DNSC 6315, Machine Learning 2 - Group 14
Meghana Nekkanti, Karim Alami, Henry Wan, Akshay Thirumalreddy

Airline Data

Background

Our project focuses on predicting future customer satisfaction based on the given information about their flight. Our information is based on customer feedback from an unnamed airline and given the name Invistco Airlines for the project.

The project aims to identify what factors result in satisfied customers and to use this information to convert neutral and dissatisfied customers into satisfied ones.

Exploratory Data Analysis

There are approximately 130K rows of data on the columns listed above. Of the 130K rows of data, there are 393 missing values in the "Arrival Delay in Minutes' column. These are for when there is no delay in the arrival time. We have filled them with 0. Aside from those, the data is complete.

The passenger demographics indicate a majority in the 20 to 60 age bracket, suggesting targeted services and marketing strategies for this group. Female passengers slightly outnumber males. Customer satisfaction levels reveal 82% loyalty, emphasizing the need for retention strategies. Business class passengers show the highest satisfaction (48%), followed by economy class (45%) and economy plus class (7%). Satisfaction levels are similar between genders, implying no significant differences. Business travelers (69%) are more satisfied than personal travelers (31%). The age group 35-44 has the highest satisfaction (23.57%), followed by 45-54 (20.20%), highlighting areas for improvement in the youngest and oldest age groups. Online boarding and inflight entertainment receive favorable ratings, indicating overall satisfaction. Departure/arrival time convenience is rated well, with most customers finding the timing acceptable.

Modeling

In our predictive modeling analysis, we determined that all the variables present in the data effectively predict customer satisfaction.

We split the data into 60-40 splits of training and testing to ensure optimal prediction. 'Satisfaction was chosen as the dependent variable, and the rest were used as independent variables for prediction.

Our rigorous evaluation process involved testing various models - Logistic Classifier, Decision Tree Classifier, Bagging Classifier, Random Forest Classifier, Gradient Classifier and K Nearest Neighbour Classifier. We assessed their accuracy and area under the curve (AUC) to identify the model that consistently delivered the highest accuracy across the data. This model was then selected for further evaluation based on subsets of our data, including 'Type of Class', 'Age Group', 'Gender', 'Type of Travel', and 'Customer Type'.

For the business decision, we chose the model for each group of subsets based on the highest accuracy. Then, we performed recursive feature elimination to remove the weakest features. Then, we trained the model again—the model we chose for the particular group based on accuracy and AUC mean—using the selected features.

We then choose the top 3 most important features (independent features that have more effect on dependent variables) and increase their satisfaction by 1 level; we then predict using the changed data. To compare these predictions with the initial test data, we count the number of initially dissatisfied passengers who are now predicted as satisfied.

Performance Evaluation

We evaluated the performance of the models based on accuracy and AUC mean. Accuracy measures the proportion of correctly classified instances out of the total number of instances. AUC measures the performance of a classification model across all possible classification thresholds. Both of them range between 0-100%.

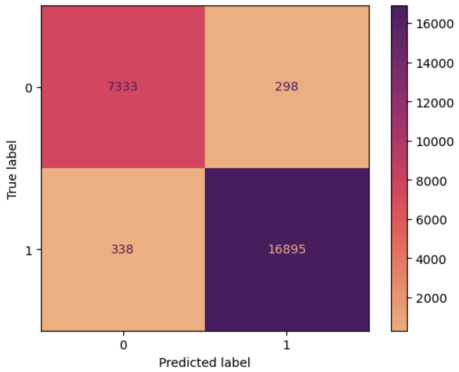
Entire Dataset

The Random Forest Classifier works the best for the entire dataset. It has an accuracy of 95.96% and an AUC mean of 99.31%.

Business Class

The Random Forest Classifier works the best for this subset. It has an accuracy of 97.19% and an AUC mean of 99.55%. It correctly classifies 24,228 observations out of 24,864 observations.

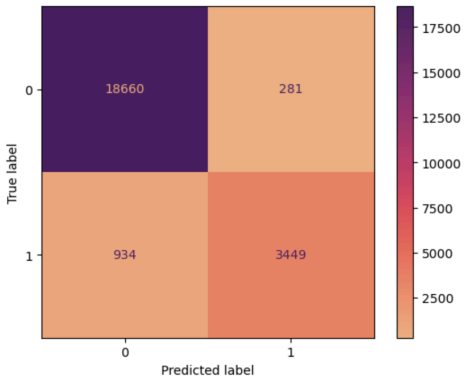
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	99.55	0.02	97.19	0.09
2	Bagging Classifier	99.25	0.07	96.79	0.15
4	Gradient Boosting Classifier	99.09	0.05	95.53	0.24
1	Decision Tree Classifier	96.46	0.63	95.77	0.18
0	Logistic Regression	88.19	0.89	81.67	0.62
5	KNN	73.12	0.20	72.60	0.29



Economy Class

The Random Forest Classifier works the best for this subset. It has an accuracy of 94.51% and an AUC mean of 98.12%. It correctly classifies 22,106 observations out of 23,321 observations.

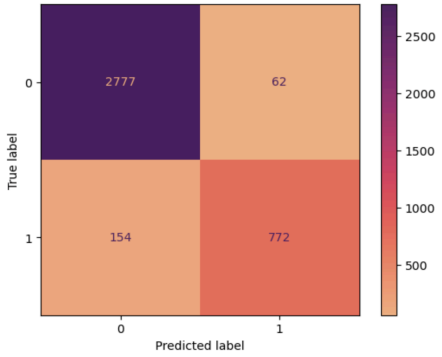
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	98.12	0.11	94.51	0.17
4	Gradient Boosting Classifier	98.03	0.05	94.03	0.27
2	Bagging Classifier	97.35	0.08	94.56	0.14
1	Decision Tree Classifier	95.85	0.28	94.13	0.14
0	Logistic Regression	81.10	0.44	86.33	0.58
5	KNN	66.12	0.30	81.80	0.21



Economy Plus Class

The Gradient Boosting Classifier works the best for this subset. It has an accuracy of 94.01% and an AUC mean of 98.22%. It correctly classifies 3,549 observations out of 3,765 observations.

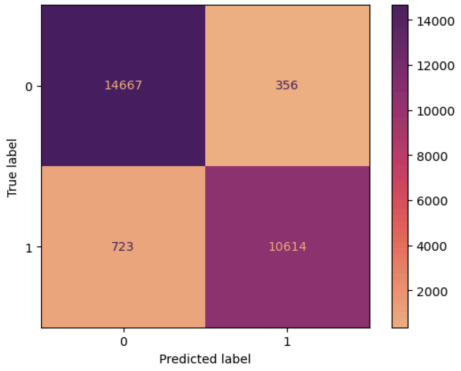
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
4	Gradient Boosting Classifier	98.22	0.14	94.01	0.11
3	Random Forest Classifier	97.97	0.03	93.66	0.35
2	Bagging Classifier	97.01	0.17	93.85	0.22
1	Decision Tree Classifier	95.53	0.41	92.99	0.41
0	Logistic Regression	84.19	0.89	84.77	0.73
5	KNN	60.12	0.65	73.34	0.30



Female Data

The Random Forest Classifier works the best for this subset. It has an accuracy of 95.85% and an AUC mean of 99.23%. It correctly classifies 25,281 observations out of 26,360 observations.

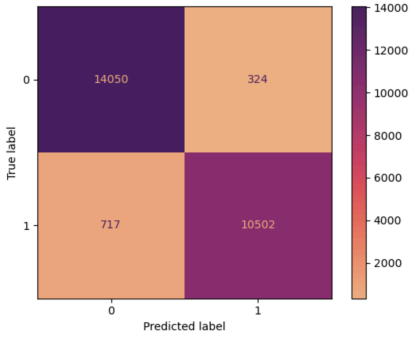
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	99.23	0.03	95.85	0.16
2	Bagging Classifier	98.85	0.09	95.58	0.20
4	Gradient Boosting Classifier	98.78	0.05	94.14	0.05
1	Decision Tree Classifier	96.65	0.11	94.68	0.08
0	Logistic Regression	86.07	0.23	80.57	0.34
5	KNN	76.36	0.54	71.54	0.39



Male Data

The Random Forest Classifier works the best for this subset. It has an accuracy of 95.69% and an AUC mean of 99.22%. It correctly classifies 24,552 observations out of 25,593 observations.

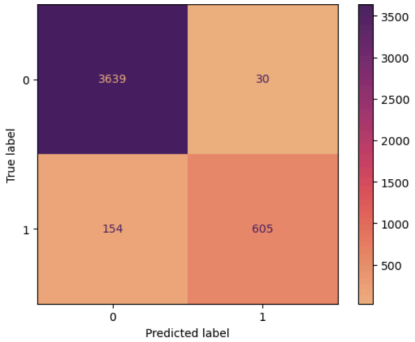
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	99.22	0.03	95.69	0.10
2	Bagging Classifier	98.83	0.06	95.48	0.12
4	Gradient Boosting Classifier	98.80	0.02	94.22	0.04
1	Decision Tree Classifier	96.32	0.20	94.36	0.17
0	Logistic Regression	87.99	0.51	81.46	0.85
5	KNN	76.63	0.11	71.47	0.10



Age group 1 (6-18)

The Gradient Boosting Classifier works the best for this subset. It has an accuracy of 95.23% and an AUC mean of 98.46%. It correctly classifies 4,244 observations out of 4,428 observations.

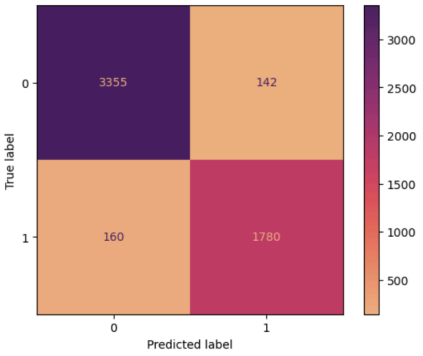
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
4	Gradient Boosting Classifier	98.46	0.21	95.23	0.29
3	Random Forest Classifier	98.22	0.19	95.08	0.32
1	Decision Tree Classifier	97.69	0.35	94.47	0.17
2	Bagging Classifier	97.42	0.33	94.60	0.37
0	Logistic Regression	81.14	0.33	86.28	1.02
5	KNN	65.50	1.14	81.78	0.85



Age group 2 (19-24)

The Random Forest Classifier works the best for this subset. It has an accuracy of 94.57% and an AUC mean of 98.81%. It correctly classifies 5,135 observations out of 5,437 observations.

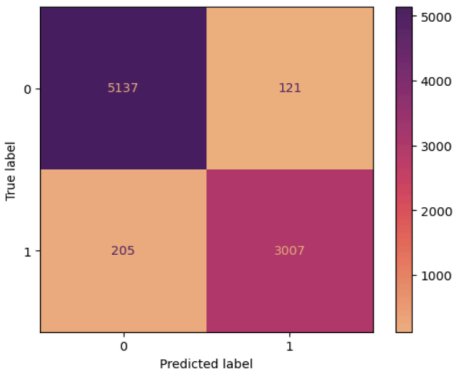
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	98.81	0.06	94.57	0.11
4	Gradient Boosting Classifier	98.80	0.08	94.15	0.17
2	Bagging Classifier	98.13	0.27	93.42	0.41
1	Decision Tree Classifier	96.46	0.48	93.29	0.53
0	Logistic Regression	84.45	0.62	81.30	0.57
5	KNN	68.74	0.74	68.08	0.55



Age group 3 (25-34)

The Random Forest Classifier works the best for this subset. It has an accuracy of 95.69% and an AUC mean of 99.17%. It correctly classifies 8,144 observations out of 8,470 observations.

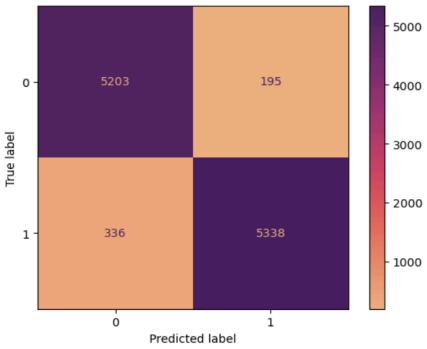
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	99.17	0.04	95.69	0.07
4	Gradient Boosting Classifier	99.00	0.08	94.41	0.22
2	Bagging Classifier	98.76	0.09	95.26	0.22
1	Decision Tree Classifier	96.40	0.27	94.32	0.25
0	Logistic Regression	88.52	0.85	83.88	0.76
5	KNN	73.09	0.30	71.11	0.33



Age group 4 (35-44)

The Random Forest Classifier works the best for this subset. It has an accuracy of 94.97% and an AUC mean of 98.85%. It correctly classifies 10,541 observations out of 11,072 observations.

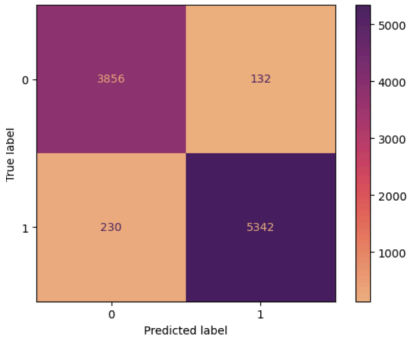
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	98.85	0.07	94.97	0.23
4	Gradient Boosting Classifier	98.56	0.06	93.79	0.19
2	Bagging Classifier	98.36	0.10	94.79	0.33
1	Decision Tree Classifier	96.65	0.21	94.15	0.15
0	Logistic Regression	90.80	0.34	82.92	0.41
5	KNN	76.41	0.44	70.89	0.42



Age group 5 (45-54)

The Random Forest Classifier works the best for this subset. It has an accuracy of 95.59% and an AUC mean of 99.20%. It correctly classifies 9,198 observations out of 9,560 observations.

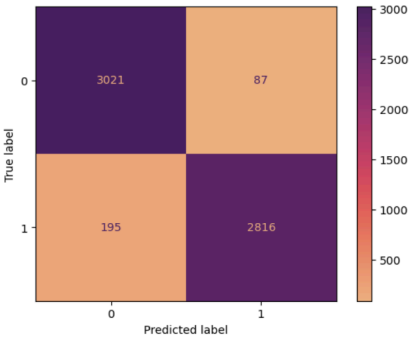
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	99.20	0.12	95.59	0.38
4	Gradient Boosting Classifier	98.98	0.15	94.84	0.26
2	Bagging Classifier	98.90	0.22	95.45	0.32
1	Decision Tree Classifier	97.21	0.38	94.77	0.27
0	Logistic Regression	93.29	0.58	86.20	0.68
5	KNN	77.63	0.54	71.59	0.74



Age group 6 (55-64)

The Gradient Boosting Classifier works the best for this subset. It has an accuracy of 95.49% and an AUC mean of 99.19%. It correctly classifies 5,837 observations out of 6,119 observations.

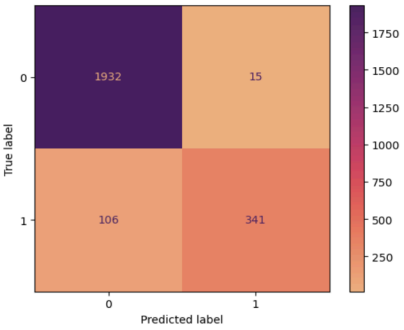
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
4	Gradient Boosting Classifier	99.19	0.12	95.49	0.40
3	Random Forest Classifier	99.07	0.18	95.55	0.42
2	Bagging Classifier	98.79	0.08	95.42	0.21
1	Decision Tree Classifier	95.98	0.11	94.45	0.22
0	Logistic Regression	93.41	0.18	87.34	0.31
5	KNN	76.55	0.06	70.95	0.24



Age group 7 (>65)

The Gradient Boosting Classifier works the best for this subset. It has an accuracy of 94.76% and an AUC mean of 98.25%. It correctly classifies 2,273 observations out of 2,394 observations.

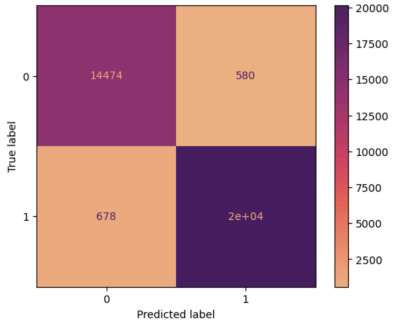
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
4	Gradient Boosting Classifier	98.25	0.18	94.76	0.22
3	Random Forest Classifier	97.55	0.29	94.62	0.35
2	Bagging Classifier	96.87	0.59	94.40	0.36
1	Decision Tree Classifier	93.48	1.86	93.54	0.62
0	Logistic Regression	83.89	0.92	87.46	0.42
5	KNN	61.17	0.60	79.99	0.41



Business Travel

The Random Forest Classifier works the best for this subset. It has an accuracy of 96.25% and an AUC mean of 99.37%. It correctly classifies 30,474 observations out of 31,732 observations.

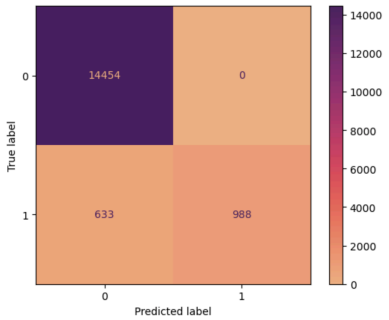
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	99.37	0.06	96.25	0.15
2	Bagging Classifier	98.93	0.15	95.80	0.20
4	Gradient Boosting Classifier	98.75	0.08	94.26	0.22
1	Decision Tree Classifier	95.39	0.12	94.98	0.20
0	Logistic Regression	85.83	1.64	78.15	1.89
5	KNN	77.71	0.37	71.74	0.08



Personal Travel

The Gradient Boosting Classifier works the best for this subset. It has an accuracy of 95.80% and an AUC mean of 97.83%. It correctly classifies 15,442 observations out of 16,075 observations.

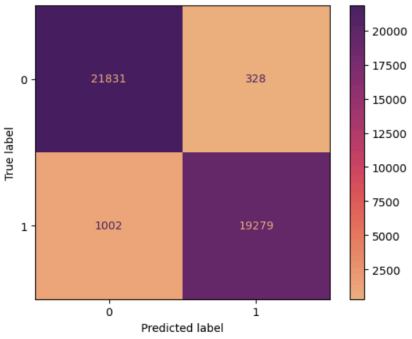
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
4	Gradient Boosting Classifier	97.83	0.07	95.80	0.32
3	Random Forest Classifier	97.64	0.06	95.72	0.34
2	Bagging Classifier	96.55	0.25	95.32	0.24
1	Decision Tree Classifier	79.37	1.20	95.80	0.32
0	Logistic Regression	74.35	1.73	90.36	0.51
5	KNN	58.49	0.62	89.76	0.27



Loyal Customer

The Random Forest Classifier works the best for this subset. It has an accuracy of 96.67% and an AUC mean of 99.46%. It correctly classifies 41,110 observations out of 42,440 observations.

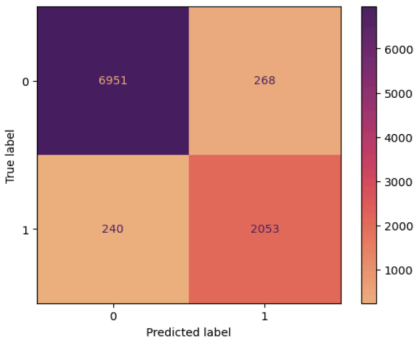
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest Classifier	99.46	0.04	96.67	0.12
2	Bagging Classifier	99.16	0.07	96.37	0.05
4	Gradient Boosting Classifier	99.12	0.06	95.17	0.16
1	Decision Tree Classifier	97.24	0.26	95.70	0.08
0	Logistic Regression	88.61	0.24	81.93	0.43
5	KNN	79.89	0.21	73.37	0.26



Disloyal Customer

The Gradient Boosting Classifier works the best for this subset. It has an accuracy of 94.03% and an AUC mean of 98.17%. It correctly classifies 9,004 observations out of 9,512 observations.

	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
4	Gradient Boosting Classifier	98.17	0.08	94.03	0.09
3	Random Forest Classifier	97.81	0.14	93.77	0.21
1	Decision Tree Classifier	97.47	0.15	93.84	0.16
2	Bagging Classifier	97.12	0.09	93.12	0.28
0	Logistic Regression	80.39	0.45	83.96	0.57
5	KNN	69.45	0.12	77.83	0.17



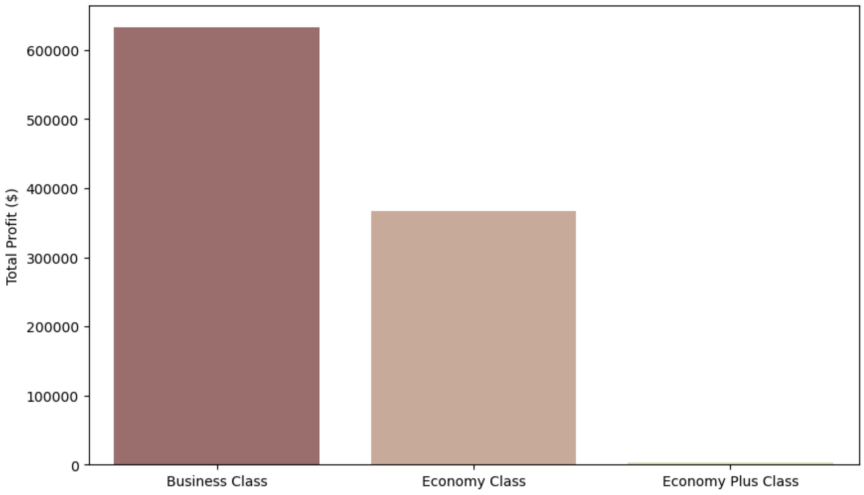
Business Decision

We are selecting the best-performing model for each subset and identifying the top 3 most essential features of that model. For each of these features, we plan to increase the satisfaction rating by one and predict how many passengers who were initially dissatisfied will be predicted as satisfied based on the test data. Assuming the predictions are accurate (since most model accuracies are above 93%), we will calculate each scenario's cost, profit, and cost-benefit.

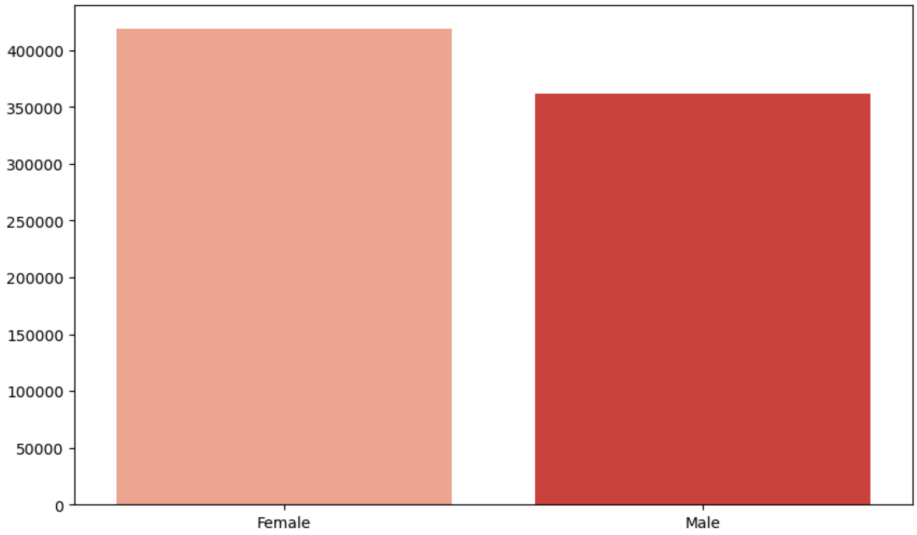
We assigned the profit per turned (dissatisfied in the test but prediction with the increased satisfaction predicts satisfied) by $500.We assigned the following costs (by unit increase) for each variable:

1. Flight Distance: $400
2. Inflight Wifi Service: $80
3. Ease of Online Booking: $60
4. Gate location: $140
5. Food and Drink: $100
6. Online Boarding: $70
7. Seat Comfort: $110
8. Inflight Entertainment: $50
9. On-board Service: $30
10. Leg-room Service: $500 (Bigger plane impractical)
11. Baggage Handling: $20
12. Check-in Service: $20
13. Inflight Service: $40
14. Cleanliness: $20

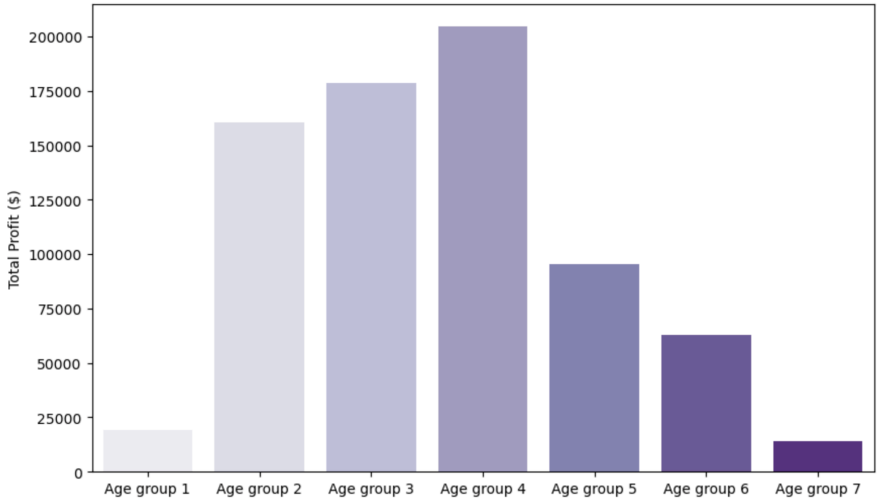
Total Profit over different Classes



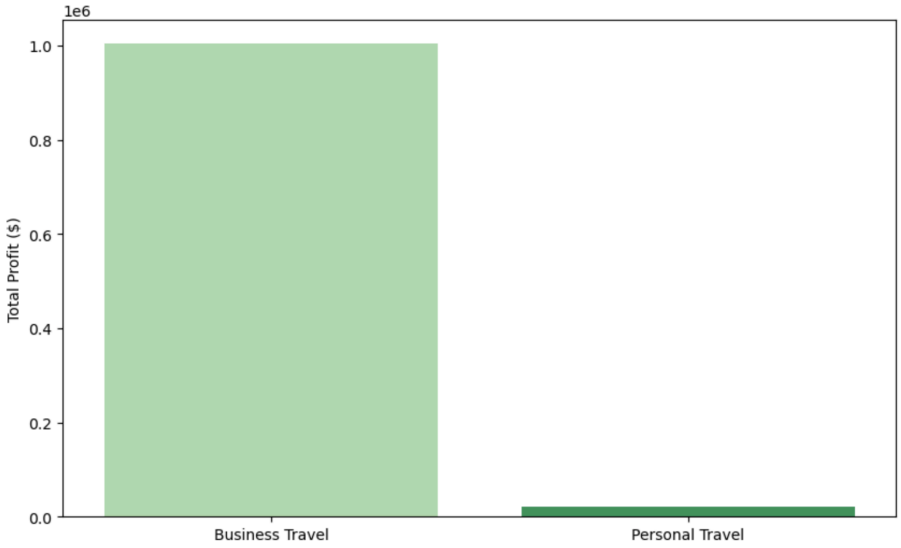
Total Profit over different Genders



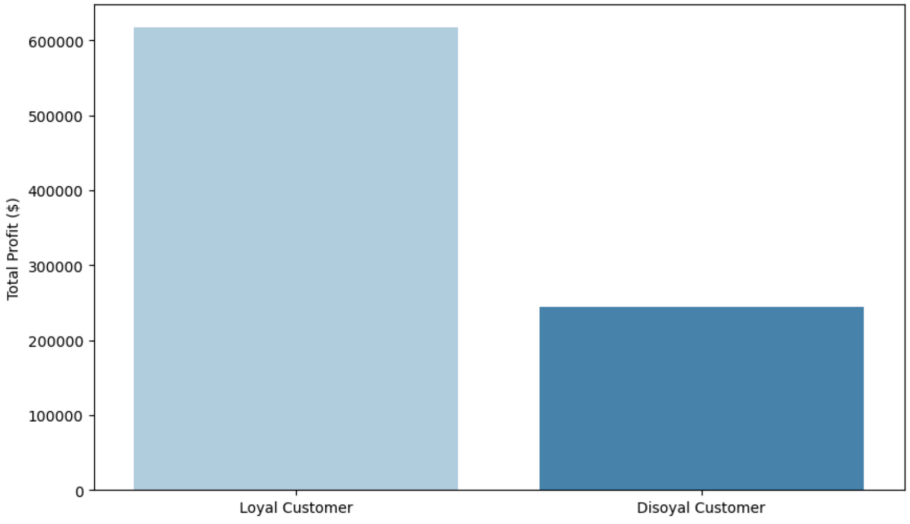
Total Profit over different Age groups



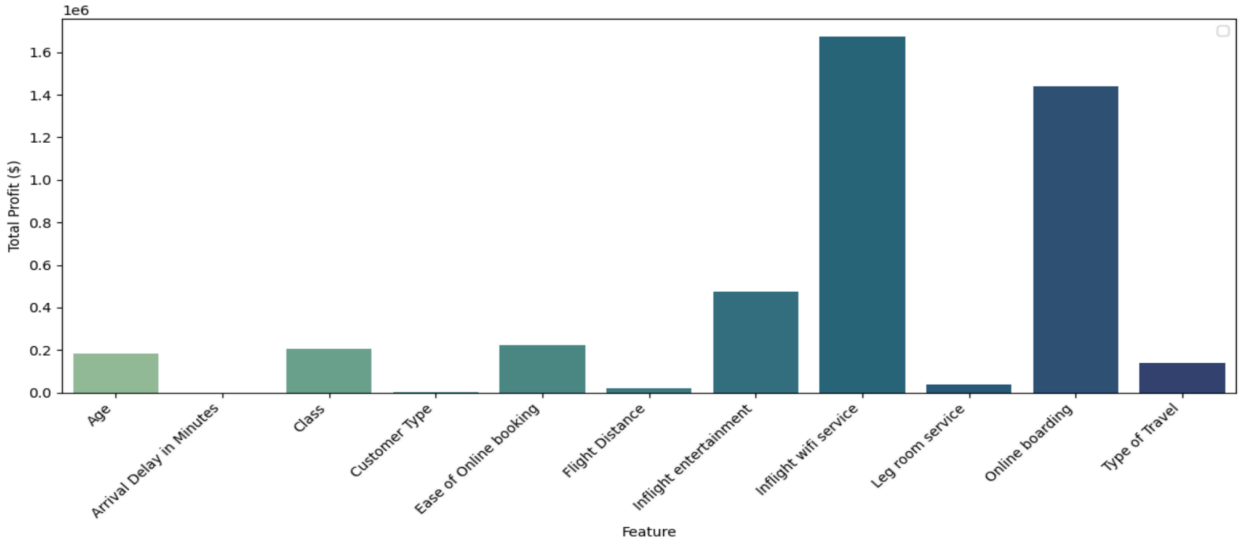
Total Profit over different Types of Travel



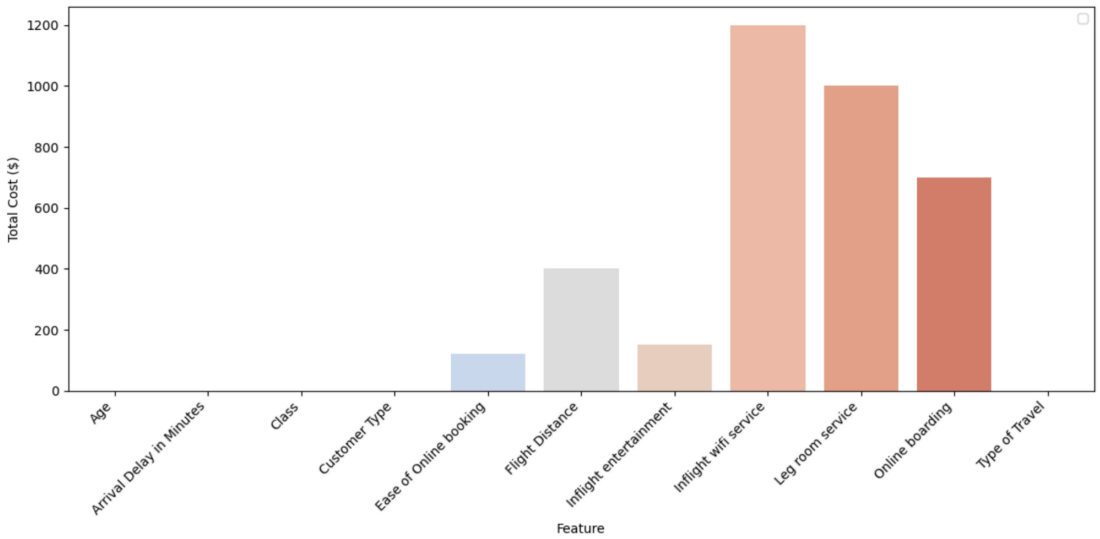
Total Profits over different Types of Customers



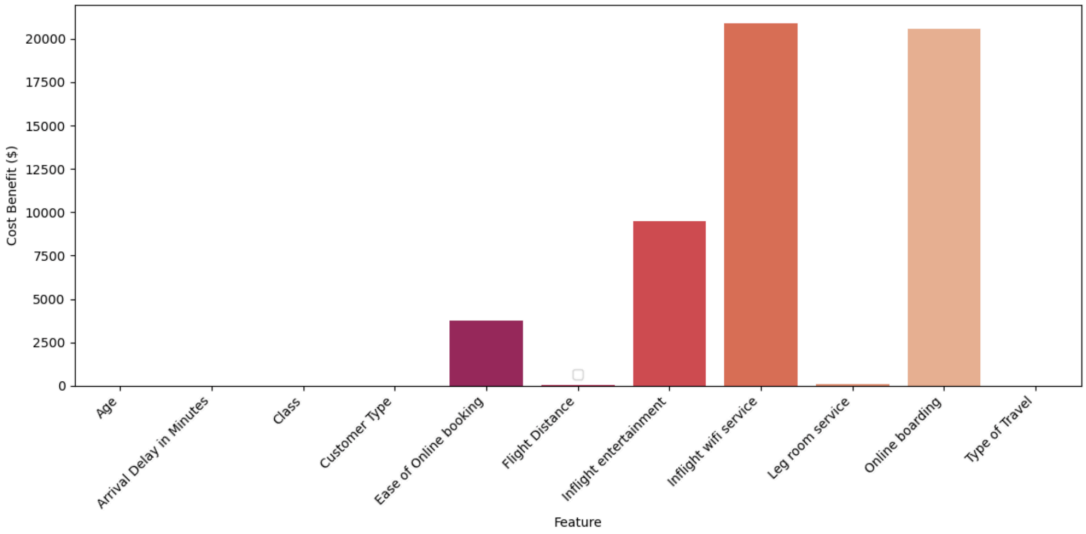
Total Profits over different features



Total Costs over different features



Total Cost Benefit over different features



Conclusion and Recommendations

Based on the findings and analysis of all our models, we have reached several key conclusions regarding the dataset.

The two best-performing models for our selected subsets are the Random Forest Classifier and the Gradient Boosting Classifier. They were consistently outperforming the other models in terms of Accuracy and AUC and so we formulated our business decision using these models for the respective subsets.

In conducting the Decision Performance, we can see that Inflight wifi service and Online Boarding are the two most profitable features in terms of cost-benefit, with both yielding over $20,000 if increased, followed by Inflight entertainment and Ease of Online Booking, yielding around $10,000 and $3,500 respectively. Hence, we would recommend that Invistco Airlines prioritizes improving these features by directing more of their budget towards increasing Inflight wifi service and entertainment, and making Online booking/boarding more available, as this would make more customers satisfied, and consequently increase future profits and revenue.

Alongside this, we can also make some recommendations based on the preliminary data analysis and general observations. Invistco could focus on improving satisfaction levels for subgroups with high dissatisfaction rates. For example, in the eco group, over 60% of customers are dissatisfied. They can do this by offering tailored services and amenities catering to their needs and preferences. Invistco could also consider implementing targeted retention strategies for disloyal customers to improve their satisfaction and prevent churn. They can do this by offering discounts and deals/offers to new customers.

Limitations

When analyzing and interpreting our data, we encountered some limitations that made our analysis slightly more difficult. Although the dataset was rich with features (25 columns), it would have been helpful to have more information in the dataset such as ticket price and Inflight service/food costs as this would have made calculating the Business Decision easier with given values rather than assumed values. Another limitation was the fact that Satisfaction was in the form of a categorical variable (Binary) where customers were either grouped as satisfied or neutral/dissatisfied. One can say that it would have been more practical for this variable to be numerical in the form of a 1-10 score measuring customer satisfaction as this would have made the analysis more detailed and thorough in terms of knowing how much customer satisfaction increased out of 10 rather than the “Satisfied” or “Neutral/Dissatisfied” classification.

Author Contributions

All team members collaborated equally on all the parts of the project.