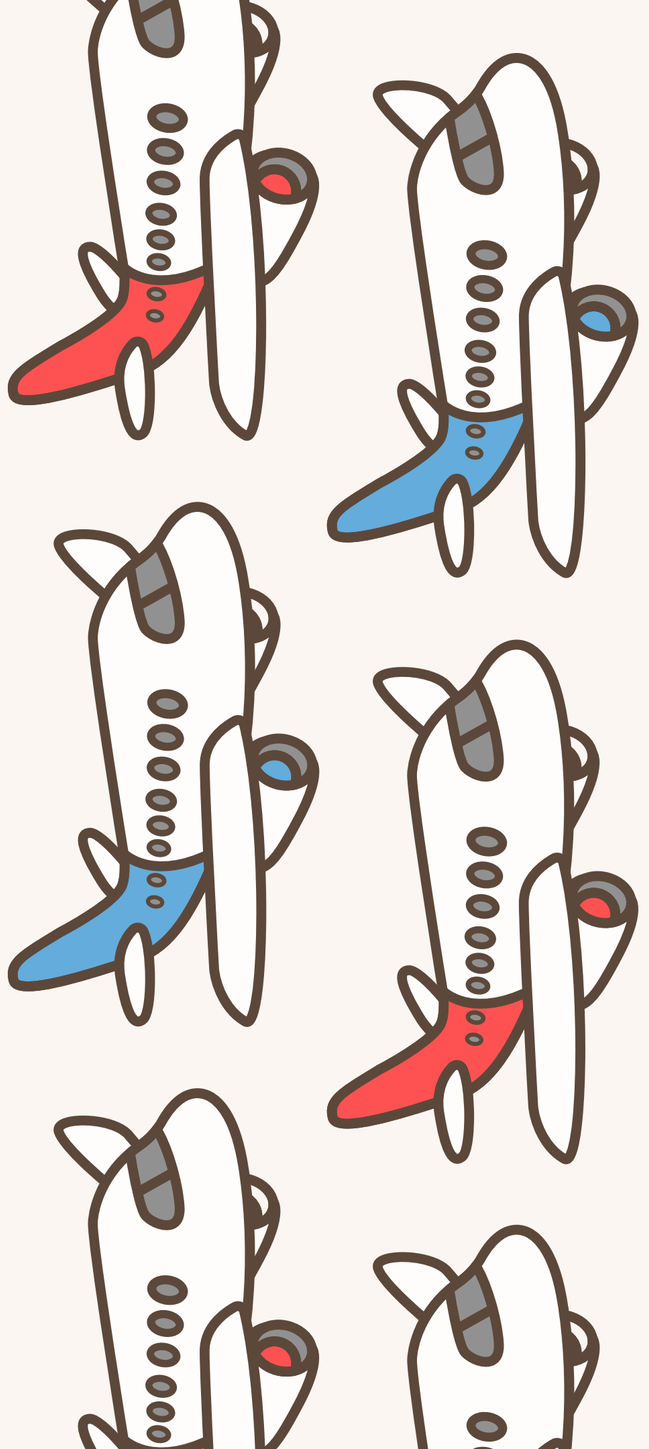
Airline Passenger Satisfaction Analysis

Modeling Satisfaction

Data Splitters



Research Question

Given structured passenger survey input, predict whether a customer is ***“satisfied”*** or ***“neutral/ dissatisfied”***.

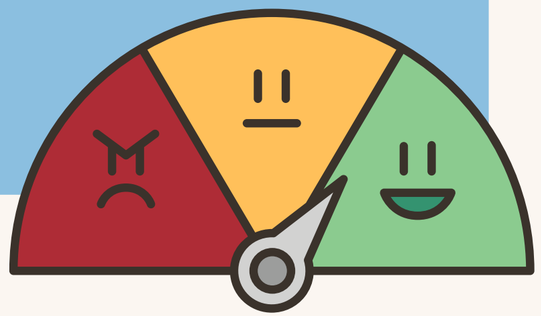
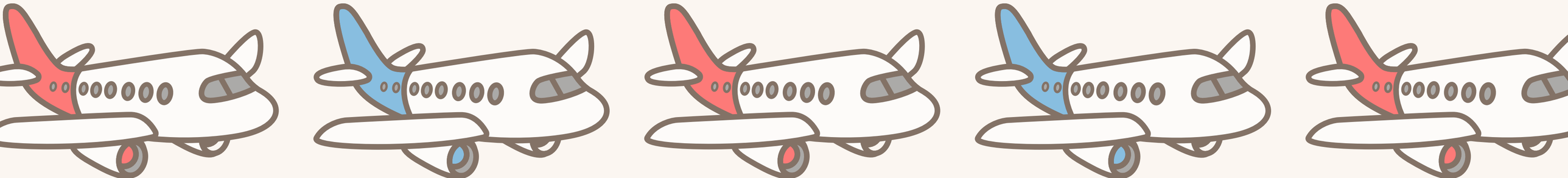
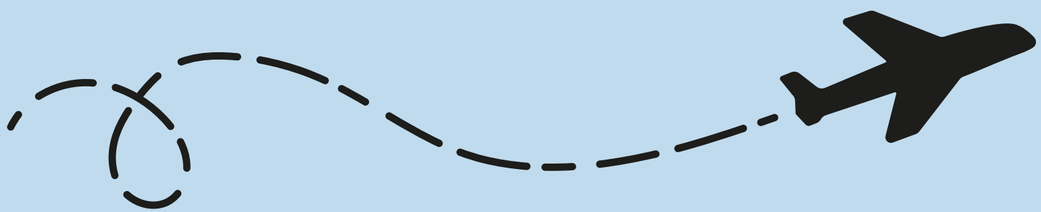


Table of Contents



The Importance of This Topic

Passenger satisfaction is a key driver of an airline to enhance competitiveness. Understanding which factors influence satisfaction is crucial to improve service quality, reduce customer churn, and strengthen loyalty programs.



01

Enhancing Customer Experience

- retain loyal passengers
- intervene and improve service quality

02

Data-Driven Decision Making

- collect large volumes of survey data
- turn raw data into actionable insights

03

Targeted Service Improvement

- identify service factors
- help allocate resources

04

Competitive Advantage in the Airline Industry

- higher loyalty and repeat purchase behavior
- market share and company reputation

Related Work:

Machine Learning Approaches to Airline Passenger Satisfaction

01

Model Benchmarking - Focus on Prediction

- Compared various ML models to identify the most accurate approach
- Proposed hybrid pipelines to improve predictive performance (RF-RFE-LR pipeline proposed by Jiang in 2022)

02

Rise of Interpretability & Causal Analysis

- Emphasized feature importance analysis and identify key variables including ***type of travel, inflight Wi-Fi quality, customer loyalty status, and online boarding efficiency***
- Recent shift toward causal inference, evaluating true causal effects of digital services

Positioning of Our Work

- Extend prior literature by **combining XGBoost + SHAP**
- Achieve both strong predictive performance and high interpretability
- Move beyond accuracy-only focus to **provide actionable insights** for airlines

Background of Dataset

01

Base analysis on the Airline Passenger Satisfaction dataset from ***Kaggle*** (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction)

02

The dataset is provided split into **training set** (*103,904 examples*) and **test set** (*25,976 examples*), contain a **total of 129,880** passenger survey records

03

The dataset has **24 columns** and describe ***passenger's demographic*** attributes, ***flight characteristics***, ***service ratings***, and ***satisfaction outcome***.

04

This dataset provides a high degree of variability, making it well-suited for **supervised learning**, allowing models to learn complex **nonlinear relationships**.

Data Cleaning Steps

To ensure the dataset is reliable for modeling, we perform data cleaning to ***remove noise and irrelevant information***, ***handle missing or inconsistent records***, ***covert raw data into learning-ready formats***.

Clean data leads to more trustworthy insights and stronger predictive performance.

01 Removal of Non-informative Columns

Drop ‘id’ and ‘satisfaction’ columns

02 Handling Missing Values

Impute missing values in “Arrival Delay in Minutes” column using the ***median*** of the training set and test set respectively

03 Encoding Categorical Variables

- **One-Hot Encoding:** nominal features
- **Ordinal Encoding**: ordinal features
- **Label Encoding**: target variable

04 Feature Normalization

Apply **Z-score** normalization using **StandardScaler** to standardize continuous variables

Model Training & Evaluation

01

Logistic Regression

Standard logistic regression model is trained.

Accuracy: 0.871
Precision: 0.868
Recall: 0.834

02

LASSO

Feature selection, reduce overfitting, and select optimal regularization strength

Accuracy: 0.871
Precision: 0.868
Recall: 0.833

03

RFECV

To identify optimal number and best subset of features

Accuracy: 0.870
Precision: 0.866
Recall: 0.833

Results from the LASSO Regression Model

01

Best C: 0.359

02

Optimal number of features: 15

03

Best features:

1. Inflight wifi service
2. Departure/Arrival time convenient
3. Ease of Online booking
4. Online boarding
5. On-board service
6. Leg room service
7. Baggage handling
8. Checkin service
9. Inflight service
10. Cleanliness
11. Departure Delay in Minutes
12. Arrival Delay in Minutes
13. Type of Travel_Personal Travel
14. Customer Type
15. Class

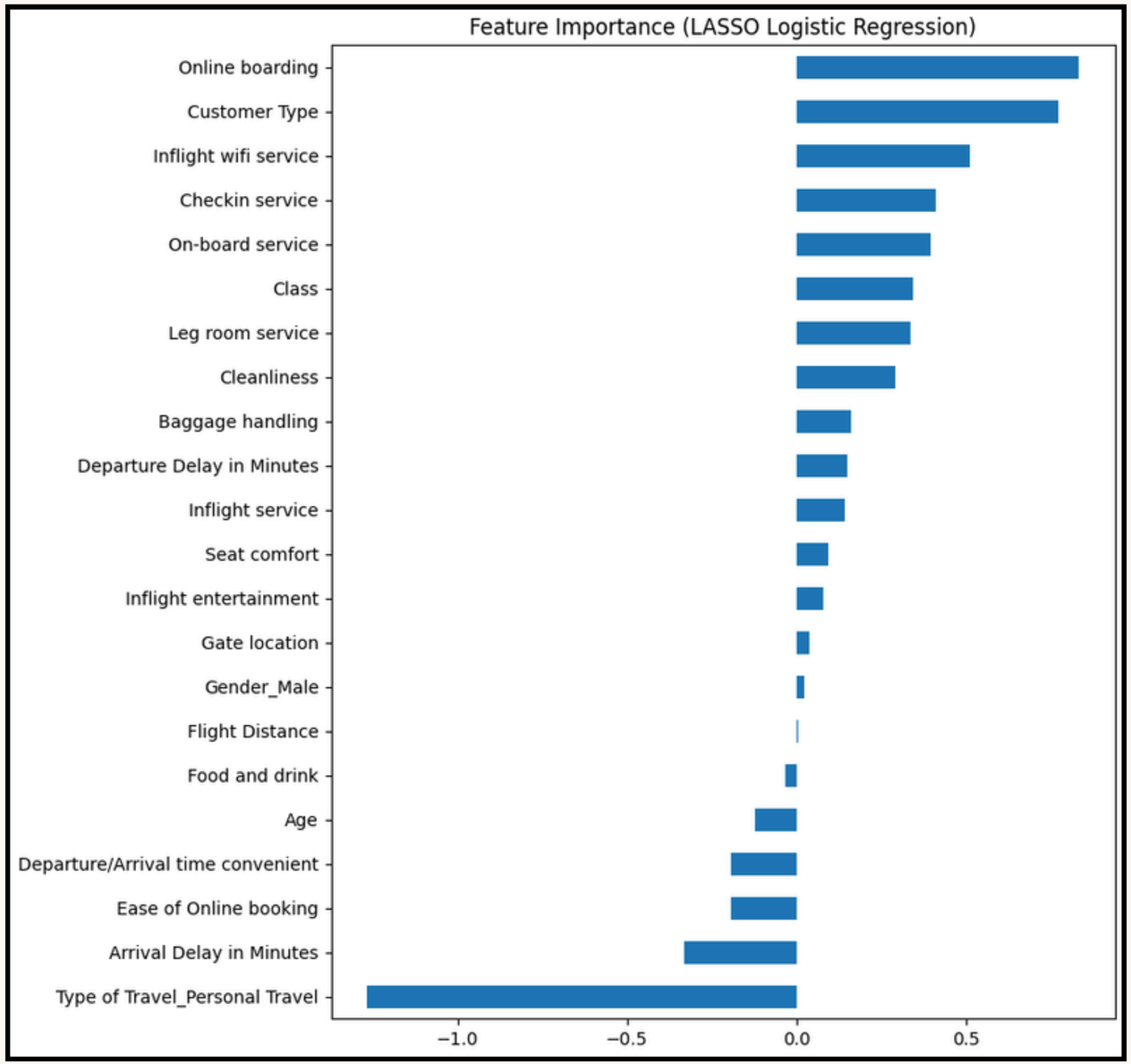
Feature Importance from LASSO

A higher rating for online boarding indicates a higher probability of customer satisfaction

Longer arrival delays indicates a lower probability of customer satisfaction

Assumes a linear model

Partially valid. Explains simple trends



Model Selection

01

Logistic Regression

Standard logistic regression model is trained (L2), since it preserves all features.

Accuracy: 0.871
Precision: 0.868
Recall: 0.834
MSE: 0.098
MAE: 0.202
RMSE: 0.313

02

Random Forest

Random Forest model is trained with 200 trees and a maximum depth of 20.

Accuracy: 0.963
Precision: 0.972
Recall: 0.943
MSE: 0.029
MAE: 0.074
RMSE: 0.171

03

XGBoost

XGBoost model is trained with 200 boosted trees and a maximum depth of 6 for each of the trees.

Accuracy: 0.964
Precision: 0.974
Recall: 0.944
MSE: 0.026
MAE: 0.054
RMSE: 0.161

XGBoost

01

XGBoost is the best model

- Highest accuracy, precision, and recall
- Lowest MSE, MAE, and RMSE

02

Hyperparameters Used:

- Evaluate model performance using logarithmic loss
  - eval_metric = ‘logloss’
- Already encoded data in preprocessing
  - use_label_encoder = False

03

Best Parameters Found:

- Learning Rate = 0.1
- Maximum Depth = 6
- Number of Boosted Trees = 200

04

XGBoost Shortcomings

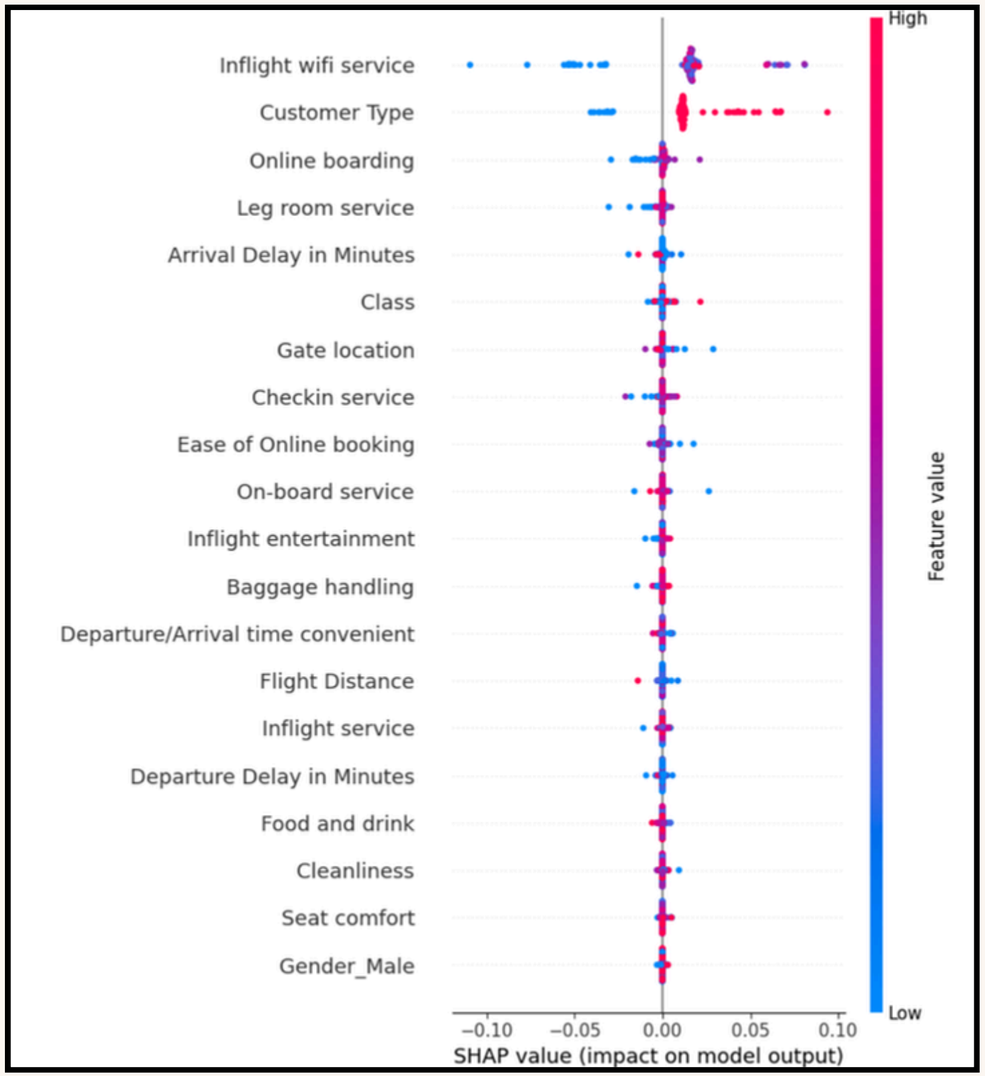
- XGBoost is not an easily interpretable model, but we want to show how our features impact satisfaction
- We used SHAP to overcome this issue

Key Findings From XGBoost + SHAP analysis

Top 5 Features Influencing Passengers Satisfaction

by the magnitude of their impact on model output

1. Inflight wifi service (0.027167)
2. Customer type (0.019843)
3. Online boarding (0.002438)
4. Leg room service (0.001400)
5. Arrival delay in minutes (0.001272)

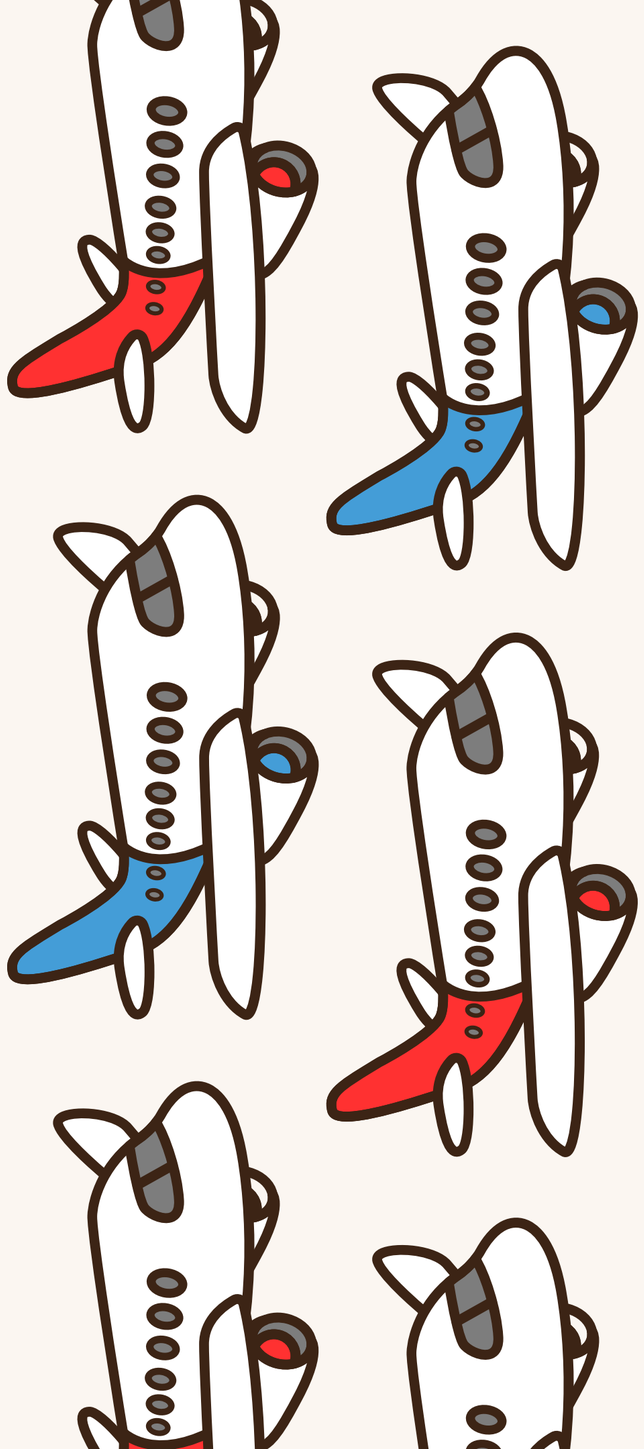


Actionable Recommendation

- Upgrade inflight Wi-Fi and offer tiered service options
- Enhance loyalty programs to improve membership conversion
- Simplify online boarding for a smoother check-in experience
- Improve seat comfort and legroom through better cabin layout
- Strengthen on-time performance and proactive delay communication

Thank you very much!

Data Splitters



References

Hayadi, B. H., Kim, J. M., Hulliyah, K., & Sukmana, H. T. (2021). Predicting airline passenger satisfaction with classification algorithms. International Journal of Informatics and Information Systems, 4(1), 82-94.

Jiang, X., Zhang, Y., Li, Y., & Zhang, B. (2022). Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model. Scientific reports, 12(1), 11174.

Mirthipati, Tejas. "Enhancing airline customer satisfaction: A machine learning and causal analysis approach." arXiv preprint arXiv:2405.09076 (2024).

Mirzahossein, Hamid, and Soheil Rezashoar. "Feature Importance Analysis of Optimized Machine Learning Modeling for Predicting Customers Satisfaction at the United States Airlines." Machine Learning with Applications (2025): 100734.

Nurdina, A., & Puspita, A. B. I. (2023). Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis. Journal of Information System Exploration and Research, 1(2).