**Research Methodology Practical Work**

**Student Details**

**Name:** Manasi Upadhyay  
**Student ID:** 5146271  
**Subject:** MSc IT (Cloud Computing)   
**Date:** 3rd January 2025

**1. Introduction**

**Problem Statement:**   
High customer churn rates pose a significant challenge for subscription-based businesses, leading to substantial revenue losses and increased customer acquisition costs. Traditional methods of predicting churn often lack precision and fail to generalize across industries. Leveraging machine learning techniques presents an opportunity to enhance prediction accuracy and inform retention strategies, ultimately reducing churn rates and boosting profitability

**Objective:**   
The primary objective of this research is to develop and evaluate machine learning models to predict customer churn with high accuracy. By incorporating feature engineering and ensemble learning techniques, the study aims to.

1. Identify critical factors influencing customer churn.
2. Demonstrate the effectiveness of advanced machine learning models over traditional predictive methods.
3. Provide actionable insights to reduce churn and improve customer retention strategies.

**Research Plan:**

**Research Type:**

* **Descriptive research to analyze historical patterns in customer data.**
* **Experimental research to test and validate predictive models.Research Design:**

**Approach:** Mixed (quantitative for model performance and qualitative for feature selection insights).

**2. Literature Review**

**1. "Comparative analysis of logistic regression and decision trees for churn prediction."**

**Insight: Logistic regression is simple but less accurate; decision trees provide better interpretability but may overfit.**

**2. "Applications of ensemble methods in e-commerce."**

**Insight: Techniques like boosting improve prediction reliability.**

**3. "Challenges in deploying real-time churn prediction systems."**

**Insight: Lack of frameworks for operational integration.**

**4. "Deep learning models for customer retention in telecom."**

**Insight: High computational requirements limit scalability.**

**5. "Leveraging feature engineering for churn prediction."**

**Insight: Feature engineering can improve accuracy but demands domain knowledge.**

**Research Gaps:**

1. **Limited emphasis on real-time prediction and integration with operational systems.**
2. **Underexploration of ensemble techniques such as XGBoost, stacking, and boosting in churn prediction.**
3. **Lack of generalized models for diverse industries.**

**Reviewed Studies:**

1. **Comparative analysis of logistic regression and decision trees.**
2. **2. Applications of deep learning, despite high resource needs.**
3. **3. Ensemble methods in e-commerce for reliable predictions, similar to Kubernetes’ resource allocation strategies.**

**3.Proposed Methodology:**

**1. Data Preprocessing:** Cleaning: Handle missing values (e.g., mean, median, or KNN imputation) and remove outliers (IQR or Z-score methods).

**Feature Engineering:** Derive meaningful features like historical churn likelihood and average session duration. Adopt one-hot encoding for categorical variables.

**Scaling:** Normalize numerical features using Min-Max scaling or standardization.

**2. Model Selection:**

**Baseline Models:** Logistic regression for simplicity and decision trees for interpretability.

**Advanced Models:** Random forests and XGBoost for robust performance. Stacking models for integrating predictions, analogous to Kubernetes' cluster federation.

**3. Hyperparameter Tuning:**

Perform grid search and cross-validation to optimize learning rates, max depths, and subsampling parameters. For instance:

Learning rate: [0.01, 0.1, 0.2].

Max depth: [3, 5, 7].

Subsample: [0.6, 0.8, 1.0].

**4. Model Evaluation:**

Metrics include accuracy, precision, recall, F1-score, and AUC-ROC, ensuring comprehensive model assessment.

**5. Statistical Analysis:**

**Chi-Square Test:** Evaluate significance of categorical features like subscription type.

**ANOVA:** Analyze numerical variables like monthly charges.

**Correlation Analysis:** Address multicollinearity in features.

**6. Implementation Plan:**

Conduct exploratory data analysis using Python libraries to uncover patterns and correlations.

Train models with 70% of the data and validate using k-fold cross-validation.

**4.Tools and Resources:**

1. **Software**: Python libraries like pandas, numpy, scikit-learn, XGBoost, and seaborn.

2. **Hardware**: Leverage high-performance computing, similar to Kubernetes’ resource efficiency models, via AWS or Google Colab GPUs.

**5.Anticipated Challenges and Solutions:**

**1.** **Imbalanced Datasets:** Address with oversampling (SMOTE) or undersampling techniques.

**2.** **High-Dimensional Data:** Use recursive feature elimination for efficient feature selection.

**3.** **Real-Time Adaptation:** Draw parallels with Kubernetes’ real-time resource scheduling to improve the system’s adaptability.

**6. Anticipated Findings and Outcomes**

**1. Findings:**

Ensemble models (e.g., XGBoost) are expected to outperform traditional methods.

Key churn drivers include subscription type, billing frequency, and usage patterns.

**2. Business Implications:**

Enable precise, targeted retention campaigns.

Reduce customer acquisition costs by retaining high-value customers.

**7.Conclusion:**

By leveraging advanced machine learning techniques and borrowing ideas from orchestration frameworks like Kubernetes, this research aims to address critical challenges in churn prediction. Ensemble models, robust feature engineering, and real-time adaptability will provide actionable insights, enabling businesses to proactively manage customer retention. Future exploration will focus on integrating real-time predictions and expanding across diverse industries.

**Resource Index:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Resource No.** | **URL** | **Title** | **Year** |
| 1. Research Gate | https://www.kaggle.com | Telecom Churn Dataset | 2022 |
| 2. Research Gate | https://www.researchgate.net | Ensemble Methods in Prediction | 2023 |
| 3.IEEE | https://www.ieee.org | Deep Learning and Challenges | 2023 |

**7. References**

1. Kaggle (2023). Telecom Churn Dataset.

2. ResearchGate (2022). Comparative Analysis of Logistic Regression and Decision Trees.

3. Springer (2023). Ensemble Learning for Prediction.

4. Arxiv (2024). Real-Time Predictive Models.

5. IEEE Xplore (2023). Challenges in Deep Learning for Customer Retention.