C03051: Master of Analytics
CRICOS Code: 075277F
13389465 Masters Thesis: Analytics
December 2023

A Study on

Approaches To Developing Customer Risk Profile For Underwriting Using Knowledge Graph And Multilabel Classification

Khanh Van Nguyen

School of Computer Science
Faculty of Engineering & IT
University of Technology Sydney
NSW - 2007, Australia

A Study on

Approaches To Developing Customer Risk Profile For Underwriting Using Knowledge Graph And Multilabel Classification

A thesis submitted in partial fulfilment of the requirements for the degree of

Master by Research

in

Analytics

by

Khanh Van Nguyen
Under the supervision of Prof. Guandong Xu and Dr. Huan Huo.

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

December 2023

AUTHOR'S DECLARATION

I, *Khanh Van Nguyen* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Master in Analytics, in the *School of Computer Science*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia, is my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note:
Signature removed prior to publication.

[Khanh Van Nguyen]

DATE: 1st Dec, 2023

PLACE: Sydney, Australia

Abstract

As the insurance industry expands into new markets and mass customers, the volume of data required to process with regards to a customer's mortality risk has become significantly exponential. To alleviate this issue, the companies in the insurance industry have adopted insurance technology (InsurTech), in the industrial project connected to this study, the company implemented an underwriting rule engine (URE), which functions as a rule-based system simplifying the underwriting rules to a hardcoded decision tree model. This tree flow model guides users through a series of questions, ultimately reaching the final outcome leaf as either a preimium loading or exclusion code. However, as this process is hardcoded, it becomes less optimal and lacks personalisation for individual customers. To tackle these issues, this study proposes the development of an Underwriting Knowledge Graph (UKG) that integrates all customer information received with the existing underwriting manual and explainable exclusion code system to formulate personalised customer risk profiles. The UKG utilises interconnected information pertaining to the customers, including historical data, and explainable exclusion to create customer risk profile, while offering valuable insights and potential correlations with specific exclusion codes. The study's deep dive into insurance domain-specific requirements, current research limitation within automated underwriting has led to the creation of the first-ever UKG, trained on real-life underwriting data thanks to the collaboration with our Australian industry partner. In addition to the UKG, the study also introduces a semi-automated novel method for maintaining the UKG. This method factors in the multi-label classification nature of the data outcome to provide explainable exclusion. The UKG as a data structure provides a comprehensive understanding of the insurance ecosystem, facilitates a representation of information on customer risk profiles, and enables explainable exclusion classifications. While this study focuses on its application in insurance, the UKG and application of graph databases hold promises in enhancing risk assessment and decision-making for other personalised services beyond the realm of life insurance.

This study explores the development of customer risk profiles in life underwriting by integrating underwriting knowledge graphs and explainable exclusion multi-label classification. Traditional underwriting processes, though semi-automated, still heavily rely on manual intervention, leading to inconsistencies and biases. Hence, this research identifies challenges such as handling missing values, adopting a more data-driven approach, and ensuring model explainability. The proposed methodology involves constructing underwriting knowledge graphs, implementing multi-label classification for explainable exclusions, and providing transparency on feature impact. Contributions include empirical application of real-life data, semi-automated knowledge graph construction, and transparency enhancement in underwriting rules. This thesis is structed to chapters covering data requirements, related works, methodology, and discussion of findings, laying the groundwork for future research in underwriting data innovation.

DEDICATION

To my loved ones, my family, my friends and my supervisors who have supported me in multiple ways and motivate me to complete this journey ...

Acknowledgments

This Master in Analytics journey has been dynamic yet turbulent for me with many new concepts to comprehend at once as well as exciting opportunities of working in industry projects while completing my degree. This study and this thesis can only be completed thanks to the ongoing support from my supervisors, my family as well as my friends here in Sydney. Words can be used to describe the scientific and technological terminologies used in this thesis, yet no amount of words can express fully my gratitude and appreciation to the people that provided help and support along the way.

Specifically, I would like to put this section in dedication to my supervisors and collaborators from UTS and DSMI Lab, naming Professor Guandong Xu, Doctor Huan Huo, Doctor Hongxu Chen and Doctor Md Rafiqul Islam. Their continuous support, insightful guidance and great knowledge sharing throughout the journey has allowed me to proceed with confidence and finding the right track with this study. Specifically, I would like to thank Doctor Huan Huo personally for her immense support and guidance both in the academic aspect via the feedback during our meetings and personal growth aspect. I would also like to express my gratitude to the Faculty of Engineering and Information Technology (FEIT) and Zurich Financial Services for the financial support via scholarship that allows me to complete this course.

In addition, I would like to thank my peer candidates at the lab for providing valuable and timely comments on my research progress, as well as the anonymous reviewers that has provided constructive feedbacks to help me adjust my research direction, further improve my research paper and identify the lacking in my methodologies.

LIST OF PUBLICATIONS

RELATED TO THE THESIS :

1. Nguyen, K. V., Islam, M. R., Huo, H., Tilocca, P., & Xu, G. (2023, June). Explainable exclusion in the life insurance using multi-label classifier. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE. **Refer to Chapter 4**

2. Nguyen, K. V., Islam, M. R., Huo, H., Tilocca, P., & Xu, G. (2023, October). Underwriting Knowledge Graph Construction, Maintenance and Its Application on Explainable Exclusion. In The 10th International Conference on Behaviour and Social Computing (BESC 2023). IEEE. **Refer to Chapter 5**

Abbreviations

1. AI: Artificial intelligence
2. AIKG: Auto-insurance knowledge graph
3. InsurTech: Insurance technology
4. ID: Identity
5. KG: Knowledge graph
6. ML: Machine learning
7. OWL: Web ontology language
8. QAR: Quality assurance report
9. RDF: Resource description framework
10. SVM: Support vector machine
11. UKG: Underwriting knowledge graph
12. UP: Underwriting process
13. URE: Underwriting rules engine

Table of Contents

List of Publications ix

List of Abbreviations xi

List of Figures xvii

List of Tables xix

1 Prologue 1

1.1 Background and Motivation 1

1.1.1 Insurance Sector and Life Underwriting 1

1.1.2 Current Industry Headings 2

1.1.3 Underwriting Process and Underwriting Rules Engine 2

1.2 Current Challenges 6

1.3 Research Problems 7

1.4 Research Objectives 7

1.5 Research Contributions 8

1.6 Thesis Organisation 8

2 Related Works 11

2.1 Automated Underwriting Process Approaches 11

2.2 Automated Underwriting Customer Risk Profile 14

2.3 Domain-specific Knowledge Graph 14

2.4 Knowledge Graph Link Prediction Using Multi-label Classification 16

2.5 Explainability in Artificial Intelligence 16

2.6 Summary 16

3 The Underwriting Data 19

3.1 Domain Requirements 19

3.2 Data Set Overview 20
3.3 Data Preliminary Analytics 20
3.3.1 Customers With Exclusions Applied 21
3.3.2 Feature Value Frequencies 23
3.4 Algorithmic Problem 23

4 Explainable Exclusion Using Multilabel Classification 25
4.1 Background and Motivation 25
4.2 Key Contribution 25
4.3 Model Methodology 26
4.4 Evaluation and Explainability 28
4.4.1 Quality assurance report (QAR) 28
4.4.2 Shapley Additive Explanations (SHAP) 29
4.5 Empirical Experiment 29
4.5.1 Data collection and processing 29
4.5.2 Experimental setup 30
4.5.3 Evaluation metrics 30
4.5.4 Evaluation 30
4.6 Results Analysis 31
4.6.1 Explainable machine learning analysis 33

5 Knowledge Graph For Underwriting 37
5.1 Background and Motivation 37
5.2 Key Contribution 38
5.3 Methodology 39
5.3.1 Ontology Design 39
5.4 Construction of Knowledge Graph 42
5.4.1 Exclusion Classification and Link Prediction 42
5.4.2 Automated knowledge graph creation, update and maintenance . . 44
5.5 Results Analysis and Discussion 44

6 Risk Profile Using Knowledge Graph And Multilabel Classification 47
6.1 Background and Motivation 47
6.2 Key Contribution 48
6.3 Revised Methodology 48
6.3.1 Automated UKG Construction 49

6.3.2 Explainable Exclusion 49

6.3.3 Rules Identification and Knowledge Graph Maintenance 50

6.4 Experiment Setup 51

6.5 Results Analysis 52

6.5.1 Customer risk profile using UKG 53

6.5.2 Automatic rules identification after graph adjustments 54

6.5.3 Updated explainable exclusion from graph mining 55

6.6 Summary 59

7 Discussion of Approaches Taken, Use Cases and Future Work 61

7.1 Results Finding 61

7.2 Use Cases 61

7.2.1 Recalculating life table 61

7.2.2 Adding a new customer 63

7.2.3 Adding a new rule 65

7.2.4 Recalculating graph link weight 67

7.2.5 Standardise and transpose UKG into RDF and OWL structure . . 67

7.2.6 Insurance Fraud Prediction 68

7.3 Current Limitations and Future Headings 68

7.3.1 Current Limitations 68

7.3.2 Use of Exclusion Code within Underwriting Knowledge Graph . . 69

7.3.3 Future Headings 70

8 Conclusion 71

Bibliography 73

LIST OF FIGURES

FIGURE **Page**

1.1 URE questionnaire structure 4
1.2 Exclusion decision process 5

2.1 Taxonomy of automated underwriting approaches 12
2.2 AIKG model architecture [70] 15

3.1 Top 20 exclusion codes 21
3.2 Applications of 317 exclusion codes on trauma cover policies 21
3.3 Percentage on total number of questions for most popular question groups (excluding basic details) 23

4.1 Model methodology for explainable exclusion [63] 27
4.2 Top 15 most common exclusion codes - SHAP summary report 34
4.3 Driving factors of the highest applied exclusion code (MHEX) - SHAP detailed report 34

5.1 Underwriting Ontology Workflow Design 38
5.2 Node layers of UKG 39
5.3 Link prediction 41

6.1 Model Methodology 48
6.2 Experiment setup 52

7.1 Proposed Validation Framework 69

List of Tables

Table **Page**

2.1 Key studies: different methods for life insurance label classification in the insurance industry. 13

3.1 Features of sample data attributes 20

3.2 Features of sample data attributes 22

3.3 Preprocessed demo data set for multi-label classification. 24

4.1 Precision, recall, and F-score and hamming loss of the multi-label classification model [63] 31

4.2 Description of quality assurance outcome 32

4.3 Sample of QAR results 33

5.1 Simplified example of knowledge graph triples constructed for UKG in a case of lumbar spine exclusion 43

5.2 Changes between top 20 impact attribute list when using different benchmarks for exclusion code SPIN-LSS 45

6.1 Revised relationships in UKG 49

6.2 Customer risk profile captured by the UKG for a sample customer 53

6.3 Graph edges after each adjustments 54

6.4 Attribute rankings before and after updated UKG for exclusion code MEDI-EYB 57

6.5 Attribute rankings before and after updated UKG for exclusion code MEDI-GOS 59

CHAPTER 1

PROLOGUE

1.1 Background and Motivation

Ever since the 16th century, the insurance industry in general and underwriting profession in particular has undergone substantial evolution, continuously expanding into diverse markets and products. A key aspect of life insurance is the underwriting process, where underwriters assess an applicant's risk to determine their premium payments. Traditionally, this process involved manual, paper-based assessments, making it both labor-intensive and prone to inconsistencies. With the rise of digitalisation and InsurTech, automated underwriting processes and tools like Underwriting Rules Engines (URE) have emerged to streamline the risk assessment process. However, the challenge remains in balancing the need for automation to wrangle millions records of data from multiple sources with the need for human expertise to ensure risk-minimised and unbiased decisions. This chapter explores the intricacies of the life insurance underwriting process, current industry trends, and the challenges posed for the automation process.

1.1.1 Insurance Sector and Life Underwriting

Insurance is one of the oldest industries in the world tracing back to the 16th century with marine insurance integrated to expand the international trade network, from then on it has gone through multiple transformations of product and market expansion with high suggestions to be added into the economic history [33]. Life insurance, as one of the insurance products, [3]. The process of manual life underwriting, as described by Aggour et al. [2], is a heavy paper-based process with multiple documents to be considered to determine an applicant's mortality risk, turning this into a risk management calculation with the output being the customer's premium payment. Since the majority of insurance companies' premium revenue is then spent back on claims [48], underwriters bear the burden of accurately assessing and managing applicant risks through initial and ongoing policy evaluations. The decision-making process follows the rule the underwriting manual, with the risk mitigation output exhibited via two common practices, by applying a penalty (loading) to the standard premium or process the cover with certain constraints (exclusions).

1.1.2 Current Industry Headings

In line with other major economic industries, the insurance industry also has taken digitalisation and technology solutions to automate and improve their existing products and processes to allow for its extension into lower-income brackets and less developed markets [50]. This expansion not only has transitioned insurance from a privileged product to a necessity with easy access, but also extended the customer base, meaning more data and sophisticated risks to process for an already highly personalised product as life insurance. This leads to the inevitable creation of the InsurTech (Insurance Technology) ecosystem [38], including multiple insurance-prone adoptions of artificial intelligence (AI) to develop automated decision-making systems to alleviate current insurance in general as well as underwriting problems specifically [54]. These systems are meant to assist underwriters and insurance experts with understanding the complex layers of information derived from underwriting data points and making decisions on an application to reduce resources put into the procedure. However, compared to banking and finance with regards to economic applied academia fields, insurance has received little research attention [33], as well as numerous suggested models still portray limitations, such as their incapability to adjust to ongoing changes in insurance underwriting policies and their lack of transparency in offering explainability in automated lending decisions [54].

1.1.3 Underwriting Process and Underwriting Rules Engine

1.1.3.1 The Underwriting Process

The traditional underwriting process for insuring a person starts with a customer sending an application to the insurance company for the human underwriters to review [2]. Underwriters then determine a customer's risks based on the rules from the underwriting manual in combination with leveraging the experts' knowledge through medical records and personal experience from previous assessed cases. However, Biddle et al. [5] has pointed out a possibility for inconsistencies in different underwriters' decisions, which can lead to inaccurate rate classification. In particular, while applying the same rules from the underwriting manual, the knowledge gaps among underwriters along with the long period of time to fully process the application, with some over 100 pages, can potentially prompt inconsistencies among underwriters. The risk increases with applications requiring more than one exclusion codes applied.

1.1.3.2 Underwriting Rules Engine

Many carriers have applied an underwriting rules engine (URE) with its code based directly from the underwriting manual to tackle the aforementioned disparities [2]. The URE contains a rule base and is set to return a pre-defined result when certain set of conditions are triggered [54], including decision for exclusion and loading depends on how specific the rules and conditions are.

Figure 1.1 portrays a simplified look of the underwriting rules engine structure with multiple question lines (depicted as L) tailed by multiple questions (depicted as Q) belonging to each group along with their answers (depicted as A), from which a specific outcome can be determined, in this case, an exclusion code. Multiple question answers can lead to one exclusion code to be applied, meaning that this result is not fixed to one specific answer.

1.1.3.3 Automated Underwriting Process With URE

The typical exclusion code decision-making process with the use of URE or any insurance decision support system follows the three-step process: data capture, automated decision-making, and human underwriter review, as portrayed in Figure 1.2. The breakdown of each step is as below:

- Data capture: This is the process of gathering customer information from the customer disclosure survey, which is the equivalent of the previously mentioned application. A customer can fill this survey by themselves or with the help of a financial adviser. The survey should be carefully structured so that data capture can be standardised and pre-processed before applying any automated model.

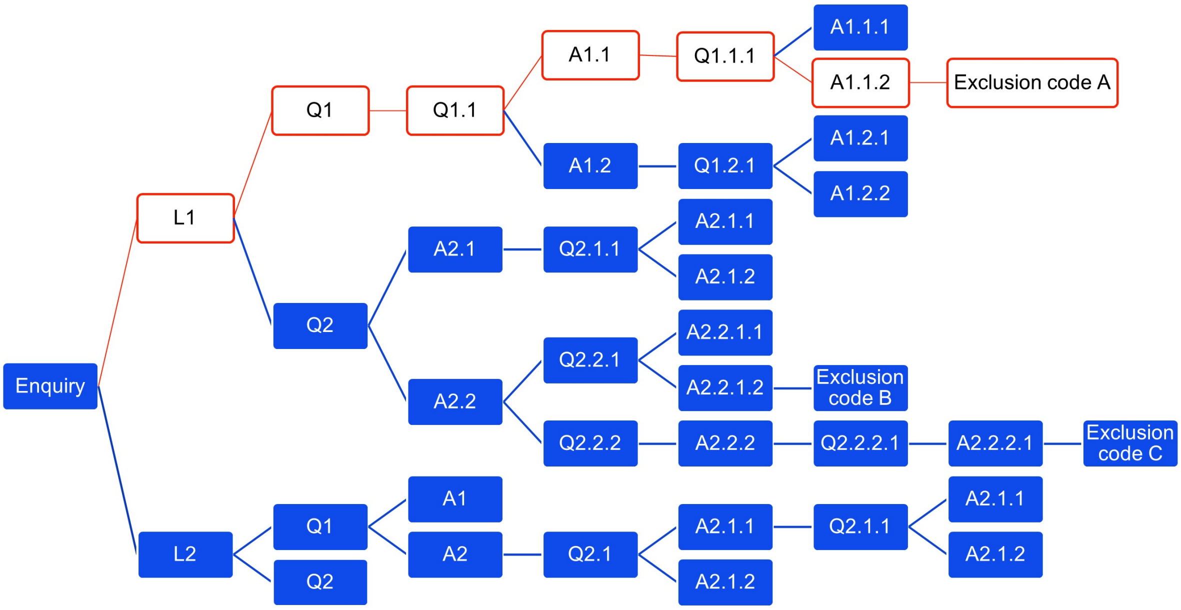


Figure 1.1: URE questionnaire structure

- Automated decision-making: This step is highly enhanced with the Insurtech applications. Particularly, in the case with the insurance industry partner that I have been working with while completing this study, the URE is structured as a multiple tree-based flow diagrams in alignment with the customer disclosure survey, meaning that each question is a node and customer-provided answers are used as branches guide to the follow-up questions. Each "tree" leave is a hard-coded outcome of a chain of choices and answers precedent to it, with this outcome being either to approve, decline or require more information from this application, along with which exact exclusion code should be applied.

- Human underwriter review: In this step, the human underwriters review the output of the automated decision-making model comparing to the survey data collected in the first step to make a decision on the application. The underwriters have full control over whether to follow the model results or override the model with their personal expertise when it comes to exclusion code decisions.

1.1.3.4 Examination of Current Process

The examination of existing processes raises a concern, despite the help of the semi-automated decision-making system as URE, human underwriters must manually input new decision rules and review both customer disclosures and model decisions to finalise the outcome. Although this marks the importance of human expertise within the field, this further confirms the heavy reliance of the existing process on human accuracy, as well as the potential heavy-biased results from existing rules without a full picture of customer risks. The review of the current 3-step business process has prompted my three data-specific concerns as below:

- Identifying domain prerequisites for exclusion analysis and determining how they can be met using a large dataset,
- Designing a multi-label classifier system to capture label-feature relationships, identify opportunities, and compare exclusion classification results between the rules engine and underwriters,
- Developing an approach to explainability in mapping exclusion decisions within the decision-making process.

These questions are further elaborated later in section 1.2.

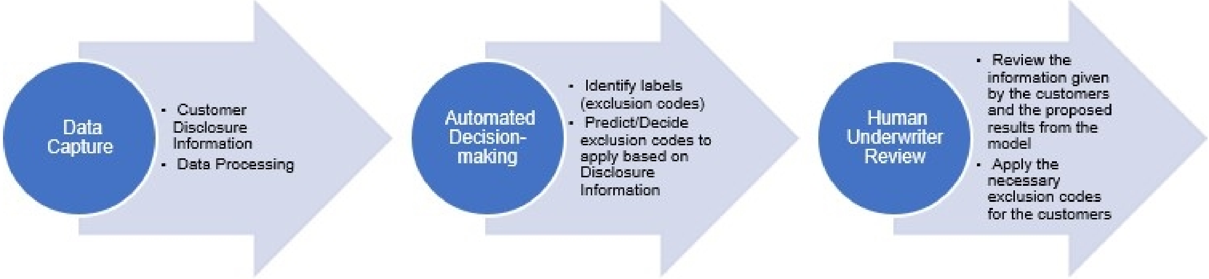

Data Capture
• Customer Disclosure Information
• Data Processing
Automated Decision-making
• Identify labels (exclusion codes)
• Predict/Decide exclusion codes to apply based on Disclosure Information
Human Underwriter Review
• Review the information given by the customers and the proposed results from the model
• Apply the necessary exclusion codes for the customers

Figure 1.2: Exclusion decision process

1.2 Current Challenges

Our investigation into industry practices reveals several challenges in the exclusion classification process. Despite the help of the automated rule system, human underwriters must still review paper-based information. This undermines the aim of a streamlined process where underwriters can easily review customer disclosures and clearly link exclusion codes to risk-related responses. With the questionnaire constantly being updated, the volume of information requiring manual review will continually increase. Embrechts and Wuthrich [13] highlight the necessity for effective optimization tools and customer portfolio analysis over time to sustain long-term financial guarantees in the industry, a goal that remains unfulfilled. The automated rule system resolved classification inconsistencies and standardised decisions, but its simplified logic led to sub-optimal risk classification. The strict rules and lack of personalised underwriting in the URE might not be comparable for the futre of the insurance industry, which must handle larger data volumes and adapt to non-standard situations with insurance rolling out with more product types covering expanded markets and lower-income branches. This oversimplification, although effective for commonly used exclusion codes in routine situations, results in the inadvertent neglect of less common exclusion codes, increasing underwriting risks due to their infrequent consideration [5, 23].

Combining with the results from the data preliminary analysis conducted above, our three current challenges are identified as below:

- More focus needed to understand the domain-specific requirements of the life underwriting (i.e. explainable exclusion code) from a data perspective,

- Handling a large number of missing values from the data survey,

- Handling the trade-off between accuracy and transparency: the current black-boxing of machine learning models on automated underwriting models make it difficult to determine which features have an impact on the results, whereas when applying a tree-diagram approach, the task becomes sub-optimal and potentially hard-coded.

Thus, the following system requirements, in addition to the challenges above, should be put into consideration when developing a decision-making tools for analysing exclusion:

- The system should allow applicable and optimisable changes to tackle the updates in data and business for the insurance industry.
- The system should provide an overview of a customer,Äôs risk profile based on the information they provided.
- The system should incorporate explainable exclusion classification based on respective circumstances.
- The system should provide explainability or transparency on an exclusion classification decision made by the model.

One point to be made clear in the study is that the designed system aims at improving the existing UP and assisting human underwriters as well as insurance companies in managing and decision-making when it comes to customer risks and relations, not replacing them entirely.

1.3 Research Problems

Based on the highlighted requirements and the defined gaps from current challenges, our research problem can be concluded into three main issues:

- **RP1**: How can we handle missing values of existing data survey to account for new data coming in and less frequent label?
- **RP2**: How can we better represent a more data-driven approach to identify inferred relationships when the model result and rules are black-boxed?
- **RP3**: How do we represent the explainability and transparency of the model?

The research problems identified here are further explored in Chapter 2 and formularised in Chapter 3.

1.4 Research Objectives

The following objectives have been determined from the problems above:

- **RO1**: Knowledge graph (KG) construction for a general view of all attributes linked to a customer with specific cases to account for less popular fields

- **RO2**: Semi-automated knowledge graph construction and rules identification for capturing domain knowledge.
- **RO3**: Provide transparency on which features have an impact on the link prediction result based on quality assurance report (QAR) and graph mining.

The justification for the choice of modelling (KG) is elaborated in Chapter 5.

1.5 Research Contributions

The contribution of this study is in alignment with the objectives stated above. As mentioned in previous section, this study is based heavily around the empirical application of the data set, with the contributions elaborated as:

- **RC1**: Construction and application of the underwriting knowledge graph (UKG) based on the two years worth of real-life data from industry to provide a more in-depth investigation into practical use.
- **RC2**: Propose a process to semi-automated knowledge graph construction and rules identification for capturing domain knowledge from data analytics and engineering.
- **RC3**: Propose an approach to reveal transparency (explainability) on which features have an impact on the link prediction result based on quality assurance report (QAR) and graph mining to assist with the maintenance and adjustments of underwriting rules.

The remaining parts of the thesis provide a view on how the study's objectives and contributions have been achieved.

1.6 Thesis Organisation

The first chapter of my thesis has introduced the current life underwriting landscape at the beginning of the study and the challenges derived from both the business and research aspects. The remaining chapters of my thesis are divided into seven parts to detail the work that has been carried out to further investigate and determine a solution for the currently identified challenges. The content of the following chapters are described as below:

Chapter 2: The Underwriting Data - This chapter portrays the domain requirements for constructing a model that caters to the life underwriting data as the backbone of this research study. It highlights the key objectives and limitations faced with regards to conducting an industry-based research to further explore the data set used within this research study.

Chapter 3: Related Works - This chapter provides an overview on the existing studies within related domains to identify the current research progress on constructing a model for life underwriting. In this chapter, comparison among existing studies is highlighted to determine the success and notable findings as well as the existing research gaps of automated underwriting methodologies. The key takeaways from this section shapes the formation of the proposed methodology for this study.

Chapter 4: Explainable Exclusion Using Multilabel Classification - This chapter portrays the first approach taken for explainable exclusion using multilabel classification. This section details the empirical experiment conducted and the result comparison between multiple multilabel classification models when applying to a real life underwriting data set to identify the most suitable model for the task with suggestions on further analysis in future studies. At the time of the study, this is the first application of multilabel classification to resolve the problem of explainable exclusion in life underwriting.

Chapter 5: Initial Knowledge Graph For Underwriting - This chapter portrays the second approach taken to alleviating the problem identified in section 2 using knowledge graph as a base. This chapter details the construction process, including the construction of nodes and edges of the knowledge graph that caters to the underwriting data and potential use cases from this initial underwriting knowledge graph. To the best of my knowledge, at the time of proposal, this is the first attempt of constructing an underwriting knowledge graph within the research field.

Chapter 6: Risk Profile Using Knowledge Graph and Multilabel Classification - This chapter presents the proposed methodology of this study. The proposed methodology is a combination of the explainable exclusion using multilabel classification with the use of knowledge graph with adjusted weight calculation inspired by the Jaccard similarity [24]. This includes the proposal of a semi-automated process as a part of the methodology for underwriting rules identification and maintenance using the adjusted weight and comparison with subgraphs of customers with similar features recorded.

Chapter 7: Discussion of Approaches Taken and Future Work - This chapter concludes the thesis with discussion on achievements and limitations throughout the study. These key points united serves as a baseline direction for future research in the field of applications of data innovations within underwriting sector and highly personalised services with similar data set structure.

CHAPTER 2

RELATED WORKS

2.1 Automated Underwriting Process Approaches

Artificial intelligence has been "trending" and applied by many industries in the last decade as simulating human intelligence via the application of machine [29, 30]. Particularly with the insurance industry, business needs regarding underwriting and claim assessments can be assisted significantly with machine learning and predictive analysis as part of AI, which further increases its popularity and adaptation. Previous studies have shown that the assistant of AI allows speeding up the process of underwriting with enhanced risk selection to improve pricing strategies [31, 40]. Within an industry characterised by its personalisation of policies, the application of machine and AI makes life underwriting faster, giving the green light for different insights to be drawn from the client's data [43, 44]. Research conducted over recent years on the integration of technology in the insurance industry has introduced various models to accelerate the underwriting process, leveraging the use of traditional machine learning (ML), deep learning (DL), and ensemble learning. Figure 2.1 presents a taxonomy of the most prevalent AI applications in the insurance sector, organized by their specific tasks. From an extensive review of the literature and this taxonomy, we identified that automated underwriting encompasses two primary risk classification tasks:

- categorizing applicants based on the decision outcome;
- identifying claiming patterns to detect potential anomalies.

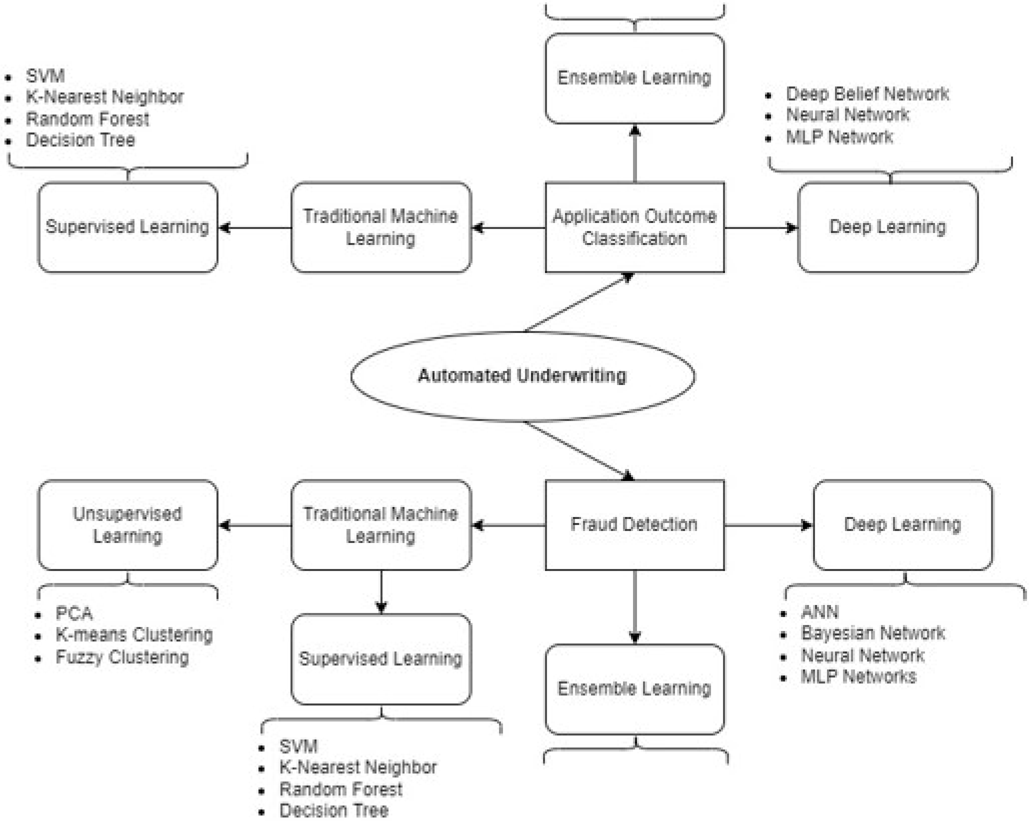


Figure 2.1: Taxonomy of automated underwriting approaches

Based on the identified tasks above, exclusion analysis can be deducted as a detail-specific subcategory of classification underwriting with explicit categories of risks entail. However, upon our literature review, a gap of coverage on this subtopic can be observed in current research studies.

Table 2.1 outlines key studies exploring various ML approaches to automate the underwriting process. It reveals that few studies have successfully addressed exclusion classification [5] [47]. Arora and Vij [4] proposed a neuro-fuzzy network in 2021 to classify applicants into five risk levels for premium calculation. However, their work lacks validation of the experiment and detailed results on how these premiums are calculated. Contemporary state-of-the-art machine learning models, naming Support Vector Machines (SVM), random forest algorithms, and Naive Bayes classifiers, have been adopted to categorise risk levels in insurance applications[22], but there has been minimal effort to innovate these models or focus on exclusion code classification. Despite the acknowledgment of the influence of exclusion codes on underwriting decisions [27], this aspect was addressed as an ancillary consideration rather than being the central

References	Key purposes	Model	Exclusion classification
Arora et al. [4]	Applying artificial neuro-fuzzy network to classify insurance application to risk classes for premium calculation	Hybrid Neuro-Fuzzy Network	No
Joram et al. [26]	Developing a knowledge-based system to output insurance decisions (risk classes)	Knowledge-based System	No
Hutagaol et al. [22]	Speed up the UP with ML application	SVM, Random Forest and Naive Bayes	No
Kavscelan et al. [27]	Applying non-parametric data mining technique to classify insurance claim sizes and occurrences	Support Vector Regression, Kernel Logistic Regression	No
Biddle et al. [5]	Applying ML methods to classify the application of exclusions in life insurance	Logistic Regression, XGBoost and Recursive Feature Elimination	Yes
Mourmouris et al. [47]	Assign a score applications to classify them to risk classes	Multi-criteria Decision-making analysis	Yes

Table 2.1: Key studies: different methods for life insurance label classification in the insurance industry.

focus. Biddle et al. [5] investigated the exclusion classification process using XGBoost, but they only managed to concentrate on the top 20 exclusion codes, leaving out less common ones and not utilising the entire underwriting dataset. From the literature available, it is evident that most studies on underwriting automation focus on classifying risk levels. Although exclusion codes are referenced, their classification has not constituted the primary research focus. To the best of my knowledge, our collaborative research [63] represents the inaugural study to adapt multi-label classification framework for explainable exclusion classification within the realm of life insurance.

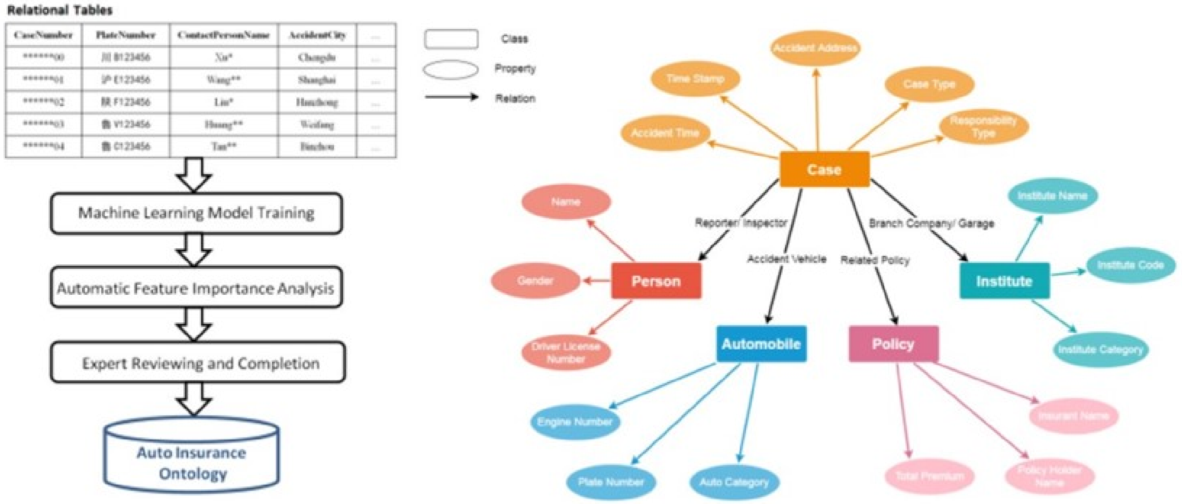
2.2 Automated Underwriting Customer Risk Profile

At the time of this literature review, the prevailing method for developing a customer risk profile involves classifying applicants into specific categories based on the risk levels indicated by their underwriting data [27]. Table 2.1 has determined the existing efforts in categorising risk level within the insurance industry. While this method maintains simplicity, it may overly simplify the risk levels associated with an application and obscure the decision-making logic. A comprehensive customer risk profile should retain a degree of complexity, utilising more advanced risk representation systems. Additionally, the process for creating this risk profile should support scalability and incorporate guidelines from human underwriters' manuals. This research study aims to address these existing gaps via the use of multi-label classifications to maintain the complexity of customer risks presented by exclusion codes, as well as investigating the underwriting rules manual knowledge by applying knowledge graph to study underwriting data.

2.3 Domain-specific Knowledge Graph

The examination of the current UP has driven this research study towards the creation of an underwriting ontology or underwriting knowledge graph, which, while not applied in a prevalent way within the insurance sector, has been proffered by numerous scholars for health risk profiling. Wu et al.[69] defined a knowledge graph as an aggregation of knowledge points rendered in graph form, with nodes and edges symbolising data entities and their interrelations. This definition interprets that a knowledge graph can be any graph-based data model, including resource description framework (RDF) datasets, semantic web knowledge bases, ontologies, and multi-relational graphs comprising nodes and entities[66].

Research on knowledge graphs can be categorised into two main areas: techniques for constructing a KG and the applications of KGs [72]. The use of KGs for underwriting falls under domain-specific KGs. These are specialized applications within the broader field of KGs, with successful implementations in areas such as medicine [14, 17, 53, 55, 67], cybersecurity [25, 49], finance [11, 12, 35], education [6, 18], music [56], and religion [68]. The deployment of Knowledge Graphs (KGs) for customer profiling within the finance and insurance sectors has been advocated by prominent consulting firms such as McKinsey and Deloitte [8, 9] in recent years due to its application potentials. Nonetheless, this subject has not yet garnered substantial attention within the domain of insurance and



Relational Tables

CaseNumber	PlateNumber	ContactPersonName	AccidentCity	...
******00	川 B123456	Xu*	Chengdu	...
******01	沪 E123456	Wang**	Shanghai	...
******02	陕 F123456	Liu*	Hanzhong	...
******03	鲁 V123456	Huang**	Weifang	...
******04	鲁 C123456	Tan**	Binzhou	...

Figure 2.2: AIKG model architecture [70]

underwriting-specific research.

Zhang et al. [70] asserted that they were pioneers in applying KGs and creating an ontology in the insurance sector, specifically for fraud prediction. Figure 2.2 demonstrated their proposed model architecture, wherein they constructed an Auto Insurance Knowledge Graph (AIKG) through the extraction of knowledge from a relational insurance database. Subsequently, they employed link prediction within the knowledge graph to discern individual fraud cases and to identify "fraud gangs", characterised by multiple fraud cases exhibiting resembling attributes. Although their system successfully applied KGs to the insurance field, it did not provide a comprehensive risk profile for customers with multiple claims, raising difficulties in applications into life insurance. Additionally, more work is needed in feature selection and feature engineering to reduce computational demands and enhance graph scalability.

In his survey over domain-specific knowledge graph in 2021, Abu-Salih [1] have raised several concerns over the construction and quality of existing domain-specific knowledge graph, first being the unstandardisation and undisclosure of algorithm in knowledge graph construction and data capturing process, second being incorrect facts captured within the graph, and third being the imcompleteness of the graph due to embedding. These problems will be addressed throughout the study.

2.4 Knowledge Graph Link Prediction Using Multi-label Classification

Multi-label classification is explored for its potentials in addressing exclusion prediction and the significant improvements it demonstrates when combined with link prediction [52, 71]. Martinez, Berzal, and Cubero's work integrates multi-label classification into the classifier-based link prediction task [41]. Although multi-label classification has previously been applied to link prediction, most current uses are based on deep learning and focus on unstructured social network data [7, 39, 52, 65, 71]. This presents an opportunity to apply multi-label classification for link prediction in knowledge graphs using machine learning algorithms on relational databases.

2.5 Explainability in Artificial Intelligence

Within the context of AI, explainability, interpretability and transparency can be considered semi-synonymous with subtle differences. Interpretability refers to understanding the reasoning behind an AI-based decision-making system, whereas explainability involves the AI system's ability to provide a clear explanation of how it reached a decision [34, 46]. Certain ML/AI models, such as decision trees and logistic regression, are deemed interpretable; however, their nested non-linear architectures can be opaque and less accurate in comparison with intricate black-box models like deep neural networks or ensemble models. Numerous interpretation methodologies have been proposed, including tree interpreters for random forests and deep decomposition for neural networks [46]. These methods range from formalising interpretability mathematically to offering visual explanations or enhancing task performance through algorithm-generated explanations [61]. In fields like insurance and credit scoring, numerous studies have introduced various ML models, but they often overlook the explainability of decisions made by these models [54]. Considering that these methodologies are tailored to specific models, additional research is imperative to formulate novel approaches for the explainability of ML models specifically and AI models in general.

2.6 Summary

Exclusion codes, which represent the risks associated with an insured customer, are a crucial component in constructing a domain-specific customer risk profile for life insurance. The comprehensive literature review in Chapter 2 has unveiled numerous limitations within this domain. In this section, an evaluation is presented aligning with the prerequisites outlined in Chapter 1 to determine whether the current methodologies satisfy the method requirements.

The proposed method should be applicable and optimisable to deal with the ever-increasing changes in data in the insurance industry. Applicability and optimisation to handle evolving data in the insurance industry comes in as the first prerequisite for these approaches. Most studies in the insurance industry meet this requirement. Various ML and DL models have been parameterised and optimised for the necessary tasks. Although the basic industry needs are addressed, there is still room for improvement, particularly in the area of explainable exclusion.

The proposed method should provide a better way to overview a customer's portfolio (based on their information) and potential risks. Although the necessity of maintaining a customer profile portfolio has been acknowledged, the existing methodologies for constructing customer risk profiles for underwriting and insurance purposes remain overly simplistic. Despite efforts to introduce KG applications in insurance, a domain-specific KG for life insurance has yet to be developed, with a lack of a classification systems catered for sophisticated data dimensions. The earliest public research application of a knowledge graph (KG) within this domain was in auto insurance. Despite the successful construction of the AIKG [70] model, there remains an imperative for advancements in the domains of graph feature engineering and the optimisation of link prediction performance.

The proposed method should have the ability to classify applications into multiple exclusion codes based on their respective circumstances. There has been a constrained focus on the researched automated classification and application of exclusion codes to determine risk factors within the insurance sector, notwithstanding the fact that this constitutes a critical component of the underwriting process. Existing studies referring to exclusion code classification reveal gaps that future research should address, particularly in considering more infrequent exclusion codes. Insufficient information to predict exclusion codes can disrupt the procedure of curating a customer risk profile.

The proposed method should be able to highlight the explanation for an exclusion classification made by the model. Substantial potential for enhancement resides in model explainability, particularly given that the prevailing interpretability techniques are confined to being model-specific. The lack of studies on exclusion code classification has also resulted in little focus on explainable exclusion. This requirement has not been adequately met.

Overall, the literature review conducted bestows invaluable insights into ongoing existing studies on customer risk profiles utilising knowledge graphs and multi-label link prediction. It is evident that current studies on explainable exclusion analysis and the construction of customer risk profiles is circumscribed. A limited number of studies have applied advanced data analytics methodologies to address the pragmatic requisites of the insurance industry, and none have undergone empirical validation on substantial data sets. These current limitations delineated above serve as a directive to formulate the research questions and objectives of this study.

Chapter 3

The Underwriting Data

3.1 Domain Requirements

A domain-specific knowledge graph tailored for life insurance ought to possess the capability to comprehend and be applicable to the extensive use cases prevalent in the UP. As discussed earlier, the primary use case of such a system entails the precise assessment of a customer's risk through the classification of exclusion codes. For instance, a customer exhibiting a heightened propensity for incurring substantial and frequent claims related to arthritis, predicated on preexisting knee health conditions, is more inclined to be excluded from payouts pertinent to osteoarthritis. Therefore, it is crucial to understand how life insurance data is collected, structured, and utilized to determine exclusion decisions. This requires knowledge of the attributes relevant to the underwriters' decision-making process, potential dilenma, data analysis, and the decision itself.

The UP typically relies on underwriting standards and the expertise and experience of underwriters. The rules take into account various personal factors gathered from multiple sources, such as customer survey questionnaires, medical records, and employment history, depending on their importance to the decision. However, this highly customised data demands extensive feature engineering before its incorporation to a knowledge graph system. Specifically, underwriting data is specifically derived from customer underwriting survey questionnaires, which filter subsequent inquiries based on the customer's prior responses. As previously stated in Section 1.1.3.2, the streamlined view of the questionnaire reveals that questions are aggregated, directing customers only to pertinent questions based on their answers. This process results in numerous unanswered questions, creating data fields with over 75% missing values when translated into a tabular dataset. This was raised in Section 2.6 as a limitation to the efficacy of current ML-based risk classification models.

3.2 Data Set Overview

The research methodology and experiment focus on applying real-world life insurance data from a prominent Australian insurance company. The data provided by our industry partner includes four datasets: customer disclosure information, ID linkage information, policy number information, and exclusion codes with subcodes. These datasets are merged into a single primary dataset using the ID linkage information, where each row is uniquely identified by a policy number, as a customer may have multiple policies. Table 3.1 provides a sample view of the attributes in the dataset.

Attribute	Type	Description
Customer age	Number	Age in years
Gender	String	2 value depicting biological sex
Employment status	String	Categories of employment status
Occupational status	Boolean	Is the customer employed?
Smoking status	Boolean	Does the customer smoke?
Disclosure questionnaires	String	Various categories on different types

Table 3.1: Features of sample data attributes

The dataset is categorised into four types: categorical, date, numeric, and multi-valued data. Attributes containing personally identifiable information (e.g., email, name, phone number) are excluded to protect customer privacy and reduce model noise, as these attributes do not contribute to risk identification. Free text attributes (often found as clarifications to specific question nodes) are also removed, as they have minimal impact on the exclusion classification process.

3.3 Data Preliminary Analytics

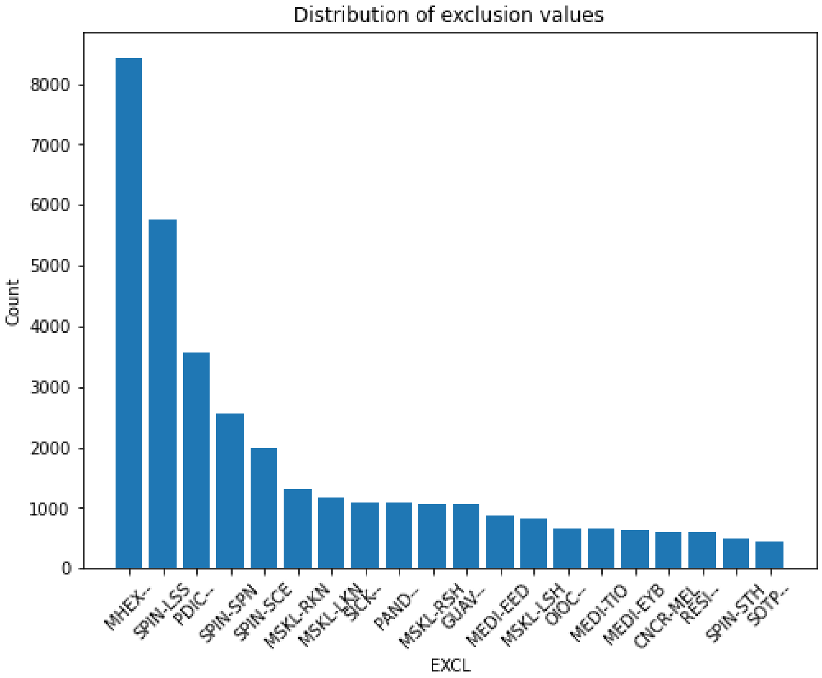
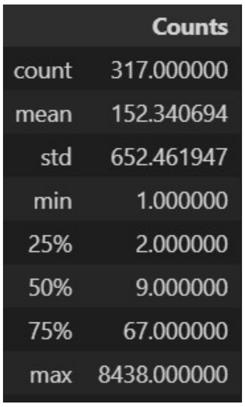


Figure 3.1: Top 20 exclusion codes



	Counts
count	317.000000
mean	152.340694
std	652.461947
min	1.000000
25%	2.000000
50%	9.000000
75%	67.000000
max	8438.000000

Figure 3.2: Applications of 317 exclusion codes on trauma cover policies

In order to resolve the data-specific question detailed in Section 1.1.3.4, an understanding of the data set should be prioritised. This study is based on the real life underwriting data set that is provided by the University's industry partner over the course of 3 years (2019-2021). Prior to approaching the methodologies, I have conducted preliminary data analysis for an overview of the data set in line with the business problem. This section details the key points result from the process, serving as a deep-dive into the data set to further elaborate the research problem portrayed in Section 1.4.

3.3.1 Customers With Exclusions Applied

One of the hypotheses for this study is most of the exclusion codes are specific health related problems. Figure 3.1 portrays the distribution of exclusion codes applied on the policies existing in the data set, with 13 out of 20 exclusion codes focus on detailing health concerns, which is in alignment with our hypothesis. The preliminary

Exclusion Name	Description
MHEX	Mental Health Exclusion
SPIN-LSS	Spine - lumbosacral spine
PDIC	Pre-disability income
SPIN-SPN	Spine - spine
SPIN-SCE	Spine - cervical spine
MSKL-RKN	Musculoskeletal - right knee
MSKL-LKN	Musculoskeletal - left knee
SICK	Sick leave offset clause
PAND	Pandemic illness
MSKL-RSH	Musculoskeletal - right shoulder
GUAV	Guaranteed agreed clause
MEDI-EED	Medical condition - any disease or disorder of either or both ears including deafness
MSKL-LSH	Musculoskeletal - left shoulder
OIOC	Ongoing income
MEDI-TIO	Medical condition - any disease or disorder of the inner ear (cochlea, vestibule and semi circular canals)
MEDI-EYB	Medical condition - any disorder of either or both eyes including blindness
CNCR-MEL	Cancer-specific - melanoma or other skin cancer
RESI	Residential clause
SPIN-STH	Spine - thoracic spine
SOTP	Second occupation TPD

Table 3.2: Features of sample data attributes

data analysis result also displays that 48292 policies out of a total of 550776 policy results provided has had an exclusion code applied, which puts the estimates on roughly 8.8%.

Specifically for trauma data set, 12% of customers have more than 1 exclusion code applied to their application. Over the course of 2 years, more than 65% of exclusion codes are applied less than 10 times, aligning with the data description of exclusion codes shown in Figure 3.2. with 75% of exclusions fall into the 1-67 range while the remaining 25% has feature that counted up to 8438 times. This trend can also be observed with other cover types.

3.3.2 Feature Value Frequencies

Similarly, a simple feature value counts after preprocessing portrays that 98% of feature columns have more than 75% missing value cells. This is due to the setup of the survey, with questions in the original questionnaire grouped into multiple subcategories. Figure 3.3 shows that even the highest ranking group of questions only accounts for nearly 5.5% of the total number of questions in the questionnaire.

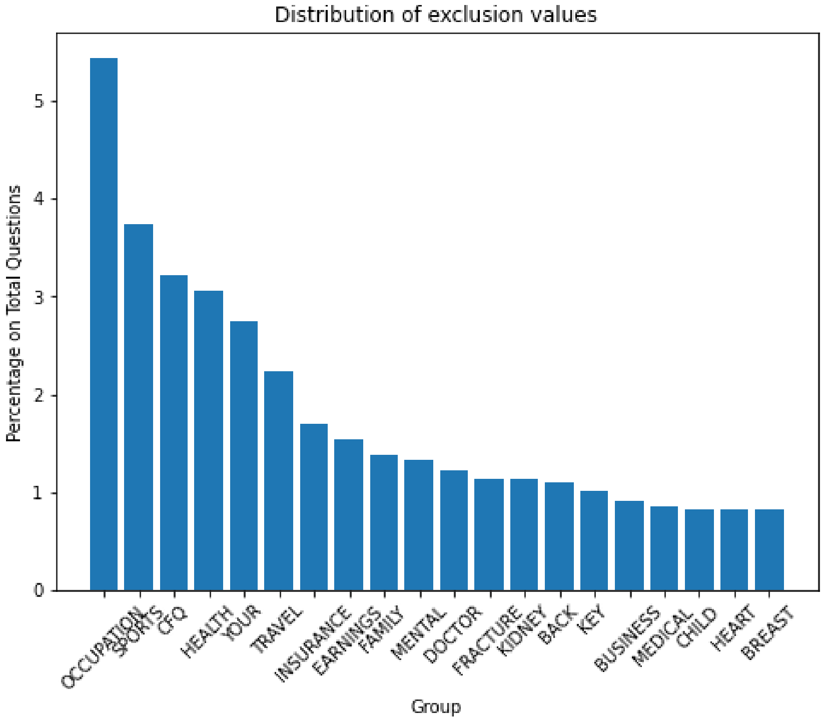


Figure 3.3: Percentage on total number of questions for most popular question groups (excluding basic details)

3.4 Algorithmic Problem

The requirements are translated into a data problem based on the findings from the literature review and the specific needs of the insurance partner. Considering z_i as an instance in the dataset (representing a customer), the following definitions apply:

- $X = (X_1, X_2, ... X_m)$ represents the attributes derived from customer disclosure questions used to classify exclusion codes.,

- $X_i = (X_{i_1}, X_{i_2}, ...X_{i_m})$ is the set of attribute values associated with the learning instance,
- $Y = (Y_1, Y_2, ...Y_n)$ is the list of labels (exclusion codes),
- $Y_i = (Y_{i_1}, Y_{i_2}, ...Y_{i_n})$ is the list of probabilities that each exclusion code is applied to customer z_i,
- k is the number of exclusion codes applied to customer z_i.

Thus, the research problem simplifies to developing an algorithm to classify $z_i \in Z$ into $y_k \subset Y$ based on X_i. Identifying the y_k sub-list of possible labels for each customer is the core of our multi-label classification phase. Table 3.3 demonstrated a formulated dataset used in this study for multi-label classification.

	X_1	X_2	...	X_m	Y_1	Y_2	...	Y_m
z_1	X_{1_1}	X_{1_2}	...	X_{1_m}	Y_{1_1}	Y_{1_2}	...	Y_{1_n}
z_2	X_{2_1}	X_{2_2}	...	X_{2_m}	Y_{2_1}	Y_{2_2}	...	Y_{1_n1}
...	...	...	...	...	...	...	...	...
z_N	X_{N_1}	X_{N_2}	...	X_{N_m}	Y_{N_1}	Y_{N_2}	...	Y_{N_n}

Table 3.3: Preprocessed demo data set for multi-label classification.

This problem forms the first approach to tackle customer underwriting risk profile which is elaborated in Chapter 4.

CHAPTER 4

EXPLAINABLE EXCLUSION USING MULTILABEL CLASSIFICATION

4.1 Background and Motivation

From the domain requirements and results captured in Section 3, along with the lack of focus pointed out in Section 2, multilabel classification is the first approach taken upon the dataset to alleviate the existing issues identified. The data chosen for this approach includes the categorical and binned data due to several reasons including:

- The majority of free text questions are encoded for being customer identified information,
- The free text questions for additional information are usually precedented with one or multiple categorical questions above to identify whether more information are required, hence, we assume that a general free text question will not be triggered in the URE unless a directly related question is called previously.
- These processed data types can be transformed to fit our multi-label classification model.

4.2 Key Contribution

The main contributions of this approach can be summarized as follows:

- Gathered domain-specific requirements for analyzing insurance datasets for exclusion classification, laying the groundwork for a deeper exploration of domain requirements in our research.
- Utilised a comprehensive two-year dataset from a reputable Australian insurance company to address exclusion analysis, aligning with the study's aim of using a data-driven approach to automate the process.
- Employed four multi-label classifiers (binary relevance, classifier chains, label powerset, and ensemble learning) along with various ML techniques to tackle the exclusion problem and determine the most effective model for explainable exclusion. This model serves as a baseline for future analysis and improvement leading to the construction of a UKG as mentioned in section 1.1.
- Validated the results by comparing different metrics and involving human underwriters to assess the proposed model's performance, ensuring human experts are part of the evaluation and quality control process.

4.3 Model Methodology

The objective is to determine the appropriate exclusion code(s) to be assigned to a policy, thereby mitigating insurance risks. Given the preference for considering multiple exclusion codes, we have suggested the use of a multi-label classification method. Multi-label classification operates similarly to single-label classification, but with the distinction of accommodating multiple target labels [60]. In this study, the experiment implements four multi-label classification algorithms utilising five primary classifiers: Multinomial Naive Bayes, Support Vector Classification, Logistic Regression, Random Forest, and Decision Tree.

Binary Relevance: Binary Relevance algorithm transforms multi-label classification problem to multiple independent single-label binary classifications. As the classes are considered independent from one another, Binary Relevance can use any binary classifiers as base learner [45]. However, it is worth noting that the original form of Binary Relevance do not recognise possible dependency among class labels [36].

Classifier Chains: Similar to Binary Relevance, Classifier Chain transforms the multi-label classification into multiple binary classifiers for each label [10], though the main difference is the algorithm forms a series of label relevances using the predictions of previous label classifiers, hence linking all binary classifiers to a chain. The improvement

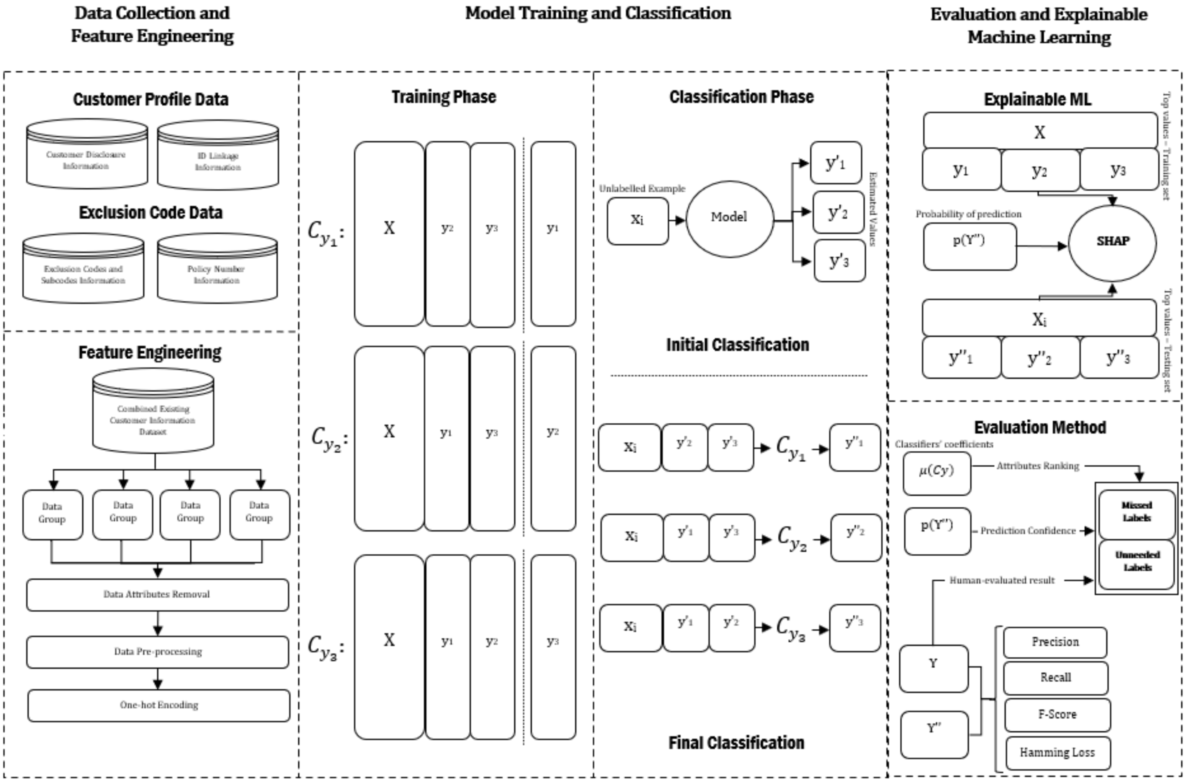


Figure 4.1: Model methodology for explainable exclusion [63]

of Classifier Chain in comparison to Binary Relevance comes with its ability to construct a correlation system among labels [51].

Label Powerset: Label Powerset takes into consideration potential label correlations to construct unique label set from existing labels provided. Each set of labels is transformed into a class in Label Powerset classification, from which the problem is converted into a multi-class classification with the output being the most probable class from the label classes [16]. In comparison to Binary Relevance and Classifier Chains, Label Powerset sits right in the middle, having the advantages of both previous algorithms. The setback of Label Powerset comes only when the problem complexity increases, meaning there are more labels to group and consider.

Ensemble Learning: Ensemble Learning method follows the process of learning the chosen base classifiers and determining weights for each classifier [36]. Different classifier applied as base classifier for Ensemble Learning might focus on different steps of the process.

4.4 Evaluation and Explainability

In the evaluation process, standard metrics are considered such as precision, recall, F-score, and hamming loss. A novel evaluation-explanation technique called Quality Assurance Reports (QAR) is introduced in this study, which involves identifying "missing" and "unneeded" labels. Shapley Additive Explanations (SHAP) [37] is also leveraged to interpret the model outcomes and explicate the manner in which various features impacted the exclusion classification[64]. SHAP leverages coefficients in the context of linear models or feature importance within tree-based models for each exclusion code, thereby revealing and visualising the particular impact of each attribute to the model's classification results.

4.4.1 Quality assurance report (QAR)

Due to a large number of missing values in this real-life data set, the model could be biased towards more common and non-specific data attributes (those with fewer missing values), and low scorings from the four standard evaluation methods are expected. Thus, QAR is introduced for human underwriter reviews and evaluation.

QAR is computed using the classification probability and feature importance of each label's classifier for every data row in the testing set. This process entails generating a compendium of differential comparison values for each individual row by deducting the actual label results, represented as binary outcomes (0s and 1s), from their corresponding classification probabilities, from which "missed" and "unneeded" values can be identified based on a predefined threshold. Specifically, by denoting Y'' as the classified values, which can assume binary outcomes (0 or 1), $p(Y'')$ as the classification probability, encompassing a continuum from 0 to 1, Y as the authentic label values, also binary (0 or 1), and t as the threshold parameter, spanning a range from 0 to 1:

- A label y is considered "missed" if $p(y'') - y \geqslant t$ (4)
- A label y is considered "unneeded" if $p(y'') - y \leqslant t$ (5)

The primary contributing factors for each missed or unneeded label are hierarchically ranked according to their feature importance in the classification of that specific data record. This importance value is obtained either from the coefficients (linear algorithms) or from the feature importance metrics (tree-based algorithms) for each label's classifier. Within our context, instances devoid of affirmative responses ("YES" answers), indicated by the absence of 1s in the data attributes for missed labels, and instances devoid of negative responses ("NO" answers), indicated by the absence of 0s in the data attributes for unneeded labels, are excluded from consideration.

4.4.2 Shapley Additive Explanations (SHAP)

SHAP is a comprehensive framework used to interpret a model's decision by assessing the importance of each feature in a specific prediction [37]. In this study, SHAP's visualisation features assist in recognising and prioritising feature combinations influencing decisions for individual exclusion codes and a set of the 15 most common exclusion codes. While QAR explains decisions at the granular level of individual data records, SHAP provides a hierarchical ranking of features across multiple rows classified under identical exclusion labels. This hierarchical feature ranking can be measured against pre-existing underwriting rules to enhance explainability and inform refine the decisions in future underwriting.

4.5 Empirical Experiment

4.5.1 Data collection and processing

This research methodology and experiment focus on applying real-world life insurance data from a prominent Australian insurance company. The datasets, spanning from 2019 to 2021, include customer disclosure information, policy details, and the exclusion codes applied based on customer responses. This dataset is then divided into subsets based on the types of coverage the customers are seeking (such as trauma, total permanent disability - TPD, disability, and term) to help manage the data effectively as well as customising for specific covers. This categorisation helps streamline the processing and reduces the computational burden by focusing on specific cover types.

Each cover type dataset is further divided into four sub-datasets, resulting in a total of four datasets per cover type. To consolidate the data, all four sub-datasets are merged into a single primary dataset using ID linkage information, with the policy number serving as the unique identifier for each row. This approach is illustrated in Figure 4.3, and a sample of the data attributes is provided in Table 3.1.

For each type of cover, the subset is divided into four groups based on data types: categorical, date, numeric, and multi-valued. To protect customer privacy, columns containing personally identifiable information (e.g., email, name, phone number) are removed. Additionally, columns containing free text are excluded, as they have minimal impact on the exclusion categorization process. The model's input data mainly consists of columns with binary values (0 and 1).

4.5.2 Experimental setup

Twenty multi-label classification models are developed by substituting five base classifiers (multinomial Naive Bayes, Support Vector Classifier (SVC), logistic regression, random forest, and decision tree) into the four multi-label classification algorithms (binary relevance, classifier chain, label powerset, and ensemble learning). These models are constructed and parameterised utilising predefined functions available within the sk-multilearn library [57].

Beyond the standard metric evaluation, QAR is applied investigations into instances where the model output diverges from human underwriter decisions. For example, a threshold of 97% can be established, whereby any classification by the model with a confidence score exceeding 97% that contradicts an underwriter decision will be flagged for further examination.

4.5.3 Evaluation metrics

The evaluation process employs four widely recognised evaluation metrics prevalent in the assessment of multi-label classification: precision, recall, f-score, and hamming loss. Accuracy score was initially considered for the evaluation process, yet later removed due to the false positives affecting the result.

4.5.4 Evaluation

The evaluation process is conducted by involving human underwriters. They review the model scores, QAR outputs, cross-reference additional applicant materials, and compare the final decision in their system. This process helps determine the accuracy of exclusion label identification by the model and assesses their satisfaction with the QAR results.

Algorithm	Binary Relevance					Classifier Chain				
Classifier	MNB	SVC	LG	RF	DT	MNB	SVC	LG	RF	DT
Precision	0.26	0.62	0.80	0.93	**0.63**	0.26	0.94	0.80	0.93	0.61
Recall	0.11	0.46	0.41	0.32	**0.56**	0.11	0.11	0.41	0.32	0.56
F-Score	0.15	0.53	0.54	0.47	**0.60**	0.15	0.20	0.54	0.48	0.59
Hamming Loss	0.002	0.002	0.001	0.001	**0.001**	0.001	0.001	0.001	0.001	0.001

Algorithm	Label Powerset					Ensemble Learning				
Classifier	MNB	SVC	LG	RF	DT	MNB	SVC	LG	RF	DT
Precision	0.35	0.89	0.76	0.88	0.57	0.26	0.93	0.80	0.92	0.59
Recall	0.45	0.06	0.41	0.40	0.48	0.10	0.10	0.43	0.34	0.54
F-Score	0.08	0.11	0.54	0.56	0.52	0.15	0.19	0.56	0.49	0.56
Hamming Loss	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 4.1: Precision, recall, and F-score and hamming loss of the multi-label classification model [63]

4.6 Results Analysis

This section presents the outcomes of the experiments conducted. Initially, accuracy were proposed as an evaluation metric. However, accuracy only provides an overall measure of model confidence, which aggregated the results without considering the huge gap among detailed accuracy of each label. Therefore, the evaluation metrics opted in the use of precision, recall, and F-measure, with a focus on the F-score, to ensure the reliability of our classification. Table 4.1 encapsulates the comprehensive results of our models subjected to cross-evaluation. Generally speaking, a similar trend of Hamming loss scoring below 0.01 amongst all models signified that the majority of labels were accurately classified in terms of their relevance.

During the course of the experiments, logistic regression models consistently exhibited commendable baseline performance, showcasing uniformity across various algorithms (approximately 0.54). The experimentation revealed that multinomial Naive Bayes models yielded a substantially greater Hamming loss score compared to other conuterparts, indicating subpar performance in the assigned task. It is advisable to refrain from employing Support Vector Classifier (SVC) models for this task, as they struggle to manage a considerable volume of missing data. Although initially considered for their promising scores on smaller datasets with low missing value rates, the SVC

SN	Dataset	# features	# labels	Total size	# missed exclusion	# un-needed exclusion
1	Term	8161	85	(57900,8248)	1221	3351
2	Trauma	6643	125	(16971,6770)	35	347
3	TPD	6105	182	(33289,6289)	347	1350
4	Disability	12785	206	(21598,12993)	442	1551

Table 4.2: Description of quality assurance outcome

models failed to function properly on the full dataset, as evidenced by the significant gap between their precision and recall results in comparison with their counterparts.

Tree-based models emerged as the preeminent performers among the array of models assessed. Random forest consistently demonstrated precision values ranging from 0.88 to 0.93 across diverse algorithms. Nonetheless, the recall scores and F-scores obtained when training on the entire dataset were comparatively deficient. Conversely, although the decision tree could have exhibited heightened precision, its F-scores surpassed those of its counterparts. Discrepancies in the scores derived from distinct evaluation methodologies stemmed from data attributes plagued by an excess of 75% missing values. This underscores the exigency for future endeavors to enhance this study through the implementation of advanced feature selection techniques and techniques for handling missing values.

Tree-based models (random forest and decision tree) achieved the highest scores among the models tested. Random forest consistently achieved a precision of 0.88 to 0.93 across multiple algorithms. However, the recall scores and F-scores obtained when training on the entire dataset were relatively low. On the other hand, while the decision tree could have performed better in terms of precision, its F-scores were higher than those of the other models. Variations in the scores from different evaluation methods were due to data attributes having more than 75% missing values. This suggests the need for future work to improve this study by implementing advanced feature selection and missing value handling techniques.

Using QAR, the exclusion labels were identified as missed or unnecessary by the models. A summary of this information for the four cover type datasets can be found in Table 4.2, while Table 4.3 provides a sample outcome of the QAR, intended for review by human underwriters. IDuring the third-party evaluation phase employing model scoring and QAR analysis, the outcomes emanating from the best performing model were deemed applicable for implementation into production by the human underwriters.

SN	Missed/Unneeded Exclusion Reasons	Label	Conf.	Outcome
1	MotorCar One Off - No, Motor Car Prof - No, MotorCar Marshall - No, MotorCar Intl - No, MotorCar Type1 - Circuit racing, MotorCar Circuit - Single seater, MotorCar Drag - Other, MotorCar Circuit - Saloon cars, Doctor Main Last - 1-2yrs, MotorCar Type1 - Rallying.	PSTM-MCR	.99	Unneeded
2	Travel Region Malaysia - Yes, OCCUPATION - Rehabilitation consultant, POSTCODE 2096, OCCUPATION CONFIRM - Unspecified occupation - A3 - Office based occupations that involve no manual or field work, Naevus Cancer - Yes, Earnings Which Year - No, EMPLOY STATUS CONFIRM - Employee (permanent or contractor), Back Occupation - Yes, QV Summary Semp - No, Back Location - Neck (Cervical Spine).	TRAC-MAL	.98	Unneeded
3	Melanoma Spread - No, Melanoma Recur - No, SkinCancer Confined - Yes, Recent Ix Disclosed - No, Melanoma Removed - Yes, CANCER - Yes, Additional Info - No, Melanoma Surgery - Yes, Melanoma OtherTx - No, Melanoma Treat Ceased - Yes	CNCR-MEL.	.98	Unneeded
4	WHICH HEARING - Hearing impairment, InsHx CompanyBE - Asteron, MH Lastsym - 2-3yr, PRODUCT - Business Expense, Snow Comp - No, BE Complete now - No, Earnings Employee Last Bonus 20 - No, InsHx Purpose LIFE - Personal protection, Deafness Date Resolve - 1-2yrs, Occ Sick Leave 10 days or less.	MEDI-EED	.85	Missed
5	PRODUCT SUITE - Active, BackDisc Location - Lower back (Lumbar Spine), BackDisc Sep Occ - More than twice or continuously, CVComb Await - No, MH Medication ongoing - No, MH Medication ongoing - Yes, Occ Duties Hazard2 - No, Knee OneOff - More than one episode, TRAVEL PAST - No, Administration - No.	SPIN-LSS	.96	Missed

Table 4.3: Sample of QAR results

4.6.1 Explainable machine learning analysis

In addition to using QAR for explaining our models' outcomes, as mentioned in Section 4.4.2, this experiment uses SHAP to explore the relations and magnitude of the features for the top 15 exclusion codes. The SHAP summary report, as shown in Figure 4.2, hierarchically orders the data attributes pivotal in influencing the classification of the top 15 exclusions within the trauma dataset. This facilitates the identification of features serving as the principal driving forces within their corresponding question

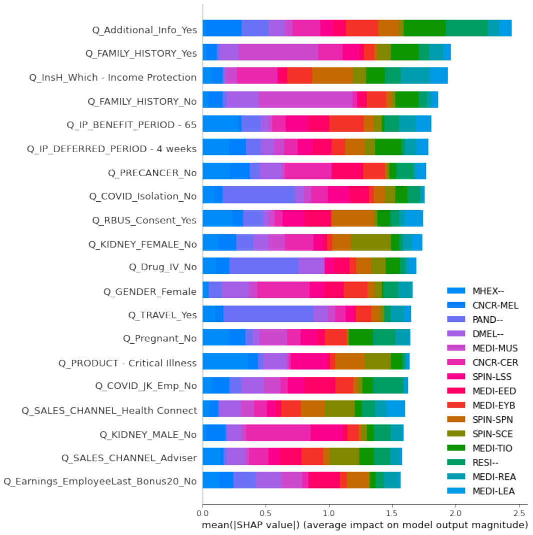


Figure 4.2: Top 15 most common exclusion codes - SHAP summary report

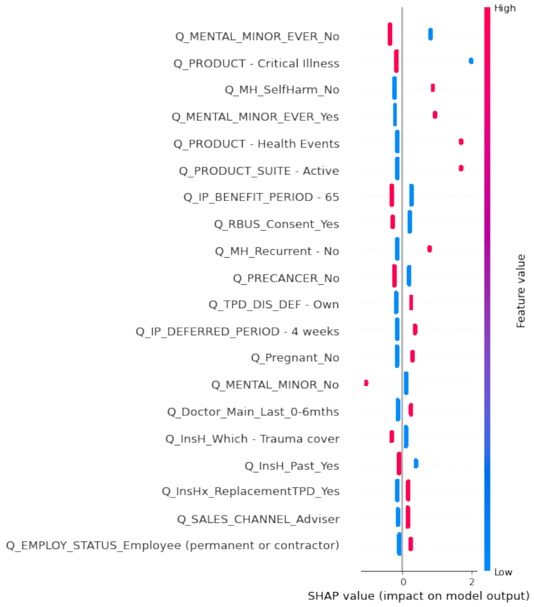


Figure 4.3: Driving factors of the highest applied exclusion code (MHEX) - SHAP detailed report

groupings. In detail, the SHAP detailed report provides an enhanced understanding of a designated target label and enables explainability even for less frequently occurring exclusions. Figure 4.3 specifies the pivotal questions instrumental in the classification of the exclusion code label.

Deciphering the outcomes of the model necessitates a comprehensive analysis SHAP results combining with QAR results. While the QAR furnishes the primary rationales underlying the classification of an exclusion code as either "missed" or "unneeded," SHAP visualisations offer supplementary insights into the same exclusion code by showing the distribution of values for each individual question. These attributes empower human underwriters to discern the pivotal factors within each unique record while simultaneously overseeing overarching patterns across records within the same label category.

In alignment with previous research studies [5, 42], this experiment's findings substantiate the assertion that tree-based classifiers, including random forest and decision tree algorithms, remain the optimal choice for exclusion classification endeavors. Notably, within the purview of our investigation, decision tree classifier engined binary relevance multi-label classification model surpasses all alternative models in terms of performance metrics. Additionally, the decision tree's feature importance contributes to model transparency and explainability. This notably facilitates the integration of QAR, where erroneous model classifications manifesting as false positives and false negatives are construed as missed exclusions and unneeded exclusions, respectively. The hierarchy of feature importance engenders a systematic ranking of attributes instrumental in shaping the model's classification decisions. Analysing The foremost rationales ranked by each exclusion classifier can reveal noisy features, setting a foundation for future feature engineering work. Future application of this study could implement a weighting metric for features, assigning less weight to noisy features and more weight to features directly correlated with the exclusion code. In addition to the selection of the classifier itself, the findings also suggest that the algorithmic framework enveloping the classifiers exerts a substantial influence on the outcome, particularly discernible with certain classifiers. This is evident in the performance discrepancies observed with SVC-based models when integrated with the binary relevance algorithm, resulting in an F-score of 0.54, as opposed to alternative algorithms yielding F-scores spanning a range from 0.11 to 0.20.

Observing Table 4.1, even with the best-performing model, the F-score recorded is 0.60. it becomes apparent that even with the most optimal model, the recorded F-score stands at a modest 0.60. This underwhelming performance can be ascribed to one of the principal challenges encountered throughout the experimental proceedings, stemming from the arduous task of addressing missing values within the original dataset. Owing to the bespoke nature of the existing system, each inquiry is meticulously tailored to suit the unique circumstances of individual customers. This approach means the majority of inquiries in the survey are dynamically filtered based on the responses provided by the customer, consequently leading to a plethora of queries featuring null values.

Approximately 75% of the customer disclosure columns contain missing values, although the information in these columns determines which exclusion codes to apply. Another issue is that infrequent exclusion codes are applied fewer than 10 times over 3 years, creating a historical data problem. This impacts model accuracy as the recall score does not match the precision score. These challenges in data preparation, including data cleaning and feature selection, present opportunities for future studies to focus on data wrangling and classification with limited values. Approximately 75% of the columns pertaining to customer disclosures harbor missing values, notwithstanding their pivotal role in dictating the applicability of exclusion codes. Additionally, another complication arises from the infrequent application of exclusion codes, with some codes being invoked fewer than 10 times over a span of 3 years, thereby engendering a historical data conundrum. This predicament exerts a palpable impact on the accuracy of the model, as evidenced by the incongruity between the recall and precision scores. These inherent challenges intrinsic to data preparation, including data cleansing and feature selection, consequently present opportunities for prospective future investigations to focus on the intricacies of data manipulation and classification within the confines of limited values.

CHAPTER 5

KNOWLEDGE GRAPH FOR UNDERWRITING

5.1 Background and Motivation

Considering the current limitations identified in the literature review, domain requirements as well as the limitations on missing values highlighted in Chapter 4, the Underwriting Knowledge Graph is proposed to address these challenges. This methodology is chosen due to several critical reasons:

- Insurance underwriting data contains varied features and is tailored for each customer, requiring a flexible and adaptable approach,
- The structure of the dataset adheres to a unique feature pattern, indicating knowledge graphs as an intuitive and effective methodology for encapsulating multidimensional data relationships,
- The personalised essence inherent within the dataset is ideal for the application of knowledge graphs. Building the knowledge graph from preexisting data facilitates the discernment of individualised connections for each customer and the exploration of regulatory frameworks through the prism of graph mining techniques.

In summary, a knowledge graph stands as a formidable instrument for managing complex insurance data, modeling convoluted relationships, and providing bespoke insights tailored to individual clientele. The representation of multilabel relationships within the knowledge graph improvesthe precision and pertinence of our analysis, enhancing

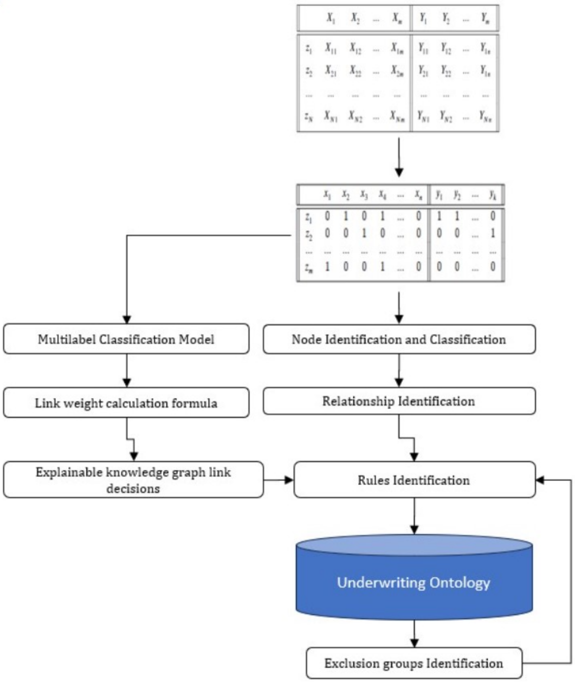

Multilabel Classification Model
Node Identification and Classification
Link weight calculation formula
Relationship Identification
Explainable knowledge graph link decisions
Rules Identification
Underwriting Ontology
Exclusion groups Identification

Figure 5.1: Underwriting Ontology Workflow Design

the methodology's robustness and insightfulness. The following sections captures the exploration into building a knowledge graph for underwriting.

5.2 Key Contribution

This section of the research study contributes directly to the research contribution RC1 highlighted in Chapter 1. From the initial construction, I propose the inception to the link prediction task with the use of metrics taken directly from the multilabel classification model to resolve research problem RP3, hence contributing to research objective RC3. This can also be portrayed as using multilabel classification for knowledge graph link prediction, which has yet to be investigated within the industry sector. In the following subsections, the details of this application and improvement are explained.

5.3 Methodology

5.3.1 Ontology Design

This section describes the entire process of building the UKG, integrating underwriting data with the underwriting rules manual. This methodology involves three primary stages: 1) constructing the initial graph, 2) applying explainable exclusion classification for predictive linkage, and 3) identifying rules and perpetually refining the knowledge graph.

5.3.1.1 Initial Graph Construction

Underwriting data undergoes collection via customer survey questionnaires, subsequently undergoing preprocessing and one-hot encoding to convert categorical attributes into binary format, thereby denoting the presence (1) or absence (0) of each attribute. The result processed dataset assumes a tabular format, with individual customers represented in rows and attributes represented in columns. Each entry is associated with a distinct policy number, thus constituting the customer node within the knowledge graph. Attributes are depicted as nodes, and associations between customers and attributes are forged based on the binary values enshrined within the dataset.

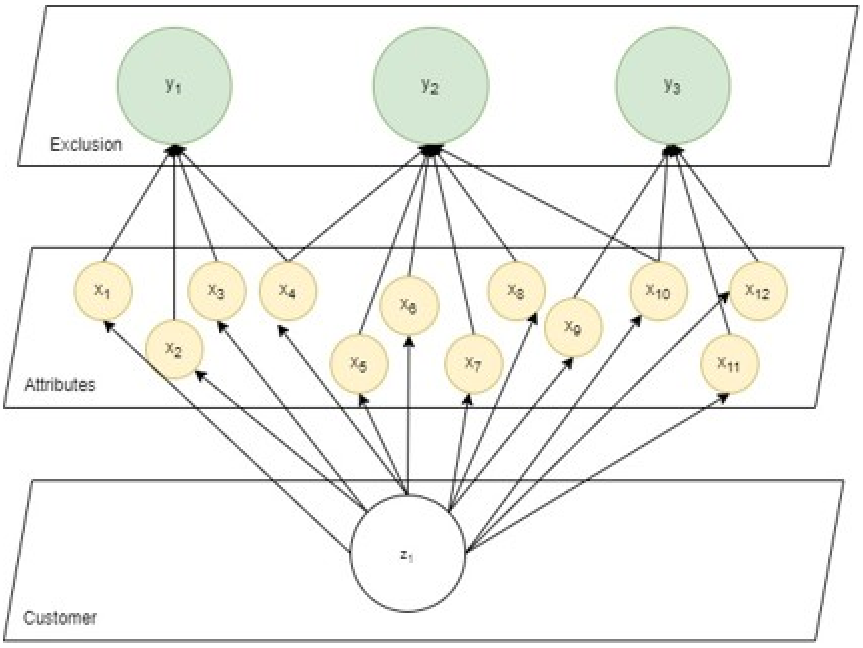


Figure 5.2: Node layers of UKG

A pivotal requisite of the UKG entails the portrayal of underwriting data attributes alongside the representation of underwriting rules distilled from the underwriting manual. Figure 5.2 illustrates the three node types in this knowledge graph - customer, attributes, and exclusion, defined as follows:

- The customer node uniquely identifies each customer;
- Attribute nodes emerge through the process of one-hot encoding the underwriting customer disclosure dataset, with each attribute endowed with characteristics signifying the rule groups to which they pertain in the underwriting manual; these properties serve as a reference point used as a benchmark to compare the top 20 attributes that influence the decision to assign a label;
- The exclusion node mirrors the target label exclusion codes enumerated in the underwriting manual.

The graph includes three types of relationships that are crucial:

- The customer-attributes relationship represents the personalised attributes pertinent to an individual's risk profile.
- The attributes-exclusions relationship delineates the attributes contributing to the classification of one or more exclusions, which is imperative for discerning rules. While most of these associations can be inferred from the tabular dataset, comparisons between attribute nodes and exclusion nodes must also be made against rules stipulated in the underwriting manual. The links between attribute nodes and label nodes are initialised with a predetermined weight, reflecting the initial significance of the attribute to the decision-making process.
- The customer-exclusions relationship takes precedence in our link prediction task. This connection is classified into three categories: "potential" links, "accurate" links, and "outlier" links, serving as a foundation for establishing a benchmark and adjusting the weight of links for any novel predictions integrated into the knowledge graph.

5.3.1.2 Link Prediction for multilabel classification

The aim of this phase is to ascertain the suitable exclusion code(s) to be assigned to a customer's policy, thereby mitigating insurance risks for the company. This entails training the processed dataset utilising a multi-label classification methodology. Multi-label classification operates akin to single-label classification, but it accommodates multiple target labels concurrently [60]. The dataset is partitioned into subsets, each corresponding to a distinct label within the original dataset, enabling the classifier model to glean insights and identify key attributes influencing the classification for that particular label.

The training phase yields n classifiers, each aligned with one of the n target labels (exclusion codes) in our dataset. These classifiers, denoted as C_y in Figure 4.3 above and Figure 5.3 below, are term label classifiers. uring the classification phase, the trained model predicts the labels to be assigned to an unlabelled record by initially estimating the applicable exclusion codes, followed by refining these predictions using the label classifiers.

Figure 5.3 illustrates how the multi-label classifier addresses the link prediction problem within the KG. The UKG also utilises the explainable results from the QAR to identify prospective link prediction opportunities. The QAR outcomes aid in identifying "potential" links (representing missed exclusions), "accurate" links (the exclusion classifications by human underwriters align with the model results), and "outlier" links (indicating unnecessary exclusions). Essentially, the current predicament can be conceptualized as a graph feature problem, with $z \in Z$ symbolises the customer node, $x_z \subseteq X$ symbolises an attribute node for z, and $y_z \subseteq Y$ symbolises an exclusion node for the

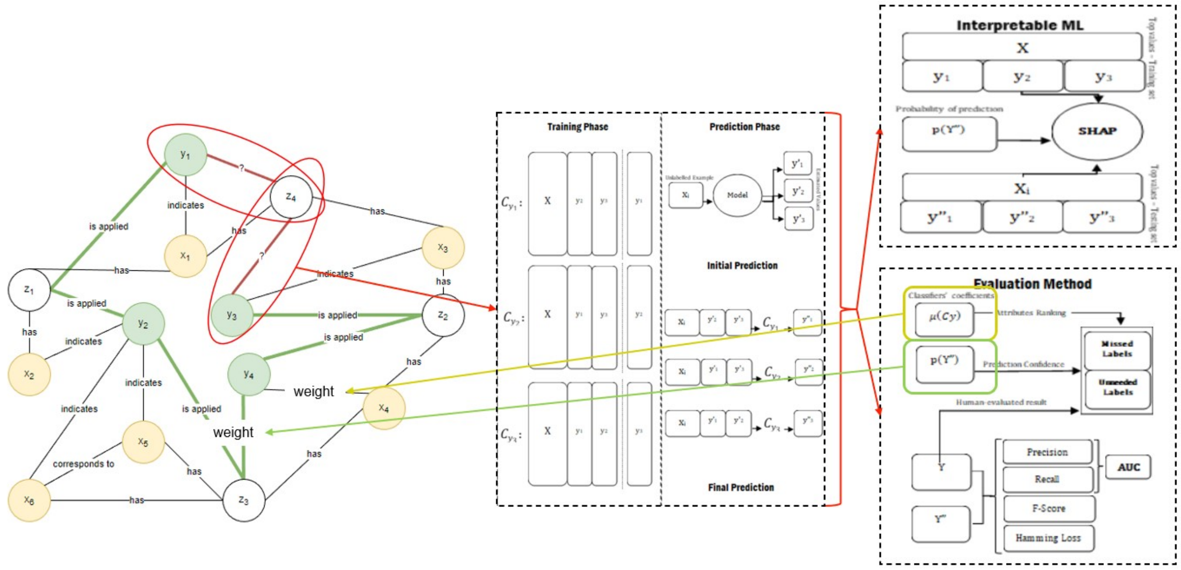


Figure 5.3: Link prediction

customer z. Consider a set of relationships as (h,r,t):

$$z_x = f(h_z, r, t_x) \quad \text{with } r \text{ as a constant} \tag{5.1}$$

$$x_y = f(h_x, r', t_y) \quad \text{where } r' = \mu(C_y) \tag{5.2}$$

$$z_y = f(h_z, r'', t_y) \quad \text{where } r'' = p(y) \tag{5.3}$$

The link weight ω_{zy} is calculated as:

$$\omega_{zy} = \lim_{p(y) \to p(y_\delta)} \frac{\sum_n (f(h_z, r, t_{x_n}) + f(h_{x_n}, \mu(C_y), t_y))}{f(h_z, p(y), t_y)} \tag{5.4}$$

In other words, for the first attempt of the UKG construction, the metrics calculated from the model in Chapter 4 is applied directly as the weight for each link with a slight moderation using the QAR results. In particular, the proximity of a new customer-exclusion association to a classified link (represented as δ) determines the likelihood of this new link being assigned a weight akin to that of the classified link. For instance, upon detecting a resemblance to an "outlier," the original weight of the link undergoes adjustment and reduction according to the provided equation.

5.4 Construction of Knowledge Graph

5.4.1 Exclusion Classification and Link Prediction

The initial graph is constructed using the results assimilated from running the multi-label classification model from the previous study [63] in Chapter 4. The initial graph includes three classes of nodes as stated in previous section:

- Customer: each instance of the customer node is identified by enquiry ID under the assumption that each customer has 1 enquiry ID,
- Attributes: instances of attribute nodes are represented by appending question name to the answer given to include information included in each enquiry under the same attribute class,

- Exclusion: instance of this are exclusion nodes applied to the customer.

Based on the relationship identified through the enquiry data, three types of triples are added in with its weight:

- (customer, has, attribute)
- (attribute, leads to, exclusion) with $\omega_{leads-to} = \mu(C_y)$
- (customer, receives, exclusion) with $\omega_{receives} = p(y)$

Table 5.1: Simplified example of knowledge graph triples constructed for UKG in a case of lumbar spine exclusion

Source	Link	Weight	Destination
Customer 1	has	0	BackDisc Location - Lower back (Lumbar Spine)
Customer 1	has	0	BackDisc SepOcc - More than twice or continuously
BackDisc SepOcc - More than twice or continuously	leads to	0.85	SPIN-LSS
Customer 1	receives	0.9	SPIN-LSS

In particular, after incorporating QAR results to the relationships of the graph, the (customer, receives, exclusion) triple is then divided into three subtypes:

- (customer, receives accurate, exclusion): links that have been previously identified by the underwriters that the model result agrees with.
- (customer, receives potential, exclusion): links based on missed report on the QAR, meaning the exclusions that the model has picked up that was not identified by the underwriters.
- (customer, receives outlier, exclusion): links previously identified by underwriters that the model disagrees with.

5.4.2 Automated knowledge graph creation, update and maintenance

The outcome of the multi-label classification, comprising the training set, testing set, and QAR outcomes, is employed to ascertain the guidelines for establishing the UKG. By using a customised Python script that leverages the NetworkX library, the encoded dataset is loaded into the graph with predefined links. A custom function then converts the graph into RDF triples for standardized application, enabling a semi-automated approach to building a knowledge graph compared to the traditional Web Ontology Language (OWL) process. Prior to being utilised in the construction of the UKG, the QAR findings necessitate reviews and validation from human underwriters.

Upon the addition of each fresh entry to the knowledge graph post-creation, the link weight is computed employing equation 5.4 in conjunction with the existing entry possessing the highest match score. Following the adjustment of the weights for the new entries, all corresponding attribute-label links are revised to generate an updated roster of attributes exerting the most pronounced influence on a specific label.

At this stage, the initial graph includes three layers of nodes, as depicted in Figure 5.2. Figure

5.5 Results Analysis and Discussion

This section presents the findings from our experiments, which involved two primary tasks:

- Evaluating the multilabel link prediction task using QAR to establish a benchmark for identifying new rules,
- Verifying the benchmark results through the knowledge graph update method.

Table 5.2 provides an example of our evaluation. This process suggests that in order for the rules identification and update process to commence, the new record z and its most akin existing record δ must achieve a minimum similarity score of **0.8** to uphold the existing rules derived from the underwriting manual. Setting the benchmark too low (e.g., 0.6) results in merely 7 attributes from the top 20 impact list being recognised as relevant to the exclusion code, which corresponds to the initial list. On the other hand, if the benchmark is too high (above 0.9), the top 20 impact list shows little change, potentially reinforcing existing biases in the UKG and limiting its adaptability. This pattern was

Table 5.2: Changes between top 20 impact attribute list when using different benchmarks for exclusion code SPIN-LSS

Benchmark	Attributes similar to initial top 20 impact	Attributes related to label (out of 20)
0.6	7	7
0.7	11	14
0.8	17	18
0.9	18	20
0.95	20	20

observed in 96 out of 108 exclusion codes evaluated. Labels deviating from this pattern were flagged for review by human underwriters, offering transparency into the decision-making process. The periodic updating of the top 20 impact list ensures the graph retains its dynamism and adaptability to diverse scenarios. Any newly incorporated attribute associated with the exclusion code is retained within the list; otherwise, it is flagged for potential rule discovery by underwriters, integrating human expertise into the process and making UKG a supportive tool for decision-making and updates to the underwriting manual.

However, because of the predetermined initial weight, the method for calculating link weight cannot be promptly applied to newly added attribute nodes within the graph. This means new attribute nodes are less likely to be considered in the impact attribute list until sufficient historical records are trained and loaded into the graph. This issue, which depends heavily on the availability of historical data, suggests that multiple updates may be necessary for a new attribute to be considered impactful. Further study and development of the UKG could focus on addressing this challenge.

Another problem that was found after this study case is the unreliability of the model, as the heavy reliance of the link weight on the model result, the case of an overfitted model may create systemic false positive results, which goes against the issues identified in Chapter 2. Moreover, this methodology has yet to provide a way for underwriting rules identification.

CHAPTER 6

Risk Profile Using Knowledge Graph And Multilabel Classification

6.1 Background and Motivation

Based on the previous approaches detailing in Chapter 4 and 5, the UKG is refined to alleviate the identified issues in order to establish a customer risk profile. The proposed methodology in this chapter is a combination of the advantage points from existing approaches to further enhance the UKG for the rules identification process.

One key factor to note is the definition of risk profile in this study is different from previous researches. Differ to prvious previous research studies that aim to classify customers into overall risk classes, the risk profile in our study aims to pinpoint which particular aspects contribute to the customer's risks, based on the preliminary statistics on exclusion code application mentioned in Chapter 3, those with exclusion codes are in high risks factor with regards to their exclusions. In other words, instead of ranking a customer based on an overview high risk and low risk, we are ranking customers with regards to the specific exclusion cases they fall into. This objective is achieved by combining the explainable exclusion QAR introduced in Chapter 4 as a part of the reasoning process and updating the link weight calculation method for rules identification and suggestions.

6.2 Key Contribution

Although the application of proposed methodology for this use case can be assigned to the risk classification task of automated underwriting, this study elevates from existing researches with a more sophisticated risk classification system based on explainable exclusion. In comparison to the existing approach of classifying customers into high-low risk classes, the categorisation of the classification process using UKG focuses on the exclusion codes, providing a more sophisticated and personalised view into which exact risks should be paid attention to.

This methodology assist in resolving the existing research problems in several ways:

- The UKG at this stage should have a view of the attributes linked to the customers as well as the exclusion code that is applied to them (divided into three types: accurate, potential, outlier). This aligns to research contribution RC1.
- Multiple additional nodes and relationships are defined in comparison to the existing initial graph to assist with the auto-construction of the inferred relationships to represent the business rules to resolve RO2.
- Provide transparency on which features have an impact on the link prediction result based on QAR and graph mining in regards to RO3.

6.3 Revised Methodology

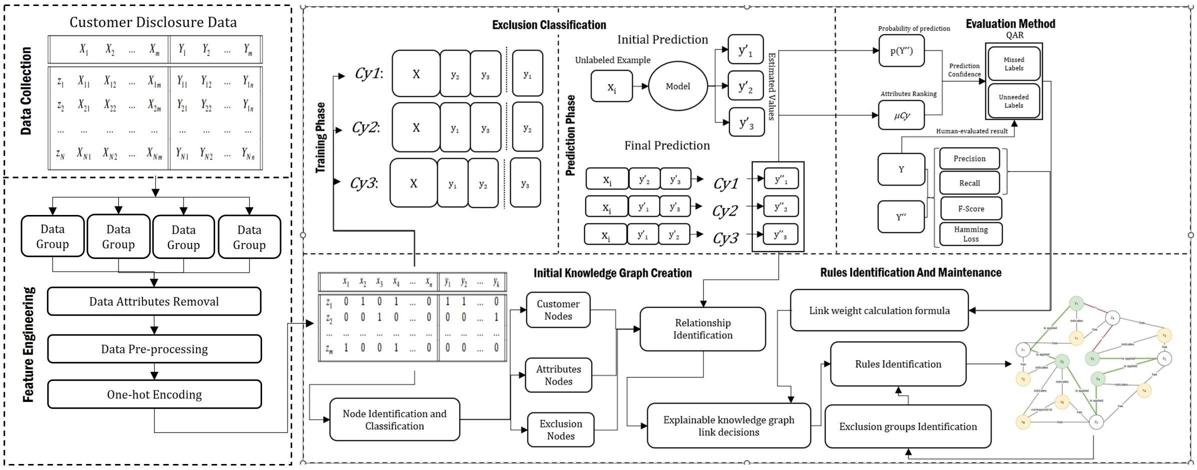


Figure 6.1: Model Methodology

6.3.1 Automated UKG Construction

Following the construction of the initial UKG from previous section, multiple new nodes and relationships have been added in to deeper layers of the UKG. Table 6.1 depicts the current triples within the UKG. In the revised version, attribute nodes in the previous model are broken down into question and question answer to better represent the question groups. Predicting and calculating the link weight for the relationships between question answer - exclusion code and customer - exclusion code remain the priority.

Table 6.1: Revised relationships in UKG

Source	Link	Destination
Customer	gives answer	Question answer
Question answer	is an answer of	Question
Question	belongs to	Question group
Question answer	**leads to**	**Exclusion code**
Exclusion code	belongs to	Exclusion type group
Exclusion code	is applied to	Policy number
Policy number	is provided to	Customer
Customer	**has**	**Exclusion code**

At this stage, the process of creating nodes and links within the UKG is enhanced by separating the data to multiple subsets. Each of these subsets are then tranformed into subgraphs parallelly by using multiprocessing for faster runtime prior top join multiple subgraphs together to create one centralised knowledge graph.

6.3.2 Explainable Exclusion

This phase is a direct application of the explainable exclusion presented in Chapter 4. Section "Exclusion Classification" in 6.1 provides further details on how the multi-label classifier tackles the link prediction issue within the UKG. Prior to integrating the multi-label classification outcome into the UKG, we generate the explainable exclusion result utilising the QAR. In essence, the QAR utilises the classification probability of each label in a data row from the testing set and the feature importance determined by each label's classifier to identify the "missed" and "unneeded" labels within a multi-label classification, based on a specific threshold. Subsequently, these are translated into "potential" and "outlier" links within our UKG.

6.3.3 Rules Identification and Knowledge Graph Maintenance

This section remains the most delicate part of the automation process. Traditionally, both domain-specific knowledge graphs rules reasoning generally and underwriting rules identification requires business expertise and manual insertion, providing concerns as well as room for development.

The rules identification analysis in this investigation concentrates on two primary focus:

- Establishing a baseline reference point for rule identification based on the underwriting manual, whereby all freshly formulated rules link weight are compared with this benchmark within an allowed threshold to ensure consistency with established guidelines.

- Employing a weight calculation approach to prioritise nodes that substantially impact link predictions, bridging the knowledge gaps in comprehending the significance of each attribute to a customer's risk profile.

Given the necessity for the UKG to be adaptable and capable of seamlessly integrating new data, a pioneering algorithm is proposed for maintaining the knowledge graph. This includes updating both new and existing nodes and refining the rules analysis to highlight potential pattern-based rules deemed significant to underwriters.Upon the addition of a new customer record, the knowledge graph undergoes updates to incorporate the associated attributes and relationships of the new customer node. This process also entails revising existing nodes, recalibrating the weights of their relationships to mirror the evolving relevance of attributes to each customer's risk assessment. Consequently, the knowledge graph serves as a continuously evolving tool, perpetually enhancing and fine-tuning the rules identification process as new data is assimilated.

The previous experiment shows that coefficients and feature importance alone is **insufficient** to identify the link should a model be overfitted (which is highly likely with the case of multilabel classification). Hence, the formula adjustments are introduced to further improve the results while still preserving the the transparency of the output. The adjustments construction is inspired and derived from Bayesian statistical theory [32, 32] to account for the ever-evolving potentials of the UKG through each new data batch load. The inherent semi-automated mining functions of the UKG's graph database structure can also be leveraged using this traditional approach.

6.3.3.1 Weight adjustments: attributes to exclusion code

Consider the following values:

- $\mathbb{FI}(X_z \to Y_z)$ as the feature importance of attribute x_z for each label y_z from our multilabel classification model,
- n as the number of similar neighbors of customer z
- $\rho_{(x\to y)} = \frac{|z\in Z:\{z\to y, z\to x\}|}{|z\in Z:\{z\to x\}|}$ as the probability of a customer answering a question having the exclusion

We can formulate the calculation as:

$$\omega_{xy} = \rho_{(x\to y)} \times \mathbb{FI}(x_{1\to n} \to y_{1\to n}) \quad \text{if } \mathbb{FI} > 0 \tag{6.1}$$

6.3.3.2 Weight adjustments: customer to exclusion code

With:

- $J_{1\to k}(z) = \text{range}\left(\lim_{1\to k}\left(J(z, z'_k)\right)\right)$ as the Jaccard similarity score range of customer z
- Top k most similar customers based on features existence to identify $Z(z_\delta) \subset Z$ as most similar customers to the existing
- $\rho_{(z\to y)}$ as prediction probability from the model

Then:

$$\omega_{zy} = \text{range}\left(\lim_{1\to k}\left(\frac{\sum_1^k \omega_{z_\delta y}}{k} \times P_{zy}\right)\right) \quad \text{if } P_{zy} > 0 \tag{6.2}$$

Using this weight adjustment, instead of mapping a new customer to the most similar existing customer, a group of k most similar customers are established for comparison.

6.4 Experiment Setup

Figure 6.2 outlines the process flow for the UKG set up across three main phases: Data Collection and Preprocessing, Exclusion Classification and Initial Graph Creation, and Rules Identification and Maintenance. The experiment employs a structured, tripartite framework to synthesize and refine a knowledge graph from customer survey datasets

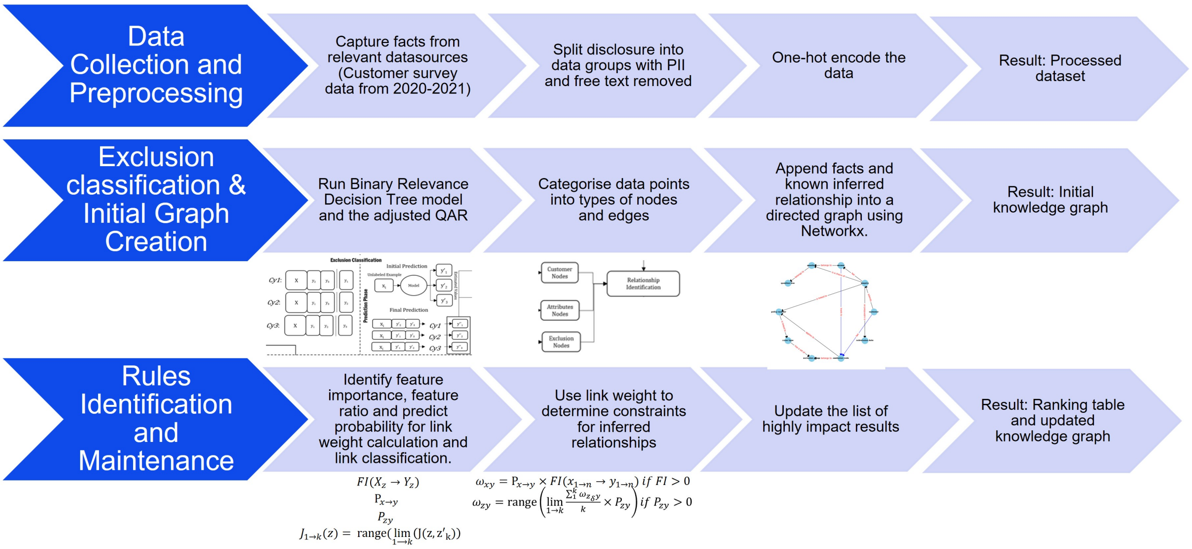

Data Collection and Preprocessing
Capture facts from relevant datasources (Customer survey data from 2020-2021)
Split disclosure into data groups with PII and free text removed
One-hot encode the data
Result: Processed dataset
Exclusion classification & Initial Graph Creation
Run Binary Relevance Decision Tree model and the adjusted QAR
Categorise data points into types of nodes and edges
Append facts and known inferred relationship into a directed graph using Networkx.
Result: Initial knowledge graph
Exclusion Classification
Initial Prediction
Final Prediction
Customer Nodes
Attributes Nodes
Exclusion Nodes
Relationship Identification
Rules Identification and Maintenance
Identify feature importance, feature ratio and predict probability for link weight calculation and link classification.
Use link weight to determine constraints for inferred relationships
Update the list of highly impact results
Result: Ranking table and updated knowledge graph
$FI(X_z \rightarrow Y_z)$
$\mathrm{P}_{x \rightarrow y}$
P_{zy}
$J_{1 \rightarrow k}(z) = \mathrm{range}(\lim_{1 \longrightarrow k} (\mathrm{J}(\mathrm{z}, \mathrm{z}'_{\mathrm{k}})))$
$\omega_{xy} = \mathrm{P}_{x \rightarrow y} \times FI(x_{1 \rightarrow n} \rightarrow y_{1 \rightarrow n})\ if\ FI > 0$
$\omega_{zy} = \mathrm{range}\left(\lim_{1 \rightarrow k} \frac{\sum_1^k \omega_{z_\delta y}}{k} \times P_{zy}\right) if\ P_{zy} > 0$

Figure 6.2: Experiment setup

spanning 2020-2021. The initial phase, Data Collection and Preprocessing, involves the extraction of customer information from the data as well as removing personally identifiable information to uphold data privacy norms, followed by the application of one-hot encoding to facilitate computational data wrangling for a structured dataset. The second phase, Exclusion Classification and Initial Graph Creation phase utilise the data from the first phase to run the Binary Relevance Decision Tree multilabel classification model, along with creating the QAR metric. In parallel, the structured one-hot encoded data is also used for the creation of an initial knowledge graph. The final stage, Rules Identification and Maintenance, is characterized by a rigorous analytical regime that evaluates feature importance, ratio, and predictive probability, instrumental in calibrating link weight for nuanced link classification within the graph. The result of this phase is a maintained UKG, of which data can be mined and extracted for evaluation.

6.5 Results Analysis

As the UKG can be utilised as a graph database, to properly evaluate its functions for customer risk profiles, three main points are examined in alignment with the research objectives identified in section 1.5. The angles of examination analysis are determined below:

- Customer risk profile using classified exclusion codes and their attached link types (aligns with **RO1**).
- Rules identified via automatic rules identification after graph adjustments (aligns with **RO2**).
- Updated explainable exclusion after graph adjustments (aligns with **RO3**).

These points are elaborated in the following subsections.

6.5.1 Customer risk profile using UKG

Customer ID	Exclusion code	Classification	Link weight
a8c92c4d-0b82-4431-ae6b-72d7081c01ba	MSKL-RAN	accurate	1
a8c92c4d-0b82-4431-ae6b-72d7081c01ba	PSTM-SCU	accurate	1
a8c92c4d-0b82-4431-ae6b-72d7081c01ba	PDIC	accurate	1
a8c92c4d-0b82-4431-ae6b-72d7081c01ba	MEDI-TIO	potential	0.82

Table 6.2: Customer risk profile captured by the UKG for a sample customer

The use of exclusion codes linked to customer nodes in UKG to portray an overall view of a customer's risk profile is paramount. This categorisation aids in refining the assessment of the risk associated with a customer using their relevance and accuracy - ranging from directly applicable ("accurate link") from the human underwriter decisions, to those that might be speculative ("potential") or even unneeded ("outlier") as specified in Chapter 6. The examples provided in Table 6.2 represents the exclusion codes and link weights connected to a customer as their risk profile, from which the risks related to a customer are portrayed not just on a blanket-level of risk classification but detailed into the different categories of the exclusion code, which allows for multiple risk factors to be shown and ranked. For existing customers, all links have been verified by human underwriters with accurate links imported from the historical data, as well as potential and outlier links imported from QAR. To put it into perspective, the UKG assists in managing retained customer with updated link weights for existing links, creating an improved personalised ranking system for existing customers. Whereas in the case of a new customer, as stated previously in Section 1.2, UKG's provided ranking and QAR aims to assist underwriters in their decision-making, meaning that the new links resulted from the UKG should be presented to human underwriters before the final decision outcome.

6.5.2 Automatic rules identification after graph adjustments

Secondly, the impact of modifications within the knowledge graph on the hierarchy of rules is another vital area of focus. Adjustments to the graph can lead to shifts in the ranking of rules, which in turn might alter the interpretation and application of these rules in risk profiling. This dynamism necessitates continuous monitoring to ensure that the rule application remains valid and reflective of the current data structure and insights.

Link weight serves as a pivotal metric in bench-marking the underwriting knowledge graph. By analysing the distribution and magnitude of link weights between nodes, we can identify potential rule patterns and correlations. Higher link weights signify stronger relationships between nodes, indicating significant factors influencing underwriting decisions. By scrutinizing the link weight distribution, we aim to uncover hidden insights and refine decision-making processes.

Definition	Accurate Link	# of Potential Link	# of Unnecessary Link
Initial KG	Inferred relationships formed either from preliminary data analytics or graph mining	NA	NA
After graph mining adjustment		NA	NA
After multilabel model adjustment	Edges predicted accurately by the model	Edges predicted as missed	Edges predicted as unneeded
After link weight adjustment	Add/remove edges based on weight benchmark	Add/remove edges based on weight benchmark	Add/remove edges based on weight benchmark

Table 6.3: Graph edges after each adjustments

From the previous study conducted (refer to Section 5), a benchmark is required for the suggested rules to be implemented. To remove statistical coincidence, question answers that are connected to general questions groups (such as basic details, questions marked as general) will be excluded from this process.

6.5.3 Updated explainable exclusion from graph mining

Lastly, identifying the specific attributes that most significantly influence the decision to exclude a customer from services is crucial. For a more granular understanding, examining both commonly and rarely invoked exclusion codes provides valuable insights. For instance, a common health-related exclusion code might be routinely applied and serve as a standard metric for exclusion. In contrast, a mental health exclusion code applied less frequently - such as less than ten times in the past two years - could indicate a nuanced criterion that requires specific conditions to be met. Analysing these attributes helps in pinpointing the decisive factors in customer risk profiling, thus enabling more targeted and effective risk management strategies.

In our previous research in Section 5, the initial UKG weight is calculated using only the model's classifiers' confidence ($\mu(C_y)$), which is prone to misleading results should the model be overfitted. This section focuses on comparing the results between the two versions of the UKG for room of improvement among common and less common exclusion codes. The classification of common and less common exclusion codes is based primarily on frequency statistical analysis, meaning an exclusion code is defined as common or less common based on how often they are applied within the course of 2 years.

6.5.3.1 Handling Common Exclusion Codes

As common exclusion codes are those that are frequently applied, the data available for these groups are more in-depth for UKG attribute-exclusion evaluation. Table 6.4 puts the top 20 attributes related to exclusion code MEDI-EYB (medical - eyeball) coming from the initial UKG in Section 5 in comparison with the updated UKG. The initial UKG link weights are normalised from the model's classifiers' feature importance that ranks over 6000 data attributes, whereas the updated UKG uses the formula 6.1 presented above. In this case, the updated UKG results portrayed more attributes directly related to visions, with first 12 having the link weight of 1, meaning that the multi-label classification model ranked these attributes as unimportant even though the majority of customers having exclusion code MEDI-EYB assigned have these attributes. Another point worth noting is that 75% of the top 20 attributes related to MEDI-EYB all comes from the question group "WHICH VISION", with additional attributes on benign paroxysmal positional vertigo (BPPV) receiving the maximum link weight, whose connection to nystagmus - a condition in which the patient shows involuntarily repetitive rapid eye movements - has been proven [15, 28].

Initial UKG Ranking	Attribute	Link Weight (scaled)	Updated UKG Ranking	Attribute	Link Weight
1	WHICH VISION - Keratoconus	1	1	WHICH VISION - Partial loss of vision	1
2	WHICH VISION - Glaucoma Mild	0.74	2	WHICH VISION - visual snow	1
3	VISION - Yes	0.17	3	WHICH VISION - vitelliform macular dystrophy	1
4	CYSTS GROWTHS - Yes	0.11	4	WHICH VISION - Macular degeneration	1
5	WHICH VISION - Optic neuritis	0.11	5	WHICH VISION - Papilloedema	1
6	WHICH VISION - Ocular Albinism	0.09	6	WHICH VISION - Loss of vision	1
7	FAMILY HISTORY - Don't Know	0.08	7	WHICH VISION - Optic atrophy	1
8	Back Physio Date - 5+yrs	0.08	8	WHICH VISION - Congenital caratact	1
9	ALCOHOL ADVICE - No	0.07	9	WHICH BACK PAIN - Mild muscular pain in back	1
10	WHICH VISION - Corneal transplant	0.07	10	WHICH HEARING - Benign paroxysmal positional vertigo	1
11	Cystit Last Symptom - 5+yrs	0.07	11	Cataract Operation When - 1-2years	1
12	WHICH BACK PAIN - Ache in back	0.07	12	Sarcoid Diagnosis When - 2-3 years	1

13	WHICH VISION - High pressure in eye	0.07	13	WHICH VISION - Glaucoma Mild	0.74
14	WHICH VISION - Detached lens	0.07	14	WHICH VISION - Blepharitis	0.50
15	RECENT TREATMENT - No	0.07	15	WHICH VISION - Uveitis	0.50
16	GHQ Diagnosis Date - 5+yrs	0.07	16	Haematur Last - 3-5years	0.50
17	Osteoa Sx Date - 5+yrs	0.07	17	WHICH VISION - Dry Eyes	0.33
18	MENTAL SERIOUS - No	0.06	18	WHICH VISION - Central serous retinopathy	0.33
19	KIDNEY FEMALE - No	0.01	19	WHICH VISION - Inflammation of eye	0.33
20	CONGENITAL - No	0.01	20	WHICH VISION - Keratoconus	0.27

Table 6.4: Attribute rankings before and after updated UKG for exclusion code MEDI-EYB

Considering the potential data saturation created by the popularity of the common exclusion codes, question group nodes have been added to the underwriting knowledge graph to bring more in-depth layers to the attributes connected to these codes. The view of question group alongside the attributes itself, especially in health exclusion codes, provides a higher level view of which potential factor groups are contributing to the exclusion code assignment, similar to a diagnosis process. This approach enables a more granular analysis of attributes associated with highly prevalent exclusion codes, facilitating targeted risk assessment and decision-making.

6.5.3.2 Handling Less Common Exclusion Codes

For less popular exclusion codes, we apply a similar evaluation process by comparing the initial top 20 most popular attributes tied to the exclusion code with the updated top 20 exclusion code attributes. This analysis provides insights into shifting trends and emerging risk factors associated with less frequently encountered exclusion codes. By identifying changes in attribute distributions, underwriters can adapt their strategies accordingly and mitigate potential risks. Table 6.5 shows the top 20 attributes contributed to the exclusion code MEDI-GOS (medical - esophagitis), which is applied less than 20 times over the last 2 years.

Initial UKG Ranking	Attribute	Link Weight (scaled)	Updated UKG Ranking	Attribute	Link Weight
1	WHICH LUNG - Nose blockage	1	1	WHICH LUNG - Nose blockage	0.5
2	Occupation Change - 0-6mths	0.83	2	WHICH BACK PAIN - Scoliosis	0.07
3	Chest Pain Last Symptom - 1-2yrs	0.17	3	WHICH HEART - Chest pain	0.004
4	NA	0	4	GHQ Treatment Date - 1-2yrs	0.004
5	NA	0	5	GHQ Diag Date - 1-2yrs	0.03
6	NA	0	6	Cholesterol Test Date - 6-12mths	0.02
7	NA	0	7	GHQ Symptom Date - 1-2yrs	0.02
8	NA	0	8	Mental Health Medication Last - 3-5yrs	0.02
9	NA	0	9	Mental Health Talk Therapy Last - 1-2yrs	0.01
10	NA	0	10	Shoulder Treatment Date - 5+yrs	0.006
11	NA	0	11	Shoulder Symptoms Date - 5+yrs	0.005
12	NA	0	12	Occupation Employed Start - 1-2yrs	0.004
13	NA	0	13	WHICH SKIN - Eczema	0.004

14	NA	0	14	Occupation Change - 0-6mths	0.003
15	NA	0	15	Heart Problems - Yes	0.003
16	NA	0	16	Back Last Sx - 1-2yrs	0.003
17	NA	0	17	Bowel - Yes	0.003
18	NA	0	18	Precancer Fem - Yes	0.002
19	NA	0	19	Cancer - Yes	0.002
20	NA	0	20	Chest Pain Last Symptom - 1-2yrs	0.01

Table 6.5: Attribute rankings before and after updated UKG for exclusion code MEDI-GOS

A recurring pattern that can be seen among less common exclusions in the initial UKG is that most of these exclusions only has less than 5 attributes with feature importance greater than 0, meaning the results might be highly skewed and fairly unreliable for rules prediction. This is a point of improvement from the updated UKG that resolved the limitation on less common exclusion mentioned in Section 1.2. In particular, the initial UKG only list three seemingly unrelated answers to attribute to the exclusion code, whereas the updated UKG provides a better overview of attributes with attributes related to back pain, heart pain and mental health, all are either symptoms related to esophagitis conditions or a condition that esophagitis has been found to be connected to [20, 21, 58, 59]. As the link weight is calculated using formula 6.1 based on the probability of a customer answering this question having the exclusion code, due to the limited samples provided for this exclusion code, it is mathematically understandable that the link weights are significantly lower than their common counterpart while keeping the same logic to provide better ranking results.

6.6 Summary

The analysis of exclusion codes in risk profiling identifies key attributes influencing customer exclusions. Common exclusion codes, such as those for health issues, have more data available, allowing detailed insights into contributing factors, like vision-related conditions. Less common codes are more difficult to analyse due to limited data, but updates to the Underwriting Knowledge Graph (UKG) improved attribute ranking and relevance. This refined approach enhances decision-making by enabling a clearer understanding of both prevalent and rare exclusion codes, leading to better-targeted risk management strategies. This method follows the logic structure of

CHAPTER 7

Discussion of Approaches Taken, Use Cases and Future Work

7.1 Results Finding

The study is based on a case study of our industry partner using real-life insurance dataset from 2019-2021, with each iteration trying to resolve the limitations found in the previous one. Comparing to previous attempts to build a personalise reliable customer risk profile using underwriting data, the UKG has been able to resolve the limitations set out in Section 1.3 and 2.6, with the most apparent improvement in explainable exclusion and customer risk profile stated in Section 6.5. The upgraded version incorporates refined algorithms and data processing techniques to enhance the accuracy and reliability of link weights. By comparing the two versions, we can assess the effectiveness of these enhancements and identify areas for further improvement.

7.2 Use Cases

7.2.1 Recalculating life table

In the dynamic landscape of insurance underwriting, the integration of an underwriting knowledge graph presents a transformative approach to refining life tables, a cornerstone for risk assessment and premium determination. This innovative use case involves recalculating life tables by leveraging the nuanced link weights between various attributes and exclusion codes within the graph. Essentially, each attribute - such as age, medical history, or lifestyle choices - and its associated exclusion codes are interconnected within the graph with specific weights that signify their impact on mortality risk. By analyzing these weighted links, insurers can dynamically adjust life tables based on real-world data and emerging trends, ensuring more accurate, personalized, and fair pricing for policyholders. This methodology not only enhances the precision of risk assessment but also allows for more granular insights into the factors affecting longevity, thereby refining the actuarial models that underpin the entire insurance industry.

7.2.2 Adding a new customer

Algorithm 1 Pseudo Code: Adding a new customer

```
Input: New customer records with question answers $X$ into the knowledge graph
Output: Calculated link weights between customers and exclusion codes

1. Load new customer records with question answers $X$ into the knowledge graph.
2. For each customer $c$ in the knowledge graph:
     a) For each question answer $qa$ of $c$:
          i. If $qa$ node does not exist in the graph:
               A. Create a new question answer node $qa$ and connect it to the customer
                  node and related question group.
          ii. Connect the customer node to existing attribute nodes if the $qa$ node
              already exists in the graph.
3. Find customers with the most similar record based on similarity in question
   answers.
4. Calculate link weights using the formula:
     • Weight adjustments: customer to exclusion code
     • Input: Jaccard similarity score range $J_{1\to k}(z)$, top $k$ most similar customers
       $Z(z_\delta)$, prediction probability $\rho(z \to y)$
5. For each customer $z$:
       $\omega_{zy} = \text{range}\left(\lim_{1\to k}\left(\frac{\sum_1^k \omega_{z_\delta y}}{k} \times P_{zy}\right)\right) \quad \text{if } P_{zy} > 0$
6. Create a link between the customer node and the exclusion code node.
7. Create links between the customer node and existing exclusion code nodes based
   on QAR results.
8. Return calculated link weights between customers and exclusion codes.
```

Incorporating a new customer with unique attributes and a new exclusion code into an underwriting knowledge graph represents a significant advancement in personalising insurance policies and streamlining the underwriting process. This use case highlights the graph's capability to assimilate new information seamlessly, enabling insurers to tailor their risk assessment and policy pricing with unprecedented precision. When a new customer's data enters the system, bringing along its attributes (question answers), UKG dynamically integrates this information using Algorithm 1, expanding its nodes and edges to reflect these additions. This expansion not only enriches the graph with fresh insights but also aids in identifying patterns and correlations among attributes and risks. Consequently, insurers can offer policies that are more accurately priced according to the individual's unique risk profile, fostering a more equitable and efficient underwriting process. This adaptability ensures that the knowledge graph remains a robust, evolving tool that continuously enhances the accuracy and fairness of insurance underwriting.

For the application of Algorithm 1, let:

- n be the number of customers in the knowledge graph.
- q be the number of question answers per customer.
- k be the number of most similar customers considered for weight adjustments.
- m be the number of exclusion codes.

The computational complexity for each part of the algorithm is as follows:

- Adding customer records: $O(n \times q)$
- Finding similar customers: $O(n^2 \times q)$
- Calculating link weights: $O(n \times k)$

Based on the analysis above, the most expensive part in terms of computational complexity within the function lies in finding similar customers, which can be a point of improvement for future iterations.

7.2.3 Adding a new rule

Algorithm 2 Pseudo Code: Adding a new rule

```
Input: New customer records with question answers $X$ and exclusion codes $Y$ into the knowledge graph
Output: Calculated link weights between attributes and exclusion codes

1. Load new customer records with question answers $X$ and exclusion codes $Y$ into the knowledge graph.
2. For each customer $z$ in the knowledge graph:
   a) For each question answer $x$ of $z$:
      i. If $x$ node does not exist in the graph:
         A. Create a new question answer node $x$ and connect it to the customer node and related question group.
      ii. Connect the customer node to existing attribute nodes if $x$ node already exists in the graph.
   b) For each exclusion code $y$ of $z$:
      i. If $y$ node does not exist in the graph:
         A. Create a new exclusion code node $y$ and connect it to the customer node and related exclusion group.
      ii. Connect the customer node to existing exclusion code nodes if $y$ node already exists in the graph.
3. Calculate link weights using the following formula:
   • For each attribute $x$ connected to exclusion code $y$:
        $\omega_{xy} = \rho(x \rightarrow y) \times FI(x_{1\rightarrow n} \rightarrow y_{1\rightarrow n}) \quad \text{if } FI > 0$
4. Return calculated link weights between attributes and exclusion codes.
```

The addition of a new rule based on the graph link weight using Algorithm 2, calculated by the probability of an attribute being linked to an exclusion code, showcases the dynamic and adaptive nature of underwriting knowledge graphs in refining insurance processes. This use case illustrates how the graph's architecture facilitates the incorporation of new underwriting rules as relationships or links between specific attributes and exclusion codes. When the graph algorithm identifies a statistically significant weight or probability that associates a particular attribute with an exclusion code, it can autonomously generate a new rule. This rule then becomes a critical part of the decision-making framework, guiding underwriters in evaluating risks more accurately. For instance, if data analysis reveals that a certain lifestyle choice (attribute) has a high probability of leading to a specific health condition (exclusion code), the knowledge graph automatically formalizes this correlation into a new underwriting rule. This process not only enhances the precision of risk assessment but also ensures that the underwriting rules evolve in tandem with emerging data trends, maintaining the relevance and effectiveness of the underwriting process. One additional note, to reduce the risk of data bias coming from limited sample data, before an exclusion is automatically updated using the graph function, there should be at least 5 data records of that particular exclusion code prior to adding in the graph.

For the application of Algorithm 2, let:

- n = number of customers in the knowledge graph.
- q = number of question answers per customer.
- m = number of exclusion codes per customer.
- a = number of attributes related to question answers.

The complexities for each step are as follows:

- **Load new customer records**: $O(n)$
- **Loop through each customer and their question answers**: $O(n \times q)$
- **Loop through each customer and their exclusion codes**: $O(n \times m)$
- **Calculate link weights**: $O(n \times q \times m)$

The total complexity of Algorithm 2 is $O(n) + O(n \times q) + O(n \times m) + O(n \times q \times m)$, with the potential to over-expand in the worst-case scenario that all customers answered all questions given. However, as presented in the preliminary findings in Section 3.3.2 as well as the questionnaire structure in Section 1.1.3.2, this is an impossible edge case.

7.2.4 Recalculating graph link weight

As mentioned in Chapter 6, following the concept of Bayesian theory[62], the graph requires recalculation after each new batch load of data to incorporate new attributes and update link weights. This continuous re-calibration ensures that the graph remains accurate and up-to-date with the latest information. The use of a graph structure in this model assists in reducing computational power requirements comparing to rerunning the whole system every update, as only the links attached to impacted nodes need to be recalculated. This selective updating process enhances efficiency, allowing for a more organised system update schedule. By alternating between data batch load updates and exclusion classification functions for new customers, the system maintains its performance and adaptability, ensuring both accuracy and efficiency in its operations.

7.2.5 Standardise and transpose UKG into RDF and OWL structure

As mentioned previously in Chapter 2, many domain-specific knowledge graphs are not up to par with the standardised format due to its way of construction. Although the UKG is built using Python and networkx [19] for a smoother calculation function incorporation, the graph itself after construction can be transposed to OWL for standardised preservation using Algorithm 3 below:

Algorithm 3 Pseudo Code: Update OWL Graph with NetworkX Graph Data

Input: UKG in NetworkX form
Output: UKG in OWL standard

```
1. For each node in UKG:
    a) If not $owl$.search(iri="*".node."*"):
        i. $new \leftarrow$ create(node, (Thing))
        ii. $owl$.append($new$)
2. For each edge in $nx$.edges:
    a) Let source, target ← edge
    b) If not $owl$.search(iri="*".source."_to_".target."*"):
        i. $new \leftarrow$ create_object_property(source, target)
        ii. $source \leftarrow owl$.search(iri="*".source."*")[0]
        iii. $target \leftarrow owl$.search(iri="*".target."*")[0]
        iv. add_relation($source$, $new$, $target$)
3. Return null.
```

7.2.6 Insurance Fraud Prediction

As a benefit of the structure of the graph, finding insurance fraud claims clusters can be made easier by tracking fraud gangs via clusters of fraud cases linked to certain doctor practices, which has already provided as an attribute node in the knowledge graph. This is not within the scope of this research, however, the idea has been proposed by [70] in the sub-field of auto insurance.

7.3 Current Limitations and Future Headings

7.3.1 Current Limitations

Due to the time limitation for Masters' degree of this study, the potentials of the UKG in its various use cases have yet to be fully explored. These limitations include:

- **Limitation in fully automated process for new rules identification:** As stated in Chapter 5, this model still relies on a fixed benchmark (currently set at 0.8 for statistical reliance - proven by manual review of the results) in order for a link between an attribute and an exclusion to be established, though this problem has been partially resolved with the adjustment on UKG link weight calculation in Chapter 6 via the use of probability statistics and Bayesian theory [32]. This indicates that the model's current functionality is limited at rules identification and suggestions, hence more industrial research efforts into model evaluation and automated reasoning are required for the model to truly achieve the state of fully automated identification.

- **Rules evaluation:** The model at its current state relies on the companies' underwriters' expertise (in the form of utilising ML and data analytics for pattern recognition from historical decisions) without other sources of truth for verification. That being said, as the scope set out for this study did not aim to replace human underwriters, this limitation can potentially be resolved by either having the UKG's suggested rules identified through changes in the top 20 rankings should be regularly reviewed by underwriting experts for regular maintenance. The evaluation framework can be as simple as presented in Figure 7.1.

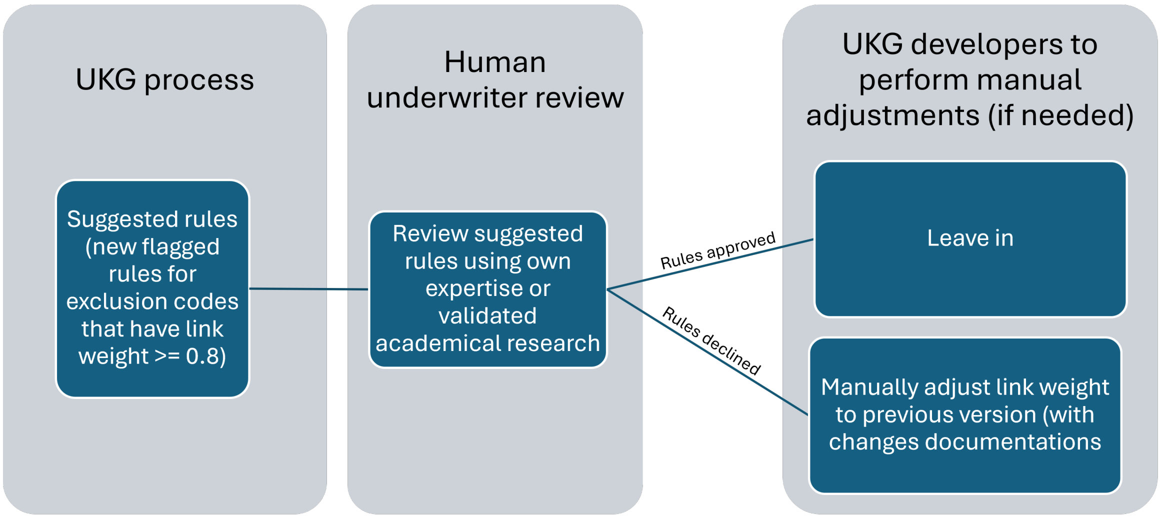

UKG process
Human underwriter review
UKG developers to perform manual adjustments (if needed)
Suggested rules (new flagged rules for exclusion codes that have link weight >= 0.8)
Review suggested rules using own expertise or validated academical research
Rules approved
Rules declined
Leave in
Manually adjust link weight to previous version (with changes documentations

Figure 7.1: Proposed Validation Framework

7.3.2 Use of Exclusion Code within Underwriting Knowledge Graph

Insurance data represents one of the most comprehensive and personalized datasets available on an individual, encapsulating myriad details that paint a vivid portrait of a person's life and potential risks. Given its depth and scope, the application and interpretation of exclusion codes within the UKG framework is pivotal for the nuanced classification and segmentation of this data. Particularly in the context of health exemptions, it is advisable to leverage these codes not merely as tools for exclusion but as instrumental guides to identifying potential health risks a customer may face. By doing so, insurance policies can be combined with health recommendations and products aimed at mitigating these risks at an early stage. This approach marks a paradigm shift towards utilising machine learning and data-driven methodologies not just for enhancing the insurance company's risk assessment and mitigation strategies but also for significantly improving the care provided to customers and further drive the personalisation effort. Through the repurposed use of exclusion codes, insurance entities can transcend traditional benefits, offering a more holistic and preventative care model as well as personalised policies that prioritises the well-being of the insured, thereby fostering a more nurturing and protective insurance ecosystem. This usage of UKG and explanable exclusions can be harnessed to deliver mutual benefits to both insurance companies and their customers, enhancing care while optimising risk management.

In events of a pandemic such as the case of COVID-19, the UKG can be utilised for a view of multiple exclusion codes to identify potential mental health issues with the customer risk profile for early forecast of customers who are likely to make a claim (those with high ranking for multiple combination exclusion codes). This can help the insurance industry to better their personalisation experience with early risk intervention efforts.

7.3.3 Future Headings

The current version of the UKG adopts a simpler method to maintain accuracy in order to effectively reducing the risk of "hallucination" or errors in data interpretation. This streamlined approach ensures a decently reliable output, leaving more room for improvement on improving link weight calculations. Enhancing these calculations is crucial for advancing the process of automated reasoning, particularly in the identification of rules within the knowledge graph. Additionally, since the UKG functions as a graph database, there is significant potential to establish and explore a wider range of use cases for this model. Investing efforts in these areas will enable us to unlock new applications and drive further advancements in data processing and knowledge representation, ultimately expanding the utility and impact of the UKG across various domains.

Chapter 8

Conclusion

In conclusion, this thesis explored innovative approaches to enhance customer risk profiling in life underwriting by leveraging an underwriting knowledge graph and multi-label classification for explainable exclusions. The study addressed significant challenges in traditional underwriting processes, such as handling missing values and ensuring model transparency. By constructing a UKG based on real-life data and introducing a semi-automated process for rules identification and maintenance, the research provided valuable insights into customer risk profiles and potential correlations with exclusion codes. The proposed methodology demonstrated the effectiveness of integrating a knowledge graph structure with multi-label classification to improve accuracy and transparency in underwriting decisions.

This study serves as one of the continuous efforts to set the steppingstone in the race to put technological models that have been created and researched within academia momentarily to real-life practice. Due to time limitations of my degree, my research barely scraps the surface of the full potential of knowledge graph application within this field, as the purpose is to explore potential approaches and to produce a functioning model within the time frame allowed. That being said, future research should focus on expanding use cases of the UKG, refining the algorithms for better performance and further the process of automated reasoning in knowledge graph, ultimately contributing to the broader application of data-driven approaches in the insurance industry and other personalised services. Though the application of the model and its results eventually relies on the insurance industry and companies themselves, the risk profile created from the UKG should be used to drive the personalised efforts to offer personalised packages instead of outright exclusions for customers, and additionally, thanks to the comprehensive view that insurance data covered on a person's life, including but not limited to their health, lifestyle and income, the risks interpreted from the model's exclusion code results can serve as an early-on warning for customers as well as healthcare personnel partners to provide and cover precaution treatments and first-aid practices for prevention of potential risks.

Bibliography

[1] B. Abu-Salih, *Domain-specific knowledge graphs: A survey*, Journal of Network and Computer Applications, 185 (2021), p. 103076.

[2] K. S. Aggour, P. P. Bonissone, W. E. Cheetham, and R. P. Messmer, *Automating the underwriting of insurance applications*, AI magazine, 27 (2006), pp. 36–36.

[3] T. L. Alborn, *Regulated lives: life insurance and British society, 1800-1914*, University of Toronto Press, 2009.

[4] N. Arora and S. K. Vij, *A hybrid neuro-fuzzy network for underwriting of life insurance.*, International Journal of Advanced Research in Computer Science, 3 (2012).

[5] R. Biddle, S. Liu, P. Tilocca, and G. Xu, *Automated underwriting in life insurance: Predictions and optimisation*, in Australasian Database Conference, Springer, 2018, pp. 135–146.

[6] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, *Knowedu: A system to construct knowledge graph for education*, Ieee Access, 6 (2018), pp. 31553–31563.

[7] Z. Chen, M. Chen, K. Weinberger, and W. Zhang, *Marginalized denoising for link prediction and multi-label learning*, in Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[8] M. . Company, *Future of enterprise data management: Data products and knowledge graphs*, 2021.

[9] Deloitte, *Knowledge graphs for financial services*, 2020.

[10] K. Dembczynski, W. Cheng, and E. Hüllermeier, *Bayes optimal multilabel classification via probabilistic classifier chains*, in ICML, 2010.

[11] S. A. Elnagdy, M. Qiu, and K. Gai, *Cyber incident classifications using ontology-based knowledge representation for cybersecurity insurance in financial industry*, in 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), IEEE, 2016, pp. 301–306.

[12] ——, *Understanding taxonomy of cyber risks for cybersecurity insurance of financial industry in cloud computing*, in 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), IEEE, 2016, pp. 295–300.

[13] P. Embrechts and M. V. Wüthrich, *Recent challenges in actuarial science*, Annual Review of Statistics and Its Application, 9 (2022), pp. 1–22.

[14] P. Ernst, C. Meng, A. Siu, and G. Weikum, *Knowlife: a knowledge graph for health and life sciences*, in 2014 IEEE 30th International Conference on Data Engineering, IEEE, 2014, pp. 1254–1257.

[15] J. M. Furman and S. P. Cass, *Benign paroxysmal positional vertigo*, New England Journal of Medicine, 341 (1999), pp. 1590–1596.

[16] D. Ganda and R. Buch, *A survey on multi label classification*, Recent Trends in Programming Languages, 5 (2018), pp. 19–23.

[17] T. Goodwin and S. M. Harabagiu, *Automatic generation of a qualified medical knowledge graph and its usage for retrieving patient cohorts from electronic medical records*, in 2013 IEEE Seventh International Conference on Semantic Computing, IEEE, 2013, pp. 363–370.

[18] C. Grévisse, R. Manrique, O. Mariño, and S. Rothkugel, *Knowledge graph-based teacher support for learning material authoring*, in Colombian Conference on Computing, Springer, 2018, pp. 177–191.

[19] A. Hagberg, P. Swart, and D. S Chult, *Exploring network structure, dynamics, and function using networkx*, tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[20] J. Harman, *Oesophagitis*, British Medical Journal, 1 (1952), p. 941.

[21] R. Hewett, C. Alexakis, A. Farmer, J. Ainley, V. Chhaya, J. Hayat, A. Poullis, and J.-Y. Kang, *Effects of eosinophilic oesophagitis on quality of life in an adult uk population: a case control study.*, Diseases of the Esophagus, 30 (2017).

[22] B. J. Hutagaol and T. Mauritsius, *Risk level prediction of life insurance applicant using machine learning*, International Journal of Advanced Trends in Computer Science and Engineering, 9 (2020).

[23] M. R. Islam, S. Liu, I. Razzak, M. A. Kabir, X. Wang, P. Tilocca, and G. Xu, *Mhivis: Visual analytics for exploring mental illness of policyholders in life insurance industry*, in 2020 7th international conference on behavioural and social Computing (BESC), IEEE, 2020, pp. 1–4.

[24] G. Ivchenko and S. Honov, *On the jaccard similarity test*, Journal of Mathematical Sciences, 88 (1998), pp. 789–794.

[25] Y. Jia, Y. Qi, H. Shang, R. Jiang, and A. Li, *A practical approach to constructing a knowledge graph for cybersecurity*, Engineering, 4 (2018), pp. 53–60.

[26] M. K. Joram, B. K. Harrison, and K. Joseph, *A knowledge-based system for life insurance underwriting*, International Journal of Information Technology and Computer Science, 3 (2017), pp. 40–49.

[27] V. Kašćelan, L. Kašćelan, and M. Novović Burić, *A nonparametric data mining approach for risk prediction in car insurance: a case study from the montenegrin market*, Economic research-Ekonomska istraživanja, 29 (2016), pp. 545–558.

[28] J.-S. Kim and D. S. Zee, *Benign paroxysmal positional vertigo*, New England Journal of Medicine, 370 (2014), pp. 1138–1147.

[29] I. Kose, M. Gokturk, and K. Kilic, *An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance*, Applied Soft Computing, 36 (2015), pp. 283–299.

[30] M. Kraus, S. Feuerriegel, and A. Oztekin, *Deep learning in business analytics and operations research: Models, applications and managerial implications*, European Journal of Operational Research, 281 (2020), pp. 628–641.

[31] W. D. Larson and T. M. Sinclair, *Nowcasting unemployment insurance claims in the time of covid-19*, International journal of forecasting, 38 (2022), pp. 635–647.

[32] P. M. Lee, *Bayesian statistics*, Oxford University Press London:, 1989.

[33] M. Lengwiler, *History of insurance in a global perspective: A novel research agenda*, 2023.

[34] Z. C. Lipton, *The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.*, Queue, 16 (2018), pp. 31–57.

[35] J. Liu, Z. Lu, and W. Du, *Combining enterprise knowledge graph and news sentiment analysis for stock price prediction*, (2019).

[36] H.-Y. Lo, S.-D. Lin, and H.-M. Wang, *Generalized k-labelsets ensemble for multi-label and cost-sensitive classification*, IEEE Transactions on Knowledge and Data Engineering, 26 (2013), pp. 1679–1691.

[37] S. M. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, Advances in neural information processing systems, 30 (2017).

[38] T. Lynn, J. G. Mooney, P. Rosati, and M. Cummins, *Disrupting finance: FinTech and strategy in the 21st century*, Springer Nature, 2019.

[39] X. Ma, S. Tan, X. Xie, X. Zhong, and J. Deng, *Joint multi-label learning and feature extraction for temporal link prediction*, Pattern Recognition, 121 (2022), p. 108216.

[40] K. Maehashi and M. Shintani, *Macroeconomic forecasting using factor models and machine learning: an application to japan*, Journal of the Japanese and International Economies, 58 (2020), p. 101104.

[41] V. Martínez, F. Berzal, and J.-C. Cubero, *A survey of link prediction in complex networks*, ACM computing surveys (CSUR), 49 (2016), pp. 1–33.

[42] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, *Machine learning for financial risk management: A survey*, IEEE Access, 8 (2020), pp. 203203–203223.

[43] D. McGlade and S. Scott-Hayward, *Ml-based cyber incident detection for electronic medical record (emr) systems*, Smart Health, 12 (2019), pp. 3–23.

[44] Y. Mita, R. Inose, R. Goto, Y. Kusama, R. Koizumi, D. Yamasaki, M. Ishikane, M. Tanabe, N. Ohmagari, and Y. Muraki, *An alternative index for evaluating amu and anti-methicillin-resistant staphylococcus aureus agent use: A study based on the national database of health insurance claims and*

specific health checkups data of japan, Journal of Infection and Chemotherapy, 27 (2021), pp. 972–976.

[45] E. Montañes, R. Senge, J. Barranquero, J. R. Quevedo, J. J. del Coz, and E. Hüllermeier, *Dependent binary relevance models for multi-label classification*, Pattern Recognition, 47 (2014), pp. 1494–1508.

[46] G. Montavon, W. Samek, and K.-R. Müller, *Methods for interpreting and understanding deep neural networks*, Digital signal processing, 73 (2018), pp. 1–15.

[47] J. Mourmouris and T. Poufinas, *Multi-criteria decision-making methods applied in health-insurance underwriting*, Health Systems, (2022), pp. 1–33.

[48] I. O. Pappas and A. G. Woodside, *Fuzzy-set qualitative comparative analysis (fsqca): Guidelines for research practice in information systems and marketing*, International Journal of Information Management, 58 (2021), p. 102310.

[49] Y. Qi, R. Jiang, Y. Jia, R. Li, and A. Li, *Association analysis algorithm based on knowledge graph for space-ground integrated network*, in 2018 IEEE 18th International Conference on Communication Technology (ICCT), IEEE, 2018, pp. 222–226.

[50] S. Rawat, A. Rawat, D. Kumar, and A. S. Sabitha, *Application of machine learning and data visualization techniques for decision support in the insurance sector*, International Journal of Information Management Data Insights, 1 (2021), p. 100012.

[51] J. Read, B. Pfahringer, G. Holmes, and E. Frank, *Classifier chains for multi-label classification*, Machine learning, 85 (2011), pp. 333–359.

[52] G. Rizos, S. Papadopoulos, and Y. Kompatsiaris, *Multilabel user classification using the community structure of online networks*, PloS one, 12 (2017), p. e0173347.

[53] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, *Learning a health knowledge graph from electronic medical records*, Scientific reports, 7 (2017), pp. 1–11.

[54] S. Sachan, J.-B. Yang, D.-L. Xu, D. E. Benavides, and Y. Li, *An explainable ai decision-support-system to automate loan underwriting*, Expert Systems with Applications, 144 (2020), p. 113100.

[55] L. Shi, S. Li, X. Yang, J. Qi, G. Pan, and B. Zhou, *Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services*, BioMed research international, 2017 (2017).

[56] A. Swartz, *Musicbrainz: A semantic web service*, IEEE Intelligent Systems, 17 (2002), pp. 76–77.

[57] P. Szymański and T. Kajdanowicz, *A scikit-based Python environment for performing multi-label classification*, ArXiv e-prints, (2017).

[58] J. Tack, A. Becher, C. Mulligan, and D. Johnson, *Systematic review: the burden of disruptive gastro-oesophageal reflux disease on health-related quality of life*, Alimentary pharmacology & therapeutics, 35 (2012), pp. 1257–1266.

[59] T. Taft, E. Kern, M. Kwiatek, I. Hirano, N. Gonsalves, and L. Keefer, *The adult eosinophilic oesophagitis quality of life questionnaire: a new measure of health-related quality of life*, Alimentary pharmacology & therapeutics, 34 (2011), pp. 790–798.

[60] E. A. Tanaka, S. R. Nozawa, A. A. Macedo, and J. A. Baranauskas, *A multi-label approach using binary relevance and decision trees applied to functional genomics*, Journal of biomedical informatics, 54 (2015), pp. 85–95.

[61] E. Tjoa and C. Guan, *A survey on explainable artificial intelligence (xai): Toward medical xai*, IEEE transactions on neural networks and learning systems, 32 (2020), pp. 4793–4813.

[62] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, et al., *Bayesian statistics and modelling*, Nature Reviews Methods Primers, 1 (2021), p. 1.

[63] K. Van Nguyen, M. R. Islam, H. Huo, P. Tilocca, and G. Xu, *Explainable exclusion in the life insurance using multi-label classifier*, in 2023 International Joint Conference on Neural Networks (IJCNN), IEEE, 2023, pp. 1–8.

[64] N. N. Vo, S. Liu, X. Li, and G. Xu, *Leveraging unstructured call log data for customer churn prediction*, Knowledge-Based Systems, 212 (2021), p. 106586.

[65] M. Wang, L. Qiu, and X. Wang, *A survey on knowledge graph embeddings for link prediction*, Symmetry, 13 (2021), p. 485.

[66] Q. Wang, Z. Mao, B. Wang, and L. Guo, *Knowledge graph embedding: A survey of approaches and applications*, IEEE Transactions on Knowledge and Data Engineering, 29 (2017), pp. 2724–2743.

[67] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., *Drugbank 5.0: a major update to the drugbank database for 2018*, Nucleic acids research, 46 (2018), pp. D1074–D1082.

[68] T. Wu, C. Gao, G. Qi, L. Zhang, C. Dong, H. Liu, and D. Zhang, *Kg-buddhism: The chinese knowledge graph on buddhism*, in Joint International Semantic Technology Conference, Springer, 2017, pp. 259–267.

[69] T. Wu, G. Qi, C. Li, and M. Wang, *A survey of techniques for constructing chinese knowledge graphs and their applications*, Sustainability, 10 (2018), p. 3245.

[70] L. Zhang, T. Wu, X. Chen, B. Lu, C. Na, and G. Qi, *Auto insurance knowledge graph construction and its application to fraud detection*, in The 10th International Joint Conference on Knowledge Graphs, 2021, pp. 64–70.

[71] Y. Zhao, L. Li, and X. Wu, *Link prediction-based multi-label classification on networked data*, in 2016 IEEE First International Conference on Data Science in Cyberspace (DSC), IEEE, 2016, pp. 61–68.

[72] X. Zou, *A survey on application of knowledge graph*, in Journal of Physics: Conference Series, vol. 1487, IOP Publishing, 2020, p. 012016.