"What Makes Flying Enjoyable?"

Using Machine Learning to Analyze Which Factors are Linked with Airline Passenger Satisfaction

Estella Dentinger and Brittany Nguyen

15.095 Machine Learning Under an Optimization Lens

Fall 2022

I. Problem Statement

As the use of data to improve customer satisfaction has skyrocketed for industries varying from retail to hospitality, it comes at no surprise that analytics has found its way into the airline industry, too. Customer satisfaction surveys are very common, but there is room for improvement in how they are analyzed and used for making decisions. Perhaps one motivation for customer surveys is to be able to report how happy customers are. Knowing that the majority of customers who took the surveys are satisfied is a nice finding to report to senior leaders, especially if satisfaction increases over time. Using machine learning and optimization, specifically OCTs and interpretable clustering, however, provides an edge. Figuring out which aspects of the flight customers care most about, in addition to uncovering airline passenger archetypes is a powerful tool for making business decisions. Moreover, customer focus groups are timely and costly, but interpretable clustering can provide similar insights with less time and money. Overall, global airlines comprise an estimated $785 billion dollar industry, and thus a clear understanding of customer satisfaction is key.

In this project, we begin with a dataset which contains data regarding factors which affect the passenger flight experience on an individual customer level. Such variables include demographic data about the passenger (age, gender, etc.), flight data (class of seat, flight distance, gate location), and satisfaction levels with various experiences (wifi service, seat comfort, ease of online boarding, etc.). From the data, we defined a binary classification problem and built a variety of machine learning models in order to predict whether a customer would be “neutral or dissatisfied” versus “satisfied” with the overall flight experience. The chosen models were Logistic Regression, CART, Random Forests, XGBoost, and Optimal Classification Trees. Interpretable clustering was also performed at the end to analyze groups of customers. The performance of each method was compared. All computations were performed using Julia.

Our knowledge is that harnessing analytical techniques to predict customer satisfaction is of utmost importance for airlines to make the right allocations and decisions to improve the overall flight experience which they offer.

II. Data

The dataset obtained for this project was from Kaggle. The data is given in the form of both a training and testing set which had dimensions 103,904 rows by 25 columns and 25,976 rows by 25 columns, respectively. This data was merged into one dataframe under the conclusion that having more observations would make for more robust models trained on a wider variety of customer scenarios. It is also important to note that the data is obtained from a customer satisfaction survey, and thus the inherent subjectivity of the responses was taken into account when making conclusions.

The 22 independent predictors (three columns were related to customer ID and thus not used) are each predictive of whether a customer was satisfied with their flight versus neutral or dissatisfied. Some contained demographic information regarding the gender, customer type, and age of the passenger. Others centered around the flight. These were type of travel (personal versus business), class of the seating, flight distance, inflight wifi service, ease of online booking, departure and arrival convenience, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on-board service, leg room service, baggage handling, check in service, inflight service, cleanliness, departure delay in minutes, and arrival delay in minutes.

The final column was satisfaction (either "neutral or dissatisfied" or "satisfied"). A table of all variables and their descriptions is provided in the appendix in Table A.

As for the values of the different variables, the demographic variables (gender, customer type, age) had String data types and so did the class of seating. Meanwhile, the other flight-related variables had Integer or Float values ranging from 0 through 5 with 1 denoting a very bad experience and 5 denoting a very good experience (0 denotes "not applicable"). The departure delay in minutes and arrival delay in minutes variables ranged between 0 and approximately 1,500 each. The dependent variable column for satisfaction was a String type with two values: "satisfied" and "neutral or dissatisfied."

III. Methodology

The methodology utilized for this project can be summarized into several sections: data cleaning, splitting into training, validation, and test sets on which to fit the machine learning models, and finally modeling with cross validation to choose optimal hyper parameters. Results were then obtained for each model in the form of train AUC, validation AUC, train accuracy, and validation accuracy. The "best" model was deemed the one with the most optimal set of metrics in the validation step, and was fit on the testing data for final results.

A) Data Cleaning

Firstly, all necessary packages for the data cleaning and machine learning methods were added and loaded within Julia: DataFrames, CSV, Pkg, Random, StatsBase, LinearAlgebra, Plots, DecisionTree, MLDataUtils, and ScikitLearn. After both the training and testing sets which were provided were read into Julia, they were merged into dataframe "all_data" and features which contained a space in their naming convention were re-named for convenience. To prepare the variables for our models, we one-hot encoded all categorical variables in our data which included gender, customer type, class, and type of travel. The dependent variable, satisfaction, was level-encoded with 1 representing "satisfied" and 0 representing "neutral or dissatisfied." Following the dummy variable encoding, the original columns were dropped from the "all_data" data frame so that only the encoded columns remained for those variables. For the categorical variables that had two or more levels, we dropped one to avoid multicollinearity. There were a total of 393 rows that had a missing value, all from the arrival delay in minutes variable. We removed these 393 rows since this only made up less than 0.5% of the total data, so there is very little chance this would affect our findings. Finally, potential class imbalance was checked for, but considering that the data was so large and that the imbalance was minimal, we were advised at office hours to continue forth with splitting.

B) Training/Validation/Testing Split

We split our data 70/15/15 into training, validation, and test sets, respectively. We used k-fold cross validation and stratified sampling where k = 5 folds to split our data. This was to help protect against our results being subject to a specific randomized sample or split. For each model, after calculating the AUC and accuracy on the training and validation sets for each of the 5 folds, we averaged them to report our out-of-sample performance. We compared the out-of-sample performance on the validation sets to choose the best model from the list below, then predicted performance on the test set with this chosen model as our "best" model.

C) CART

To train the CART models we used grid search to search over max depth = 1:5, min bucket = 10, 20, and criterion = gini, entropy, misclassification for each of the 5 sets of training and validation data. We also decided to set localsearch=false, and set a seed for reproducibility. With the best model (combination of hyperparameters selected) chosen by the classifier, we used the IAI.fit_cv function to get the tree for each fold, where the validation criterion was AUC. The IAI.fit_cv function was fitted onto both the training and validation data partitions. One of these trees, Figure A, can be found in the appendix. The most important features from this model are "Online Boarding", followed by "Inflight Wifi Service" and "Type of Travel - Personal". These are also the same variables used in the first few splits of the tree. Table B in the appendix contains the full feature importance list. The IAI.score() function was employed to compute the AUC and accuracy for both the training and validation sets, for each of the 5 folds.

D) Logistic Regression

To train the Logistic Regression models we used the glm function in R for each of the 5 sets of training and validation data. There were only two variables that were highly correlated with each other (above 0.8) which were arrival delay in minutes and departure delay in minutes. We removed the departure delay in minutes variable to address multicollinearity. Similar to the rest of our models, we calculated the AUC and accuracy for both in-sample and out-of-sample (training and validation sets) for each of the 5 folds and took the average for later comparison. These are shown in the results section below.

E) Random Forest

To train the Random Forest models we used the same procedure from the CART model with the addition of searching over the number of trees = 20, 25, 50, 100, and 150. With the best model chosen by the classifier, we used the IAI.fit() function to get the output for each fold, where the validation criterion was AUC. The most important features from this model are "Online Boarding", "Inflight Wifi Service", and "Class - Business". Table C in the appendix contains the full feature importance list.

F) XGBoost

To train the XGBoost models we used grid search to search over max depth = 1:5, min bucket = 10, 20, and num_estimators 20, 25, 50, 100, and 150 for each of the 5 sets of training and validation data. With the best model chosen by the classifier, we used the IAI.fit function to get the output for each fold, where the validation criterion was AUC. The most important features from this model are "Online Boarding", "Inflight Wifi Service", and "Type Travel - Personal". Table D in the appendix contains the full feature importance list.

G) OCT

To train the OCT models we used the same grid search procedure from the CART model. Once again, all three criterion were utilized (gini, entropy, misclassification). With the best model chosen by the classifier, we used the IAI.fit function to get the tree for each fold, where the validation criterion was AUC. One of these trees, Figure B, can be found in the appendix. The most important features from this model are "Inflight Wifi Service", "Travel Type - Personal", and "Online Boarding. The first split in the tree was on "Class - Business", followed by "Ease of Online Booking" and "Inflight Wifi Service". We found that it was interesting to have different top importance variables than features chosen for the first splits of the tree, although this is a common phenomenon. Table E in the appendix contains the full feature importance list.

H) Rationale on Choosing the "Best" Model

The table containing the performance of each model is shown below in the results section. XGBoost performed the best on all criteria, so we choose this as our best model. OCT came in second, followed by Random Forest, CART, then Logistic Regression. While XGBoost had the highest accuracy and AUC values, its results are not as interpretable as with tree-based OCT and CART. We then implemented the interpretable clustering method we learned in class to get more intuition on the groups of passengers and what was important to them in their satisfaction ratings.

IV. Results of Binary Classification Models

	Train AUC	Val AUC	Train Accuracy	Val Accuracy
CART	0.968	0.968	0.905	0.905
Logistic Regression	0.926	0.929	0.874	0.878
Random Forest	0.973	0.974	0.915	0.916
XGBoost	**0.997**	**0.997**	**0.970**	**0.971**
OCT	0.976	0.977	0.926	0.927

*XGBoost performed the best, so we ran this model on the testing data with test AUC = 0.995 and test accuracy = 0.963.

V. Interpretable Clustering

For the entire dataset, we ran K-means clustering and labeled each point with the cluster it belongs to. Using a scree plot to choose the optimal number of clusters, in addition to running the model for different values of k, we settled on k = 10 as this yielded the most interpretable clusters. The scree plot can be found as Figure C the appendix. Most specifically we chose the value of k which minimized the intra-cluster distance, meaning that the clusters were "tightly" formed and thus interpretable. Then, we fed the results of clustering through an OCT with 10 classes to get interpretation on each cluster. The OCT from this clustering is Figure D in the appendix. The findings from this analysis are similar to what we could get from focus groups, which are costly in time and money. Meanwhile, the OCT method harnesses the interpretability of tree-based algorithms to easily understand customer segments. The clusters we have defined are as follows:

Cluster 1 (14% of all passengers):

- Business class flyers who gave positive scores for inflight entertainment, inflight wifi service and gave overall positive satisfaction scores
- Passenger Archetype: Business traveler who is overall satisfied when they can do their work and have entertainment on their flight

Cluster 2 (7%):

- Passengers who gave positive scores for seat comfort and departure/arrival time convenient, but gave negative scores for inflight entertainment and baggage handling
- Passenger Archetype: “Basics of flight were good, but the add-ons could have been better”

Cluster 3 (10%):

- Passengers who gave positive scores for ease of online booking, but negative scores for baggage handling, seat comfort, and inflight entertainment
- Passenger Archetype: “Booking my flight was good, but my experience on the flight was negative”. Perhaps these customers had cheap tickets with lower quality service.

Cluster 4 (11%):

- Passengers who gave negative scores for inflight entertainment, seat comfort, and ease of online booking
- Passenger Archetype: Dissatisfied customer

Cluster 5 (16%):

- Business class flyers who gave positive overall satisfaction scores despite giving negative scores for inflight wifi service. Gave positive scores for inflight entertainment
- Passenger Archetype: Business traveler who was dissatisfied with the wifi, but the inflight entertainment made up for it; also it is possible that these travelers gave overall positive scores since they are “busy” business flyers and are more prone to quickly filling out the survey just to get it done

Cluster 6 (8%):

- Passengers who gave positive scores for seat comfort, but negative scores for inflight entertainment and departure/arrival time convenience
- Passenger Archetype: “Departure/arrival time was inconvenient and the inflight entertainment was subpar, but at least the seats were comfortable”

Cluster 7 (7%):

- Passengers who gave negative scores for inflight entertainment, seat comfort, but positive scores for ease of online booking and baggage handling
- Passenger Archetype: “Good ease of online booking and baggage handling, but flight itself was uncomfortable and unenjoyable”

Cluster 8 (10%):

- Passengers who gave overall negative score, but positive scores for inflight entertainment and departure/arrival time convenience
- Passenger Archetype: Despite good inflight entertainment and departure/arrival time, gave an overall negative score for some other reason. Airlines should dig into this cluster. Perhaps one source of further investigation would be looking more closely at what time of year these flights occurred. Influence from the media, politics, etc. may be impacting overall customer sentiment during that period.

Cluster 9 (7%):

- Passengers who gave negative overall score, negative score for departure/arrival time convenience and positive score for inflight entertainment
- Passenger Archetype: "Despite the good inflight entertainment, the departure/arrival time being inconvenient was a dealbreaker and led to my poor ratings"

Cluster 10 (10%):

- Passengers not in business class who gave overall positive satisfaction score and positive score for inflight entertainment
- Passenger Archetype: Non-business class traveler who is overall satisfied when they have inflight entertainment. It would be interesting to further investigate what age range these passengers represent; for instance younger travelers prioritize inflight entertainment while older travelers may prioritize comfort.

The interpretation from these clusters is valuable for airlines to make decisions. Firstly, the OCT only made splits on 8 out of the 23 variables. This is much easier than trying to make sense of the results from K-means alone, which doesn't identify and limit to the most important variables. Thus, the airline can know to focus on these 8 variables when distinguishing between customer archetypes. While these 10 clusters are interpretable, they don't give insight into what the overall satisfaction score is for some clusters. Some clusters were very clear, such as cluster 1. Cluster 7, however, has mixed reviews and we don't know what were or weren't deal breakers in determining the overall score. Additionally, the 10 clusters may be formed largely upon whether or not the reviews were positive or negative, and thus filtering first would provide more granularity into what other features are defining each cluster besides the rating. To address this, we ran additional interpretable clustering models when filtering on passengers who gave positive versus negative overall satisfaction scores. These are Figures E and F in the appendix. The clusters we have defined are below:

Passengers who gave **negative** overall satisfaction (57% of all passengers):

- Cluster 1: Negative score was for departure/arrival time convenience despite having positive inflight entertainment (20% of all passengers who gave negative overall satisfaction)
- Cluster 2: Negative score was for some other feature for personal travelers despite having positive inflight entertainment and departure/arrival time convenient (24%)
- Cluster 3: Negative score was for some other feature for business travelers despite having positive inflight entertainment and departure/arrival time convenient (14%)
- Cluster 4: Negative scores were for inflight entertainment and inflight service (16%)

- Cluster 5: Negative score was for inflight entertainment despite having positive inflight service (26%)

Passengers who gave **positive** overall satisfaction (43% of all passengers):
- Cluster 6: Positive scores were for inflight wifi service, seat comfort, and baggage handling (28% of all passengers who gave positive overall satisfaction)
- Cluster 7: Positive scores were for for some other feature despite having negative inflight wifi service (7%)
- Cluster 8: Positive scores were from baggage handling despite having negative inflight wifi service (34%)
- Cluster 9: Positive scores were from inflight wifi service and seat comfort despite have negative baggage handling score, OR positive scores from some other feature despite having negative inflight wifi service and baggage handling (17%)
- Cluster 10: Positive score from inflight wifi service despite having negative seat comfort (13%)

Of course, these interpretable clusters don't provide causality for what was the determining factor in giving a positive or negative overall score for each case, but they give us some information on what is not a deal breaker. For example, passengers in clusters 7 and 8 (41% of all positive scores, 18% of all passengers) rated inflight wifi service negatively but still gave an overall positive satisfaction score. This suggests that passengers don't always need wifi service to be fully satisfied. Alternatively, passengers in clusters 2 and 3 (38% of all negative scores, 21% of all passengers) rated departure/arrival time convenience positively but still gave an overall negative satisfaction score. This suggests that a convenient flight time is not always enough to make a passenger's experience satisfactory. These two clusters also represent two segments of travelers (business versus leisure) thus it illustrates that convenient flight time is not always a priority regardless of reason for travel.

VI. Conclusions and Recommendations

In conclusion, this project illustrates the power of utilizing machine learning, especially under an optimization lens through the use of Optimal Classification Trees, to provide insights in the airline industry. The first phase of the project was focused on data cleaning and modeling methodology. As communicated in Section IV above, the "best" model which was chosen from the validation step was XGBoost as it obtained the highest training AUC, validation AUC, train accuracy, and validation accuracy (all greater than or equal to 0.97 with AUCs of ~0.99 each). As this high performance is typically expected from boosted algorithms when compared to simpler regression or tree-based methods, they lack interpretability. For instance, simply reading off an AUC curve and understanding numerical metrics is telling of predictive power, but it is not intuitive. Meanwhile, the visual component of trees enables even non-technical audiences to understand what points have been classified with what parameters. It is simple to follow points from the highest node of a tree through each variable split, until the leaf nodes are fully formed. Thus, even though XGBoost was labeled our "best" model due to its high predictiveness, it may not generally be the model of choice for a scenario which prioritizes transparent algorithms (even with a slight dip in performance). Such a circumstance may deem the OCT or Random Forest model the best, because they are more communicative to general audiences than XGBoost yet achieve AUC values and accuracies above 0.90.

Secondly, our use of Optimal Classification Trees in order to more accurately understand the results of interpretable clustering was our second attempt at showing how interpretability matters over performance in some cases. While the K-Means algorithm produced 10 relatively well-defined clusters, all variables were used and the connection between each cluster and overall rating became murky. It was unclear as to which feature was the primary motivating factor for each cluster's positive or negative rating. Filtering the clustering through the optimal classification tree produced a more subsetted list of 8 variables the model deemed as important in making splits, as opposed to the previous 23 in K-means. Additionally, filtering on positive and negative reviews further informed us as to which factors (ease of online booking, wifi service, arrival or departure time, etc.) were more influential in predicting customer satisfaction.

Thus, as throughout the class we have been encouraged to try new cutting-edge methods which have proven results over traditional methods, we utilized both CART and Random Forest tree-based methods as well as OCT. The use of CART and Random Forest showed that despite having slightly lower performance metrics, interpretability provides value over more complex methods such as XGBoost. Secondly, our use of OCT as a way to understand the results of traditional K-Means clustering led to a more specific analysis of customer segments.

There are several future directions of study we would recommend to airlines following this study, as outlined below:

1) **Understanding the Timing of Surveys**:

As mentioned previously, certain clusters give overall positive or negative reviews despite opposite ratings for several flight-related factors. What else could be influencing these ratings? Airlines should take a deep dive into the surrounding political, socioeconomic, and global scene at the time during which these surveys were sent. Perhaps it is even the time of year, tourist versus holiday season, or time of day during which the surveys were completed which are accountable for these small differences.

2) **A NLP Approach** :

If a column had been added to this dataset with real comments from flyers, it would have been interesting to use a "bag of words approach" or similar NLP packages within Julia and Python to understand what terms and n-grams are most associated with positive versus negative ratings for different clusters of customers. This could be a future focus for airlines.

3) **Pre versus Post Covid Differentiation**:

Is there a difference in what factors customers value most before versus after the pandemic? As the data used in this survey is pre-Covid data, perhaps more digitized features of flying are now more popular than they were in the time the survey was conducted. As an example, passengers may now value the flexibility of online booking and check-in features more than they did before.

VII. Team Member Contributions

Overall, we feel we both contributed equally and worked together very well on this project. We worked both together and at times individually, but had back and forth collaboration where we made sure to communicate our work so the other could understand and edit as needed. To outline each of our contributions:

1. Project Proposal (researching potential projects, gathering data, writing proposal)
    a. Brittany and Estella
2. Classification Methods
    a. Data Cleaning: Brittany and Estella
    b. Running Models: We took turns running models
3. Interpretable Clustering
    a. Estella
4. Interpreting Results
    a. We both contributed to writing up our intuitions on our findings
5. Report
    a. Problem Statement: Brittany and Estella
    b. Data, Methodology, Data Cleaning: Brittany
    c. Conclusions and Recommendations: Brittany
6. Presentation
    a. Brittany and Estella
7. Office Hours Attendance
    a. We both attended all three TA's office hours for project help :)

VIII. References

*https://www.ibisworld.com/#:~:text=The%20market%20size%2C%20measured%20by,to%20increase%2057.6%25%20in%202022.

*kaggle dataset:
https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?select=train.csv

IX. Appendix

Table A: List of Variables in Data

Variable	**Description**	**Type**
Gender	Gender of passenger	Categorical (Male, Female)
Customer Type	The loyalty of passenger	Categorical (Loyal customer, Disloyal customer)
Age	Age of passenger	Numeric
Type of Travel	Purpose of flight of passenger	Categorical (Personal travel, Business travel)

Class	Travel class of passenger	Categorical (Business, Eco, Eco plus)
Flight Distance	Flight distance of journey	Numeric
Inflight Wifi Service	Satisfaction level of the inflight wifi service	Numeric (0-5)
Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient	Numeric (0-5)
Ease of Online booking	Satisfaction level of online booking	Numeric (0-5)
Gate location	Satisfaction level of Gate location	Numeric (0-5)
Food and drink	Satisfaction level of Food and drink	Numeric (0-5)
Online boarding	Satisfaction level of online boarding	Numeric (0-5)
Seat comfort	Satisfaction level of Seat comfort	Numeric (0-5)
Inflight entertainment	Satisfaction level of inflight entertainment	Numeric (0-5)
On-board service	Satisfaction level of On-board service	Numeric (0-5)
Leg room service	Satisfaction level of Leg room service	Numeric (0-5)
Baggage handling	Satisfaction level of baggage handling	Numeric (0-5)
Check-in service	Satisfaction level of Check-in service	Numeric (0-5)
Inflight service	Satisfaction level of inflight service	Numeric (0-5)
Cleanliness	Satisfaction level of Cleanliness	Numeric (0-5)
Departure Delay in Minutes:	Minutes delayed when departure	Numeric (0-5)
Arrival Delay in Minutes	Minutes delayed when Arrival	Numeric (0-5)

Satisfaction	Airline satisfaction level	Categorical (Satisfaction, neutral or dissatisfaction)

Table B: CART Variable Importance

Variable	**Importance**
Online_boarding	0.482727
Inflight_wifi_service	0.223625
typetravel_personal	0.178261
Inflight_entertainment	0.060761
Checkin_service	0.0197665
class_business	0.0150307
customertype_loyal	0.0129642
Gate_location	0.00421701
Arrival_delay_min	0.00172048
Ease_of_online_booking	0.000765304
Food_and_drink	0.000111631
Leg_room_service	5.0403e-5

Figure A: CART Tree

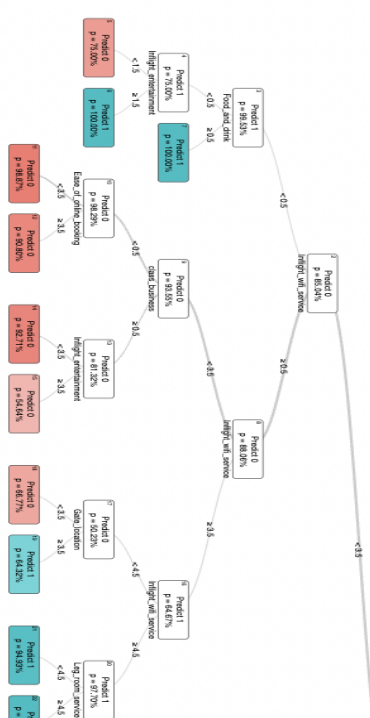

Table C: Random Forest Variable Importance

Variable	**Importance**
Online_boarding	0.257269
Inflight_wifi_service	0.14928
class_business	0.115534
typetravel_personal	0.0977713
Inflight_entertainment	0.0736051
class_eco	0.0636958
Leg_room_service	0.0380548
Ease_of_online_booking	0.0363912

Seat_comfort	0.0311836
customertype_loyal	0.0232694
Onboard_service	0.0229542
Flight_Distance	0.0184831
Cleanliness	0.0165245
Inflight_service	0.0153041
Baggage_handling	0.0109719
Checkin_service	0.009634
Age	0.00733594
Food_and_drink	0.00439113
Departure_Arrival_time_convenient	0.00341095
Arrival_delay_min	0.00151462
Gate_location	0.00145648
gender_female	0.00132647
Departure_delay_min	0.000638725

Table D: XGBoost Variable Importance

Variable	**Importance**
Online_boarding	0.35169
Inflight_wifi_service	0.215996
typetravel_personal	0.139441
customertype_loyal	0.043114
class_business	0.0412133
Inflight_entertainment	0.0290764
Checkin_service	0.0276262

Inflight_service	0.0264801
Seat_comfort	0.0196986
Baggage_handling	0.0185239
Onboard_service	0.0139471
Gate_location	0.0132233
Leg_room_service	0.0131854
Age	0.0127157
Cleanliness	0.0120723
Flight_Distance	0.00630237
Arrival_delay_min	0.00556138
Departure_Arrival_time_convenient	0.00372298
Ease_of_online_booking	0.00289734
Departure_delay_min	0.00176928
Food_and_drink	0.00100102
class_eco	0.000590426
gender_female	0.000152016

Figure B: OCT Tree

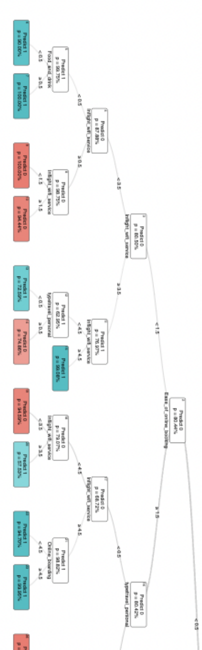

Table E: OCT Variable Importance

Variable	Importance
Inflight_wifi_service	0.337777
typetravel_personal	0.170421
Online_boarding	0.12248
class_business	0.112664
Inflight_entertainment	0.110659

customertype_loyal	0.0970871
Leg_room_service	0.00994696
Gate_location	0.00987267
Inflight_service	0.00956211
Cleanliness	0.00479431
Checkin_service	0.00455681
Age	0.00317777
Seat_comfort	0.0028937
Arrival_delay_min	0.00265054
Ease_of_online_booking	0.00102157
Food_and_drink	0.000434826

Figure C: K-means Scree Plot - No Filters

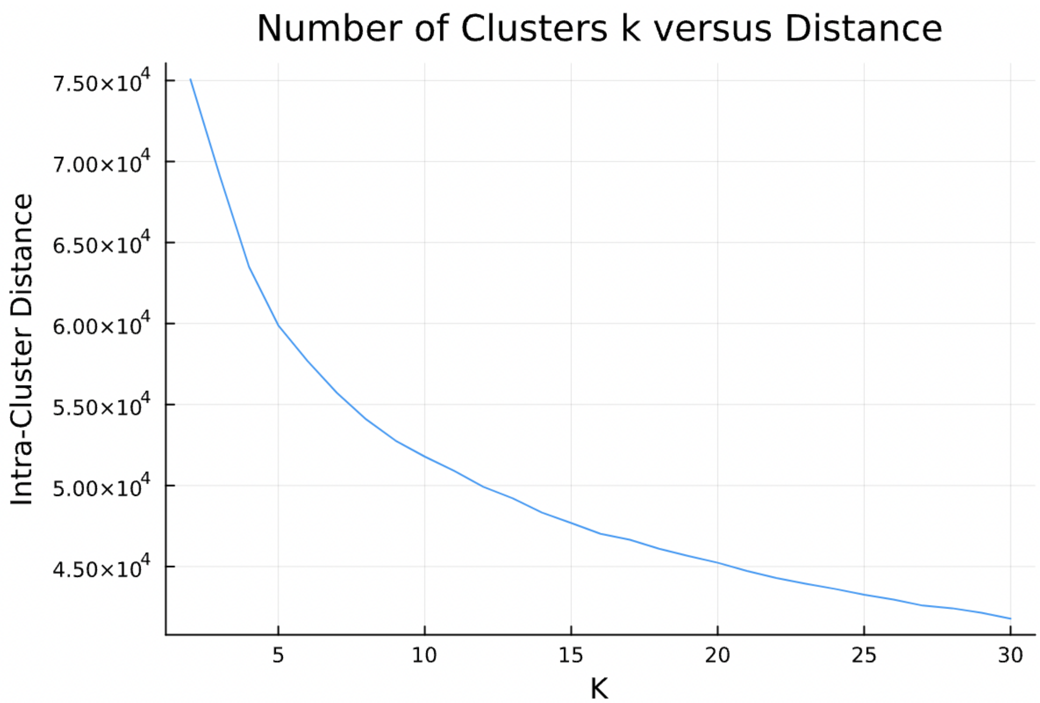

Figure D: Interpretable Clustering - No Filters

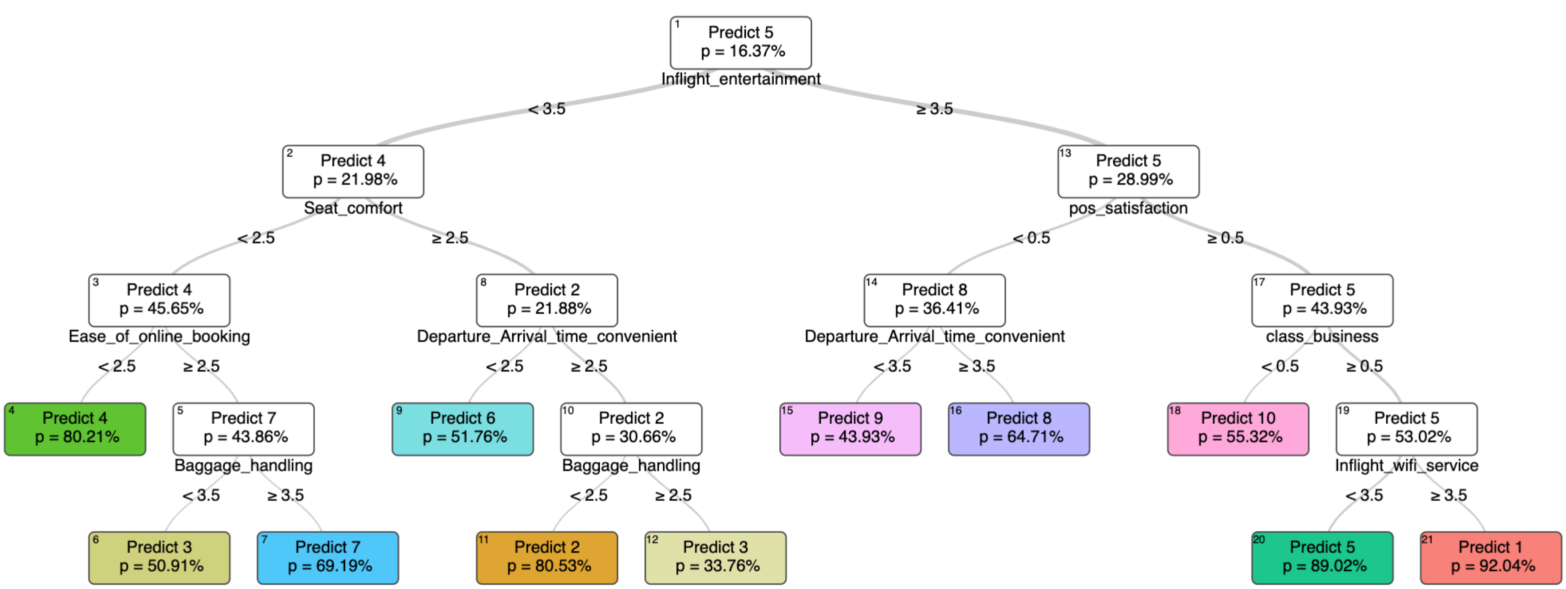

Figure E: Interpretable Clusters: Only Negative Overall Scores

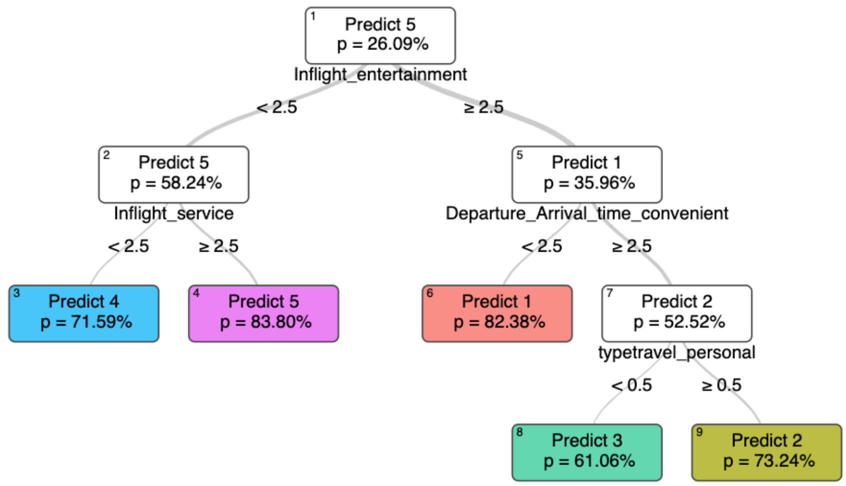

Figure F: Interpretable Clusters: Only Positive Overall Scores