AIRLINE PASSENGER SATISFACTION

Team Project : Group 1

Date : 26th September 2023

AGENDA

01

Problem Statement & Feature Description

02

Story Telling

03

Data Preprocessing

04

Baseline Model

05

Hyperparameter Tuning

06

Feature Selection

07

Model Fit Check

08

Conclusion

1:01

PROBLEM STATEMENT & FEATURE DESCRIPTION

To develop a predictive model to determine passenger satisfaction in the airline industry based on a comprehensive set of traveler-related and service-related factors.
This model aims to uncover the primary drivers of satisfaction and assist airlines in enhancing the overall travel experience.

This dataset contains a satisfaction survey of passengers who travel in flights

1:01

Customer Related Features

1. Gender
2. Customer Type
3. Age
4. Type of Travel
5. Class

Services Related Features

1.Inflight WIFI service

2.Departure/Arrival time convenient

3.Ease of Online booking

4.Gate location

5.Food and drink

6.Online boarding

7.Seat comfort

8.Inflight entertainment

9.On-board service

10.Leg room service

11.Baggage handling

12.Check in service

13.Inflight service

14.Cleanliness

Flight Related Features

1.Flight Distance

2.Departure Delay in Minutes

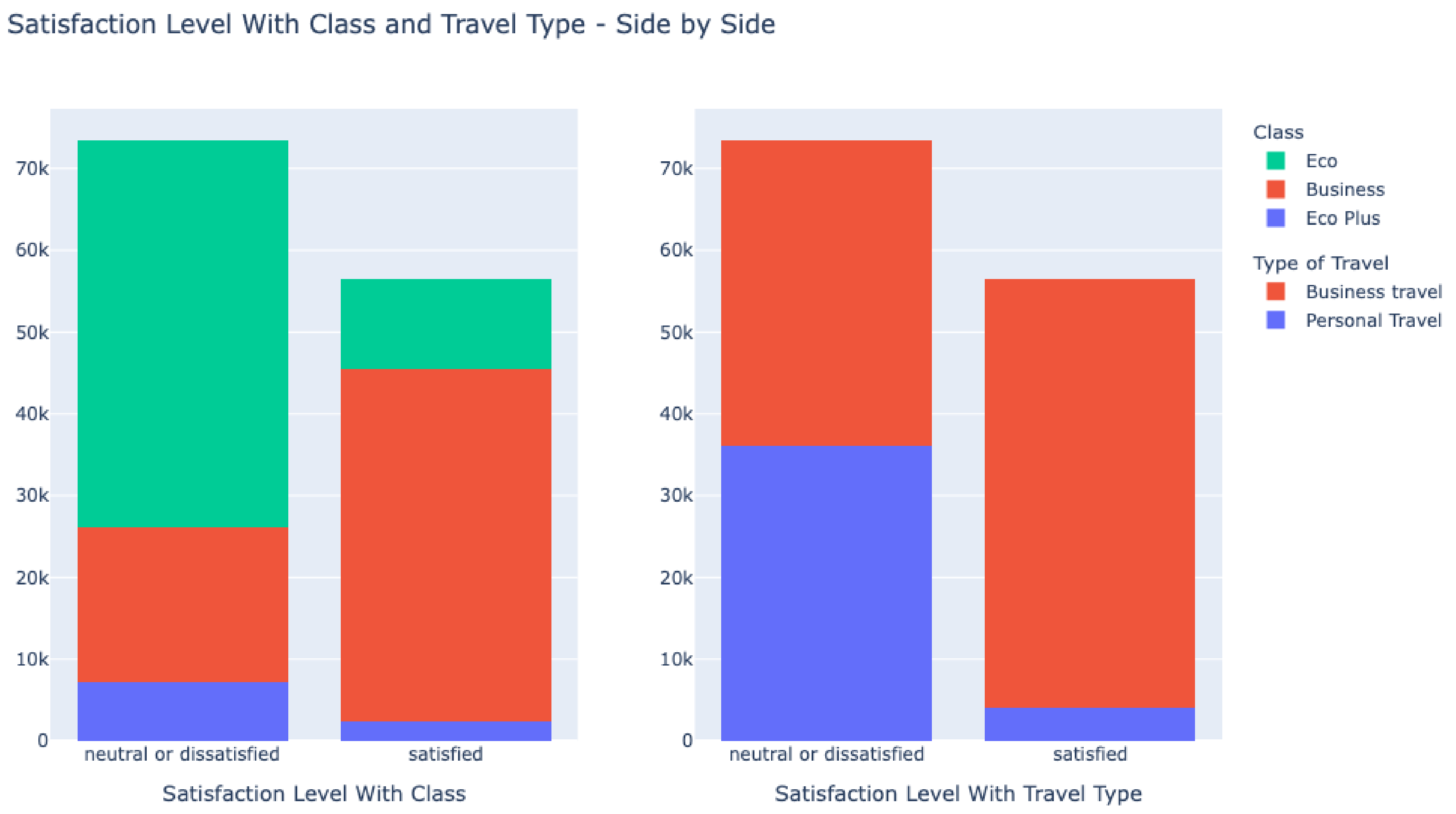
3.Arrival Delay in Minutes

Target Feature

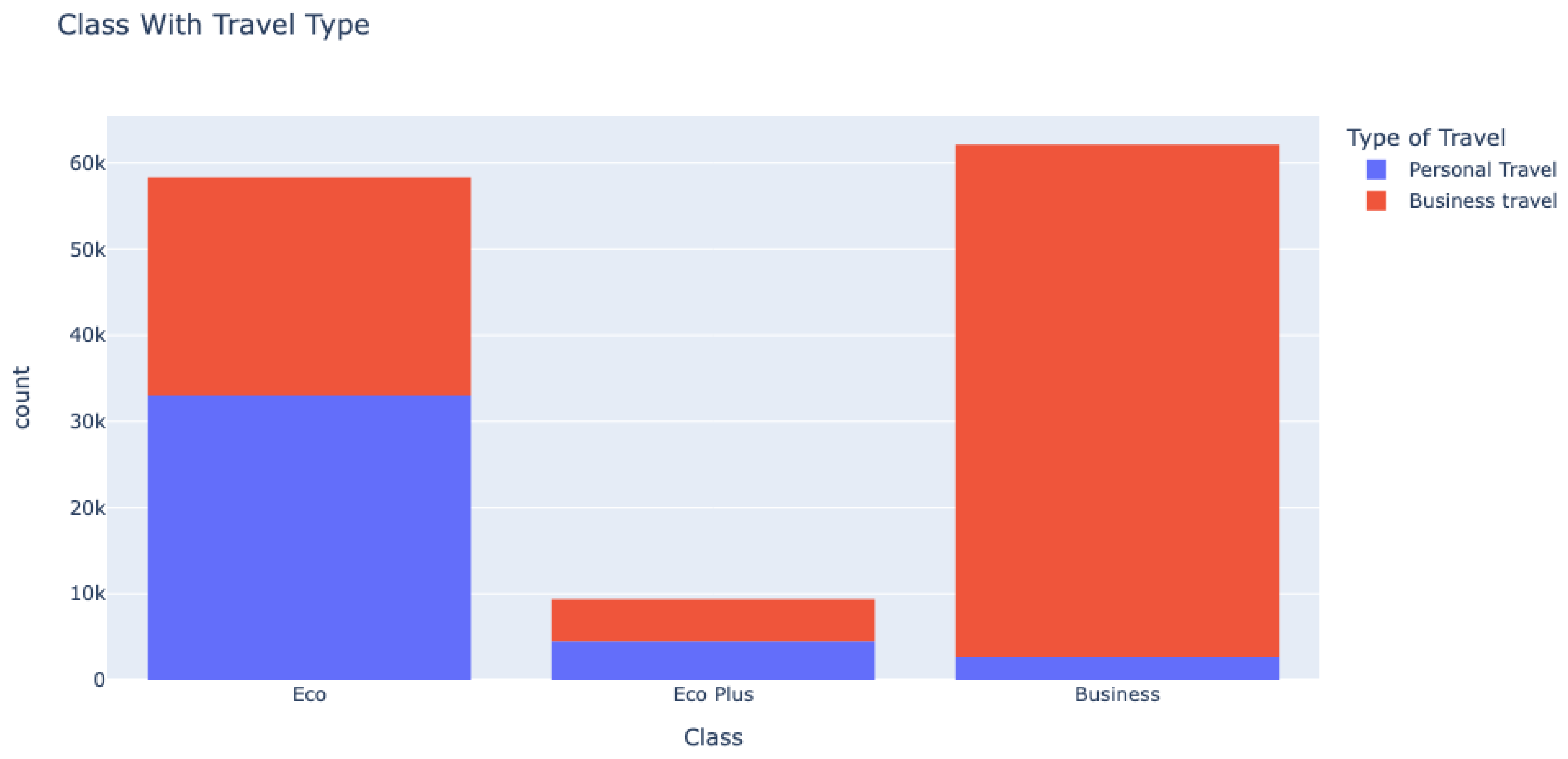
1.Satisfaction

1:01

STORY TELLING



Satisfaction Level With Class and Travel Type - Side by Side



Class With Travel Type

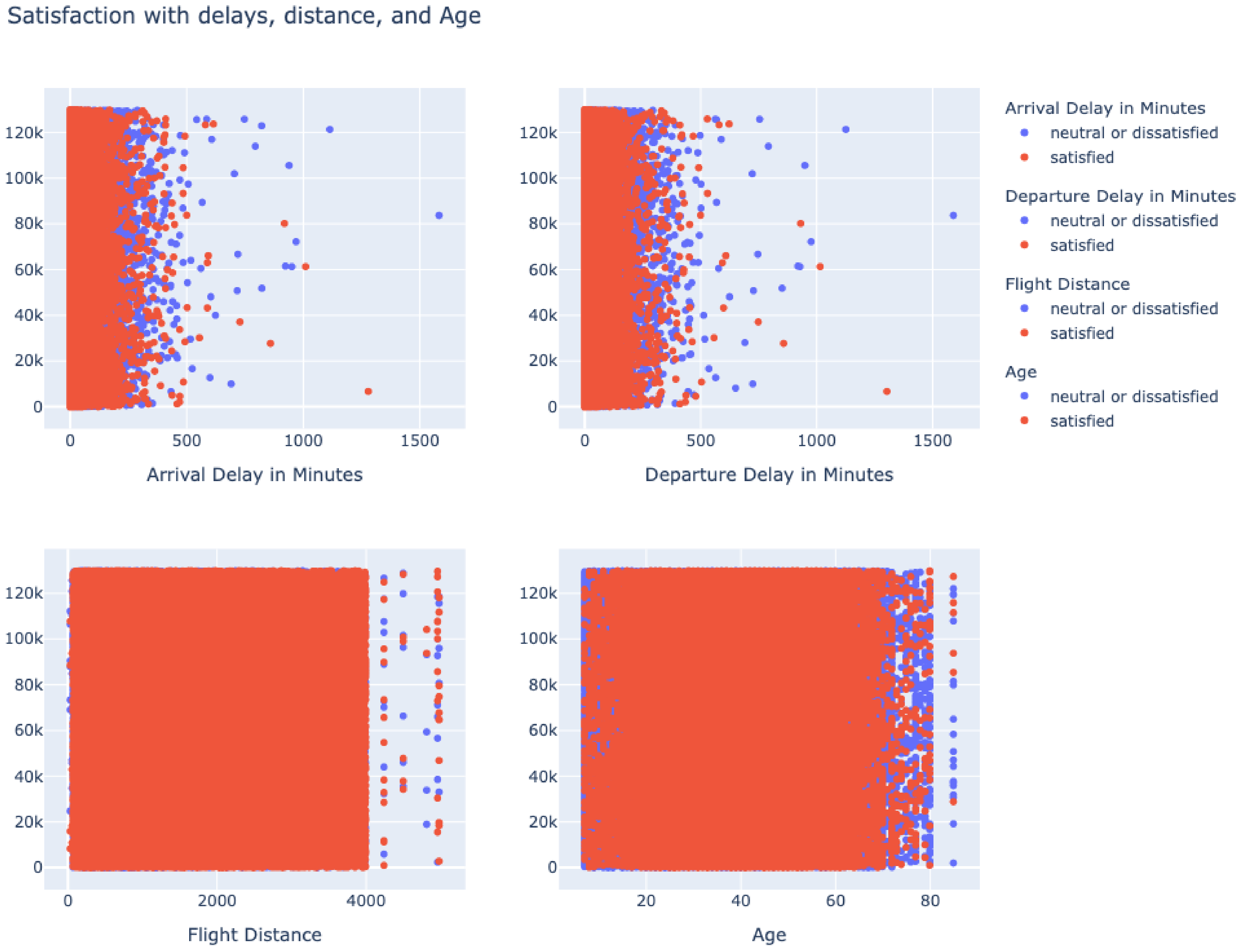
"Business travelers in Business class exhibit high satisfaction, reflecting the company's effective catering to this segment."

"Dissatisfaction is notable among Economy class personal travelers, emphasizing the need to enhance their experience for improved satisfaction and loyalty."

"Dissatisfaction among Economy class business travelers suggests an opportunity to promote Business class benefits while addressing cost concerns."

1:01

STORY TELLING



"Longer delays lead to a noticeable dip in satisfaction. Punctuality is key!" (>250min)

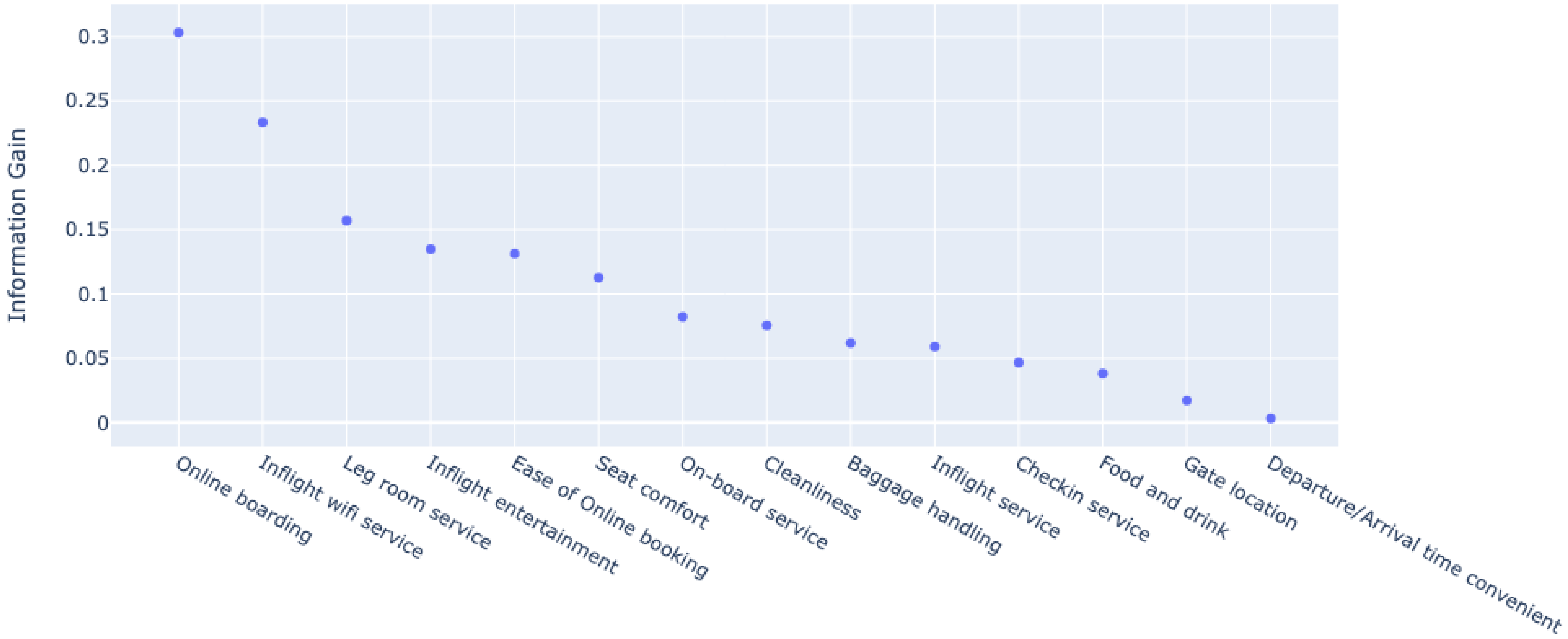
"Endurance Test: Flights exceeding 4000km tend to result in dissatisfaction. Comfort and services for long-haul need reevaluation."

"Age plays a role: Younger and Older travelers may have higher expectations or vice-versa."
(<20 and >60)

1:01

STORY TELLING

Importance of services over satisfaction

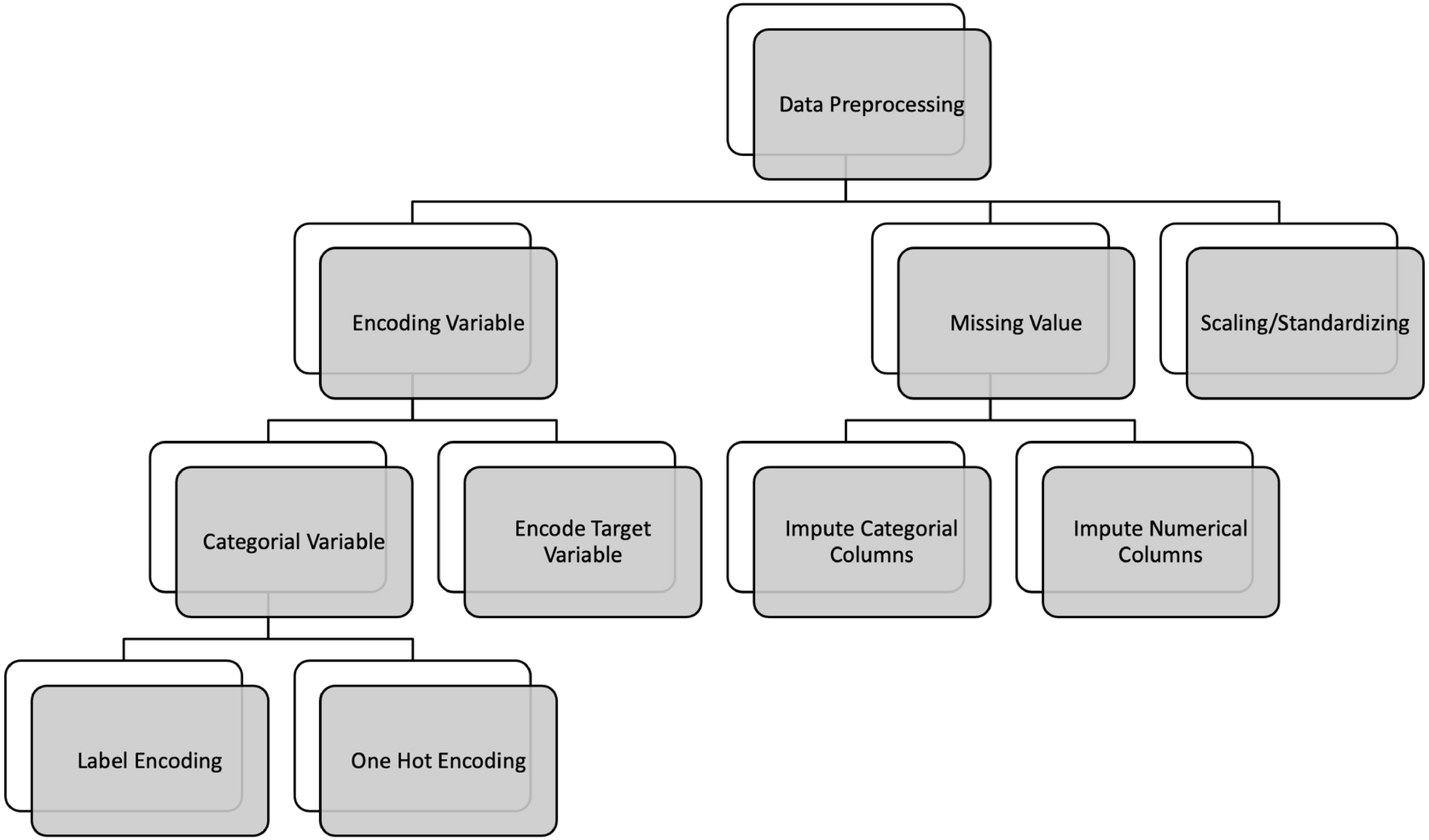


"Addressing these key services, even with small incremental changes, can lead to a significant leap in overall satisfaction."

- Excelling in boarding and inflight WiFi can greatly elevate passenger satisfaction.
- Comfort during the journey, especially legroom, plays a pivotal role.
- Entertainment and seamless booking processes are integral to a complete experience.

1:01

DATA PRE-PROCESSING



1:01

BASELINE MODEL

	Model	Accuracy	Precision	Recall	F1 Score
1	**Logistic Regression**	0.874192	0.869698	0.835637	0.852327
2	**DecisionTree Classifier**	0.946412	0.936628	0.940280	0.938451
3	**RandomForest Classifier**	**0.962119**	**0.971099**	**0.940812**	**0.955716**
4	**Stochastic Gradient Descent**	0.870958	0.853124	0.849194	0.851155
5	**Support Vector Machine**	0.944949	0.947025	0.925040	0.935903

1:01

HYPER PARAMETER TUNNING USING GRID SEARCH

Model	Params Grid	Best Params

Logistic Regression	{
	"C": [0.5, 1, 5, 10],
	"max_iter": [500, 1000]
	}
		{'C': 0.5, 'max_iter': 500}

DecisionTree Classifier	{'max_depth': [2, 20], 'min_samples_leaf': [2, 10, 100, 1000], 'criterion': ['gini','entropy', 'log_loss'], 'max_leaf_nodes': [10, 100, 1000], 'min_impurity_decrease': [0.000001, 0.0001, 0.001, 0.010], 'splitter': ['best', 'random']}	{'criterion': 'log_loss', 'max_depth': 20, 'max_leaf_nodes': 1000, 'min_impurity_decrease': 0.0001, 'min_samples_leaf': 2, 'splitter': 'best'}

RandomForest Classifier	{
	"n_estimators": [400, 500],
	"criterion": ["gini", "entropy", "log_loss"],
	"max_depth": [20, 25, 32],
	"min_samples_split": [1, 2]
	}
		{'criterion': 'log_loss',
		'max_depth': 25,
		'min_samples_split': 2,
		'n_estimators': 500}

Stochastic Gradient Descent	{
	"loss": ["hinge", "log_loss"],
	"penalty":["l2", "l1", "elasticnet"],
	"alpha": [0.0001, 0.001, 0.1,0.5]
	}
		{'alpha': 0.001,
		'loss': 'hinge',
		'penalty': 'l2'}

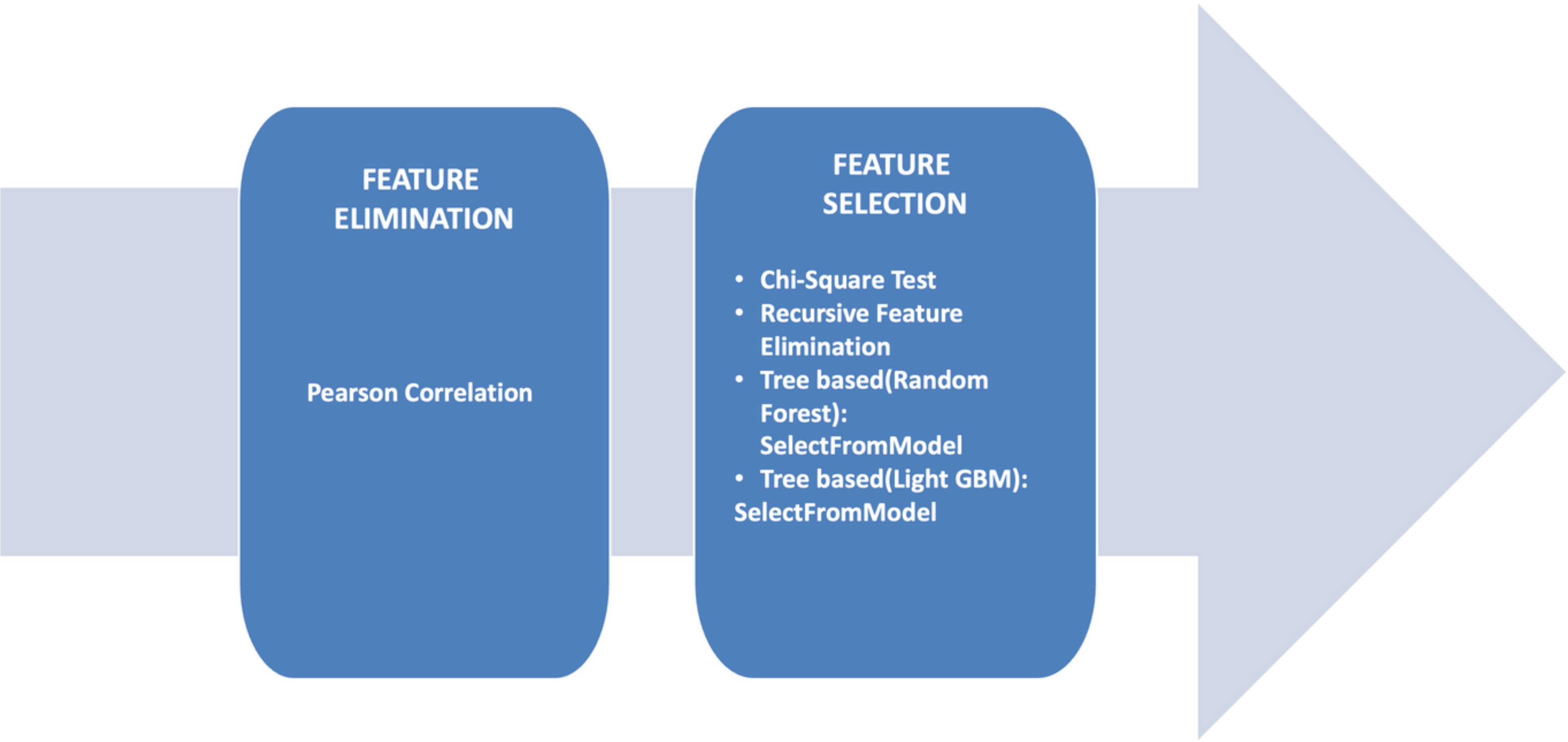
Support Vector Machine	{
	"C": [1, 5, 10],
	"kernel": ["linear", "rbf"],
	}
		{'C': 10,
		'kernel': 'rbf'}

1:01

HYPER PARAMETER TUNNING RESULTS

Model	Accuracy	Precision	Recall	F1_score
Basline Logistic Regression	0.874115	0.870015	0.835017	0.852157
Baseline DecisionTree Classifier	0.946682	0.936591	0.940989	0.938785
Baseline RandomForest Classifier	0.962388	0.971808	0.940723	0.956013
Baseline Stochastic Gradient Descent	0.876732	0.880531	0.828726	0.853843
Baseline Support Vector Machine	0.945180	0.946244	0.926458	0.936246
Parameter Tuned Logistic Regression	**0.874076**	**0.869730**	**0.835283**	**0.852158**
Parameter Tuned DecisionTree Regression	**0.957499**	**0.969217**	**0.931774**	**0.950126**
Parameter Tuned RandomForest Classifier	**0.962465**	**0.972678**	**0.940014**	**0.956067**
Parameter Tuned Stochastic Gradient Descent	**0.876771**	**0.896129**	**0.810296**	**0.851054**
Parameter Tuned Support Vector Machine	**0.955035**	**0.959742**	**0.935761**	**0.947600**

FEATURE SELECTION METHODS

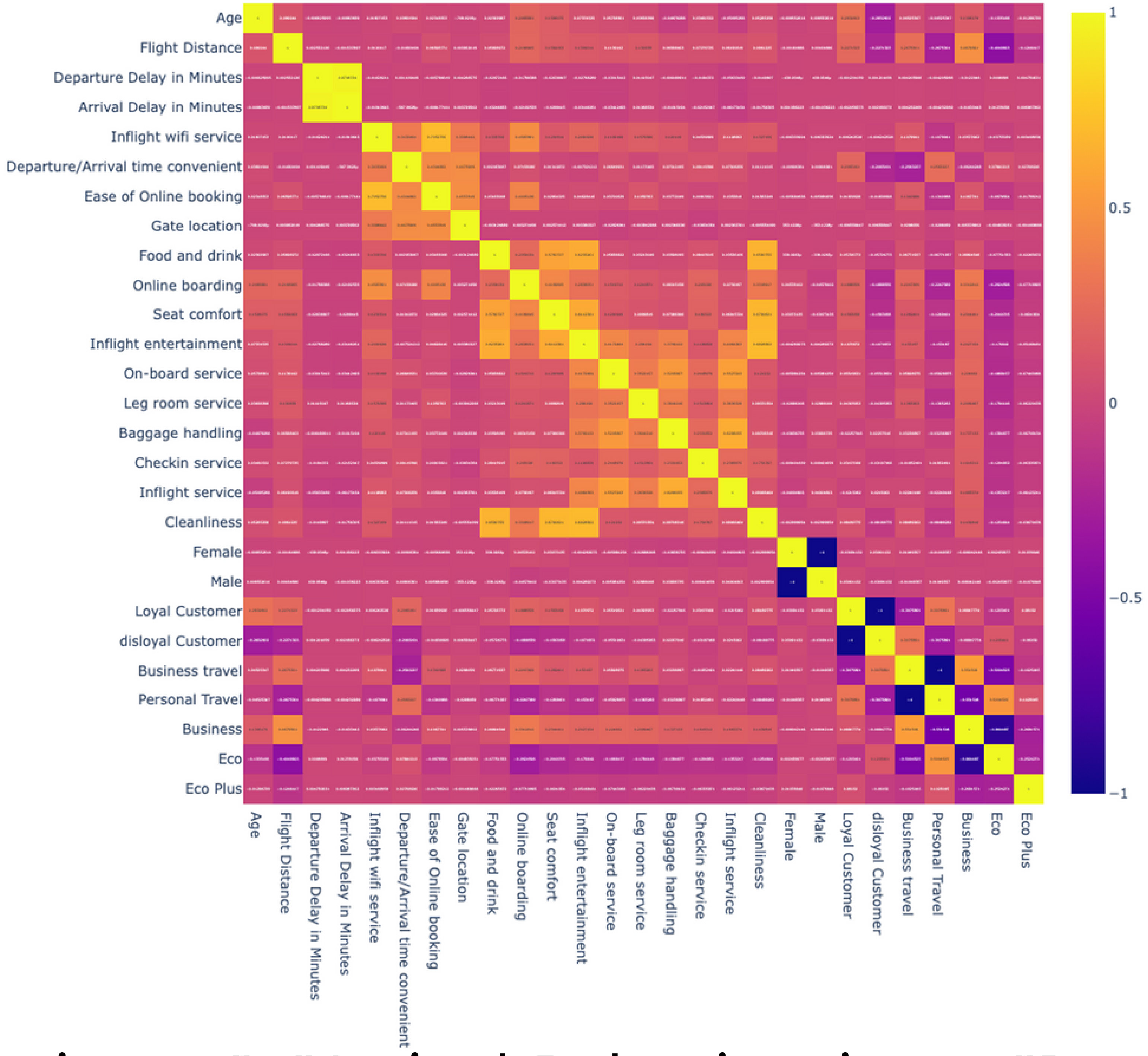


FEATURE ELIMINATION

Pearson Correlation

FEATURE SELECTION

- Chi-Square Test
- Recursive Feature Elimination
- Tree based(Random Forest): SelectFromModel
- Tree based(Light GBM): SelectFromModel



Highly correlated columns: ["Departure Delay in minutes", "Arrival Delay in minutes"]

A high correlation between independent variables can diminish model accuracy because both columns essentially convey redundant information. Therefore, it is advisable to remove one of these correlated columns.

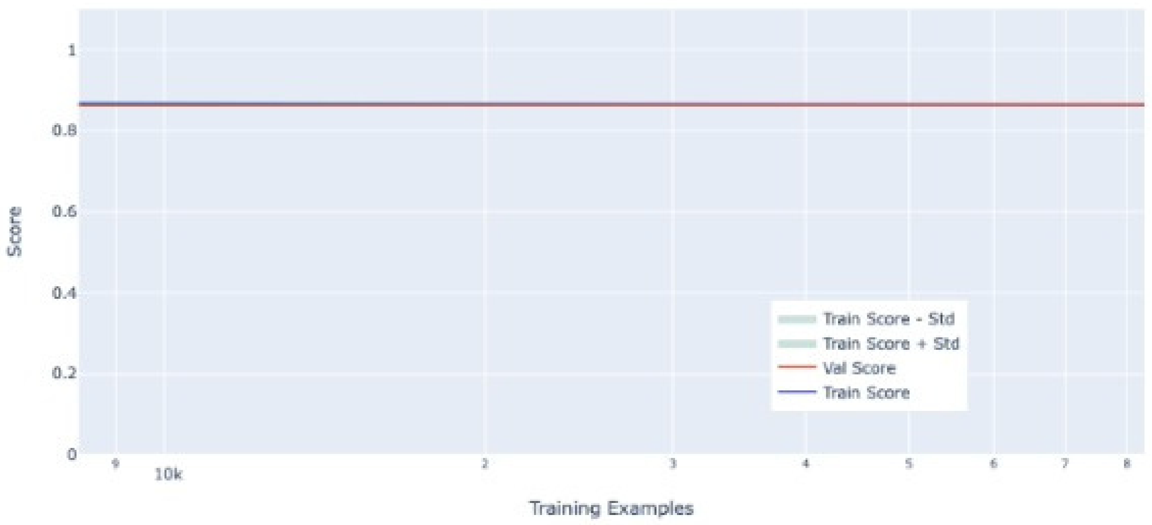
1:01

Describes the actual age of the passenger	Customers Who give repeated Business	WI-FI service available inside the flight
The Process of transporting Passengers bag	Travellers going to abroad other than business purpose	Electronic Boarding Pass that allows travellers to checkin online
Flight distance refers to the distance to travel	a class of seating on an airplane	Seat with extra legroom
It is the area where passengers board to the aircraft	Travellers going to abroad for a business purpose	
Services provided in the flight during the transit	the entertainment available to aircraft passengers during a flight	

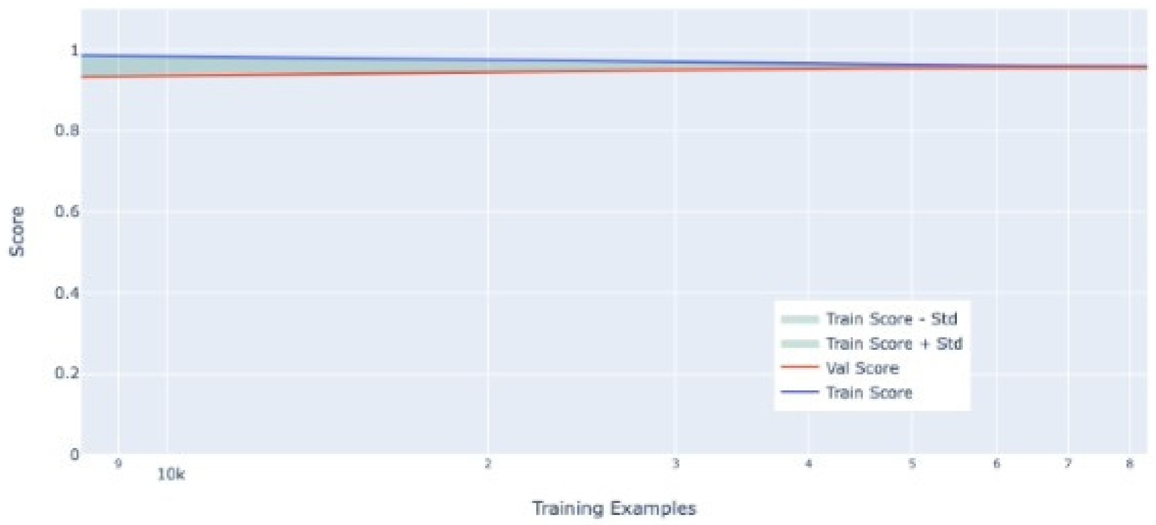
1:01

CHECKING THE FIT OF THE MODEL (OVERFIT/UNDERFIT/JUST RIGHT)

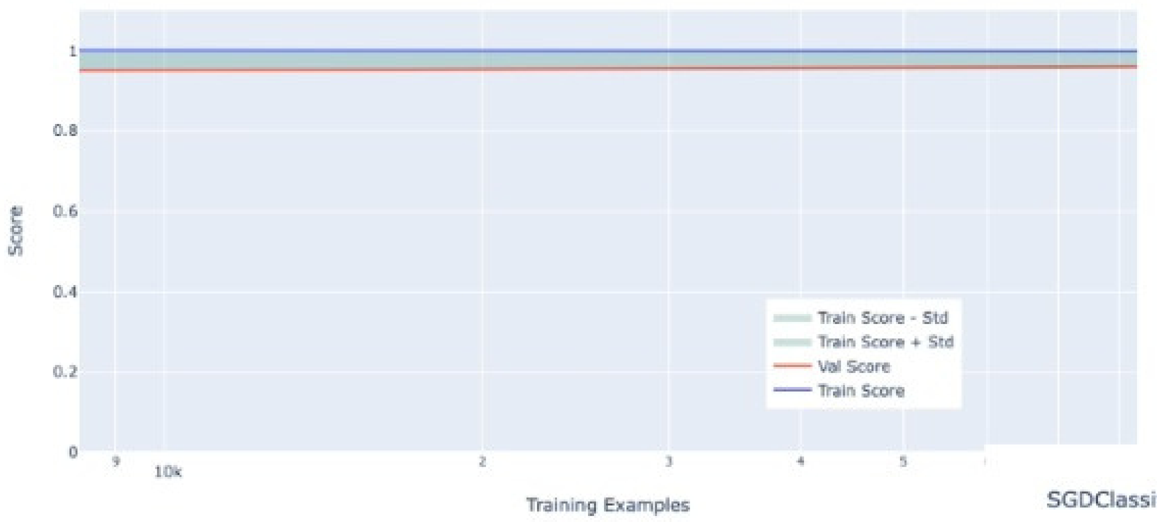
LogisticRegression Learning Curve



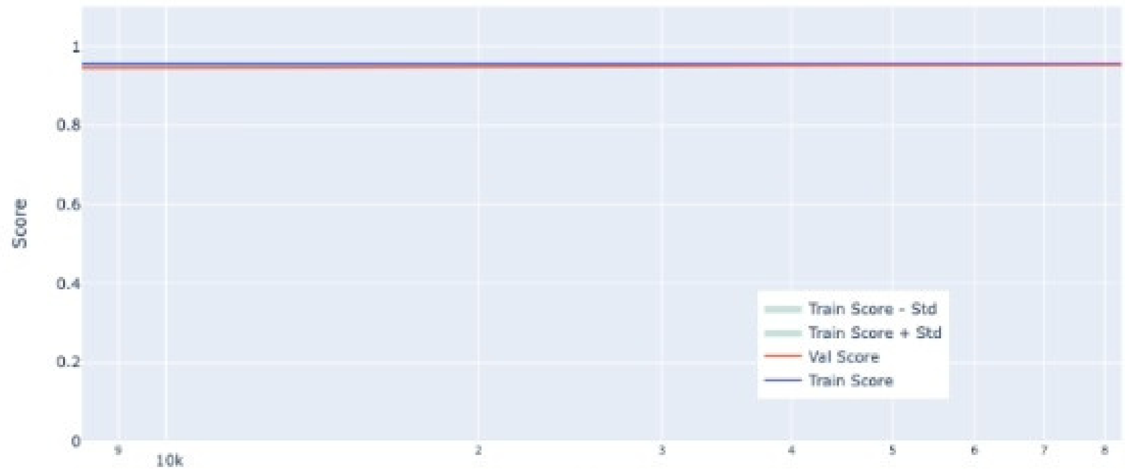
DecisionTreeClassifier Learning Curve



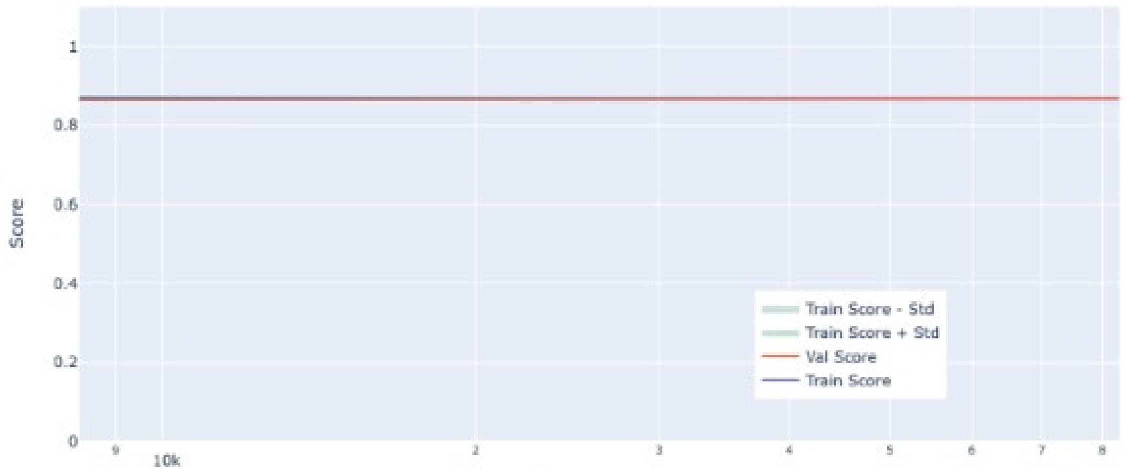
RandomForestClassifier Learning Curve



Support Vector Machine Learning Curve



SGDClassifier Learning Curve



1:01

CONCLUSION

Model	Accuracy	Precision	Recall	F1_score
Parameter Tuned RandomForest Classifier	0.962465	0.972678	0.940014	0.956067
Baseline RandomForest Classifier	0.962388	0.971808	0.940723	0.956013
Parameter Tuned DecisionTree Regression	0.957499	0.969217	0.931774	0.950126
Final RandomForest Classifier	0.957422	0.967230	0.933635	0.950135
Parameter Tuned Support Vector Machine	0.955035	0.959742	0.935761	0.947600
Final Support Vector Machine	0.952302	0.957305	0.931774	0.944367
Baseline DecisionTree Classifier	0.946682	0.936591	0.940989	0.938785
Baseline Support Vector Machine	0.945180	0.946244	0.926458	0.936246
Parameter Tuned Stochastic Gradient Descent	0.876771	0.896129	0.810296	0.851054
Baseline Stochastic Gradient Descent	0.876732	0.880531	0.828726	0.853843
Basline Logistic Regression	0.874115	0.870015	0.835017	0.852157
Parameter Tuned Logistic Regression	0.874076	0.869730	0.835283	0.852158

- Our tuned RandomForest model excels in all four metrics.
- The reduced-feature tuned RandomForest Classifier offers comparable performance with efficiency benefits, making it the top choice for this dataset.

In conclusion, the RandomForest model is the best choice for this dataset due to its high accuracy and absence of overfitting.

1:01