VIETNAM GENERAL CONFEDERATION OF LABOUR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**HOANG QUOC VINH – 519H0259**

**NGUYEN ANH KHOA – 519H0303**

**HANDLING IMBALANCED DATA IN THE PREDICTION OF CUSTOMER CHURN**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**COMPUTER SCIENCE**

**HO CHI MINH CITY, 2024**

VIETNAM GENERAL CONFEDERATION OF LABOUR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**HOANG QUOC VINH – 519H0259**

**NGUYEN ANH KHOA – 519H0303**

**HANDLING IMBALANCED DATA IN THE PREDICTION OF CUSTOMER CHURN**

**UNDERGRADUATE THESIS OF/ PROJECT REPORT OF**

**COMPUTER SCIENCE**

Advised by

**Prof. Trinh Hung Cuong**

**HO CHI MINH CITY, YEAR 2024**

**ACKNOWLEDGMENT**

I sincerely thank

*Ho Chi Minh City, day month year 20*

*Author*

*(Signature and full name)*

This thesis was carried out at Ton Duc Thang University.

Advisor: ............................................................................................

............................................................................................

*(Title, full name and signature)*

This thesis/project report is defended at the Undergraduate Thesis Examination Committee was hold at Ton Duc Thang University on … /…/……

Confirmation of the Chairman of the Undergraduate Thesis Examination Committee and the Dean of the faculty after receiving the modified thesis (if any).

**CHAIRMAN DEAN OF FACULTY**

**…………………………. ………………………………**

DECLARATION OF AUTHORSHIP

I hereby declare that this thesis was carried out by myself under the guidance and supervision of ………………………………………..; and that the work and the results contained in it are original and have not been submitted anywhere for any previous purposes. The data and figures presented in this thesis are for analysis, comments, and evaluations from various resources by my own work and have been duly acknowledged in the reference part.

In addition, other comments, reviews and data used by other authors, and organizations have been acknowledged, and explicitly cited.

**I will take full responsibility for any fraud detected in my thesis**. Ton Duc Thang University is unrelated to any copyright infringement caused on my work (if any).

*Ho Chi Minh City, day month year*

*Author*

*(signature and full name)*

**XỬ LÍ DỮ LIỆU KHÔNG CÂN BẰNG TRONG DỰ ĐOÁN KHÁCH HÀNG RỜI ĐI**

**TÓM TẮT**

(Time New Romans – 13)

**HANDLING IMBALANCED DATA IN THE PREDICTION OF CUSTOMER CHURN**

**ABSTRACT**

(Time New Romans – 13)

**TABLE OF CONTENT**

[Chapter 1. INTRODUCTION 1](#_Toc172808608)

[1.1 REASON FOR CHOOSING TOPIC 1](#_Toc172808609)

[1.2 SCOPE AND OBJECTIVES 1](#_Toc172808610)

[CHAPTER 2. HANDLE THE IMBALANCED DATA PROBLEM 2](#_Toc172808611)

[2.1 CLASS IMBALANCE PROBLEM 2](#_Toc172808612)

[2.1.1 Problem Definition 2](#_Toc172808613)

[2.1.2 Real-World Example 2](#_Toc172808614)

[2.2 FEATURE SELECTION 2](#_Toc172808615)

[2.3 EVALUATION METRICS FOR IMBALANCED CLASSES 2](#_Toc172808616)

[2.4 TECHNIQUES FOR HANDLING IMBALANCED CLASSES 2](#_Toc172808617)

[CHAPTER 3. LITERATURE REVIEW 3](#_Toc172808618)

[CHAPTER 4. METHODOLOGY 4](#_Toc172808619)

[CHAPTER 5. EXPERIMENTAL RESULTS 5](#_Toc172808620)

[CHAPTER 6. CONCLUSION 6](#_Toc172808621)

[REFERENCES 7](#_Toc172808622)

**LIST OF FIGURES**

[Hình 2.1: Scaled Dot-Product Attention 3](#_Toc142677545)

LIST OF TABLES

[Bảng 4.1: Thống kê kiểu thực thể trong tập VLSP 2016 4](#_Toc142677565)

ABBREVIATIONS

|  |  |
| --- | --- |
| BERT | Bidirectional Encoder Representations from Transformers |
| GEC | Grammatical Error Correction |
| MLM | Masked Language Model |
| NLP | Natural Language Processing |
| NSP | Next Sentence Prediction |

# Chapter 1. INTRODUCTION

## REASON FOR CHOOSING TOPIC

Customer churn is an important term which refers to the rate of customers dropping off a company’s products or cancelling subscription within a specific period of time. This has become a crucial priority in businesses’ operational performance, especially in subscription business models. The ability of a company to turn new customers into repeat buyers and prevent them from switching to a rival competitor is called customer retention. This process requires more marketing expenses, time and effort than retaining existing customers. Therefore, high-rate level of customer churn can negatively impact revenue, profitability, and overall business growth. This topic is significantly relevant in today’s competitive business environment. Those companies can apply effective customer churn prediction strategies to minimise churn, are the winners. Understanding the context, the aim of this research is to identify existing and new efficient methods to handle imbalanced data which is a obstacle in predicting churn techniques.

## SCOPE AND OBJECTIVES

To demonstrate ‘Handling imbalanced data in customer churn’ topic, this research focuses on the telecommunications sector and a publicly available dataset.

# CHAPTER 2. HANDLE THE IMBALANCED DATA PROBLEM

## 2.1 CLASS IMBALANCE PROBLEM

### 2.1.1 Problem Definition

Class imbalance is one of the challenges in machine learning and data mining fields. This problem occurs when the instances of one class significantly outnumber the instances of other classes. The class has larger number of instances called the majority class, while the smaller one is minority class. The imbalance can lead to poor predictive results of machine learning models because of high accuracy for the majority class and poor accuracy for the minority class.

### 2.1.2 Real World Applications

* Mail Services: detect spam or not.

## 2.2 FEATURE SELECTION IN CLASS IMBALANCE PROBLEM

## 2.3 EVALUATION METRICS FOR IMBALANCED CLASSES

## 2.4 TECHNIQUES FOR HANDLING IMBALANCED CLASSES

# CHAPTER 3. LITERATURE REVIEW

# CHAPTER 4. METHODOLOGY

# CHAPTER 5. EXPERIMENTAL RESULTS

# CHAPTER 6. CONCLUSION

# REFERENCES

Tiếng Việt

…

Tiếng Anh

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762