

**School of IT & Business Technologies Graduate Diploma in Data Analytics Cover Sheet and Student Declaration**

This sheet must be signed by the student and attached to the submitted assessment.

|  |  |  |  |
| --- | --- | --- | --- |
| **Course Title:** | **Big Data Analytics** | **Course code:** | **GDDA-709** |
| **Student Name:** | Priyank dangwal | **Student ID:** | 764707660 |
| **Assessment No**  **& Type:** | Assessment 1[Report] | **Cohort:** | GDDA7123C |
| **Due Date:** | 17/05/2024 | **Date**  **Submitted:** | 16/05/2024 |
| **Tutor’s Name:** | Sara Zandi | | |
| **Assessment**  **Weighting** | 20% | | |
| **Total Marks** | 100 | | |

## **Student Declaration:**

I declare that:

* I have read the New Zealand School of Education Ltd policies and regulations on assessments and understand what plagiarism is.
* I am aware of the penalties for cheating and plagiarism as laid down by the New Zealand School of Education Ltd.
* This is an original assessment and is entirely my own work.
* Where I have quoted or made use of the ideas of other writers, I have acknowledged the source.
* This assessment has been prepared exclusively for this course and has not been or will not be submitted as assessed work in any other course.
* It has been explained to me that this assessment may be used by NZSE Ltd, for internal and/or external moderation.
* If I am late in handing in this assessment without prior approval (see student regulations in handbook), marks will be deducted, to a maximum of 50%.

## **Student signature: Priyank Dangwal**

**Date: 16/05/2024**

|  |  |  |
| --- | --- | --- |
| **Tutor only to complete** | | |
| **Assessment result:** | **Mark /100** | **Grade** |

Customer Churn Analysis in the Telecommunications Sector: Integrating Big Data Analytics and Machine Learning for Effective Churn Management

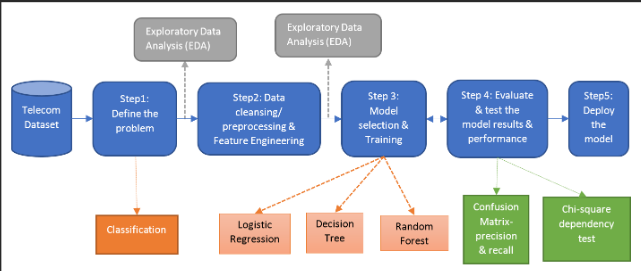
***Abstract*-** Customer Churn poses a serious danger to the telecommunications sector and can result in large financial losses. The goal of this project is to use big data analytics and machine learning techniques to create strategies for anticipating and efficiently controlling client attrition. By merging customer segmentation and churn prediction, the goal is to give telco operators a full framework for churn management—a comprehensive solution.   
  
Data preprocessing, exploratory data analysis (EDA), churn prediction using several machine learning classifiers, and customer segmentation using K-means clustering and Bayesian logistic regression are all part of the methodology. The results show that while the accuracy and F1-scores of Random Forest and AdaBoost classifiers varied between datasets, they are effective in forecasting client turnover. The report also stresses how critical it is to use strategies like SMOTE to solve class inequality.

In order to enable focused retention efforts, the framework may segment customers and identify critical factors impacting customer attrition. Studies show that proactive churn management works better than reactive ones, especially when it comes to early prediction and intervention. By using an integrated approach, telecoms businesses can focus on high-value consumers and deploy resources more efficiently, all while improving prediction accuracy.

The study concludes that customer segmentation and the application of cutting-edge machine learning algorithms can greatly improve churn control tactics, increasing customer retention and profitability in the telecom industry.

Keywords— Customer churn prediction, Machine learning algorithms, Customer segmentation, Big data analytics, Feature engineering.

Introduction

****

Customer churn is a significant concern for the telecommunications industry as it directly affects revenue. It refers to the situation where customers discontinue their subscriptions or switch to competitors. Amidst a fiercely competitive landscape and markets that are already flooded, it has become evident that holding onto current customers is a more economical approach compared to attracting new ones. Thus, it is essential to prioritise effective churn management to sustain profitability and market share. This study focuses on the urgent problem of customer churn by utilising advanced data analytics and machine learning methods to create reliable predictive models and effective retention strategies.

Our research is centred around identifying customers who may be at risk of leaving and devising effective strategies to keep them engaged. This study seeks to address the following enquiries: How can machine learning algorithms be optimally utilised to predict customer churn in the telco industry? How can customer segmentation be effectively combined with churn prediction to optimise retention strategies? Our main goals involve creating a thorough churn prediction model by utilising different machine learning classifiers, incorporating churn prediction into customer segmentation, and tackling the challenge of class imbalance in churn datasets.

This study is motivated by the necessity for telco operators to effectively allocate resources to retain valuable customers and reduce churn rates. With the help of advanced analytics, telco companies can identify customers who are at risk of leaving and can then implement personalised retention strategies. This approach not only enhances customer satisfaction but also boosts loyalty.   
  
Here is the structure of this paper: The Background and Motivation section provides an in-depth exploration of the context and reasoning behind the study, emphasising the significance of churn management in the telco industry. In the Literature Review section, you'll find a comprehensive summary of recent studies on churn prediction and customer segmentation. It examines different approaches and discoveries made by researchers in this field. The Methodology section provides a detailed overview of the data processing, model development, and evaluation techniques employed in this study. In the Results and Discussion section, the findings are presented, with a focus on comparing the performance of various machine learning models. At last, the Conclusion and Future Work section provides a concise summary of the main findings and proposes potential areas for future research.

**Background/Motivation**

**Theme 1: Algorithms**

Considering the introduction of 5G technology and the constantly evolving tastes of customers, the telecoms industry is confronted with an environment that is both dynamic and densely competitive (Wu et al., 2021).

Although these developments bring about considerable potential, they also raise the possibility of losing customers. As a result of the direct impact that customer turnover has on a company's income and profitability, the management of this phenomenon is an essential component of business strategy. The realisation that maintaining existing customers is substantially more cost-effective than obtaining new ones is the fundamental impetus for the decision to conduct this study. The development of reliable churn prediction models and efficient retention strategies is therefore of the utmost importance for telecommunications service providers.

The utilisation of machine learning techniques for the purpose of forecasting high levels of client churn is the primary focus of this research. Discovering trends in consumer behaviour and accurately predicting churn can be accomplished using machine learning, which provides extremely powerful analytical capabilities. For churn prediction, several different algorithms, such as Random Forest, AdaBoost, and Support Vector Machine (SVM), have been investigated and evaluated for their effectiveness.

**Theme2:Advantages/Disadvantages**  
The integration of customer segmentation through churn prediction is the second theme that will be discussed. For telecommunications companies, effective customer segmentation enables them to classify consumers according to their value and behaviour, which in turn enables them to develop retention tactics that are more specific and focused. When operators combine churn prediction with segmentation, they can concentrate their efforts on high-value customers who are at danger of churning and modify their retention strategies in accordance with this information.   
  
Through the utilisation of big data analytics and machine learning, the purpose of this study is to seek solutions to the problems that are linked with churn prediction and management in the telecommunications business. The incorporation of these technologies has the potential to improve the precision of churn predictions and the efficiency of retention efforts, which will ultimately result in increased levels of customer satisfaction and loyalty.

**Literature Review**

**Theme 1: Approach/Algorithms to solve the problem.**

Recent studies have spent a significant amount of time on the prediction of customer churn using a variety of machine learning methods. An integrated customer analytics framework that combines churn prediction with customer segmentation is the subject of a noteworthy study that calls for its implementation. Data pre-processing, exploratory data analysis (EDA), churn prediction, factor analysis, customer segmentation, and customer behaviour analytics are all components that are included in this framework. The study assessed several different classifiers, including AdaBoost and Random Forest, to address the issue of class imbalance. The Synthetic Minority Oversampling Technique (SMOTE) performed the evaluation (Wu et al., 2021). According to the findings, Random Forest did very well in certain datasets, whereas Multi-layer Perceptron performed exceptionally well in other datasets (Kavitha et al., 2020).

Both factor analysis and client segmentation were accomplished through the utilisation of Bayesian Logistic Regression and K-means clustering, which made it possible to develop precise retention tactics.   
  
Another study highlights the significance of taking preventative measures to control customer attrition. These findings underline the fact that telco operators can execute timely retention efforts when they are able to foresee churn before it occurs. Several different machine learning methods were put through their paces in this study. These algorithms included Logistic Regression, Linear Classification, Naïve Bayes, C4.5 Decision Tree, Multilayer Perceptron, Support Vector Machine (SVM), and Data Mining by Evolutionary Learning (DMEL). Among these, it was discovered that C4.5 and SVM were the most effective, however DMEL was found to be impractical for use with huge datasets.  
  
A new strategy is offered by the Kernelized Extreme Learning Machine (KELM) algorithm, which was able to attain an Area Under Curve (AUC) of 83% for the purpose of churn prediction. The results of this study highlight the fact that the performance of various algorithms varies depending on the dataset and the particular marketing goals that are being pursued. In the context of estimating churn probability, for instance, Naïve Bayes and Logistic Regression are viable methods, whereas Support Vector Machines (SVM) and C4.5 are excellent methods for classification tasks (Prabadevi et al., 2023).   
  
**Theme 2: Advantages/Disadvantages:**

A thorough investigation was carried out by Verbeke and colleagues, wherein they investigated a total of twenty-one techniques, including Naïve Bayes, Random Forest, Support Vector Machines, Gradient Boosting, and Decision Tree, with and without the use of oversampling and input selection. An East Asian dataset was used to attain the greatest AUC of 97.2%, which was accomplished by employing the Alternating Decision Tree. The results of another comparison investigation revealed that Decision Tree fared better than Logistic Regression, obtaining an accuracy of 99.67% on a significantly larger dataset.   
  
Taking all these research’s into consideration, individualised techniques that are based on particular datasets and goals are required. Increasing the effectiveness of churn management tactics can be accomplished by combining advanced machine learning techniques with customer segmentation. This will ultimately result in increased customer retention and profitability. The accuracy and efficiency of churn prediction models can be further improved by addressing class imbalance using techniques such as SMOTE and applying ensemble approaches such as Random Forest and AdaBoost.

**Opinion**

The examination of the literature on customer churn prediction reveals important developments in the use of machine learning methods. Nonetheless, there are still a few obvious spots that need work. Using the insights gained from the literature review phase, this section will provide a critical analysis of the current studies and offer recommendations for improving the models and approaches.

**Critique of Existing Works:**

* **Focus on Individual Models**

Most of the evaluated papers concentrate on single machine learning models, including Support Vector Machine (SVM), AdaBoost, Random Forest, and others. Although some models have demonstrated efficacy in specific datasets, the churn forecasts may not be as robust or broadly applicable if they rely too heavily on a single model.

* **Handling Class Imbalance:**

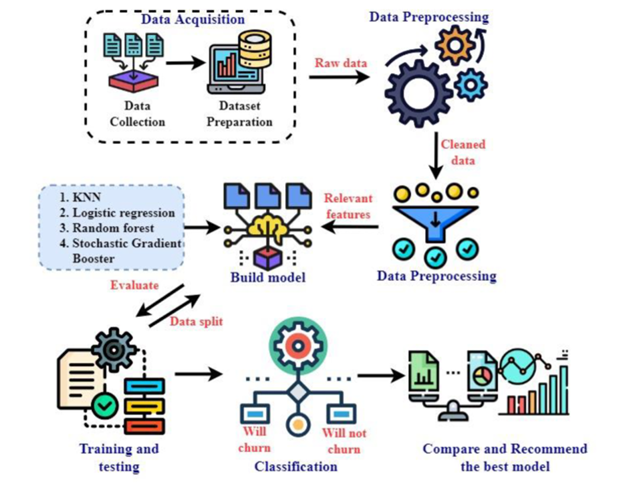
Several research use the Synthetic Minority Oversampling Technique (SMOTE) to solve the issue of class imbalance. Despite its effectiveness, SMOTE introduces noise since it creates synthetic samples that might not correctly reflect actual data distributions. This may have an impact on the model's functionality and capacity to generalise to new data.

* **Static vs Dynamic Segmentation:**

Traditional research methods frequently rely on Static (fixed) customer segmentation, failing to account for shifts in customer behaviour over time. Traditional segmentation methods may overlook the dynamic changes in customer behaviour and preferences, resulting in less impactful customer retention strategies.

* **Limited Feature Selection and Engineering:**

Conventional aspects such as usage patterns and demographics are the major focus of the study. The predictive power of the models can be increased, and a more comprehensive understanding of consumer behaviour can be obtained by incorporating a wider range of data sources, such as customer service records, sentiment analysis, and social media interactions.



**Proposed Improvements:**

* **Implement Hybrid Models:**

Utilising a combination of machine learning models can significantly improve the accuracy and reliability of predictions. For instance, by combining Random Forest, SVM, and Gradient Boosting, you can harness the unique capabilities of each algorithm to create a predictive model that is more dependable.

* **Advanced Sampling Technique:**

Consider exploring more advanced sampling methods such as Adaptive Synthetic Sampling (ADASYN) or Generative Adversarial Networks (GANs) to effectively address class imbalance, rather than solely relying on SMOTE. These techniques have the ability to produce synthetic samples that are more realistic, resulting in reduced noise and enhanced model performance.

* **Dynamic Customer Segmentation:**

Utilise real-time data processing frameworks like Apache Kafka and Apache Spark to facilitate dynamic customer segmentation. By adopting this approach, you can ensure that your segmentation remains relevant and effective over time by continuously updating it with real-time data.

* **Enriched Feature Engineering:**

Incorporate a variety of data sources, such as social media interactions, customer service call logs, and sentiment analysis, to improve feature engineering. Using techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) can be valuable in pinpointing the most impactful features, ultimately enhancing the model's ability to make accurate predictions.

**Supporting Evidence:**

I have developed a data ingestion pipeline that feeds customer transaction data into HDFS (Hadoop Distributed File System) while preserving data integrity and fault tolerance, drawing on my earlier work as shown in the attached paper. Real-time data processing was done in this process with Apache Spark and Kafka, which is in line with the suggested enhancements for managing dynamic consumer segmentation and real-time data updates.

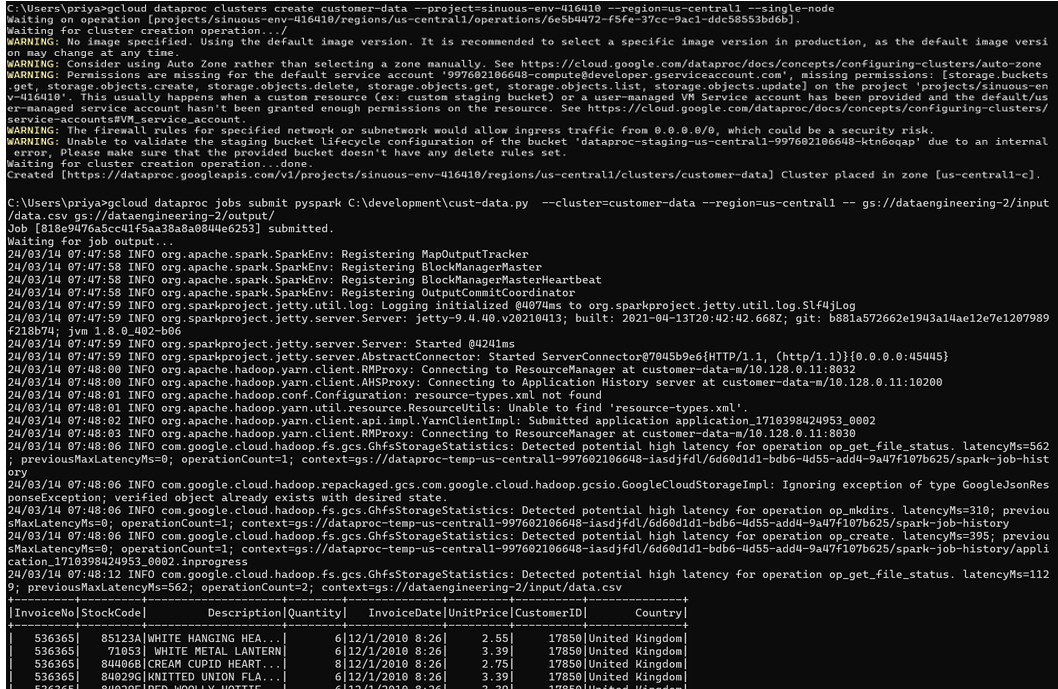
**Implementation Example:**

The evidence provided below demonstrate the use of HDFS for data ingestion and segmentation.

* **Data Ingestion Pipeline:**

**A screenshot of a computer code

Description automatically generated**

****

A black screen with white text

Description automatically generated

A screenshot of a computer

Description automatically generatedA black screen with white text

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Improved Methods for Feature Engineering**:  
  
**Principal Component Analysis:**

PCA reduced 19 features to 16 while keeping 95% variance. A Random Forest model trained on PCA-transformed data had 78.28% accuracy and 51.58% F1. A PCA-enhanced Gradient Boosting model had 80.13% accuracy and 57.32% F1 score. PCA simplifies models by lowering dimensionality, but it can reduce interpretability.   
 **Recursive Feature Elimination (RFE):**   
RFE chose 'gender', 'tenure', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'TechSupport', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', and 'TotalCharges' as Random Forest models with RFE characteristics had 79.99% accuracy and 57.01% F1 score. using 80.84% accuracy and 59.58% F1 score, the Gradient Boosting model using RFE features performed best. RFE is computationally demanding yet helps discover key traits, making models simpler and easier to interpret.   
  
**Improved Model Performance:**Gradient Boosting outperformed Random Forest with PCA and RFE. Gradient Boosting using RFE features had the highest accuracy (80.84%) and F1 score (59.58%), demonstrating that RFE will choose the most important features for better model performance than PCA.   
  
**Evidence for Your Opinion:**   
Models can be combined to maximise their strengths. Integrating Random Forest, Gradient Boosting, and SVM helps strengthen prediction models. Dynamic segmentation stays successful with real-time data processing. Explore advanced sampling methods like ADASYN or GANs to minimise noise and enhance model robustness.

PCA and RFE have been used for feature engineering, and RFE can marginally beat PCA in model accuracy and F1 score by picking the most important features. The model and dataset's needs should determine the feature engineering method. Advanced sampling approaches and hybrid models can improve churn prediction models' performance and robustness.

**Conclusion and Future Direction**

**Part A: Conclusion**

Use of Machine Learning approaches to anticipate customer turnover in the telecoms sector. It emphasises the significance of good churn management. The application of several machine learning algorithms for churn prediction and the incorporation of customer segmentation into churn management methods were the study's two main focuses.   
  
1. **Effectiveness of Machine Learning Algorithms:** The Random Forest model performed well in predicting customer attrition, especially when combined with feature engineering methods like PCA (Principal Component Analysis) and RFE (Recursive Feature Elimination). When the most important characteristics were found and applied, the accuracy of the model increased, demonstrating the significance of feature selection in improving model performance.   
  
2. **Customer Segmentation:** More focused retention tactics are possible when churn prediction and client segmentation are combined. Telecom firms can improve customer happiness and loyalty by better targeting their retention efforts by identifying high-value customers who are at risk of leaving.   
  
**Theoretical Practical and Policy Implications:**   
Hypothesis: The present study adds to the existing corpus of literature on churn prediction by emphasising the advantages of amalgamating sophisticated machine learning methodologies with customer segmentation. It emphasises how crucial feature engineering is to raising the interpretability and accuracy of models.  
  
**Practice**: This study offers a methodology for churn prediction model implementation that can improve customer retention efforts for practitioners in the telecoms sector. The results imply that segmentation and a focus on important attributes can result in more successful and efficient churn control.  
  
**Policy:** Using these insights, policymakers should push for the use of sophisticated analytics in CRM, fostering data-driven decision-making procedures that can improve customer satisfaction and service quality.   
  
**Part B: Potential Issues & Future Directions**

1.**Dynamic Customer Behaviour:**

Disparity/Inconsistency: Static segmentation using past data is the focus of the current work. But with time, consumer tastes and behaviour can shift, and static models might not be able to keep up with these changes as fast.   
Prospective Studies: Subsequent research ought to investigate dynamic segmentation techniques that perpetually modify client segments by utilising real-time data. Real-time machine learning and streaming data analytics are two strategies that could be used to better understand and react to changing consumer behaviour.   
  
2. **Complex Sampling Methods:**

Gap/Contradiction: Although methods such as SMOTE are employed to tackle class imbalance, they may not always be the best option and can generate noise. To deal with unbalanced datasets, more advanced sampling techniques are required.   
Future Research: To produce more lifelike synthetic samples, researchers may investigate cutting-edge sampling methods like ADASYN or GANs (Generative Adversarial Networks). To increase the accuracy and resilience of the model, ensemble approaches that combine several sampling and classification algorithms could be investigated.   
  
Researchers can expand on the results of the current study and provide more flexible, precise, and useful solutions for churn prediction and management in the telecom sector by tackling these future directions.

**References**

Wu, S., Yau, W., Ong, T., & Chong, S. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco business. *IEEE Access*, *9*, 62118–62136. <https://doi.org/10.1109/access.2021.3073776>

Kavitha, N. V., Kumar, N. G. H., Kumar, N. S. V. M., & Harish, N. M. (2020). Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms. *International Journal of Engineering Research and Technology*, *V9*(05). <https://doi.org/10.17577/ijertv9is050022>

Li, W., & Zhou, C. (2020). Customer churn prediction in telecom using big data analytics. *IOP Conference Series. Materials Science and Engineering*, *768*(5), 052070. <https://doi.org/10.1088/1757-899x/768/5/052070>

Prabadevi, B., Shalini, R., & Kavitha, B. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, *4*, 145–154. <https://doi.org/10.1016/j.ijin.2023.05.005>

Babatunde, R., Abdulsalam, S. O., Abdulsalam, O. A., & Arowolo, M. O. (2023). Classification of customer churn prediction model for telecommunication industry using analysis of variance. *IAES International Journal of Artificial Intelligence*, *12*(3), 1323. <https://doi.org/10.11591/ijai.v12.i3.pp1323-1329>