Machine Learning Project Report

Team 3

Name	Section	BN	Load
Mostafa Magdy	2	23	Adaboost/SVM
Omar Said Mohammed	1	29	Preprocessing
Nour Ayman	2	30	Logistic Regression
Abdelrahman Mohamed Mahmoud Abdelfatah	1	24	Decision Tree

Problem

This dataset contains an airline passenger satisfaction survey. By investigating which factors are most correlated with passenger satisfaction (or dissatisfaction), we can predict satisfaction levels to help airlines improve their services.

Motivation

Understanding passenger satisfaction enables airlines to enhance service quality and increase customer retention. By analyzing key influencing factors, we can build predictive models to proactively identify dissatisfied passengers and improve the overall travel experience.

Evaluation Metrics:

1. **Accuracy, Precision, Recall, and F1-score to measure classification performance**
2. **Confusion Matrix to analyze false positives/negatives and ensure balanced predictions**

Dataset characteristics

Airline Passenger Satisfaction
Dataset Size: +100k row
Number of features: 25

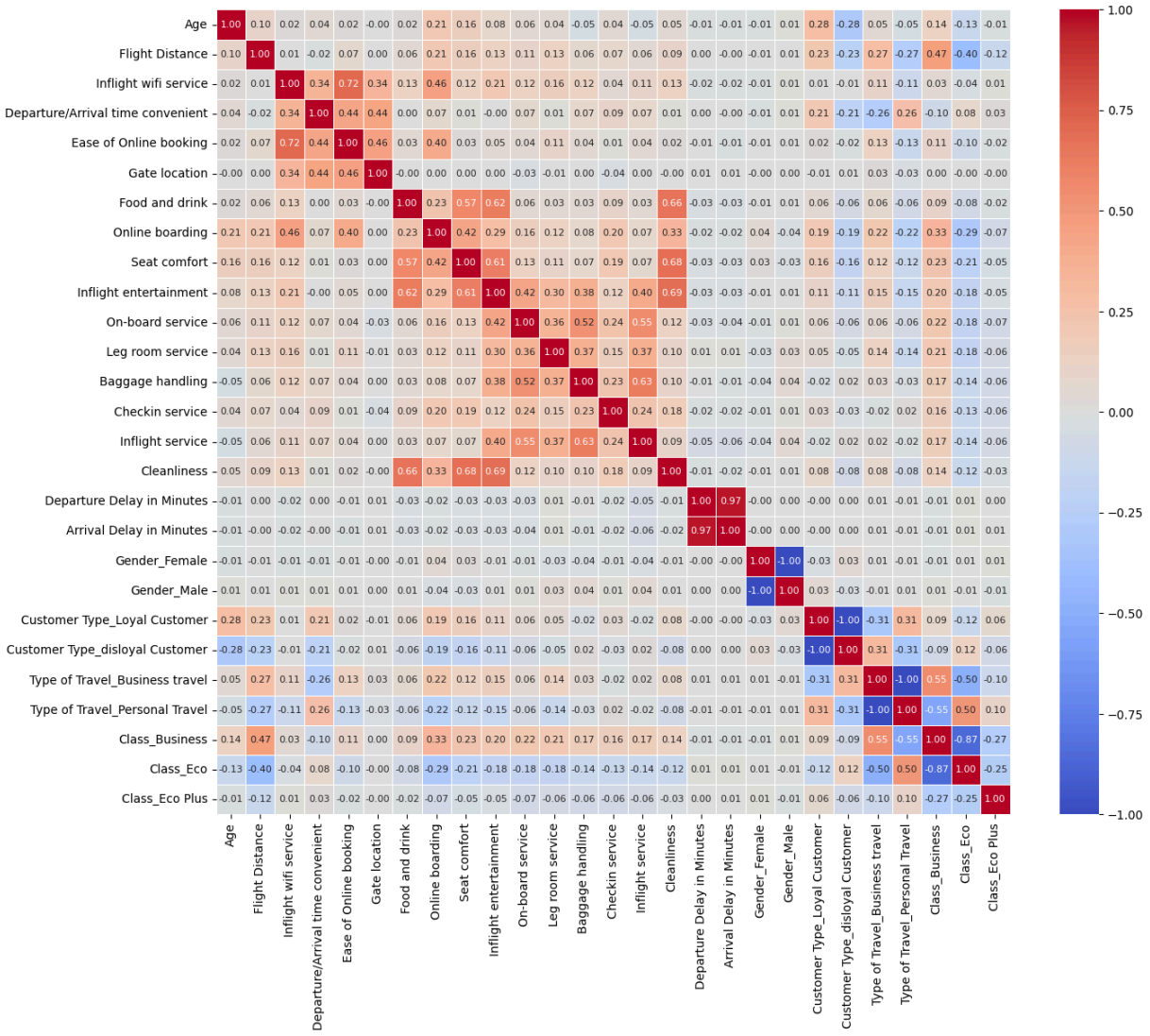
Data Preprocessing

1. Handling Missing Values

- Identifying missing (NULL) values (found) `Arrival Delay in Minutes` has 310 null values.
- The dataset contains a feature `Departure Delay in Minutes` which may the cause of the arrival delay
- Using `IterativeImputer` to fill missing arrival delay values based on departure delays.
- The imputer basically learns a relationship between departure delays and arrival delays to estimate missing arrival delays.

2. Remove Correlated Features

We visualized the correlation between features with a heatmap



Correlation Matrix Insights

- It is visible that Arrival Delay highly positively correlates with Departure Delay.
- No need to keep both of the features, it is enough to keep one of them.

Note: The highly negatively correlated features are categorical ones which were turned into numericals.

3. Removing Outliers

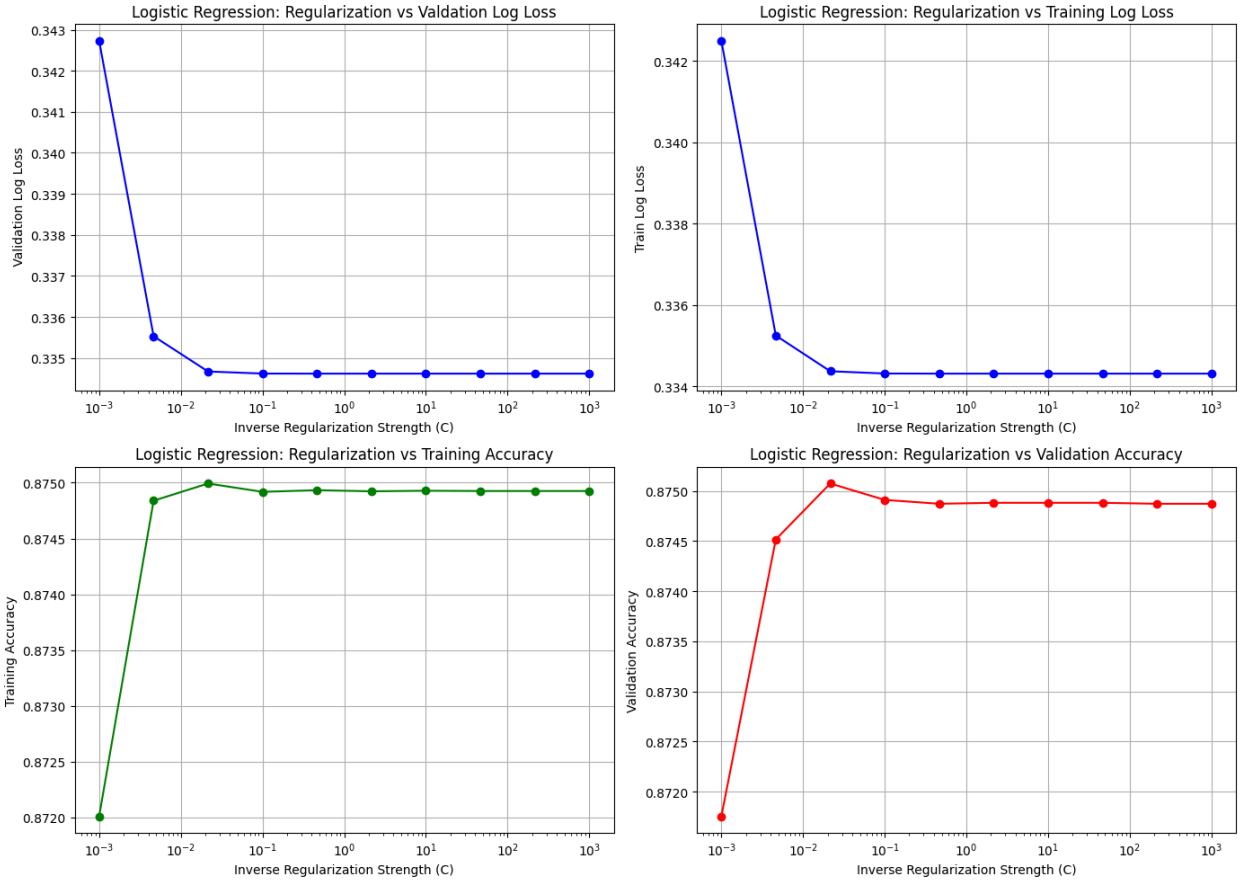
- Using Mahalanobis distance to identify outliers in numerical features
- Calculating the covariance matrix and its inverse
- Setting a threshold using the chi-square distribution (95th percentile)
- Identifying the outliers based on threshold and removing them.

Models

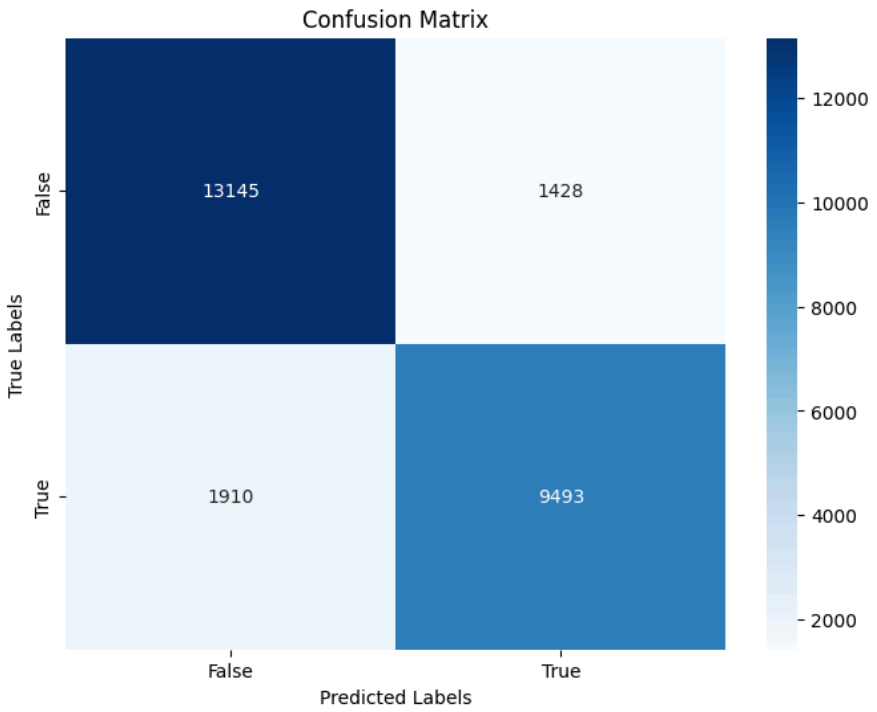
1. Logistic regression

- Trained Logistic regression model with L2 regularization and max iterations=10000.
- Tuned the model for proper C inverse regularization strength
- Used K-fold Cross validation (k=5) as a validation metric.

Results and Graph



- **Training Accuracy (F1-score) = 0.88**
- **Testing Accuracy (F1-score) =0.87**
- **Confusion Matrix for Test data**



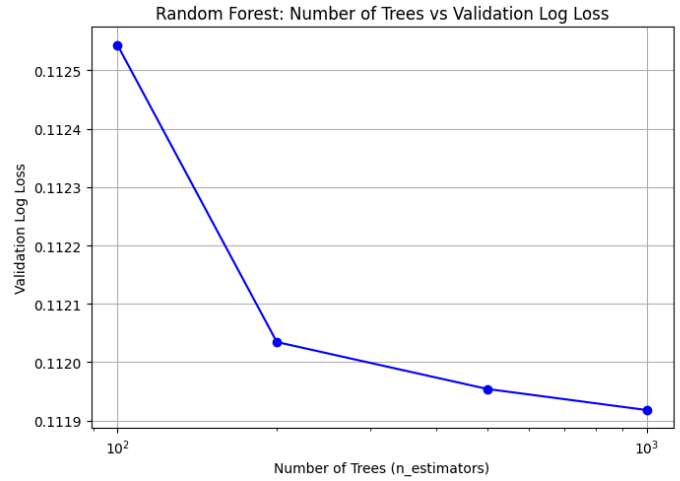
2. SVM

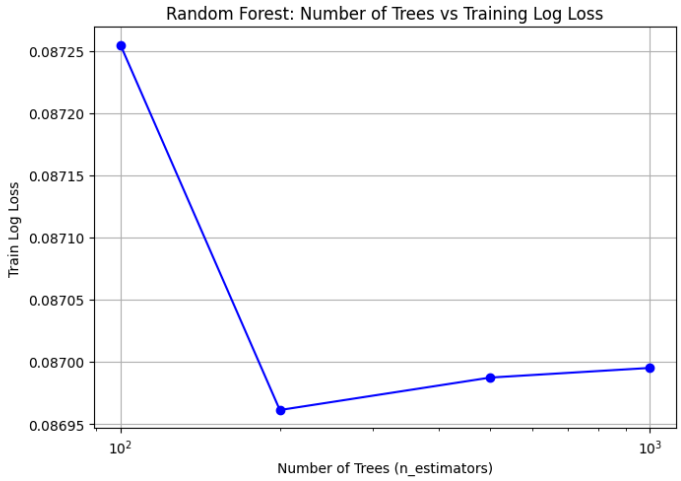
- Trained the model using k-fold cross-validation, tuning hyperparameters C (regularization) and gamma (influence of individual samples).
- Attempted training with an RBF kernel (non-linear), but it was too slow.
- Switched to a linear kernel, but the training loop was getting slower over the time as hyperparameter search took over 12 hours in the final iteration.
- Best validation accuracy: 87%, which is similar to logistic regression.

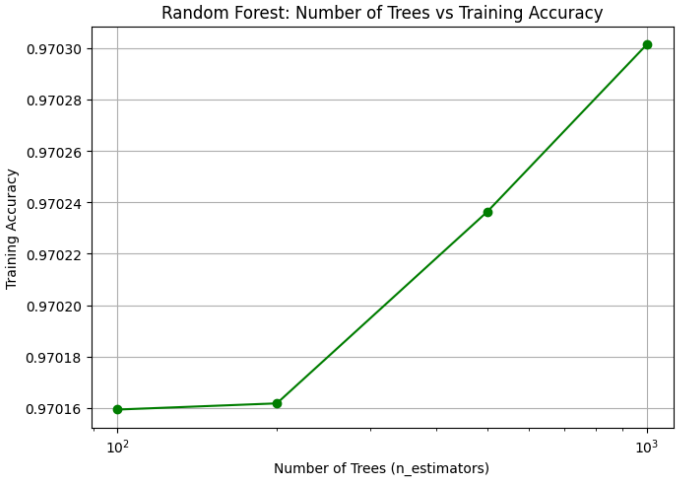
3. Random Forest

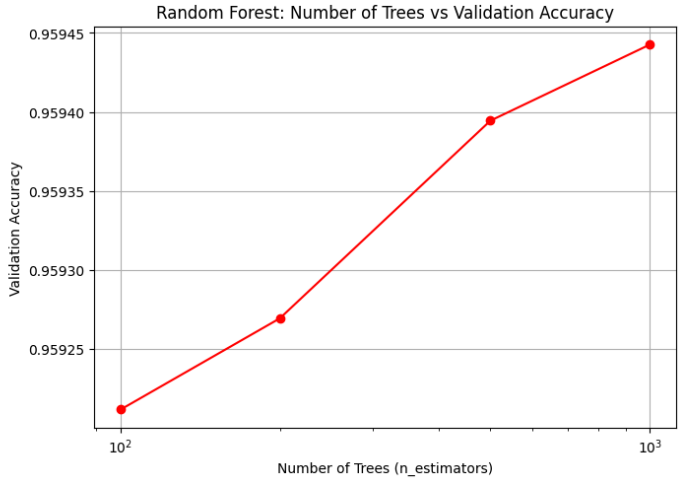
- Trained Random Forest model with `max_depth=20` and `min_samples_leaf=5` and `class_weight=Balanced`
- Used K- Fold cross validation and tuned the number of estimators as a hyperparameter.

Results and Graphs

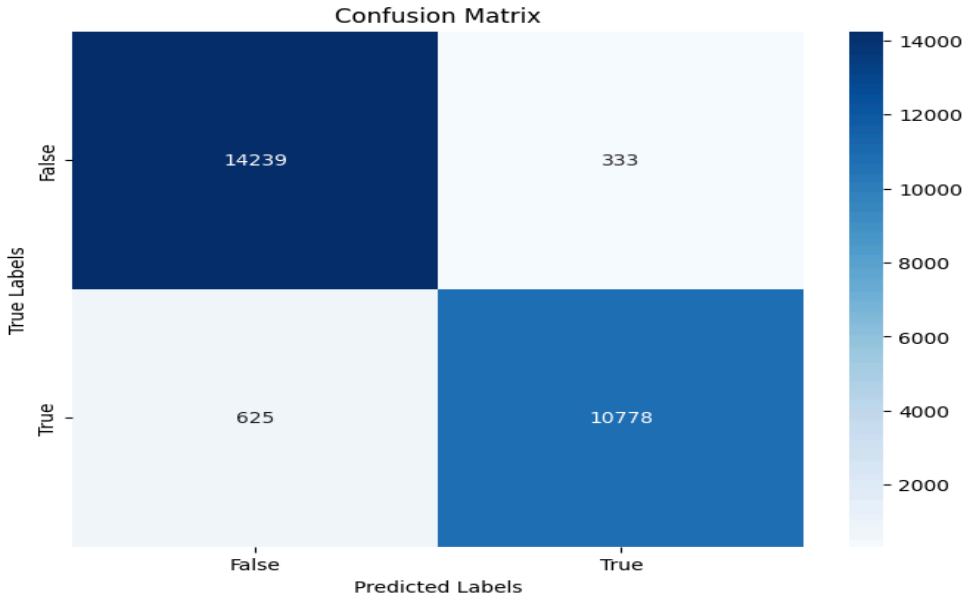








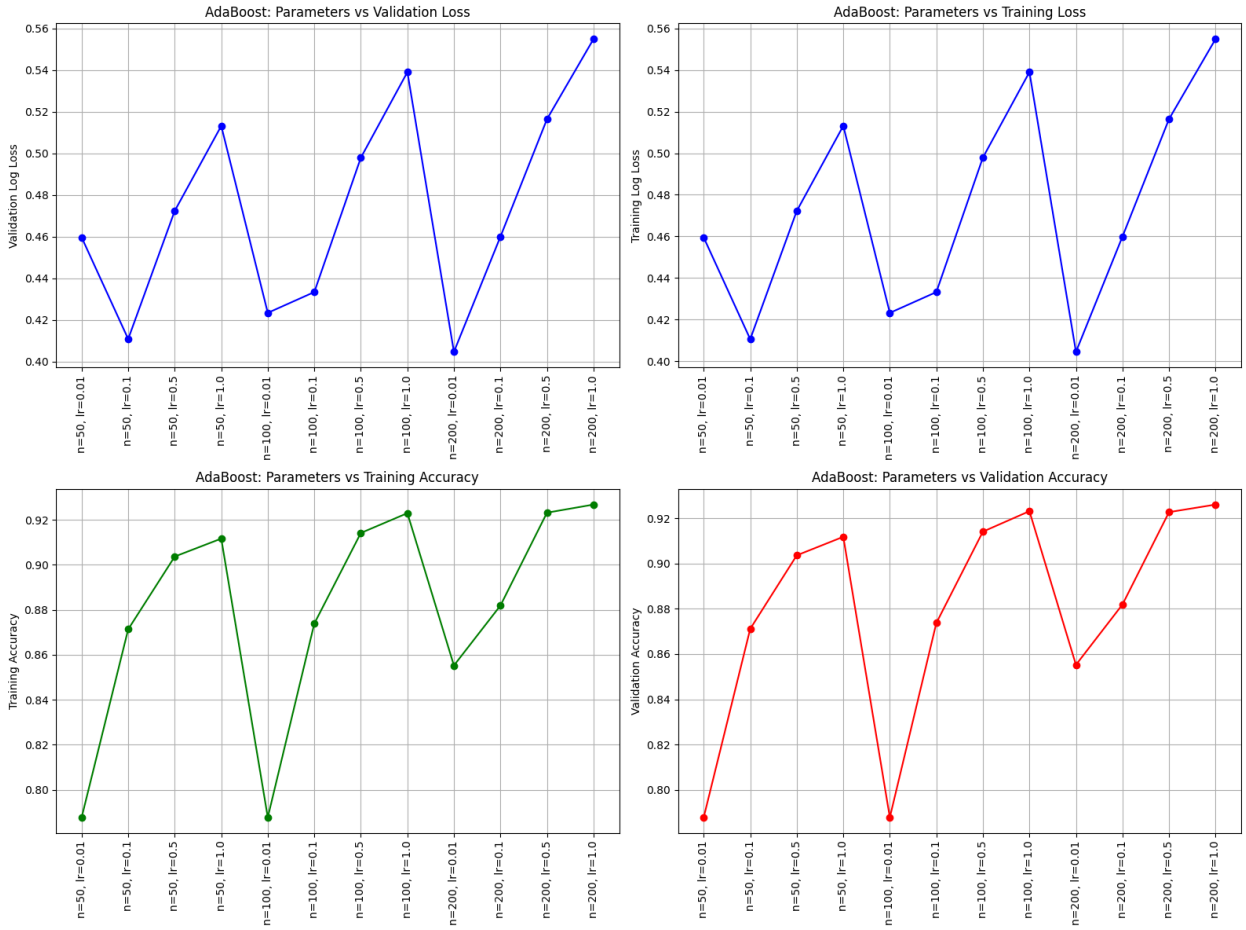
- Training Accuracy (F1-Score)=0.99
- Test Accuracy (F1-Score)=0.96
- Test Data Confusion Matrix



4. Adaboost Logistic Regression

- Utilized Ensembled Techniques for the linear models as the normal logistic regression does not perform too well on the data (86% test accuracy).
- Used LogisticRegression as the base estimator for the AdaBoostClassifier with max_iter=1000
- Tuned The number of used estimators and the learning rate

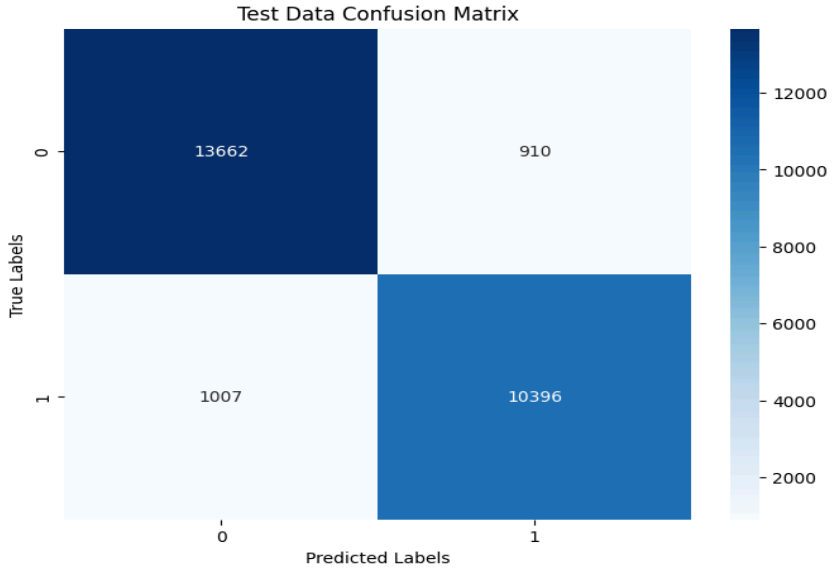
Graphs and results



Training accuracy (F1-Score) =0.93

Test accuracy (F1-Score)=0.93

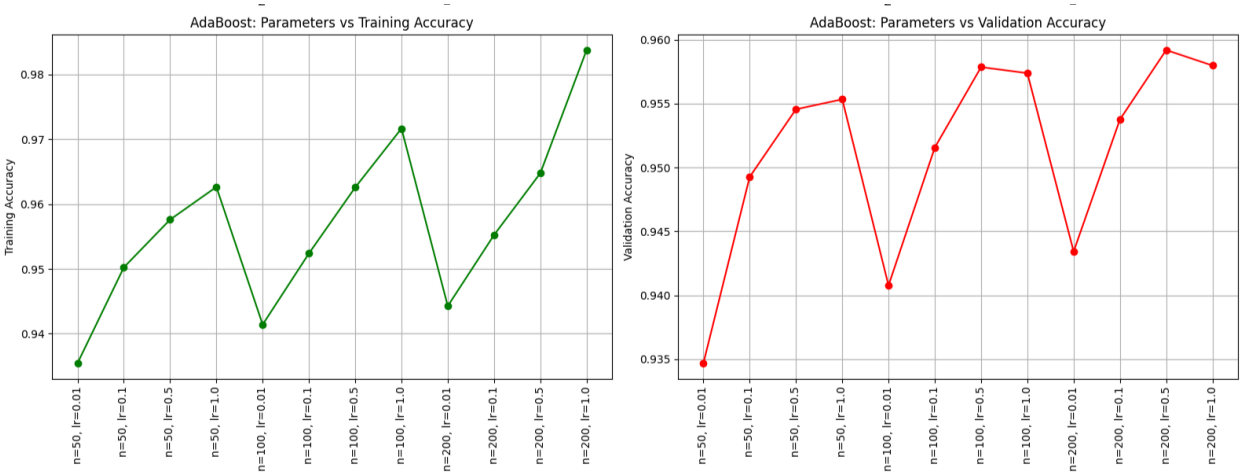
Test Confusion Matrix

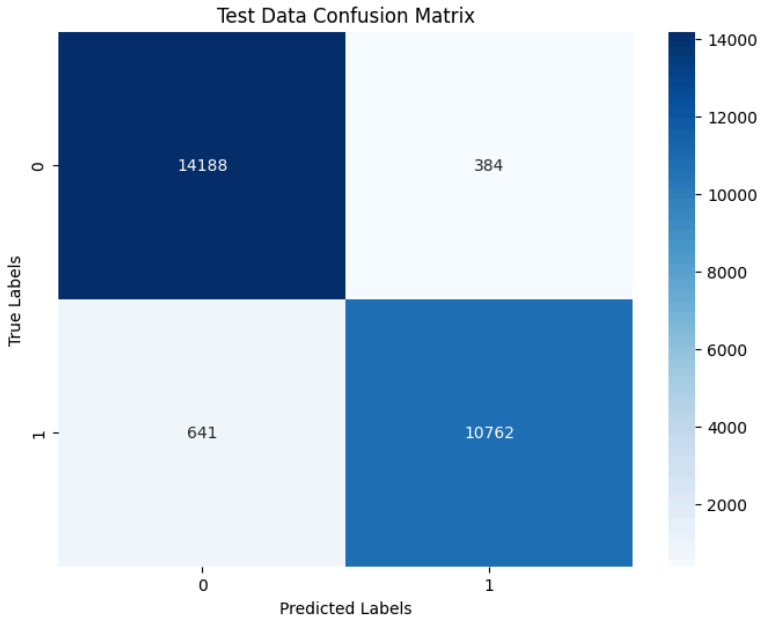


5. Adaboost (Decision Stumps and Tree depth=6)

- Tested decision stumps to determine if ensemble techniques with non-linear models (decision trees) outperform standard decision trees.
- Fine-tuned key parameters, including learning rate and n_estimators.
- Achieved an accuracy of `0.93` with decision stumps and `0.96` with deeper trees (`max_depth=6`).
- Further increasing max_depth might improve results, but current performance appears near optimal.

Graphs and results





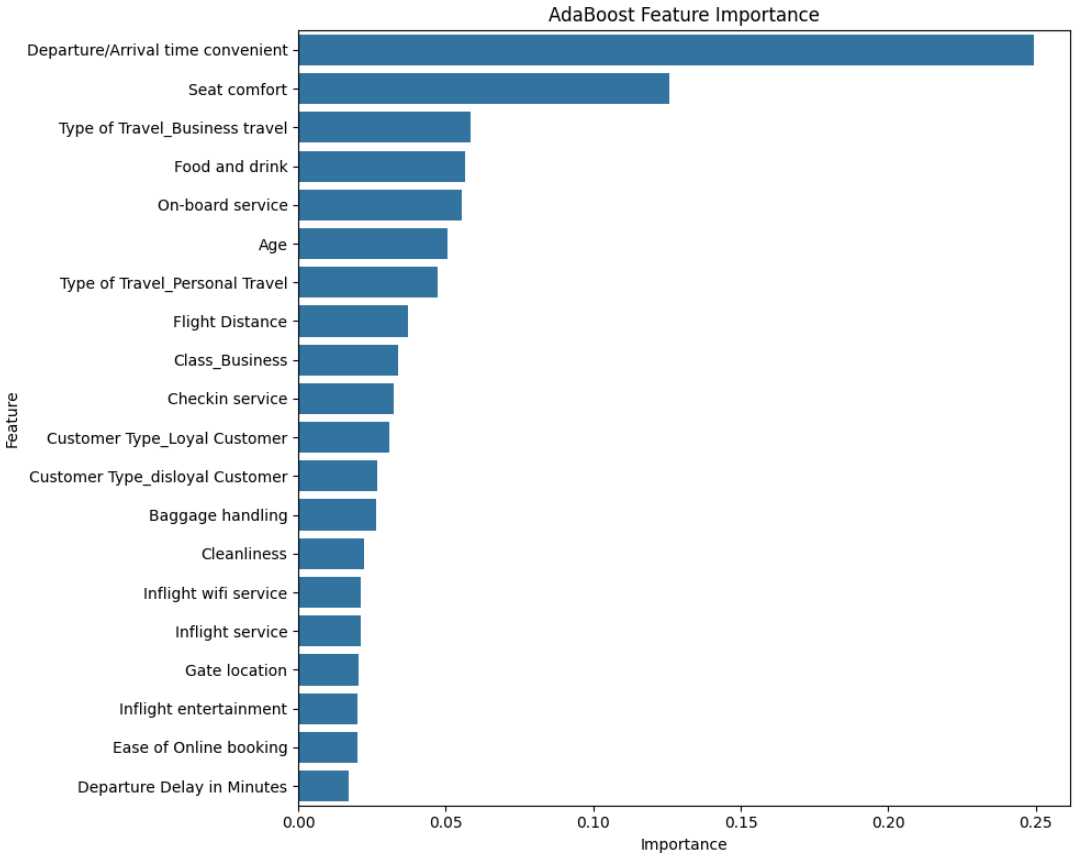
Conclusion

Non linear models best fit the function of Prediction Customer satisfaction. Linear models such as (logistic regression and linear SVM) has moderate accuracy.

This suggest

1. Complex Non-Linear Relationships: The relationship between passenger features and satisfaction cannot be captured by a linear decision boundary. The data appears to have complex, non-linear patterns.
2. Hierarchical Patterns: The superior performance of tree-based models (Random Forest and AdaBoost with decision trees) suggests the data contains hierarchical patterns and distinct customer segments that respond differently to different features.

Here are the best model Important Features



The results suggest that departure/arrival time convenience, seat comfort, and food/drink quality have the strongest impact on customer satisfaction. This aligns with logical expectations, indicating that airlines should prioritize improving these areas to enhance customer satisfaction.