Enhancing Automotive Customer Segmentation through Comprehensive Imputation Techniques

Vincent Tian (vinnyt@mit.edu)
Meredith Gao (mered405@mit.edu)

Fall 2023

Contents

1 Introduction

An automotive company aims to expand into new markets by introducing its current product lineup (P1, P2, P3, P4, and P5). Extensive market research indicates similarities between the behavior of the new market and the existing one. In their current market, the sales team has categorized customers into four segments (A, B, C, D) and implemented targeted outreach and communication for each segment, resulting in remarkable success. The company intends to apply the same successful strategy to the new markets, where they have identified 2627 potential customers.

Our task at hand is to assist the manager in accurately predicting the appropriate segment for these new customers.

2 Data Source

2.1 Description

The dataset [4] is provided on Kaggle and consists of 11 features and 8068 customers in the training dataset. It has insights into several parameters, for example, gender, marital status, age and profession. The distribution of the dependent variable is quite balanced: A (24%), B (23%), C (24%), D (28%). You can find the dataset *here*.

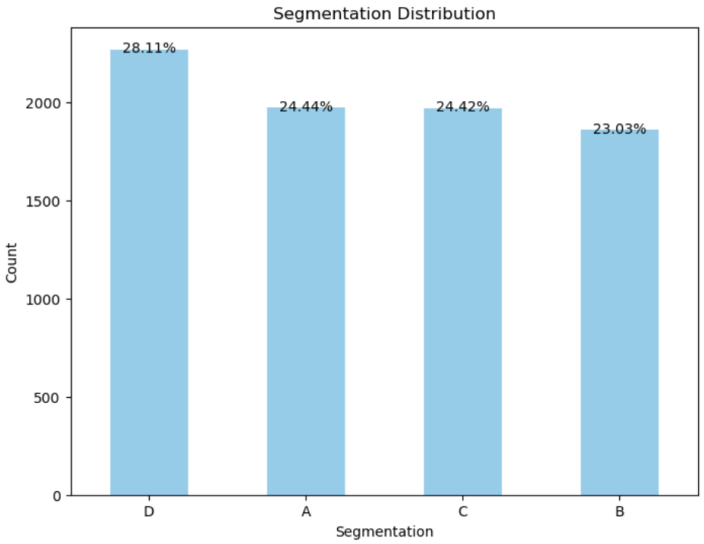

Figure 1: Class distribution in training set

2.2 Challenges

The primary challenge we face is the presence of missing values in both our training and testing datasets, which could be attributed to incomplete record-keeping. 60% of our features have missing values: *Ever_Married*, *Graduated*, *Profession*, *Work_Experience*, *Family_Size* and *Var_1*. Among these, *Work_Experience* has the maximum percentage of missing values of 10.3%, which isn't sufficiently high to justify the removal of the entire column. Similarly, the percentage of rows containing missing values is 17.4%, which isn't low enough to warrant exclusively discarding rows that have missing data.

To address this challenge, our approach involves experimenting with various imputation methods. Following that, we partitioned a validation set and trained a Random Forest model on each imputed dataset to evaluate their performance on the validation data. To ensure thoroughness, we applied all imputation methods to all the models, enabling us to conduct a comprehensive analysis to determine which combination of imputation method and model resulted in the highest accuracy. Details are discussed in the following.

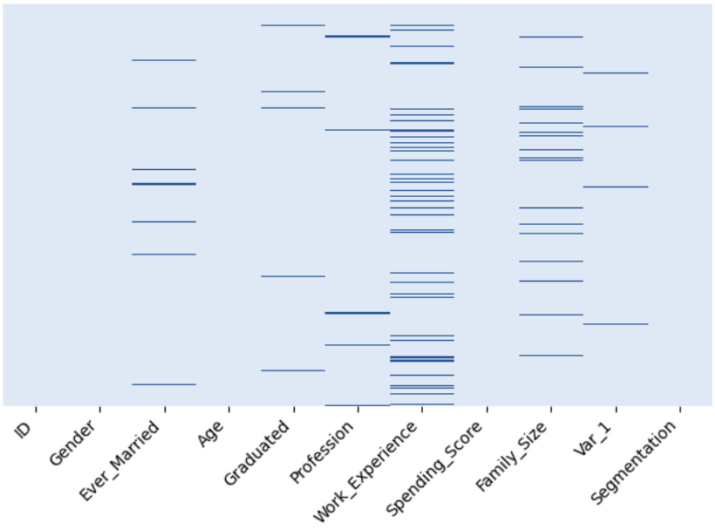

Figure 2: Missing data in training dataset

3 Imputation

We investigated four imputation methods: *mean* impute, *opt_knn*[1], *opt_svm*[1], *knn*.

We first cast categorical and integer columns: *Gender*, *Ever_Married*, *Graduated*, *Profession*, *Spending_Score*, *Family_Size*, *Var_1*, *Segmentation* to categorical types, then we stratified sampled 20% from the whole training data to be our validation set. We used *IAI.ImputationLearner* to train the four imputation models on the rest of the training data to impute both train and validation features. Finally, we utilized imputed data to train a base Random Forest model and evaluated validation accuracy to compare the performances of each imputation method. Training and validation performances for all four imputation methods are listed below.

Method	Training accuracy	Validation accuracy
mean	0.569	0.531
opt-knn	0.573	0.535
opt-svm	0.570	0.537
knn	0.570	0.532

Table 1: Training and validation accuracy for imputation methods.

We observed that all four imputations have similar performances while *opt_svm* has a slight edge. We decided to impute our data with all four methods to see if any of the methods could lead to a significant test accuracy.

Therefore, we used the whole training data to train each of the imputation models and then used the imputation model to impute train and test features. In this way, we obtained complete train and test data.

4 Data Pre-Processing

The initial step of data pre-processing involved an examination for collinearity among features, as it can significantly influence the performance of models like Logistic Regression or XGBoost. With just three numerical features in the dataset, it appears that there isn't a substantial correlation among them.

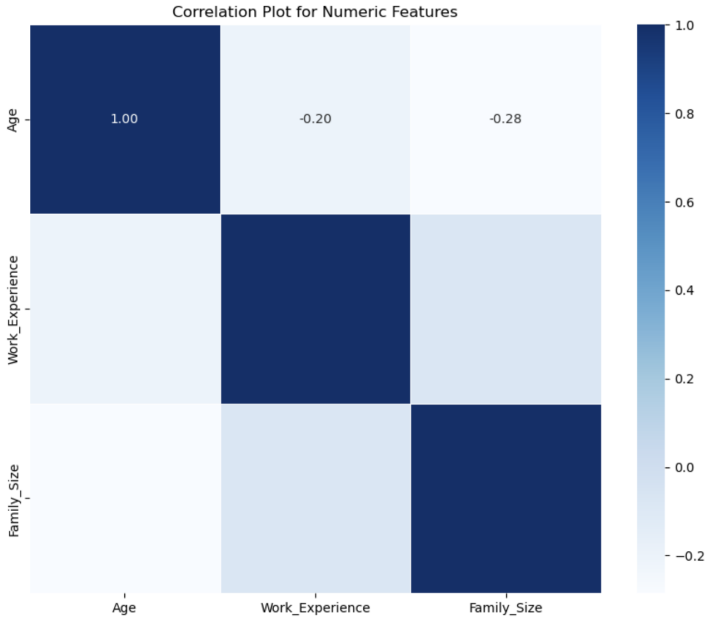

Figure 3: Correlation Plot

Additionally, there is significant variability within the *Age* and *Work_Experience* features. To address this, we applied min-max scaling, which transformed these features to a uniform scale ranging between 0 and 1.

To address the categorical features, the following transformations were applied:

- *Gender* - converted to a boolean variable, representing 'Male'.
- *Ever_Married* - converted to a boolean variable.
- *Graduated* - converted to a boolean variable.
- *Spending_Score* - values 'Low', 'Average', or 'High' were label encoded to 0, 1, and 2, respectively, to preserve the ordering relationship.
- *Profession* - Utilized One-Hot Encoding.
- *Var_1* - Utilized One-Hot Encoding.

5 Methods

5.1 Metrics

As our dataset is not unbalanced, accuracy is our metric of choice.

5.2 Modelling

The second objective of this project is to evaluate the performance of OCT (Optimal Classification Tree) in comparison to traditional models like Logistic Regression, CART, and Random Forest, as well as the state-of-the-art XGBoost model. The baseline model in this case entails predicting the most common class in the training set, which is Segmentation 'D'.

While maximizing model accuracy is a priority, interpretability holds equal significance in this project. The reason for this emphasis is our goal to provide clear explanations to the automotive company regarding the characteristics of customers within each of the four segmentations. This interpretability will enable them not only to gain insights into their customer base but also to devise effective marketing strategies for reaching potential customers.

While the ideal scenario would involve interpretable models like CART or OCT achieving the best performance, we may need to consider trade-offs if less interpretable yet more complex methods such as Random Forest or XGBoost yield superior results.

6 Results

6.1 Test Accuracy

These results are derived from the testing set and are based on the dataset imputed using either the *opt_knn*, *opt_svm*, *mean*, or *knn* method. All models are compared against the baseline, which involves predicting the most common segmentation, 'D'.

Notably, all models outperformed the baseline by approximately 100%, indicating the presence of discernible data relationships that can be effectively modeled.

A closer examination reveals that consistently using the *opt-knn* method for imputation yielded the best-performing models, with the Random Forest model achieving the highest overall performance, which improved from baseline model by 127.34%. In contrast, the *opt-svm* imputation method performed the least favorably, with Logistic Regression being the sole exception among the models.

Additionally, it is evident that utilizing the *mean* and *knn* imputation methods yielded similar results for Logistic Regression, Random Forest, and OCT. However, XGBoost and CART exhibited significantly improved performance when using the *knn* imputation. This observation aligns with expectations, as *knn* imputation offers greater complexity compared to *mean* imputation.

Unfortunately, the more interpretable models, CART and OCT, did not emerge as the top-performing models in this analysis. Nevertheless, they can still prove valuable in gaining insights into the characteristics of various customer segments.

Model	opt-knn	opt-svm	mean	knn
Baseline	0.278	0.278	0.278	0.278
Logistic Regression	0.527	0.531	0.523	0.525
Random Forest	0.632	0.595	0.600	0.596
XGBoost	0.569	0.566	0.574	0.593
CART	0.538	0.527	0.527	0.539
OCT	0.545	0.546	0.544	0.547

Table 2: Testing accuracy for different classification methods for various imputation methods.

6.2 Feature Importance

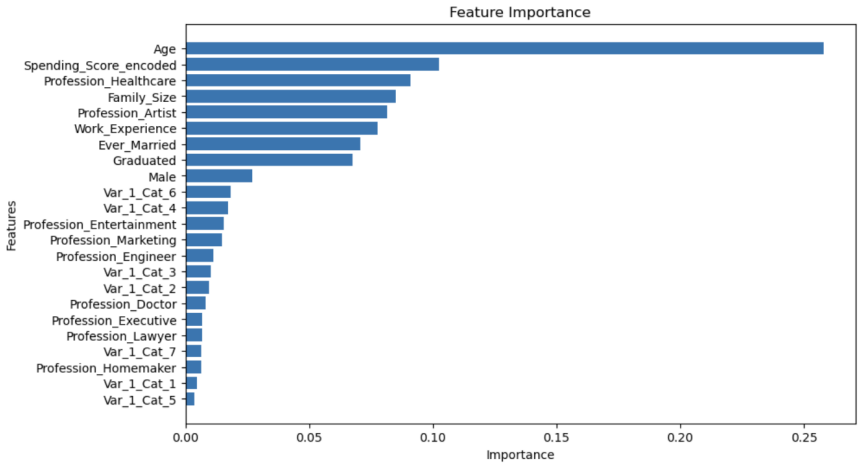

Figure 4: Feature Importance for Random Forest using opt-knn imputation

From the feature importance graph [3] we can see that the most important features to determine customer segmentations are *age*, *spending score* and *profession*.

Age: Being the most influential feature suggests that different age groups are likely to belong to different customer segments. For instance, younger customers might be more inclined towards certain types of purchases or interests compared to older customers.

Spending Score: This implies that the way customers are scored based on their spending habits is crucial for segmentation. A high spending score might indicate a segment with more disposable income or a propensity for luxury goods, while a lower score might indicate a more cost-conscious segment.

Profession (Healthcare, Artist, etc.): Different professions may correlate with distinct spending patterns, interests, or needs, which is why they appear as significant features for customer segmentation.

6.3 SHAP

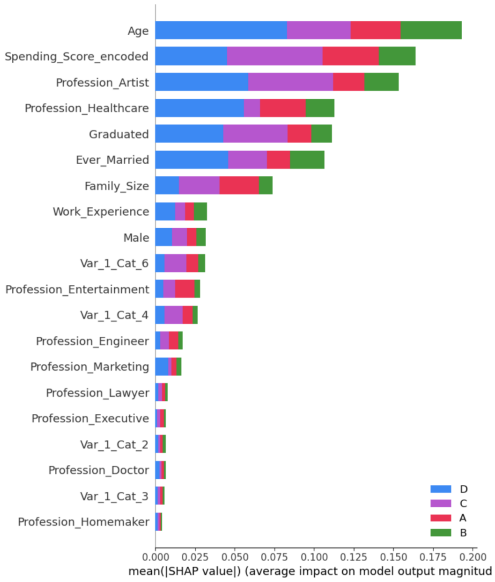

Figure 5: SHAP summary plot for Random Forest using opt-knn imputation

The SHAP value graph [2] indicates the impact of various features on the prediction of a customer segmentation model. Each bar's length and colour show how a particular feature affects the segmentation across different customer groups, A, B, C, and D. Here are some interesting insights from the graph, which align with our insights from the above feature importance graph.

Age: Strongly influences all segments, with the greatest impact on segment D.

Profession: Various professions have varying effects on different segments. For example, the 'Profession_Artist' and 'Profession_Healthcare' have a strong influence on segment D, possibly indicating a unique set of preferences or financial behaviours associated with these professions within this segment.

Graduated and Ever_Married: These features show a significant impact on certain segments but not others. Marital status and educational background may correlate with lifestyle choices that are important for segmentation.

Family_Size and *Work_Experience*: These have smaller but still noticeable effects across segments. Larger family sizes might dictate the need for different products and services, while work experience might correlate with income levels and spending power.

Male: This feature has an influence, suggesting that gender may play a role in segmenting customers, possibly due to differences in product preferences or spending habits.

6.4 Interpretation of Segmentations

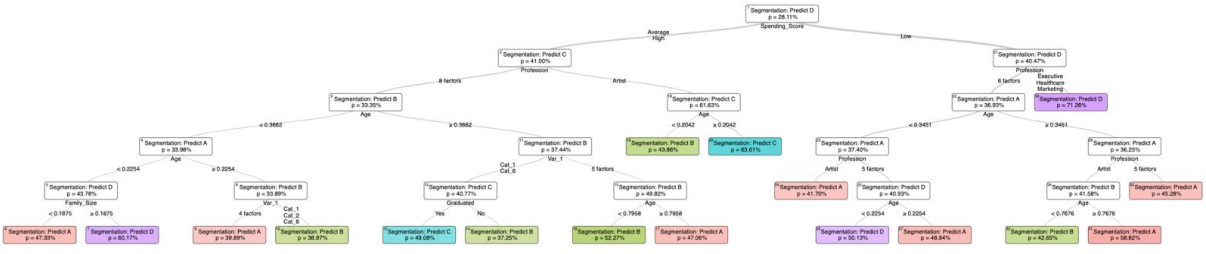

Figure 6: Optimal Classification Tree interpretation on CV set imputed by oct_svm

Segmentation A: Older customers with higher spending scores, especially Executives, or mid-aged Artists with moderate spending scores

Segmentation B: Generally older customers with lower spending scores, and younger Artists with higher spending scores

Segmentation C: Educated customers with lower spending scores

Segmentation D: Younger customers with low spending scores and small family sizes, or high spending individuals in Healthcare or Executive roles

The OCT result suggests a model that classifies customers into four segments based on a combination of spending score, age, family size, and profession, which is similar to the set of top important features generated from Random Forest. Spending score is a primary differentiator at the highest level, with professional role and age acting as subsequent discriminators. Customer segments are also influenced by whether the customer is graduated, which anonymized category Var_1 the customer is in, and family size, particularly in the lower spending score categories. Customers with higher spending scores are more distinctly segmented by their profession and age. Additionally, Artists always have unique characteristics that distinguish them from other professions. The probabilities associated with each segment provide

insight into the model's confidence in its predictions and suggest varying levels of homogeneity within each segment.

7 Conclusion

In conclusion, our analysis emphasizes the significance of imputation methods in addressing missing data within the customer segmentation dataset. By experimenting with various imputation strategies: opt-knn, opt-svm, mean, and knn, our study has underscored the impact of data completeness on the performance of classification models. The nuanced comparison of these methods revealed that while the opt-svm method provided a marginal advantage in validation accuracy, the opt-knn method consistently enhanced model performance on the test set, particularly with the Random Forest model.

This exploration of imputation techniques has been crucial in refining the dataset, which in turn, has facilitated the accurate segmentation of potential customers. It demonstrates the importance of methodical pre-processing in predictive modeling and offers a methodical approach for similar challenges in future projects. The project ensures the integrity of the segmentation analysis, enabling the automotive company to extend its tailored outreach strategies to new markets with confidence.

The analysis on the interpretation from Random Forest and Optimal Classification Tree (OCT) has greatly improved our understanding, highlighting important features of each customer group. Insights from the OCT show that spending levels, age, family size, and jobs are important in telling customers apart. This clear understanding is very useful for the company to plan specific marketing strategies for different groups. Combining proper imputation methods with this kind of analysis, the project provides strong conclusions that help with the company's plans to enter new markets and offer a useful approach for future projects in understanding customer groups.

8 Individual Contribution

Both team members made equal contributions to various aspects of the project.
Meredith's primary focus was on data imputation and OCT training:

- exploratory data analysis
- stratified sampling to get training and validation sets to evaluate each imputation models
- mean impute, opt-knn, opt-svm, knn imputation training
- OCT classification model training
- OCT's interpretation of each segmentation

Vincent managed data preprocessing and was responsible for the majority of the modeling work:

- exploratory data analysis
- data pre-processing
- training classification models: Baseline, Logistic Regression, Random Forest, XGBoost, and CART
- determine feature importance and SHAP values from the random forest results

Additionally, we collaborated on exploring and evaluating imputation models, analyzing the results collectively.

References

[1] Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research*, 18(1):7133–7171, 2017.

[2] Idit Cohen. Explainable ai (xai) with shap: Multi-class classification problem. 2023. Accessed: [2023].

[3] Terence Shin. Understanding feature importance and how to implement it in python. 2023. Accessed: [2023].

[4] Kaushik Suresh. Customer segmentation. `https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation/data`, 2023. Accessed: [2023].