Research Title.

An Investigation into Machine Learning Methods that Most Accurately Predict Customer Churn in the Telecom Industry: A British Context.

Abstract.

In today's highly competitive telecommunications industry, predicting and understanding customer churn is crucial for companies to develop effective retention strategies and ensure profitability. This study intends to identify customers most likely to cancel their subscriptions by analysing and forecasting customer turnover. To achieve this, the project will explore machine learning algorithms such as *Decision Tree*, *Random Forest*, *Gradient Boosted Machine Tree*, *Logistic Regression*, and *Extreme Gradient Boosting* to develop a predictive churn model to forecast customer churn in the telecom industry. The performance of these algorithms will be evaluated using metrics such as F1 score and ROC-AUC (receiver operating characteristic curve's area user curve) to ensure the best results. The primary objective of this research is to produce a predictive churn model that can accurately assess customer churn rate in the telecom industry. With this model, companies can proactively retain their customers and improve profitability. By understanding the factors and behaviours contributing to customer churn, telecom companies can stay ahead of the game and ensure their success in the highly competitive telecommunications industry.

Research Problem.

The project intends to significantly contribute to the telecom industry by developing an effective churn prediction model using machine learning techniques. Recent advancements in machine learning (ML) offer promising avenues for enhancing churn prediction models by leveraging large datasets and identifying intricate

patterns that elude traditional statistical methods. However, applying ML techniques in the British telecom sector's churn prediction still needs to be explored, with existing studies providing limited insights into their effectiveness, adaptability, and practical implementation challenges. This research aims to bridge this gap by systematically investigating the application and performance of various ML techniques in predicting customer churn within this context with additional social network analysis parameters.

Research Question.

How effective are machine learning techniques in predicting customer churn in the telecom industry?

Research Aims and Objectives.

This research proposal aims to explore predictive models using machine learning techniques that can accurately assess the customer churn rate of telecommunication companies. The specific objectives of the project are as follows:

1. To analyse the factors and reasons contributing to customer churn in the telecom industry.
2. To build a churn prediction model based on customer service usage history.
3. To develop a new way of feature engineering and selection to enhance the performance of the churn prediction model (Ahmad et al., 2019).
4. To evaluate the performance of different machine learning algorithms (such as Decision Tree, Random Forest, Gradient Boosted Machine Tree, and Extreme Gradient Boosting) in predicting customer churn in the telecom industry.
5. Incorporate social network analysis features into the churn prediction model to enhance its performance further.(Ahmad et al., 2019).

Research Key Literature.

Several studies have focused on using machine learning techniques for customer churn prediction in the telecommunications industry. They found that machine learning algorithms, such as Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Bayesian algorithm, Ensemble learning, Sample-based weight optimisation, and Neural Network, among others, could provide accurate customer churn predictions.

The paper by Ahmad et al. (2019). titled "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform" provides a comprehensive analysis of the importance of customer churn prediction in the telecom industry and highlights machine learning techniques for this purpose. The paper compares the effectiveness of various machine learning algorithms, including the Decision Tree, Random Forest, GBM, and XGBOOST algorithms. The XGBoost algorithm was the most effective, with 92% precision (oversampling), which further analysed the likelihood of churn in specific users. Additionally, the paper presents the concept of incorporating Social Network Analysis features in the churn prediction model, which further enhances its performance.

Another study aimed to predict customer churn and identify factors based on past service usage history (Krishnaveni et al., 2022). They compared the performance of four machine learning algorithms, including Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting. The most prevalent finding from these is that the number of churned customers is significant in the sixth and seventh months but declines in the eighth month. The models were evaluated based on accuracy, but using the F1 score and ROC-AUC would have been a better judge of model performance.

The churn Prediction Model Improvement Using Automated Machine Learning with Social Network Parameters paper published by Marin, M. & Goran, K. 2022, is another crucial paper in this research work. This paper discusses telecom churn prediction using Automated Machine Learning (AutoML) and Social Network Analysis (SNA). The authors propose an enhanced AutoML model with two new social attributes, the Churn Influence of Neighbour Distance One and the Churn Influence of Neighbour Distance Two, to measure the social influence of users who have already churned. Their experiment shows that directly connected users significantly influence telecom churn likelihood more than those indirectly connected. The authors developed two models with SNA measures and two without SNA measures, and the results indicate that incorporating SNA measures into models significantly enhances their accuracy.

The critical papers used in this research include Customer Churn Prediction Using Four Machine Learning Algorithms Integrating Feature Selection and Normalization in the Telecom Sector by Aldalan, A. & Almaleh. A. (2023)., Predicting Customer Churn in Insurance Industry Using Big Data and Machine Learning by Nagaraju, J. et al. (2023).

Research Methodology and Design.

The proposed methodology for this research involves utilising machine learning techniques for churn analysis in the telecommunications industry. Specifically, the following steps will be taken:

1. Data Collection: Gather a dataset from telecommunications companies that includes relevant variables such as customer demographics, usage patterns, service complaints, and churn status.

2. Data Preprocessing: Clean and preprocess the collected data by handling missing values, encoding categorical variables, and scaling numerical variables.
3. Feature Selection: Use methods such as correlation analysis, chi-square tests, and recursive feature elimination to identify the most relevant features for churn prediction.
4. Model Development: Implement and evaluate multiple machine learning algorithms, including Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting, to develop predictive models for customer churn.
5. Model Evaluation: Assess the performance of the developed models using evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve.
6. Cross-validation: Employ cross-validation techniques such as K-fold, Hold-out, and Monte Carlo to validate the performance of the models and ensure their generalizability.
7. Results Analysis: Analyse the results of the developed models and compare their performance to determine the most accurate algorithm for customer churn prediction in the telecommunications industry.
8. Conclusion: Summarise the research findings and discuss their implications for the telecommunications industry.

Challenges and Limitations.

Predicting customer churn in the telecommunications industry is a complex task. Various challenges and limitations come with it. Some challenges include dealing with imbalanced datasets, where the number of churned customers is much smaller than that of non-churned customers. Another challenge is handling the high

dimensionality of the data due to many features. Additionally, dealing with noisy or incomplete data can take time and effort. Moreover, the reliability of the predictive models can be affected by external factors such as changes in market conditions or customer behaviour.

In order to improve the accuracy and performance of customer churn prediction models, future research could investigate the use of advanced machine learning techniques, such as deep learning or ensemble methods. Furthermore, incorporating additional external data sources, such as customer social media activity, could offer further insights and enhance the predictive power of the models.

Ethical Consideration and Risk Assessment

The study will thoroughly examine ethical considerations and potential risks in predicting customer churn in the telecom industry.

Ethical considerations: The study will strictly follow customer privacy and data protection regulations when using customer information for churn prediction. It will also address potential biases from historical customer data and ensure fairness in implementing the churn prediction model.

Potential risks: While there is a risk of misclassifying customers as churners or overlooking those likely to churn, the study will incorporate thorough validation and evaluation processes to ensure the accuracy and reliability of the churn prediction model.

The study will also confidently consider the potential consequences of relying solely on machine learning algorithms for churn prediction. The study will combine human expertise with algorithmic predictions to account for unforeseen events and capture all relevant factors, resulting in more reliable results. Furthermore, the study will also

confidently consider the potential impact on customer trust and loyalty when using churn prediction algorithms.

Description of Artefacts

The artefacts generated from this study would include a predictive churn model using machine learning algorithms, such as Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting.

Timelines of proposed activities

The proposed activities for this study would include the following timelines:

- Week 1-2: Background Research and Literature Review: Review existing literature on customer churn prediction in the telecom industry and machine learning techniques for predictive modelling.
- Week 3-4: Data Collection: Gather a dataset from telecommunications companies that includes relevant variables such as customer demographics, usage patterns, service complaints, and churn status.
- Week 5-6: Data Preprocessing: Clean and preprocess the collected data by handling missing values, encoding categorical variables, and scaling numerical variables.
- Week 7-8: Feature Selection: Use methods such as correlation analysis, chi-square tests, and recursive feature elimination to identify the most relevant features for churn prediction.
- Week 9-10: Model Development: Implement and evaluate multiple machine learning algorithms, including Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting, to develop predictive models for customer churn.

- Week 11-12: Model Evaluation: Assess the performance of the developed models using evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve.
- Week 13-14: Cross-validation: Employ cross-validation techniques such as K-fold, Hold-out, and Monte Carlo to validate the performance of the models and ensure their generalizability.
- Week 15: Results Analysis: Analyse the results of the developed models and compare their performance to determine the most accurate algorithm for customer churn prediction in the telecommunications industry.
- Week 16: Conclusion: Summarise the research findings and discuss their implications for the telecommunications industry. Prepare the final report and presentation for dissemination.
- Week 17: Writing: Complete a full thesis draft and meet with the supervisor to discuss feedback and revisions
- Week 18-20: Revision: Redraft based on feedback, Get supervisor approval for final draft, proofread, print, bind and submit.