UNIVERSITI TEKNOLOGI MALAYSIA

RESEARCH DESIGN AND ANALYSIS IN DATA SCIENCE

CUSTOMER CHURN PREDICTION IN E-COMMERCE INDUSTRY USING RANDOM FOREST ALGORITHM

PRESENTER: SOH JOEN SHIUAN (MCS241028)
SUPERVISOR: DR SHAHIZAN BIN OTHMAN

Video Presentation Link

https://www.youtube.com/watch?v=oHS1FFeaRh8

INTRODUCTION

Transformation of business model from retail sales to online sales leads to customer churn trend.

Why analyze Customer Churn?

- A lot of similar platform exist in the market, making the customer have more than 1 option to go.
- Investigate the key reason that causes the customer to leave the platform

Refinement of Marketing Strategies

Marketing campaigns tailored to user needs would significantly increase the customer interest in continuous support.

Revenue Loss Reduction

Management would have more budget to improve operation flow rather than allocating the budgets to customer retention.

Problem Background

Research faces challenges with incomplete data collection, particularly lacking comprehensive environmental and behavioural data.

Current approaches rely on manual detection that identifies churn after it occurs, limiting prevention opportunities.

Previous studies using call tree methods identified specific customer segments but have limited scope. (Nagaraj P et al. 2023)

Research shows focus on specific business sectors like telecommunications and insurance but lacks broader applicability.

PROBLEM STATEMENT

“ Declining e-commerce customer daily time indicates potential customer churn, resulting in reduced transactions and commission losses across the platform ecosystem.

Management’s current assumption that active users equal purchasing customers is contradicted by evidence, proving that there is no direct correlation between customer activity and actual transaction behaviour. ”

RESEARCH AIM

To develop, compare, and enhance a predictive model that uses classification algorithm, by labelling the customer as churn and not churn, providing actionable insights in improving sales revenue and customer retention rate.

RESEARCH QUESTION

RESEARCH QUESTION 1

What are the steps to preprocess the customer churn dataset?

RESEARCH QUESTION 2

How does the known characteristics affect the customer churn rate?

RESEARCH QUESTION 3

How does the predicted result give insights on customer churn?

RESEARCH OBJECTIVE

RESEARCH OBJECTIVE 1

To preprocess the customer churn prediction data, leading to cleaned data for model training.

RESEARCH OBJECTIVE 2

To identify the key relevant attributes that affects the customer churn rate using Correlation Coefficient Matrix.

RESEARCH OBJECTIVE 3

To develop a machine learning model using Random Forest Algorithm that predicts the potential churning customers, visualizing the results in dashboard.

LITERATURE REVIEW

Customer Churn Prediction in E-Commerce Industry

- Customer tends to leave a platform due to better offers provided by the competitors.
- 2 Types of Churning Customers exist in E-Commerce
- Voluntarily Churn: Customer leave with initiative (Frederick Fagerholm, 2022)
- Involuntarily Churn: External factors force customer to leave (Nayema Taskin, 2023)

LITERATURE REVIEW

Customer Churn Prediction in E-Commerce Industry

- Behavioural Analysis stated that equity instantiation leads to revenge behaviour. (Emilio José Montero Arruda Filho and Alexis de Araújo Barcelos, 2020)
- Financial loss and organization reputation becomes worse.
- Traditional statistical method requires manual intervention, while Machine Learning algorithms does not require human intervention to predict the potential churning customers accurately. (Asad Khattak et. Al, 2023)

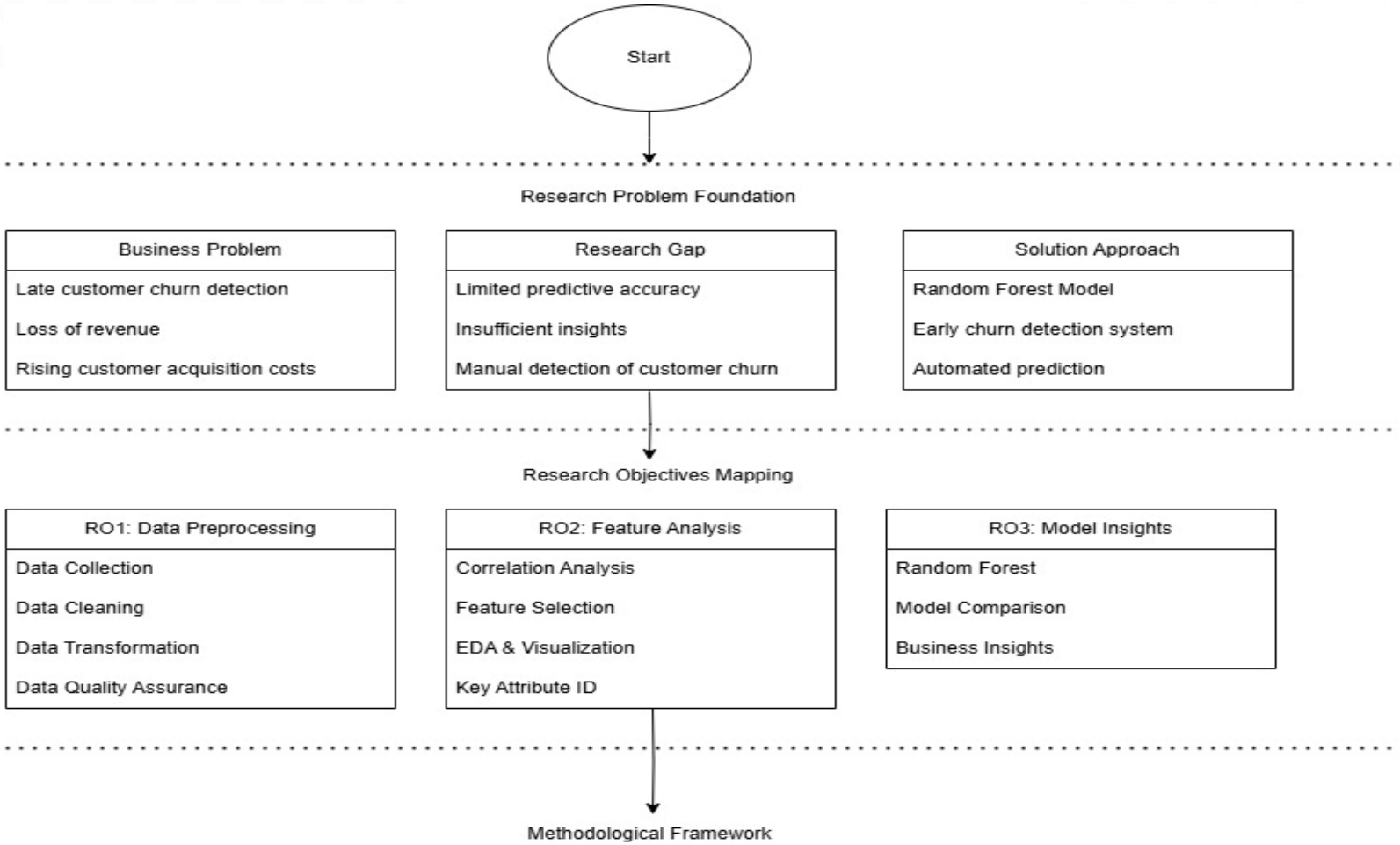
Comparison of dataset in past studies

References	Experiment	Strength	Limitation
Abdulrahman Alshamsi, 2022	Customer Churn in E-Commerce Sector	Include preferred login device and satisfaction metrics Include churn flag for labelling	Imbalanced dataset
Rehka Yadav, 2024	Machine Learning Insights into E-Commerce Churn	Comprehensive customer behaviour and purchase history Large dataset (250,000 entries with 13 columns)	Duplicated features exist (Customer Age and Age columns)
Mukun Chang, 2025	Customer Churn Prediction based on E-Commerce Live Streaming	Revealed non-intuitive insights, such as satisfaction, that is not always correlated to churn	Focus on live-streaming e-commerce segment only

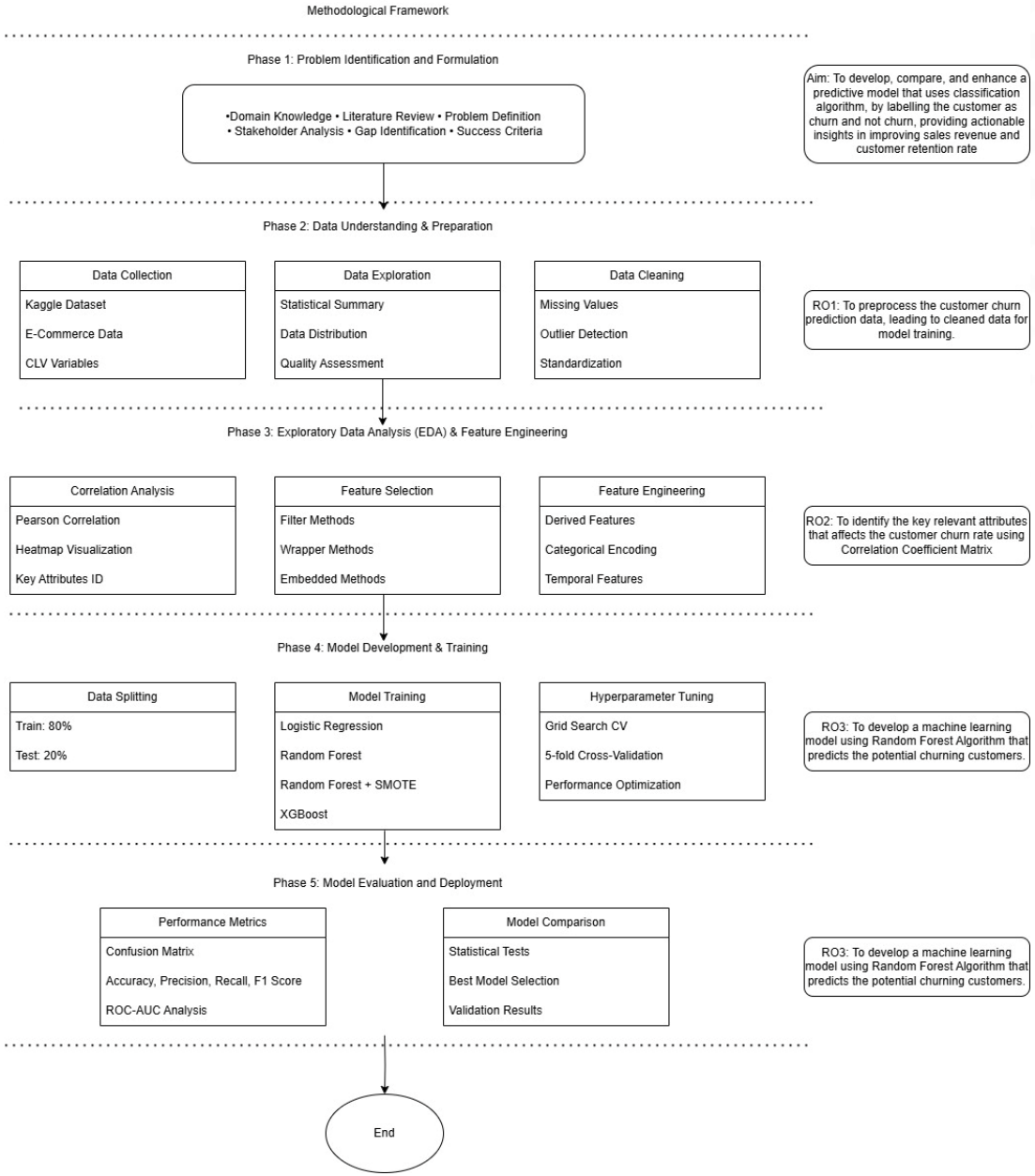
Comparison of ML Algorithms in past studies

Model	Strengths	Limitations	Citation
Logistic Regression	Effective for binary classification problems Provides probability estimates Less prone to overfitting with regularization	Limited to linear decision boundaries Require more data for stable estimation	Ahmed (2024)
Random Forest	Reduce overfitting via ensemble averaging	Increasing trees cause longer prediction time More computational resources required Biased to dominant class if not tuned properly	Swetha P, Dayananda R B (2020)
Random Forest + SMOTE	Improve imbalanced dataset performance Achieve higher accuracy and recall for churn class	Artificial Data Noise More computational resources than pure Random Forest	Hafiz Ma'ruf (2021)
XGBoost	Combine bagging and boosting for higher accuracy output	Extra parameter tuning required Model becomes more complex, increasing maintenance difficulty	Yashkumar Burnwal and Dr R.C.Jaiswal (2023)

RESEARCH METHODOLOGY



RESEARCH METHODOLOGY

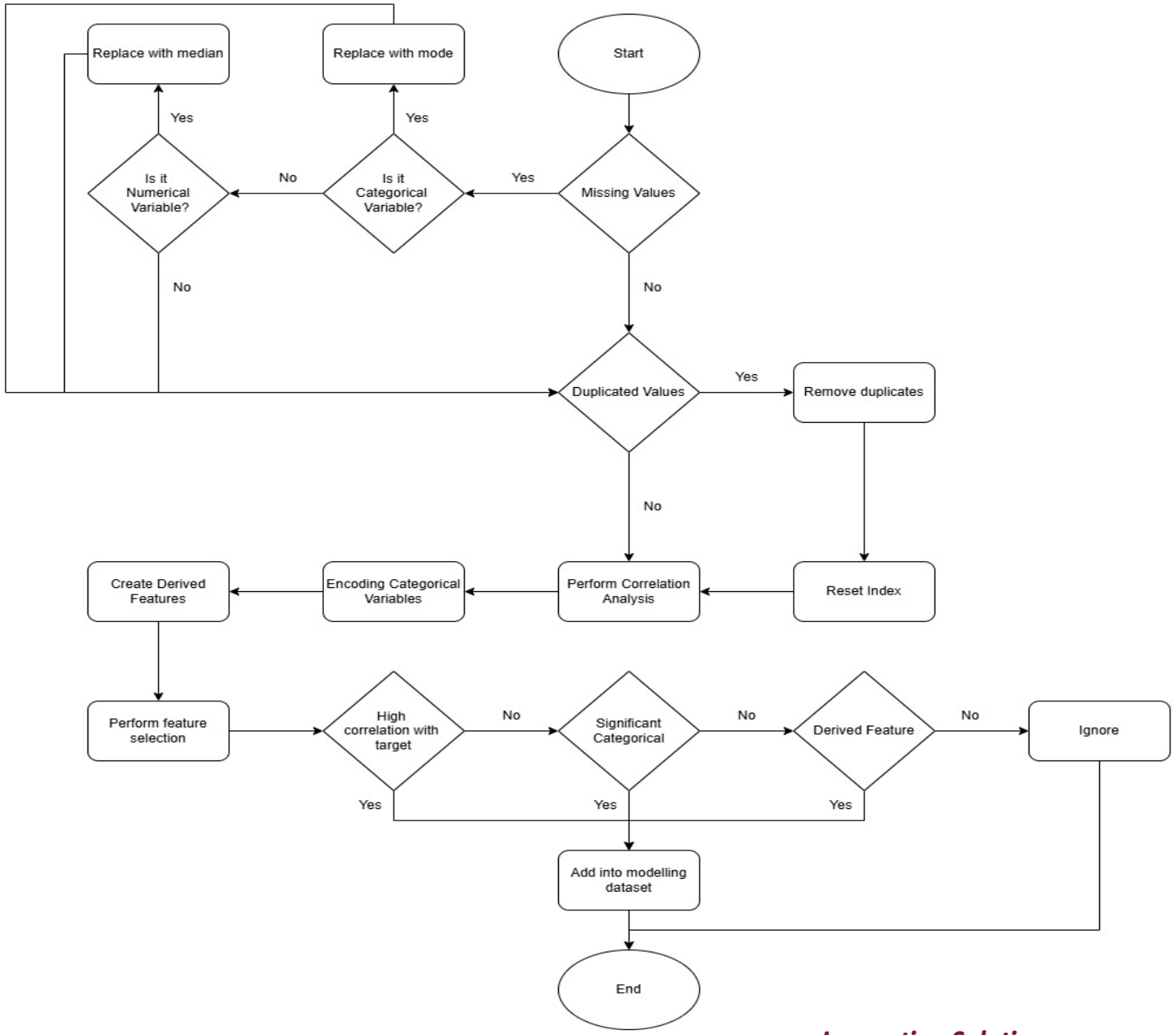


Dataset

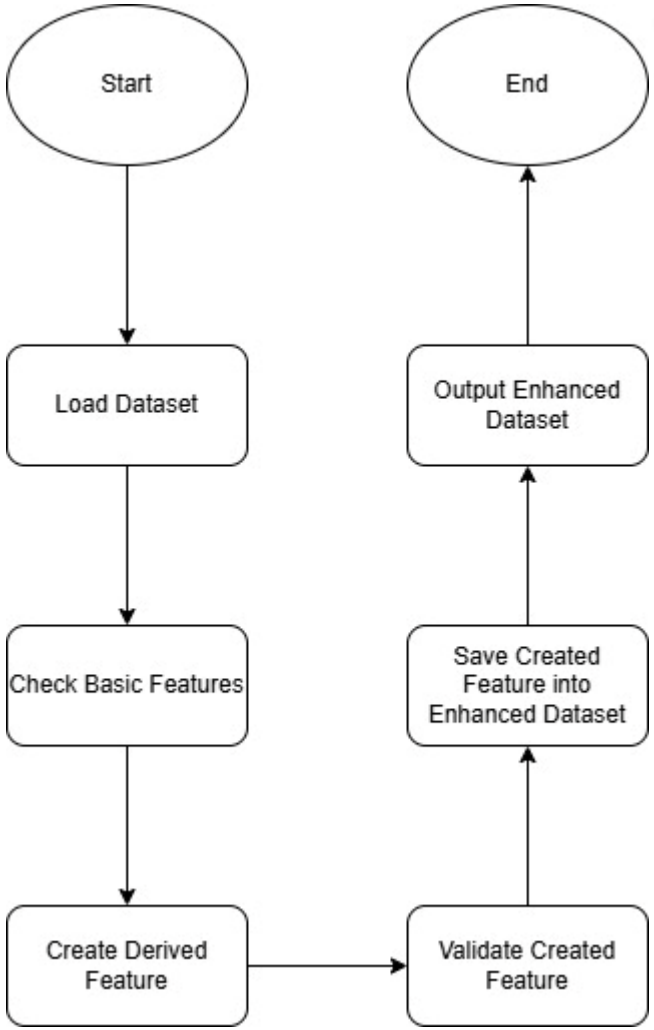
CustomerID	Churn	Tenure	PreferredLoginDevice	CityTier	WarehouseToHome	PreferredPaymentMode	Gender	HourSpendOnApp	NumberOfDeviceRegistered
50001	1	4	Mobile Phone	3	6	Debit Card	Female	3	3
50002	1		Phone	1	8	UPI	Male	3	4
50003	1		Phone	1	30	Debit Card	Male	2	4
50004	1	0	Phone	3	15	Debit Card	Male	2	4
50005	1	0	Phone	1	12	CC	Male		3
50006	1	0	Computer	1	22	Debit Card	Female	3	5
50007	1		Phone	3	11	Cash on Delivery	Male	2	3
50008	1		Phone	1	6	CC	Male	3	3
50009	1	13	Phone	3	9	E wallet	Male		4

Dataset

PreferedOrderCat	SatisfactionScore	MaritalStatus	NumberOfAddress	Complain	OrderAmountHikeFromlastYear	CouponUsed	OrderCount	DaySinceLastOrder	CashbackAmount
Laptop & Accessory	2	Single	9	1	11	1	1	5	160
Mobile	3	Single	7	1	15	0	1	0	121
Mobile	3	Single	6	1	14	0	1	3	120
Laptop & Accessory	5	Single	8	0	23	0	1	3	134
Mobile	5	Single	3	0	11	1	1	3	130
Mobile Phone	5	Single	2	1	22	4	6	7	139
Laptop & Accessory	2	Divorced	4	0	14	0	1	0	121
Mobile	2	Divorced	3	1	16	2	2	0	123
Mobile	3	Divorced	2	1	14	0	1	2	127



DATA PREPARATION FLOW



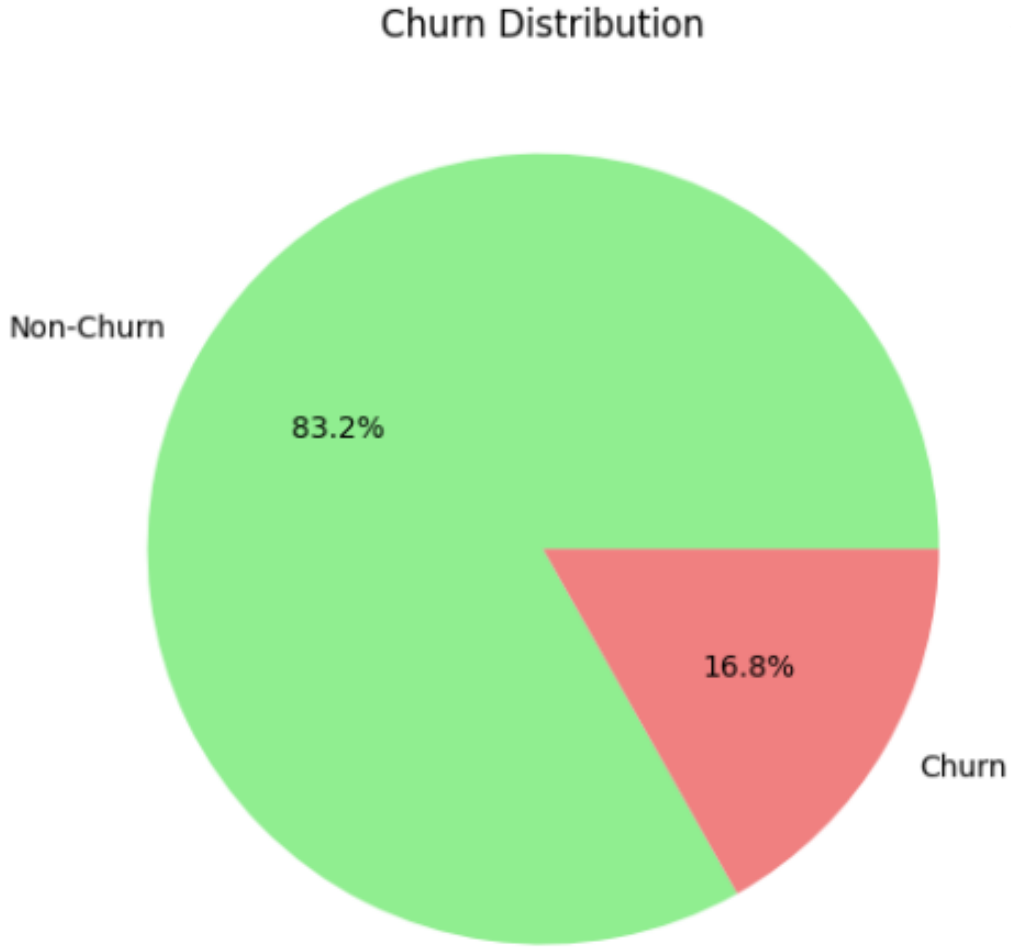
DATA DERIVATION FLOW

Metric	Count
Original Features	23
Encoded Categorical Features	5
New Derived Features	6
Final Feature Count	33
Selected Important Features	16

DATA DERIVATION SUMMARY

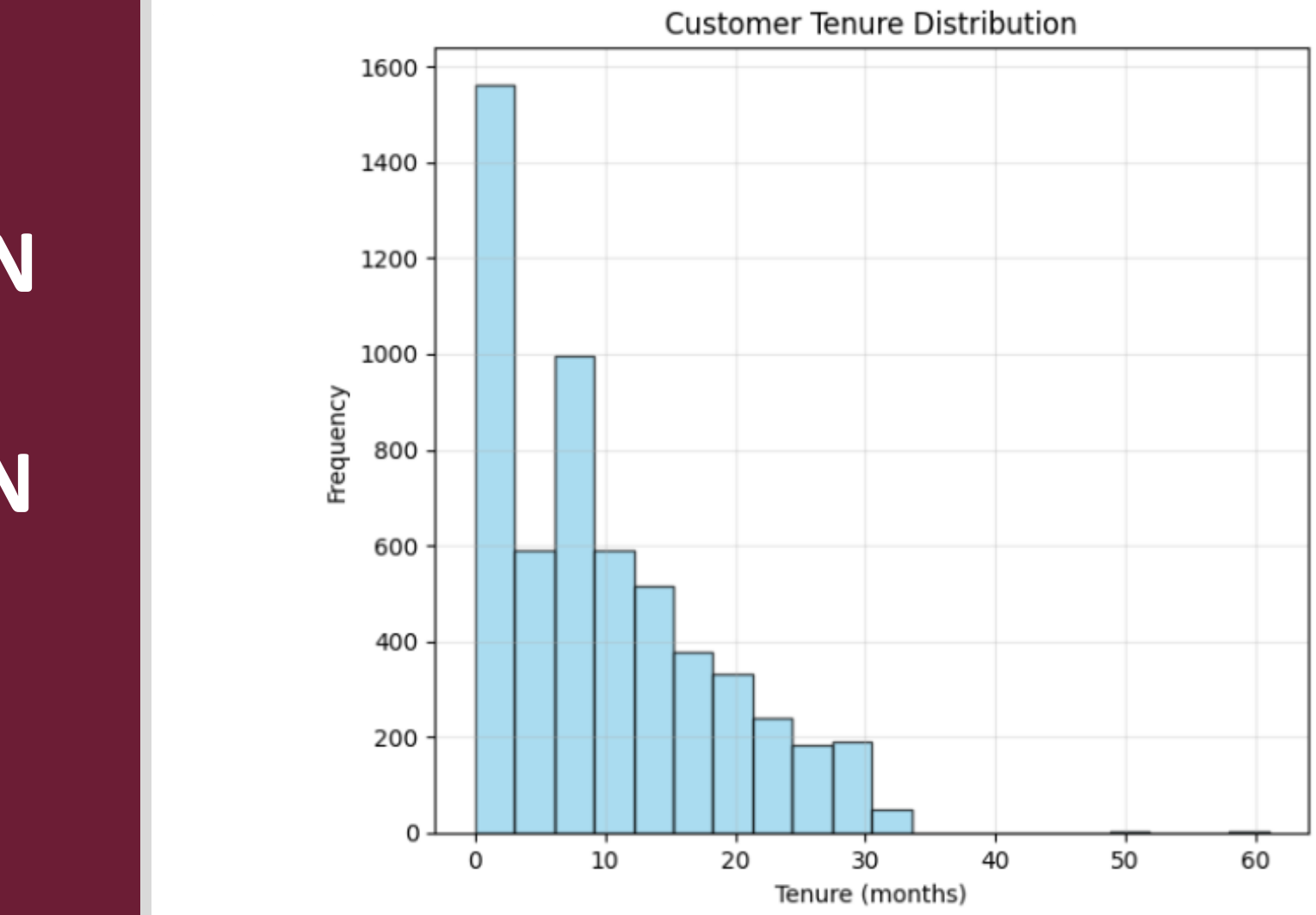
RESEARCH DESIGN AND IMPLEMENTATION

UNIVARIATE ANALYSIS



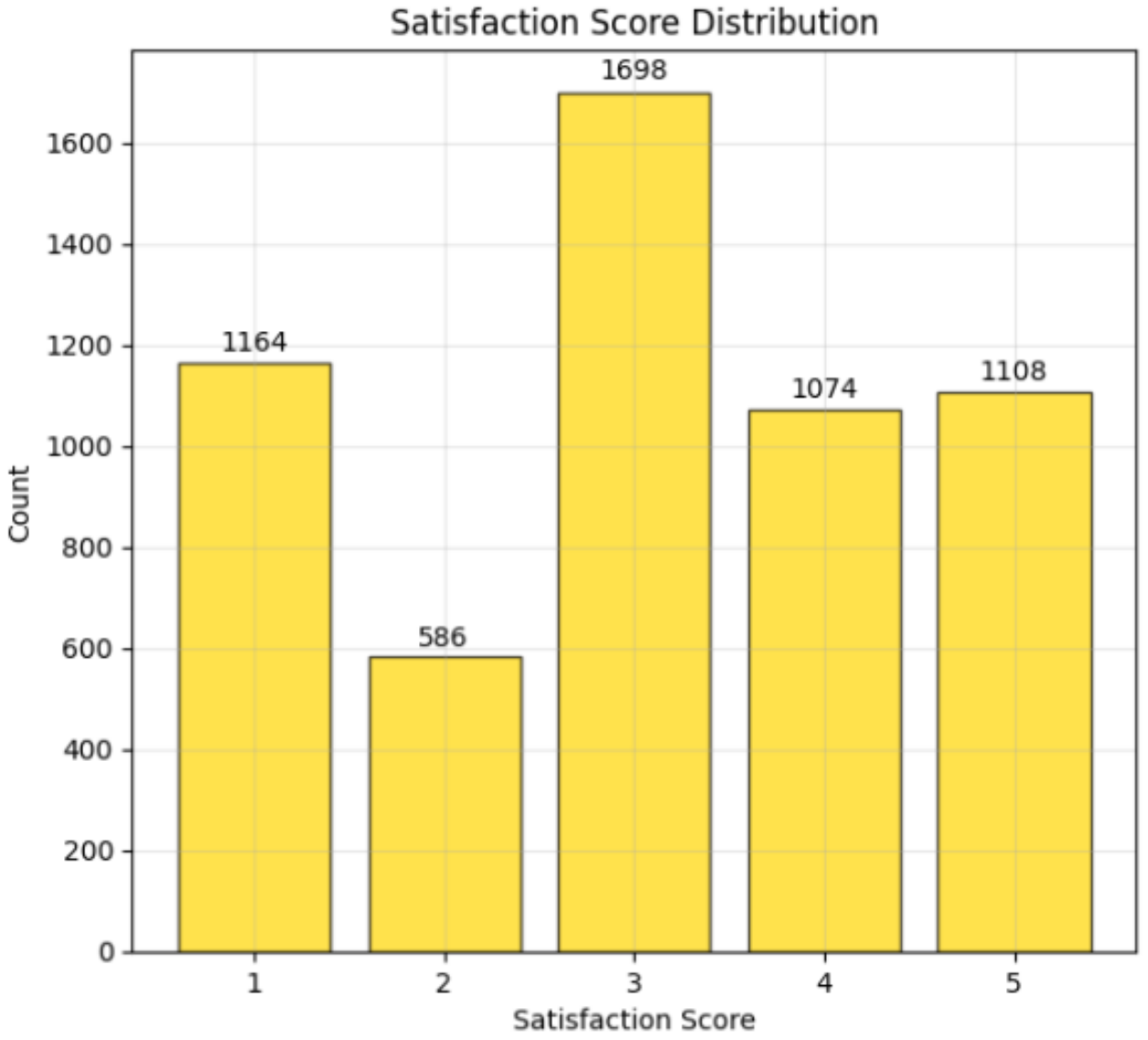
RESEARCH DESIGN AND IMPLEMENTATION

UNIVARIATE ANALYSIS



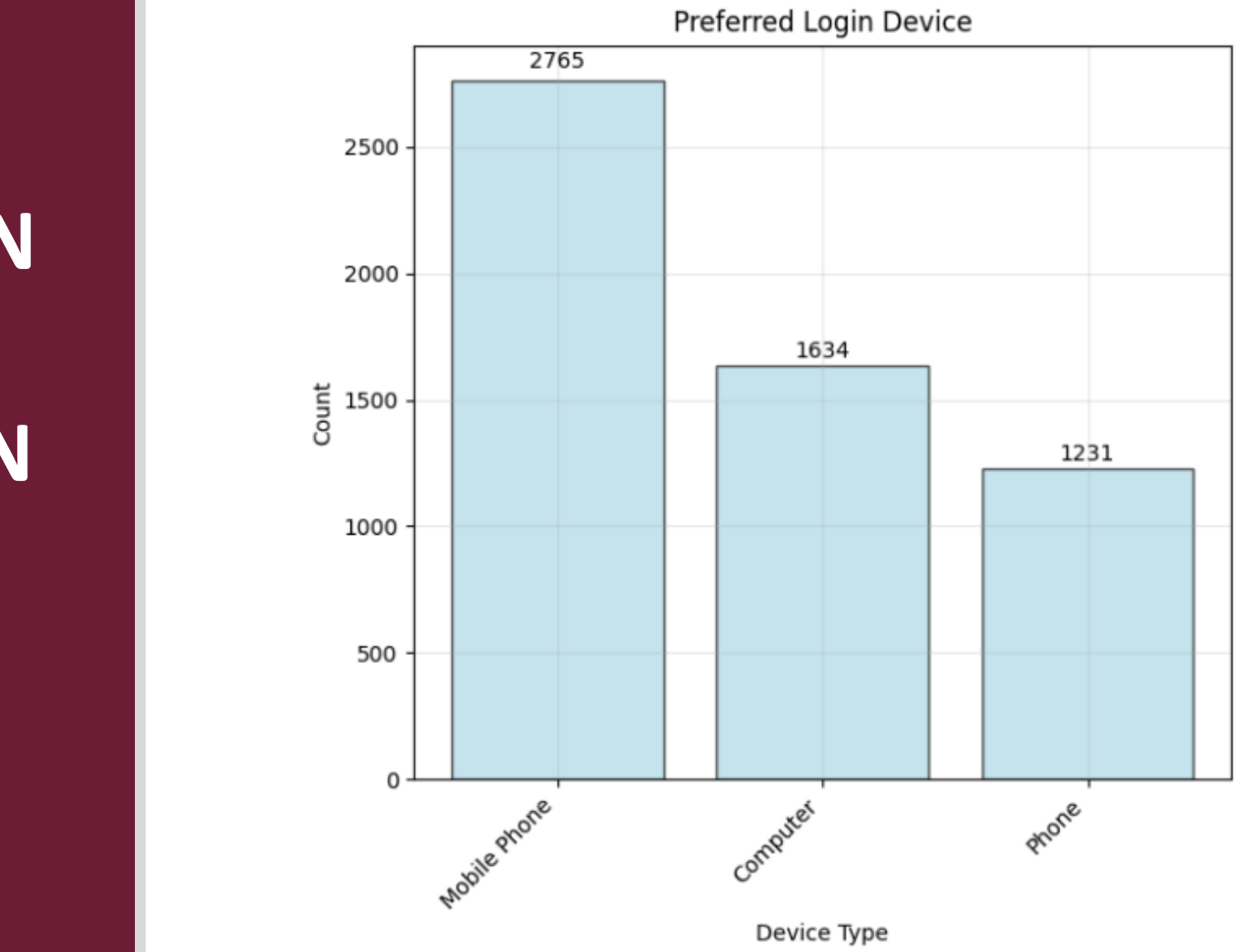
RESEARCH DESIGN AND IMPLEMENTATION

UNIVARIATE ANALYSIS



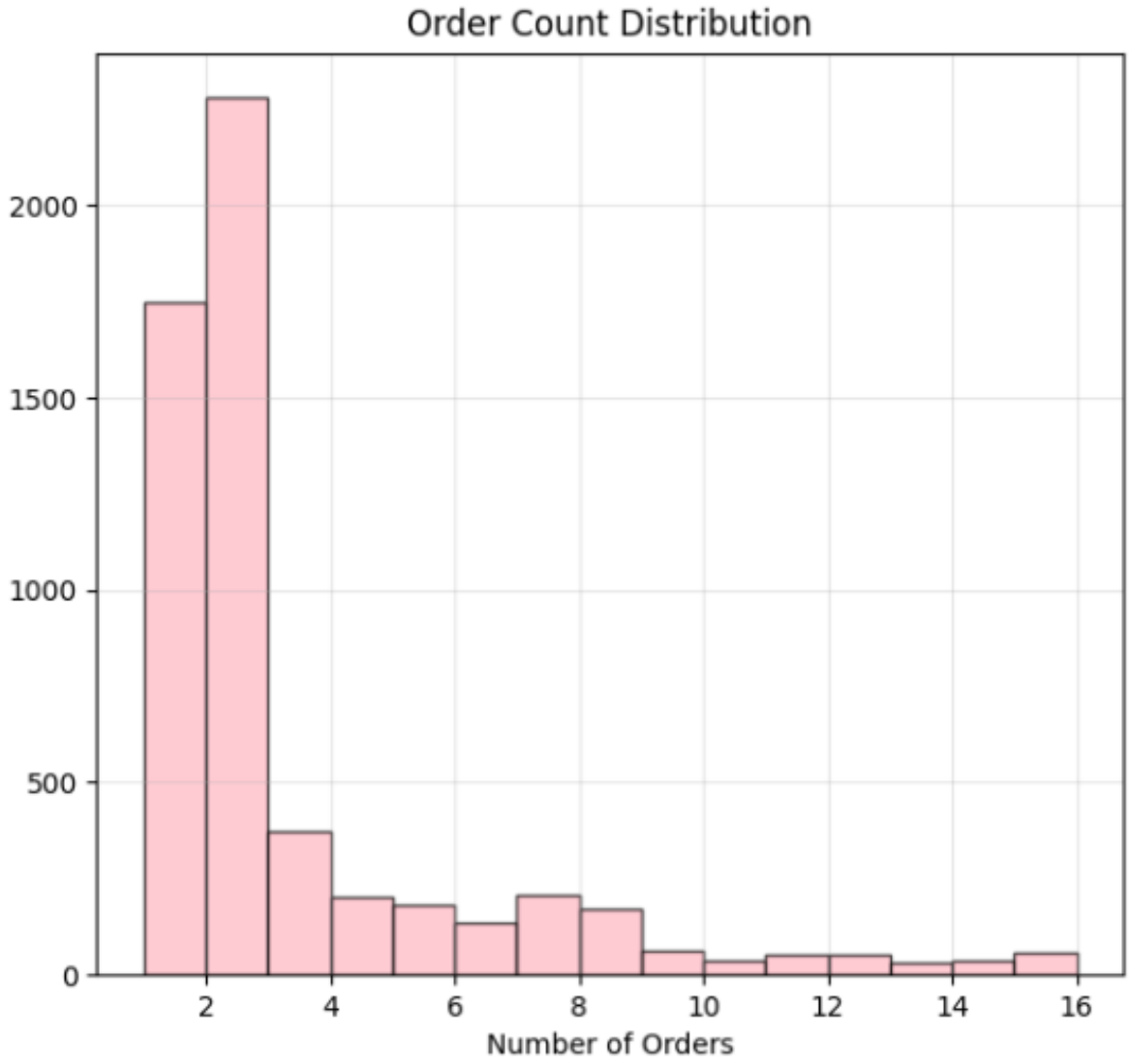
RESEARCH DESIGN AND IMPLEMENTATION

UNIVARIATE ANALYSIS



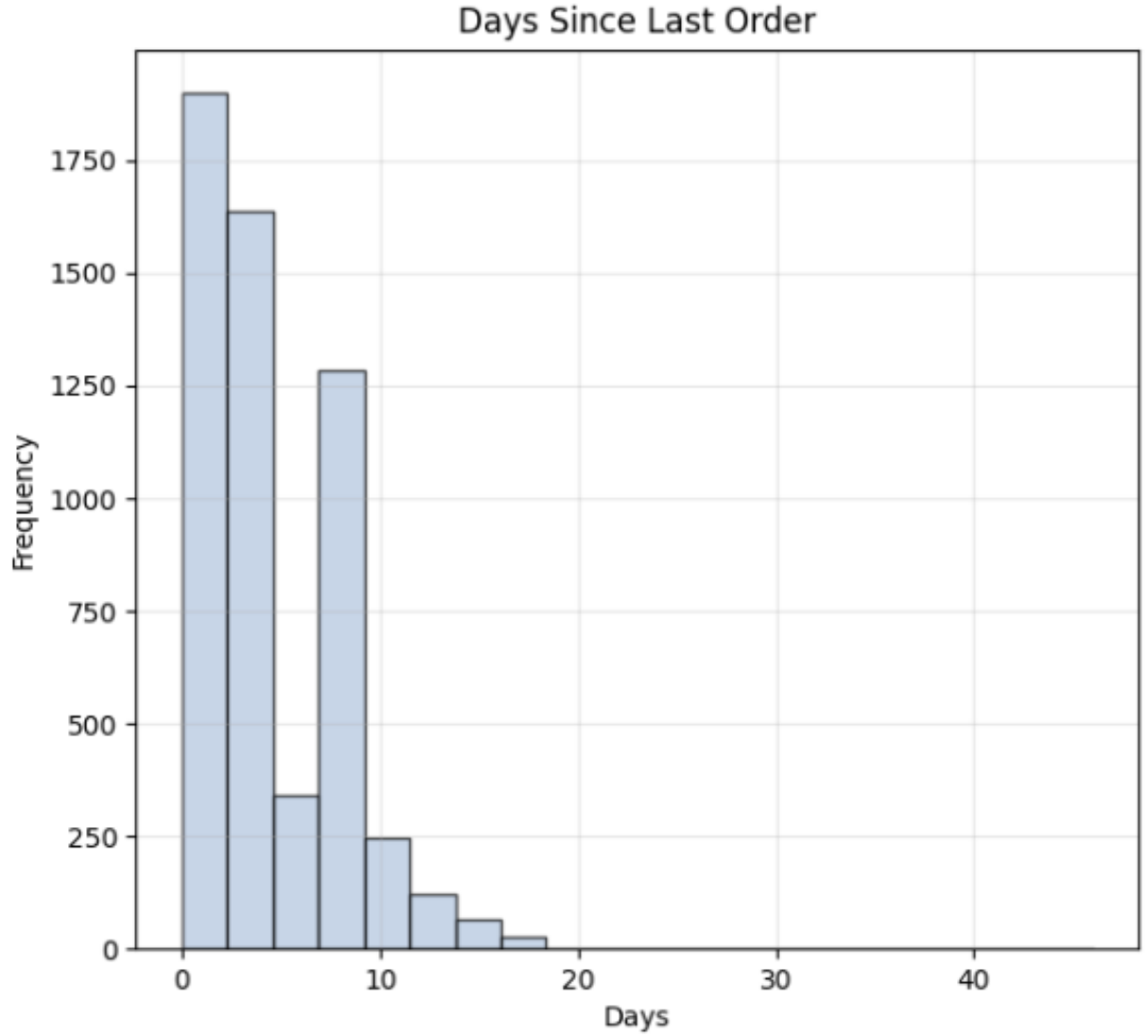
RESEARCH DESIGN AND IMPLEMENTATION

UNIVARIATE ANALYSIS



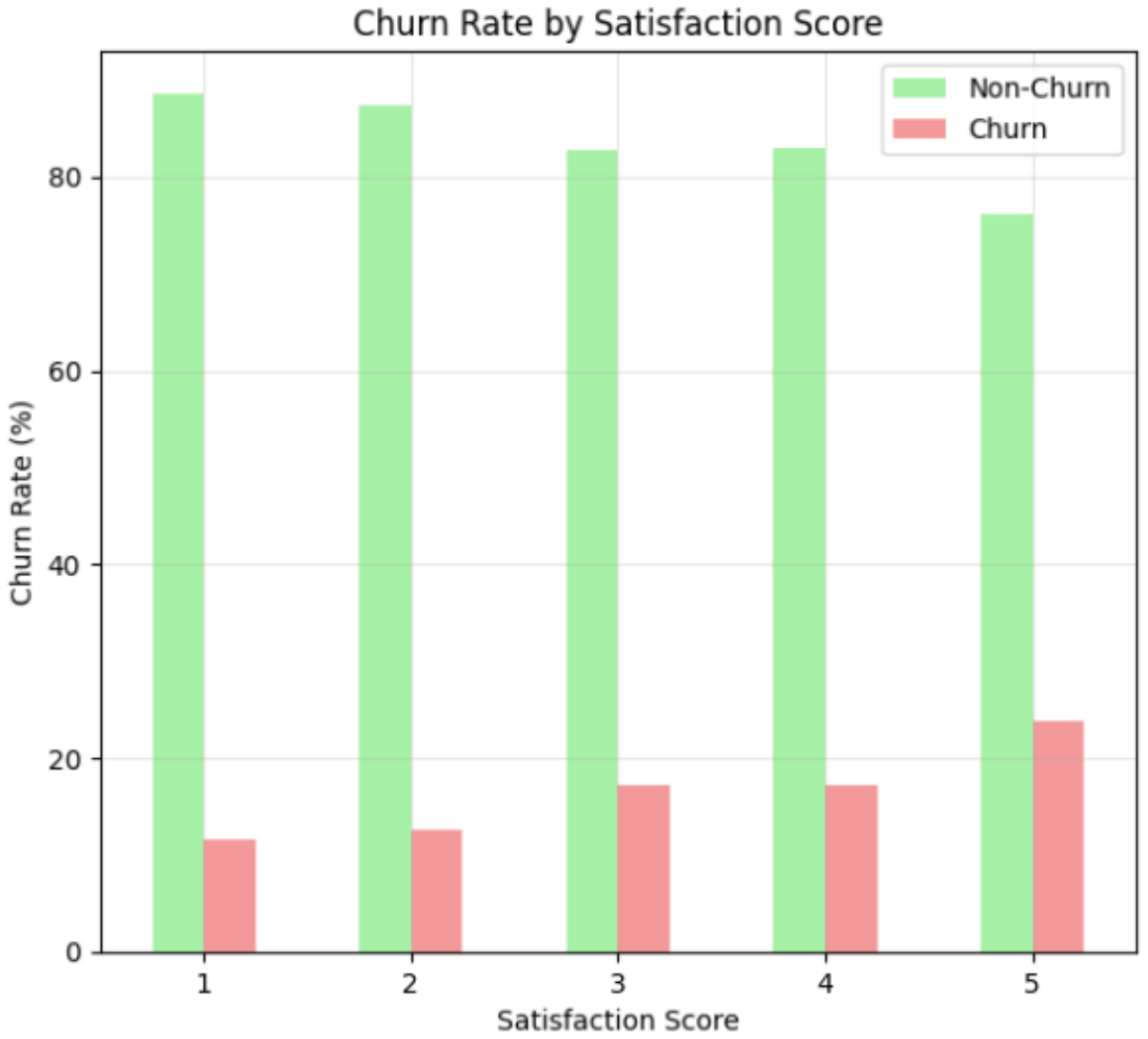
RESEARCH DESIGN AND IMPLEMENTATION

UNIVARIATE ANALYSIS



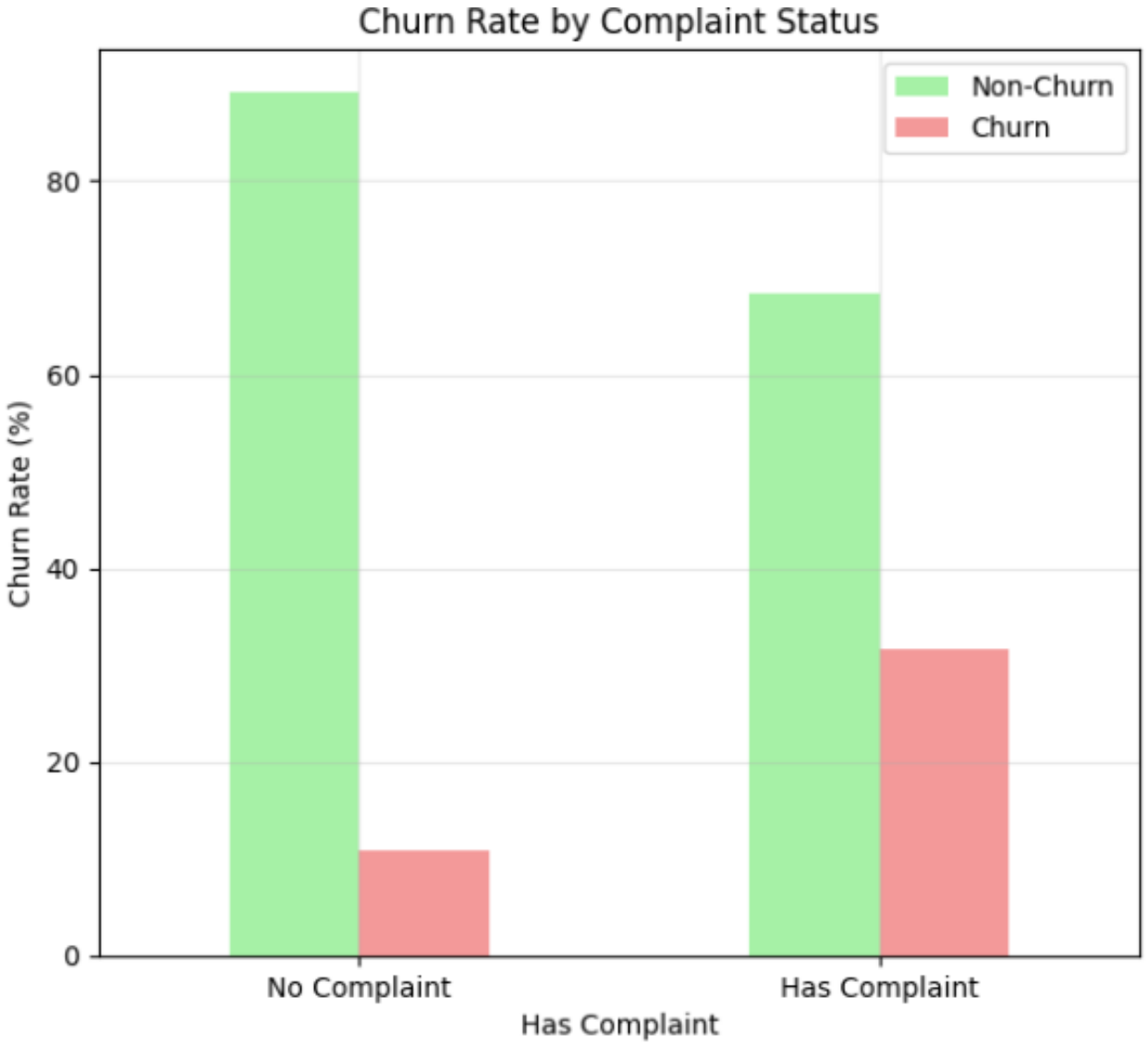
RESEARCH DESIGN AND IMPLEMENTATION

BIVARIATE ANALYSIS



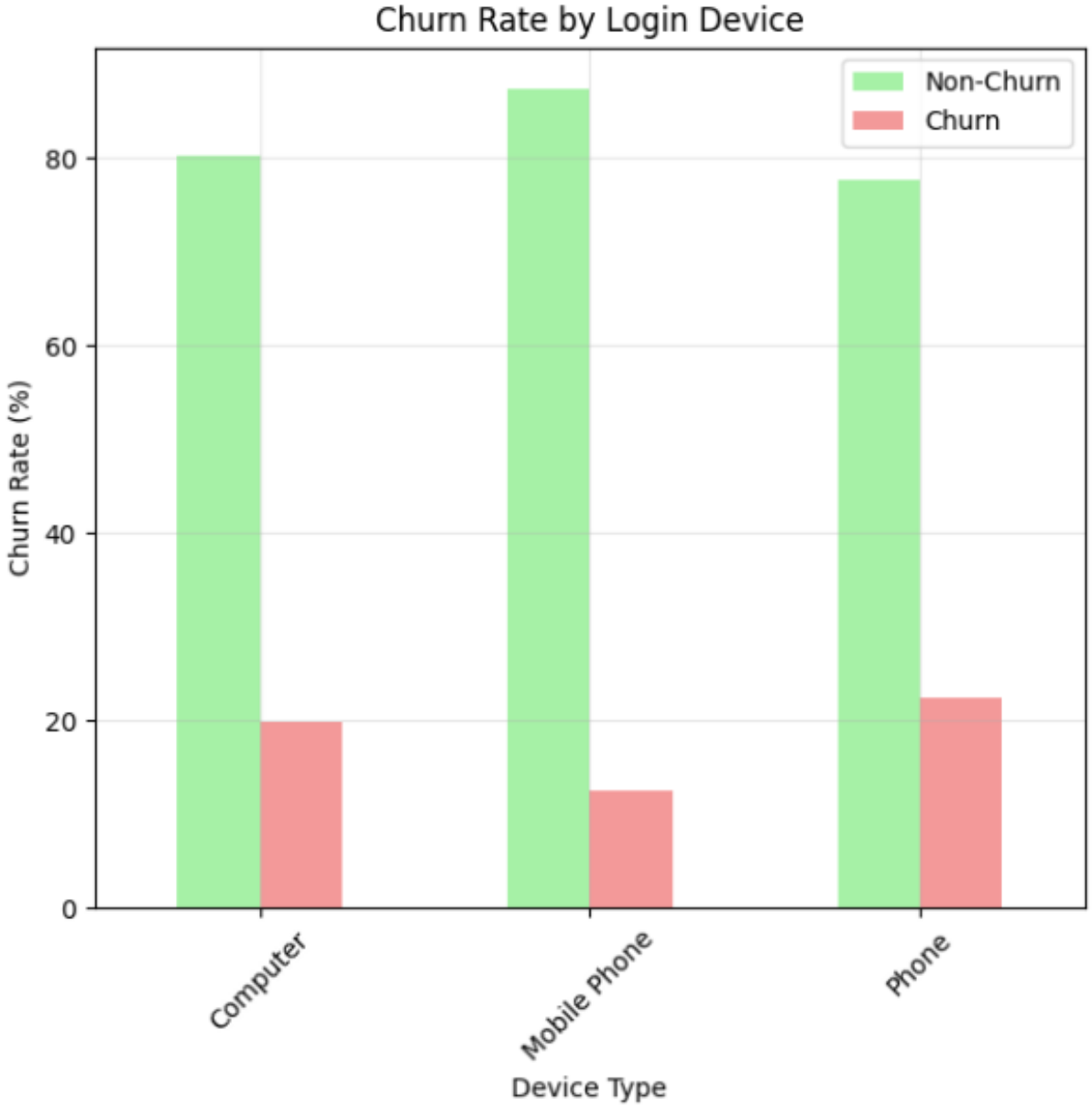
RESEARCH DESIGN AND IMPLEMENTATION

BIVARIATE ANALYSIS



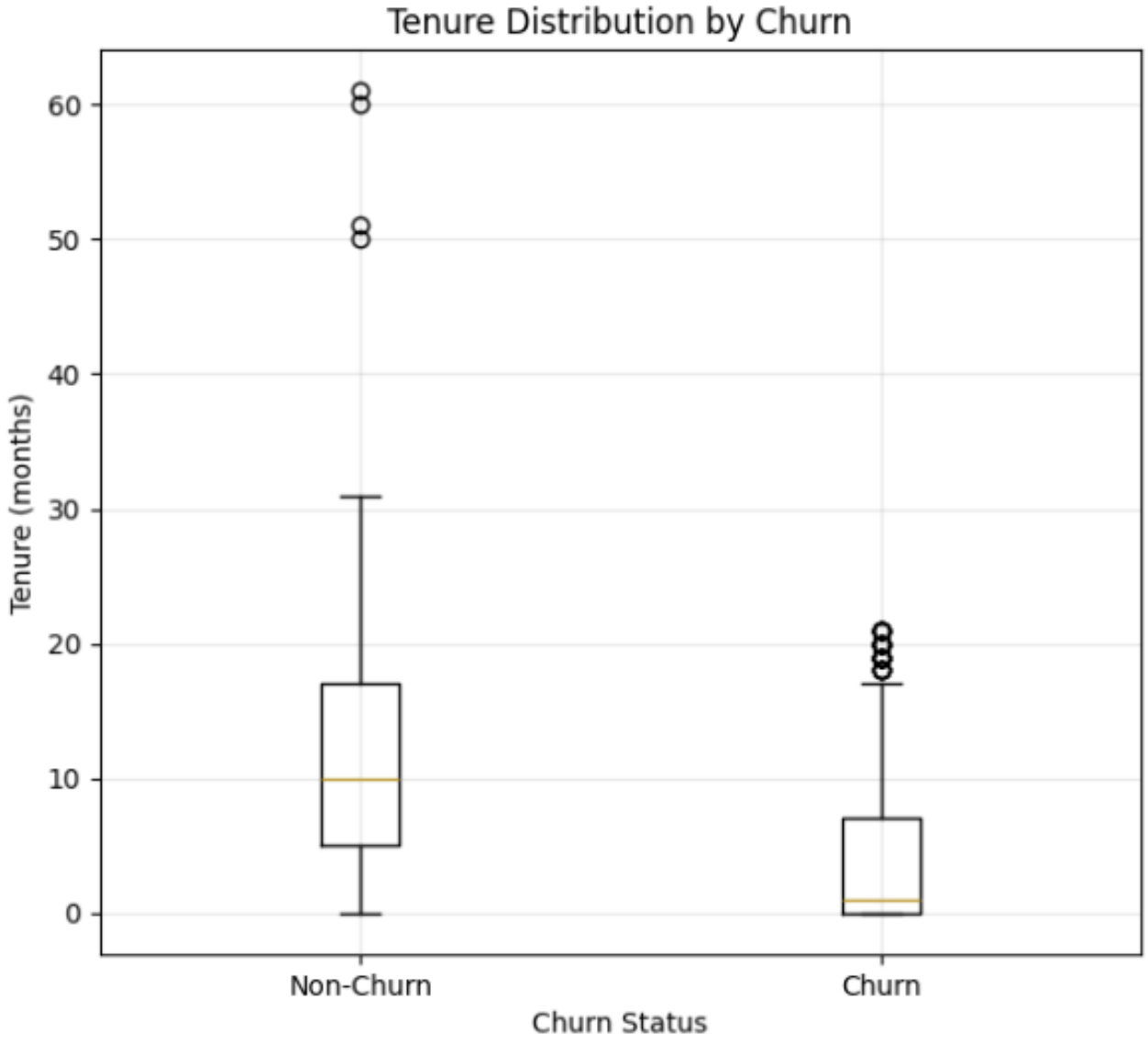
RESEARCH DESIGN AND IMPLEMENTATION

BIVARIATE ANALYSIS



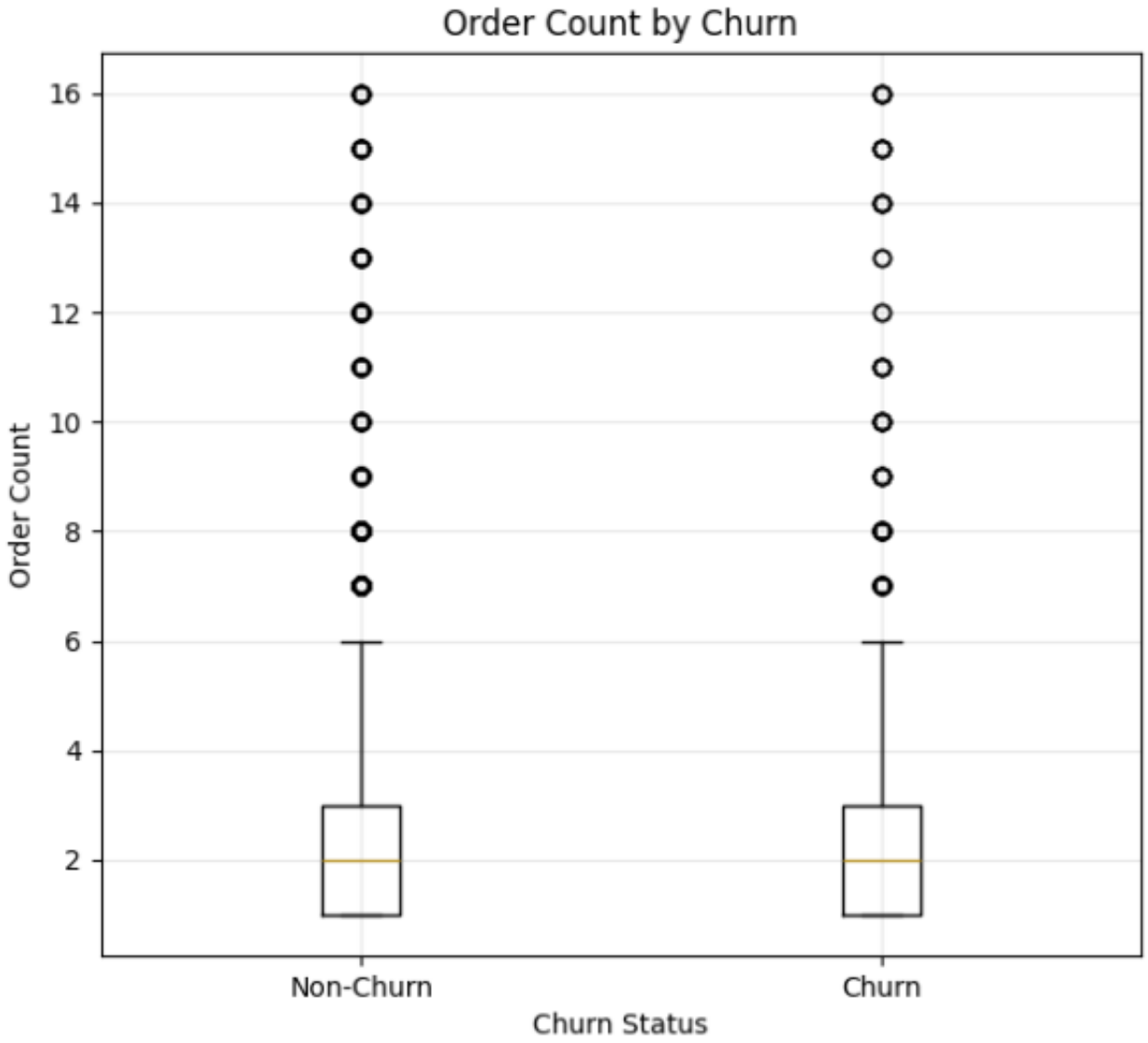
RESEARCH DESIGN AND IMPLEMENTATION

BIVARIATE ANALYSIS



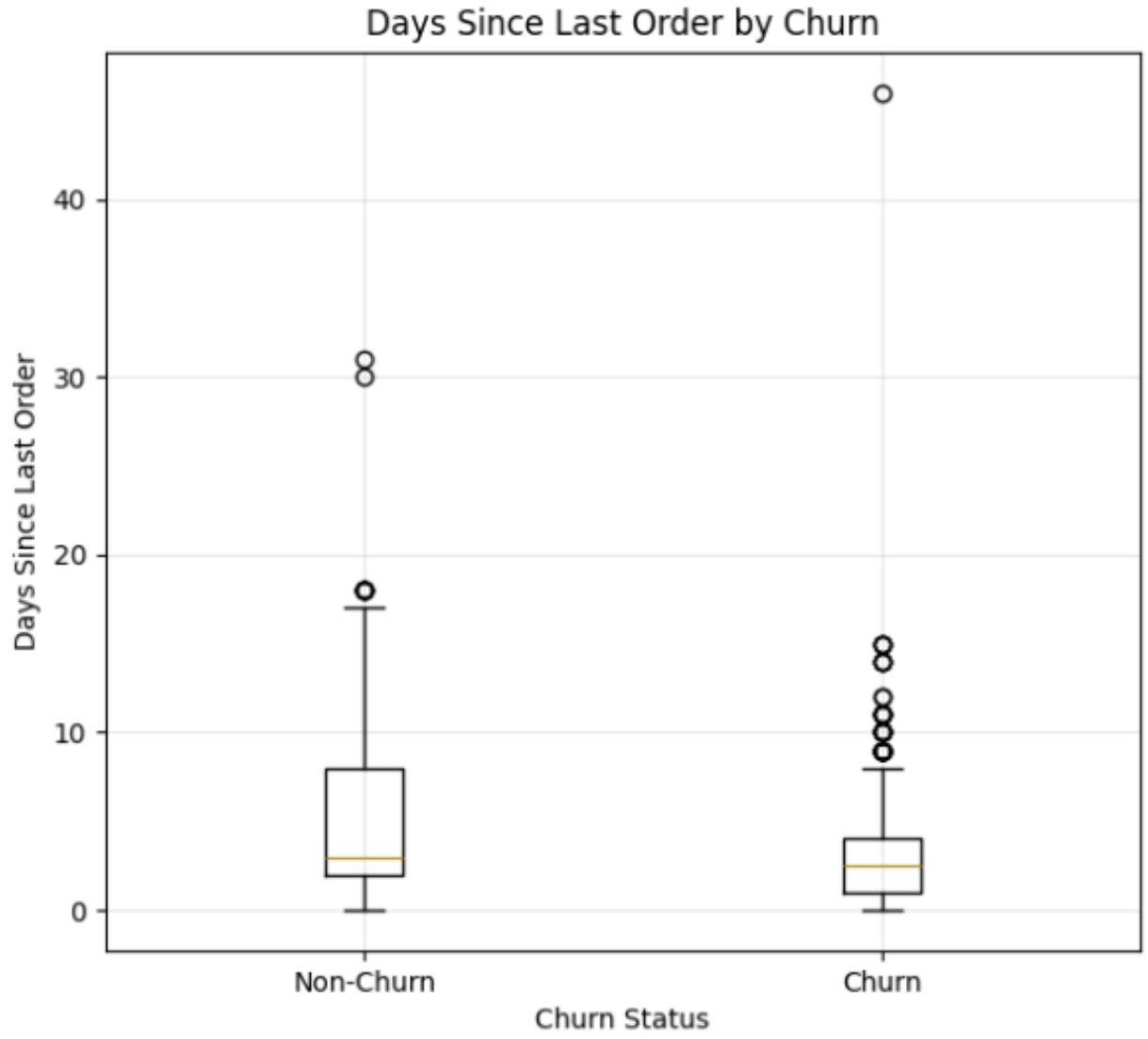
RESEARCH DESIGN AND IMPLEMENTATION

BIVARIATE ANALYSIS



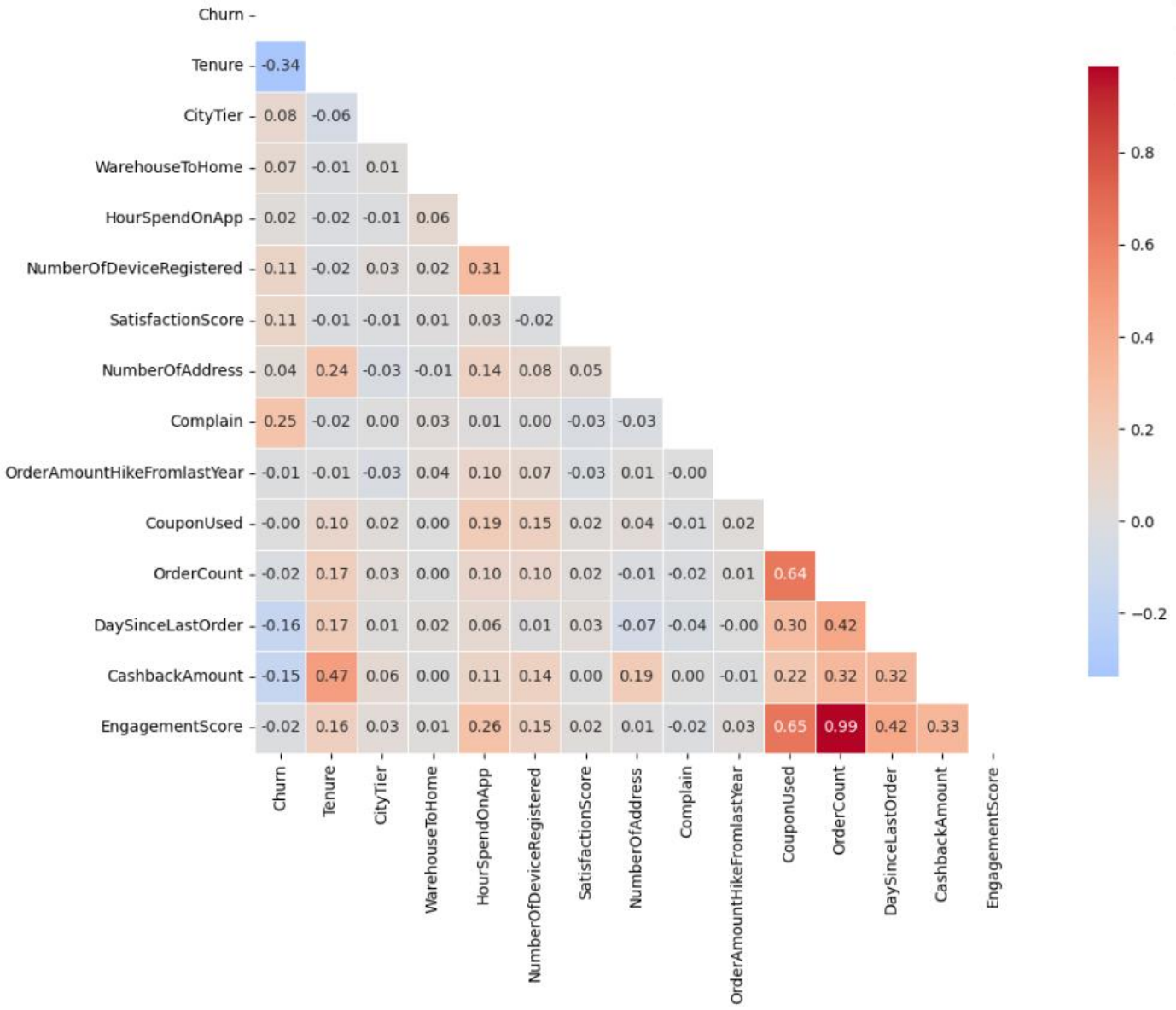
RESEARCH DESIGN AND IMPLEMENTATION

BIVARIATE ANALYSIS



RESEARCH DESIGN AND IMPLEMENTATION

CORRELATION HEATMAP



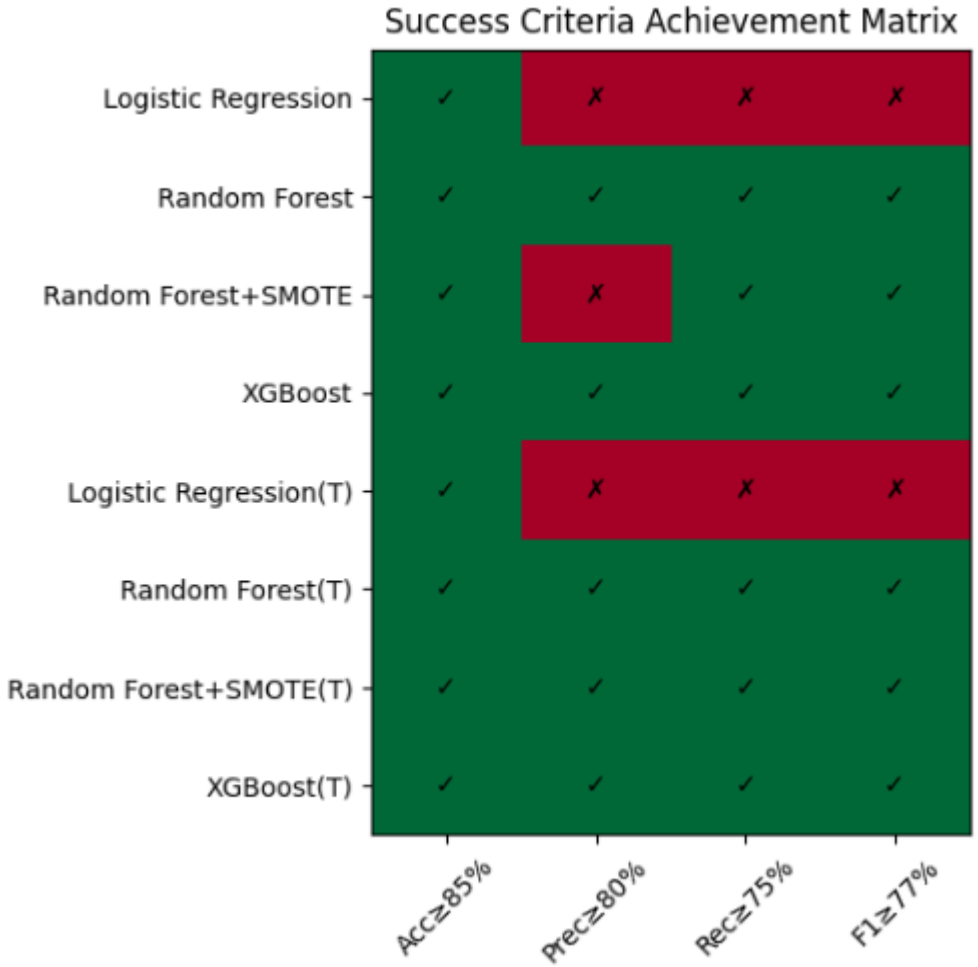
Model Performance Comparison (Before Hyperparameter Tuning)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Time (s)
XGBoost	0.9458	0.8107	0.8836	0.8456	0.9662	2.90
Random Forest	0.9476	0.8652	0.8148	0.8392	0.9643	3.42
Random Forest + SMOTE	0.9378	0.7960	0.8466	0.8205	0.9626	5.07
Logistic Regression	0.8792	0.6803	0.5291	0.5952	0.8730	0.15

Model Performance Comparison (After Hyperparameter Tuning)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Time (s)
Random Forest (Tuned)	0.9520	0.8649	0.8466	0.8556	0.9801	772.16
XGBoost (Tuned)	0.9449	0.8068	0.8836	0.8434	0.9643	111.27
Random Forest + SMOTE (Tuned)	0.9414	0.8187	0.8360	0.8272	0.9698	1287.92
Logistic Regression (Tuned)	0.8792	0.6803	0.5291	0.5952	0.8729	0.21

Success Criteria Achievement Matrix



DISCUSSION AND FUTURE WORK

Discussion

- Best Model Before Tuning: XGBoost
- Performance (F1-Score): 0.8456
- Best Model After Tuning: Random Forest
- Performance (F1-Score): 0.8556
- Key Insight: Behavioural factors > Demographics
- Critical Factor: Tenure (Strong negative correlation -0.338)
- Business Impact: High ROI potential via retention strategies

DISCUSSION AND FUTURE WORK

Future Work

- Larger dataset volume
- Application of Deep Learning Models
- Real Time Implementation and Monitoring
- Enhanced Feature Engineering with external Data Sources

THANK YOU

univteknologimalaysia utm.my utmofficial