Assignment 1 – Data Exploration and Feature Extraction

Due Date: 11PM Sun 10 Sep 2023

Word limit: ~1,500 words (excluding spaces, tables, and references)

Submission

The assignment should be completed in groups and submitted through learnonline. In case that you are not able to find a team, you are also welcome to submit your individual work. Please make sure that you complete all the parts of the assignment.

Assignment Requirement

The Healthcare Fraud Detection dataset (https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis) you will be working with is a rich and comprehensive set of data that provides detailed information about healthcare providers, their billing practices, patient demographics, and procedures performed. This data, while intricate, is indispensable for building a robust model to identify fraudulent activities in the healthcare industry.

This dataset is typically derived from healthcare insurance claims and could include various pieces of information, including but not limited to:

- **Provider Information**: This includes unique identifiers for healthcare providers, their location, their specialty, and perhaps information about their patients' demographics.
- **Billing Information**: Detailed information about the medical services billed to the insurance company. This may include the type of service, the diagnosis, the cost of the service, and the date of service.
- **Patient Demographics**: Information about the patients like their age, gender, medical history, and location.
- **Procedural Information**: Information about the procedures performed, which can include procedure codes, the medical necessity of the procedures, and the type of procedures (such as surgery, consultation, or test).
- **Fraud Labels**: This is crucial for training predictive models. In some datasets, certain claims or providers might be labeled as fraudulent based on past investigations or confirmed fraud cases. This forms the basis for supervised learning approaches.

Given the diversity and depth of information in this dataset, it indeed presents a significant level of complexity. However, this also means there is great potential for discovering insightful patterns and trends that could help identify fraudulent behavior. Your task will be to apply your data mining skills to explore this data, engineer relevant features, and build a model that can predict potentially fraudulent providers.

Remember that healthcare fraud detection is not just about identifying billing anomalies; it is also about understanding the behaviors and patterns that might indicate a deeper problem. Therefore, a thorough exploratory data analysis and thoughtful feature engineering will be pivotal in your approach.

Assignment 1 should include three parts, as outlined below.

Part 1: Introduction (5 marks)

In this section, you should detail the objectives of this assignment. Discuss the significance of detecting healthcare fraud and how it affects different stakeholders, including patients, healthcare providers, and insurance companies. Elucidate how data mining techniques could be instrumental in identifying potential fraud, thus helping to curtail financial losses and improve healthcare services.

Part 2: Related Work (8 marks)

In this part, you should review a minimum of three academic papers that have addressed the problem of predicting healthcare fraud or similar challenges. For each paper, construct a focused narrative that draws out the primary themes and findings. Instead of merely listing the papers and their content, interweave a storyline that connects the insights from these papers to your assignment's objectives.

- Point: Begin each paragraph by establishing a key point derived from the paper(s). This could revolve around a common methodological approach (such as the use of SVM for fraud detection), a prevalent feature set, or a significant problem that researchers have striven to address in the realm of healthcare fraud.
- Evidence: Support your point with evidence extracted directly from the paper(s). This includes specific findings, statements, results, or observations made by the authors. By doing this, you demonstrate your understanding of the papers and provide concrete backing for the points you're making.
- Relevance: Conclude each paragraph by connecting the evidence back to your work. Explain how the insights gleaned from the paper(s) have informed your approach to this assignment, particularly in terms of feature extraction from the dataset.

In structuring your review in this way, you will be engaging deeply with the literature, fostering a coherent narrative, and clearly demonstrating the influence of prior work on your current assignment.

Part 3: Data Exploration and Feature Engineering (12 marks)

This part is devoted to exploring the healthcare fraud dataset and creating new features that could be helpful in predicting fraudulent behavior. Your exploration should involve statistical analysis of the data - looking at the means, medians, ranges, standard deviations, to name a few, of different features.

Visualizations, such as histograms, scatter plots, or box plots, can be very useful to understand the data better and to find any patterns, trends, or anomalies in the data.

Correlation analysis will help you determine if there are any features in the data that are strongly related to each other. This can be very useful when deciding which features to include in your predictive model, as features that are very strongly correlated can often be reduced to a single feature.

In feature extraction, you will process and transform the raw data into features that can be used in your model. This could involve combining several raw features into a single derived feature, or it could involve applying some kind of mathematical transformation to a feature to make it more useful for your model.

Lastly, assemble a CSV file that consists of your chosen features, as well as the target variable - in this case, an indicator of whether or not a provider is fraudulent. This will serve as the input to the predictive model that you'll develop in the next phase of the project.

Marking Criteria

High Distinction

To achieve a high distinction, your submission should comprehensively address all three assignment sections, demonstrating a clear understanding of course material. It should include a robust inspection of data supported by relevant, high-quality figures, each accurately labelled and accompanied by in-depth analysis of observed trends. A critical review of existing work is expected, including a detailed explanation of how it guided the extraction of the features you incorporated in your analysis. For each feature, provide clear explanations and summary statistics. The narrative should be clear, concise, and coherent. Your work should exhibit comprehensive knowledge of the concepts through detailed descriptions, insightful explanations, and well-articulated discussions. A CSV file with the final feature set should also be submitted.

Distinction

To secure a distinction, your work should competently address all three assignment sections, demonstrating a clear understanding of the course topics. It should include an analysis of the data supported by relevant figures, accurately labelled, and supplemented with a thorough explanation of observed trends. A summary of prior work should be provided, clearly illustrating how it influenced the features you selected. For each feature, provide detailed explanations and summary statistics. In demonstrating a well-considered understanding of the concepts, your descriptions and explanations should be thoughtful and articulate. Ensure any feedback from previous reports is addressed. Submit a CSV file containing the final feature set.

Credit

To attain a credit, your work should adequately cover all three assignment sections. Data inspection should include basic plots like histograms, supported by relevant figures and analysis of observed trends, all accurately labelled. Reference relevant literature in part 2 and provide a list of features used in your work. Your submission should demonstrate a sound understanding of the course concepts through clear descriptions and explanations. Finally, submit a CSV file with your final feature set.

Pass

To achieve a pass, you need to address at least two sections of the assignment. In doing so, you should demonstrate an understanding of course concepts through clear descriptions, and a conscientious effort to design a meaningful set of features.

Academic integrity

You are expected to reference and cite all resources mentioned using a selected referencing convention (e.g., UniSA Harvard, or APA). If you used Generative AI, please state to what extent and how did the use of tools such as ChatGPT or Claude impact your assignment.

Extensions

Extensions for assignments are available under the following conditions:

- permanent or temporary disability, or
- compassionate grounds

In all cases, documentary evidence (e.g., medical certificate, road accident report, obituary) must be presented to the Course Coordinator. A medical certificate produced on or after the due date will not be accepted unless you are hospitalized.

If you apply for extension within 24 hours before the deadline, you must see the course coordinator in person unless you are in an emergency like being admitted in a hospital.