Table of Contents

Introduction

In the world of competitiveness, airline companies around the globe are emphasizing on offering unique experience and comfort to ensure customer satisfaction. In the midst of aviation sector growth, there is no alternative for customer satisfaction. So, an Aviation Operator has to provide keen focus on customer behavior and innovate unique values through service that can be key differentiators in the crowded skies.

Business Problem and Objective

To better understand customer relationships with satisfaction and ensure holistic consistency, Aviation Operator is looking forward to offering distinctive offerings throughout the value chain. They acknowledge that any unsatisfied customer will not only switch their preference because of recent experience but they also can influence potential customers through word of mouth, reviews and social media. On the other hand, identifying and appreciating satisfied customers can go a long mile when it comes to brand loyalty and implementing positive influence in the market.

The main objective here is to improve passenger satisfaction for Aviation Operators. Aviation Operator can pave their way forward through 3 possible objectives:

- **Precision:** If the aim is to avoid false positives that means steering clear from predicting incorrectly passengers are satisfied when they are not then precision is the ideal metric.

- **Recall:** If the focus is to avoid false negatives, that means Aviation Operators want to identify all satisfied customers and target them for loyalty programs instead of misclassifying dissatisfied customers as a satisfied one.

- **F1 Score:** Instead of Precision or Recall, this could be ideal if the purpose is to explore the harmony of precision and recall and balance two.

In this analysis, Aviation Operators should be looking towards **precision** because if recall is considered for evaluation then there is a chance of classifying dissatisfied customers as satisfied customers. However, the risk of losing satisfied customers for customer referral programs is the trade-off here. This would incur high cost when it comes to profitability and long-term growth of Aviation Operators.

Methodology

To achieve the objective, we will incorporate the data-driven approach through Machine Learning for predicting Customer Satisfaction for Aviation Operators.

Dataset: Dataset consists of 103,905 samples and 24 variables reflecting diversified attributes related to passenger's journey and experience.

Feature: The dataset made of 24 attributes both categorical and numerical data such as Gender, Age, Type of Travel, Arrival delay etc.

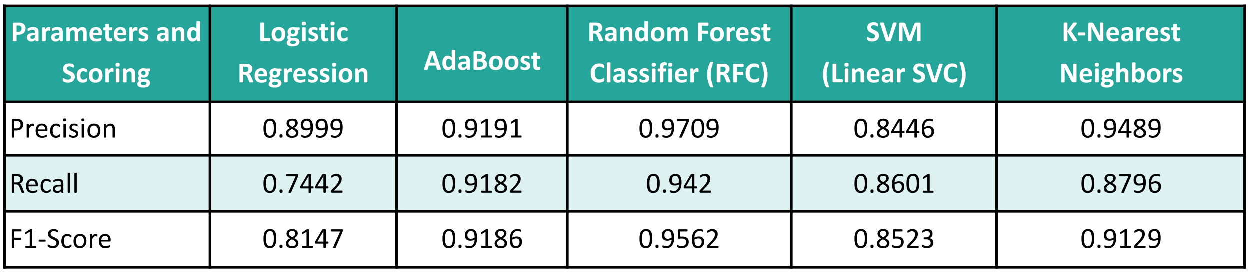
Target Variable: 'Satisfaction' is the primary variable of the dataset and this is categorized as 'satisfaction', or 'neutral or dissatisfied'.

Methodology: Several machine learning algorithms to build predictive models:

- **Logistic Regression:** Since target variable 'Satisfaction' is categorical (satisfied or not satisfied), logistic regression is an ideal framework to implement because it is suitable for binary classification problems to detect the customer 'Satisfied' or 'neutral or dissatisfied'.
- **Random Forest:** Random forest is capable of handling the mixture of categorical and numerical features with its ensemble approach. This dataset requires higher predictive power for complex relationships across flight delay time, cleanliness, etc.
- **AdaBoost:** This algorithm is used to adjust the weights of the classifiers and instances for considering complex and non-linear relationships between features and the target variables. Despite being a weak learner, Adaboost can precisely predict based on specific important features.
- **Support Vector Machine:** Linear SVC is chosen for its effectiveness in binary classification, high-dimensional feature handling across customer experience, and model interpretability, which align well with the needs of the satisfaction dataset.
- **K-Nearest Neighbors:** This algorithm used to make predictions based on similar historical examples to identify customer satisfaction patterns locally and capture actionable insights which are not assumptions.

Predictive Model Performance

Machine Learning Algorithms implemented to find the best model for predicting Satisfaction.



Parameters and Scoring	Logistic Regression	AdaBoost	Random Forest Classifier (RFC)	SVM (Linear SVC)	K-Nearest Neighbors
Precision	0.8999	0.9191	0.9709	0.8446	0.9489
Recall	0.7442	0.9182	0.942	0.8601	0.8796
F1-Score	0.8147	0.9186	0.9562	0.8523	0.9129

Figure: Aviation Operator Satisfaction Dataset Model Performance

Random Forest

Working Algorithm: Random Forest Classifier bulbs a forest of decision trees, each trained on random sunsets of the data while using bootstrapping, and considers a random subset of features splitting at each node which ensures diversity among trees and minimizes overfitting. When prediction happens, each of the trees in the forest votes for a specific class and the class with majority votes becomes the model's prediction for classification tasks.

The Best Model RFC & Caution: Mentioned performance metrics, RFC holistically outperformed others in predicting customer satisfaction since it performed well in Precision, Recall and F-1 Score which reflects high reliability and a balanced performance. However, RFC can be computationally intensive due to a larger number of trees, so RFC needs to be tested in real-world settings to ensure its performance metrics before full deployment with an alignment with the business goals in securing maximum accuracy in positive prediction.

RFC Hyperparameter Tuning: RFC classification__n_estimator is 800 which represents the total number of trees in the forest. Each tree is going to be trained on the bootstrap sample of data and produce individual predictions. Random forest will aggregate all predictions and deliver results. For this dataset, 800 is the optimal one after doing the analysis back and forth. However, increasing the number of trees can make the model robust but also increase the computational cost and provide marginal performance gain.

KNN

Working Algorithm: Initially, it keeps the training dataset and compares the distance between the training dataset and its examples. Afterwards, it identifies the nearest neighbors to the training dataset. The main idea is that similar data points have similar labels. Moreover, This algorithm reflects as a nonparametric method which means it uses its proximity to make predictions through statistical methods based on ranks, signs etc.

New Algorithm Advantage & Disadvantage: KNN can be a handful for generating insights in the local environment so it could help capturing clusters of similar customers for targeted marketing. Though KNN performed well, it did not outperform other models. However, KNN’s computation time could become vulnerable if the Aviation Operator is growing and predicting from large datasets.

KNN Hyperparameter Tuning: From a range of 3, 5, 7 closest items, when it comes to predict customer satisfaction, 5 nearest neighbors is considered to be an ideal value for the model according to GridSearch. That means if it requires to classify whether a customer is satisfied or not then for a particular passenger the 5 nearest neighbors of the dataset like age, ticket price, flight duration etc. then it will determine how many of them are satisfied and unsatisfied and based on the majority new data point will be labeled as ‘Satisfied’ or ‘Dissatisfied’.

Logistic Regression

Working Algorithm: The main objective is to predict a binary outcome based on analyzing one or more independent variables and find the relationships between two data factors. If any customer is dissatisfied in the current dataset, logistic regression will detect the pattern to predict based on relevant variables which influence the final decision of the previous customer.

Affordable Logistic Regression: Logistic regression could be the relevant parametric model for this scenario due to interpretability. Though it did not perform well in recall so if the business objective is to capture accurate positive rate then it could be an ideal option due to high precision.

LR Hyperparameter Tuning: alpha: 0.001, eta0: 0.001, l1_ratio: 0.5, max_iter: 100

Alpha: To avoid overfitting 0.001is the optimal value by weakening the regularization.
Eta0: Learning rate is smaller which means the model learns slower for better convergence.
L1_ratio: Equally benefited from sparsity inducing property (L1) and shrinking property (L2).
Max_iter: The number of passes over training data to control iteration for the algorithm unless it converges before that.

Adaboost

Working Algorithm: This algorithm classifies data by integrating weak learners into a strong learner. It weights instances in the training dataset based on previous classification accuracy. One of the objectives of this algorithm is to train models until smaller errors return by weighting approach through adding larger weights to incorrectly categorized points.

Consistency of Adaboost: Apart from all algorithms adaboost performance is consistent across all metrics in terms of Precision, Recall and F-1. This reflects a balanced prediction and indicates low trade-off between precision and recall. So, it actually defines that both 'Satisfied' and 'Neutral or dissatisfied' are handling well in this model.

Adaboost Hyperparameter Tuning: Classification__n_estimators is 70 which means 70 models are built sequentially in the ensemble to provide different sets of weights or important values to each data point which made errors from the earlier estimator.

SVM

Working Algorithm: To perform classification, Linear SVC increases the distance by attempting to find a hyperplane between classified examples in order to find a best fit hyperplane. It provides better performance with a large number of samples.

Moderate Performer Linear SVC: While it is not the top performer, it could be a potential model for the future, if Aviation Operators want to strike a balance between minimizing false positives through identifying true positives. Additionally, it is also suitable for large datasets and efficient storage capability since SVM only needs Support Vector instead of the entire dataset.

SVC Hyperparameter Tuning: Classification__C is 0.01 reflected as a cost parameter and emphasizes a larger margin but allows some misclassification to avoid overfitting due to the noise or overlapping classes.

All models hyper parameter tuning & performance scores can be founded in the appendix*

Cost-Benefit Analysis

Based on predictive model performance, Random Forest Classifier (RFC) will be implemented to predict customer satisfaction for Aviation Operators and also provide a comparison scenario without any classifier.

- Aviation operator has 800,000 customers
- Of these, 18% are dissatisfied, 144,000 customers expected to dissatisfied annually
- For each dissatisfied customer, the airline incurs a loss of €700
- To draw in a new customer in place of a dissatisfied one, it costs €450
- A special offer to win back dissatisfied customers costs €250

Without the Model: If the airline didn't use a model, they might present every dissatisfied customer with a special deal to win them back, or just absorb the loss from their dissatisfaction.

- **Cost from Dissatisfaction:**

  Projected number of dissatisfied customers = 144,000
  Loss from each dissatisfied customer = €700
  Total loss without model = 144,000 customers * €700 per customer = €100,800,000

- **Cost of Special Deals to Counter Dissatisfaction:** If they opted to give a deal to each customer,

  Total cost of deals = 800,000 customers * €250 per customer = €200,000,000

  However, Providing special offers to everyone isn't feasible because not every customer will come back, so Aviation Operators should focus on dissatisfaction costs.

With the Model: The positive label here is 'Dissatisfied Customer' instead of 'Satisfied Customer' since it is more essential when it comes to improving service quality and customer retention. That means,

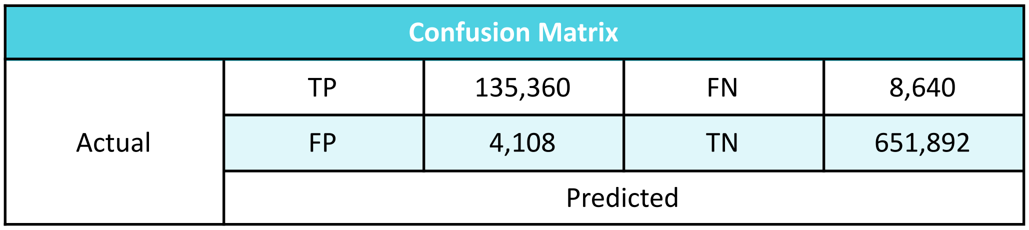
- **True Positive:** Identify a satisfied customer as a satisfied customer.
- **True Negative:** Identify a dissatisfied customer as a dissatisfied customer.
- **False Positive:** Identify a dissatisfied customer as a satisfied customer.
- **False Negative:** Identify a satisfied customer as a dissatisfied customer.

- **Calculating True Positives (TP) and False Negatives (FN):** With a recall of 94%, the number of actual dissatisfied customers correctly identified is,
  TP= Recall*Actual Number of Dissatisfied Customer
  TP = 0.94 * 144,000 = 135,360
  FN = Actual Number of Dissatisfied Customer - TP
  FN = 144,000 - 135,360 = 8,640

- **Finding False Positives (FP):** With precision at 97%, the equation becomes,
  Precision = TP/(TP+FP)
  0.97 = 135,360 / (135,360 + FP)
  FP = (135,360 / 0.97) - 135,360 = 4,108

- **True Negatives (TN):** From our total, TN + FP gives us the number of satisfied customers,
  TN+FP = Total Customers - Expected to Dissatisfied
  TN + 4,108 = 800,000 - 144,000
  TN = 656,000 - 4,108 = 651,892

  The model lets the airline target only those predicted to be dissatisfied with a special offer.



Confusion Matrix				
Actual	TP	135,360	FN	8,640
	FP	4,108	TN	651,892
	Predicted			

- **Cost of Special Deals (Based on Precision):**

  Predicted dissatisfied customers (TP + FP) = 135,360 (TP) + 4,108 (FP) = 139,468
  Cost of special deal per customer = €250
  Total special deal costs with model = 139,468 customers * €250 = €34,867,000

- **Loss from Dissatisfaction After Offers (Recall):** Some customers will remain dissatisfied since the model isn't perfect

  Number of overlooked dissatisfied customers (FN) = 8,640
  Loss from each = €700
  Total loss with model = 8,640 customers * €700/customer = €6,048,000

- **Revenue Loss due to False Positives:** Special deals are given to 4,108 satisfied customers,
  Cost of unneeded offers: 4,108 customers * €250/customer = €1,027,000

- **Cost to Bring in New Customers to Replace Dissatisfied Ones:** Companies will be able to retain some, but not all. We still need replacements for the rest,
  Number still dissatisfied (FN) = 8,640
  Cost for each new customer = €450
  Total replacement cost = 8,640 customers * €450/customer = €3,888,000

- **Total loss due to False Negative:**

  Total Loss with model + Total Replacement Cost = €6,048,000 + €3,888,000 = €9,936,000

- **Costs using the predictive model:**

  Total model cost = Special deal cost + Loss from dissatisfaction + Cost of unneeded deals + Replacement cost
  Total model cost = €34,867,000 + €6,048,000 + €1,027,000 + €3,888,000
  Total model cost = €45,830,000

- **Cost-Benefit Evaluation:**

  Without Model:
  Loss from dissatisfaction = €100,800,000

  With Model:
  Total with predictive model = €45,830,000

  Net Savings:
  Net savings = Cost without model - Cost with model
  Net savings = €100,800,000 - €45,830,000
  Net savings = €54,970,000

By employing the predictive model, the airline might save roughly €54,970,000 from dissatisfaction-related costs. This basic analysis centers on direct costs tied to customer dissatisfaction and retention tactics.

Conclusion

In the competitive landscape of the aviation industry, maintaining high levels of customer satisfaction is imperative for sustaining growth and profitability. Aviation Operator's initiative to leverage machine learning to predict and enhance customer satisfaction has proven to be a valuable strategy based on the analysis and the predictive model performance data provided.

The use of a Random Forest Classifier has shown to be the most promising approach among the tested algorithms, delivering a precision score of 0.9709, which indicates a high level of accuracy in predicting true positives. This precision is crucial for Aviation Operator's strategy, which prioritizes correctly identifying satisfied customers to foster brand loyalty and capitalize on positive word-of-mouth marketing, while minimizing the associated misclassification cost of dissatisfied customers as satisfied ones.

By incorporating this predictive model, Aviation Operator has significantly reduced its potential churn-related losses from $80 million to $12.037 million. This is not only a testament to the power of machine learning in operational settings but also highlights the importance of precision in customer satisfaction models for cost-effective decision-making.

In this context, the trade-off of potentially losing out on some satisfied customers for referral programs due to a lower recall score is deemed acceptable when compared to the substantial cost savings and the positive impact on the airline's bottom line. The predictive model not only helps in retaining customers but also provides strategic insights into the customer base, which can be leveraged for tailored marketing campaigns and improved service offerings.

By continuing to refine their predictive capabilities and align them with customer-centric strategies, Aviation Operators can ensure a consistent and satisfying experience for its passengers, which is paramount in today's hyper-competitive airline industry.

Reference

- Kaggle, 2020. Airline Passenger Satisfaction Dataset. [data] Available at: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction [Accessed 2 November 2023].

- Liaw, A. and Wiener, M., 2002. Classification and Regression by randomForest. R News, 2(3), pp.18-22.

- CAPA - Centre for Aviation, 2021. Airlines: Customer satisfaction during and post COVID-19. [online] Available at: https://centreforaviation.com/analysis/reports/airlines-customer-satisfaction-during-and-post-covid-19-558795 [Accessed 2 November 2023].

- Marr, B., 2019. How Is AI Used In Healthcare - 5 Powerful Real-World Examples That Show The Latest Advances. [online] Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2019/01/22/how-is-ai-used-in-healthcare-5-powerful-real-world-examples-that-show-the-latest-advances/ [Accessed 2 November 2023].

- Saito, T. and Rehmsmeier, M., 2015, November. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. In PloS one, 10(3), p.e0118432.

- Brownlee, J., 2016. How to Implement k-Nearest Neighbors in Python. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/ [Accessed 3 November 2023].