# Predicting Customer Satisfaction in E-Commerce Using Machine Learning and Neural Networks

Abhishek Malaviya | Ashfi Ansari | Aaryan Kaushalya | Tabish Khan

# Agenda

1. Introduction
2. Dataset
3. Research Question & Methodology
4. EDA & Preprocessing
5. Modeling & Evaluation
6. Interpretation & Conclusion

# Introduction

E-commerce success relies heavily on customer satisfaction. Our research explores using machine learning and neural networks to predict customer satisfaction in online retail. Accurate prediction helps businesses improve user experience, build loyalty, and make better decisions.

# About the Dataset

Source: Kaggle — Brazilian Online Retail Dataset
Period: 2016 to 2018
Customers: 91,466 unique
Records: 113,194 rows (after merging)
Features: 44 columns (after merging)

1. order_payments_dataset
2. customers_dataset
3. geolocation_dataset
4. order_items_dataset
5. orders_dataset
6. products_dataset
7. sellers_dataset
8. category_name_dataset
9. order_review_dataset

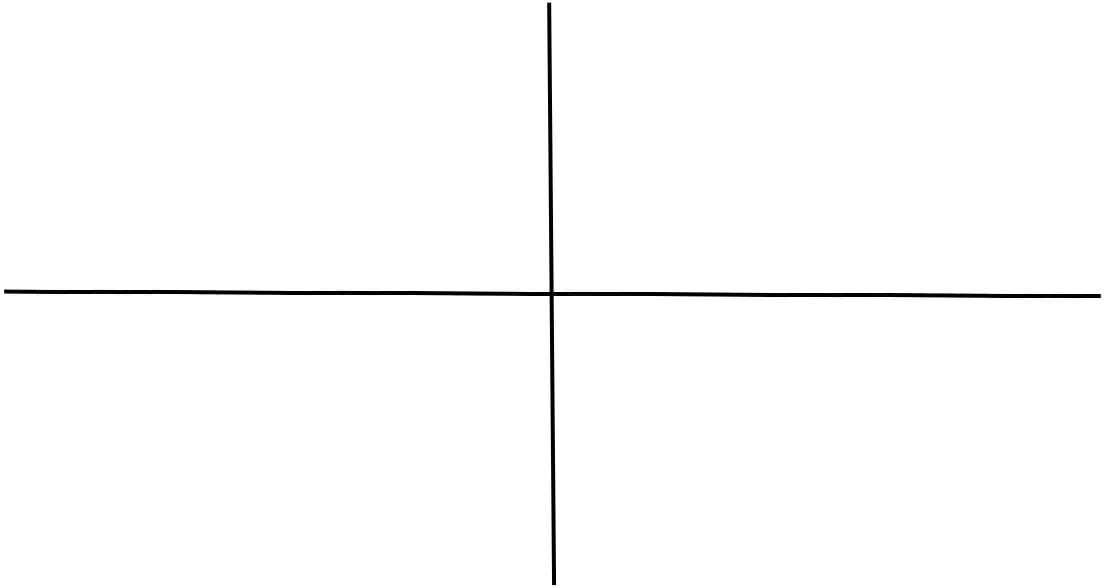
# Research Question

What machine learning methods can best predict customer satisfaction in e-commerce, using various data like delivery details, customer reviews and purchase history?

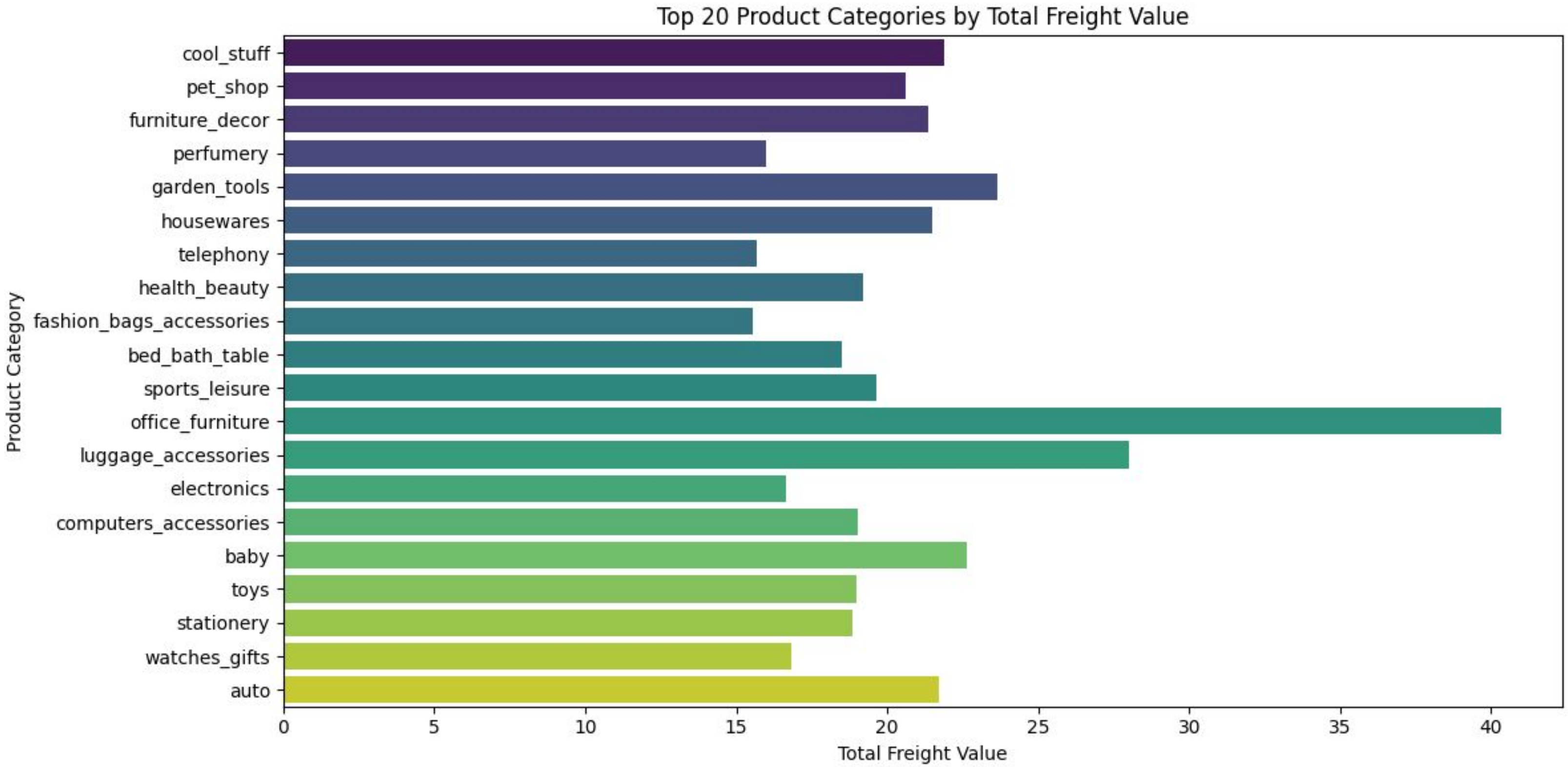
# Methodology

1. Introduction
2. Data Understanding
3. Exploratory Data Analysis
4. Data Preparation
5. Feature Engineering
6. Handling Imbalanced Data
7. Modeling
8. Evaluation Metrics

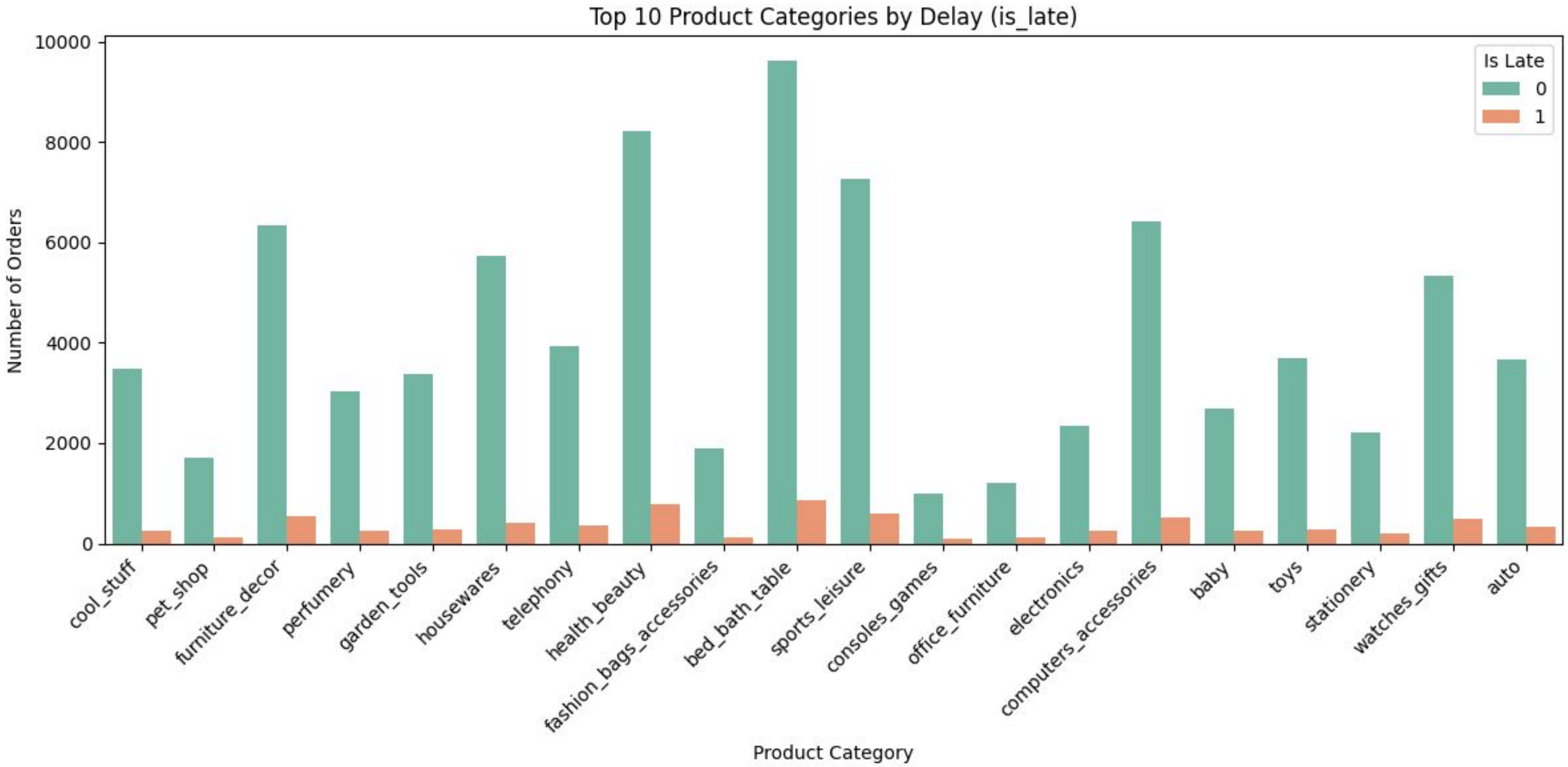
# Literature Review



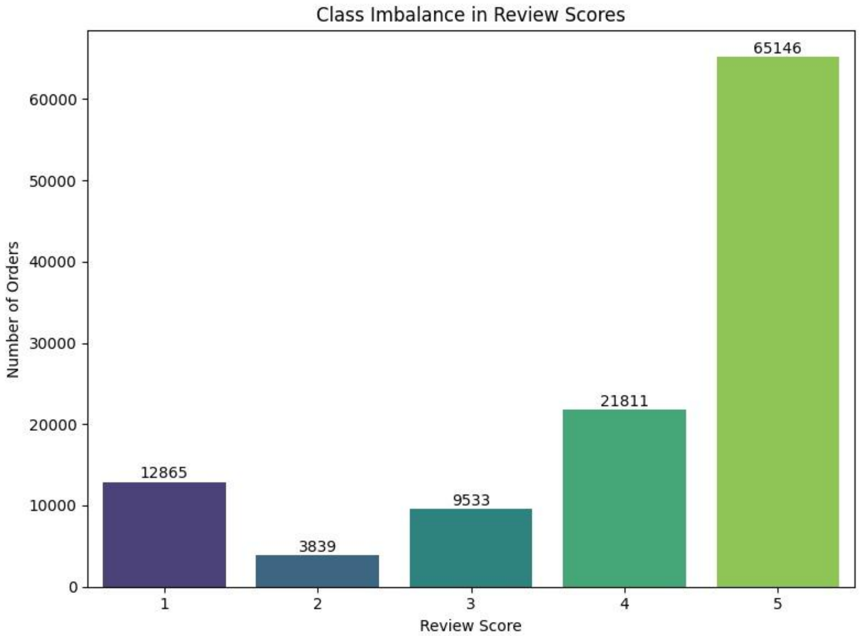
# Exploratory Data Analysis

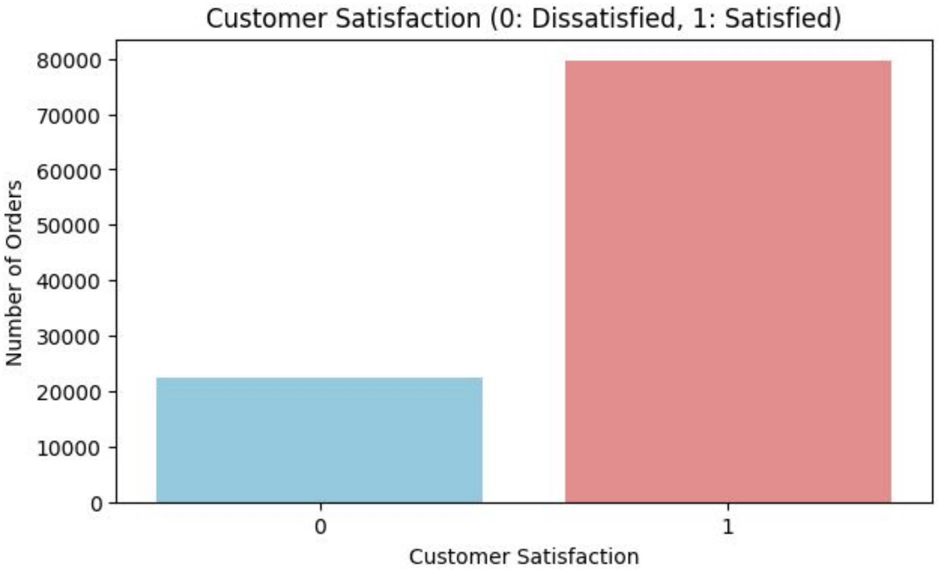


# Exploratory Data Analysis



# Handling Imbalance

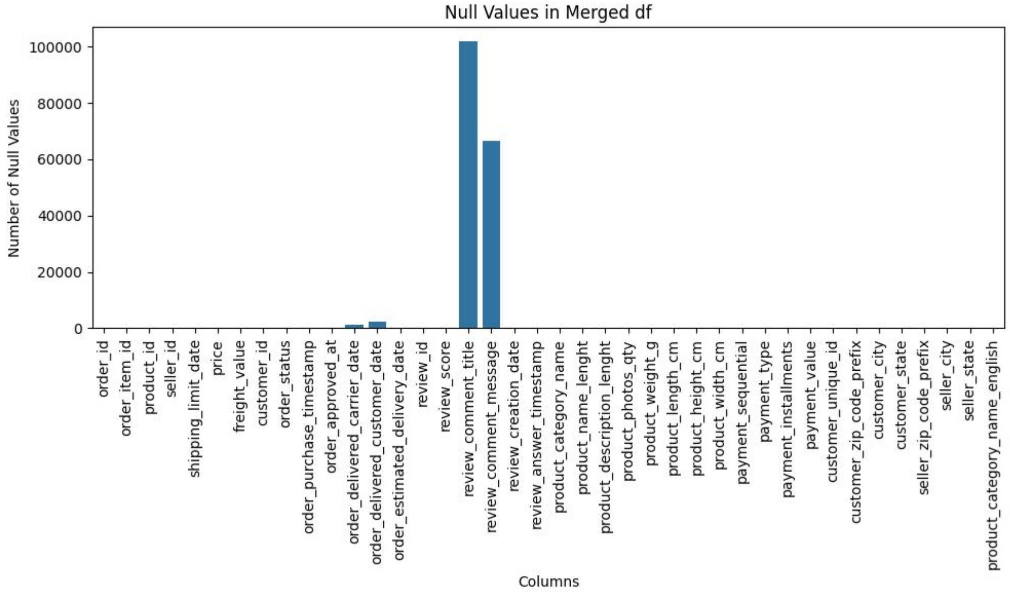



- Applied Class Weighting - Assign higher weight to minority class (Dissatisfied customers)
- Binary Target Transformation - Converted review_score → Satisfied (≥4) vs Dissatisfied (≤3)
- Stratified Train-Test Split - Ensure same satisfied/dissatisfied ratio in train & test sets
- Weighted Metrics - Fair evaluation despite imbalance

# Processing Nulls, Duplicate and Dataset specific Issues

- Dropped rows with missing order_delivered_customer_date (~2%) → critical for delivery analysis.
- Missing product_weight_g, product_length_cm, product_height_cm, product_width_cm → dropped (only 1 row each).
- Missing review_comment_title & review_comment_message → kept as is (textual data, not affecting modeling).



# Feature Engineering for Forecasting

**Target Variable (y):**

Customer Satisfaction (review_score)

- Binary classification:
  - *Satisfied* → Review Score ≥ 4
  - *Dissatisfied* → Review Score ≤ 3

**Feature Variables (X):**

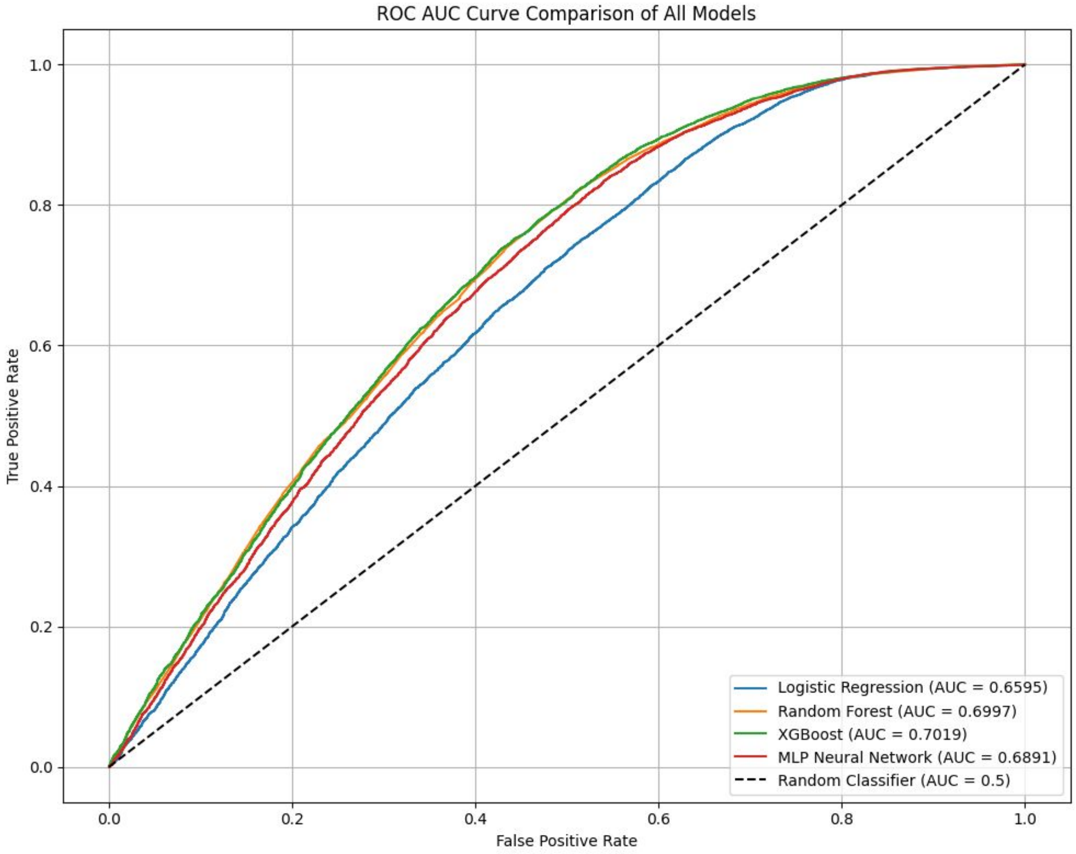
**Payment Value** - Total payment made by customer
**Order Freight Ratio** - Shipping cost relative to product price
**Freight Value** - Shipping cost
**Delivery Days** - Days taken to deliver the product
**Product Category** - Encoded category of the product
**Payment Type** - Mode of payment (encoded)
**Is Late** - Flag indicating late delivery (1 = Late)

**Product Dimensions** - Weight, Height, Length, Width of product
**Product Description Length** - Text length of product description
**Product Photos Quantity** -No. of product images

# Modelling

**Models Trained:**

- **Logistic Regression -** Baseline model, easy to interpret
- **Random Forest -** Robust, interpretable, good with tabular data
- **XGBoost -** High performance, handles class imbalance and complex patterns
- **MLP (Neural Net) -** Captures non-linear relationships

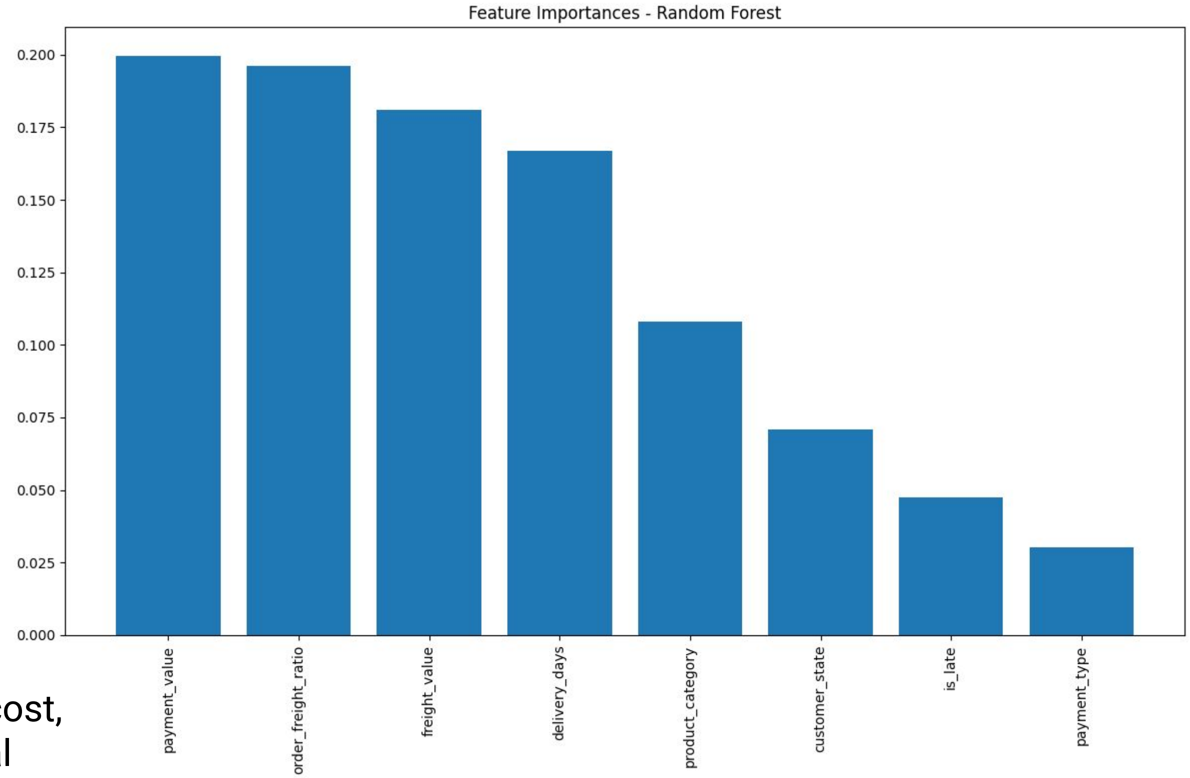


# Results : Model Comparison

| Model | Accuracy | Weighted Precision | Weighted Recall | Weighted F1 | ROC-AUC |
|---|---|---|---|---|---|
| XGBoost Classifier | 72.78% | 75.54% | 75.78% | 73.90% | 70.19% |
| Random Forest Classifier | 80.57% | 78.22% | 80.57% | 77.40% | 69.97% |
| Logistic Regression | 72.37% | 73.39% | 72.37 | 72.85% | 65.95% |
| MLP Neural Network | 73.10% | 75.13% | 73.10% | 73.97% | 68.91 |

# Model Interpretation

**Payment_value** - If payment_value is high and delivery is on time, the customer is likely to be satisfied.
**Order_freight_ratio** - Impact of shipping cost relative to product price
**Freight_value** - Direct influence of shipping charges
**Delivery_days** - Faster deliveries drive higher satisfaction
**Product_category** - Certain product categories affect satisfaction patterns




***Business Takeaway***: Pricing, shipping cost, and delivery speed are the most critical factors affecting customer satisfaction.

# References


**[1]** J. Zhang and Y. Li, ”Review rating prediction framework using deep learning,” in Proc. IEEE Conf. Big Data, 2019.
**[2]** H. Kim and B. Lee, ”Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches,” Journal of Retail Analytics, vol. 12, no. 3, pp. 22–29, 2021.
**[3]** M. A. Khan and S. Ahmad, ”Prediction of Product Rating based on Polarized Reviews using Supervised Machine Learning,” Int. J. Comput. Appl., vol. 175, no. 2, pp. 34–38, 2020.
**[4]** N. Gupta and A. Singhal, ”Sentiment Analysis: Predicting Product Reviews' Ratings using Online Customer Reviews,” in Proc. ICCCNT, 2019.
**[5]** S. Patel and V. Mehta, ”The Analysis and Prediction of Customer Review Rating Using Opinion Mining,” IJERT, vol. 9, no. 8, pp. 90–95, 2020.
**[6]** H. He and E. A. Garcia, ”SMOTE based class-specific extreme learning machine for imbalanced learning,” Neurocomputing, vol. 73, no. 10–12, pp. 185–192, 2010.
**[7]** S. S. Islam and M. S. Haque, ”Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets,” IJCSMC, vol. 7, no. 4, pp. 15–25, 2018.
**[8]** R. Chawla and S. Bhatia, ”Oversampling Techniques for Diabetes Classification: a Comparative Study,” Health Inf. Sci. Syst., vol. 8, no. 12, 2020.
**[9]** J. Branco, L. Torgo, and R. P. Ribeiro, ”Survey of Predictive Modeling on Imbalanced Domains,” ACM Comput. Surv., vol. 49, no. 2, 2017.
**[10]** A. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, ”An approach for classification of highly imbalanced data using weight ing and undersampling,” Knowl.-Based Syst., vol. 25, no. 1, pp. 13–26, 2012.
**[11]** M. Brown and T. Lee, ”Understanding Customer Satisfaction Factors: A Logistic Regression Analysis,” J. Serv. Res., vol. 18, no. 3, pp. 261–273, 2015.
**[12]** S. Liu and Y. Chen, ”XGBoost Method Based on Telecom Customer Sat isfaction Enhancement Strategy,” IEEE Access, vol. 7, pp. 10007–10015, 2019.
**[13]** A. Roy, K. Dutta, and M. Das, ”Comparative analysis of classification algorithm: Random Forest, SPAARC, and MLP for airlines customer satisfaction,” Int. J. Comput. Appl., vol. 182, no. 18, pp. 1–5, 2018.
**[14]** Y. Zhou and M. J. D. Powell, ”Predicting the helpfulness of online reviews using multilayer perceptron neural networks,” Expert Syst. Appl., vol. 124, pp. 291–301, 2019. [15] OpenAI, “Chatgpt,” https://chat.openai.com, 2023, accessed: 20th Febru ary 2025. Used for summarization and content refinement.