CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the research context in the chain-of-thought approach, stepping up from the first step to the last step of the process execution. The Research Framework consists of 5 different stages. They are Problem Identification and Formulation, Data Understanding and Preparation, Exploratory Data Analysis (EDA) and Feature Engineering, Model Development and Training, and Model Evaluation and Deployment. This framework shows the direction to complete the solution implementation based on the Data Science Life Cycle approach.

3.2 Research Framework

The Research Framework consists of 5 different stages. They are Problem Identification and Formulation, Data Understanding and Preparation, Exploratory Data Analysis (EDA) and Feature Engineering, Model Development and Training, and Model Evaluation and Deployment. The research objectives are mapped with 3 of the 5 stages. Stage 2 (Data Understanding and Preparation) preprocess the customer churn data, ensuring the data is cleaned for model training process. Stage 4 (Model Development and Training) develops a machine learning model based on Random Forest algorithm that predicts the potential churning customers. Stage 5 (Model Evaluation and Deployment) starts measuring model performance using performance measures, model comparison and business intelligence.

3.2.1 Phase 1: Problem Identification and Formulation

The initial stage sets the research foundation through extensive acquisition of domain knowledge and problem contextualization. The feedback phase involves four components, that is, acquisition of knowledge through extensive exploratory review of literature, followed by problem definition in which the business issue is clearly defined. Then, stakeholder analysis is used to identify the significant stakeholders that are affected by customer churn. Next, gap identification is used to determine the existing research gaps that still have not resolved by previous researcher. Additionally, the success criteria are established to determine the outcomes that are quantifiable in metric form for the research study. This phase allows a proper understanding of customer churn phenomenon and its business implications before going into technical section.

3.2.2 Phase 2: Data Understanding and Preparation

The second phase revolves around data collection from different sources, mainly from Kaggle datasets. Data exploration involves performing statistical aggregations, assessing the data distribution, and performing data quality checks to verify the characteristics of the dataset. The preparation steps which is known as data cleaning involve handling missing values, outlier detection, and data normalization. This phase ensures research objective 1 to be achieved at the end, that cleans the existing dataset for model training, which would be further executed in the phase 4.

3.2.3 Phase 3: Exploratory Data Analysis (EDA) and Feature Engineering

This phase conducts rigorous data analysis to find the patterns and relationships that lies within the customer churn dataset. Correlation Analysis employs Pearson correlation coefficients, heatmap visualizations, and identification of key attributes to understand the relationship between each variables. Feature selection is used to filter methods, wrapper methods, and embedded methods to choose the predictor that is most relevant in predicting

customer churn. Feature Engineering extract features through categorical encoding and temporal feature extraction. These procedure provide insights into customer behaviour patterns and refine feature sets for machine learning model development, which would be useful in phase 4.

3.2.4 Phase 4: Model Development and Training

Phase 4 employs machine learning algorithms to develop customer churn prediction models. Before developing customer churn prediction model, it is required to split the dataset into 3 different sets, that is, training, validation and testing. Training constitutes 70%, while validation and testing sets constitutes 15% each respectively. This approach ensures the model evaluation result could be guaranteed at the end. Model training employs various algorithms limited to Logistic Regression, Random Forest (SMOTE) and Random Forest (XGBoost) to enable a comparative analysis from different perspectives. Hyperparameter Tuning utilizes Grid Search Cross Validation and performance optimization techniques to enhance model accuracy, which requires multiple try-and-error approach to find the best training value for the parameter. This phase directly addresses the Research Objective 2, that focuses on designing a machine learning model for predicting the customers that has potential to churn in the future.

3.2.5 Phase 5: Model Evaluation and Deployment

Phase 5 constitutes model evaluation and providing actionable business insights to the relevant stakeholders. Performance metrics evaluation include confusion matrix analysis, accuracy, F1-Score, and ROC-AUC analysis to evaluate the model performance. Model comparison involves statistical testing, best model selection, and result validation across multiple algorithms. Business Intelligence components include stakeholder visualization by dashboard creation, actionable insight development to guide business strategy, and development of recommendations for anti-churn initiative programs. This phase clearly matches with Research Objective 3, that evaluates the model performance and identifying churning customers according to various performance metrics, enabling proactive business actions to minimize financial impact as much.

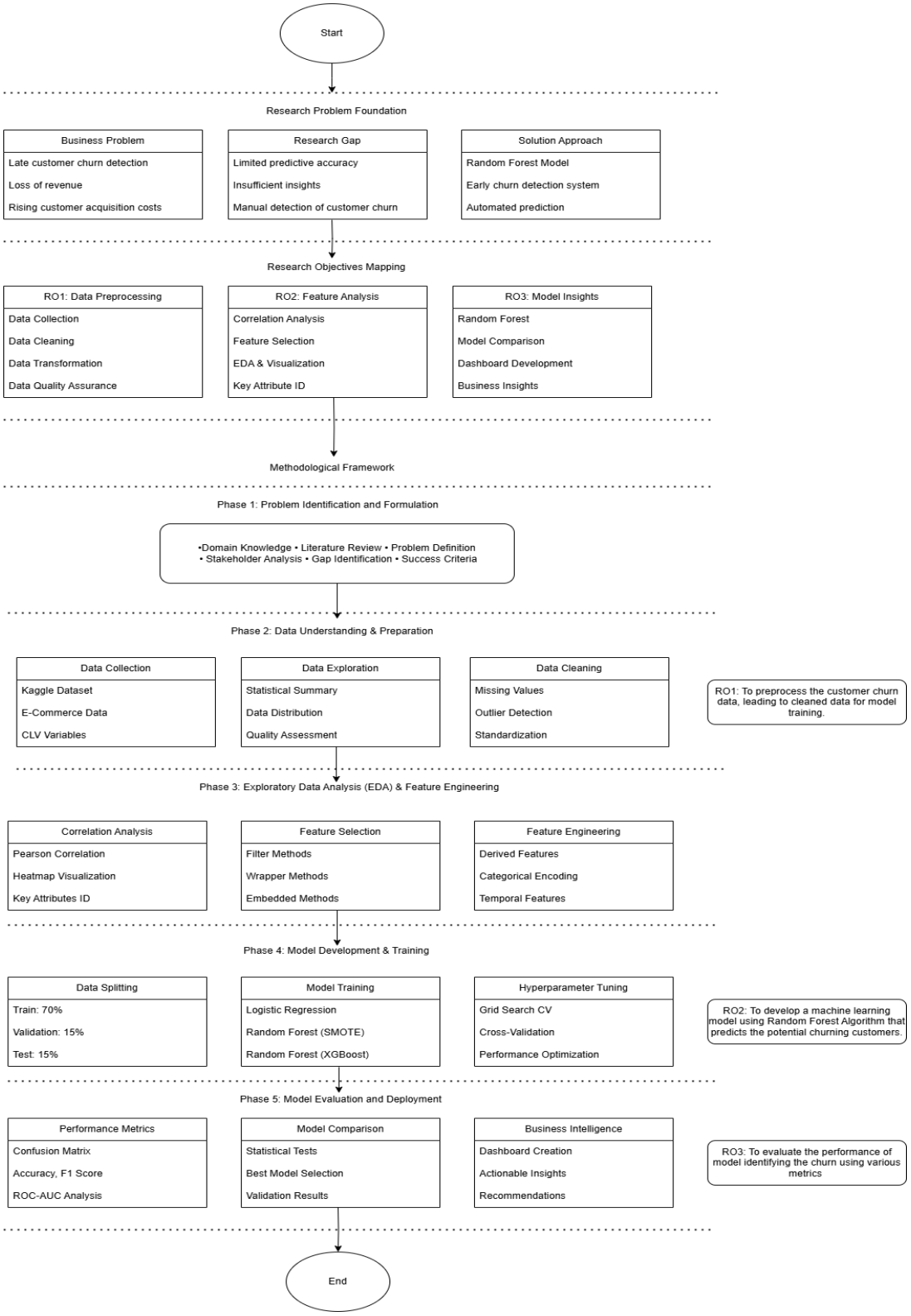

Figure 3.1 Research Framework

3.3 Phase 1: Problem Identification and Formulation

The stage of identification and formulation process begins from systematic review of customers' churn behaviours in e-commerce environments. It is built upon the extensive literature review to craft an explicit research direction. This phase focuses more on the alignment between academic research and practical applications, ensuring that the research would be valuable for both theoretical understanding and industry practice. This process is iterative, as it is refined by looping into multiple cycles. Each cycle shows an improvement in domain insights and understanding of stakeholder requirements.

Domain knowledge acquisition commenced with intensive literature review across numerous databases and publication types. The search strategy was iterative and systematic in approach, starting from keyword searches according to the topic, followed by narrowing down the scope of searching, ending up with specific research topic and research gaps. This literature review recognized several recurring problems in past research, including non-standard evaluation procedures, inadequate business context, and insufficient consideration of model interpretability requirements. The results in the previous research provide a clear direction on the gaps that needs to be resolved in the current research.

The formulation of problem definition involved critical discussion of technical and business issues surrounding customer churn in e-commerce contexts. As compared to typical sectors such as telecom or banking, e-commerce contexts present uniquely challenges that include higher transaction frequency, cross category products, and multichannel complexity of customer interactions. These characteristics necessitated adaptation of standard churn prediction methods while maintaining sufficient rigor to yield correct conclusions.

Stakeholder analysis revealed a number of stakeholders with potentially conflicting needs and success factors. E-commerce management groups primarily seek strategic advice to guide customer retention planning, with a need for actionable forecasting with measurable business effects. But there is limited technical expertise to make them opt for explainable models rather than exclusive accuracy-optimizing approaches. Customer success and marketing teams need high-precision targeted intervention capabilities to minimize resource wastage, while they need near-real time predictions to succeed in campaign execution. Data

science teams are seeking a robust, sustainable predictive models with high accuracy and consistency over different data conditions, constrained by computational performance and system integration requirements.

The mapping of these stakeholder requirements revealed several fundamental areas that needs to be negotiated carefully in methodology design. Sometimes the stakeholder requirements contain disagreement that needs to be addressed out for multiple verification and refinement. The classic trade-off between interpretability and model accuracy was found to be particularly applicable, with technical teams desiring predictive capability and business teams desiring understandability for decision-making. Similarly, precision versus recall preference trade-offs varied between stakeholder groups, requiring flexible evaluation strategies that could accommodate different optimization criteria.

Research gaps were identified systematically via the literature review, and several key areas of research were revealed. Possibilities for algorithmic enhancement included limited exploration of Random Forest variants specifically tuned to e-commerce churn scenarios, no comparison of SMOTE and XGBoost approaches in realistic business contexts, and limited holistic evaluation frameworks that consider both statistical and business performance measures. These gaps provided sufficient justification for the comparative approach in this research.

Dataset and evaluation limitations were another visible gap area, with most existing studies employing proprietary or limited datasets that do not allow reproducibility. Inconsistent utilization of evaluation metrics across studies bars interesting comparison, while limited temporal validation procedures fail to reflect real-world deployment environments. These limitations influenced the decision to utilize publicly available datasets and implement stringent evaluation protocols.

Setting the success criteria involved making finer balancing acts between technical excellence and meaningful value creation. Technical requirements were taken from literature review findings with minimum 85% overall precision requirements, greater than 80% precision targets to guarantee low false positive intervention costs, recall thresholds greater than 75% to detect sufficient churn cases to have business influence, and F1-score targets greater than 77%

as balanced performance measures. These needs were complemented by business impact criteria based on the effectiveness of intervention, resource optimization, revenue protection, and proof of scalability.

Problem formulation was tested and validated via several channels for research quality and practical use. Consultation with academic supervisor guaranteed external input towards research rigor and contribution potential, while validation of industry context via existing case studies and reports guaranteed practical use. Methodological consistency checks against existing data science frameworks guaranteed best practice adherence, while systematic verification guaranteed research questions directly solve identified gaps.

3.4 Phase 2: Data Understanding and Preparation

Data preparation and knowledge is a foundational step essential to having direct influence on all subsequent analysis and modeling work. Strategy deployed here is based on lessons learned from literature review, namely addressing the data quality issues and class imbalance issues which are ever-present issues of concern in previous studies. This step entailed systematic probing of accessible datasets, careful checks of the properties of the data, and intensive preparation procedures to present optimal conditions for model construction.

Data collection process involved identifying suitable datasets that support realistic representation of e-commerce customer behavior without compromising reproducible research availability. After careful evaluation of a number of available datasets, the Kaggle e-commerce customer churn dataset was selected based on its extensive coverage of relevant customer attributes, sufficient sample size for pertinent analysis, and public access that supports research reproducibility. This dataset contains 5630 rows and 20 columns. It contains 2 sheets, that is Data Dict and E Comm. Data Dict describes the data that is contained within each column whereas E Comm displays all the customer data with relevant attributes. The data source link is as follows https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

Exploratory data analysis started with intensive examination of structural characteristics like sample size, feature space, types, and general quality indicators. The database consists of customer records across several attribute categories portraying different aspects of customer interaction with the e-commerce website. Demographic attributes like Gender, CityTier, MaritalStatus, NumberOfAddress, and WarehouseToHome provide information regarding customer background properties that could potentially lead to churn likelihood. Behavioral attributes like Tenure, HourSpendOnApp, OrderCount, CashbackAmount, CouponUsed, PreferredOrderCat, PreferredPaymentMode, and DaySinceLastOrder are indicative of actual customer interaction behavior and transactional activities.

Experience-driven features include SatisfactionScore, Complain, and OrderAmountHikeFromLastYear that record dimensions of customer satisfaction and service experience directly related to churn propensity. Technology profile features such as PreferredLoginDevice and NumberOfDeviceRegistered provide information on customer trends of technology uptake likely to be related to engagement levels and chances of retention. Target variable Churn provides the binary class outcome necessary to build supervised learning models.

Data quality evaluation will identify a number of real-world dataset challenges, including missing values, possible outliers, and issues with class imbalance. The class imbalance issue was especially marked, with churned customers making up about 16.8% of the overall sample. Although this imbalanced ratio is more favorable than the severe imbalances found in some published literature reports, it nonetheless must be treated with caution in model development so that proper learning can take place from minority class instances.

Missing value analysis will provide data completeness patterns within different attribute categories, with systematic missing patterns in some of the attributes that demand special treatment. The treatment included thorough examination of missing value mechanisms for the purpose of determining if data were missing completely at random, missing at random, or missing not at random. Analysis will inform the selection of appropriate imputation methods that minimize bias introduction without sacrificing underlying data relationships.

Outlier detection employed a mix of statistical techniques like interquartile range analysis, z-score calculation, and business logic checks at the domain level. It checked the outliers very carefully to distinguish real extreme values that indicate genuine customer behavior from faulty data points that must be repaired or removed. It was a balancing act between having absolute data cleaning and keeping genuine behavioral variance that assists model learning.

Data transformation and standardization operations ensured consistent format and scale for all attributes without sacrificing underlying relationships and patterns. Categorical variables were suitably treated for proper encoding with techniques selected based on cardinality and target variable relationship. Numerical variables were examined for characteristics of distribution and transformed appropriately to satisfy modeling assumptions at the expense of interpretability.

The final readiness dataset is the outcome of careful quality improvement processes undertaken to optimize conditions for building models while preserving the inherent nature of customer behavior patterns. Quality validation processes ensured effective resolution of issues encountered with problems identified in data while guaranteeing dataset representativeness and integrity. The readiness dataset is used as the foundation for exploratory analysis and building all models to be undertaken later on, with clear documentation of preparation processes ensuring reproducibility and transparency of research work.

3.5 Phase 3: Exploratory Data Analysis and Feature Engineering

This phase will establish a comprehensive understanding of customer behavior patterns and relationships within the dataset through systematic statistical analysis and feature optimization. The EDA process will employ both univariate and multivariate analysis techniques to uncover hidden patterns that influence customer churn decisions, while feature engineering will transform raw data into optimized inputs for machine learning models.

3.5.1 Exploratory Data Analysis Methodology

For Univariate Analysis, the procedure will begin with one-variable exploration to learn about properties of the distribution and data quality for every attribute. Numerical variables will be examined through descriptive statistics like mean, median, standard deviation, skewness, and kurtosis to check shapes of the distribution and ascertain possible data quality issues. Box plots and histograms will be developed to see the distribution and identify outliers or anomalies. Categorical variables will be examined with frequency distributions and bar charts to understand the relative proportions across categories and identify if there are class imbalances in individual features.

Regarding Bivariate Analysis, correlation between every feature and target variable (churn) will be examined to identify the most important predictors. For numerical variables, Pearson correlation coefficients will be employed to quantify linear relationships with churn outcomes via correlation analysis. Statistical significance tests will be employed to check if the strength of these relationships is significant. For category variables, chi-square tests of independence will be conducted to test for association with churn behavior, with contingency table analysis being included as a supporting tool to identify patterns of distribution across categories.

For Multivariate Analysis, multicollinearity tests and dependencies of features that can impact model performance will be detected by extensive correlation analysis among all numeric variables. Correlation heatmaps will be generated to visualize the relation matrix and identify the groups of highly correlated features. Cross-tabulation will be applied in examining interaction between two or more categorical variables and their joint effect on churn outcomes. Statistical tests will be employed to verify the significance of any perceived relation and guarantee firm feature selection decisions.

3.5.2 Feature Engineering Methodology

For Categorical Encoding, categorical data shall be mapped to numerical values suitable for machine learning algorithms. The encoding method shall be chosen based on each categorical variable's cardinality and relationship with the target variable. Low-cardinality

nominal variables will be processed through one-hot encoding for categorical uniqueness maintenance, and ordinal variables will be processed through label encoding for retaining natural ordering relationships. High-cardinality categorical variables will be examined for target encoding methods if appropriate.

Feature Scaling and Normalization Numerical features will be examined for scale differences and rescaled to optimize model performance. Min-max scaling shall be applied to features with bounded ranges, while standardization (z-score normalization) shall be applied to features with normal distributions. The scaling approach would be determined by the distribution features of individual variables as well as the requirements of selected machine learning algorithms.

Feature Generation Derived Feature Generation New features will be created from domain knowledge and exploratory discovery for better predictability. Ratio-based features will be created to identify relationships between similar variables such as cashback-to-order ratios or engagement-to-tenure correlations. Temporal features will be created from date-based variables to extract seasonality and recency effects. Interaction features will be created for variables with significant combined effects on churn outcomes.

Regarding feature selection approach, a hybrid approach involving filter, wrapper, and embedded techniques will be applied to identify the most informative features to predict churn. Filter techniques will employ statistical measures like mutual information and correlation coefficients to measure features on a basis of individual predictive power. Wrapper techniques will employ recursive feature elimination with cross-validation to measure feature subsets based on model accuracy. Embedded techniques will employ regularization techniques to select features automatically while training the model.

3.5.3 Tools and Implementation Framework

Python shall be utilized as the foundation platform for feature engineering and EDA. Pandas library will be utilized for data transformation and manipulation operations, whereas NumPy will be utilized for doing computations. Matplotlib and Seaborn libraries will be

utilized to do statistical plots and visualizations to discover patterns. Scikit-learn will provide feature engineering tools and statistical test functions.

Statistical significance tests will be used to validate all EDA results to ensure good conclusions. Data leakage protection will be verified through feature engineering transformation documentation and testing. Generalizability of feature selection will be promoted by using cross-validation techniques. Business interpretability will be tested for engineered features in order to maintain model explainability for stakeholder reporting.

The outcome of this stage will be an enriched dataset with enriched features that detect the most impactful customer behavior patterns to forecast churn, supported by thorough statistical analysis documenting the relationships and trends in the customer data.

3.6 Phase 4: Model Development and Training

Customer churn prediction model evaluation requires aggregate metrics that capture both statistical accuracy and business usefulness. Variability in the evaluation approaches among studies from MVL suggests that comparison is impossible.

Model development strategy seeks to utilize algorithms specifically designed to binary classification problems in order to ensure methodological appropriateness in customer churn prediction. Linear regression was deliberately excluded from this research as it produces unbounded continuous outputs that cannot be meaningfully interpreted as probabilities in binary classification contexts. The selected algorithms provide comprehensive coverage of both statistical and ensemble learning approaches while maintaining suitability for the binary nature of churn prediction.

Logistic Regression serves as the statistical baseline method with the sigmoid function formulation $p(y=1|x) = 1/(1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)})$, where β_i represents the change in log-odds per unit increase in feature x_i, e^{β_i} provides the odds ratio for feature x_i, and $|\beta_i|$ indicates the strength of effect. This formulation naturally constrains outputs between 0 and 1,

providing interpretable probability estimates for churn likelihood. Key hyperparameters include regularization strength, solver algorithms, maximum iterations, and penalty terms that control model complexity and convergence behavior.

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \quad (3.1)$$

Where:
βi = change in log-odds per unit increase in feature xi
eβi = the odds ratio for feature xi
|βi| = the strength of effect

Random Forest is the case of the ensemble approach with the classification prediction function $\hat{y} = \text{mode}\{T_1(x), T_2(x), ., T_n(x)\}$ with T_1 through T_n being individual decision trees trained on bootstrap samples and the final prediction being the majority vote among all trees. This approach tackles the binary classification problem by aggregating single tree predictions, and it naturally outputs class probabilities by vote ratios. Key hyperparameters of Random Forest include n_estimators (number of trees), max_depth (limit of tree depth), min_samples_split (minimum samples required for splitting nodes), min_samples_leaf (minimum number of samples in leaf nodes), and max_features (the maximum number of features to consider at each split).

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), ., T_n(x)\} \quad (3.2)$$

Where:
βi = change in log-odds per unit increase in feature xi
eβi = the odds ratio for feature xi
|βi| = the strength of effect

The approach also considers more sophisticated Random Forest implementations to address specific churn prediction issues. Random Forest with SMOTE is intended to address class imbalance by generating synthetic instances of churned customers before training, which

could improve minority class detection. Random Forest with XGBoost integration applies gradient boosting techniques to optimize prediction accuracy with iterative error correction. These implementations facilitate robust testing of ensemble techniques while monitoring approaches compatible with binary classification operations.

Data splitting operations adhered to best practices while being tuned into the idiosyncrasies of the churn prediction task. Data splitting was conducted into training (60%), validation (20%), and testing (20%) sets using stratified sampling to maintain class distribution equality across all subsets. This splitting ratio leaves sufficient training data for model learning and reserves sufficient validation and test samples to allow robust performance assessment using cross-validation processes. The stratification ensures that the issue of imbalance in classes is presented equally in all data subsets.

Both algorithms' training procedures were designed with extreme caution to maximize performance with equal comparison conditions. Logistic Regression training employed regularization techniques to prevent overfitting without losing the ability to interpret the model. The strength of the regularization was determined through cross-validation procedures that provided an optimal compromise in bias-variance trade-off for the binary classification task. The application of a sigmoid activation function guarantees that all predictions fall within the appropriate probability range of 0 to 1, and outputs are directly interpretable as probabilities of churn.

Random Forest models needed to have hyperparameter choice handled carefully, especially the number of trees, depth, and minimum samples per leaf. These had a critical impact on model accuracy and computational cost, so they needed systematic tuning using grid search cross-validation approaches. Optimizing between prediction accuracy and computation cost, final models would still be feasible for deployment cases.

Use of SMOTE solved the class imbalance issue by synthetically oversampling the minority class with artificial churned customer examples to balance the training set. The SMOTE parameters like k-nearest neighbors and sampling strategy were optimized to generate realistic synthetic examples to encourage model learning without the introduction of artifacts.

Validation procedures were designed in a meticulous way to prevent synthetic examples from augmenting overfitting against artificial patterns rather than enhancing models' generalization.

XGBoost application employed gradient boosting technique to enhance Random Forest performance through iterative correction of error. The parameters of boosting like learning rate, maximum depth, and regularization terms had to be set with utmost care so as to achieve maximum performance without excessive overfitting. The iterative nature of gradient boosting necessitated monitoring of validation performance so as to set optimal stopping points in addition to limiting complexity.

Hyperparameter tuning applied systematic grid search techniques in combination with cross-validation to determine the optimal parameter combinations for each algorithm. This was computationally demanding but guaranteed that the performance comparisons reflect the best achievable results for each strategy and not inferior default settings. The optimization procedures applied both accuracy metrics and computational speed factors to determine practically implementable parameter values.

Model validation procedures extended beyond simple accuracy measurement to include comprehensive evaluation of prediction stability and generalization capability. Cross-validation approaches assessed performance consistency across different data subsets, while temporal validation procedures evaluated model stability over time-based splits that better reflect deployment scenarios. These validation approaches provide confidence that observed performance differences reflect genuine algorithmic advantages rather than random variation or overfitting.

3.7 Model Evaluation and Deployment

The comparison between Linear Regression, Logistic Regression, and Random Forest model has been made in previous Chapter. The selected model would be based on Logistic Regression and Random Forest algorithm. The reason of selection would be mentioned in the next paragraphs.

Linear Regression is not selected as it only predicts a continuous numerical value. Customer Churn Prediction requires a binary outcome or the probability of churning, which is denoted as 0 for no and 1 for yes. Linear Regression outputs are unbounded, which indicates that the result would be in decimal form, for example, 0.5, 1.3, that is meaningless to the probability requirement. It does not make sense to the probability approach.

Logistic Regression is selected because the method itself is designed for predicting the binary outcomes. At the same time, it generates a Sigmoid-Shaped curve that captures the non-linear relationship between features and probability of churn. Logistic Regression is built upon Bernoulli distribution, matching the binary nature of the target variable. Probability score is well-calibrated, that is highly interpretable and useful for ranking customers by risk of churning.

Random Forest is selected due to the nature of handling binary classification problems without modification. It could handle non-linearity and complex interactions. It gives a higher predictive power due to combination of multiple decision trees. No strict assumptions such as linearity, normality of errors, or homoscedasticity exist.

Each model performance would be evaluated using the metrics as mentioned below, that is Confusion Matrix, Accuracy, F1 Score, ROC-AUC Analysis, and External Factors and Environmental context. The details would be described in detail in next subtopics.

3.7.1 Confusion Matrix

Confusion matrix is a table that is used to predict the performance of classification model using tabular form. The confusion matrix is important, because it reveals the mistakes that has been done by the model, specifically narrow down the error scope. Before going to the performance metrics, it is important to understand the role played in Confusion Matrix as

shown below. 4 key attributes (True Positive, True Negative, False Positive, False Negative) are used to predict the customer churn status.

Table 3.1: Confusion Matrix table

	Actual Churn (Yes)	Actual Non-Churn (No)
Predicted Churn (Yes)	True Positive (TP)	False Negative (FN)
Predicted Non-Churn (No)	False Positive (FP)	True Negative (TN)

True Positive (TP) indicates that model successfully predicted the customer as churn before the customer start to churn. True Negative (TN) indicates that the model predicted customer as not churn, the customer is not churn in reality. False Positive (FP) indicates that the model predicted the customer churn but the customer did not churn. False Negative (FN) indicates that the model predicted the customer as non-churn, but the customer will be going to churn in the future.

3.7.2 Accuracy

Accuracy is the metric derived from Confusion Matrix, as stated in 3.7.1. It measures the overall correctness of the classification model. The formula is shown below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.3)$$

Where:

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

3.7.3 F1 Score

F1 Score is a harmonic mean of precision and recall. Precision quantifies the reliability of positive class predictions. It measures the proportion of correctly identified instances among all instances predicted as positive. Recall measures the completeness of positive class identification. It calculates the proportion of actual positive instances correctly predicted by the model. The formula for Precision, Recall, and F1-Score is shown below.

$$Precision = \frac{TP}{TP + FP} \quad (3.4)$$

$$Recall = \frac{TP}{TP + FN} \quad (3.5)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.6)$$

Where:

TP = True Positive

FP = False Positive

FN = False Negative

3.7.4 ROC-AUC Analysis

ROC (Receiver Operating Characteristics) is a graphical analysis tool for evaluating binary classifiers across all possible decision thresholds. It visualizes the tradeoff between sensitivity (True Positive Rate) and 1-specificity (False Positive Rate) as the classification threshold varies.

AUC (Area Under the Curve) calculates the area under the ROC curve, collapsing the ROC's 2D information into a single number. It quantifies the ROC curve performance into a single metric ranging from 0 to 1. The formula and the value of AUC is shown below. AUC estimates the probability that the model rates a randomly chosen churning customer higher than a randomly chosen non-churning customer. The greater the AUC, the stronger is the model discrimination power, which is critical to identify risk customers before they actually churn.

Table 3.2: Performance of AUC Value

AUC Value	Performance
0.5	Random Classification (No discriminative ability)
0.6-0.7	Poor
0.7-0.8	Acceptable
0.8-0.9	Excellent

0.9-1.0	Outstanding

3.7.5 External Factors and Environmental Context

The e-commerce marketplace is very competitive with constant promotional campaigns, seasonal sales, and market entrants. Such outside market forces have the potential to create abrupt alterations in customer behavior not contained in historical training sets. Competitor actions such as deep discounting of prices or improved service alternatives can introduce unexpected churn patterns.

Macro-economic factors like levels of inflation, consumer spending power, and recessions play major roles in determining customer purchasing behavior and loyalty. In times of economic uncertainty, customers will become price-sensitive and migrate to other platforms if alternative offers are available, thereby affecting model performance.

Abrupt technological advancement, user interface revamps, mobile app revamps, or site overhauls may influence customer satisfaction and retention patterns. The model should provide room for potential short-term spikes in churn rates following major platform overhauls.

Seasonal trends in customer behavior characterized by holidays, festivals, back-to-school periods, and culture events are prevalent for e-commerce customer behavior. The model's predictions should be compared against these time-of-year variations to avoid misattributing natural seasonal churn changes.

Churn patterns can radically vary across different demographic groupings, geographic regions, and cultural settings. The model's performance must be tested across different customer segments to preserve fairness in accuracy of prediction and avoid bias in retention strategies.

3.8 Mapping between Research Phases, Questions, Objectives, Activities, and Deliverables

The research stages are charted together with questions, activities, and deliverables as indicated in the table below.

Table 3.3: Mapping between Research Phases, Questions, Objectives, Activities, and Deliverables

Research Phase	Research Questions	Activities	Deliverables
Phase 1: Problem Identification and Formulation	1.How to accurately predict the churning customers before churn starts to happen using machine learning algorithms?	1. Define research problems 2. Investigate the dataset from previous researcher. 3. Investigate the machine learning techniques done by the previous researcher. 4. Investigate the model performance with the predefined metrics.	Chapter 1 and 2
Phase 2: Data Understanding and Preparation	1.What are the datasets that can be used to predict churning customers?	1. Find out the limitations of the dataset from the previous researcher. 2. Select the dataset	Chapter 2

		that is suitable in this case.	
Phase 3: Exploratory Data Analysis (EDA) and Feature Engineering	1. How to explore the data attributes from the selected dataset? 2. How to find the key attributes that influences the customer churn rate?	1.Find out the attribute of the dataset using EDA techniques. 2. Application of ensemble learning techniques to find out the best individual predictor	Chapter 4
Phase 4: Model Development and Training	1.How to train and develop the model?	1.Develop and train the model using the cleaned dataset from Phase 3, together with selected machine learning algorithms.	Chapter 5
Phase 5: Model Evaluation and Deployment	1.How to define the best model from the experiment?	1.Evaluate the model performance with metrics.	Chapter 5
Conclusion		1.Conclude Research Contributions 2. Research Limitations and provide	Chapter 5

		recommendation for future improvements	

3.9 Summary

Chapter 3 demonstrates the overall steps that needs to be executed according to Data Science Life Cycle. The 5 main phases included Problem Identification and Formulation, Data Understanding and Preparation, Exploratory Data Analysis (EDA) and Feature Engineering, Model Development and Training, and Model Evaluation and Deployment. The 5 phases mentioned in this Chapter act as a framework to be executed in the next chapter. The next step Chapter would be developing the model and perform model training using the cleaned dataset. The steps execution details would be described in chain-of-thought process.