The AI underwriter - an active learning approach towards more transparent, fair predictive underwriting for life insurance.

Paul Trayers

8907021

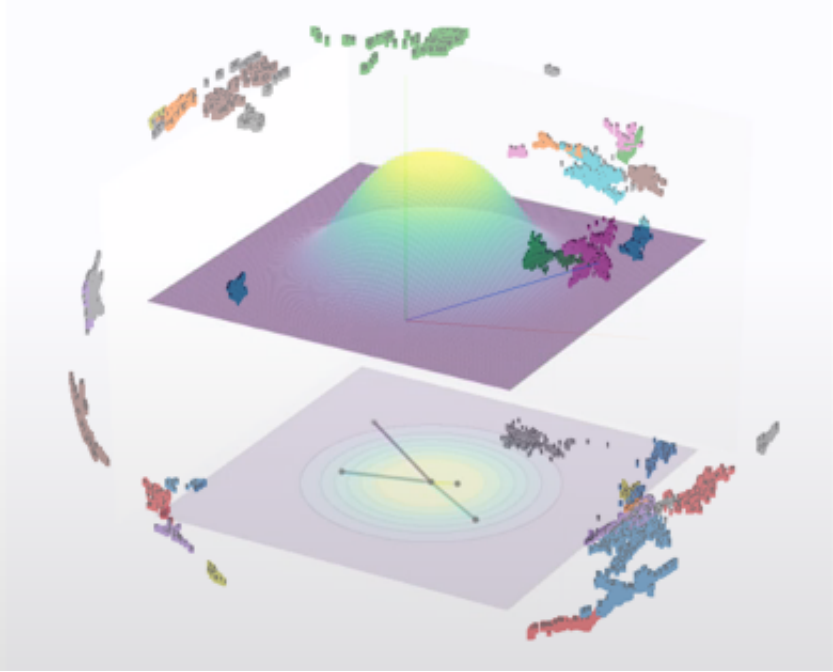
Department of Computer Science and Information Systems

Faculty of Science and Engineering

University of Limerick

Submitted to the University of Limerick for the degree of

Magister Scientiæ(MSc) 2023/2024



[1]Prudential Insurance dataset cluster visualisation blended with image of an isograph ren-

1. Supervisor: Dr. Martin Cunneen
Dept. of Accounting & Finance
Kemmy Business School
University *of* Limerick
Ireland

dered by ChatGPT.

Abstract

(Background)

(Objective) In this thesis, we explore the application of unsupervised learning methods to identify meaningful clusters within historical insurnance data samples. Our primary goal is to develop a framework that enables underwriters as domain experts to assign risk class designations to these automatically identified clusters, thereby enhancing human oversight, interpretability and utility of the clustering results for the prediction of future life application risk class. We focus on HDBSCAN and UMAP clustering algorithms, to uncover inherent patterns and groupings in the data.

(Methods)

(Results) The effectiveness of these methods is evaluated using the Prudential Insurnance dataset, demonstrating their potential to reveal hidden structures that are not immediately apparent through traditional analysis techniques and potential to classify risk based on these structures.

(Conclusion) By incorporating expert underwriter input, we bridge the gap between black-box predictive underwriting techniques coupled with limited explainability techniques and practical, real-world classification needs, providing a robust tool for data-driven decision-making processes for risk-sensitive applications.

The results indicate XXX ..that our approach not only can improve the accuracy of classification (TBC) ..and offers a classification method where predictions can be explained in a mathematically rigourous way

but also facilitates human oversight and a more intuitive understanding of complex data landscapes.

(Contribution) This has significance for high-risk and risk-sensitive fields like insurance where leveraging machine learning techniques has been because of lack of sufficient causal explainability.

...

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other Irish or foreign examination board.

The thesis work was conducted from 2023 to 2024 under the supervision of Dr. Martin Cunneen at University of Limerick.

Limerick, 2024

Acknowledgements

dedication

Contents

List of Tables

LIST OF TABLES

List of Figures

LIST OF FIGURES

1

Introduction

[Guidance text commented here]

1.1 Modern Insurance Underwriting

First, we introduce to the field of life insurance underwriting and comment on the state of automation in underwriting decision-making today.

Today, life insurance underwriting best practice typically blends automated rule-based engines with referrals to traditional underwriting for more complex applications (Wang, 2021). The observation of the author having seven years of experience working within the Reinsurance sector is that the job of maintaining and managing rules governing automation becomes more difficult as the number and complexity of product offerings increase. As shown in this research, in recent years research has emerged demonstrating the potential of machine learning techniques to instead directly predict underwriting decisions, with a number of these models being successfully deployed commercially by major insurance carriers such as MassMutual (Maier et al., 2019). While the level of precision that such approaches enable at individual level provide opportunities to drive efficiencies and profitability we examine here the implications this introduces as as potential threat to social solidarity one of the key original purposes of insurance. This research examines the potential of approaches that can address this issue

still availing of the benefits of modern techniques AI automated or underwriter-augmented decision making that keeping human-in-the loop oversight in critical decision making also prioritising interpretability of models as a means towards the transparency that forms a prerequisite to a more ethical and fair automated underwriting decision making.

1.2 Thesis Structure

The remaining chapters of this dissertation are as follows:

Chapter Two provides a comprehensive review of the literature concerning the intersection of insurance and its social functions, particularly mutualisation. It explores the methodologies by which underwriting assesses and classifies risk and investigates the implications of increasing automation in underwriting processes. The review examines contemporary research on predictive techniques using supervised machine learning, examining their impact on mutualisation. It also examines the notions of actuarial fairness, bias, and unfair discrimination. Furthermore, the chapter surveys the governance and regulatory frameworks in the United States and the European Union with respect to consumer protections.

Chapter Three introduces the data set with an Exploratory Data Analysis (EDA).

Chapter Four describes the methodology for the design and execution of the experiments using supervised and unsupervised machine learning techniques.

Chapter Five presents the results from experiments.

Chapter Six draws conclusions and evaluates the results from experiments. Finally, it suggests possible future works.

A series of documents have been included in the Appendix section of this dissertation. These are:

- *Appendix A* outlines . . .
- *Appendix B* presents . . .
- *Appendix C* includes . . .

Available with this dissertation are two Jupyter Notebook files containing the following items:

- *Python Notebook*: contains an Exploratory Data Analysis (EDA) of the Prudential data set and application of supervised machine learning techniques on this dataset to predict risk classes and unsupervised clustering approaches to the dataset.

2

Literature Review

2.1 Introduction

Background

The purpose of this review of the literature is to introduce the central topics related to the research, including the key literature related to these topics. The review begins with a background on the insurance industry and its ongoing transformation through automation. Next, the focus is more specifically on machine learning and its adoption in insurance underwriting, and after surveying the regulatory environment with regards to AI especially in the area of decisioning, moving into literature around risks to the key insurance concept of "mutualization" from modern methods and approaches.

The context is to ensure social fairness for consumers through transparency within the context of the technological advances that machine learning offers in the field of life insurance underwriting. The research question is focused on whether modern unsupervised learning techniques can be leveraged to offer models that encourage fairness through improved levels of transparency while maintaining performance characteristics comparable with "back-box" supervised methods. As such, a critical analysis of the literature here aims to identify gaps that can justify the need for such research.

Purpose of the review

Evaluate the impacts of the wider availability of personal data and the acceleration of machine learning in the field of automated underwriting. Secondly, this review aims to evaluate the impact on fairness and the extent to which regulation and governance have mitigated any threats.

Scope of the review

The scope of the review will be insurance, focusing on life insurance in particular and spanning the insurance industry evolution from traditional manual underwriting, subsequent advancement through the adoption of technologies such as rules-based systems for increased automation and efficiency towards modern efforts to adopt machine learning methods to drive further automation of underwriting decisions. The scope of the literature will be global, but with a focus on the United States, since the largest market for life insurance is based in the United States, where there is also greater availability of personal information that may be included in underwriting decision making.

Review Methodology

The following methods were used to select and analyse the literature. First, selecting key themes for each section, various sources including Google Scholar, UL Library, scholar.google.com, semanticscholar.org, connectedpapers.com, elicit.com, litmaps.com, scite.ai were searched to identify key relevant and impactful papers both historically and from the last five years; these were downloaded and collated with Zotero before reviewing adding relevant references to the narrative discussion and critically analyses in the relevant sections. The approach is to critically analyse the literature with a view to identifying and analysing gaps that are consistent with the research question.

2.2 Origins and Ethos

Here we briefly trace the origins of insurance mutualisation and how this is preserved from a governance, regulatory, and compliance standpoint.

In 18th and 19th century Britain, insurance was viewed as a form of gambling, leading to regulatory measures like the Gambling Act of 1774 and the Annuity Act of 1777 (Tapan Biswas, 1997). Different types of insurance, such as life and fire insurance, developed different market characteristics. Life insurance, being a long-term contract and form of saving, required companies to establish reputation and trust. In contrast, fire insurance, typically short-term, demanded less trust (Tapan Biswas, 1997). The life insurance market primarily targeted the wealthy and middle class as they could afford to save and plan for their families' futures.

In fact, the development of life insurance in this period was marked by moral controversies and cultural changes. Initially condemned as sacrilegious in the US, life insurance gradually gained acceptance as a means of financial protection (V. Zelizer, 1979). In England, early life insurance societies offered mutual economic protection and promoted reforming ideals, but also faced challenges in distinguishing between prudential insurance and gambling on lives (Geoffrey Clark, 1997). The industry grappled with the assessment of both physical and moral risk, particularly with regard to certain groups such as Jews and Irish (R. Pearson, 2002).

The rise of insurance coincided with the dismantling of feudal solidarity and the emergence of individualism, reflecting a capitalist ethos (F. Ewald, 2019). As the industry evolved, it developed new moral technologies to segregate legitimate from illicit motives to protect and restrict members' proprietary rights over policies (Geoffrey Clark, 1997). This transformation shaped the culture of life insurance in England from 1695 to 1775 (T. Alborn, 2000).

We see that the insurance market underwent significant changes during the 18th and 19th centuries, reforming its image as gambling on lives to being considered as a means of mutual financial protection. In addition to the moral hazards and ethnically based discrimination pointed above and its origins according to Halpérin is that its origins are less "founded on a sense of solidarity" and "common security" than in the spirit of financial gain (F. Ewald, 2019).

2.3 Underwriting Automation (first generation - 1980 to 2010)

This section traces the origins of underwriting automation in insurance from the "first-generation" of expert decision support systems before more recent adoption of machine learning models.

With the advent of expert systems in the 1980s attempts were underway to automate and enhance decision-making processes in insurance underwriting. Early research focused on developing prototypes and examining the feasibility of applying expert systems to determine applicant insurability based on mainframe, rule-based systems and utilising LISP and PROLOG at an unamed Midwestern insurer (Gary A. Wicklund and R. Roth, 1987). Subsequent studies went beyond rule-based systems to explore various nascent artificial intelligence techniques, including fuzzy logic, evolutionary algorithms, and even neural network techniques. Risk classification and claim cost prediction was implemented using k-means clustering and heuristic methods to group policy holders (A. C. Yeo et al., 2003); (K. Aggour et al., 2005); (K. Aggour et al., 2005).

These systems showed promise in automating underwriting tasks, with one implementation at Genworth achieving a significant automation rate close to 20% on LTC product applications at Genworth (K. Aggour et al., 2005). Researchers were also investigating the workflow side with the integration of web services and alerts to enhance workflow automation and exception handling in underwriting processes (Raymond C. M. Lee et al., 2007); indeed, having actively participated in such programmes during the past seven years, the author can speak to such workflow automation and optimisation efforts ongoing in major insurers to this day. Later, neural network models also emerged tentatively being proposed as support tools for determining premium rates in property and casualty insurance to address limitations of traditional interpolation methods (Chaohsin Lin, 2009).

Although relatively sparse during this period, the research reflects a focus on researching rule-based and fuzzy systems for underwriting decision-making. However, based on our observation of the state of automation in large insurers globally in the past decade, the reality is more likely that industry adoption of such emerging technologies would have significantly lagged transformations

in financial domains such as banking which underwent dramatic transformation during this era. More likely, this period was still very much marked by traditional manual underwriting methods, with automation being introduced more around process automation to support the manual underwriter, digitisation of documentation, etc. with relatively little automated decisioning.

2.4 Evidence-based risk assessment

In this section, we trace the evolution and emphasise the importance of medical and other evidential data in modern insurance underwriting.

Underwriting has evolved from data collection at a population level or ad hoc at an individual level to a much more systematic gathering of the critical medical and other related data that underpins individual life underwriting decisions. Milano (2001) proposed a comprehensive, rule-based risk assessment framework that moved toward implementing a much more systematic, personalised, evidence-based risk assessment than traditional empirical methods and reliance on generic population data, thus improving the quality and consistency of risk selection. This represented an evolution of insurance underwriting towards a more evidence-based, systematic, and competitive approach compared to traditional empirical methods.

Milano (2001) noted at that time the limited availability of good quality evidential data and the need for training and interpretation of such data. Of course availability of data has become much less of an issue with increased access to digitized personal health records as provided by third-party data vendors to insurers especially in US including the more recent availability of personal Electronic Health Records (EHR) and uptake by insurer wanting to leverage these as part of risk assessment.

Klein (2013) details the multiple data sources that supplement the typical application risk assessment in US begineeing with the tele-interview questionaire now generally evolved to an online format. Additional evidences range from reports and data on fluids (blood and urine and oral) to MIB check for any existing conflicting applications (Medical Information Bureau), doctor APS reports (Attending Physician Statement), MVR (Motor Vehicle Record) reports and data

from state DMVs, Pharamceutical prescription records (Rx) in addition to other medical reports such as EKG, Chest x-rays etc.

Some of these evidences may be requested by default on application initiation, and others may be requested depending on the disclosures made by the application during the interview. The insurer decides the circumstances in which evidence data are requested. These may be ordered automatically or manually depending on the specific characteristics of an application and the overall level of automation implemented in the underwriting workflow.

2.5 Risk Classification

Here, we outline the reasons for the critical role of classification in risk assessment and the central role played by the underwriter in evaluating and classifying risks to preserve the health of the insurer portfolio.

Underwriters analyse information on insurance applications to determine whether a risk is acceptable and will not result in an early claim using underwriting guidelines and typically working with medical doctors and other specialists to assess the risks. Risks are divided into different classes (standard, substandard, preferred, etc.) based on the likelihood of a claim that allows the insurer to charge appropriate premium rates (Macedo, 2009).

The underwriter must evaluate and select risks to maintain a homogeneous portfolio of risks for the insurer. As Macedo (2009) points out, this is based fundamentally on the Central Limit Theorem principle that a large enough number of similar risks will exhibit a normal (Gaussian) distribution. Therefore, risks should have a high degree of correlation to behave in a predictable manner. However, different insurance companies may accept different risk profiles across their product lines. In these cases, the underwriter may accept the risks under different conditions (terms), such as charging an extra premium (loading), applying exclusions, or imposing waiting periods.

However, the goal is to maintain the necessary homogeneity of risks required by the insurer in the overall portfolio. As Macedo (2009) identifies, this is crucial because insurance companies can only create value by reducing the volatility of claims if their portfolio consists of homogeneous risks. arguing that this involves

understanding not just the biophysical risk characteristics of the insured but the "moral risk" including applicant's reputation, financial position, etc. The latter point begins to move away from what are considered traditional data sources mentioned in the section on *Evidence-based risk assessment* and toward more newer data sources, which are more controversial including criminal and credit records in addition to those shared from the world of personal electronic devices either with express consent (personal fitness trackers) or not (social media networks).

2.6 Predictive Underwriting

Leaving behind traditional rule-based expert systems, this section explores the emergence of machine learning models leveraging the dramatic increases in availability of personal data, compute and advances in AI and machine learning techniques. Ultimately, we would like to identify what formats and contexts have these been most successfully deployed for life insurance underwriting decision making.

2.6.1 Earlier Developments

Predictive underwriting is not only a phenomenon that arises in the context of recent advances in AI. In the early 1990s Nikolopoulos and Duvendack (1994) described the application of machine learning to life insurance decision making, first comparing and then combining evolutionary learning with classification tree techniques to build a knowledge base of rules for an expert system to determine the expiration of life insurance policies.

In the early 2000s, data mining and knowledge discovery techniques were emerging to produce tools for decision support and risk assessment reminiscent of the modern machine learning of today (Apté et al., 2002). These were aimed at using large volumes of high-dimensional data to develop automated, scalable analytics that could supplement or replace traditional human-expert approaches. Predictive models were used to set competitive premiums while managing risk, avoiding overcharging low-risk policyholders and undercharging high-risk ones (Apté et al., 2002).

2.6.2 Modern Predictive Underwriting

This section traces the rise of predictive models in underwriting decision automation during more recent times arising from dramatically increased access to personal data and advances in AI and machine learning techniques.

Recent research has explored the application of machine learning techniques to automate and enhance risk assessment in life insurance underwriting. Here, we outline the aims, techniques, and results from key studies relating to application of machine learning to underwriting decision support.

Boodhun and Jayabalan (2018) aimed to enhance risk assessment comparing multiple algorithms, including Multiple Linear Regression, Artificial Neural Network, REPTree, and Random Tree classifiers, to predict applicant risk levels. Using the publicly Prudential Dataset, they found that REPTree performed best with Correlation-Based Feature Selection producing a lowest mean absolute error (MAE) of 1.5285 and root-mean-squared error (RMSE) of 2.027, while Multiple Linear Regression excelled when combined with Principal Components Analysis for dimensionality reduction with the lowest MAE of 1.6396 and RMSE of 2.0659.

Biddle et al. (2018) similarly implemented and evaluated classical techniques this time to predict the application of exclusions using a dataset provided by a leading Australian life insurer covering an 8-year period from 2009 and 60,000 thousand individual applications. The researchers noted challenges with the data exhibiting sparsity in the questionnaire responses due to the conditional-branching structure, as well as extreme class imbalance in the application of exclusions, with over 1,000 different exclusions present. Implementing and evaluating techiques including Logistic Regression, XGBoost, and Recursive Feature Elimination. The study found that both Logistic Regression with L1 regularization and XGBoost performed well in predicting exclusions, with XGBoost using significantly fewer features to achieve similar accuracy, suggesting it as the better model for the task.

Levantesi and Pizzorusso (2019) aimed to investigate the ability of machine learning techniques to improve the accuracy of some standard stochastic mortality models in the estimation and forecasting of mortality rates. This study used tree-based machine learning techniques (decision tree, random forest, and gradient

boosting) to calibrate the machine learning estimator parameter that was then applied to standard mortality models. The study showed that the implementation of these machine learning techniques, based on features such as age, sex, calendar year, and birth cohort, leads to a better fit of the historical data compared to the original standard mortality models. The key novelty of the paper was that machine learning was used as a complement to standard stochastic mortality models, rather than as a substitute, in order to both improve model fit and forecasting, while also trying to create a bridge between the data-driven machine learning approach and the theoretical mortality modelling.

(Maier et al., 2019) aimed to use a large historical dataset of life insurance applicant data at MassMutual to develop a mortality prediction model using machine learning techniques, designing a novel evaluation framework, this forming the core of an algorithmic underwriting system at MassMutual. The key techniques used were Survival modelling using the Cox proportional hazards model and random survival forests to predict mortality risk. The results showed that the random survival forest model outperformed traditional underwriting, yielding a 6% reduction in claims in the healthiest pool of applicants. This algorithmic underwriting system reduced the time to issue policies by 25% and increased customer acceptance by more than 30% for offers made with a light manual review, saving millions of dollars in operational efficiency while driving the decisions behind tens of billions of dollars of life insurance benefits.

The group at MassMutual Maier et al. (2020) followed a year later with an improved version of the original model the aim of developing a "high-resolution mortality and life score" to serve as a primary driver of an algorithmic underwriting systems for life insurance, that would embrace transparency in terms of methodology sufficient to build trust with consumers and regulators. The research used a comprehensive dataset of 1.5 million MassMutual records over 20 years. Random survival forest model to directly estimate the cumulative hazard function and derive a standardised life score. Similarly to the original study Maier et al. (2019), a random survival forest model was used to directly estimate the cumulative hazard function and derive a standardised life score. Although the results were comparable with the original study, this study had more focus on transparency by implementing a SHAP framework to generate additive feature

contributions to explain individual life score predictions. MassMutual also developed a consumer-facing transparency tool called MyLifeScore to demonstrate how various factors drive individual risk.

Hutagaol and Mauritsius (2020) also aiming to examine how machine learning can help life companies determine the risk level of prospective applicants, implemented classical machine-learning techniques of Support Vector Machine (SVM) using different linear and non-linear kernels), Random Forest and Naive Bayes, implementing on the Prudential dataset. Results found Random Forest achieving the highest precision (0.85) over SVM (0.72) and Naive Bayes (0.49).

Wang (2021) constructed predictive machine learning models to predict underwriting decisions for life and health insurance applications using reinsurer real-world dataset of 29k records covering a 3-year period from 2017. Uniquely, the solution included machine learning techniques such as natural language processing and clustering analysis to process data including free-text descriptions of impairments and occupations.

Models were trained comparing the performance of various machine learning algorithms including Random Forests, Decision Tree, Gradient Boosting, Extreme Gradient Boosting, Bagging, AdaBoost, Support Vector Machine, Stochastic Gradient Descent, K Nearest Neighbors, and Ordinal Logistic Regression. The best performing algorithm was Extreme Gradient Boosting (XGB), achieving 94% accuracy on the training set and 71% accuracy on the test set. It was noted that this was a significant improvement over rules-based engines that can only process about a third of applications. In terms of explainability, feature importance ranking from XGB provided underwriting insights, such as BMI being a key predictor. Overall, the study concluded the potential of predictive modelling to handle complex cases over rules-based engines.

Sahai et al. (2022) also focused on Machine Learning (ML) techniques in underwriting decision making have saved time and improved operational efficiencies user-friendly cause-and-effect explanation of model's predictions. The research compared performances between tree-based classifiers including Decision Tree, Random Forest and XGBoost. The XGBoost classifier performed best with an AUC value (0.86) and F1-score (above (0.56) on the validation data, followed by Random Forest with AUC value (0.84) and f1 score (0.53). The research also

focused on interpretability of these techniques applying SHAP to the more complex (bck-box) models XGBoost and neural networks and 'Feature Importance' to models including Logistic Regression and tree-based models such as Decision Tree and Random Forest. (Dataset details not available at time of writing)

Recently, Varadarajan and Kakumanu (2024) surveyed the work of researchers attempting to determine the optimum machine learning model to enhance the accuracy of predicted policy issue decisions and to determine the strategies used to arrive at individual risk predictions. The findings were that the XGBoost model was the most effective and has the advantage of immunity to missing values according to the research. A notable omission is Maiers work on the MassMutual dataset, which speaks to a predictive model implemented in an underwriting system at a major US insurer. Varadarajan and Kakumanu (2024) proposes an increased focus on class-balancing, cross-validation and hyperparameter optimisation on the customer dataset, as not having received much attention in prior literature. However, since these are normally taken as standard it seems more likely that these activities have been included to some extent but perhaps have not explicitly mentioned.

2.6.2.1 Analysis

First, we will summarise some of the most pertinent aspects of the studies outlined in the previous section.

The increasing availability of data and advances in machine learning now means that underwriting automation can significantly speed up the processing of applications Boodhun and Jayabalan (2018). A common theme in the research is the use of supervised machine learning techniques to create predictive models from a publicly available dataset such as that published by Prudential (2016) and used by Boodhun and Jayabalan (2018) and Hutagaol and Mauritsius (2020) or proprietary historical underwriting datasets such as those used by MassMutual (Maier et al., 2019). Data sets include biophysical, occupational, insurance history and medical characteristics; however, Maier et al. (2019) refers to the increasing availability of nontraditional sources such as financial, public records and even wearables but does not include any of these in the data set. In

addition, traditional sources such as drug prescriptions (Rx) and motor vehicle records (MVR), which Maier et al. (2019) acknowledge would improve accuracy, are also omitted from this study. On the other hand, Boodhun and Jayabalan (2018) state that the Prudential dataset has almost sixty thousand applications with 128 attributes, however, it is unclear whether all of the expected traditional sources are included since values are normalised and field names anonymised e.g. `Medical_History_41`. We can conclude that all models examined may be incomplete to a significant degree in terms of missing some of the fundamental data sources typically utilised as part of the life underwriting process by insurance carriers in the North America region in the author's experience.

Commentary

A common theme in the literature is the research, comparison, and in some cases ultimately the adoption of machine learning techniques as a core element of the underwriting decision. The studies can be categorised into distinct groups based on what they predict;

- Predict the outcomes of the risk classification decisions in the underwriting application.
- Predict outcomes of underwriting application decisions other than risk classification, such as risk such as loadings, exclusions, and other terms.
- Predict mortality to provide a life risk score.
- Predict other aspects of the policy lifecycle, such as policy termination.

In the experience of the author with respect to the third item above, such risk scoring models can relate to, for example, smoking risk or overall mortality that typically serve as input to the underwriting evaluation (automated or manual). While such scores are useful, the authors observation is that insurer evaluations typically include risk score models when available as an input factor as part of a holistic view of evidences evaluated from multiple sources including biophysical, medical, and avocational to make a comprehensive assessment of an

applicant's risk. The implications of a trend towards risk-based scoring in terms of transparency and fairness are discussed separately later.

Sahai et al. (2022) notes the desire for explainability as a "user-friendly cause-and-effect explanation of model's predictions" being useful to stakeholders, financial institutions and regulators.

Ultimately, the aim, regardless of technique, is by such adoption to save time and improved operational efficiencies in the underwriting process, as achieved at MassMutual. However, Maier et al. (2020) also speaks to the "technical and business challenges" to get to full implementation in a production environment.

2.6.3 Industry adoption

Today, best practice in life insurance underwriting typically blends automated rule-based engines with referrals to traditional underwriting for more complex applications (Wang, 2021). The observation of the author having seven years of experience working within the reinsurance sector is that the job of maintaining and managing rules governing automation becomes more onerous for carriers as the number and complexity of product offerings increases. In recent years, research has emerged showing the potential of machine learning techniques to directly predict underwriting decisions, with a number of these models successfully being commercially deployed by major insurance carriers such as MassMutual (Maier et al., 2019). However, the authors experience is that the field continues to be dominated more by rules and model hybrids with a primary rules mechanism governing underwriting decisions while incorporating model results as an element of that decision making.

2.6.4 Risks to Mutualisation

The insurance industry relies from a commercial perspective on customer segmentation while simultaneously fulfilling the social function of mutualisation. As Charpentier et al. (2015) details, traditionally insurance relies on mutualising risks among policyholders, creating the need for homogeneous risk pools. Segmentation helps insurers achieve homogeneity by using proxy variables such as age and location to classify risk levels. The challenge is to balance segmentation

and mutualisation in a competitive market as more fine-grained segmentation risks introducing increasing variability and potential risk of market exit to the insurer.

Algorithmic prediction and big data are potentially "disruptive" to the traditional insurance model, promising to allow highly personalised insurance policies and premiums based on individual behaviour and risk profiles, rather than pooled risk (Cevolini and Esposito, 2020). Charpentier et al. (2015) concurs pointing out that this increased capacity for personalised data analysis enabled by modern machine learning and AI threatens to disrupt the traditional insurance model as insurers embrace the ability to predict individual risk more accurately and price accordingly. Cevolini and Esposito (2022) indicates use of telematics and behavioural data in motor insurance as an example of the industry moving from an actuarial valuation model to a more individualised behavioural model that allows insurance companies to better predict individual risk exposure, rather than relying on proxy variables such as gender, age, etc. This enables personalised pricing, potentially reducing cross-subsidisation between different risk groups.

The demise of risk pooling with increased focus on individualisation, which according to Cevolini and Esposito (2020) risks an end to mutualisation, might ironically be accelerated by consumer demand for personalised pricing. This risks introducing a new form of discrimination and exclusion as insurers refuse to cover higher-risk individuals, which conflicts with the social solidarity function Cevolini and Esposito (2020).

Actuaries should have a responsibility to balance technical with social considerations in risk pricing. Cevolini and Esposito (2020) argues that the shift from pooled risk to individualised prediction in insurance could fundamentally transform the social function and meaning of insurance, with wide-ranging consequences still largely unexplored. The advent of InsurTech and the availability of detailed personalised behavioural data creates asymmetry, as the insurer knows more about the policyholder than they might know about themselves raising privacy concerns despite the benefits. (Cevolini and Esposito, 2020).

Cevolini and Esposito (2022) speaks about the evolution in such an environment of "behavioral tribes" where risks are no longer pooled but rather individualized raising concerns about fairness and discrimination. Cevolini and Esposito

(2022) references the ongoing debate around whether this is more or less fair than traditional actuarial approaches. The use of behavioral data also provides opportunities for a changing more proactive role for insurance companies with policyholders perhaps a "coaching" role in trying to improve policyholder behavior and reduce risk, rather than just reacting to claims, potentially transforming the traditional business model and relationship between insurers and policyholders Cevolini and Esposito (2022)

2.6.5 Regulation and Compliance

This section describe key legal regulation relating to AI, specifically around predictive models and algorithmic decision making focused on EU and US as it relates to insurance and underwriting. A primary focus is to determine how current machine learning predictive models align with the transparency requirements set by various legislations.

2.6.5.1 United States

US Federal

At the federal level, the Health Insurance Portability and Accountability Act (HIPAA) Rights (OCR) intersects with regulations around algorithmic and automated decision making, particularly in the context of protecting patient personal health information (PHI) and ensuring that fair practices in healthcare algorithms and automated systems must comply with stringent standards. HIPAA through granting patients rights over access to their health information implies that there must be transparency in how patient data is used in algorithmic decision-making. HIPAA further intersects with efforts to regulate algorithmic fairness as system must not discriminate based on protected health characteristics (Rights , OCR).

In the authors experience, HIPAA compliance is already well embedded as part of the life insurance application and underwriting lifecycle with U.S. insurers during which the applicant is typically requested to consent to the terms of HIPAA to release access to applicant medical data.

The American AI Initiative Act (AII) Sen. Cantwell (2024) provides a comprehensive approach to regulating algorithmic and automated decision making. By

establishing standards, promoting transparency, ensuring fairness, and encouraging human oversight, the Act claims to aim to foster trustworthy and ethical AI systems that protect individual rights and promote innovation while preventing algorithmic discrimination Sen. Cantwell (2024). The Act also includes establishment of an Artificial Intelligence Safety Institute and tasking tasked (NIST) with developing a voluntary risk management framework to ensure AI systems are trustworthy.

The Act requires proactive and continuous efforts to ensure that AI systems do not produce biased outcomes based on protected characteristics although focus on development of voluntary standards and metrics. The Act includes provision for human alternatives and oversight in AI systems ensuring mechanisms for human intervention in automated decision-making. This is an aspect that is a focus on in this research as we define a process that involves the underwriter oversight as an inherent part of the process.

US States

The Colorado State Department of Insurance adopted regulation *Notice of Adoption - New Regulation 10-1-1 Governance and Risk Management Framework Requirements for Life Insurers' Use of External Consumer Data and Information Sources, Algorithms, and Predictive Models — DORA Division of Insurance* (n.d.) requiring insurers to remediate any unfair discrimination detected when using external consumer data and information sources (ECDIS) and any associated algorithms/models. The regulation requires from insurers a framework comprised of a gap analysis, compliance roadmap and risk assessment rubric to prioritize high-risk use cases of ECDIS. The insurer must further establish a cross-functional governance group to oversee the operation of the model as well as document policies, processes and procedures related to the full lifecycle of ECDIS from design, development and testing through to operation. Noteably specific documentation is required around racial bias testing. As noted by *The Final Colorado AI Insurance Regulations: What's New and How to Prepare* (n.d.) the requirement to remediate detected discrimination could raise concerns about unintended consequences.

This is significant legislation in being the first to specifically address predictive models in insurance imposing specific governance and risk management requirements on insurers, and frameworks to prevent and remediate algorithmic discrimination in addition to regular auditing of their AI models. This includes documenting policies and procedures, ensuring senior management accountability, and addressing consumer complaints about AI system usage.

Colorado's AI legislation sets a precedent in the US for comprehensive regulation of high-risk AI systems, strengthens consumer protection, and fair practices in consequential decision-making processes.

Other States

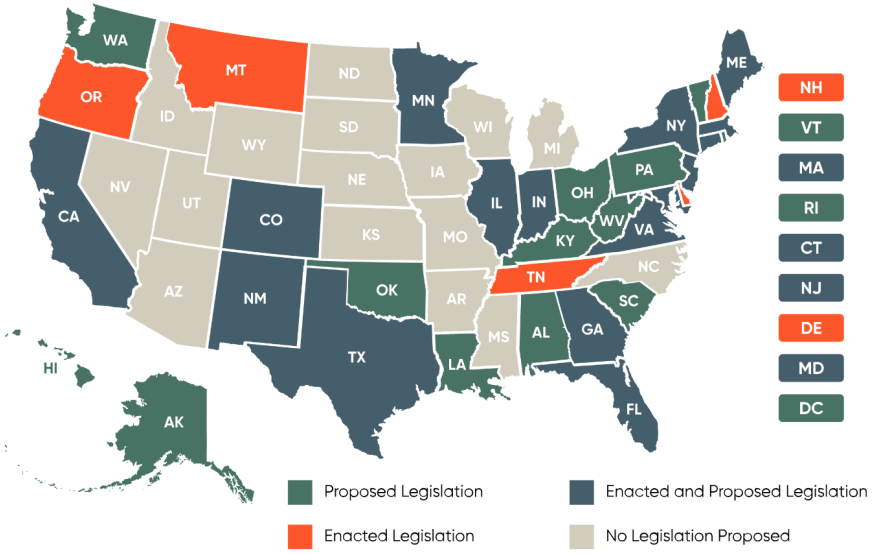


Figure 2.1: US State AI legislation status.Paisner (2024)

2.6.5.2 Europe

GDPR Article 22: Automated Individual Decision-Making

The General Data Protection Regulation (GDPR) *Art. 22 GDPR – Automated Individual Decision-Making, Including Profiling* (n.d.), implemented by the European Union in May 2018, is one of the most comprehensive data protection

regulations globally. Article 22 of the GDPR specifically addresses automated individual decision-making, including profiling. It covers data subjects right not to be subject to automated decisions or "decisions made solely on automated processing", including profiling, without express consent, with non-compliance having potential for large fines (*Art. 22 GDPR – Automated Individual Decision-Making, Including Profiling*, n.d.).

EU Artificial Intelligence Act (AIA)

The European Union is also working on the Artificial Intelligence Act (AIA) *High-Level Summary of the AI Act — EU Artificial Intelligence Act* (n.d.), aiming to create a legal framework to regulate AI classifies AI systems into four risk categories. High-risk AI systems, such as those used in critical infrastructure, including transportation, energy), law enforcement, and healthcare (*High-Level Summary of the AI Act — EU Artificial Intelligence Act*, n.d.).

Insurance underwriting falls under the category of "Access to and Enjoyment of Essential Private Services and Public Services." AI systems used in insurance underwriting assess the risk profiles of individuals and determine the terms and pricing of insurance policies. Given the significant impact these decisions can have on individuals' financial stability and access to essential services, they are likely to be classified as high-risk under the EU AIA. High-risk systems must adhere to strict requirements, around transparency, accountability, bias mitigation, human oversight, and robustness. For this research in addition to performance we will evaluate the model under these criteria.

[paragraph from INS5103 work commented out - review and integrate]

2.6.5.3 International

The OECD Principles on AI call for AI systems to be transparent and accountable European Parliament. Directorate General for Parliamentary Research Services. (2019)and The Council of Europe is drafting a legal framework to ensure the development and use of AI respects human rights, democracy and the rule of law *ECHR - Homepage of the European Court of Human Rights - ECHR - ECHR / CEDH* (n.d.).

2.6.5.4 Analysis

Considering the rapid acceleration in the sophistication of machine learning models, we find that balancing the benefits of precise risk assessment with the social goals of insurance becomes more complex.

In general, while there is growing awareness of the need for algorithmic accountability and transparency, comprehensive legislation is still emerging. Existing laws provide some protections; for example, The US Equality Act 2010 Tobin (2024) prohibits discrimination in the provision of services, which could apply to algorithmic systems. However, more targeted regulation may be required to fully address the challenges posed by algorithmic decision-making systems.

For example, in addition to ensuring that individuals are not subject to bias or unfair discrimination, the Colorado State Dept of Insurance legislation *Notice of Adoption - New Regulation 10-1-1 Governance and Risk Management Framework Requirements for Life Insurers' Use of External Consumer Data and Information Sources, Algorithms, and Predictive Models — DORA Division of Insurance* (n.d.) requires insurers to prove and monitor this on an ongoing basis and to take remedial action where it might occur.

In some regions, regulations are more permissive of using certain types of individual data that can end up being used in the underwriting decision. However, this varies greatly by jurisdiction (e.g., stricter in EU with GDPR). In the US, in the experience of the authors that insurers have direct access to a wide range of medical and behavioural data on the individual from third-party data vendors, a survey of these "traditional" data types is provided in **?**. This is provided subject to the applicant's consent to the terms of the HIPAA, typically provided as part of the application process.

It could be argued that regulators should also intervene to ensure fair access to insurance. This is pertinent in the context of the trend towards individualisation and the potential for individuals to lose out based on an apparent trend away from social solidarity of risk pooling towards more precise individual risk scoring.

2.7 Research Gap

Introduction

The life insurance industry has generally been considered a laggard in terms of technology adoption. Today, life insurance underwriting blends rule-based inference engines with referral to manual underwriting processes for complex and high-value cases. Fully automated application decision rates, also known as straight-through processing (STP), vary widely among insurers depending on client base profile, face amount, maturity of the rules base but typically vary between 30% to 80%.

Availability of large historical underwriting data sets at insurers, a competitive environment, and the desire to increase STP rates means a high level of interest in the potential of predictive models to improve STP rates. However, as adoption has been slow, this research contends that two significant contributing factors are lack of sufficient transparency in such models and their potential to undermine rather than augment the role of the underwriter.

Current State of Research

Contemporary literature in relation to insurance is relatively sparse, presumably relating to commercial sensitivities. A common theme in the research is the use of supervised machine learning techniques to create predictive models from the publicly available Prudential data set (2016) such as the research published by Boodhun and Jayabalan (2018) and Hutagaol and Mauritsius (2020) or more rarely, the results are published based on insurer proprietary underwriting datasets such as those published by MassMutual (Maier et al., 2019).

A significant study was MassMutual training data that spanned 15 years to implement a mortality model and develop a life score metric with which to evaluate applications, reporting an increase of 30% in appliant uptake using the automated system and continuing to implement it fully, resulting in savings of millions of dollars over two years Maier et al. (2019). Wang (2021) at Lloyds of London is a

researcher in this space notable for a focus on XAI. Wang trained a series of supervised models on reinsurer data with the best result from XGBoost with a test accuracy of 0.81 and using SHAP and LIME to explain decisions. Varadarajan and Kakumanu (2024) recently surveyed the work of nine researchers attempting to determine the optimum machine learning model to enhance the accuracy of predicted policy issue decisions and to determine the strategies used to arrive at individual risk predictions. All but one used the same Prudential data set with majority finding that either RandomForest or XGBoost being the most effective model, with a notable absence of focus on explanability result from the research. The origins of insurnance are rooted in mutualisation of risk after reforming an initial more negative image associated with "gambling on lives" in the 18th and 19thC (F. Ewald, 2019) being replaced with the principles of social solidarity and common security. The initial adoption of automation did not disrupt that focus on rules-based systems and decision support systems that support the manual underwriter. However, contemporary research focused on individual risk scoring models threatens this principle.

Identification of Gaps

It is recognised by Maier et al. (2019) that "topics related to fairness and transparency of complex models are equally crucial to study" However, explainability appears to be somewhat of an afterthought in much of the literature. Maier et al. (2019) admits the weakness of feature importance techniques such as SHAP that while "these quantitative contributions can serve as an explanation, ..these methods may not directly generate actionable explanations".

Contemporary research focused on individual risk scoring models threatens the traditional principle of risk mutualisation but is not much discussed in the contemporary technical literature, which appears instead to strive towards greater model accuracy. The reality is that insurers are unlikely to adopt such models in the decision making systems unless interpretability and explainability are also sufficient.

The limitation of previous studies is the lack of publicly available data sets, with

the Prudential data set being used in the majority of studies, obfuscated to the extent that it is unclear to properly evaluate performance results. There is a clear gap in literature in the application of unsupervised techniques to this dataset, either to identify new features that may augment the performance of existing supervised techniques but more particularly in the context of building a potentially inherently interpretable visualisable model that is more explainable in terms of the reasoning behind individual decisions.
The dangers of the trend to individual risk score based to mutualisation is exposed in Cevolini and Esposito (2020) but no research endeavours to actually address this question in the related research studies.

Relevance of Identified Gaps

The gap that exists in relation to transparency relate to an emphasis on performance metric such as accuracy at the expense of interpretability and explainability.
The gap that exists in relation to a drift away from mutualisation towards a more profit-driven model that may ultimately undermine insurance markets and disenfranchise more vulnerable consumers who are no longer able to avail of insurance, being squeezed out by insurer models that target them as no less profitable as individuals. An approach that reaffirms the notion of risk categorisations in predictive underwriting models while aiming to achieve comparable performance with typical supervised models can speak to Responsible AI.
The gaps that exist in underwriter decision support as opposed to underwriting automation can also speak to Responsible AI by retaining the pivotal oversight role of the underwriter remaining the key decision maker by being able to visualise and confirm the risk categories within the data before assigning risk assessments to the categories. This paradigm is more likely to be adopted than the former, which appears to trend towards greadually replacing the underwriter function. The visualisation capability should rather empower the underwriter, allowing potential discovery in the data.
The lack of data sets presents a challenge for any new research. Running the standard supervised techniques on a new proprietary data set would contribute

significant value to the research. If this is not possible then running the standard.....

Research Focus

This research intends to explore the application of contemporary unsupervised learning techniques on the Prudential data set having the aim of providing a means for underwriters to visually fine-tune the identification of clusters so that they can closely align with predefined risk categories. The research will evaluate how closely the results from the use of such cluster identification can match the results from the supervised methods. The research will evaluate how risk classification predictions on unseen applications can be interpreted and explained in a user-friendly way that is potentially useful to underwriters and customer. Further, the research will evaluate how this application of advanced clustering algorithms can be used to explain predictions in a robust mathematical way that can withstand regulatory and legal scrutiny. The research will apply supervised methods as a reference for evaluation in comparison with unsupervised methods. A key hypothesis is that application of unsupervised learning techniques for underwriting risk classification can provide models suitable for development into tooling for use by underwriters to augment understanding of the data, reviewing and applying risk categorisations for use by automated decisioning system. Another hypothesis is that these models are inherently more interpretable from the perspective of being visualisable and explainable from the perspective of having known mathematical techniques underpinning the definition of the clusters. This contrasts with the supervised approaches, which are effectively black-boxes in terms of and so rely on post hoc explainability techniques which lack sufficient mathematical rigour in explaining decision rationale.

Conclusion

The research community has explored different methods to enhance the interpretability and explainability of AI systems, ranging from formal definitions of interpretability to visual explanations and strategies to improve task performance based on generated explanations. However, the literature also highlights a gap in

focus on the explainability of machine learning models decisions, indicating the need for further research to develop new strategies tailored to specific fields of application (Gunning et al., 2019, p. 3). This research aims to contribute to this gap by developing a novel strategy aimed at enhancing explainability AI systems in the domain of life insurance.

¿¿¿¿¿¿¿¿¿¿

¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡

Additional sections for Literature Review or Analytical Background section..

2.7.1 Fairness in Predictive Models

2.7.1.1 Definition and Importance of Fairness

2.7.1.2 Common Biases in Predictive Models

2.7.1.3 Techniques to Ensure Fairness

2.7.2 Interpretability and Explainability in AI Models

2.7.2.1 Definition and Importance

2.7.2.2 Existing Frameworks and Techniques

2.7.2.3 Challenges and Solutions

¿¿¿¿¿¿¿¿¿¿¿¿¿¿¿¿ ¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡

,

3

Analytical Background

3.1 Dimension Reduction

Dimension reduction techniques can be key to simplifying complex datasets while retaining their essential structure. Here, we discuss two key frameworks as outlined by McInnes (2018): **Matrix Factorisation** and **Neighbour Graphs**, covering a range of algorithms and methodologies.

3.1.1 Matrix Factorisation

Matrix factorisation contains a broad class of techniques from topic modelling to word embeddings, PCA, other probabilistic techniques that all revolve around expressing a data matrix X as the product of two smaller matrices U and V:

$$X \approx UV$$

where X is an $N \times D$ matrix, U is an $N \times d$ matrix, and V is a $d \times D$ matrix. The goal is to minimise the reconstruction error subject to certain constraints, leading to various algorithms.

According to McInnes (2018) what we define $\approx$ approximately equal to mean is a notion of error or loss defined as minimising the loss between our original data and our reconstructed data - minimising the following yields all of the many different matrix factorisation techniques.

$$\text{Minimise} \quad \sum_{i=1}^{N}\sum_{j=1}^{D} \text{Loss}(X_{ij}, (UV)_{ij}), \quad \text{subject to constraints}$$

The following is a sample of some of the key factorisation techniques (Nanga et al., 2021) McInnes (2018);

Principal Component Analysis (PCA)

PCA minimizes the sum of squared errors without constraints:

$$\text{Minimise} \quad \sum_{i=1}^{N}\sum_{j=1}^{D} (X_{ij} - (UV)_{ij})^2$$

PCA is widely used for reducing dimensionality while preserving as much variability as possible by solving an eigenvalue/eigenvector problem.

Sparse PCA

Sparse PCA adds constraints to make the results more interpretable by limiting the number of non-zero entries in U. This method enhances the interpretability by ensuring that each representation is a linear combination of a small number of archetypes.

$$\text{Minimize} \quad \sum_{i=1}^{N}\sum_{j=1}^{D} \left(X_{ij} - (UV)_{ij}\right)^2$$

$$\text{Subject to} \quad \|U\|_2 = 1 \quad \text{and} \quad \|U\|_0 \leq k$$

It turn out that if we constrain to a single archetype per row, it results in K-Means clustering, this results in matrix U having one non-zero entry in each row, with the index in the row being the cluster label, and the archetypes representing the K-Means cluster centroids, although it should be noted that this is not how K-Means is normally computed.

k-Means

If we constrain to a single archetype per row, it results in K-Means clustering.

$$\text{Minimise} \quad \sum_{i=1}^{N}\sum_{j=1}^{D}(X_{ij}-(UV)_{ij})^2 \quad \text{subject to} \quad \|U\|_2 = 1 \quad \text{and} \quad \|U\|_0 = 1$$

This results in matrix U having one non-zero entry in each row, with the index in the row being the cluster label, and the archetypes representing the K-Means cluster centroids. Although this is not how K-Means is normally computed, it shows how we are heading in that direction.

Non-negative Matrix Factorization (NMF)

NMF constrains U and V to have only non-negative entries, making the factorisation more interpretable by ensuring additive combinations of archetypes.

$$\text{Minimize} \quad \sum_{i=1}^{N}\sum_{j=1}^{D}((UV)_{ij} - X_{ij}\log((UV)_{ij}))$$

$$\text{Subject to} \quad U_{ij} \geq 0 \quad \text{and} \quad V_{ij} \geq 0$$

Exponential Family PCA

This generalises PCA using probability theory, modelling data with distributions such as Poisson, and minimising the negative logarithmic likelihood.

$$\text{Suppose} \quad X \sim \Pr(\cdot \mid \Theta)$$

$$\text{where} \quad \Theta = UV$$

incorporating this information into matrix factorisation by looking at theta the product of U and V as the reconstruction not of X but of the model parameters that make X the most likely output. The loss term is the negative log likelihood of observing these data under this model we are reconstructing so that we are looking for a low-rank model that makes our data as likely as possible. We

parameterise this using the exponential family of distributions (covers Gaussian, Poisson, Gamme, Bernoulii etc. distributions). In general for an exponential family distribution.

$$-\log(P(X_i \mid \Theta_i)) \propto G(\Theta_i) - X_i \cdot \Theta_i$$

The negative log-likelihood then is proportional to this where the G function of the theta parameters is set depending on the distribution. McInnes (2018)

Latent Dirichlet Allocation (LDA)

LDA, originating from topic modelling, can be seen as a probabilistic matrix factorisation where the factors represent multinomial distributions.

$$\text{Minimize} \quad \sum_{i=1}^{N}\sum_{j=1}^{D} -(UV)_{ij} \cdot \log\left(X_{ij}\right)$$

$$\text{Subject to} \quad (UV)1 = 1 \quad \text{and} \quad (UV)_{ij} \geq 0$$

improving with latent k variable. If we make U a multinomial over a latent variable k and V provides k as a multinomial over the features, the result of multiplying out provides the parameters for a multinomial distribution

Let

$$U_{ik} = P(i \mid k), \quad V_{kj} = P(k \mid j)$$

Then

$$\Theta_{ij} = \sum_{k} U_{ik} \cdot V_{kj}$$

$$= \sum_{k} P(i \mid k) \cdot P(k \mid j)$$

$$= P(i \mid j)$$

So, the parameters on U . V can be swapped instead for constraints on U and V separately that guarantee applying the right constraints on the multiplied return. In terms of interpretability each row of U (representation) is going to be a multinomial distribution over V (archetypes) so that we can interpret a row of U (some point in our low-dimensional representation) as a distribution over

archetypes with each archetype being a distribution over our features providing a good understanding of what it is doing. Substituting, and adjustment of the constraints moves us towards the following formula.

$$\text{Minimize} \quad \sum_{i=1}^{N}\sum_{j=1}^{D} -(UV)_{ij} \cdot \log\left(X_{ij}\right) \tag{3.1}$$

$$\text{Subject to} \quad U1 = 1, \quad V1 = 1 \quad \text{and} \quad U_{ij} \geq 0, \quad V_{ij} \geq 0 \tag{3.2}$$

Probabilistic latent semantic indexing.

This formula is powerful in terms of interpretability, vectors of probabilities of archetypes, and each archetype is a vector of probabilities of features. Then applying Bayesian principles to apply a Dirichlet prior over the multinomial distribution for U and V provides LDA. Essentially, LDA is a matrix facorization in a probabilistic sense where everything is expressed in terms of multinomial distributions.

In terms of thinking about how data might be modelled, for example, a document can be represented by a bag of words, effectively a multinomial distribution. So defining a distribution model for our data we can use quite sophisticated matrix factorisation techniques that have allowed us to move from simple PCA through to much more sophisticated and powerful techniques.

3.1.2 Neighbour Graphs

Neighbour graph techniques involve constructing a graph from the data and embedding it in a low-dimensional space. An intuition provided by McInnes (2018) is that there may exist some low dimensional structure in the data that a linear approach like PCA can not find but however, drawing a graph linking points with an edge if it is in the K nearest neighbors of the point may lead to a graph uncovering the structure.

Fig. 3.1 depicts an interpretation provided by McInnes (2018) with the caveat that it does not hold in high dimensions in real data.

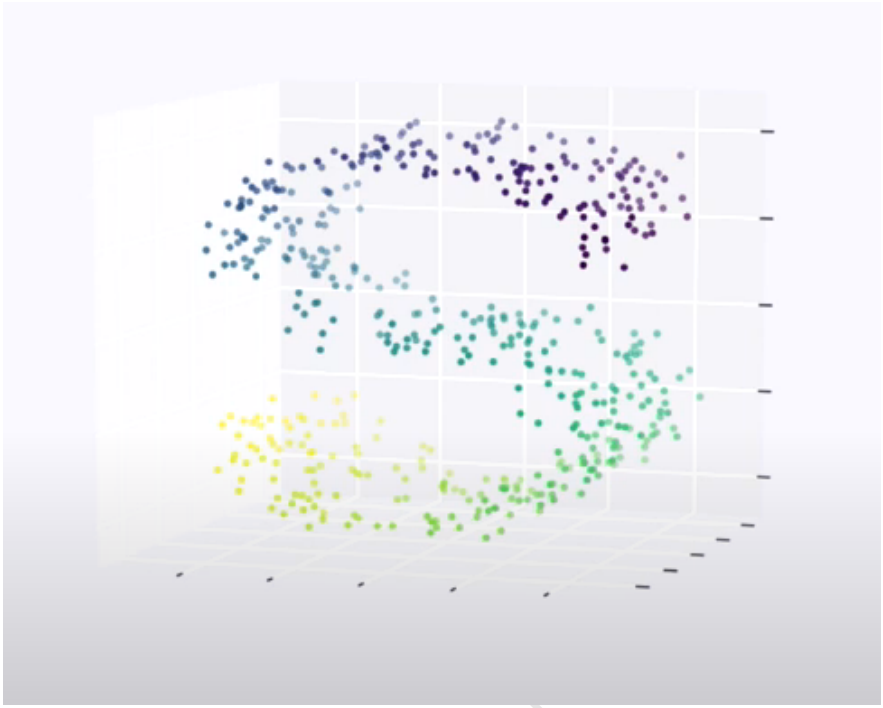


Figure 3.1: Intuition for nearest neighbour graph. McInnes (2018)

Isomap

An early approach to the neighbour graph problem, Isomap constructs a graph using k-nearest neighbours and computes the shortest paths to build a weighted adjacency matrix, which is then factorised.

Building a weighted adjecency matrix out of the graph which comprises the weight of the edge or zero if no edge;

$$A_{ij} = \begin{cases} w(i,j) & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

We can then rely on factorisation to turn the matrix into a low-dimensional representation and take the largest eigenvectors.

Spectral Embedding

This technique weights edges using a kernel and normalises the Laplacian graph, followed by factorisation using the smallest eigenvectors.

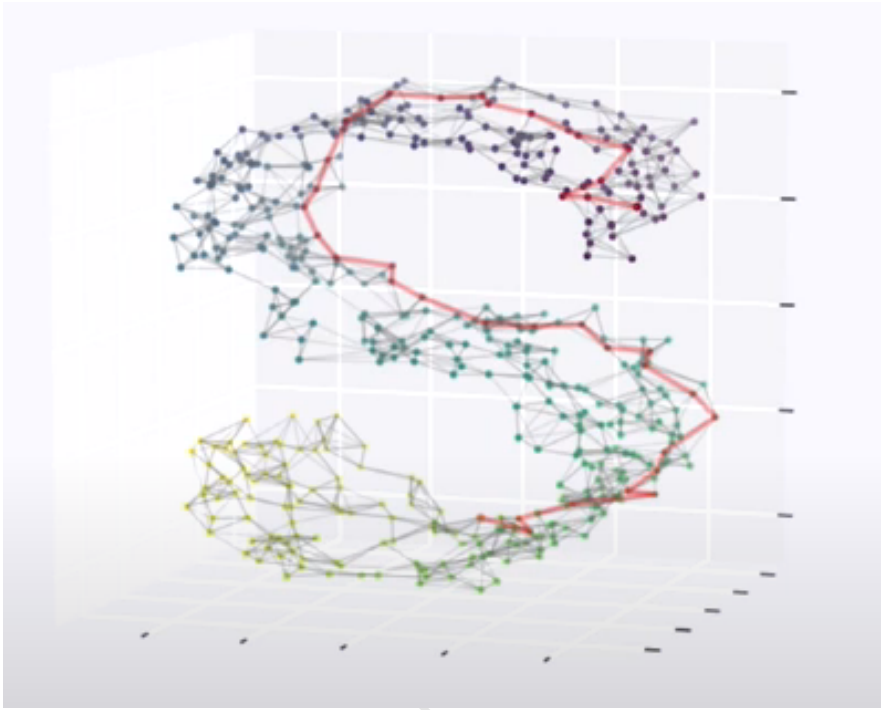


Figure 3.2: Isomap - shortest path through nearest neighbour graph. McInnes (2018)

We compute the Graph Laplacian which takes the weights attached to each edge divide out by the marginals of either the rows and the columns to normalise, resulting in a matrix which can be factored taking the smallest eigenvectors.

$$L_{ij} = \begin{cases} -\frac{w(i,j)}{\sqrt{d_i \times d_j}} & \text{if } i \neq j \\ 1 - \frac{w(i,i)}{d_i} & \text{if } i = j \end{cases} \tag{3.3}$$

Where d_i is the total weight of row i.

t-distributed Stochastic Neighbour Embedding (t-SNE)

t-SNE varies the kernel width based on data density and uses a force-directed layout to embed the graph. It minimizes the Kullback-Leibler divergence between the high-dimensional and low-dimensional data representations.

This algorithm despite being typically expressed in probabilistic terms can also essentially be considered a graph algorithm. Whereas Spectral Embedding

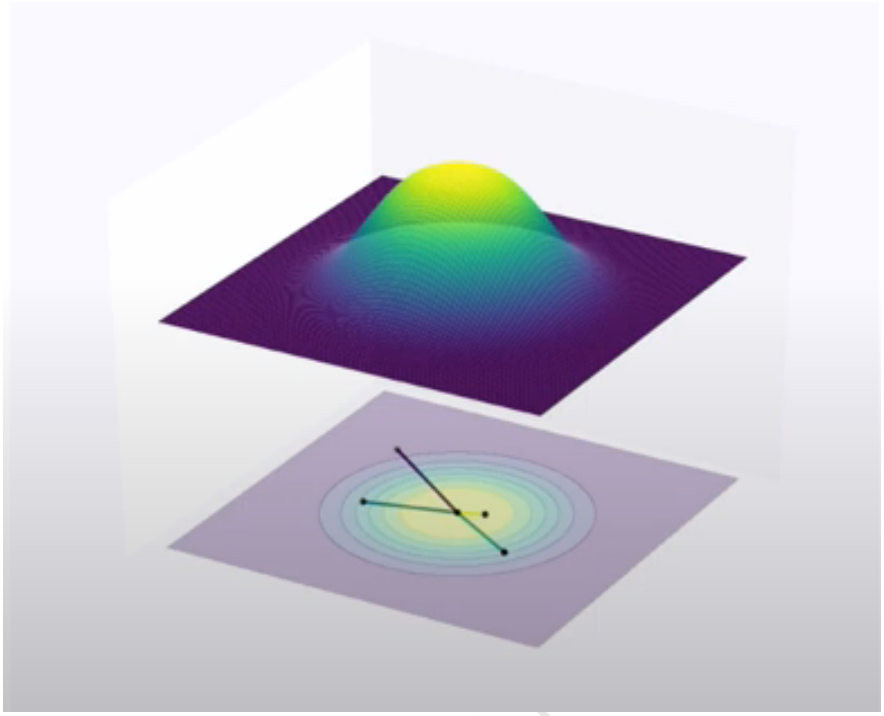
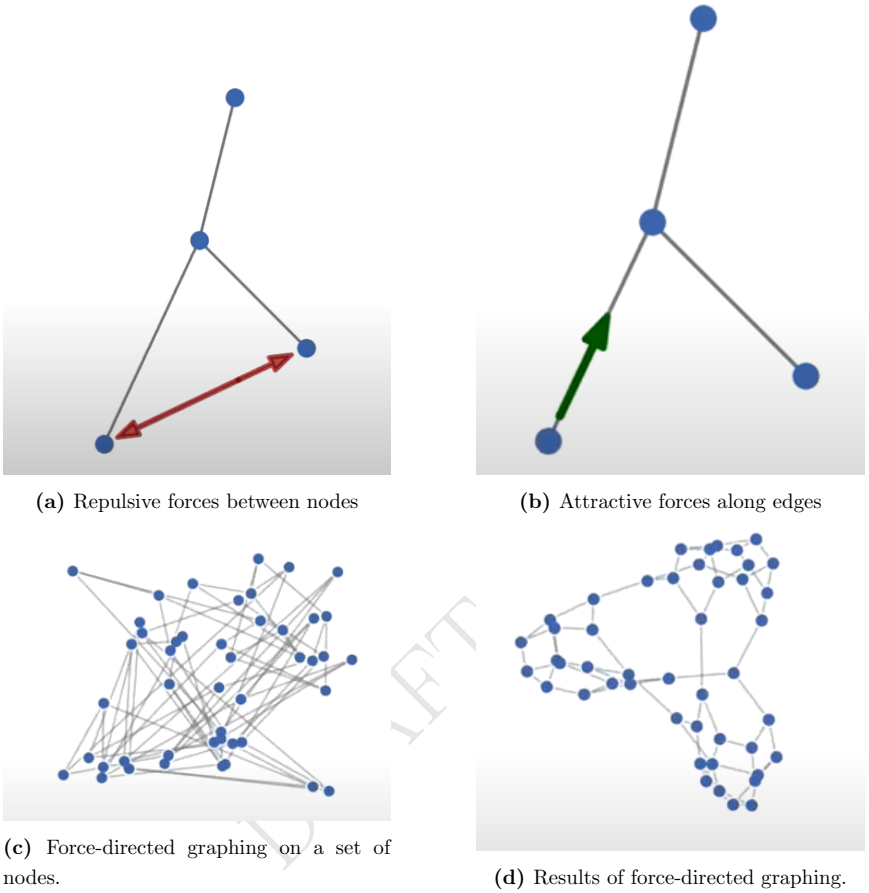


Figure 3.3: Spectral embedding. McInnes (2018)

uses a single bandwidth for the kernel across all of the data, t-SNE varies the width of the kernel according to the density of the data based on a formula. Essentially, taking K-nearest neighbours, the size of the kernel varies to match the K neighbours then normalising edges so that outgoing edge weights sum to one. Also, we symmetrise to average two edges with different weights between two points and finally renormalise so that the total edge weights in the whole graph sum to 1.

Fig. 3.4 uses a force-directed graph layout that is typically used in network science. Here, nodes have repulsive forces pushing them apart, while edges have an attractive force that pulls points together along them, with the computation of these forces influencing the algorithm. t-SNE is actively performing this type of force-directed layout with a m way of computing the forces. (McInnes, 2018)



(a) Repulsive forces between nodes

(b) Attractive forces along edges

(c) Force-directed graphing on a set of nodes.

(d) Results of force-directed graphing.

Figure 3.4: t-SNE force-directed graphing. McInnes (2018)

Uniform Manifold Approximation and Projection (UMAP)

This algorithm is in essence K-Nearest neighbours weighted according to mathematics based on category theory. Similar to t-SNE but applying a variable-sized kernel derived from topological techniques and then similarly using a force-directed layout with different derived forces. However, such topological techniques work only when the data is uniformly distributed in a manifold, and most of the data do not meet this criterion in relation to the ambient space in which the data lie. UMAP addresses this problem by varying the size of the kernel, allowing us to effectively vary the metric on the manifold to meet this requirement. However, this creates the problem of the compatibility of different topological patches with varying notions of distance and different techniques required to stitch these

patches back together. t-SNE, in contrast, for example is simply able to simply average. The core intuition of UMAP is similar to Spectral Embedding in putting a kernel over data points and looking for similarities; however, here with UMAP we must adapt to deal with these topological obstructions to get the algorithm to work in this more general case. Essentially, we can view this as scaling the kernels to fit on the right metric for the underlying topology. (McInnes, 2018)

3.1.2.1 Performance and Scaling

In terms of performance and parallelising clustering algorithm computations, one observation is that these algorithms are wildly different in terms of computational complexity. Although PCA is relatively trivial to compute since its objective function can be solved in a closed form with linear algebra. However, adding further constraints this becomes an optimization problem which is more complex and for example probabilistic constraints may result in more a non-convex optimization problem with higher computational complexity. Unfortunately, the computation complexity of these techniques is disparate, not mirroring the way we linked the techniques together above in terms of the underlying theory in the above sections. In fact addressing computational complexity may rather involve taking "short-cuts" by introducing various different concepts. (McInnes, 2018)

3.1.2.2 Conclusion

Matrix factorisation and neighbour graphs provide a comprehensive framework for understanding dimension reduction techniques. These methods, through various constraints and optimisations, offer powerful tools for simplifying complex datasets while preserving their essential structures.

3.2 Clustering

Although definitions of clustering vary Healy and McInnes (2016) describes it essentially as about finding "groups" of data that are "similar" with the meaning of this depending upon the application. Applications might range from balancing data partitions on an HPC or distributed system, summarising and exploring

data, helping to embed in vector spaces for machine learning algorithms, and finding patterns with the data. For example, the HPC example might prioritise the balancing aspect across clusters over the similarity of the data within clusters. (Healy and McInnes, 2016)

3.2.1 Centroid-based Clustering

Here and in Fig. 3.5 we review the different centroid-based clustering techniques as outlined by Healy and McInnes (2016).

k-Means

k-means essentially picks k centres in the space and iteratively moves them to minimise the distance within groups while maximising the distance between groups. One limitation is the need to pre-select k centres. As a centroid-based technique, every point is assigned to its closest center, resulting in Gaussian spheres, which may not always be appropriate. Furthermore, k-means do not handle noise, assigning every point to a cluster regardless of whether it represents an outlier.

Affinity Propagation

Each point votes on the centre that best represents it, allowing nonsymmetric similarity measures and eliminating the need to pre-select the number of clusters. However, results are sensitive to the preference parameter, which requires careful calibration. The algorithm is slow and typically ends with k-means, reintroducing the spherical cluster problem.

MeanShift

Starting with k centroids, the algorithm estimates local density around each centroid, causing them to climb to local maxima and identify high-density regions. Points in low-density areas can be viewed as noise. Despite this, MeanShift still assumes spherical cluster shapes.

Spectral Clustering

This algorithm assumes that high-dimensional data can be represented by a low-dimensional manifold. It approximates data with a k-nearest neighbour graph, using an adjacency matrix and linear algebra to project into a lower-dimensional space for clustering. However, it typically ends with k-means, retaining spherical cluster issues.

Birch

This hierarchical algorithm partitions data into trees, with leaf nodes representing groups. Running another algorithm like k-means on these groups results in larger spherical clusters. Birch is fast, has low memory consumption, supports streaming, and does not require the number of clusters to be pre-selected. However, it has a parameter that is hard to calibrate.

3.2.2 Hierarchical Clustering

Agglomerative Clustering

This algorithm merges the nearest pairs of data points iteratively, with the shape of the cluster influenced by the distance definition. Methods like single linkage produce snake-like clusters, while complete linkage and Ward's method produce spherical clusters. This technique is suitable for smaller datasets with hierarchical structures.

3.2.3 Density-based Clustering

DBSCAN

Density-based DBSCAN finds dense regions surrounded by sparser areas. It uses a fixed spherical area around each point to identify dense regions and outliers, allowing for arbitrarily shaped clusters. The performance of the algorithm depends on a well-chosen spherical area parameter.

3.2.3.1 Challenges and Quality of Clustering

3.2.3.2 Measuring Goodness of Clustering

Measures of clustering quality include intra-cluster separation, inter-cluster homogeneity, density, and cluster-size uniformity. The appropriate measure depends on both the algorithm and the application requirements. Clusters do not need to be spherical and not every data point needs to belong to a cluster.

3.2.4 Limitations of Scale

3.2.4.1 Density-based Clustering and HDBSCAN

Healy defines a cluster as a connected component of a level set of the probability density function. Creating a heat map of the estimated density and selecting contours to lasso data points into clusters can be intuitive, but computationally expensive.

3.2.5 Towards Scalability

3.2.6 Locally Approximate Density

Rather than approximating density globally, local approximations around the points are more computationally feasible. DBSCAN, with parameters 'epsilon' and 'minimum points' efficiently identifies dense regions.

3.2.7 HDBSCAN

HDBSCAN improves DBSCAN by introducing a minimum cluster size, simplifying the cluster tree. It efficiently scales to large datasets by using spatial indexing and the Dual Tree Boruvka algorithm for minimum spanning tree computations.

3.3 Conclusion

Centroid-based approaches like k-means are not always preferred. HDBSCAN often performs well and is suitable for many applications, especially when defin-

ing clusters for insurance underwriting. Clustering in high dimensions requires dimensional reduction techniques, and HDBSCAN can handle dimensions up to 50-100 effectively.

3.3.1 The Curse of Dimensionality

Clustering in very high dimensions is impractical without dimensional reduction. If the data has a low-dimensional manifold, the dimensional reduction retains meaning before clustering. HDBSCAN performs well with up to 50-100 dimensions, but struggles beyond that point. Proper standardisation of individual dimensions is essential for effective clustering.

3.4 Theoretical Framework or Models

3.4.1 Methodological Approaches

3.4.1.1 Data Preprocessing

Both Boodhun and Jayabalan (2018) and Hutagaol and Mauritsius (2020) studies preprocess the Prudential dataset using Missing At Random (MAR) and multiple imputation methods to replace missing values, omitting features with greater than 30% missing values. Boodhun and Jayabalan (2018) tested CFS and PCA dimensionality reduction techniques eventually reducing the number of attributes to twenty using PCA. Hutagaol and Mauritsius (2020) essentially follow the Boodhun and Jayabalan (2018) study also combining medical keywords into a single attribute. In regard to preprocessing there is no significant variation in Boodhun's approach. On the other hand, Wang (2021) applies natural language processing and unsupervised techniques (k-means clustering) to those fields containing free-text descriptions of medical conditions and occupations prior to applying mutual information and recursive feature elimination (RFE) techniques for feature selection. In contrast with such automated feature selection approaches, Maier et al. (2019) instead bases selection on the advice of medical and actuarial experts, performing machine learning techniques to help validate and support the process.

3.4.1.2 Modelling

Boodhun and Jayabalan (2018) implemented basic algorithms on the Prudential dataset to build predictive models using Multiple Linear Regression, REPTree, Random Tree and Multilayer Perceptron (MLP) algorithms. Next, Hutagaol and Mauritsius (2020) working with the same dataset implemented further machine learning algorithms namely Support Vector Machine (SVM) with various kernels, Random Forest and Naive Bayes determining the best model through analysis of resulting metrics including accuracy, precision, and recall metrics. Later, Wang (2021) employed ensemble methods of XGBoost (eXtreme Gradient Boosting), Random Forest, and Bagging with the aim of further improving underwriting decision predictive performance. While these studies focus largely on optimising well-known machine learning algorithms Maier et al. (2020) took a more nuanced approach focus modeling on the Random Survival Forest (RSF), discovering superior results to those from Cox statistical survival risk model and deep neural networks.

Discuss any theories or models that your research is based upon or influenced by.

3.5 Supervised Learning

3.6 Unsupervised Learning

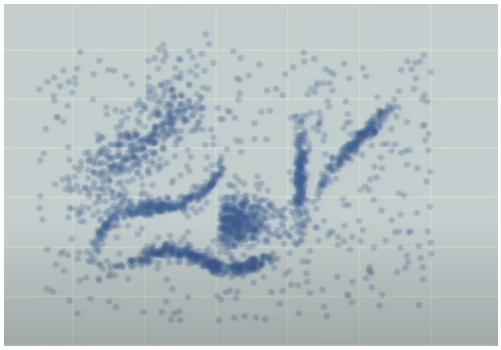
3.7 Explainable AI (XAI)

Approaches...

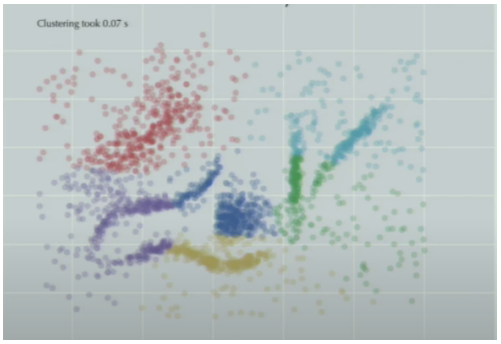
3.7.0.1 Post-hoc

3.8 Interpretable AI

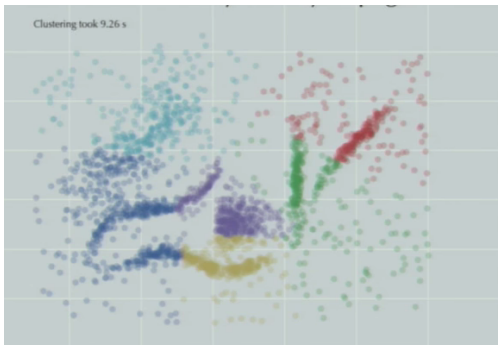
Approaches...



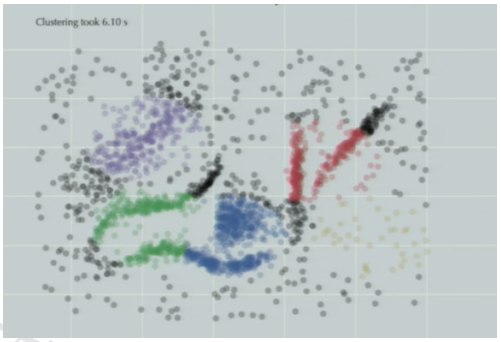
(a) Original data



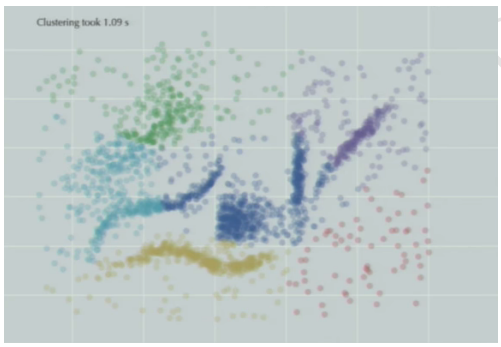
(b) k-means clustering



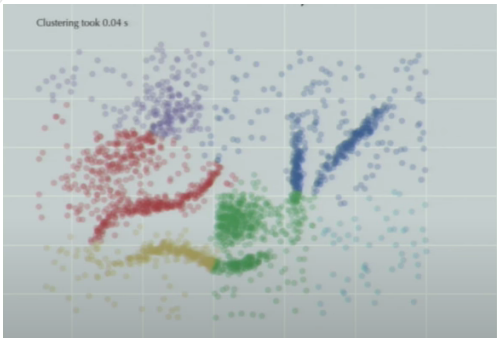
(c) Meanshift clustering



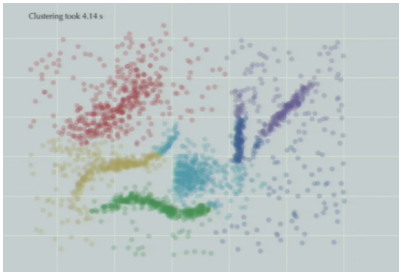
(d) Spectral clustering



(e) Birch clustering

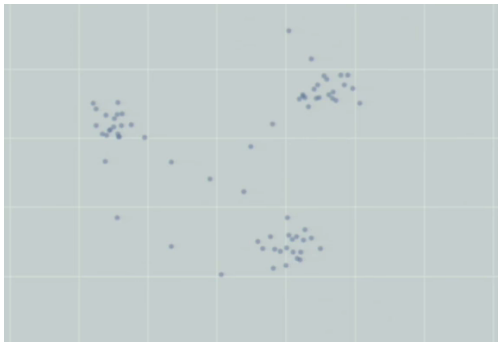


(f) Spectral clustering

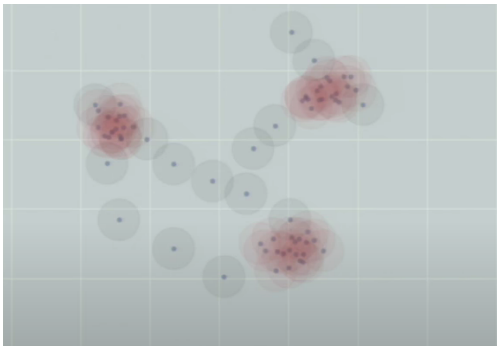


(g) Agglomorative (hierarchical)

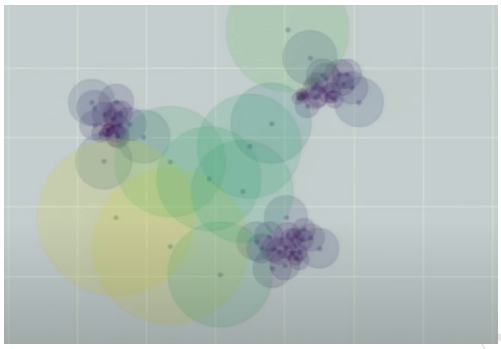
Figure 3.5: Centroid-based clustering techniques



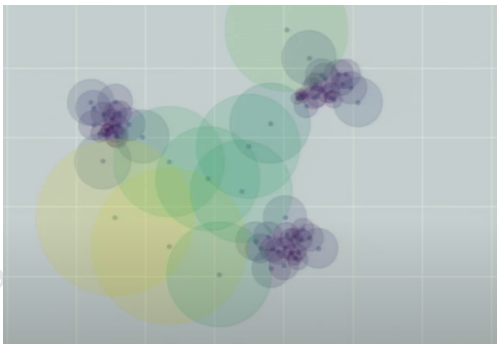
(a) Original data



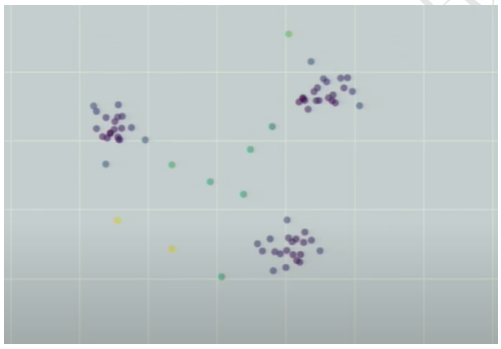
(b) Local approximate density



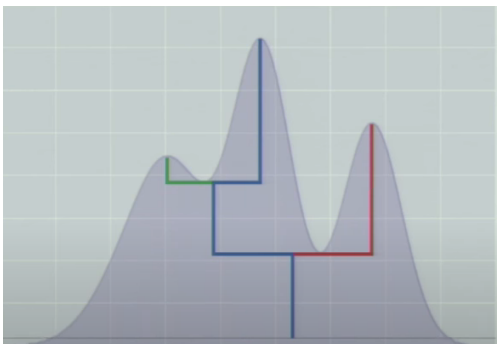
(c) Radius-based density colouring



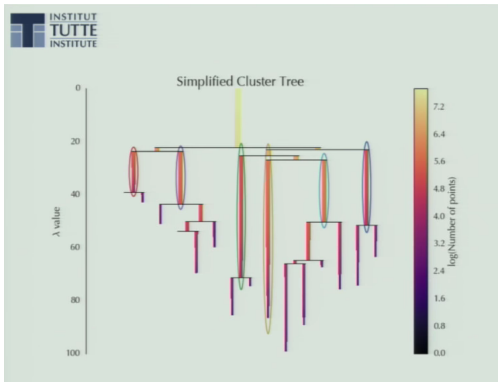
(d) Radius-based density colouring



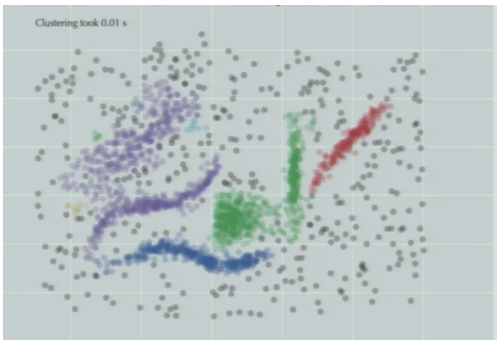
(e) Coloured points



(f) Level set connected components

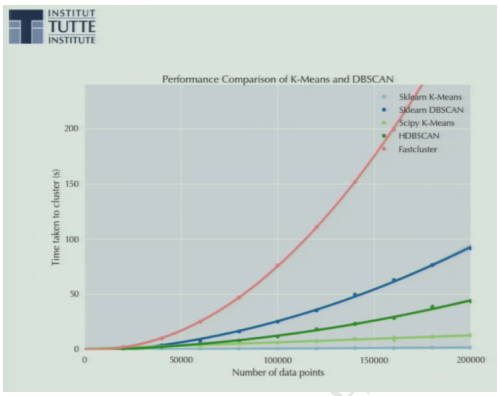


(g) Density cluster tree (simplified)

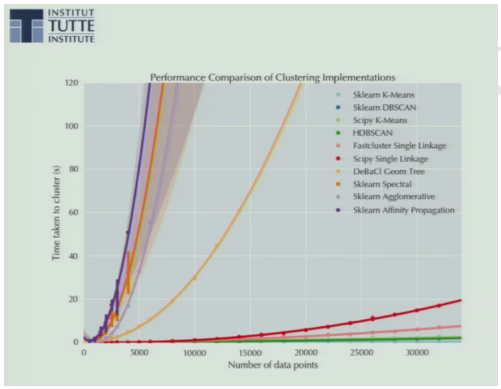


(h) DBSCAN example

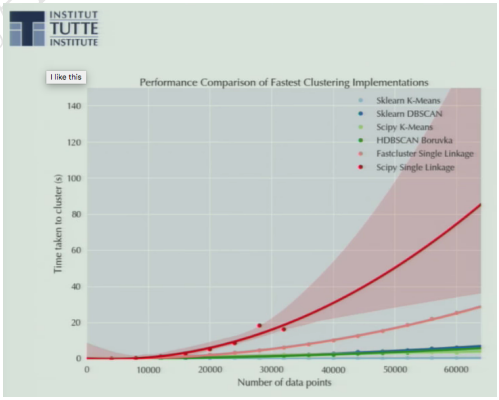
Figure 3.6: Density-based clustering process



(a) k-Means vs DBSCAN



(b) Clustering performance comparison



(c) DBSCAN example

Figure 3.7: Comparing fastest algorithm performance

4

Methodology

This section contains a detailed explanation of the research design methods, data analysis, and machine learning techniques used to conduct the research.

4.1 Introduction

4.1.1 Research Problem

The US insurance company MassMutual asserts that "it is imperative that analysts and underwriters can effectively explain why an individual applicant received a given offer". (Maier et al., 2019) This can be viewed in the context of the observation that "machine learning models [have] become increasingly opaque" Maier et al. (2019) with even simpler linear models being challenging to interpret. Boodhun and Jayabalan (2018) and Wang (2021) employ SHAP and LIME techniques to provide a degree of interpretability to model predictions. Indeed Maier et al. (2019) also adopted an approach based on the same techniques to identify the relative contribution of each feature to a decision. However, it should be noted that these techniques do not provide a complete understanding of how the model arrived at its decision, and can be difficult to interpret for nontechnical users, among other limitations.

This research question asks whether in high-risk and highly regulated applications that require a high degree of interpretability and explainability, these

aspects of undersupervised techniques can be leveraged while achieving comparable performance to supervised methods, which generally suffer from post hoc explainable methods which, while capable of faciliating a narrative about the contributing factors to a decision, lack rigour of a logically reasoned causal explanation, and as Maier et al. (2019) claim "these methods may not directly generate actionable explanations".

4.1.2 Methodology Approach

The objective here is to apply the methodology outlined in Fig. 4.1 to the Prudential Insurance dataset Montoya and Cukierski (2015) to predict insurance underwriting classes composed of unsupervised learning and human-in-the-loop (HITL) elements and to compare this with the more standard supervised learning approach from both a quantitative and qualitative standpoint.

The first few steps are common to both streams, which is Exploratory Data Analysis (EDA) to better understand the data through descriptive analytics, statistical, and plotting techniques. This is followed by pre-processing the data to ensure it is suitably cleansed and formatted, and ingestable for subsequent machine learning steps.

The approach taken on the leftmost branch *"Supervised Learning"* uses the methodology followed by Sarpal (2023) using supervised learning techniques and evaluating quantitatively based on performance metrics and also qualitatively from an explainability perspective.

In the rightmost *"Active Learning"* branch, we move to using our proposed hybrid of unsupervised learning and human-in-the-loop (HITL) based methods on the same preprocessed dataset and also evaluate performance and explainability. This time we have the objective of being able to evaluate explainability quantitatively. Finally, we will compare the results of both approaches in terms of both performance and XAI and contrast the strengths and weaknesses of the approaches.

Next, we describe the key objectives of each step in the diagram in Fig. 4.1;

4.1.2.1 Preparation Steps

- **Exploratory Data analysis on the dataset (EDA)** to explore and report the data set from a statistical perspective.
- **Data Pre-processing** prepares the data for machine learning techniques, applies feature engineering and splits the data into training, validation and testing subsets.

4.1.2.2 Supervised Learning branch

- **Model Training** trains the data on a series of supervised learning algorithms, optimizing these models through hyper-parameter searches and regularization techniques.
- **Evaluate** and rank the quantitative performance of the supervised models, ranking models based on metrics.
- **XAI** - Apply explainability techniques such as SHAP to the best model and evaluate qualitatively.

4.1.2.3 Active Learning branch (Unsupervised / HITL)

Due to the high-dimensionality (complexity) of the data, we need to apply a more advanced approach to clustering. Here, we first apply UMAP (Uniform Manifold Approximation and Projection) an advanced manifold learning algorithm for dimension reduction, in combination with HDBSCAN, a density-based clustering algorithm.

- **Dimensionality Reduction** - Reduce the dimensionality of the data to improve the performance of the clustering step using UMAP.
- **Data clustering** - HDBSCAN identifies clusters of varying densities and can classify some points as noise. Strategies to handle noise points might involve manual review and reclassification. In this case, we will treat as an unclassified category (i.e. ignore).

- **Expert Classification** - Expert underwriter (or tester) assigns risk categories to the clusters by interacting with a 3-D model of the data and clusters, updating risk classifications as appropriate within the model. In this study, we will use the Tensorboard Embeddings Projector project Abadi et al. (2015) to simulate this expert interaction. Ground truth classifications for data points will also be identified visually as a reference.

  – Although omitted from this work, this step might also include analyzing cluster distributions for any evidence of bias or unfair discrimination which could be corrected by update or omission of the associated risk category.

- **Testing** - This step will "predict" by assigning risk class to new unseen test data applying the same dimenssion reduction and clustering algorithms. Using the same unsupervised clustering algorithm(s) to determine which cluster (risk class) new sample belongs to. Ensures consistency between the clustering of historical data and the new point. Once a new sample is clustered, a risk classification is automatically assigned based on the expert assigned classification for that cluster.

- **Evaluate** the performance quantitatively comparing predicted outcomes with ground truth. We can evaluate the mathematics associated with UMAP and HDBSCAN in terms of explainability and also qualitatively evaluate visualisations in terms of interpretability.

4.1.2.4 Compare Supervised and Active Approaches

- **Compare and Evaluate** compares supervised with the active learning (unsupervised learning and HITL) approach proposed here from the perspective of performance and XAI.

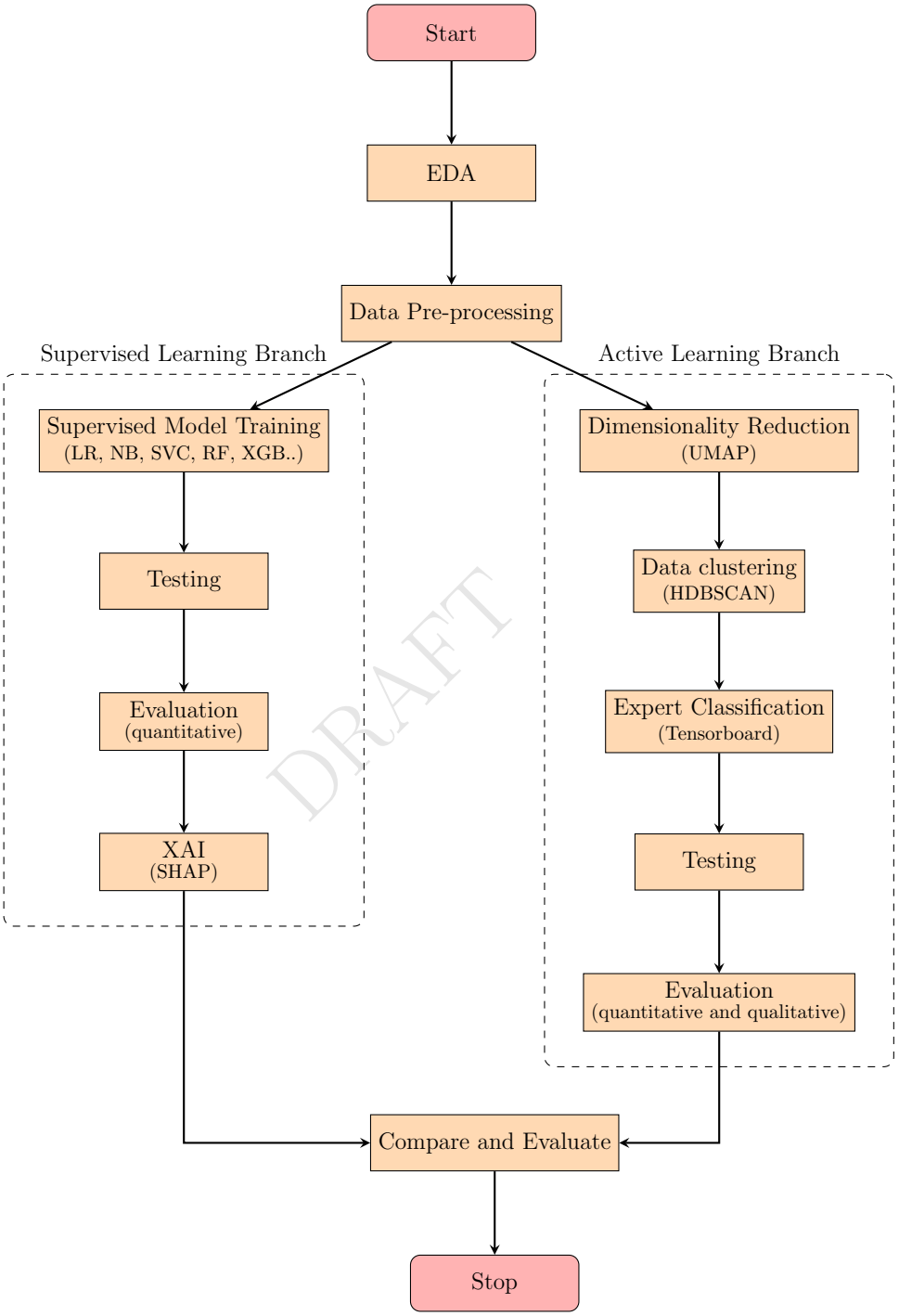


Figure 4.1: Methodology to compare supervised and active learning approaches.

4.2 Data Preparation

4.2.1 The Dataset

The data set used in this study is from Prudential Life Insurance. (Montoya and Cukierski, 2015) this being a real-world data set from the life insurance industry related to insurance applicants and their risk levels that has been anonymised for public access.

From the literature, we see that this is a popular data set for studying insurance underwriting, given the lack of publicly available data sets due to personal privacy and commercial sensitivity concerns. Varadarajan and Kakumanu (2024)

Key details about the data set are:

- Contains 59,381 applications with 128 attributes that describe the characteristics of life insurance applicants.
- Consists of nominal, continuous, and discrete features, all of which are anonymised.

The variables in the data set are:

Attribute	Description
ProductInfo1-7	7 normalized attributes concerning the product applied for
Ins_Age	Normalized age of an applicant
Ht	Normalized height of an applicant
Wt	Normalized weight of an applicant
BMI	Normalized Body Mass Index of an applicant
EmploymentInfo1-6	6 normalized attributes concerning employment history
InsuredInfo1-6	6 normalized attributes offering information about an applicant
InsuranceHistory1-9	9 normalized attributes relating to the insurance history
FamilyHist1-5	5 normalized attributes related to an applicant's family history
MedicalHistory1-41	41 normalized variables providing information on an applicant's medical history
MedicalKeyword1-48	48 dummy variables relating to the presence or absence of medical keywords
Response	The target variable, which is an ordinal measure of risk level with 8 levels

Table 4.1: Prudential Dataset - Attributes and Descriptions (Montoya and Cukierski, 2015)

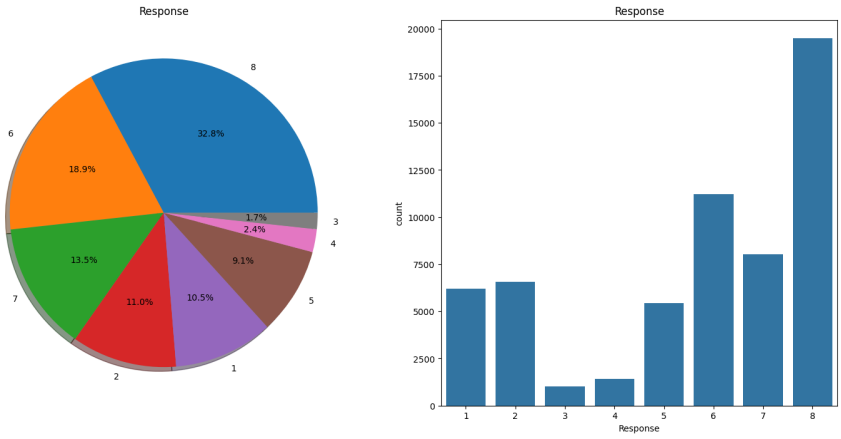


Figure 4.2: Risk category (`Response`) distribution

4.2.2 Exploratory Data Analysis (EDA)

Response Groups

In figure 4.2 the `Response` groups reveal the distribution of different risk categories among the applicants, the risk classification corresponding to the response value 8 being the largest at 33%. It is not obvious how these (obfuscated) values might map to a typical life-risk classification.

The histogram The histograms in Fig. 4.3 show the distribution of applicants across the entire data set according to their risk classification (Response).

Biophysical Information

Here, we focus on analysing the key biophysical characteristics as they relate to the Response variable (risk category). (A detailed analysis of all variables is available from the accompanying Jupyter Notebook.)

The analysis of biophysical information across various Response groups reveals distinct demographic patterns in age, height, weight, and BMI. Younger individuals are predominantly found in higher Response categories, which exhibit sharper peaks and concentrated distributions, particularly in Ins_Age and Ht. Conversely, lower Response categories show broader distributions, indicating a wider age range and variability in physical characteristics for these risk cate-

gories. The weight and BMI distributions further underscore these trends, with higher Response categories tending to cluster around lower weight and specific BMI values.

Fig. 4.3b show that younger individuals and those below the median for BMI and weight are predominantly assigned a risk classification of 8 `Response`. Further, we find from Fig. 4.2 that the largest majority of applications are also assigned this category (33%) the next largest being risk category 6 (18.9%). From these insights, we can reasonably assume that the response 8 category corresponds with *accept with standard rates* classification.

Other Information

KDE density plots were created for variables across the across the other variables in categories `Product, Employment, Insure_Info, Insurance_History, Family_History and Medical_Keywords` and the graphs analysed. With a few exceptions, a consistent trend is that they mostly appear to have a high degree of overlap across the Response variable with no significant differences in relative densities. Hence, the variations between features do not particularly help in predicting the applicant's risk rating. Consequently, we can assume that variations in these features are unlikely to individually contribute to predicting an applicant's risk rating.

Feature Interactions - Correlation

Here, we summarise key conclusions from a detailed analysis of the feature correlations shown in Fig. 4.20 provided in *Notebook 1*:

- Employment and financial impact: Employment and financial status appear to influence policy applications, as evidenced by correlations between Product Info and certain Employment Info features.
- Family history importance: The features of family history show strong correlations with age (Ins_Age) and certain columns of medical history, highlighting the importance of family medical background in risk assessment.

- Medical data interconnection: The strong internal correlations within the Medical History and Medical Keyword sets suggest a high degree of linkage in health-related data.
- Applicant profile complexity: The varied correlations across different feature sets (e.g., Applicant Info, Insured Info, Employment Info) underscore the complex nature of applicant profiles.

In conclusion, the analysis indicates significant correlations, particularly the influence of employment and financial status on policy applications, and the strong role of family history and medical data in risk assessment. Otherwise, the varied correlations across different feature sets (e.g., Applicant Info, Insured Info, Employment Info) underscore the complex interrelation of factors in insurance underwriting and the potential for improved predictive modelling.

Data Preprocessing

Here we summarise the steps taken in *Notebook1* to preprocess the data to prepare them for training.

4.2.2.1 Training Test Split

The data set partition strategy involves the allocation of 20% of the original data sets for testing and 80% for training. Of the training data, 0.25% is retained for validation.

4.2.2.2 Missing Values

We examine the training data for missing values, assuming that they are completely missing at random (MCAR) in Fig. 4.5a. This assumption simplifies the analysis by avoiding complex adjustments required for data missing at random (MAR) or missing not at random (MNAR).

Columns with more than 40% missing values in the training subset are removed to maintain data integrity, as their imputation could introduce significant bias, while others with manageable missing values are pre-processed by imputation.

Iterative imputation is applied to estimate missing values in the remaining columns, ensuring all data is machine-interpretable for subsequent processing stages.

We see in Fig. 4.5b while the distributions of three columns remain largely unchanged, `Family_Hist_4` and `Medical_History_1` exhibit additional probability density and peak splitting, necessitating caution in evaluating their significance due to potential impacts on model interpretation and accuracy.

4.2.2.3 Data Leakage

Data snooping can lead to models appearing to perform significantly better on test data than on unseen data, compromising their generalisation capabilities. (Brownlee, 2016) To mitigate this risk, we implement the following strategies:

1. Splitting the dataset into training, validation, and test sets before any preprocessing.
2. Applying data transformations solely within the training set and then to the validation and test sets.
3. Performing feature engineering exclusively on the training data and applying the engineered features to the test set without recalculating.
4. Employing cross-validation techniques that respect data structure and temporal aspects.

These practices ensure a clear separation between the training and evaluation phases, improving the accuracy and reliability of the model.

4.3 Feature Engineering

In this section, we outline how supervised and unsupervised learning techniques were employed to generate new features within our dataset, thereby enhancing the model's predictive accuracy and robustness by capturing meaningful patterns and relationships that improve generalisation to unseen data.

One-hot encoding

We applied one-hot encoding to the `Product_Info_2 column`, transforming categorical data into binary numerical representations. This conversion enables classification models to process and interpret these features effectively, as most algorithms require numerical inputs for optimal performance.

Scaling and Normalisation

Here, we perform min-max scaling on the encoded datasets to normalise the feature values between 0 and 1. This ensures that all features have variances of the same order of magnitude, preventing any single feature from dominating the objective function and thereby facilitating accurate and balanced model learning.

K-means Clustering

We employ clustering of K-means to group applicants according to commonalities, train the algorithm on the training data subset, and predict clusters for the training and test subsets. These cluster labels are then added as a feature to better help understand (and predict) risk rating assignments. As indicated by the "elbow" in the curve at k=15, further increases in k do not significantly reduce inertia; thus k = 15 is selected for `Kmeans()` initialisation.

4.3.1 Feature Selection

In this section, we apply various methods to identify the most important features for fitting each classification model, aiming to prevent overfitting. At this point, we switch to the validation data set for feature selection and model refinement to avoid data leakage from the training data set.

Mutual Information

In this section, we calculate the mutual information (MI) scores of the validation dataset using custom functions based on the `mutual_info_classif()` function from sklearn. This helps identify features that significantly reduce uncertainty in

the `Response` variable, guiding feature selection. As can be seen in Fig. 4.7a The top 5 features identified are `BMI, Wt, Product_Info_4, Medical_Keyword_15, and Medical_History_23`, indicating strong statistical dependencies with the Response variable. Features with low or zero MI scores are less useful for predictions. (Brownlee, 2020)

Multicollinearity Analysis

In this section, we conduct Variance Inflation Factor (VIF) analysis to detect multicollinearity among independent variables, which can compromise statistical inferences. High VIF scores indicate redundant features that may need to be removed during feature selection and edge cases where VIF scores approach infinity often correspond to perfectly anti-correlated dummy variables, which should be discounted. (Allison, 2012)

The results from Figure 4.7b show that certain features exhibit very high VIF scores, suggesting significant multicollinearity that could undermine the reliability of our regression coefficients. These features may need to be removed or transformed during feature selection to improve the model's statistical reliability and interpretability. Brownlee (2020)

Principal Component Analysis (PCA)

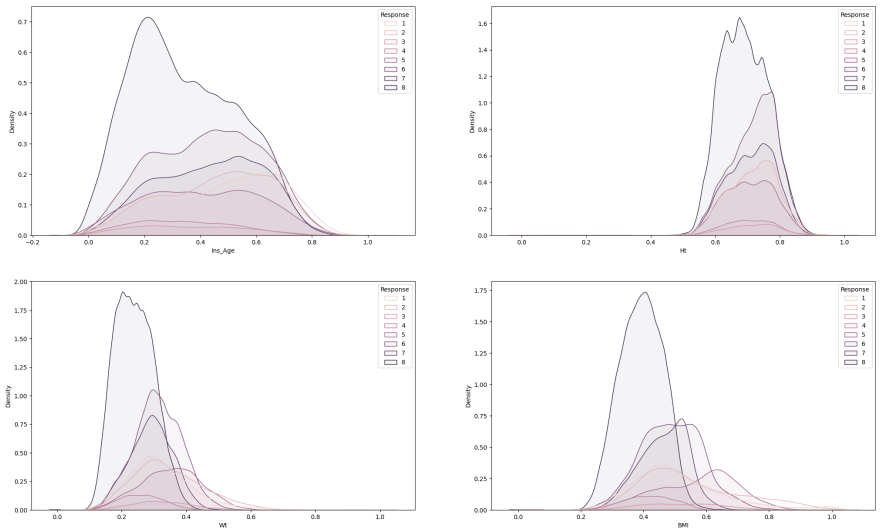
Here, we conducted Principal Component Analysis (PCA) to identify the primary sources of variation in our dataset. The analysis indicated in Fig. 4.8 that the first 40 principal components (PCs) account for over 80% of the cumulative variance in the validation dataset. This finding suggests that a lower-dimensional feature space can effectively capture the majority of the dataset's variance, potentially allowing for dimensionality reduction without significant loss of information during feature selection.

Lasso (L1) Regularisation

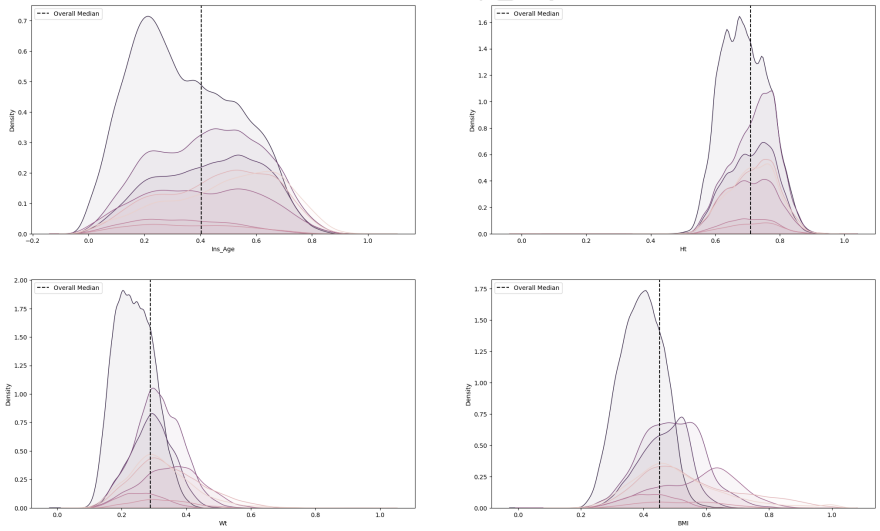
Here we performed feature selection via L1 (lasso) regularisation using the `sklearn` Python library, reducing our dataset from 128 features down to 57 features.

Conclusion

This is a summary of the Data Preparation section...

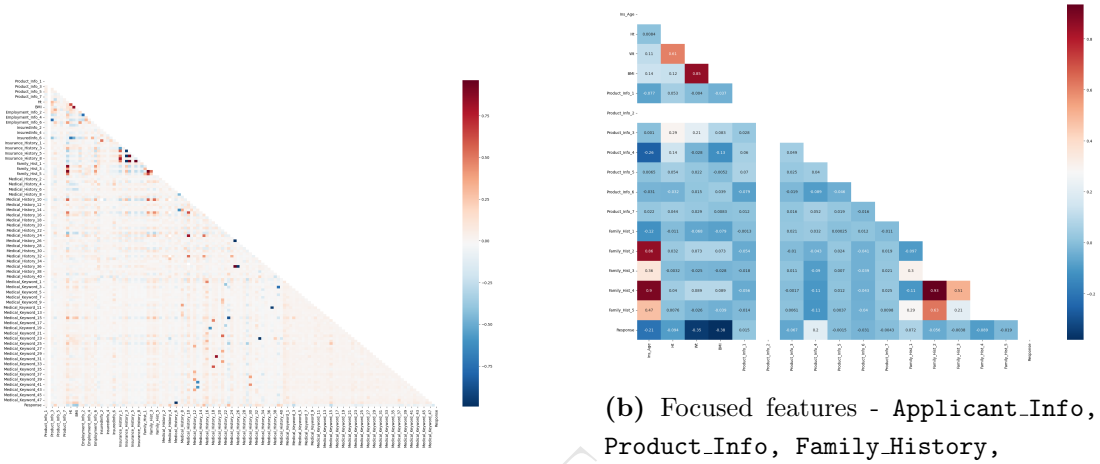


(a) Biophysical density plot per risk category



(b) Biophysical density per risk category plot - Overall median

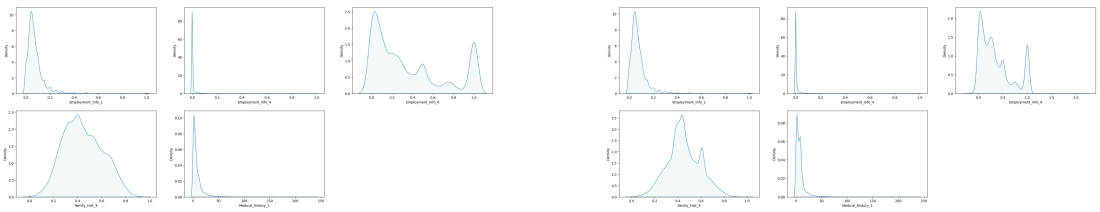
Figure 4.3: Combined diagram of Biophysical density plot and Risk category density plot with overall median



(a) All features

(b) Focused features - `Applicant_Info`, `Product_Info`, `Family_History`, `Response`.

Figure 4.4: Feature correlation heatmaps



(a) Original data

(b) After missing value replacement

Figure 4.5: Feature density plot

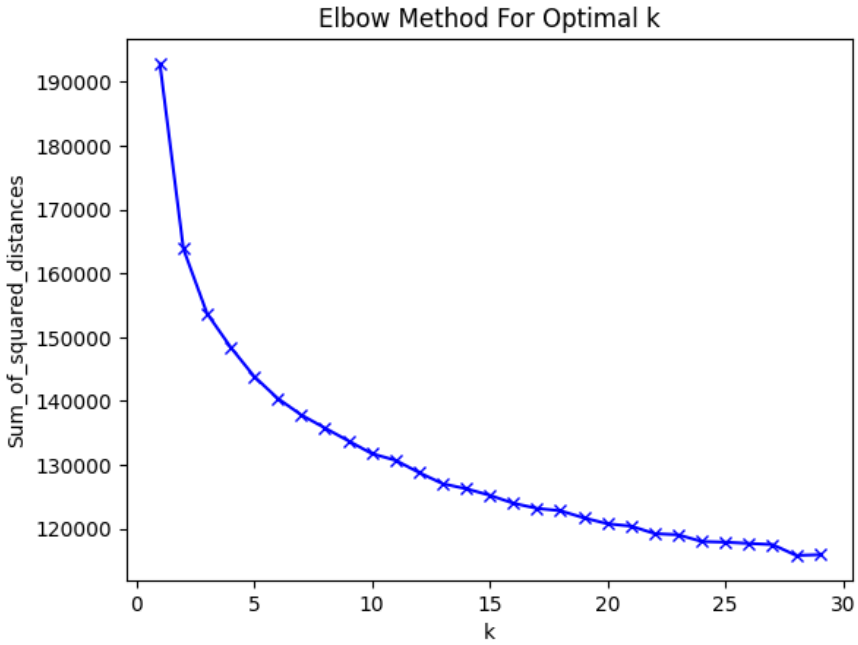
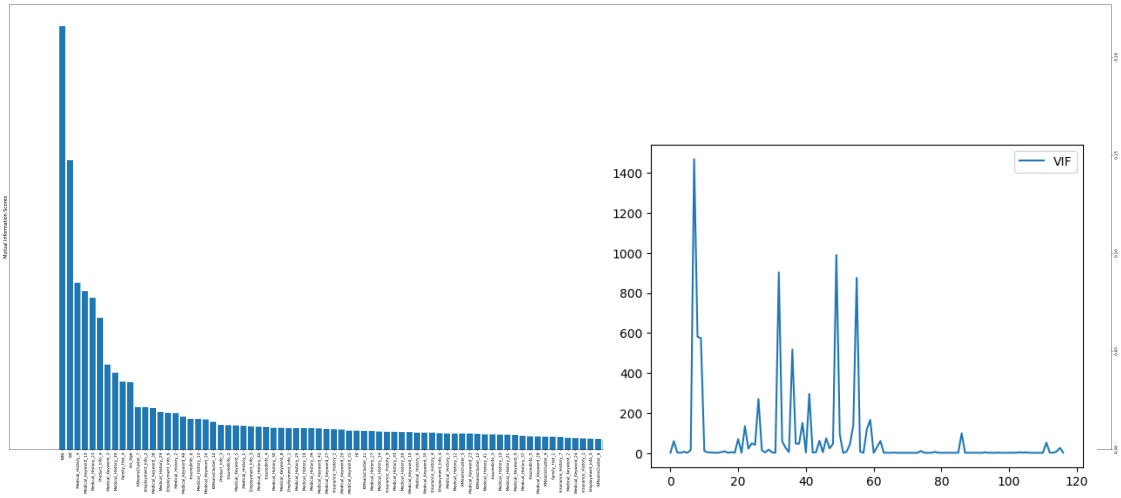


Figure 4.6: Elbow method plot

Elbow method plot illustrating the within-cluster sum of squares (WCSS) against the number of clusters (k), identifying k=15 as the optimal number of clusters where the rate of decrease in WCSS significantly diminishes.



(a) Mutual Information **(b)** Multicolinearity

Figure 4.7: Feature density plot

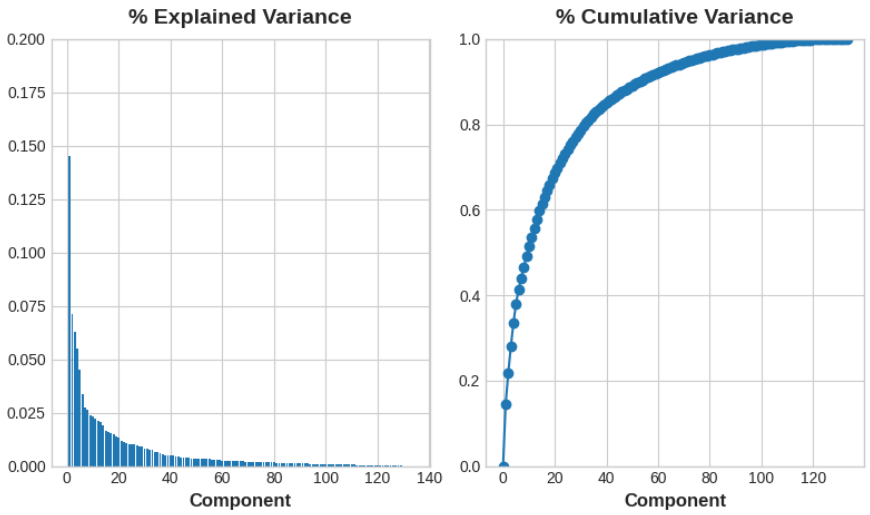


Figure 4.8: PCA Component Plots

4.4 Supervised Methods

In this section, we outline the key steps in training models using supervised methods and report on the results. A more detailed description of the algorithms used is described in @TODO section and the associated code executing the methods is contained in Section 4.1 Define Models of the Notebook.

4.4.1 Optimization (Hyperparameters)

We employed exhaustive grid-search methods to optimise several models' hyperparameters. For `Logistic Regression`, the hyperparameters include `tol`, which sets the tolerance for stopping criteria, and `C`, indicating the inverse regularisation strength. In `GaussianNB`, `var_smoothing` adjusts the Gaussian distribution's smoothness. For `SVC` with different kernels (`linear`, `poly`, `rbf`, and `sigmoid`), the `C` parameter and `tol` are optimised, with the `degree` parameter added for the polynomial kernel. `LinearSVC` similarly optimizes `C` and `tol`.

The `DecisionTreeClassifier` involves tuning `max_depth` and `max_features`, while the `RandomForestClassifier` adjusts `n_estimators`, `max_depth`, and `max_features`. `AdaBoostClassifier` optimizes `n_estimators` and `learning_rate`, with the latter influencing the contribution of each decision tree. For `GradientBoostingClassifier` and `XGBClassifier`, `learning_rate` and `n_estimators` are tuned, controlling the weight and number of boosting rounds.

Details on how these methods were applied are in the accompanying Notebook.

4.4.2 Calibrate Probabilities

The models currently at this point can provide predictions and initial probability estimates; however, further calibration is required to account for the expected distributions of the predicted classes. We performed probability calibration for all models to ensure that they provide accurate likelihoods of class membership. (Pedregosa et al., 2011)

4.4.3 Performance Results

Here, we compare the results of testing each of the machine learning algorithms, solely based in their predictive performances using a number of metrics to display how each model has performed in terms of classifying applicants into each `Response` group.

As illustrated in Figs. 4.9 and 4.10, each model exhibits a diverse range of AUC values across different classes, reflecting varying sensitivity and specificity within the dataset. Notably, the gradient boosting classifiers achieved the highest average AUC values, with Model 11 attaining a macro-average AUC of 0.84 and Model 12 a macro-average AUC of 0.83. Both models demonstrated better performance in predicting applicants for classes 3, 4, and 8, as evidenced by their high AUC scores (0.9) on the ROC curves. In contrast, their performance was slightly lower for the other classes.

The confusion matrices in Figs. 4.11 and 4.12 visualise the mapping of each model's predictions to the true labels in the test dataset. These plots illustrate whether a model is overfitted to specific classes or is sufficiently generalised to classify applicants effectively.

Generally, most models appear well-fitted, replicating the distribution of response values reasonably well. Many misclassifications occur between adjacent classes (e.g., predicting class 8 when the actual class is 7). However, some models show significant overfitting to class 8, which has the highest representation in the dataset. Model 7 (SVC with sigmoid kernel) illustrates this, as it mainly predicts class 8. Models 3 and 5 (SVCs with linear and polynomial kernels) also show this trend, though to a lesser extent.

In contrast, Models 11 and 12 (Gradient Boosting Classifier and XGBoost Classifier) generalise well, closely replicating the distribution of response values. However, these models tend to misclassify some low-risk applicants (classes 1-2) as high-risk (classes 6-8). In a business context, this could lead to unnecessary examination of low-risk applicants, increasing operational effort and time, but it would mitigate the risk of underestimating high-risk applicants.

4.4.3.1 Results Evaluation

From Table 4.2, the Gradient Boosting and XGBoost classifiers stand out as the most effective models, providing high accuracy and balanced performance across precision, recall, and F1-scores. Random Forest and Decision Tree classifiers also show strong performance, making them viable alternatives. On the other hand, models like Gaussian Naive Bayes and SVC with a sigmoid kernel demonstrate limitations. This demonstrates the need to evaluate and select the appropriate model based on the characteristics of the dataset.

Table 4.2: Performance Comparison of Models Across Key Metrics

Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
Logistic Regression	0.49	0.41	0.30	0.30	0.45	0.49	0.44
Gaussian Naive Bayes	0.43	0.27	0.24	0.23	0.36	0.43	0.37
SVC (Linear Kernel)	0.40	0.30	0.21	0.21	0.34	0.40	0.32
Linear SVC	0.48	0.44	0.30	0.30	0.46	0.48	0.44
SVC (Polynomial Kernel)	0.39	0.34	0.22	0.20	0.34	0.39	0.30
SVC (RBF Kernel)	0.41	0.30	0.24	0.23	0.35	0.41	0.34
SVC (Sigmoid Kernel)	0.33	0.40	0.13	0.07	0.37	0.33	0.18
Decision Tree Classifier	0.52	0.42	0.34	0.34	0.49	0.52	0.48
Random Forest Classifier	0.52	0.45	0.35	0.37	0.50	0.52	0.49
AdaBoost Classifier	0.47	0.36	0.27	0.26	0.44	0.47	0.39
Gradient Boosting Classifier	0.53	0.42	0.34	0.34	0.51	0.53	0.49
XGBoost Classifier	0.53	0.43	0.35	0.36	0.50	0.53	0.49

4.4.4 Explainable AI (XAI)

In this section, we analyse for the top-performing supervised model, `XGBoost`, from the perspective of explainability. We apply post-hoc techniques to illuminate the rationale behind the model's decision-making process.

4.4.4.1 Shapley Summary Plot (SHAP)

Fig. 4.13 is a SHAP summary plot (SHapley Additive exPlanations), which helps explain the relative contribution of features to the prediction outputs of our `XGBoost` model.

Reading the Plot

First we provide a brief background on reading the plot;

- **X-axis (SHAP value)** represents the impact of each feature on the model's output. Positive SHAP values indicate a positive contribution to the prediction, while negative SHAP values indicate a negative contribution.
- **Y-axis (Features)** lists the features along the Y-axis, ordered by their importance, based on the mean absolute SHAP value.
- **Colours** the colour gradient from blue to red indicates the feature value from low to high. Blue points represent low feature values, and red points represent high feature values.

Interpretation

- **Feature Importance**
  The features are ordered by their importance, with the most important feature at the top. In this plot, `Ins_Age` is the most important feature, followed by `Medical_History_23`, `Medical_History_13`, and so on.

- **Impact Direction**
  The spread of points along the X-axis shows how much impact a feature has on the prediction. For instance, `BMI` has a wide spread, indicating that it can significantly influence the model's output both positively and negatively.

- **Feature Value Impact**
  The colour of the points indicates how the value of the feature affects the prediction. For example, high values of `BMI` (in red) tend to push the prediction positively, while low values (in blue) push it negatively.

Observations

- `Ins_Age` feature has a significant impact on the model output, with both high and low values influencing the predictions in different directions.

- `Medical_History_23` and `Medical_History_13`: These features also have substantial impacts, with their high and low values affecting the model output differently.
- `BMI`: High `BMI` values (red) generally increase the model's prediction, while low `BMI` values (blue) decrease it.

Key Insights

- The plot helps understand how the model behaviour, for instance, showing age `Ins_Age` at the top of the list means overall that it is the highest contributory feature to prediction outcomes. In terms of its spread of values age also shows a wide spread of SHAP values, both positive and negative, indicating that it has a significant impact on predictions in both directions.
- From the plot we can interpret a tendency for increasing `Ins_Age` to have a more positive impact on the prediction. In general, higher ages (represented by redder dots) are more likely to contribute positively to the model output, while lower ages (bluer dots) are more likely to contribute negatively. However, this is not a strict rule, as we can see exceptions in both directions. The impact of age varies significantly across different instances, suggesting that age becomes an increasingly important factor in the prediction as it increases, but its effect is not uniform and may interact with other features.
- `Ins_Age` Both `Medical_History_23` and `Medical_History_13` are also significant features that influence the model predictions. Higher values of these features tend to have a negative impact on the prediction (blue dots), while lower values tend to have a positive impact (red dots). However, the impact of these features is not uniform, suggesting that their effects may interact with other features in the dataset.
- The SHAP summary plot for BMI reveals that higher values of BMI tend to have a positive impact on the model's predictions, while lower values tend to have a negative impact, however again there is variability here, indicating that its effect may interact with other features in the dataset.

4.4.4.2 Conclusion

SHAP provides a way to interpret complex machine learning models by assigning importance values to each feature, and the SHAP summary plot Fig. 4.13 provides a detailed view of how each feature influences the model predictions. This helps in understanding which features are most important and how their values impact the model output. SHAP values do not inherently tell us anything about the correctness of model's predictions, but only show how different features contribute to the model's prediction. Further, SHAP provides insights into correlations rather than causal relationships. In risk-sensitive applications such as insurance, for example, legal and regulatory contexts, where understanding causality can be critical, SHAP may not provide the required depth of explanation. (Huang and Marques-Silva, 2023)

4.4.4.3 Shapley Dependency Plots

SHAP dependency plots illustrate how a feature's value influences model predictions by plotting feature values on the x-axis against SHAP values on the y-axis, which represent the feature's impact. Each point corresponds to an individual data instance, and colour is often used to show the value of another feature, highlighting interaction effects. The overall shape of the plot reveals the relationship between the feature value and its impact, while vertical dispersion indicates interactions with other features.

Dependency Plot (Ins_Age vs BMI)

In Fig. 4.14a, we generated a SHAP dependence contribution plot to analyze the influence of BMI as a function of Ins_Age on the model's predictions.

We can observe a wide color gradient spread throughout the plot, indicating that the model considers a wide range of BMI values when making predictions, with no clear clustering of BMI values at specific Ins_Age levels. This suggests that BMI's effect is distributed across all ages.

At higher Ins_Age values (e.g., above 0.6 on the Ins_Age scale), the SHAP values are predominantly positive (red dots), indicating BMI having a more positive influence on the prediction (risk classification). For younger ages (e.g., below

0.3), the SHAP values are generally lower or around zero, indicating a neutral or slightly negative impact on the prediction.

Dependency Plot (BMI vs Medical_History_16)

In Fig. 4.14b, we generated a SHAP dependence contribution plot to analyze the influence of BMI as a function of `Medical_History_16` on the model's predictions.

The horizontal axis represents BMI values from the dataset, while the vertical axis shows the SHAP values for BMI.

Positive SHAP values indicate that the corresponding BMI values increase the likelihood of being classified into a certain risk category, whereas negative SHAP values suggest a decrease. A SHAP value near zero indicates that BMI has little to no impact on the classification.

The colour gradient represents the values of `Medical_History_16`, ranging from blue (low values) to red (high values). This additional dimension provides context on how `Medical_History_16` interacts with BMI to influence risk classification.

Key Observations

- For very low BMI values (below 0.3), the SHAP values are highly variable, indicating that the impact of low BMI on risk classification is influenced by `Medical_History_16`.

- For mid-range BMI values (around 0.5), the SHAP values are generally close to zero, suggesting that BMI has a neutral effect on risk classification in this range.

- For very high BMI values (above 0.8), the SHAP values are predominantly positive, indicating that high BMI increases the likelihood of certain risk classifications. The spread of colors in this range suggests that `Medical_History_16` significantly interacts with BMI to modify the risk classification.

Dependency Plot (Medical_History_23 vs Medical_History_13)

In Fig. 4.14c, we generate a SHAP dependence contribution plot to analyze the influence of `Medical_History_23` on the model's risk classification predictions, with `Medical_History_13` as the interaction feature.

The horizontal axis represents the values of `Medical_History_23`, while the vertical axis displays the SHAP values for `Medical_History_23`. Positive SHAP values indicate that the corresponding `Medical_History_23` values increase the likelihood of classification into a certain risk category, while negative SHAP values suggest a decrease. SHAP values near zero indicate minimal impact.

The colour gradient represents `Medical_History_13` values, from blue (low values) to red (high values), providing context on how it interacts with `Medical_History_23`.

Key observations include:

- `Binary Nature of Medical_History_23:` The feature is binary, taking values of either 0 or 1.

- `Medical_History_23 = 0:` SHAP values are widely spread around zero, indicating that `Medical_History_23` has minimal impact on the risk classification. The influence of `Medical_History_13` is also seen, as represented by the color gradient.

- `Medical_History_23 = 1:` SHAP values show a fairly wide range, suggesting a significant and variable influence on the risk classification. The colour spread indicates that `Medical_History_13` also plays a crucial role in modifying this impact.

4.4.4.4 Discussion - SHAP Plots in High-Risk Applications

While SHAP (SHapley Additive exPlanations) dependence plots for individual features and interaction plots between features enhance the transparency of risk models, several critical limitations persist, particularly in high-stakes domains:

Illusion of Comprehensiveness

Feature Interactions and Non-Linearities Interaction plots may reveal pairwise relationships, but higher-order interactions involving multiple features remain hidden leading to an incomplete understanding of how complex combinations of factors truly drive risk the prediction. The plots may overlook non-linear relationships if the relationship is more complex, creating a false sense of security in the model's behavior.

Interpretability Challenges

Expert Dependency While SHAP plots offer more visual clarity than raw model outputs, their interpretation still requires expertise in both the domain and the underlying model limiting their accessibility to non-technical stakeholders.

Regulatory Compliance

Narrative Gap Visual interaction plots might not provide the narrative explanation that regulators require. They may prefer or require a clear, logical storyline connecting individual feature effects to the overall risk assessment, something that SHAP plots may find difficult to provide.

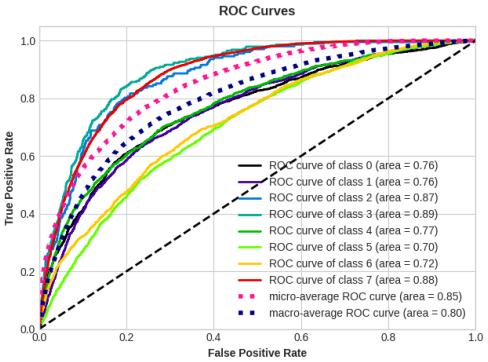
Generalisation Risk SHAP dependence and interaction plots are derived from the training data distribution. When faced with unseen that is out-of-distribution, their explanatory power diminishes, potentially leading to inaccurate desciptions of a risk classification.

SHAP values for interactions mcan be computationally expensive, particularly for high-dimensional models or large datasets. This can limit their real-time applicability in scenarios where rapid decision-making is crucial.

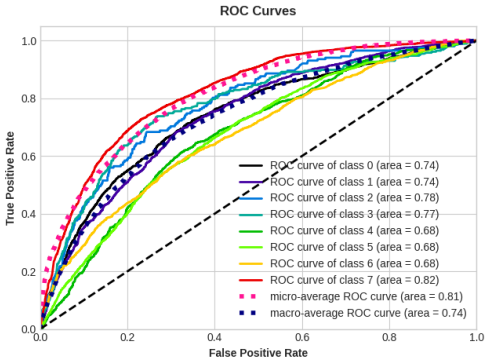
Relying solely on SHAP plots, even with interaction visualizations, might not be sufficient for a comprehensive risk assessment. Integrating them with other explainability techniques, such as counterfactual explanations or causal models, can provide a more complete picture.

In Conclusion

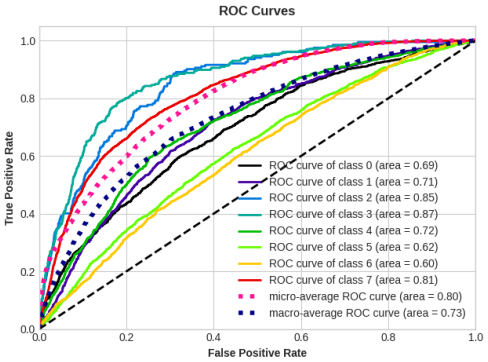
While the availability of SHAP dependence and interaction plots is a step towards greater transparency, they are not a silver bullet for the challenges of explainability in high-risk applications. To build truly trustworthy and responsible AI systems, the authors contend that we need to explore inherently interpretable models through unsupervised method to better address potential biases, and ensures that explanations are both technically rigorous and accessible to a wider range of stakeholders. This will be the focus of the next section.



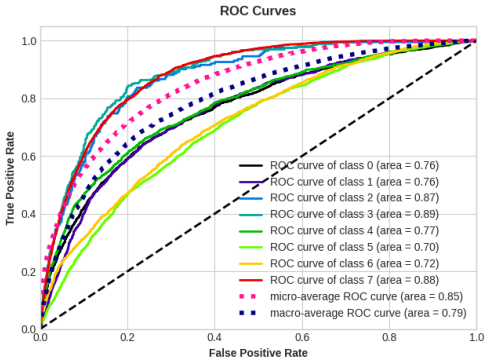
(a) Logistic Regression



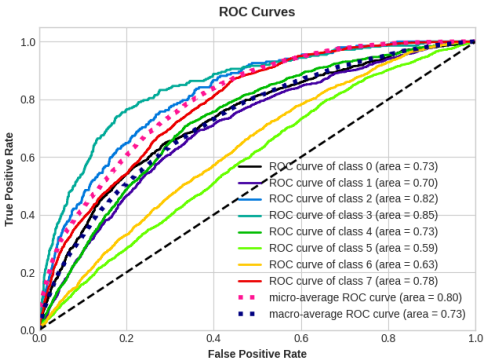
(b) Gaussian Naive Bayes



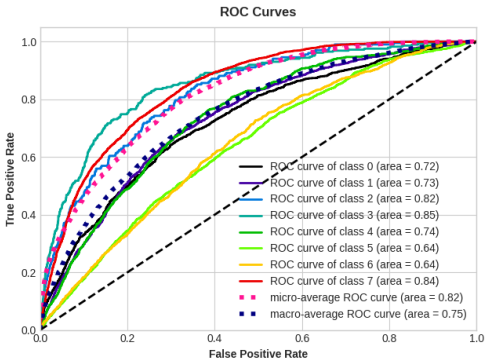
(c) SVC (Linear Kernel)



(d) Linear SVC

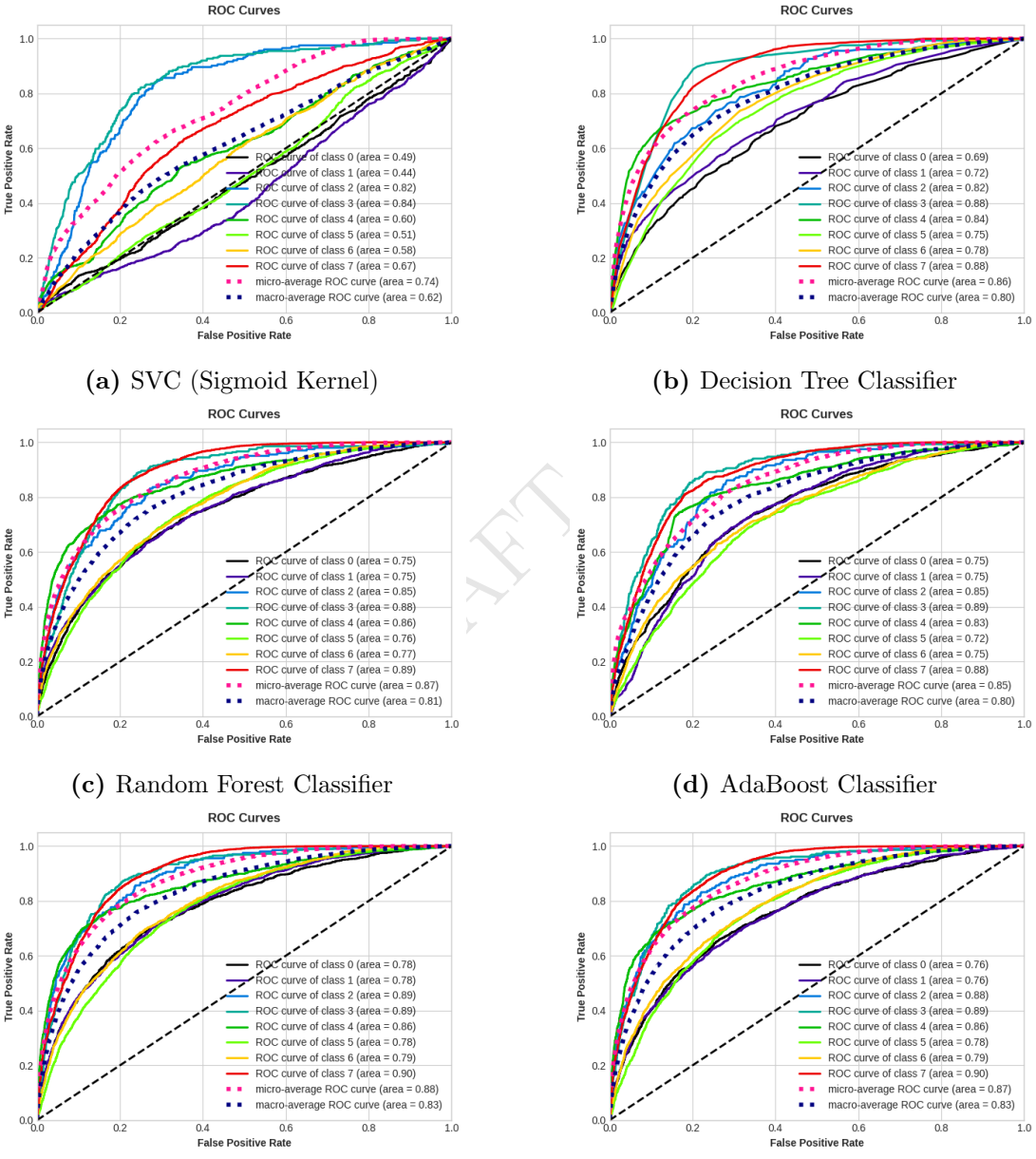


(e) SVC (Polynomial Kernel)



(f) SVC (RBF Kernel)

Figure 4.9: ROC Plots for supervised techniques (1/2)



(a) SVC (Sigmoid Kernel)

(b) Decision Tree Classifier

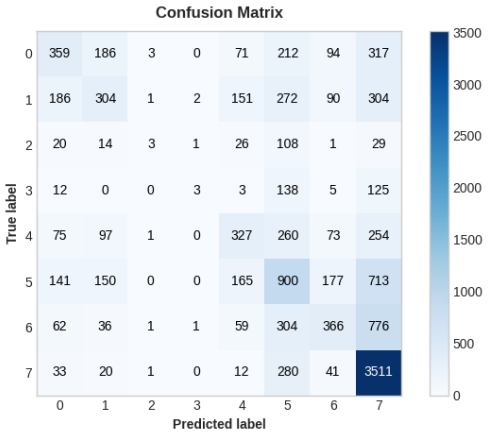
(c) Random Forest Classifier

(d) AdaBoost Classifier

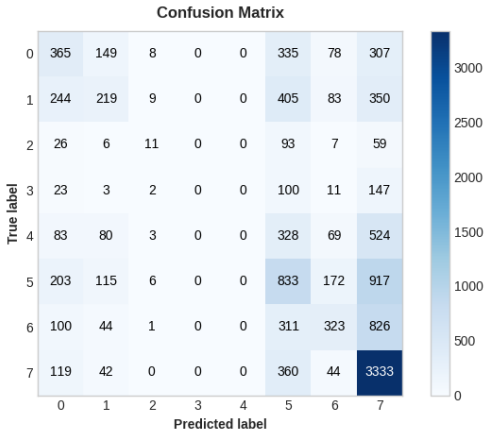
(e) Gradient Boosting Classifier

(f) XGBoost Classifier

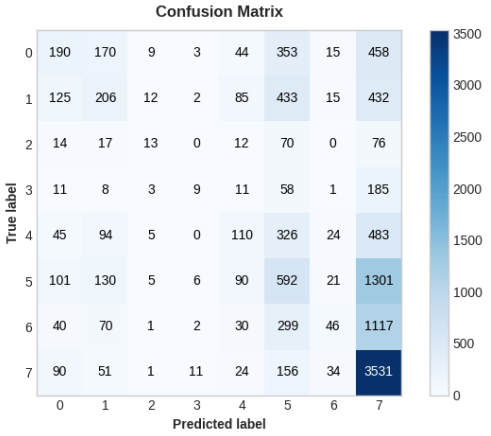
Figure 4.10: ROC Plots for supervised techniques (2/2)



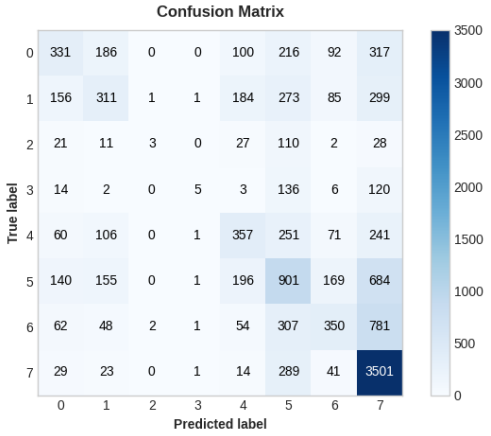
(a) Logistic Regression



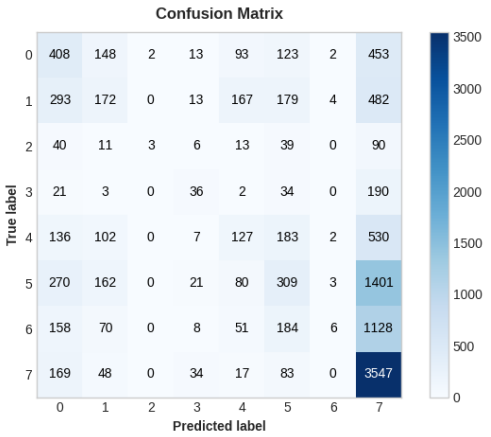
(b) Gaussian Naive Bayes



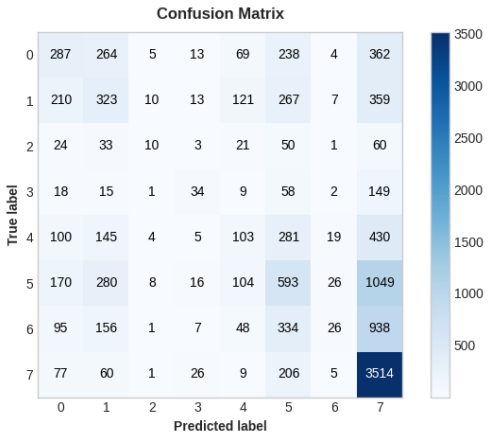
(c) SVC (Linear Kernel)



(d) Linear SVC

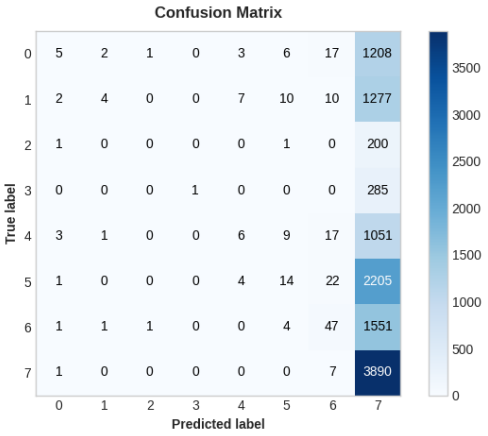


(e) SVC (Polynomial Kernel)

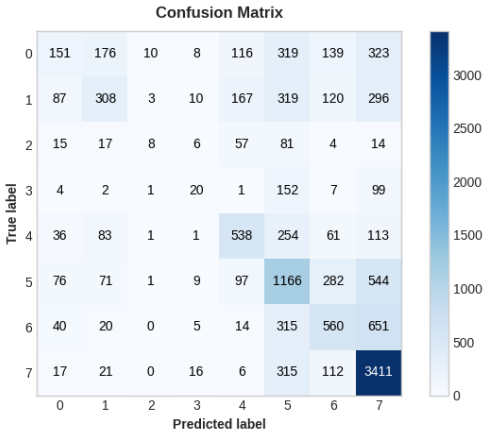


(f) SVC (RBF Kernel)

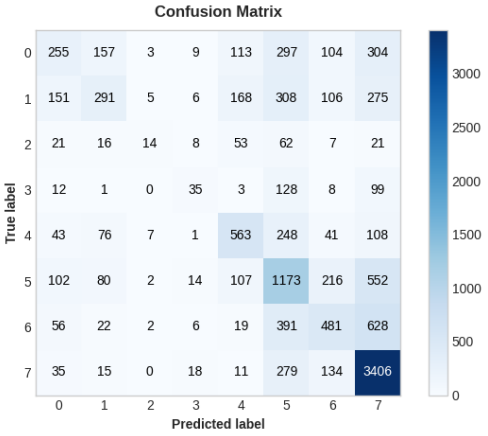
Figure 4.11: Confusion Matrices for supervised techniques (1/2)



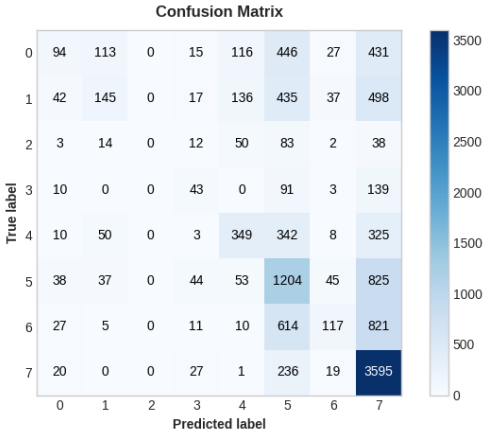
(a) SVC (Sigmoid Kernel)



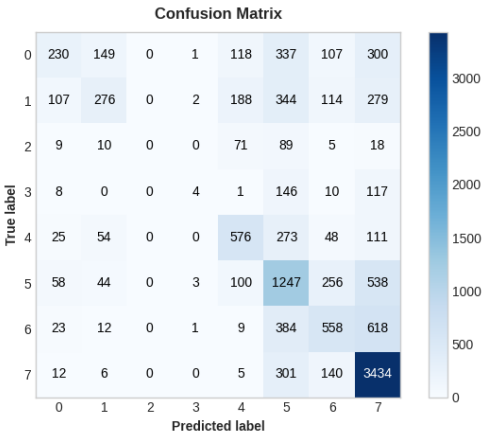
(b) Decision Tree Classifier



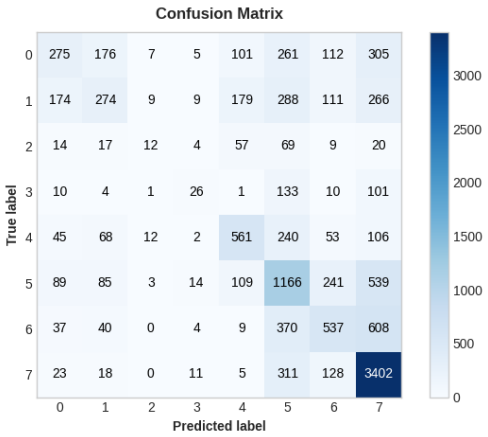
(c) Random Forest Classifier



(d) AdaBoost Classifier



(e) Gradient Boosting Classifier



(f) XGBoost Classifier

Figure 4.12: Confusion Matrices for supervised techniques (2/2)

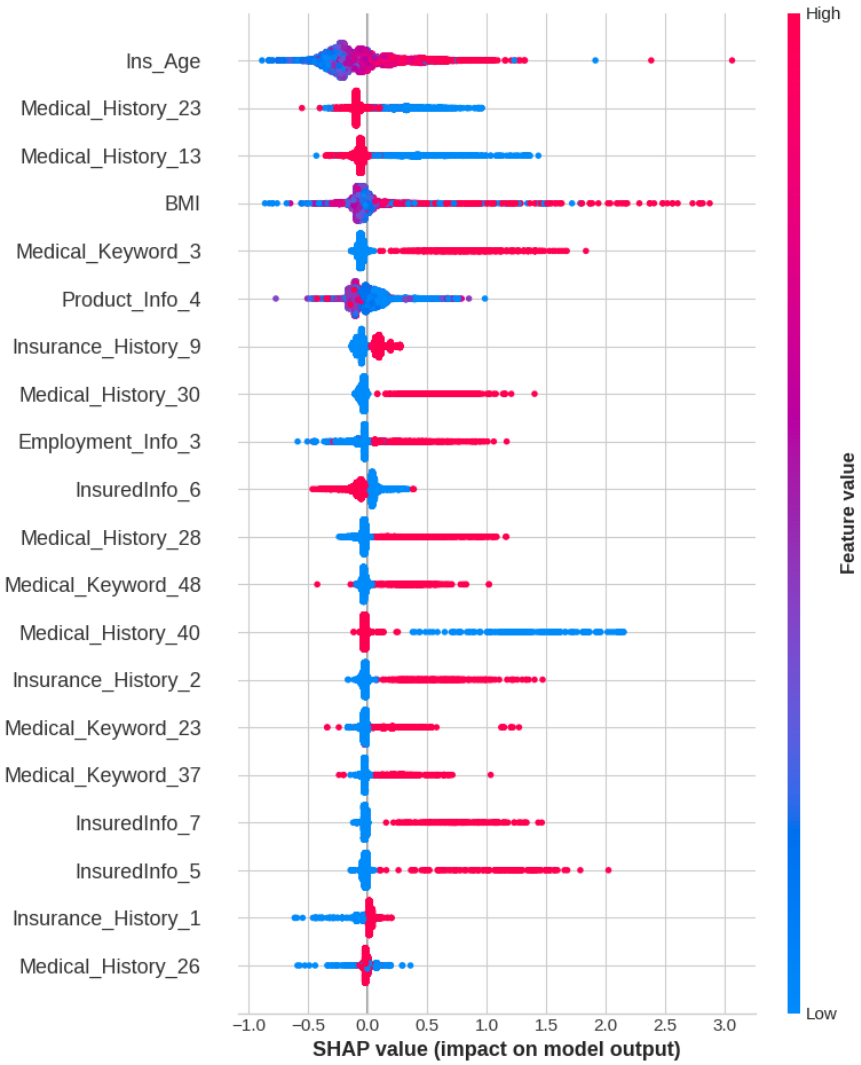
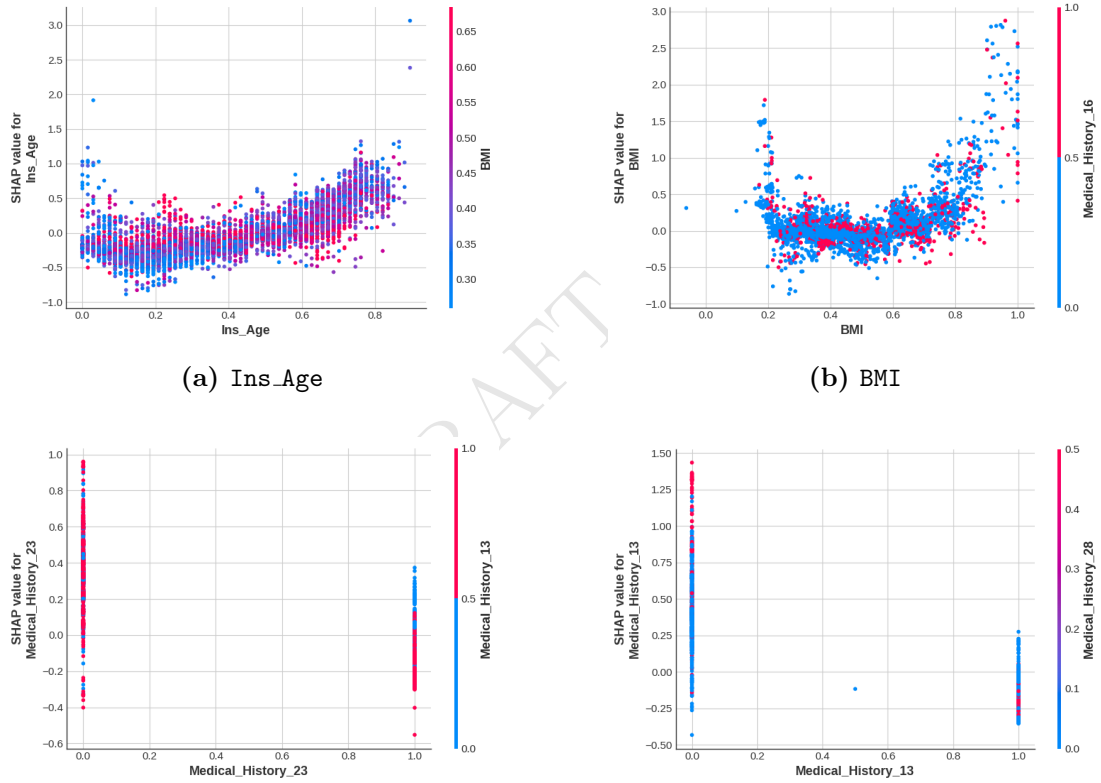


Figure 4.13: XGBoost - Shapley plot



(a) `Ins_Age` **(b)** `BMI`

(c) `Medical_History_13` **(d)** `Medical_History_12`

Figure 4.14: XGBoost - SHAP feature dependency contribution plots

4.5 Unsupervised Methods

In the previous Data Preparation section, we addressed missing values and conducted feature engineering, reducing the original Prudential dataset from 128 features to 50. This prepared the training data of 32,628 samples for various supervised techniques.

In this section, the methodology follows that described in McInnes et al. (2018) based on the underlying mathematics described by Sainburg et al. (2021).

First, we present scatter plots of 10 feature dimensions, with colours representing risk classifications. Although this visualisation provides an initial impression, it is clearly inadequate for comprehensively analysing the data.

Previously, in the Data preparation section, we preprocessed for missing values before feature engineering to reduce the original Prudential dataset from 128 down to 54 features. We used these data to train using a series of models using supervised techniques.

In Fig. 4.15 we created a scatter plot of the 10 feature dimensions, with colours corresponding to the risk classifications. Although perhaps allowing an initial impression, this approach is clearly insufficient for these data.

4.5.1 UMAP

Initial Visualisation

In Fig. 4.16, for initial visualisation, we reduce the data to 2-dimensions using UMAP. Ultimately, we will aim to cluster the data in high-dimensional space and then visualise the result of that clustering. However, first, to help frame what follows, we simply view the data coloured by the risk category that each data point represents using a different colour for each classification. We also include a version of the same plot with 1000 samples.

k-Means clustering

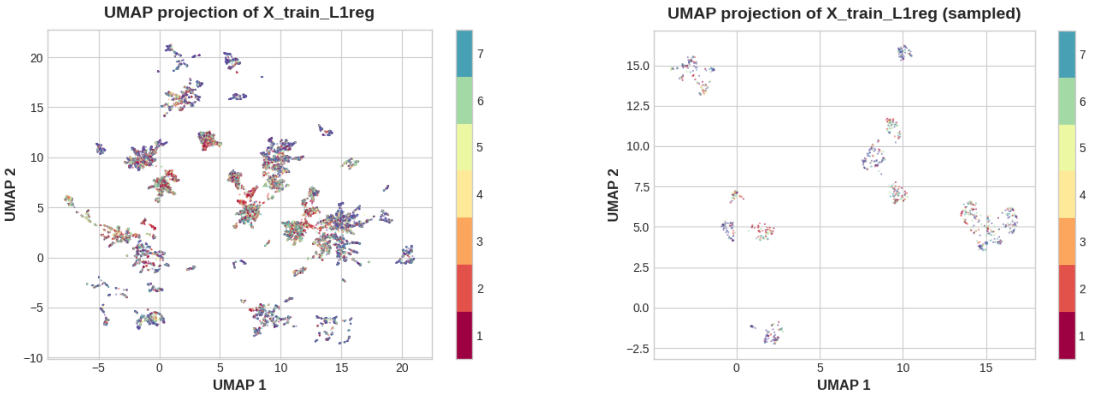
Fig. 4.17 shows the results of clustering the data using k-Means choosing k, the number of clusters to match the number of risk categories. Thus, we initialise

Figure 4.15: Pair plots (first 10 dims.), colours based on `Response` value

with the actual number of clusters we are looking for.

The results show that the clusters are not tightly coalesced, there appears to be potential as the colours are generally associated with identifiable clusters in many cases. Multiple clusters are evident, a few dense or moderately dense, and clusters are well-separated. Fig. 4.17b shows the same k-Means drawn from 1000 samples and indeed shows the same. However, the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) scores suggest little agreement between the clustering results the algorithm produced and the true labels.

This suggests that k-means clustering does not perform well in identifying the true underlying structure of the data. This is likely to be partially due to the centroid-based nature of K-Means with the assumption of spherical clusters



(a) Full dataset **(b)** 1000 samples

Figure 4.16: Prudential dataset (57 features) reduced to 2-dimensions with UMAP.

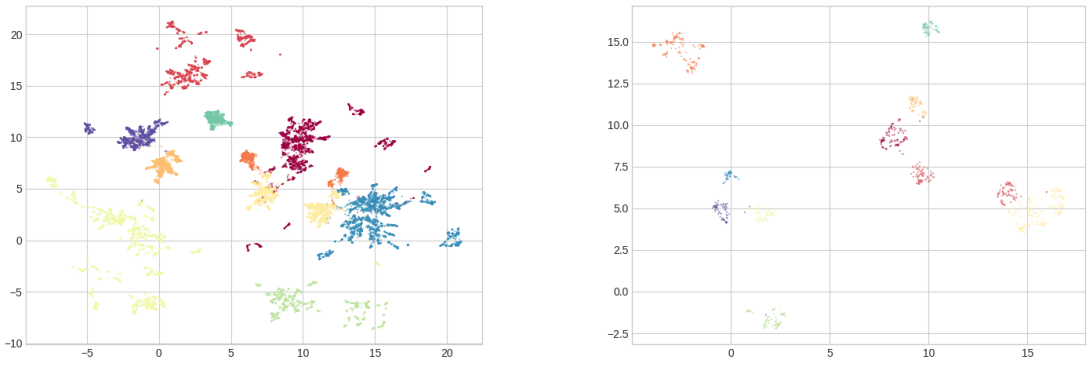
this is responsible for some of the sharp divides that K-Means puts across digit classes. Therefore, we need to investigate more advanced clustering methods and to improve on this with a density-based algorithm - Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) being one of the most advanced of these and which we will try next. (McInnes et al., 2018)

HDBSCAN

Before proceeding with HDBSCAN, for performance purposes, it can often be useful to reduce the dimensionality to around 50 or less, however, since we are already around this level through feature engineering alone we can omit this step.

An aspect of HDBSCAN is its ability to reject clustering certain points, labeling them as 'noise'. To visualise this aspect, points identified as noise are coloured grey, while the rest are coloured based on their cluster assignment. The results in Fig. 4.18 contain some grey dots particularly in the sampled plot; however, most of the plot in the full training dataset appears to have been assigned to clusters with the eight colours visible in clusters even if these are not very strongly coalesced.

The results show that similar to k-Means, the ARI and AMI scores are close



(a) Full dataset **(b)** 1000 samples

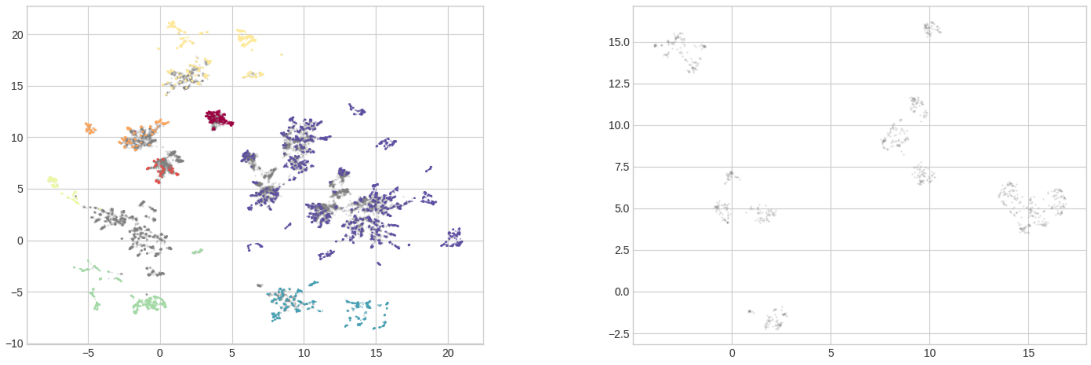
Figure 4.17: k-Means with clusters = 8.

Prudential dataset (57 dimensions), number of clusters set to match number of risk categories

to zero even when the cluster representing noise is omitted, indicating the clusters are a poor match for the ground truth clusters. However, 53% of the data were successfully clustered - so it could cluster over half of the data, however, unsuccessfully.

The challenge here is that HDBSCAN, as a density-based clustering algorithm, struggles with the curse of dimensionality; high-dimensional data require more samples to effectively represent density. By further reducing the dimensionality, we could enhance density visibility and facilitate more effective clustering by HDBSCAN. However, applying PCA for this reduction poses issues. Although PCA can reduce the 50 dimensions while preserving much of the data's variance, further reduction would significantly degrade performance due to PCA's linear nature. Therefore, more robust manifold learning is required according to Sainburg et al. (2021), which is where UMAP becomes beneficial.

Fig. 4.18 ..



(a) Full dataset **(b)** 1000 samples

Figure 4.18: HDBSCAN

Prudential dataset (57 dimensions)

UMAP

This section reports on the results of applying the combination of Uniform Manifold Approximation and Projection (UMAP) introduced by Sainburg et al. (2021) for dimension reduction followed by HDBSCAN for clustering.

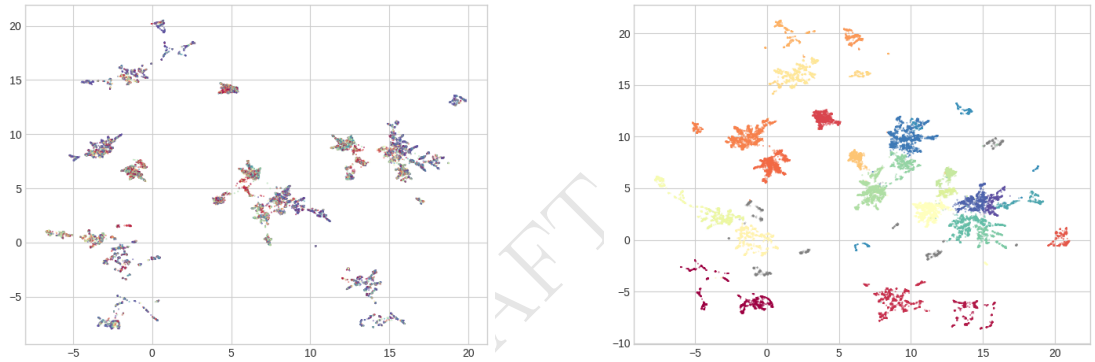
UMAP enhanced clustering

Our objective is to utilize UMAP to execute non-linear manifold-aware dimensionality reduction, enabling us to reduce the dataset to a number of dimensions low enough for a density-based clustering algorithm to function effectively. A key benefit of using UMAP is that it allows reduction to dimensions beyond just two – for instance, we can reduce to say 10 dimensions since the aim is clustering rather than visualisation, with UMAP imposing minimal performance overhead. (McInnes et al., 2018)

The challenge we have here is the complexity of the Prudential dataset and our requirement to ideally be able to reduce to two or three dimensions for purposes of visualisation and interaction by users (underwriters).

Here, we comment on setting parameters on the UMAP algorithm. When applying UMAP for dimensionality reduction, we need to choose different pa-

rameters than when using it for visualisation. Firstly, a higher `n_neighbors` value is advisable since smaller values will emphasise very local structures and are more likely to generate detailed cluster patterns that may be influenced by data noise rather than genuine clusters. For this purpose, we will double the default value from 15 to 30. Furthermore, it is favourable to set `min_dist` to a very low value. Because we aim to densely pack points together, a low value will facilitate this and also create clearer separations between clusters. Hence, we will set `min_dist` to 0. (McInnes et al., 2018)



(a) Clustering optimised **(b)** Visualization optimised

Figure 4.19: UMAP reduction to 2 dimensions. Parameterized to optimise for clustering / visualisation.

4.5.1.1 UMAP enhanced clustering - 2d visualisation

As can be seen from Fig. 4.19a we still have the general global structure, but points are slightly more densely packed together within clusters, and consequently we can see larger gaps between the clusters. However, this embedding was for clustering purposes only, so next we revert to the original embedding for visualisation purposes.

As can be observed from Fig. 4.19b we can see much clearer identification of clusters rather than just assigning the majority of data as noise. This results from the fact that we no longer have to attempt to handle the relative lack of

density in 50 dimensional space and now HDBSCAN can more cleanly discern the clusters.

However, ARI and AMI scores are still close to zero indicating the clusters are still a poor match for the ground truth clusters.

4.5.1.2 UMAP enhanced clustering - 3d visualisation

Here we repeated the same process this time reducing the dimensionality to 3 dimensions with UMAP for 3-D visualisation. We then export the embeddings for visualisation in TensorBoard.

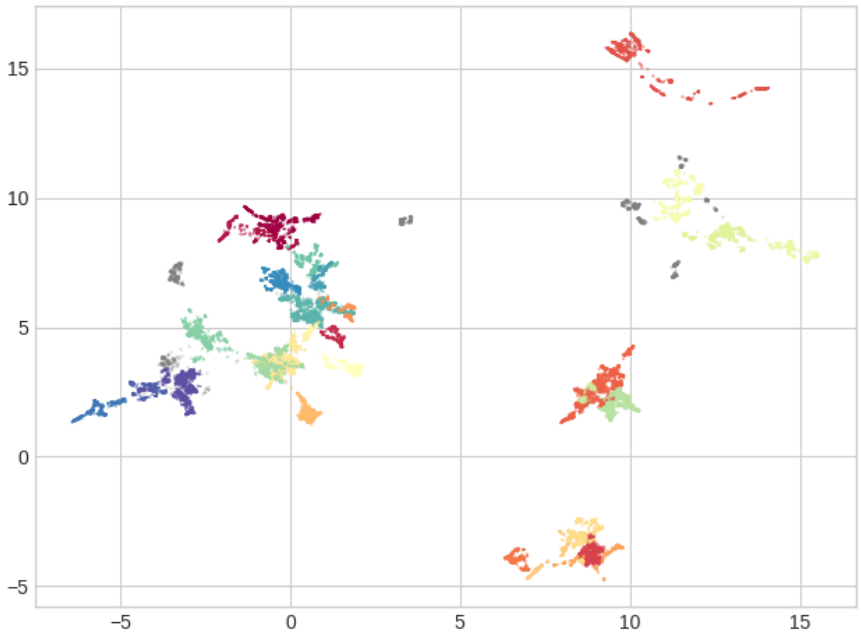


Figure 4.20: UMAP and HDBSCAN - 3d

Figs. 4.21 and 4.22 show screenshots taken from the 3-dimensional embeddings created from UMAP and HDBSCAN, exported to the Tensorflow Tensorboard interactive visualisation toolkit. (Abadi et al., 2015)

4.21 shows transformation from original data with UMAP applied to reduce dimensionality to 3 for visualization then applying HDBSCAN for clustering. Fig. 4.21d shows a snapshot of the Tensorboard interactive 3-d model model with the coloured labels indicating moderately dense and well-separated clusters.

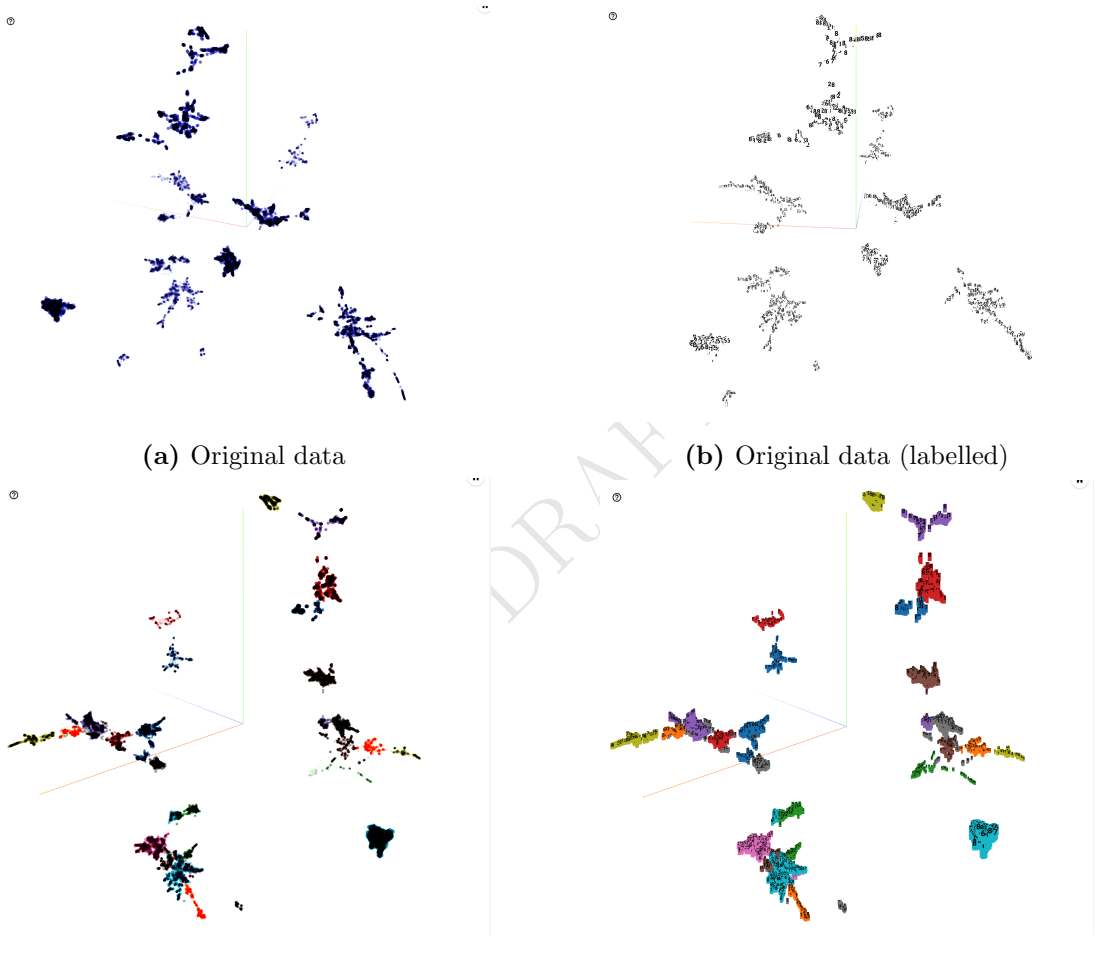
Fig. 4.21d and Fig. 4.22a show the same clusters with colours reversed as a visual aid.

Fig. 4.22b provides a snapshot of a magnified view to represent the fact that the visual model in TensorBoard is interactive, its axis can be rotated, and the viewer can zoom and pan to move to any point in the 3-d space to observe the clusters and their relationships.

Fig. 4.22c provides a snapshot of TensorBoard interactive mode where it is possible to interactively select any group of data points (e.g., cluster) and then the classifications can be updated in the tool. For example, an underwriter might identify a cluster of data points and select interactively to isolate these, then analyse further information about the associated cases in more detail, for example, checking descriptive statistics, running checks for bias and fairness and manually checking some samples. At that point, the underwriter may take the decision to set the classification label on all points, perhaps confirming what might have been the majority. Samples can also be identified by the underwriter as outliers and omitted from this labelling. In this way, the underwriter works to use the information provided by the clustering and confirm that these are true clusters so that all samples within the cluster (other than outliers) are assigned to this confirmed label. As such, the underwriter combines the identification of the cluster with other analytical and statistical methods to confirm the assignment of risk classification categories in the training data. These confirmed risk classifications will be assigned to retraining the model. The model can then be used to predict new unseen samples (life insurance applications).

4.5.1.3 Results

Overall, based on Table 4.3 none of the methods achieves a high agreement with the true class labels, indicating challenges in clustering performance for this data set. UMAP combined with HDBSCAN shows some promise, but further improvements and alternative approaches may be needed for better clustering results.



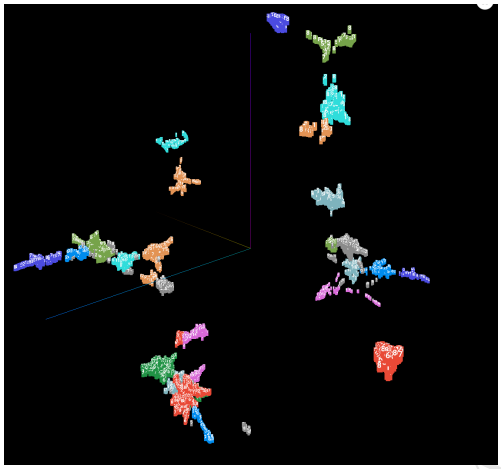
(a) Original data

(b) Original data (labelled)

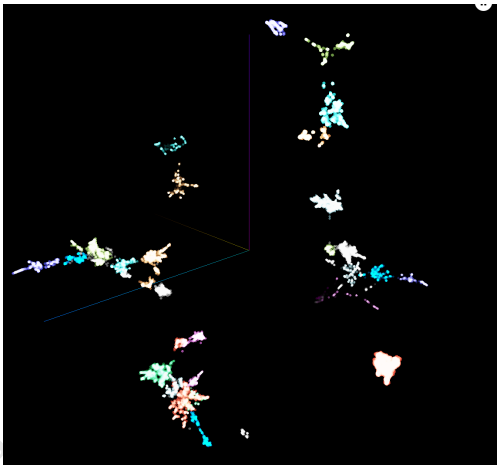
(c) UMAP then HDBSCAN

(d) UMAP then HDBSCAN (labelled)

Figure 4.21: Cluster identification with UMAP and HDBSCAN - Part 1

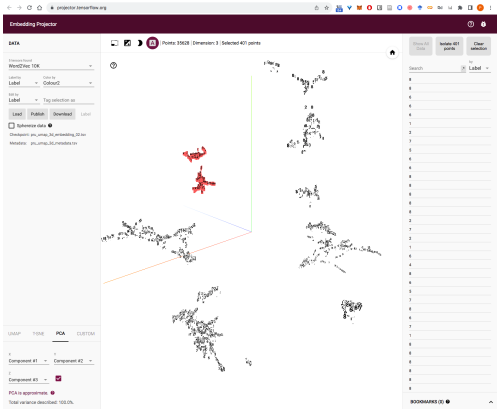


(a) UMAP then HDBSCAN (reversed)



(b) UMAP then HDBSCAN (reversed, labelled)

(c) UMAP then HDBSCAN (magnified)



(d) UMAP then HDBSCAN (manual label updates)

Figure 4.22: Cluster identification with UMAP and HDBSCAN - Part 2

Clustering Method	Adjusted Rand Index (ARI)	Adjusted Mutual Information (AMI) Score
k-Means	0.0336	0.0511
HDBSCAN	-0.0139	0.0420
UMAP / HDBSCAN	0.0208	0.0570
UMAP-2D / HDBSCAN	0.0208	0.0570
UMAP-3D / HDBSCAN	0.0189	0.0572

Table 4.3: Clustering Performance Metrics

5

Conclusions and Future Directions

The Conclusion will summarise the main finding of the research and its implications for the field. It will also highlight the limitations and suggest directions for future research.

5.1 Summary

The main problem being addressed in work is the lack of sufficient interpretability and explainability in modern machine learning techniques typically used for predictive modelling, particularly in high-risk and risk-sensitive fields such as life insurance predictive underwriting.

The rationale for the methodology follows here;

1. Identifying (coerced) clusters in complex data that align with existing classifications.

2. Coercion by means of ensembling centroid-based, density-based or manifold-based algorithms including optimisation to identify the algorithm and parameter combinations find clusters that are sufficient close to ground truth clusters (risk classification) based on metrics (e.g. ARI, ABI).

3. These algorithms have robust mathematical explanations for the clustering such that it is possible to explain in a mathematically robust way from first principles which cluster (risk classification) any sample belongs to.

5.2 Conclusions

5.3 Contributions

The focus is on high-risk and risk sensitive systems where both interpretation and causal explanation are required in addition to human review and oversight - in this case, life insurance underwriting.

The contribution is to propose a novel approach to using unsupervised learning for the purpose of adding interpretability and causal explainability to predictive underwriting that is more interpretable and more causally explainable than typical current supervised method with SHAP.

The adoption of clustering provides the interpretability concept and, where it is feasible to reduce the dataset to 2 or 3 dimensions for visualisation, this may offer a new mode towards exploring and gaining insights on the data with availability or provision of the appropriate additional tooling to the user. This visual interactive aspect can offer a window of transparency, which with tooling can support detection of anomolies, biases, and discrimination. Further, tooling can allow interventions to re-classify labels to regularise according to an identified cluster, as shown in this work.

The hope is that the work here may provide a modest beginning to a journey towards interpretability and critically, retaining human control and causal reasoning in machine learning for risk-sensitive applications.

5.4 Future Work

A key challenge here was the complexity of the Prudential dataset. Further experiments could focus on acquiring additional industry datasets or subsets of the Prudential dataset for example, binary classification for example (e.g.whether

an application has been declined) to determine whether this can offer improved performance.

The complexity of the dataset also meant that reducing to two or three dimensions for visualisation purposes compromised performance results in terms of being able to identify true clusters (risk classifications). Future work would focus much more on optimisation of the unsupervised algorithms and investigate coercion techniques to identify methods of better matching the clusters to ground truth. Future work could experiment with different embedding dimension options in this regard exploring the trade-offs between accuracy (higher dimensions) and visualisation capability (lower dimensions).

Further work might include investigating other contemporary techniques for dimensionality reduction, for example, Tenenbaum et al. (2000) influential work which unlike classical techniques such as PCA and MDS, can discover the nonlinear degrees of freedom underlying complex natural observations, such as human handwriting or faces.

Future work could explore the applicability of Parametric UMAP as a dimension reduction algorithm to identify if this can offer improved performance while retaining full explainability. By default, Parametric UMAP uses a 3-layer 100-neuron fully connected neural network, however, is customisable to other architectures. Also work investigate combining multiple UMAP models.

Once a reasonable performance level is achieved a prototype system can be created based on the Tensorboard interactive environment created here, to generate feedback from insurance underwriters.

References

A. C. Yeo, K. Smith, R. Willis and M. Brooks (2003), 'A Comparison of Soft Computing and Traditional Approaches for Risk Classification and Claim Cost Prediction in the Automobile Insurance Industry'. 8

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015), 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: *https://www.tensorflow.org/* 50, 86

Allison, P. (2012), 'When Can You Safely Ignore Multicollinearity?'.
URL: *https://statisticalhorizons.com/multicollinearity/* 58

Apté, C., Liu, B., Pednault, E. P. D. and Smyth, P. (2002), Of data mining, *in* 'OF DATA MINING'.
URL: *https://api.semanticscholar.org/CorpusID:15896869* 11

Art. 22 GDPR – Automated Individual Decision-Making, Including Profiling (n.d.). 21, 22

Biddle, R., Liu, S., Tilocca, P. and Xu, G. (2018), Automated underwriting in life insurance: Predictions and optimisation, *in* 'Databases Theory and Applications: 29th Australasian Database Conference, ADC 2018, Gold Coast, QLD, Australia, May 24-27, 2018, Proceedings 29', Springer, pp. 135–146. 12

REFERENCES

Boodhun, N. and Jayabalan, M. (2018), "Risk prediction in life insurance industry using supervised learning algorithms'', *Complex Intelligent Systems* **4**, 145–154.
URL: *https://doi.org/10.1007/s40747-018-0072-1* 12, 15, 16, 24, 42, 43, 47

Brownlee, J. (2016), 'Data Leakage in Machine Learning'. 56

Brownlee, J. (2020), 'Information Gain and Mutual Information for machine learning'.
URL: *https://machinelearningmastery.com/information-gain-and-mutual-information/* 58

Cevolini, A. and Esposito, E. (2020), 'From pool to profile: Social consequences of algorithmic prediction in insurance', *Big Data & Society* **7**.
URL: *https://api.semanticscholar.org/CorpusID:225344235* 18, 26

Cevolini, A. and Esposito, E. (2022), 'From actuarial to behavioural valuation. the impact of telematics on motor insurance', *Valuation Studies* .
URL: *https://api.semanticscholar.org/CorpusID:255672863* 18, 19

Chaohsin Lin (2009), 'Using neural networks as a support tool in the decision making for insurance industry', *Expert systems with applications* . 8

Charpentier, A., Denuit, M. and Elie, R. (2015), Segmentation et mutualisation, les deux faces d'une même pièce?
URL: *https://api.semanticscholar.org/CorpusID:194139808* 17, 18

ECHR - Homepage of the European Court of Human Rights - ECHR - ECHR / CEDH (n.d.), https://www.echr.coe.int. 22

European Parliament. Directorate General for Parliamentary Research Services. (2019), *Understanding Algorithmic Decision-Making: Opportunities and Challenges.*, Publications Office, LU. 22

F. Ewald (2019), 'The Values of Insurance', *Grey Room* . 7, 25

Gary A. Wicklund and R. Roth (1987), 'Expert systems in insurance underwriting: model development and application', *Special Interest Group on Computer Personnel Research Annual Conference* . 8

Geoffrey Clark (1997), 'Life insurance in the society and culture of London, 1700–75', *Urban History* . 7

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G.-Z. (2019), 'XAI—Explainable artificial intelligence', *Science Robotics* **4**(37), eaay7120. 28

Healy, J. and McInnes, L. (2016), Clustering: A guide for the perplexed, *in* 'Proceedings of the PyData DC 2016 Conference'. Accessed: 2024-07-31. 38, 39

High-Level Summary of the AI Act — EU Artificial Intelligence Act (n.d.).
URL: *https://artificialintelligenceact.eu/high-level-summary/* 22

Huang, X. and Marques-Silva, J. (2023), 'The Inadequacy of Shapley Values for Explainability'. 69

Hutagaol, B. and Mauritsius, T. (2020), "Risk Level Prediction of Life Insurance Applicant using Machine Learning'', *International Journal of Advanced Trends in Computer Science and Engineering* **9**(2), 2213–2220.
URL: *https://doi.org/10.30534/ijatcse/2020/199922020* 14, 15, 24, 42, 43

K. Aggour, P. Bonissone, W. Cheetham and R. P. Messmer (2005), 'Automating the Underwriting of Insurance Applications', *The AI Magazine* . 8

Klein, A. M. (2013), 'Life insurance underwriting in the united states – yesterday, today and tomorrow', *British Actuarial Journal* **18**, 486 – 502.
URL: *https://api.semanticscholar.org/CorpusID:167674142* 9

Levantesi, S. and Pizzorusso, V. (2019), 'Application of machine learning to mortality modeling and forecasting', *Risks* .
URL: *https://api.semanticscholar.org/CorpusID:86513899* 12

Macedo, L. (2009), The role of the underwriter in insurance, *in* 'The role of the underwriter in insurance'.
URL: *https://api.semanticscholar.org/CorpusID:167135291* 10

Maier, M., Carlotto, H., Sanchez, F., Balogun, S. and Merritt, S. (2019), 'Transforming Underwriting in the Life Insurance Industry', *In Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, available: https://doi.org/10.1609/aaai.v33i01.33019373, Vol. 33, pp. 9373–9380.
URL: *https://doi.org/10.1609/aaai.v33i01.33019373* 1, 13, 15, 16, 17, 24, 25, 42, 47, 48

REFERENCES

Maier, M., Carlotto, H., Saperstein, S., Sanchez, F., Balogun, S. and Merritt, S. (2020), "Improving the Accuracy and Transparency of Underwriting with AI to Transform the Life Insurance Industry'', *in AI Magazine* [online], available: `https://doi.org/10.1609/aimag.v41i3.5320` [accessed: 7 May 2023] **41**(3), 78–93.
URL: *https://doi.org/10.1609/aimag.v41i3.5320* 13, 17, 43

McInnes, L. (2018), A bluffer's guide to dimension reduction, *in* 'Proceedings of the PyData New York City 2018 Conference'. Accessed: 2024-07-31. ix, 29, 30, 32, 33, 34, 35, 36, 37, 38

McInnes, L., Healy, J. and Melville, J. (2018), *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Accessed: 2024-07-29.
URL: *https://umap-learn.readthedocs.io/en/latest/index.html* 80, 82, 84, 85

Milano, A. F. (2001), 'Evidence-based risk assessment.', *Journal of insurance medicine* **33 3**, 239–50.
URL: *https://api.semanticscholar.org/CorpusID:8713848* 9

Montoya, A. and Cukierski, W. e. a. (2015), 'Prudential life insurance assessment'.
URL: *https://kaggle.com/competitions/prudential-life-insurance-assessment* vii, 48, 52

Nanga, S., Bawah, A. T., Acquaye, B., Billa, M.-I., Baeta, F., Odai, N. A., Obeng, S. K. and Nsiah, A. D. (2021), 'Review of dimension reduction methods', *Journal of Data Analysis and Information Processing* .
URL: *https://api.semanticscholar.org/CorpusID:239719601* 30

Nikolopoulos, C. and Duvendack, S. (1994), 'A hybrid machine learning system and its application to insurance underwriting', *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence* pp. 692–695 vol.2.
URL: *https://api.semanticscholar.org/CorpusID:39542620* 11

Notice of Adoption - New Regulation 10-1-1 Governance and Risk Management Framework Requirements for Life Insurers' Use of External Consumer Data and Information Sources, Algorithms, and Predictive Models — DORA Division of Insurance (n.d.), https://doi.colorado.gov/announcements/notice-of-adoption-new-regulation-10-1-1-governance-and-risk-management-framework. 20, 23

Paisner, B. B. C. L. (2024), 'US state-by-state AI legislation snapshot', https://www.bclplaw.com/en-US/events-insights-news/us-state-by-state-artificial-intelligence-legislation-snapshot.html. ix, 21

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830. 64

R. Pearson (2002), 'Moral Hazard and the Assessment of Insurance Risk in Eighteenth- and Early-Nineteenth-Century Britain', *Business History Review* . 7

Raymond C. M. Lee, Kai-Pan Mark and Dickson K. W. Chiu (2007), 'Enhancing Workflow Automation in Insurance Underwriting Processes with Web Services and Alerts', *Hawaii International Conference on System Sciences* . 8

Rights (OCR), O. f. C. (2021), 'Health Information Privacy', https://www.hhs.gov/hipaa/index.html. 19

Sahai, R., Al-Ataby, A., Assi, S., Jayabalan, M., Liatsis, P., Loy, C. K., Al-Hamid, A. M., Al-Sudani, S., Alamran, M. and Kolivand, H. (2022), Insurance risk prediction using machine learning, *in* 'DaSET'.
URL: *https://api.semanticscholar.org/CorpusID:259120664* 14, 17

Sainburg, T., McInnes, L. and Gentner, T. Q. (2021), 'Parametric umap embeddings for representation and semisupervised learning', *Neural Computation* **33**(11), 2881–2907. 80, 83, 84

Sarpal, K. (2023), 'Machine Learning for Risk Classification - KS', https://kaggle.com/code/karansarpal/machine-learning-for-risk-classification-ks. 48

Sen. Cantwell, M. D.-W. (2024), 'Text - S.4178 - 118th Congress (2023-2024): Future of Artificial Intelligence Innovation Act of 2024', https://www.congress.gov/bill/118th-congress/senate-bill/4178/text. 19, 20

T. Alborn (2000), 'Betting on Lives: The Culture of Life Insurance in England, 1695–1775. By Geoffrey Clark. Manchester, U.K.: Manchester University Press, 1999. 220 pp. Appendices, bibliography, photographs, index, maps, notes, and tables. Cloth, $69.95. ISBN 0-719-05675-6', *Business History Review* . 7

REFERENCES

Tapan Biswas (1997), 'The Insurance Market'. 7

Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000), 'A global geometric framework for nonlinear dimensionality reduction.', *Science* **290 5500**, 2319–23.
URL: *https://api.semanticscholar.org/CorpusID:221338160* 93

The Final Colorado AI Insurance Regulations: What's New and How to Prepare (n.d.), https://www.debevoise.com/insights/publications/2023/10/the-final-colorado-ai-insurance-regulations-whats. 20

Tobin, J. (2024), 'Predictive and Decision-making Algorithms in Public Policy'. 23

V. Zelizer (1979), 'Morals and Markets: The Development of Life Insurance in the United States'. 7

Varadarajan, V. and Kakumanu, V. K. (2024), 'Evaluation of risk level assessment strategies in life insurance: A review of the literature', *Journal of Autonomous Intelligence* .
URL: *https://api.semanticscholar.org/CorpusID:268561068* 15, 25, 52

Wang, W. (2021), 'Predictive machine learning for underwriting life and health insurance', *In proceedings of The Actuarial Society of South Africa's 2021 Virtual Convention*, October 19-22, 2021, [online] available: https://www.actuarialsociety.org.za/convention/wp-content/uploads/2021/10/2021-ASSA-Wang-FIN-reduced.pdf. 1, 14, 17, 24, 42, 43, 47