looking linear as the number of clusters starts to increase.

This is the point where the value of k (the elbow) is best. On the other hand, the Silhouette score is the measure at which the points lie close to other points within-clusters or across all clusters. This provides valuable insights or scores on cluster separation or quality to inform on whether the clustering requires further refinement to get a clear separation among clusters or not. Evaluating both methods, the SSE indicated 4 clusters while the Silhouette Score method after hyper-tuning indicated 3 as seen in Figures 6 and 7. Both recommendations for the ideal number of clusters were applied in the clustering algorithms, however, it was discovered that the number of clusters identified from the Silhouette method created more distinct clusters for the dataset. That said, the silhouette method will be used moving forward with the analysis.

Fig. 6. Number of Recommended Clusters based on SSE

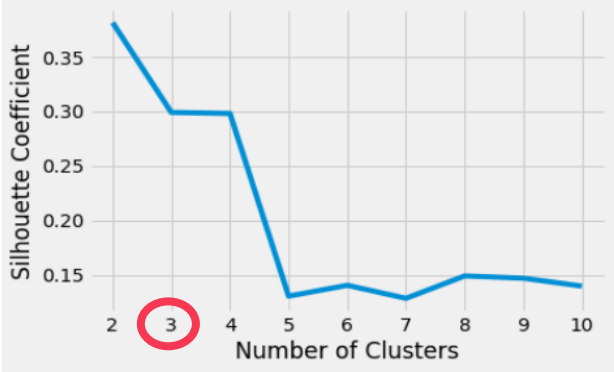


Fig. 7. Number of Recommended Clusters based on Silhouette Score

The Silhouette Score, a metric used to calculate the goodness of a clustering technique, was used as the basis of comparison for the clustering algorithms. Its value ranges from -1 to 1, with scores closer to 1 indicating well-separated clusters. The algorithms reviewed performed similarly as seen in Table 2.

Table 2. Comparison of Clustering Technique

Clustering Algorithm	Silhouette Score
Mini-Batch K-Means	0.302
K-Means	0.299
BIRCH	0.296
Agglomerative Clustering	0.296
Gaussian Mixture Model	0.289

K-Means was selected as the final method in order to include all data simultaneously in the clustering decision. The cluster distribution in Figure 8 below shows nearly 82% of customers fall within the first cluster while the other two clusters each contain approximately 9% of the population.

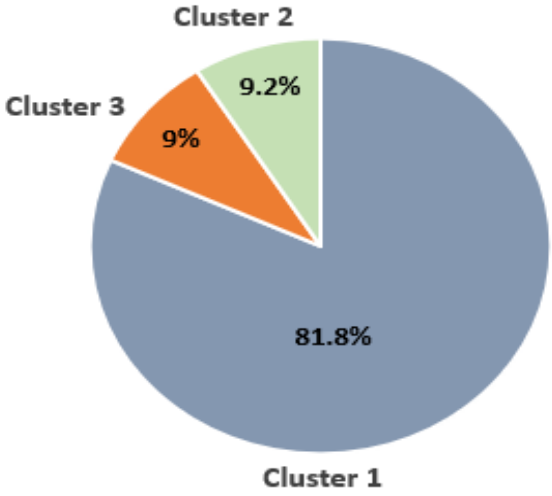


Fig. 8. Cluster Distribution

3.3 Feature Importance of Cluster Assignment to develop Customer Segment profiles

In the ensuing section, cluster assignment is analyzed to better understand the differing customer characteristics that led to cluster assignment to draw influential features for use in customer profiling. The understanding of the influential features of cluster assignment ensures the interpretability of the cluster assignment and acts as class descriptors. This is achieved by finding and assigning scores to the features that drives the cluster assignment according to relevance in the cluster by the chosen algorithm. Feature selection, or feature importance, is also the identification of variables in lower dimensional space that offer clear separation seen in the clusters. This is worth mentioning because cluster centroid also lowers features for easier interpretation [55].

To understand the underlying differences between each cluster, customer cluster assignment was added to the original dataset and classification using Random Forest Classifier was conducted and the results compared against a baseline Logistic Regression model. Random Forest was selected because of its interpretability of feature importance, which is what the analysis is seeking to understand in terms of cluster assignment to ultimately develop or create customer profiles for each cluster.

Data was separated for analysis, reserving 20% for validation, utilizing stratified shuffle split [51] to keep an even representation of classes as visualized in the cluster distribution from Figure 9 above. In addition, Synthetic Minority Oversampling Technique (SMOTE) [54] was applied to the training set to address the severe imbalances between the

classes as this synthetizes the minority classes to balance the class distribution. SMOTE transformation is important when working with an imbalanced dataset as most machine learning algorithms tend to ignore minority classes and will perform poorly, requiring modification to avoid simply predicting the majority class.

A model's hyper-parameters are parameters that can be adjusted to control the learning process and determine the best performing model for the training data. Manually setting this combination of parameters can be labor intensive, hence hyper-parameter tuning [51] is preferred to find the optimal hyperparameters of a model. The technique used in this research is RamdomizedSearchCV from scikit learn [51].

3.4 Sales Prediction

HairCo invests considerable resources in education and training campaigns as an anticipated driver to increase revenue. As seen in Figure 5 above, Education events attended versus 12 Month Net Sales indicate a weak positive correlation between education and revenue. Other data currently used for revenue prediction is limited to internal data as reflected in Table 1 above, but as discussed earlier does not include features that attempt to understand end-consumer buying behavior. In this regard, this section looks at how machine learning regression models can be used to not only unearth valuable information about the aggregated data, internal and publicly available, to determine which features are important to predict sales but also to develop models with improved predictive performance. The best regression model using the aggregated data will be compared against a model using only company provided data to measure improved model performance.

For this analysis, 20% of the data was reserved for validation. Data was normalized using Robust scaling and categorical features were one hot encoded. Linear Regression was established as the baseline model and Random Forest and Support Vector Regression (SVR) were considered as alternative models. RandomizedSearchCV was employed for hyperparameter tuning for both Random Forest and SVR.

4 Results

4.1 Classification Results

Table 3 below shows a comparison of metrics between the baseline logistic regression and Random Forest Classifier models. With a baseline accuracy of 84.6% for the logistic regression model, utilizing Random Forest improved the accuracy to 98.3%.

Table 3. Model Comparison and their Accuracies

Cluster	Model	Precision	Recall	F-1 Score	Accuracy
1	Random Forest	1.00	0.98	0.99	0.98
2		0.87	0.96	0.91	
3		0.96	1.00	0.98	
1	Logistic Regression	0.97	0.85	0.91	0.85
2		0.61	0.87	0.72	
3		0.47	0.80	0.59	

It can be noted that, as the company expects to tailor marketing and sales efforts based on cluster assignment, misclassifications are a concern. It is therefore paramount to keep False-Negatives at a minimum as much as possible hence the suggestion to focus more on maintaining a high precision rate rather than Recall rate.

The top 22 features of importance from the Random Forest Classifier model are depicted in Figure 9 below this shows approximately, 55% of the most important features used are location demographic and sociodemographic data from the publicly available data sources that were incorporated into the dataset as noted with asterisks.

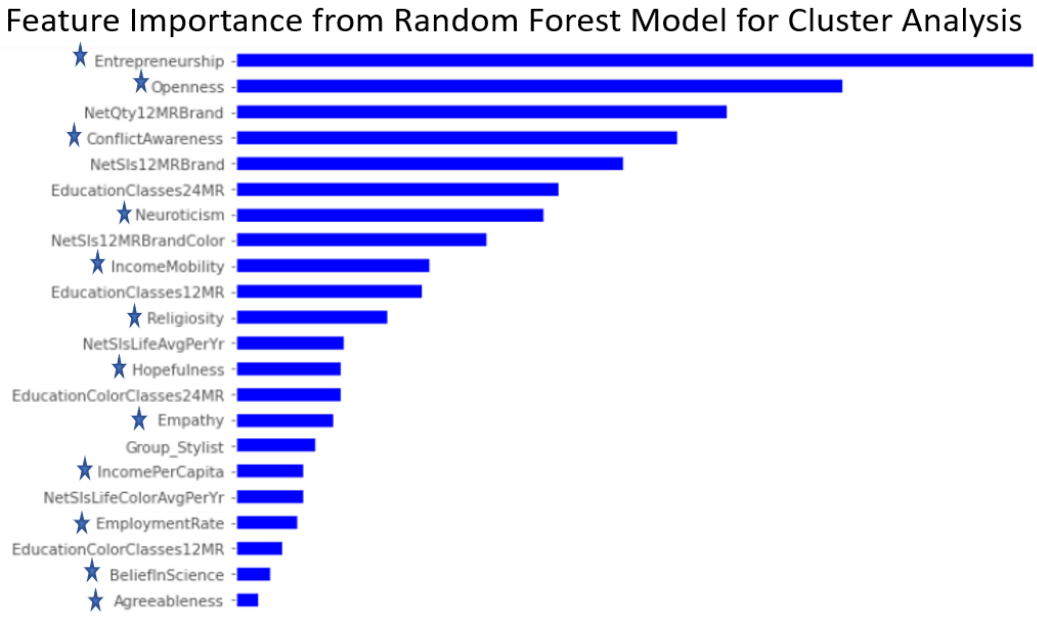


Fig. 9. Feature Importance from the Random Forest Model used for clustering

To validate the chosen features, permutation feature importance from scikit learn [51] was conducted. This step is very important because it is a model inspection technique that is useful for any fitted non-linear estimator.

The identified features were investigated for each cluster to ascertain the differences between the three groups. Based on these differences, there are unique characteristics for each cluster. Cluster 1, predominately stylists, has the highest percentage of customers accounting for 65.6% of overall sales. Clusters 2 and 3, while small, have higher levels of average sales per customer. Cluster 2, primarily salons had the highest participation rates in company sponsored education events, both product and color based, and purchased 3x the colorant products vs the other 2 clusters.

From a location perspective, some of the more interesting features were analyzed. Cluster 1 had the highest entrepreneurship and neuroticism scores, Cluster 2 tends to have the highest income per capita and employment rates, while cluster 3 had the highest scores for income mobility, belief in science and the largest population of females 30 – 44 years of age. These are shown in Table 4 below.

Table 4. Customer Profile Based on Cluster Assignment

Cluster 1	Cluster 2	Cluster 3
81.8% of Customers	9.2% of Customers	9% of Customers
68% Stylists	94% Salons	Equal Mix
65.6% of Brand Sales	25% of Brand Sales	9.4% of Brand Sales
	~3x Colorant Sales vs. Others	
	Education Participants	
Location Demographic & Sociodemographic Characteristics		
Entrepreneurship	Income Per Capita	Female Population 30-44 yrs
Neuroticism	Employment Rate	Monthly Housing Cost
		Income Mobility
		Belief In Science

4.2 Predicting Revenue

The results from the Random Forest and Support Vector Machine models were compared against the metrics of the baseline model as seen in Table 5 below. From the results, random forest was the preferred model as it had the smallest RSME and the largest R^2. This was then compared with the random forest model using only company data.

Table 5. Model & Methodology Comparison

Dataset: Company Provided & External

	Mean Absolute Error	Root Mean Squared Error	R^2
Baseline Linear Regression	2,307	3,708	67.9
Support Vector Regression	2,190	3,853	65.3
Random Forest	2,004	3,425	72.6

Dataset: Company Provided Only

	Mean Absolute Error	Root Mean Squared Error	R^2
Random Forest	2,344	3,972	63.2

Using the aggregated data results in better model performance versus only using internal company data. This can be seen in Table 5. The model solely using internal company provided data did not perform as well when comparing RMSE and R^2. This is very significant as blending company data with external data provides improved results. This establishes the foundation to find the most important features from the regression model that have the largest impact on sales.

To interpret the model and gain a better understanding of the variables that have the greatest impact on sales, feature importance was reviewed using the results from the random forest model. The 25 top features realized from the analysis can be seen in Figure 10 below. It can be seen that, among the features are features from external data sources, noted with an asterisk, which improves model performance and includes monthly median housing cost, female population groupings, income per capital, as well as median female age.

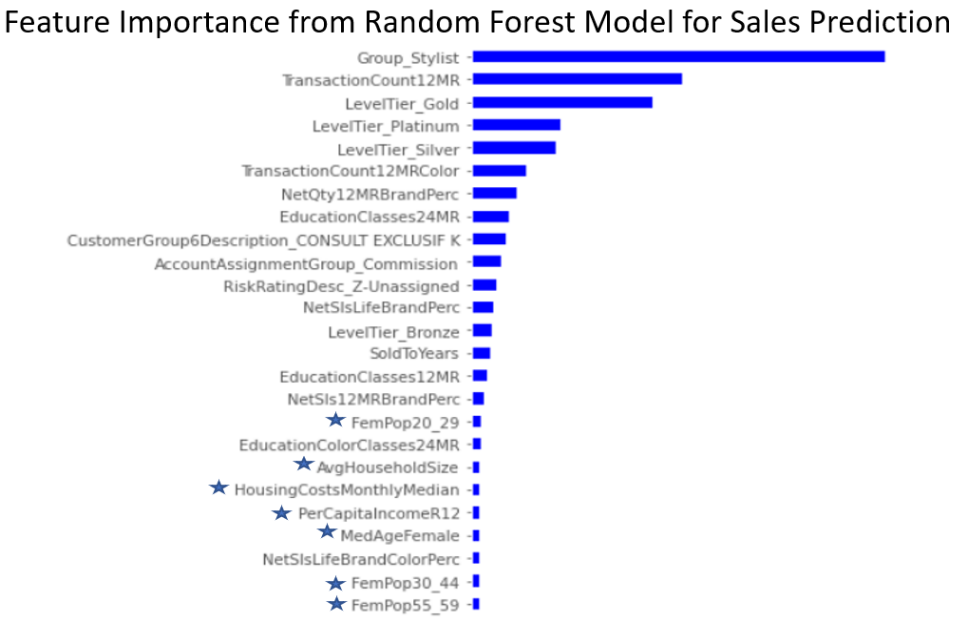


Fig. 10. Feature Importance from the Random Forest Sales Prediction Model

5 Discussion

Models represent real world scenarios to provide insights and make predictions to help guide decision making. However, many factors hinder this effort, especially in the context of inaccessible information which can prohibit the accurate determination of which information is relevant and which is not. To make inferences close to reality requires assumptions to be made which creates a risk of injecting opinions into the model which can often lead to inaccurate results and mislead decision making. Nonetheless, the findings from this research shows that the use of machine learning approaches go a long way to improve models that have been in existence or at best new ones are formed to solve more complex ones. This, when adopted by the professional hair care industry, will be very beneficial to the entire enterprise to improve business processes and maximize profit.

However, to make the best model dwells on the abundance or availability of data that is more representative of the population under study. The results from the two machine learning approaches, clustering, and regression, to solve the research problem show~~s~~ that there is significant evidence to suggest that the data favors the alternate hypothesis that aggregating HairCo's data with external data does improve~~s~~ the robustness of customer segmentation, or profiling, and improve sales prediction models currently in place.

The clustering model developed in this study drills down to the most important features that make up the three clusters to aid in customer profiling. Based on the cluster analysis, the cluster assignment describes characteristics of the segmented customers, and the information gain can be used to develop product mix and marketing strategies. For instance, as seen in cluster 1, nearly 82% of the customers are stylist with characteristics as entrepreneurship and neuroticism to mention a few. This implies that HairCo can roll out a robust marketing campaign that aligns with the entrepreneurial spirit of the stylists to help them build their business.

To average revenue per customer within this cluster, HairCo needs to constitute teams that will become account managers with more personalized skillset to manage these stylist and carter to their needs. This is important because, knowing the customers, understanding their buying behavior, and satisfying their needs give the company a competitive advantage over their fierce rivals. This brings products and services to the doorstep of the customers to achieve not only improved customer life value and retention but also a return on investment that maximizes profit.

Additionally, Cluster 2 indicates that participation in company sponsored colorant education events appears to have a direct influence on colorant sales. By increasing participation in these events for the other two clusters and educating them on colorant application and cutting-edge techniques, will hopefully increase sales. This can also improve customer brand loyalty as it is an industry belief that once a stylist is trained and certified on a particular brand's colorant products, he or she is less apt to switch brands.

In the sales prediction model, external demographic and social context features make up some of the features deemed important as predictors of revenue. Internal data features such as group stylist, transactional data, and membership tiers can be further analyzed to craft more targeted sales strategies. Other attributes such as median monthly housing cost, female population between ages 30 to 44 and 55 to 59 present a bucket list for features that HairCo should take advantage of.

Although statistical metrics present a window into how well models perform on a given dataset, care must be taken when drawing conclusion on which model to rely on. That said, statistical metrics, such as R^2, measure the proportion of variance the response variable, in this case sales, which can be explained by the explanatory variables. These metrics show how well the data fits to the regression model. The metrics used to determine the best model performance for this study are the root mean squared error and R^2. These metrics measured on the Random Forest model using the aggregated data outperformed the model solely using internal company provided data

It can be concluded that the current sales prediction model HairCo uses can be improved by incorporating publicly available location based demographic and social context data in the absence of direct data about the end consumer. In this regard, the wide pool of data discussed so far is better and relevant for revenue prediction when implemented in information systems and broadly across the enterprise in business processes that are designed to increase revenue, grow market share, and improve brand loyalty.

One key limitation worth noting is the lack of readily available industry data. Information gathering requires a high degree of domain knowledge. End consumer data is

protected and not easily attainable via public sources. Also, as much as the models realized from this study are promising, they have their own limitations. There may be additional factors not considered that could influence customer segmentation or profiling or largely impact the demand for a product beyond just sales prediction. For instance, factors such as state of the economy, seasonal changes, end consumer behavior, advertisement and efforts by the distributors can impact the sales prediction.

Another key point for future research is to endeavor for a higher level of robust data. Features that were eliminated due to high degree of missing data may provide additional insights and allow for different machine learning approaches to validate these results or gain additional insight into how the industry can leverage machine learning and data science to increase their profit and customer retention.

Ethics

There has been a quantum leap in the accuracy of decisions humans make using Machine learning algorithms. These algorithms perform complex analysis considering varied constraints to arrive at a decision. These decisions influence human behaviors and are often prone to biases and when these preconceptions become entrenched, humans who build these machine learning algorithms or models stimulate or influence their results. In that care should be taken when integrating opinion-based features into machine learning models. For an example feature such as, risk rating in this dataset is derived from sales representative or managers beliefs.

The decisions made based on these tainted results tend to have far-reaching consequences often impacting individuals and societies. For instance, demographic factors such as gender, ages, zip code as well as other features used for market segmentation modeling to predict or forecast socio-economic representations can led to discrimination in the distribution of resources. Another use of machine learning algorithms that penalizes certain demographic characteristics by assigning low scores, suppresses fair representation and minimizes feature importance. Therefore, algorithmic transparency is key to win trust in model usage when these factors are involved.

Therefore, the cost of misclassifying customers during the clustering and regression model building process could have dire consequence in which the accuracy for the model can be misleading hence could result in a huge financial loss if the model is implemented. It is therefore recommended that the model's algorithm be constantly fed with retrained data when new data becomes available to get the optimum feature importance to use. It must also be noted though data for this research was given by HairCo and the demographic part of the data was scraped from publicly available data sites.

From an ethical perspective, data scientist and machine learning experts using demographic data that informs decisions should be guided by "the Code" as espoused in the ACM Code of Ethics and Professional Conduct [49]. The code of conduct articulates the need for high professional conduct by computing professionals whose actions and inactions change the world, to act responsibly by reflecting the wider impact their work has in support

the greater good of society. In light of these ethical considerations as spelt out in "the Code", this research adhered to high professional, ethical standards as expected and exhibited competency in using data for this study. It is for these reasons that data handling and privacy were taken into serious consideration to protect all stakeholders. Also, since this study does not seek to profit in any commercial applications, the use of material and data is well within the conditions of fair use laid out by HairCo and all ethical concerns have been resolved.

6 Conclusion

In conclusion, machine learning is a growing field of statistics and computer science with endless applications across every human endeavor. The use of machine learning is breaking barriers never experienced before as well as helping expand the frontiers of the scientific and business communities.

The power of machine learning is brought to bear on an industry that is often overlooked given that data is either non-existent or is limited due to the robust competition among players and the sensitivity of data available. However, data scientists or data engineers with the right skillset can optimize data to drive the business operation that links customers directly to products that increase a company's wealth significantly. Data for this research was provided by HairCo and was subject to an NDA hence some details cannot be made public. This is a justification for future works should more data become publicly available.

As noted during cluster classification, 55% of the features deemed important by the model for cluster assignment were external data. As a future endeavor, HairCo can collect this information for potential customers in order to align them to the closest customer profile. This will allow the company to use similar marketing strategies to increase conversion rates.

The results from this work show that regardless of the business type or profession, data and statistical models can be leveraged to solve complex problems otherwise humanly impossible to improve business processes that are cost effective, will maximize profit and growth.

Acknowledgments. The authors want to extend a profound gratitude to our advisors Gordon E. Berry and Jacob Drew (PhD) for taking time out of their busy schedule to guide and work with us on this paper.

References

1. S. Kumar, "Exploratory analysis of global cosmetic industry: major players, technology and market trends," Technovation, vol. 25, no. 11, pp. 1263–1272, 2005.
2. C. Europe, "Socio-economic contribution of the European cosmetics industry," Cosmetics Europe, Brussels, 2018.
3. M. Intelligence, "Smart Wearable Market-Growth," Trends, and Forecast (2019-2024), 2019.
4. A. Ehlinger-Martin, A. Cohen-Letessier, M. Taïeb, E. Azoulay, and D. du Crest, "Women's attitudes to beauty, aging, and the place of cosmetic procedures: insights from the QUEST Observatory," Journal of Cosmetic Dermatology, vol. 15, no. 1, pp. 89–94, 2016.
5. R. Hornsey, "'The modern way to loveliness': middle-class cosmetics and chain-store beauty culture in mid-twentieth-century Britain," Women's History Review, vol. 28, no. 1, pp. 111–138, 2019.
6. P. Black, The beauty industry: Gender, culture, pleasure. Routledge, 2004.
7. G. Jones, Beauty imagined: A history of the global beauty industry. Oup Oxford, 2010.
8. J. Winship, Inside women's magazines. Pandora, 1987.
9. K. Peiss, Hope in a jar: The making of America's beauty culture. Macmillan, 1999.
10. J. A. Willett, Permanent waves: The making of the American beauty shop. NYU Press, 2000.
11. L. A. Linnan and Y. O. Ferguson, "Beauty salons: A promising health promotion setting for reaching and promoting health among African American women," Health education & behavior, vol. 34, no. 3, pp. 517–530, 2007.
12. A. Łopaciuk and M. Łoboda, "Global beauty industry trends in the 21st century," in Management, knowledge and learning international conference, 2013, pp. 19–21.
13. G. Jones, Globalizing the beauty business before 1980. Division of Research, Harvard Business School, 2006.
14. R. L. Oliver and J. E. Swan, "Consumer perceptions of interpersonal equity and satisfaction in transactions: a field survey approach," Journal of marketing, vol. 53, no. 2, pp. 21–35, 1989.
15. R. A. Spreng, S. B. MacKenzie, and R. W. Olshavsky, "A reexamination of the determinants of consumer satisfaction," Journal of marketing, vol. 60, no. 3, pp. 15–32, 1996.
16. V. A. Zeithaml, L. L. Berry, and A. Parasuraman, "The behavioral consequences of service quality," Journal of marketing, vol. 60, no. 2, pp. 31–46, 1996.
17. S. Auh, L. C. Salisbury, and M. D. Johnson, "Order effects in customer satisfaction modelling," Journal of Marketing Management, vol. 19, no. 3–4, pp. 379–400, 2003.
18. D. Gimlin, "Pamela's place: Power and negotiation in the hair salon," Gender & Society, vol. 10, no. 5, pp. 505–526, 1996.
19. L. Jacobs-Huey, From the kitchen to the parlor: Language and becoming in African American women's hair care. Oxford University Press, 2006.

20. U. J. Brown III and R. L. Beale, “Services marketing: The mediating role of customer satisfaction in the hair care industry,” Academy of Marketing Studies Journal, vol. 12, no. 1, p. 57, 2008.
21. R. L. Oliver, “Whence consumer loyalty?,” Journal of marketing, vol. 63, no. 4_suppl1, pp. 33–44, 1999.
22. A. Chaudhuri and M. B. Holbrook, “The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty,” Journal of marketing, vol. 65, no. 2, pp. 81–93, 2001.
23. D. F. Spake, S. E. Beatty, B. K. Brockman, and T. N. Crutchfield, “Consumer comfort in service relationships: Measurement and importance,” Journal of Service Research, vol. 5, no. 4, pp. 316–332, 2003.
24. M. J. Bitner, “Servicescapes: The impact of physical surroundings on customers and employees,” Journal of marketing, vol. 56, no. 2, pp. 57–71, 1992.
25. E. O. Amoakoh and M. N. Naong, “The relevance of relationship marketing model for hair salon’s competitiveness: a theoretical perspective,” Problems and perspectives in management, no. 15, Iss. 1, pp. 132–139, 2017.
26. M. Robinson, “The disadvantages of transactional marketing over relationship marketing,” Internet Marketing Training, 2012.
27. M. Cosic and M. D. Djuric, “Relationship marketing in the tourist services sector,” UTMS Journal of Economics, vol. 1, no. 1, pp. 53–60, 2010.
28. J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann, 1988.
29. Y. Horita and H. Yamashita, “Bayesian network considering the clustering of the customers in a hair salon,” Cogent Business & Management, vol. 6, no. 1, p. 1641897, 2019.
30. H. Weerasinghe and D. Vidanagama, “Machine learning approach for hairstyle recommendation,” in 2020 5th International Conference on Information Technology Research (ICITR), 2020, pp. 1–4.
31. A. C. Müller and S. Guido, Introduction to machine learning with Python: a guide for data scientists. “O’Reilly Media, Inc.,” 2016.
32. Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, “What is machine learning? A primer for the epidemiologist,” American journal of epidemiology, vol. 188, no. 12, pp. 2222–2239, 2019.
33. S. Russell and P. Norvig, “Knowledge representation,” Artificial Intelligence: A Modern Approach, 3rd ed.; Prentice-Hall: Upper Saddle River, NJ, USA, pp. 437–479, 2010.
34. S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” Emerging artificial intelligence applications in computer engineering, vol. 160, no. 1, pp. 3–24, 2007.
35. G. Hinton and T. J. Sejnowski, Unsupervised learning: foundations of neural computation. MIT press, 1999.
36. O. Chapelle, B. Schölkopf, and A. Zien, “A discussion of semi-supervised learning and transduction,” in Semi-supervised learning, MIT Press, 2006, pp. 473–478.

37. R. A. Taylor et al., "Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach," Academic emergency medicine, vol. 23, no. 3, pp. 269–278, 2016.
38. Y. Zhu, L. Zhou, C. Xie, G.-J. Wang, and T. v Nguyen, "Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach," International Journal of Production Economics, vol. 211, pp. 22–33, 2019.
39. Q. Nguyen, I. Diaz-Rainey, and D. Kuruppuarachchi, "Predicting corporate carbon footprints for climate finance risk analyses: a machine learning approach," Energy Economics, vol. 95, p. 105129, 2021.
40. E. Tarallo, G. K. Akabane, C. I. Shimabukuro, J. Mello, and D. Amancio, "Machine learning in predicting demand for fast-moving consumer goods: Exploratory research," IFAC-PapersOnLine, vol. 52, no. 13, pp. 737–742, 2019.
41. J. Wu and S. Zheng, "Forecasting of fast fashion products based on extreme learning machine model and Web search data," J. Comput. Appl, vol. 2, pp. 146–150, 2015.
42. F. L. Chen and T. Y. Ou, "Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry," Expert Systems with Applications, vol. 38, no. 3, pp. 1336–1345, 2011.
43. Y. Wang, V. Chattaraman, H. Kim, and G. Deshpande, "Predicting purchase decisions based on spatio-temporal functional MRI features using machine learning," IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, pp. 248–255, 2015.
44. H. Ponce, L. Miralles-Pechúan, and M. de Lourdes Martínez-Villaseñor, "Artificial hydrocarbon networks for online sales prediction," in Mexican international conference on artificial intelligence, 2015, pp. 498–508.
45. Y. Kaneko and K. Yada, "A deep learning approach for the prediction of retail store sales," in 2016 IEEE 16th International conference on data mining workshops (ICDMW), 2016, pp. 531–537.
46. L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 856–863.
47. M. Lakoju, N. Ajienka, M. A. Khanesar, P. Burnap, and D. T. Branson, "Unsupervised learning for product use activity recognition: an exploratory study of a 'Chatty Device,'" Sensors, vol. 21, no. 15, p. 4991, 2021.
48. I. Suwalka and N. Agrawal, "P2-374: a hybrid unsupervised clustering technique for alzheimer's disease detection," Alzheimer's & Dementia, vol. 14, no. 7S_Part_15, pp. P839–P839, 2018.
49. R. E. Anderson, "ACM code of ethics and professional conduct," Communications of the ACM, vol. 35, no. 5, pp. 94–99, 1992.
50. StatsAmerica, "Social Context," https://www.statsamerica.org/downloads/user-guides/user-guide-social-context.pdf.
51. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.
52. I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments", doi: 10.1098/rsta.2015.0202.

53. O. Pfaffel, “FeatureImpCluster: Feature Importance for Partitional Clustering,” Available online: cran. r-project. org (accessed on 4 February 2021), 2020.
54. G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 559–563, 2017.
55. D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” Stanford, 2006
56. T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: an efficient data clustering method for very large databases,”
57. D. Sculley, “Web-scale k-means clustering,” in Proceedings of the 19th international conference on World wide web, 2010, pp. 1177–1178
58. American Community Survey 2019, https://www.census.gov/programs-surveys/acs/news/updates/2019.html (accessed 0n 04/03/2022 at 19:20)