Segmenting the Travelers: A Classification-Based Study on Tour Purchase Behavior

Tara Lehr, Akku Jacob Shaji, Jigyasa Soni, Vyshali Poola

Oakland University

MIS-4560-12908 / MIS-5560-13842

Introduction to Data Science

Dr. Hongyu Gao

April 17th, 2025

Contents

Segmenting the Travelers: A Classification-Based Study on Tour Purchase Behavior

1. Project Description

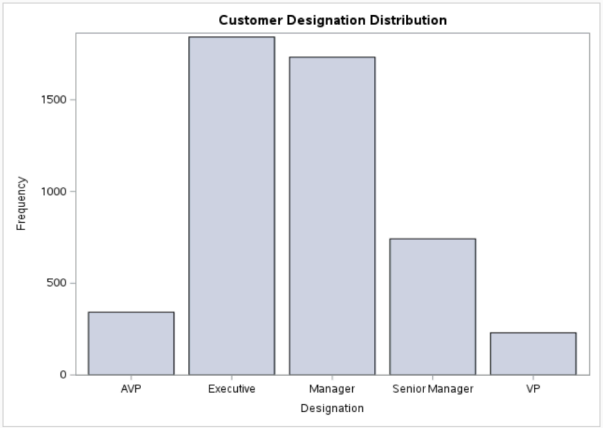
This project was developed in the context of a direct marketing scenario for a tour package company. The dataset used, tour_package.csv, contains customer demographic and interaction data, including income, age, family characteristics, and engagement with sales efforts such as the duration of promotional pitches and the number of follow-ups.

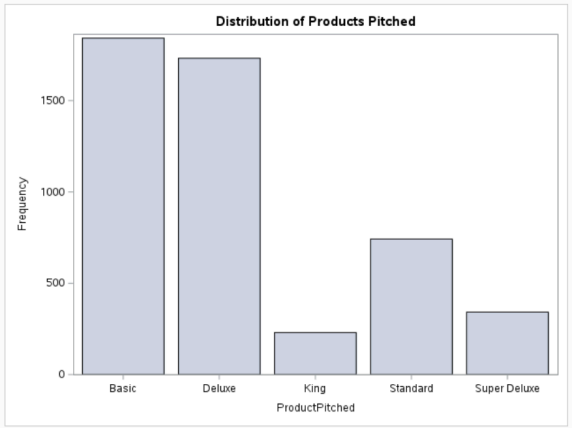
Our related business problem centers on improving marketing efficiency by identifying which customers are most likely to purchase a tour package. Since direct marketing can be costly, the company must ensure that promotional efforts are targeted toward individuals with a higher likelihood of conversion.

Furthermore, the main goal of the project is to build and evaluate predictive models that classify customers based on whether they are likely to purchase the tour package. Using classification and regression trees, logistic regression, and other supervised and unsupervised techniques, the project aims to uncover patterns in customer behavior, segment the customer base, and ultimately provide insights that can guide future marketing decisions.

2. Data Exploration and Preprocessing

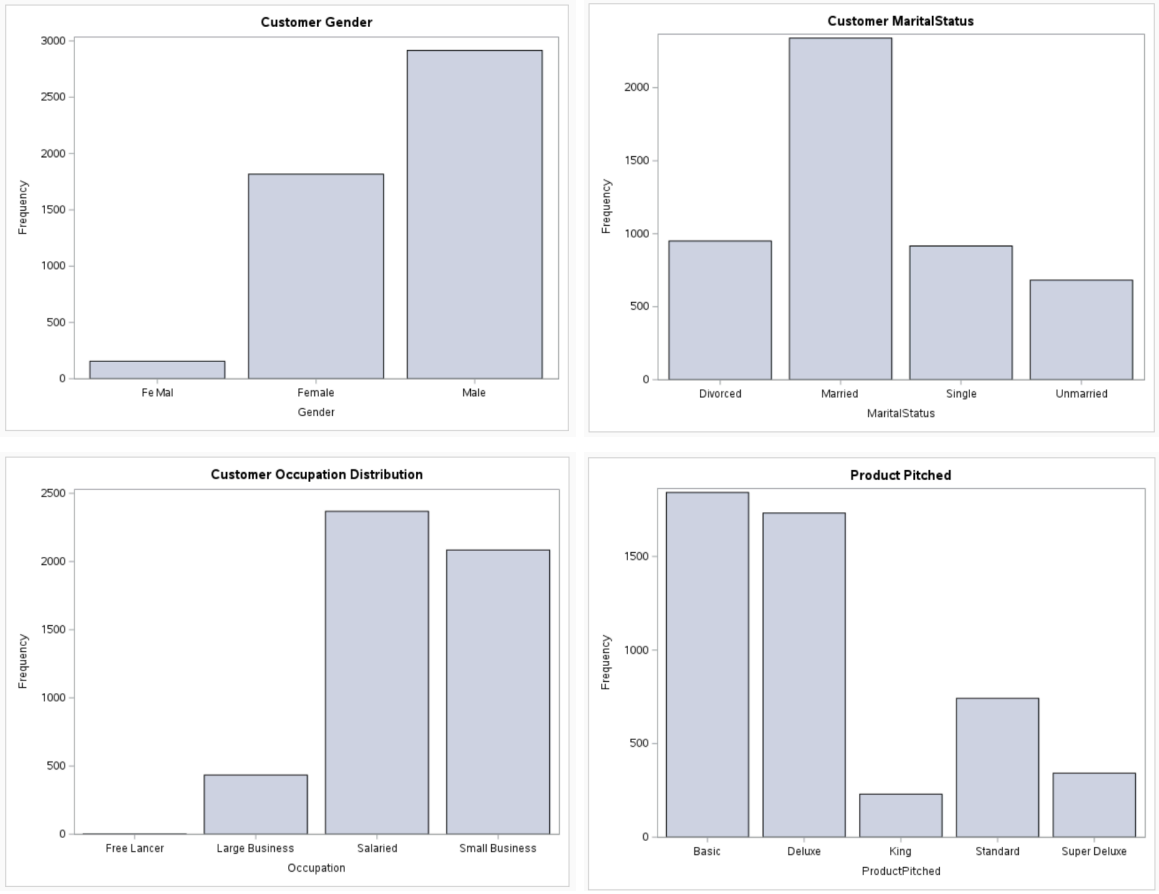
The analytical target in this study consists of ProdTaken which represents customer tour package purchase status as either one or zero. Multiple numerical variables and categorical variables form the input specifications (predictors) which include Age, MonthlyIncome, DurationOfPitch, NumberOfTrips, MaritalStatus, Gender, ProductPitched, TypeofContact, Occupation and Designation together with other fields.



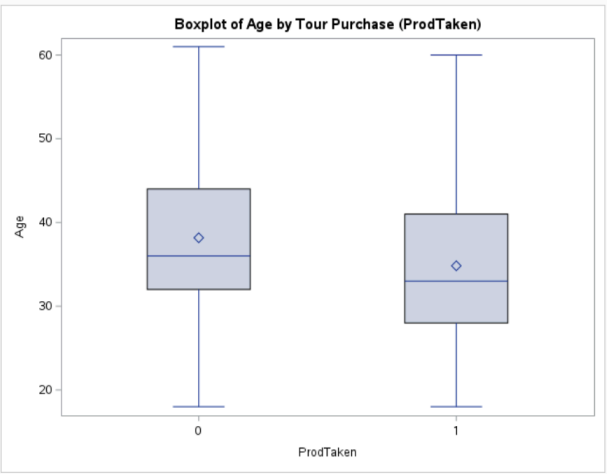


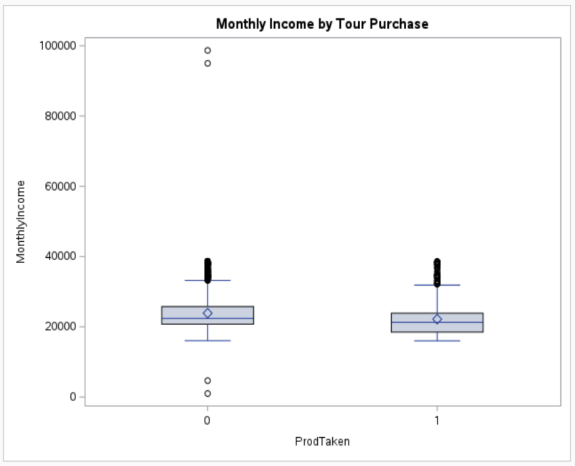
An initial data survey involved producing summary details and creating visualization displays of key data points through histograms and boxplots and frequency bar charts. The

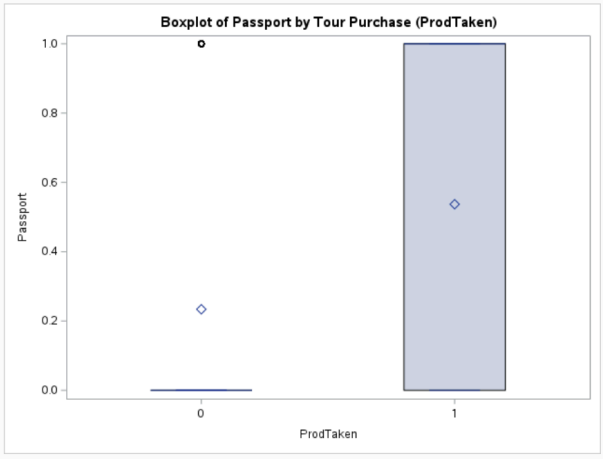
graphical representations displayed prevalent patterns including income distribution skew and pitch satisfaction measures and demonstrated that Salaried workers made up most of the sample group and Self Enquiry contacts were most common.

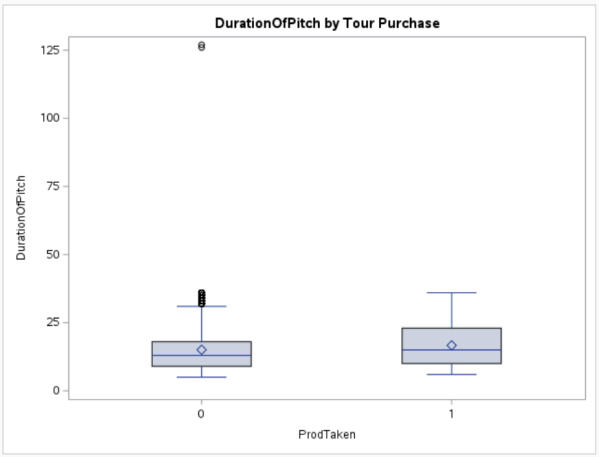


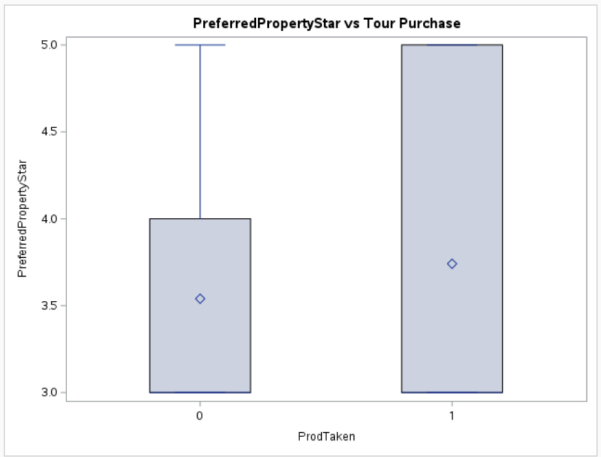
The dataset investigation showed us that Age, MonthlyIncome and DurationOfPitch contained various amounts of missing values with 226 for Age, 233 for MonthlyIncome, and 45 for DurationOfPitch. We used mean substitution to fill in missing numerical data points to keep most of the records intact though records containing essential categorical data needed for encoding were removed from analysis. The analysis revealed no cases of data errors consisting of out-of-range values or invalid categories which demanded exclusive correction procedures.











The analysis of numerical variable correlations indicated that no two variables showed high redundant relationships since their coefficients were below 0.8. Therefore, no duplicate variables needed removal. All predictive models utilized 28 final input features resulting from the transformation of MaritalStatus, ProductPitched, Designation, and Gender to dummy variables for model readiness. The dataset partition followed a 60:40 ratio selecting 2,199 cases

for training while using 734 cases for validation supported by a random seed value of 12345. All models operated under equal evaluation conditions through this partitioning which enabled a dependable performance assessment.

3. Models

The analysis presented here aims to find a response to whether a customer will purchase a tour package (ProdTaken = 1) or not (ProdTaken = 0). As a solution to this binary classification problem, we employed three supervised data mining techniques: Classification and Regression Tree (CART), logistic regression, and neural networks (NN). Because these are the methods most frequently encountered in this type of problem, they are also used by us for performance comparison across linear, non-linear, and tree-based models. To do this, we formatted a variety of models based on every technique, controlling and adjusting the parameters and increasing the complexity to increase the prediction accuracy.

3.1 Classification and Regression Tree (CART)

We implemented a Classification and Regression Tree (CART) model to predict whether a customer would purchase a tour package (ProdTaken). Two models were built using different splitting criteria:

Model	Details
1	Gini with all variables
2	Entropy with all variables

Each model was trained using a 60% training partition, and performance was evaluated using 5-fold cross-validation and a 40% test set. Cost-complexity pruning was applied to avoid overfitting.

We evaluated both models using Confusion matrix, ROC curve and AUC, Accuracy, Sensitivity, Specificity shown in Table 3.1.1.

Table 3.1.1 - Model Performance Metrics								
	Misclassification		**Sensitivity**		**Specificity**		**AUC**	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Model 1	11.83	16.14	95.96	42.58	54.86	93.27	82.06	81.98
Model 2	11.49	15.11	96.34	42.58	55.04	94.28	83.73	83.73

Based on the results, Model 2 (CART using Entropy) outperforms Model 1 (Gini) across almost every metric (Table 3.1.1). It has lower misclassification rates, slightly higher sensitivity, specificity, and precision, as well as a significantly higher AUC on both training and validation; therefore, Model 2 is considered the best model across both the training and validation sets, as it consistently outperforms Model 1 based on above criteria.

3.2 Logistic Regression

The second method used for model building is logistic regression. While the sample report explored six models using various selection strategies and sampling techniques, in this analysis, only two models were developed, as shown in Figure 3.2.1. This decision was based on efficiency and model stability. During the full model run, a quasi-complete separation warning was triggered, indicating that one or more predictors nearly perfectly separated the outcome variable, and this condition can lead to unstable coefficient estimates, poor generalization, and convergence issues, particularly in datasets with a large number of categorical dummy variables.

To address this and streamline the model-building process, a stepwise selection procedure was used. According to SAS documentation, the stepwise method combines both forward selection and backward elimination to iteratively add or remove predictors based on statistical significance (SAS Institute, 2024). Therefore, this method avoids the need to manually test multiple selection strategies while still identifying the most important variables in a statistically rigorous way.

Figure 3.2.1 Model Descriptions	
Model	**Details**
1	No selection method, used all variables (Full Model)

Figure 3.2.1 Model Descriptions	
2	Stepwise selection method, used all variables with entry/stay = 0.05

The full model (Model 1) included all available predictors, including dummies for designation, marital status, product pitched, and occupation. As mentioned, the model raised a quasi-complete separation warning and demonstrated some unstable parameter estimates. Despite this, the model performed well based on overall accuracy and AUC. However, interpretation was challenging due to the number of variables and redundancy among dummies.

Model 2 used stepwise selection to automatically identify the most statistically significant predictors. Sixteen predictors were retained, including Passport, PreferredPropertyStar, NumberOfFollowups, Designation_Executive, Gender_Male, and MaritalStatus_Single, among others. Now, compared to the full model, the stepwise model offered nearly identical performance but with fewer variables and no convergence warnings.

Figure 3.2.2 Model Performance Metrics

Model	Model Accuracy	Sensitivity	Specificity	AUC
1	84.04%	27.75%	96.92%	0.797
2	84.19%	28.02%	97.05%	0.796

Figure 3.2.2 summarizes the performance of both models. Although both achieved similar results in accuracy and AUC, Model 2 (stepwise) was chosen for final comparison with other data mining techniques due to its simplicity, stability, and interpretability.

Additionally, under-sampling techniques were not pursued in this analysis. Unlike the sample report, which addressed a 95-to-5 class imbalance, this dataset was more moderately imbalanced, with approximately 81% of observations classified as 0 and 19% as 1. Because of this, model sensitivity remained within acceptable bounds, and the lift chart showed that high-probability customers could still be effectively identified without the need for rebalancing.

3.3 Neural Networks (NN)

The third model used to determine tour package subscriptions for the tourism industry, providing customer preferences and enabling businesses to optimize their offerings was Neural network. Our objective in this analysis is to evaluate the performance of two neural network models (Model A and Model B) in predicting whether a customer will opt for a tour package

("ProdTaken"). We aim to determine which model provides better predictive accuracy and generalization ability using various performance metrics such as misclassification rate, sensitivity, and specificity.

The dataset used in both models consists of 2,933 customer observations, which were split into 2,199 for training and 734 for validation. The target variable is binary: where 1 indicates that the customer opted for the product (tour package); and 0, which indicates that the customer did not. Each model uses 28 input variables, including demographic, behavioral, and engagement data.

Table 3.3.2 Model Architecture

Model	Hidden Layers	Neurons per Layer	Optimization	Notes
A	1	5	Limited-Memory BFGS	Original configuration
B	2	8	Limited-Memory BFGS	Increased complexity

Both models utilize a Multilayer Perceptron (MLP) architecture, employing standard training procedures with a maximum of 50 iterations.

Table 3.3.3 Fit Statistics

Metric	Model A	Model B
Train Misclassification Rate	15.42%	16.23%
Validation Misclassification Rate	19.48%	20.30%
L1 Norm of Weights	116.87	41.91

Table 3.3.4 Confusion Matrix (Validation Set)

	Actual 1	**Actual** 0	**Total**
Predicted 1	Model A: 79	Model B: 88	Model A: 285 Model B: 276
Predicted 0	Model A: 72	Model B: 63	Model A: 1519 Model B: 1528

Table 3.3.5 Sensitivity and Specificity		
Metric	**Model A**	**Model B**
Sensitivity	21.7%	24.2%
Specificity	95.5%	96.0%

Model B shows slightly better sensitivity and specificity compared to Model A.

Table 3.3.6 ROC and Logistic Regression Evaluation		
Model	**AUC** (Area Under the Curve)	**95% Confidence Interval**
Model A	0.7573	(0.7301, 0.7845)
Model B	0.7336	(0.7047, 0.7625)

Model A slightly overperforms Model B as per the overall ROC area. This shows that Model A has better discrimination capability.

Table 3.3.7 Logistic Regression (Post NN predicting score)		
Statistic	**Model A**	**Model B**
Odds Ratio	64.40	106.75
Log Likelihood Improvement	240.88	217.97

Model B shows a much higher value of odds ratio, which means that it's predicted scores differentiate classes more distinctly than Model A. However, both neural network models perform similarly in terms of misclassification, but there are some nuanced differences: Model A (simpler) achieves better AUC and slightly lower misclassification where Model B (more complex) produces higher sensitivity and odds ratio.

Considering that our business goal is to accurately identify customers who will take the tour, Model B would be our preferred option/recommendation due to its higher sensitivity and stronger classification strength based on logistic regression output, despite the marginally lower AUC.

3.4 Unsupervised Learning: Clustering Models

To uncover natural customer segments, both hierarchical and partition-based clustering techniques were initially explored. The analysis began with hierarchical clustering using agglomerative methods, where each record starts in its own cluster and groups are merged iteratively. While useful for visualizing grouping tendencies, the method produced inconclusive results: the dendrogram lacked a clear cutoff, and pseudo-statistics failed to indicate a strong natural grouping structure.

As a result, the analysis shifted to the more scalable and flexible k-means clustering, implemented using the FASTCLUS procedure in SAS. K-means was selected because it minimizes within-cluster variation while maximizing separation, better known as the ratio of between-cluster to within-cluster variance. These two metrics, along with stability (consistent cluster assignment across samples), are emphasized in lecture materials as key to evaluating clustering quality. Thus, the ability of k-means to balance both goals made it the ideal approach for large-scale partitioning in this dataset.

3.4.1 Model Formulations

To determine the optimal number of clusters (k), the FASTCLUS procedure was run iteratively from k = 1 to 10, generating cluster performance metrics such as pseudo F-statistics, observed R-squared, and Cubic Clustering Criterion (CCC). These values were evaluated using the elbow method, which visually identifies the point at which additional clusters yield diminishing returns in model improvement.

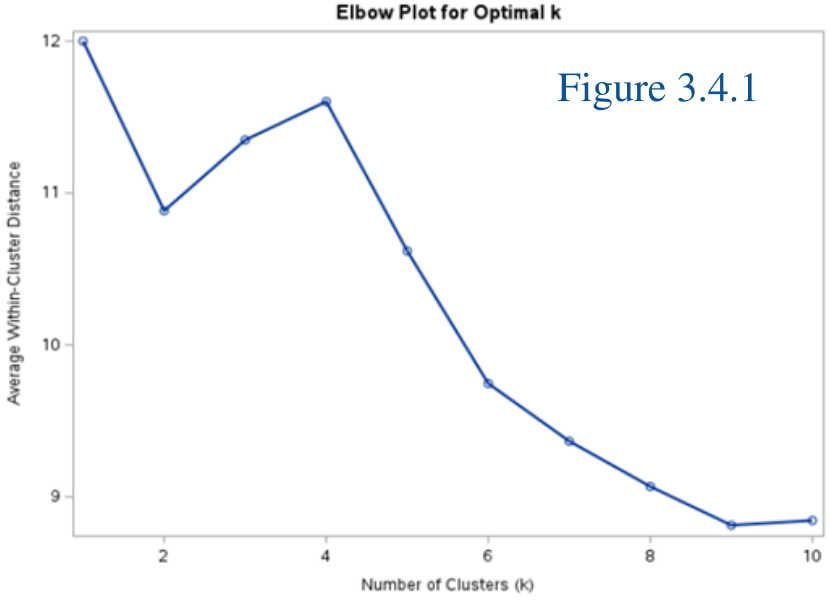


Figure 3.4.1

As shown in Figure 3.4.1, a clear inflection occurred at k = 6, suggesting that a 6-cluster solution offered the best trade-off between performance and interpretability. The model was then re-run with MAXCLUSTERS=6 to finalize cluster assignments and calculate centroids.

3.4.2 Model Specifications

The final k-means model was built using eleven standardized numeric variables, including Age, MonthlyIncome, NumberOfTrips, PitchSatisfactionScore, and other behaviorally relevant features. Variables were standardized to ensure equal influence in Euclidean distance calculations, and those with minimal variance were excluded unless they contributed to meaningful cluster differentiation.

Moreover, the final configuration used MAXITER=100 and CONVERGE=0.02 to ensure the model fully stabilized. The resulting cluster summary (Table 3.4.2) showed that most groups had well-balanced sizes (e.g., 1000+ records), while only three clusters had fewer than five observations, likely representing outliers. Importantly, centroid distances exceeded 1.98 across all clusters, signaling strong separation between customer segments.

Table 3.4.2 K-means Clustering Metrics

	Initial Model w/ k =10	**Optimal Model w/ k =6**
Pseudo F	194.72	227.01
R-Squared	0.2643	0.1886
CCC	- 52.331	- 36.182
Average Cluster Separation	≈ 4.85	≈7.21

3.4.3 Model Comparison and Recommendation

Although hierarchical clustering provided a useful starting point, it lacked the statistical support and interpretability needed for segmentation. Instead, k-means allowed for a more rigorous comparison of candidate models. As shown in Figure 3.4.2, the k = 10 model exhibited a higher R-squared (0.2643) but suffered from interpretability issues. Multiple clusters, such as Cluster 8 and 9, had only two observations, and centroid distances exceeding 11.8, suggesting extreme outliers and fragmentation of the customer base.

In contrast, the k = 6 model achieved the highest pseudo-F-statistic (227.01), a more stable CCC (-36.182), and a higher average cluster separation (≈7.21). These results, combined with large, interpretable cluster sizes and visually confirmed grouping through the elbow method, made the six-cluster model a superior choice. All that being stated, based on performance metrics, stability, and interpretability, the 6 cluster k-means model is recommended as the final segmentation solution because it offers meaningful differentiation among customer groups while minimizing noise and overfitting.

4. Results and Discussion

Our analysis used Classification and Regression Trees (CART) combined with logistic regression and neural networks models to evaluate which solution method best identifies tour package purchasing potential (ProdTaken = 1). The evaluation of the models relied on standard classification metrics (Table 4.1) which encompassed misclassification rate together with sensitivity and specificity and AUC (Area Under the Curve) according to the sample report structure.

Table 4.1 **Model**	**Misclassification**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
CART (Model 2)	15.11%	84.89%	42.58%	94.28%	83.73%
Logistic Regression (Model 2)	15.81%	84.19%	28.02%	97.05%	79.60%
Neural Network	20.30%	85.31%	24.2%	96.0%	73.36%
Clustering w/K-means	N/A	86.27%	47.65%	94.52%	86.50%

A validation misclassification rate of 15.11% together with sensitivity at 42.58%, specificity at 94.28% and an AUC of 83.73% was obtained by the best CART model which used entropy (Model 2). The created model achieved high accuracy with good detection rates of actual purchasers. The stepwise logistic regression model reached the same overall accuracy level of 84.19% but exhibited lower sensitivity at 28.02% as compared to its specificity of 97.05%. This outcome corresponds to findings in the original research report. The simplicity and robustness of logistic regression cannot replace its limited capability to identify positive cases when applying analysis for customer conversion targeting. Neural network models analyzed the relationship between bias and variance. The prediction outcomes for Model A displayed 19.48% validation misclassification rate and 21.7% sensitivity together with 0.7573 AUC. Model B (2 hidden layers, 8 neurons) delivered 24.2% sensitivity and 96.0% specificity, yet its validation AUC value reached 0.7336. The discrimination abilities of Model A were better but Model B produced superior results for actual purchaser identification which was vital for management. According to both the sample report and the evaluation sensitivity requirements were determined as crucial when dealing with class imbalance scenarios.

The segmentation of customers occurred through k-means clustering as part of unsupervised learning processes. The best cluster solution of six was chosen through combination of the elbow method together with clustering metrics (pseudo F = 227.01, CCC = -36.182). Unsupervised learning through k-means clustering detected customer groups that naturally formed due to shared behavioral and demographic aspects which included income level along with trip frequency and satisfaction with pitch support levels. These findings offer clear and actionable marketing recommendations in which travel companies should focus their direct marketing efforts on high-income, frequent travelers while also personalizing follow-up strategies for engaged but undecided customers. On the other hand, premium tour packages should be strategically offered to individuals who show a preference for luxury options, allowing marketers to align products with customer expectations.

In addition, the identified segments let marketing teams optimize their outreach tactics through specific strategies which include directing maximal efforts toward likely clusters and creating offers according to customer demographics and past activities. The research indicates that managerially speaking companies need to consider customer targeting through identifying most likely buyers and delivering tailored approaches for dealing with different customer segments. Overall, we used these models for our analysis to create a data-powered platform and optimize campaigns to achieve improved operational performance and lower the expenses while also enhancing the conversion possibilities.

5. Summary

Performing this project demonstrated beneficial knowledge about how predictive modeling and clustering functions to assist in data driven decisions within the business context. Throughout this task we gained practical skills in implementing each step of the data science lifecycle through data cleaning and transformation as well as multiple algorithm evaluation. The project taught us how to utilize CART models which generate intelligible decision rules as well as logistic regression methods for spotting important predictors and neural networks with their strong but less explainable predictive abilities. We understood how class imbalance affects model performance while learning that reading sensitivity and AUC values requires combined analysis as stated in the sample report. By using unguided learning through clustering, we reached customer segmentation thus enabling specific marketing approaches. Every business requires unique modeling solutions since decisions about approach selection hinge on business aims and evaluation metrics alongside model interpretability requirements. Through this project we have been able to improve not just our technical aptitude but have been able to learn to demonstrate the use of analytical tools to support organizational targets.

References

SAS Institute Inc. (2025, April 9). *Stepwise Logistic Regression*. From SAS Documentation: https://documentation.sas.com/doc/en/pgmsascdc/v_061/statug/statug_logistic_examples01.htm

Data Source: https://www.kaggle.com/code/yogidsba/travelpackageprediction-ensemble-techniques/input