School of Computing, Engineering & Digital Technology
Department of Computing and Games
Middlesbrough TS1 3BX

Final Year Project MSc Applied AI: Explainable AI in Bias Mitigation

An academic research paper for possible submission
to the ACM Digital Library
Submitted in partial requirements for the
degree of MSc Applied Artificial Intelligence

Submission Date: 11 May 2022

Abstract

Fairness in machine learning is growing in attention where recent methods have addressed the discrimination between particular demographic groups that contain attributes classified as sensitive that include age, gender and race. Studies have also shown if the dataset itself is biased, it can lead to unfair decision making when used in training models.

In this paper, I propose a framework combining a number of classification algorithms, fairness metrics and feature explainability methods to highlight and potentially address bias generation resulting from the training data.

Keywords: *Bias detection, Explainable Artificial Intelligence, Algorithmic Fairness, Credit Card Default, Diabetes Readmission*

Contents

.

List of Figures

List of Tables

List of Formulas

True Positive Rate (TPR)

$$\frac{True\ Positive}{True\ Positive + False\ Negative}$$

False Positive Rate (FPR)

$$\frac{False\ Positive}{False\ Positive + True\ Negative}$$

Accuracy

$$\frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

Precision

$$\frac{True\ Positive}{True\ Positive + False\ Positive}$$

F1-score

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

Chapter 1

Introduction

1.1 Background

Alongside the growing digitisation of data, there has also been a growing reliance on Artificial Intelligence (AI) systems and machine learning algorithms in the automation of decision-making processes by both private and public sectors. The technologies that utilize them are extensive and include a number of economic sectors such as finance, advertising and healthcare. AI is also beginning to have an impact on governance alongside the increase in deployment of these automated models to improve accuracy and ultimately dictate the decision-making process.

However, with the growing complexity and use of AI that is not limited to sensitive areas, it has introduced the debate surrounding bias and fairness. Human biases have been historically well-documented but these biases are beginning to pave their way into these AI systems, whether intentional or not but the results can be harmful regardless. Biased predictions, especially towards a minority or under-represented group of the population adversely affects the trustworthiness of not only the model but the organisation that utilizes it. These biases may be difficult to detect and may stem from an uncountable number of reasons; from being under proxies to confounding variables. These biases are then replicated in the learning models themselves and propagated forward – causing a vicious cycle of bias.

One of the main reasons for bias in AI is data bias, also known as “algorithmic bias”. Where collected data is based on human-based decision, this causes the data itself to be biased. This can be due to the under-representation or inclusion of sensitive relevant features that include gender, marital status and race to list a few. However, there is also a possibility that should a bias be mitigated with

respect to a certain group, a bias may arise in another group not previously considered.

A key starting point for researchers have been defining data bias and what influences it.
If the bias is not defined at the onset, it will be replicated, propagated forward and lead to skewed models with negative implications in its results. This cycle can be prevented by defining and understanding what the bias is to be able to apply the appropriate mitigation technique.

As the use of automated decision-making models are being increasingly implemented across many industries, end-users or consumers will have fairness at the front of their mind when considering using an organisation. As regulations detailing stricter guidelines are making its way into the industry as regulators are becoming increasingly aware of the issue at hand, it will be advantageous for organisations to take action now to save resources and ensure their customers are better protect and maintain trustworthiness in their relationship with the customer. Overall, it is possible to obtain competitive advantage by mitigating biases in their datasets and hence, algorithms.

A key performance indicator of an algorithm is its performance where high accuracies and low errors are achieved by statistical adjustments in addition to parameter tuning. However, this may not be of great significant if there are certain groups of individuals that may be unfairly disadvantaged as a result of unmanaged biases. It may be beneficial to use a data-centric approach to address the problem instead of only utilizing hyperparameter tuning with datasets of poor quality.

As the use of algorithms are becoming more prevalent across industries, the relevant stakeholders must be proactively aware in addressing the factors that can contribute to biases. Resolving algorithmic bias at the start will enable the averting of harmful impacts it will have on users in addition to the liabilities it can have against the creators of the algorithms. The research and understanding of

the consequences of the application of algorithms in decision making processes – intended and unintended – is growing in necessity as existing policies may not be fully sufficient to regulate the identification, mitigation and regulation on consumer impacts.

My paper presents the analysis of a number of machine learning algorithms that quantify fairness alongside bias quantifiers to demonstrate how they can be of use in combination to promote the fairness of AI applications and machine learning techniques in the industry.

1.2 Research Question

This report aims at casting light on the problem of model transparency and propose a framework addressing the main challenge of a dataset, that is how individual training data entries influence biased and inaccurate results and how identifying biased entries can be used to improve representation in the dataset to improve both fairness and accuracy.

1.3 Aims and objectives

1.3.1 Aim

This project explores the analysis of a combination of fair machine learning algorithms and bias quantifiers to identify sources of bias in the training dataset in order to mitigate bias (through updating the dataset based on the identified biased inputs). The aim is to identify and promote the fairness of AI applications and machine learning techniques in the industry.

1.3.2 Objectives

The first objective of this paper is to highlight how existing technologies influence how bias is identified in the dataset – either using evaluation metrics or bias quantifiers. A framework will be proposed as the second objective to identify bias in datasets using a combination of fairness and bias quantifiers alongside existing machine learning models. The third objective is proposing a method to produce the relationship between the input and output to identify data entries influencing inaccurate or biased results.

1.4 Solution approach

The project uses an iterative development life cycle as pictured below in Figure 1.1;

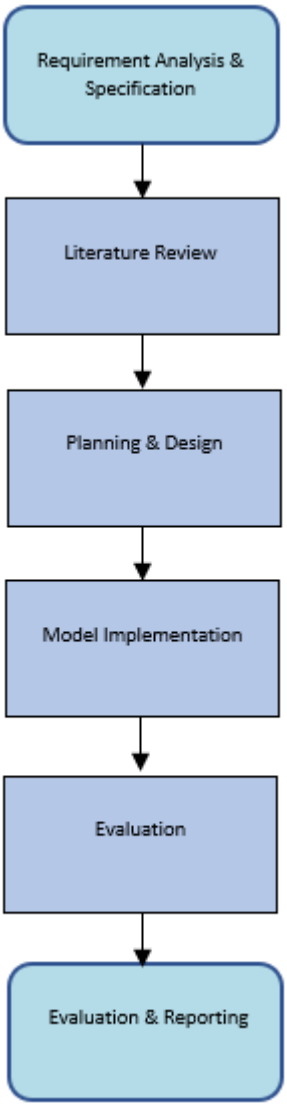

Figure 1.1. Solution Approach

Each step will be explained in detail in each subsection of the report, as summarized in the upcoming section 1.6 Organisation of the Report.

1.5 Summary of contributions and achievements

This report has achieved two out of the three objectives, that were to analyse how the two proposed classification algorithms performed in terms of their performance and fairness metrics and also by proposing a framework of algorithms and bias quantifiers to evaluate the significance of the input and their features on the output of the model at a global and local scale.
Where the third objective of proposing a method to produce the relationship between the input and output data to identify data entries influence biased and inaccurate results were not met, this will be proposed as future work to be pursued following the submission of this paper.

1.6 Report Description and Summary

Chapter 2 reviews the existing literature that covers the types of bias, how they are detected and mitigated in addition to the ethical framework relevant to it. Chapter 3 then provides a description of the methodology undertaken and introduces the performance and fairness metrics used to quantify and assess the performance of the chose machine algorithms that are “designed for the task of identifying bias”. Chapter 4 details the results of the implementation and a discussion surrounding how it answers the objectives that were set. Chapter 5 summarises the project as a whole and concludes the paper.

Chapter 2

Literature Review

This literature review discusses key aspects of Explainable Artificial Intelligence by first describing the use of AI applications in the industry then exploring the relationship data bias has with data in the industry by defining what it means, how it is detected and mitigated and finally, a literature review on the ethical frameworks surrounding fair Artificial Intelligence.

2.1 AI Application in the industry

Financial and fintech services have the opportunity to gain significant competitive advantage by addressing unfair biases in their AI systems, ahead of potentially stricter regulations being introduced in the future. In the current era of Big Data, fairness and data transparency should be at the core practices today – more so for auditability of an organisation's practices and the consumer's trust despite the potentially competitive disadvantage should an organisation publicise their in-house algorithm.

Systemic bias against a protected class can result in disparate impacts such as the denial of credit [15] or incorrect profiling for recidivism.

The discrimination in algorithm-based predictions exists in various applications including the healthcare and banking (Lee and Floridi) where decision-making models have been identified to produce biased decision for specific groups that have contain sensitive attributes. In a 2019 paper written by Brian Powers [37], it was shown that a commonly used algorithms by a number of health systems are racially biased where it has a potential harmful effect on patients when the information is used as recommendations for certain people in a professional's care.

2.2 Defining Bias

Bias in an algorithm can be sourced from an unrepresentative training data or a flawed dataset that reflects historical inequalities. These biases have the potential to result in decisions that may have a disparate effect on the groups of individuals possessing sensitive attributes although there may not have been an intention to do so.

An automated decision-making model's output that includes "bias" that will be broadly defined as outcomes that are systematically less favourable towards individuals of a specific group and also where there is no relevant difference between the groups that would justify such damage [9].

There are two main types of bias or fairness that are of concern to stakeholders – group and individual fairness. Group fairness that relates to groups of people with similar protected attributes such as gender and race. It focuses on measuring if these groups are treated similarly by the model [3].

[3] The motivation for individual fairness is that simple statistical parity between the protected groups in each outcome could be intuitively fair at an individual level.

Research results have revealed that there are algorithms that are at the risk of replicating human biases and in particular, those that affect protected groups [8]. An example is the automated risk assessment that is used by judges in the United States (U.S.) in determining an individual's bail or sentencing limit [24] such as the COMPAS algorithm. It however has the potential of generating incorrect conclusions that have resulted in adverse affects on certain groups in terms of higher bails or longer prison sentences on people of colour.

This relates to individual fairness that focuses on individual and is intuitively defined as treating similarly (Dwork et al., 2012) [3].

[3][5-18]To avoid this problem, various papers have proposed an individual level of fairness where the intuition supporting this measure is that people who are similar in profile should be given similar predictions of decisions.

2.3 Types of Bias

Algorithmic or data bias has the capability to present itself in various ways with differing degrees of consequences, as described by the examples below.

2.3.1 Training data

Should the training data for an algorithm not be as representative of certain groups as they are for others, the predictors from the model may be systematically worse for under-represented groups. An argument by Turner Lee is that under-representation of a certain group can be the result of limited diversity amongst the designers of the training dataset [13].

An algorithm with an over-representation can also skew decisions towards a particular result. A research at the Georgetown Law School showed that African-Americans had a higher likelihood of being singled out as a result of over-representation in mug-shot databases which gave rise to more possibilities of being matched falsely hence the bias [14].

2.3.2 Historical Human Biases

Human biases have historically embedded prejudices against certain demographic groups, some containing sensitive attributes. As training datasets are collected and processed using human-based decision, their use in computer models can lead to the embedded biases being amplified. The COMPAS algorithm is an example where historical prejudices within the criminal justice system have results in a higher probability of African-Americans being arrested and incarcerated.

Human biases that are left unchecked and undetected may make its way into automated decision-making processes and thus their results without the user's knowledge. It is the responsibility of the designers of the algorithm and its providers to be accountable of the implications of not debiasing their datasets and the impact that can have on an algorithm.

2.4 Detecting Bias

The first step to understanding the cause of biases is the implementation of effective algorithmic hygiene. The results however, may still be problematic even after correcting flawed datasets because the context of the dataset is relevant in the bias detection phase.

All approaches for bias detection should involve the careful management of an individual's sensitive information including any data that identifies a person possessing a protected attribute (ie. race and gender). The discussion surrounding they way algorithms are influenced by sensitive or protected attributes should include the balance between fairness and accuracy in these models. This will be further discussed in sub-section 2.7 Fairness and Accuracy Trade-Offs.

During the analysis stage of an algorithm's outputs, a comparison should be made between the outputs for differing groups of inputs, especially those that contain a sensitive attribute. This also sheds light on the distribution of error rates and discrepancies between the differing groups of inputs. Note that it may not be possible to obtain equal values between groups for all the differing error measures [20].

2.5 Measuring Bias

When analysing for model fairness, a distinction is made between the analysis for group fairness and individual fairness where the former is defined by the equality of certain statistics determined on protected attributes [3] and the latter is where individuals with similar feature background should obtain similar results (Dwork et al. 2012).

The challenge of measuring bias using the existing methods, a few examples of which were mentioned above, is dependent on an absolute mathematical condition and are not in consideration for the bias that could be already a part of the dataset. There is also limited explainability in their outcomes that are specific locally.

FairML is a Python library introduced in 2017 by Julius Adebayo [38], which will be used as part of this paper's proposed framework, was designed to audit a black-

box predictive model and produces the level of importance each input feature has on the model's output.

Another method this paper proposes using is the LIME method [39] that provides interpretability for individual predictions – or at a local level. It is used to visualise the relationship between input features and the target variable for an individual input entry. The LIME method produces the contribution of each input feature towards the prediction of the classification for an individual input entry.

2.6 Bias Mitigation

A number of research papers have been published within the scope of prediction quality disparity [40] and mitigating discrimination in a prediction's result. One of the methods is minimising the influence sensitive attributes have by reformatting the data to be independent of it (Wang and Huang, 2019, Adel et al. 2019). There has also been several methods that have been proposed to manage the bias arising from under-representation (Jian and Nachum, 2019).

In 2018, the GAN (generative adversarial network) was proposed by Fird-Adar et al. that was used in balancing the dataset by augmenting data with sample size disparity. For the relevance of this paper, these methods were not implemented in order to define the possibility of detecting and quantifying bias present in the original dataset.

In addition to providing a sense of transparency, another good practice for detecting and mitigating bias is regular and formal auditing of an algorithm in development. When performed by a third-party, it is possible to obtain an insight into the input data and the resulting model's output. Although there exists ethical frameworks rather than enforced laws at present, developers and providers should be conscious of guarding these frameworks in the algorithmic design to prevent historical discrimination from remaining prevalent.

2.7 Fairness and Accuracy Trade -Offs

In the debate between the trade-off between fairness and accuracy, there should be a focus on the evaluation of a societal definition of fairness and its potential implications. In researching the COMPAS algorithm, Corbett-Davies et al. see “an inherent tension between minimizing violent crime and satisfying common notions of fairness” [11]. There should be a conscious effort to identify methods of reducing disparities between sensitive groups without having to sacrifice a model’s performance and more so when there has to be a balance between fairness and accuracy.

The focus in the debate surrounding the trade-off between accuracy and ethics should be on the evaluation of both the societal definition of fairness and its potential social implication. The algorithm’s developers have the responsibility of identifying how inequalities between groups with sensitive attributes can be minimised without affecting the performance of the model [18].

2.8 Ethical Frameworks

The fundamental question that still comes to mind is if we can trust a computer to decide something a human would, despite the fact that historical human biases have embedded their way through a variety of decision-making applications. In response, there have been a number of efforts in the recent years in developing standards for the ethical use of Artificial Intelligence [21]. For instance, the European Union (EU) in 2019, produced the document titled “Ethics Guidelines for Trustworthy AI” and lists seven key governance principles. The two that I would focus on for this paper are (4) transparency and (7) accountability as the guidelines reflects that it would be “unethical to unfairly discriminate”. This ties in with Microsoft’s Responsible AI principles in practice [41] where three of the six principles are (1) inclusiveness, (2) transparency and (3) accountability. These five principles dictate fairness through the availability of an inclusive and equal access and design.

It is important to note however, it remains to be a challenge to strictly define and

measure fairness. Fairness is a human-based formula, so to speak, which brings us back to the question if we can trust a computer to decide something a human would.

As such, introducing and implementing automated decision-making systems into the industry, developers and providers have the responsibility of identifying an appropriate balance between their automated system's accuracy and fairness. This is to protect from perpetuating the existing prejudices and societal inequalities. As written by Lee, Resnick and Barton, algorithm developers and their providers should be asking the following question: "Will we leave some groups of people worse off as a result of the algorithm's design or its unintended consequences?" [18].

Subjecting an algorithm to the methods mentioned in the above subsections gives the opportunity to challenge the various ways fairness have been defined and improve the balance between accuracy and bias.

2.9 Summary

There are a myriad of opportunities to improve the fairness and accuracy of an algorithm from starting at the data management at the pre-processing stage to output explainability at the post-processing stage. Datasets that may not sufficiently represent certain groups could be reviewed and improved which is what this paper aims to contribute to. The understanding of the sources of bias in sensitive attributes can assist understanding where the training data lacks in representing said groups.
In order to build trust with consumers, developers and providers should consider the transparency in their algorithms to achieve auditability and set fairness as part of their model's performance to reduce the prevalence of historical human biases in the results they put out into the industry.

The literature review so far has presented that accuracy should not be the only performance indicator of a successful model.

Chapter 3

Methodology

As described in Section 1.4, the Figure 3.1 below describes the solution approach of this paper. The first section below will begin describing the requirement specification based on the literature review carried out. This will be followed by an explanation of the design and implementation phases.

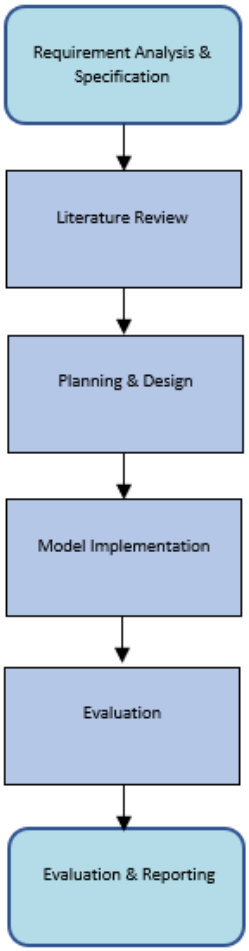

Figure 3.1 Design Flowchart

3.1 Requirement Specification

This section will detail the key steps to the methodology in the order of data acquisition and pre-processing, data analysis and model processing and the model's post-processing in terms of the selection of bias quantifiers and fairness metrics to meet the objectives of this paper.

3.1.1 Datasets

Table x contains the two binary classification datasets used in this paper. Each of the two datasets selected for this paper contains at least one identified protected attribute that could influence the outcome of a prediction.
Examples of the protected – or sensitive – attributes are "gender", "age" and "marital status".

Group fairness aims for groups to be treated similarly regardless of the presence of sensitive attributes and the same follows for individual fairness that aims for individuals with similar profiles to be treated equally by the model [25].

Dataset	No. Rows	No. Features	Protected Attribute	Label	
				1	0
Default of Credit Card [42], [44]	30,000	25	- Gender - Marital Status	Default	Not Default
Diabetes Readmission [43], [45]	101,766	50	- Race	Readmitted	Not Readmitted

Table 3.1: Details of the datasets used in this study

3.1.2 Performance Metrics Used in this Study

A performance indicator or metric is a measure of a model's behaviour. Table x shows the performance metrics utilised in this paper. During the selection of the metrics used in this paper, a number of other metrics were not included because precise implementations and definitions were not be able to be accessed.

Metric Name	Description	Ideal Value
ROC (Receiver Operating Characteristic) Curves	An ROC curve is a graph showing the performance of a classification model across all thresholds. The curve plots the True Positive Rate (TPR) and False Positive Rate (FPR) [26].	1.0
AUC (Area under the ROC curve)	The AUC measures the whole two-dimensional area under the ROC curve as shown in Figure 3.2 The AUC is an aggregate measure of performance over the possible classification thresholds. It can be interpreted as the probability of a model will rank a positive sample higher than a negative sample.	0.9 – 1.0
Accuracy	Accuracy is the number of correctly predicted outputs as a proportion of the total number of predicted outputs [27].	0.9 - 1.0
Precision	Precision is the number of correctly predicted positive outputs [28].	1.0
F1 Score	The F1 score is the combination of the precision and recall as shown in the List of Formulas.	1.0
Specificity	Specificity evaluates the ability of a model to correctly predict true negatives for each category.	1.0

Table 3.2 Description of performance metrics

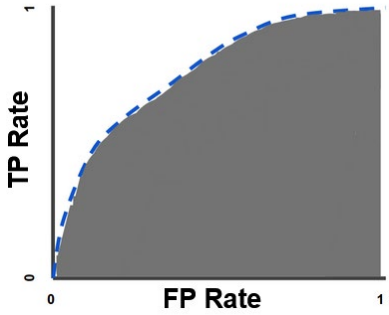

Figure 3.2 ROC Curve

The classification metrics above measures the fairness of the model's output based on the classification results. The metrics are measured based on the original dataset containing the actual labels and the predicted outputs using the confusion matrix as per Table 3.3 below.

	Positive	Negative
Predicted Positive	True Positive $\frac{True\ Positive}{True\ Positive + False\ Negative}$	False Positive $\frac{False\ Positive}{False\ Positive + True\ Negative}$
Predicted Negative	False Negative $\frac{False\ Negative}{False\ Negative + True\ Positive}$	True Negative $\frac{True\ Negative}{True\ Negative + False\ Positive}$

Table 3.3: Formulas of the confusion matrix rates

3.1.3 Fairness Metrics Used in this Study

A fairness metric is defined as the analysis of undesired bias in the models or training data [25]. Table 3.4 shows the fairness metrics selected for this paper. During the selection of the metrics used in this paper, a number of other metrics were not included because precise implementations and definitions were not be able to be accessed.

Fairness Metric	Description
FairML	FairML is a Python library [38] was designed to audit a black-box predictive model and produces the level of importance each input feature has on the model's output.
LIME	This method [39] provides interpretability for individual predictions – or at a local level. It is used to visualise the relationship between input features and the target variable for an individual input entry. The LIME method produces the contribution of each input feature towards the prediction of the classification for an individual input entry.

Table 3.4 Description of fairness metrics

3.2 Data Pre-Processing & Analysis

This section further details the selected datasets including a brief overview of general aspects of the datasets. This section further details the initial exploration and cleaning of the dataset and the results of exploratory data analysis (EDA) and feature engineering.

3.2.1 Credit Card Default Dataset

The Figure 3.3 below shows a sample of the Credit Card Default dataset.

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.payment.next.month
0	1	20000.0	2	2	1	24	2	2	-1	-1	...	0.0	0.0	0.0	0.0	689.0	0.0	0.0	0.0	0.0	1
1	2	120000.0	2	2	2	26	-1	2	0	0	...	3272.0	3455.0	3261.0	0.0	1000.0	1000.0	1000.0	0.0	2000.0	1
2	3	90000.0	2	2	2	34	0	0	0	0	...	14331.0	14948.0	15549.0	1518.0	1500.0	1000.0	1000.0	1000.0	5000.0	0
3	4	50000.0	2	2	1	37	0	0	0	0	...	28314.0	28959.0	29547.0	2000.0	2019.0	1200.0	1100.0	1069.0	1000.0	0
4	5	50000.0	1	2	1	57	-1	0	-1	0	...	20940.0	19146.0	19131.0	2000.0	36681.0	10000.0	9000.0	689.0	679.0	0

5 rows × 25 columns

Figure 3.3 Sample of the Credit Card Default Dataset

The table below describes the 25 attributes available in the dataset.

Variable	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit
SEX	Gender (1=male, 2=female)
EDUCATION	Education status (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
default.payment.next.month	Default payment (1=yes, 0=no)

Table 3.5 A description of the Credit Card Default feature variables

The following Figure 3.4 shows an overall summary of the distribution of each variable in the dataset;

	count	mean	std	min	25%	50%	75%	max
ID	30000.0	15000.500000	8660.398374	1.0	7500.75	15000.5	22500.25	30000.0
LIMIT_BAL	30000.0	167484.322667	129747.661567	10000.0	50000.00	140000.0	240000.00	1000000.0
SEX	30000.0	1.603733	0.489129	1.0	1.00	2.0	2.00	2.0
EDUCATION	30000.0	1.853133	0.790349	0.0	1.00	2.0	2.00	6.0
MARRIAGE	30000.0	1.551867	0.521970	0.0	1.00	2.0	2.00	3.0
AGE	30000.0	35.485500	9.217904	21.0	28.00	34.0	41.00	79.0
PAY_0	30000.0	-0.016700	1.123802	-2.0	-1.00	0.0	0.00	8.0
PAY_2	30000.0	-0.133767	1.197186	-2.0	-1.00	0.0	0.00	8.0
PAY_3	30000.0	-0.166200	1.196868	-2.0	-1.00	0.0	0.00	8.0
PAY_4	30000.0	-0.220667	1.169139	-2.0	-1.00	0.0	0.00	8.0
PAY_5	30000.0	-0.266200	1.133187	-2.0	-1.00	0.0	0.00	8.0
PAY_6	30000.0	-0.291100	1.149988	-2.0	-1.00	0.0	0.00	8.0
BILL_AMT1	30000.0	51223.330900	73635.860576	-165580.0	3558.75	22381.5	67091.00	964511.0
BILL_AMT2	30000.0	49179.075167	71173.768783	-69777.0	2984.75	21200.0	64006.25	983931.0
BILL_AMT3	30000.0	47013.154800	69349.387427	-157264.0	2666.25	20088.5	60164.75	1664089.0
BILL_AMT4	30000.0	43262.948967	64332.856134	-170000.0	2326.75	19052.0	54506.00	891586.0
BILL_AMT5	30000.0	40311.400967	60797.155770	-81334.0	1763.00	18104.5	50190.50	927171.0
BILL_AMT6	30000.0	38871.760400	59554.107537	-339603.0	1256.00	17071.0	49198.25	961664.0
PAY_AMT1	30000.0	5663.580500	16563.280354	0.0	1000.00	2100.0	5006.00	873552.0
PAY_AMT2	30000.0	5921.163500	23040.870402	0.0	833.00	2009.0	5000.00	1684259.0
PAY_AMT3	30000.0	5225.681500	17606.961470	0.0	390.00	1800.0	4505.00	896040.0
PAY_AMT4	30000.0	4826.076867	15666.159744	0.0	296.00	1500.0	4013.25	621000.0
PAY_AMT5	30000.0	4799.387633	15278.305679	0.0	252.50	1500.0	4031.50	426529.0
PAY_AMT6	30000.0	5215.502567	17777.465775	0.0	117.75	1500.0	4000.00	528666.0
default.payment.next.month	30000.0	0.221200	0.415062	0.0	0.00	0.0	0.00	1.0

Figure 3.4 Summary of Credit Card Default Dataset

There were no missing data points but several key points to highlight. The EDUCATION feature with the label '0' is undocumented and the labels '5' and '6' are categorised. The feature MARRIAGE with the label '0' is undocumented. All categories have the label '-2'. As they document the numbers of months of delayed payment, a payment paid duly should be labelled as '0' - as would every negative value.

Based on the target variable, 22.12% of the dataset population are expected to default (Fig. 3.6).

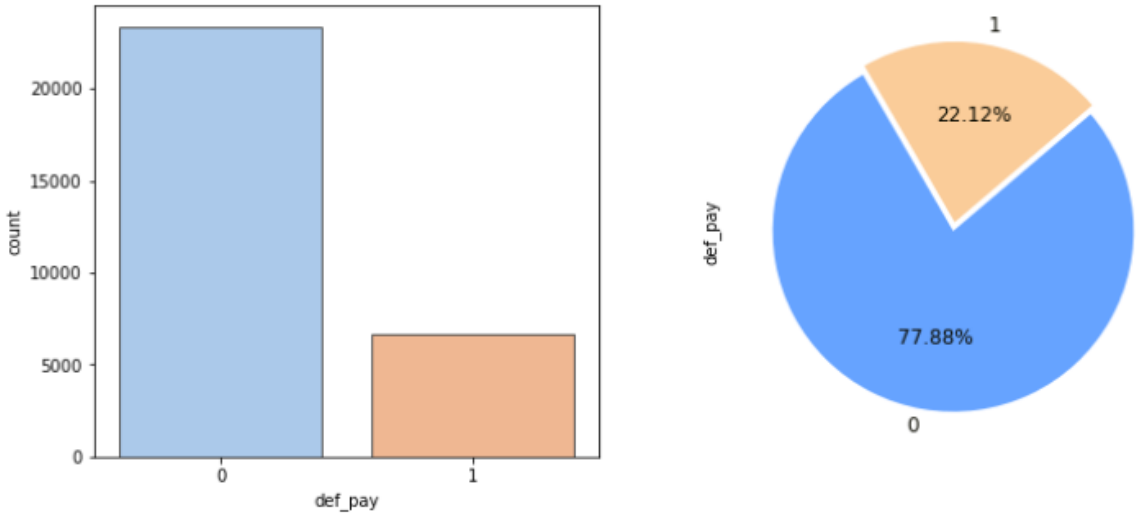

Fig 3.5 Credit Card Default Target Variable

Considering that roughly 22% of total customers will default and there are a higher number of female customers to male customers, men are 4% more likely to default in the next month.

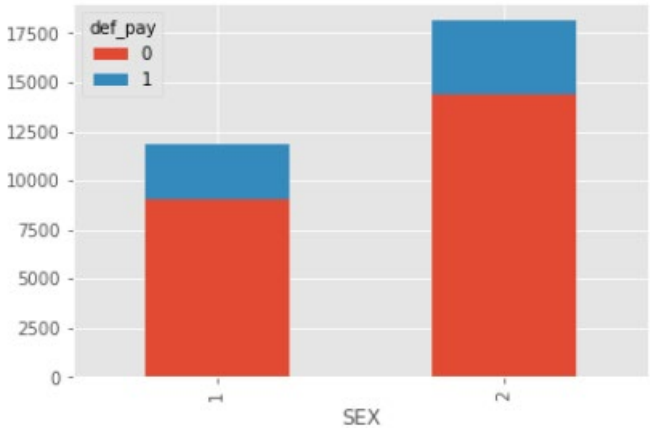

Figure 3.6 Distribution of Gender (Credit Card Default)

The other sensitive attribute that this paper looks to explore is the MARRIAGE attribute and with reference to the probability of default, there is a high chance of both single and divorced customers defaulting.

def_pay	0	1	perc
MARRIAGE			
1	10453	3206	0.234717
2	12623	3341	0.209283
3	288	89	0.236074

Figure 3.7 Distribution of Marriage Feature (Credit Card Default)

In Section 3.3 Feature Engineering, the SEX and MARRIAGE features will be combined to be evaluated as a whole to explore the relationship of sensitive attributes and the target value and determine any presence of bias in the results.

3.2.2 Diabetes Readmission Dataset

The Figure 3.8 below shows a sample of the Diabetes Readmission dataset.

	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	...	citoglipton	insulin	glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone	metformin-rosiglitazone	metformin-pioglitazone	change	diabetesMed	readmitted
0	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	...	No	No	No	No	No	No	No	No	No	NO
1	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	...	No	Up	No	No	No	No	No	Ch	Yes	>30
2	64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7	2	...	No	No	No	No	No	No	No	No	Yes	NO
3	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	...	No	Up	No	No	No	No	No	Ch	Yes	NO
4	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	...	No	Steady	No	No	No	No	No	Ch	Yes	NO

5 rows × 50 columns

Figure 3.8 Sample of the Diabetes Readmission Dataset

The table below describes the 50 attributes available in the dataset.

Variable	Description
Encounter ID	Unique identifier of an encounter
Patient number	Unique identifier of a patient
Race Values	Caucasian, Asian, African American, Hispanic, and other
Gender Values	Male, female, and unknown/invalid
Age	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
Weight	Weight in pounds
Admission type	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
Discharge disposition	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
Admission source	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
Time in hospital	Integer number of days between admission and discharge
Payer code	Integer identifier corresponding to 23 distinct values, for

	example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
Medical specialty	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
Number of lab procedures	Number of lab tests performed during the encounter
Number of procedures	Numeric Number of procedures (other than lab tests) performed during the encounter
Number of medications	Number of distinct generic names administered during the encounter
Number of outpatient visits	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Number of emergency visits of the patient in the year preceding the encounter
Number of inpatient visits	Number of inpatient visits of the patient in the year preceding the encounter
Diagnosis 1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
Diagnosis 2	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
Diagnosis 3	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
Number of diagnoses	Number of diagnoses entered to the system 0%
Glucose serum test	Result Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
A1c test result	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.
Change of medications	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
Diabetes medications	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” 24 features for medications For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter,

	“down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed
Readmitted	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission

Table 3.6: A description of the Diabetes Readmission feature variables

The following Figure 3.9 shows an overall summary of the distribution of each numerical variable in the dataset.

	count	mean	std	min	25%	50%	75%	max
admission_type_id	101766.0	2.024006	1.445403	1.0	1.0	1.0	3.0	8.0
discharge_disposition_id	101766.0	3.715642	5.280166	1.0	1.0	1.0	4.0	28.0
admission_source_id	101766.0	5.754437	4.064081	1.0	1.0	7.0	7.0	25.0
time_in_hospital	101766.0	4.395987	2.985108	1.0	2.0	4.0	6.0	14.0
num_lab_procedures	101766.0	43.095641	19.674362	1.0	31.0	44.0	57.0	132.0
num_procedures	101766.0	1.339730	1.705807	0.0	0.0	1.0	2.0	6.0
num_medications	101766.0	16.021844	8.127566	1.0	10.0	15.0	20.0	81.0
number_outpatient	101766.0	0.369357	1.267265	0.0	0.0	0.0	0.0	42.0
number_emergency	101766.0	0.197836	0.930472	0.0	0.0	0.0	0.0	76.0
number_inpatient	101766.0	0.635566	1.262863	0.0	0.0	0.0	1.0	21.0
number_diagnoses	101766.0	7.422607	1.933600	1.0	6.0	8.0	9.0	16.0
readmitted	101766.0	0.111599	0.314874	0.0	0.0	0.0	0.0	1.0

Figure 3.9 Summary of the Diabetes Readmission Dataset

The following Figure 4.0 shows an overall summary of the distribution of each variable in the dataset;

Out[55]:

Variable	#_Total_Value	#_Total_Missing_Value	%_Missing_Value_Rate	Data_Type	Unique_Value	Total_Unique_Value
race	101766	2273	0.0223	object	[Caucasian, AfricanAmerican, nan, Other, Asian...	6
diag_3	101766	1423	0.0140	object	[nan, 255, V27, 403, 250, V45, 38, 486, 996, 1...	790
diag_2	101766	358	0.0035	object	[nan, 250.01, 250, 250.43, 157, 411, 492, 427,...	749
diag_1	101766	21	0.0002	object	[250.83, 276, 648, 8, 197, 414, 428, 398, 434,...	717
miglitol	101766	0	0.0000	object	[No, Steady, Down, Up]	4
glipizide	101766	0	0.0000	object	[No, Steady, Up, Down]	4
glyburide	101766	0	0.0000	object	[No, Steady, Up, Down]	4
tolbutamide	101766	0	0.0000	object	[No, Steady]	2
pioglitazone	101766	0	0.0000	object	[No, Steady, Up, Down]	4
rosiglitazone	101766	0	0.0000	object	[No, Steady, Up, Down]	4
acarbose	101766	0	0.0000	object	[No, Steady, Up, Down]	4
troglitazone	101766	0	0.0000	object	[No, Steady]	2
glimepiride	101766	0	0.0000	object	[No, Steady, Down, Up]	4
tolazamide	101766	0	0.0000	object	[No, Steady, Up]	3
insulin	101766	0	0.0000	object	[No, Up, Steady, Down]	4
glyburide-metformin	101766	0	0.0000	object	[No, Steady, Down, Up]	4
glipizide-metformin	101766	0	0.0000	object	[No, Steady]	2
glimepiride-pioglitazone	101766	0	0.0000	object	[No, Steady]	2
metformin-rosiglitazone	101766	0	0.0000	object	[No, Steady]	2
metformin-pioglitazone	101766	0	0.0000	object	[No, Steady]	2
change	101766	0	0.0000	object	[No, Ch]	2
diabetesMed	101766	0	0.0000	object	[No, Yes]	2
acetohexamide	101766	0	0.0000	object	[No, Steady]	2
nateglinide	101766	0	0.0000	object	[No, Steady, Down, Up]	4
chlorpropamide	101766	0	0.0000	object	[No, Steady, Down, Up]	4
num_medications	101766	0	0.0000	int64	[1, 18, 13, 16, 8, 21, 12, 28, 17, 11, 15, 31,...	75
age	101766	0	0.0000	object	[[0-10), [10-20), [20-30), [30-40), [40-50), [...	10
admission_type_id	101766	0	0.0000	int64	[6, 1, 2, 3, 4, 5, 8, 7]	8
discharge_disposition_id	101766	0	0.0000	int64	[25, 1, 3, 6, 2, 5, 11, 7, 10, 4, 14, 18, 8, 1...	26
admission_source_id	101766	0	0.0000	int64	[1, 7, 2, 4, 5, 6, 20, 3, 17, 8, 9, 14, 10, 22...	17
time_in_hospital	101766	0	0.0000	int64	[1, 3, 2, 4, 5, 13, 12, 9, 7, 10, 6, 11, 8, 14]	14
num_lab_procedures	101766	0	0.0000	int64	[41, 59, 11, 44, 51, 31, 70, 73, 68, 33, 47, 6...	118
num_procedures	101766	0	0.0000	int64	[0, 5, 1, 6, 2, 3, 4]	7
number_outpatient	101766	0	0.0000	int64	[0, 2, 1, 5, 7, 9, 3, 8, 4, 12, 11, 6, 20, 15,...	39
gender	101766	0	0.0000	object	[Female, Male, Unknown/Invalid]	3
number_emergency	101766	0	0.0000	int64	[0, 1, 2, 4, 3, 9, 5, 7, 6, 8, 22, 25, 10, 13,...	33
number_inpatient	101766	0	0.0000	int64	[0, 1, 2, 3, 6, 5, 4, 7, 8, 9, 15, 10, 11, 14,...	21
number_diagnoses	101766	0	0.0000	int64	[1, 9, 6, 7, 5, 8, 3, 4, 2, 16, 12, 13, 15, 10...	16
max_glu_serum	101766	0	0.0000	object	[None, >300, Norm, >200]	4
A1Cresult	101766	0	0.0000	object	[None, >7, >8, Norm]	4
metformin	101766	0	0.0000	object	[No, Steady, Up, Down]	4
repaglinide	101766	0	0.0000	object	[No, Up, Steady, Down]	4
readmitted	101766	0	0.0000	int64	[0, 1]	2

Figure 3.10 Distribution of Diabetes Readmission variables

Based on the target variable, 11.16% of the dataset population are expected to be readmitted for diabetes (Figure 4.1).

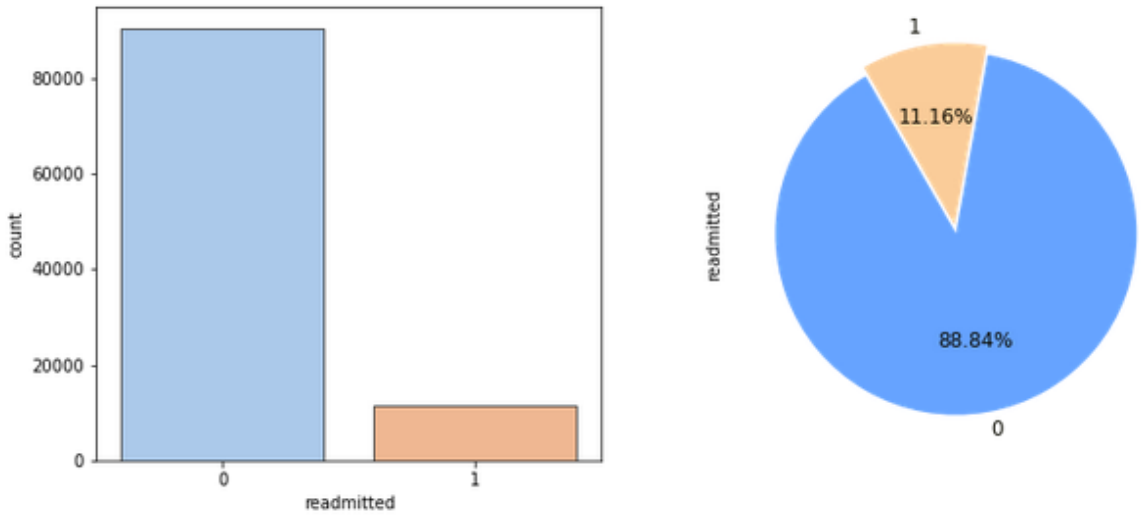

Figure 3.11 Diabetes Readmission Target Variable

From the data exploration, there is a sufficient balance between the number of entries for each gender where there are roughly 46% of male patients and 54% of female patients (Figure 4.2).

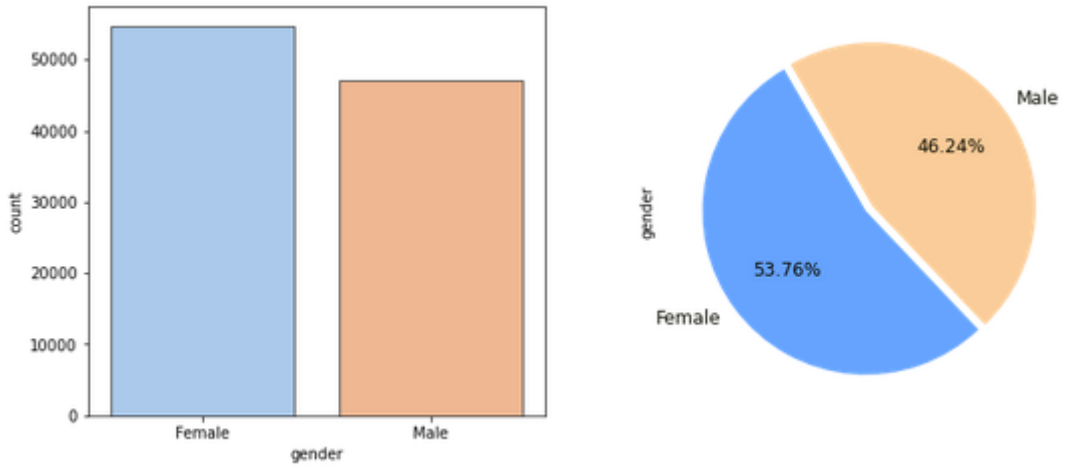

Figure 3.12 Distribution of Gender (Diabetes Readmission)

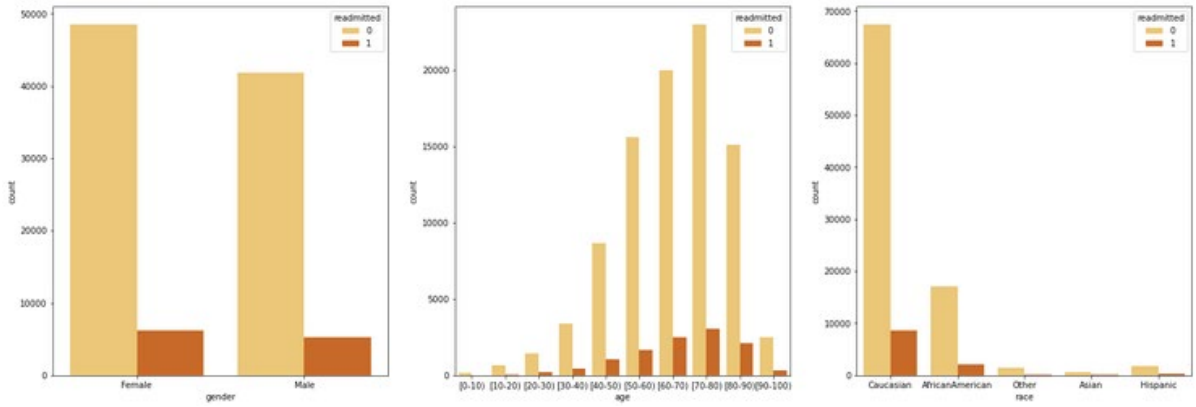

Figure 3.13 Distribution of Gender , Age & Race (Diabetes Readmission)

```
Caucasian          76452
AfricanAmerican    18772
Hispanic            2017
Other               1471
Asian                628
Name: race, dtype: int64
```

Figure 3.14 Distribution of Race (Diabetes Readmission)

Referring to Figure 3.13 and referring specifically to the designated sensitive attribute for this paper, there is a high number of Caucasians that have been admitted in comparison to other races and it is to be concluded that this is as a result of a higher number of records for said race (Figure 3.14).

3.3 Feature Engineering

This section summarises any transformation carried out on the features of each dataset.

3.3.1 Credit Card Default

As this paper explores the influence of bias as a result of sensitive attributes, feature engineering was carried out on the SEX and MARRIAGE variables to combine them into one feature. This is to enable the investigation into the reliance of the model's output on the engineered feature.

def_pay	0	1	perc
SE_MA			
1	3844	1346	0.259345
2	5068	1485	0.226614
3	103	42	0.289655
4	6609	1860	0.219625
5	7555	1856	0.197216
6	185	47	0.202586

Figure 3.15 Distribution of SE_MA variable (Credit Card Default)

As seen in Figure 3.15 above, single and divorced male customers are amongst the highest expected to default which flows from each of the features independently.

3.3.2 Diabetes Readmission

Race and gender were transformed using the *fit_transform* method as part of the Label Encoder module to scale convert the data points.

3.4 Experimental Setup

This section will describe the building of the model to train the datasets to then determine the presence of bias using a number of performance and fairness metrics.

3.3.1 Design

Figure 3.16 below shows the procedure of the training and testing of the model.

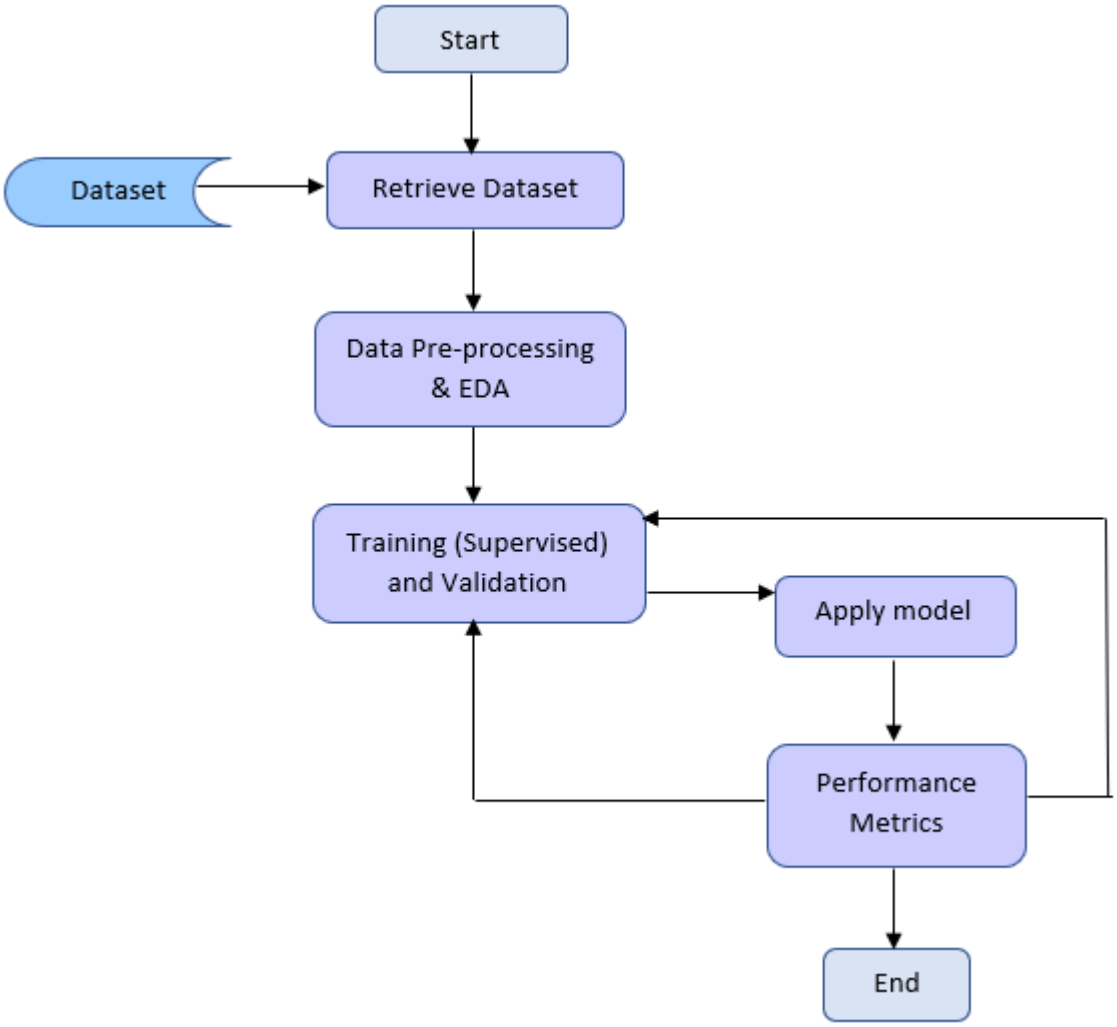

Figure 3.16 Implementation Design:

For this paper, each of the two datasets will be retrieved locally one at a time, pre-processed and passed over to the training and validation process for the model to be applied. Performance metrics will be calculated based on each model's results and the process is repeated until the objective of the paper is achieved.

3.3.2 Models

This study analyses the two datasets described in Table 3.1 using the metrics listed in Table 3.2 and Table 202. This work beings with a baseline model and hyperparameter tuning is performed to obtain the best possible model. Using optimised parameters as described in Tables 3.7 to 3.10.

Random Forest

A random forest is used to fit decision tree classifiers across various sub-samples of the dataset to control over-fitting and improves the predictive accuracy of the model using averaging [29]. Random forest is a bagging technique where the trees in random forests are run in parallel. As such, there is no interaction between these trees during building [30].

Bagging is an ensemble learning method that is can reduce the variance within a learning algorithm, particularly useful with high-dimensional data. The downside of the bagging method is a loss of interpretability of a model.

Random forests is faster to train than decision trees when working with high dimensional data. It improves bagging because by introducing splitting on a random subset of features, it decorrelates the trees. At each split of the tree, only a small subset of features are considered instead of all of it. This is important to average the variance away [31].

LGBM

LightGBM is a fast, distributed, high-performance gradient boosting framework based on the decision tree algorithm. The tree is split leaf-wise with the best fit where other boosting algorithms would split the tree by depth or level [32]. This algorithm is very fast, hence the name ‘Light’. The speedy training and reduced memory usage is the result of LightGBM’s use of histogram-based algorithms which groups continuous feature values into discrete bins.

The LightGBM algorithm was used rather than the initial XGBoost because LightGBM runs faster as the result of it splitting tree nodes one at a time in comparison to XGBoost’s splitting at a level each time.

3.3.3 Implementation

The training dataset was first divided into the training and testing datasets at a ratio of 80:20 for each of the datasets. A baseline model was calculated for a number of classification algorithms that included Naïve Bayes, K-Nearest Neighbours, XGBoost, LightGBM and Random Forest to determine the most suitable prediction model by their performance.

The Random Forest and LightGBM algorithms were chosen as a result of being some of the highest performers in accuracy. The LightGBM model was selected over the XGBoost despite their relatively close performance because of LightGBM’s ability to run at a faster pace at similar prediction performance rates.

Hyperparameter Tuning

The parameters of the estimator – Random Forest and LightGBM - were optimised by cross-validated search over parameter settings. A fixed number of parameter settings were sampled from the specified distributions. A combination of RandomizedSearchCV and GridSearchCV were used to obtain the best fit parameters, both described below.

RandomizedSearchCV

RandomizedSearchCV was initially used because there were a number of parameters to try and the training time was long. The output of the RandomizedSearchCV was used to feed into the GridSearchCV input as a smaller defined set of parameters. The parameters selected are those that maximise the score of the held-out data, according to the scoring parameter [33].

GridSearchCV

Grid Search was used to adjust the parameters in the supervised learning of the datasets to improve the generalisation performance of a model. It tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the cross-validation method. It tries every combinations of the values in the dictionary and each combination is used to evaluate the model using cross-validation. The accuracy for each combination of hyperparameters is used to find the best combination for the best performance..

Credit Card Default Dataset

Random Forest	
Parameter	Value/Quantity
bootstrap	True
max_depth	10
max_features	sqrt
min_samples_leaf	1
min_samples_split	6
n_estimators	800

Table 3.7 Random Forest Parameters (Credit Card Default)

LightGBM	
Parameter	Value/Quantity
colsample_bytree	0.5
min_child_samples	100
min_child_weight	1e-05
num_leaves	20
reg_alpha	2
reg_lambda	10
subsample	0.2

Table 3.8 LightGBM Parameters (Credit Card Default)

Diabetes Readmission Dataset

Random Forest	
Parameter	Value/Quantity
bootstrap	True
max_depth	70
max_features	Sqrt
min_samples_leaf	3
min_samples_split	6
n_estimators	500

Table 3.9 Random Forest Parameters (Diabetes Readmission)

LightGBM	
Parameter	Value/Quantity
colsample_bytree	1
min_child_samples	200
min_child_weight	0.1
num_leaves	20
reg_alpha	2
reg_lambda	10
subsample	0.2

Table 3.10 LightGBM Parameters (Credit Card Default)

The final models based on the above parameters were then implemented for each dataset and their performance and fairness metrics quantified which will be discussed in Chapter 4: Results and Discussion.

3.4 Summary

This chapter has summarised the distribution of the data, any data cleaning and transformation performed for the data to be fit for the modelling algorithms selected. As the research question surrounding this paper is how bias arises from the training dataset – specifically based on sensitive attributes – data transformation was carried out on the chosen sensitive attributes of each dataset so that their influence on the mode's outputs can be interpreted suitably to meet the objective of this paper.

Chapter 4

Results and Discussion

In this section, the results of the application of the proposed methodology on each dataset will be summarised and discussed.

4.1 Credit Card Default Dataset

The following subsections will describe the performance metrics of the optimal parameters (column highlighted) and briefly explain the feature importances based on the included feature of the module.

4.1.1 Performance Metrics

Random Forest			
	Baseline	RandomisedSearchCV	GridSearchCV
Accuracy	0.809	0.814	0.814
AUC	0.649	0.648	0.649
Recall	0.363	0.353	0.354
Precision	0.608	0.632	0.635
F1-Score	0.455	0.453	0.455
Specificity	0.934	0.942	0.943

Table 4.1 Random Forest Performance Metrics (Credit Card Default)

As seen in the GridSearchCV column, despite there being quite a high accuracy, the recall value is quite low in comparison where only 35.4% of the True Positives were correctly predicted. However, specificity was quite high where the model was able to correctly predict True Negatives for each category.

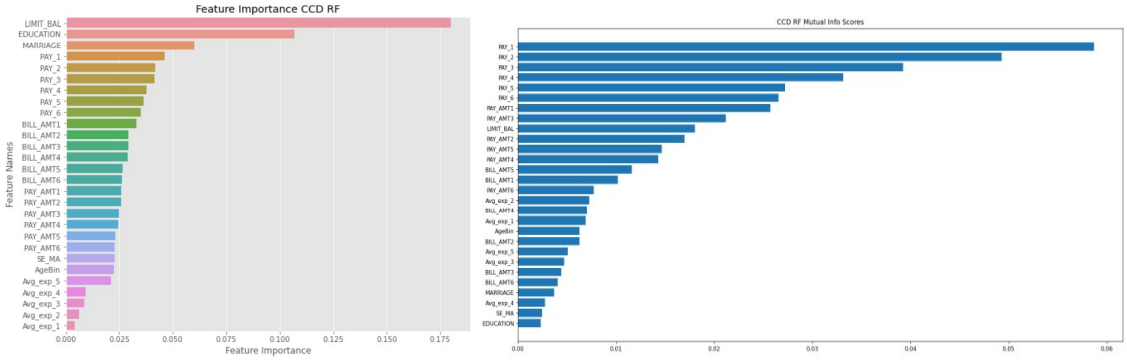

Figure 4.1 Random Forest Feature Importance (Credit Card Default)

There is a difference in the feature that is highest ranked in importance when comparing the Feature Importance method included with the module and when calculating the Mutual Information score. A limitation here and with the rest of the interpretation of feature relevance scoring was that the limited technical knowledge hindered obtaining an explanation as to why there was a difference in the feature relevance scoring.

However, in both the Feature Importance and Mutual Information scoring, it can be noted that the designated sensitive information does not play a big role in the model. The question arises then if this is sufficient to conclude that the sensitive attribute does not play a major role in the output despite there being such a low recall value.

LightGBM			
	Baseline	RandomisedSearchCV	GridSearchCV
Accuracy	0.814	0.817	0.817
AUC	0.654	0.655	0.656
Recall	0.369	0.366	0.369
Precision	0.628	0.644	0.642
F1-Score	0.464	0.466	0.469
Specificity	0.939	0.943	0.942

Table 4.2 LightGBM Performance Metrics (Credit Card Default)

As seen in the GridSearchCV column, despite there being quite a high accuracy, the recall value is still quite low in comparison where only 36.9% of the True Positives were correctly predicted. However, specificity was quite high where the model was able to correctly predict True Negatives for each category.

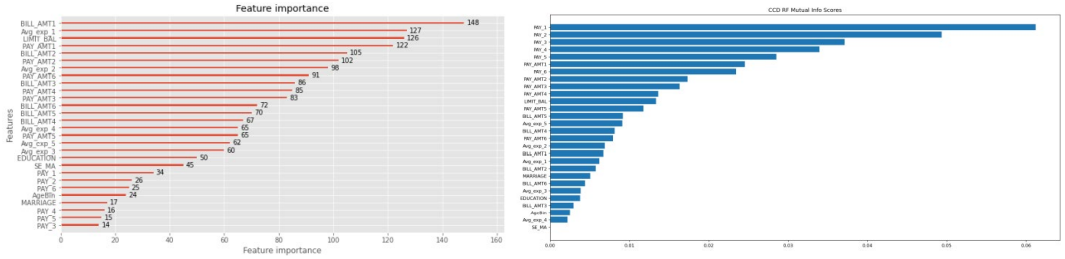

Figure 4.2 LightGBM Feature Importance (Credit Card Default)

There is again a difference in the feature that is highest ranked in importance when comparing the Feature Importance method included with the module and when calculating the Mutual Information score. However, in both the Feature Importance and Mutual Information scoring, the designated sensitive information still does not play a big role in the model.

4.1.2 Fairness Metrics

4.1.2.1 Random Forest

Global Interpretability

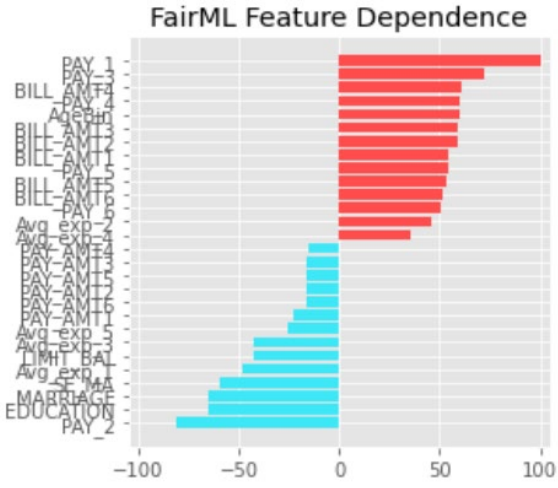

Figure 4.3 Random Forest FairML (Credit Card Default)

The FairML Feature Dependence is in-line with the Mutual Information methods results but again shows that the sensitive attributes of gender and marital status do not impact the overall model outcome.

Local Interpretability

At the local level of the mode, one input entry was chosen at random to isolate the features impacting the outcome of the entry.

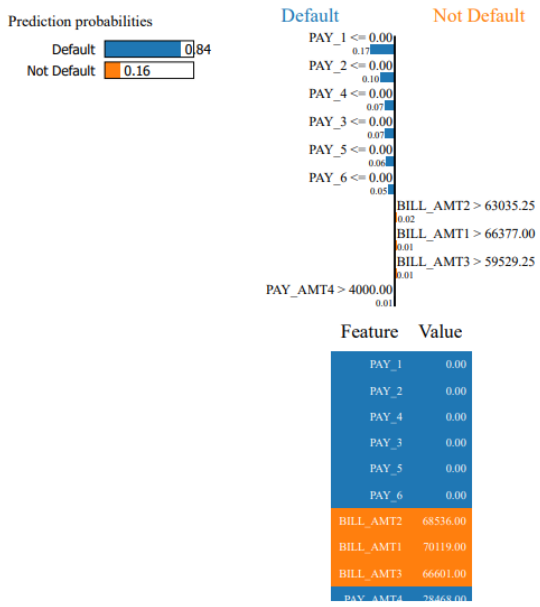

Figure 4.4 Random Forest LIME (Credit Card Default)

The LIME method for input entry #500 show a similar result as the FairML and Mutual Information where PAY_1 looks to be the feature ranking in highest in its impact on the model's output.

4.1.2.2 LightGBM

Global Interpretability

Figure 4.5 LightGBM FairML (Credit Card Default)

The FairML Feature Dependence is in-line with the Feature Importance method rather than the Mutual Information method results but again shows that the sensitive attributes of gender and marital status do not impact the overall model outcome.

Local Interpretability

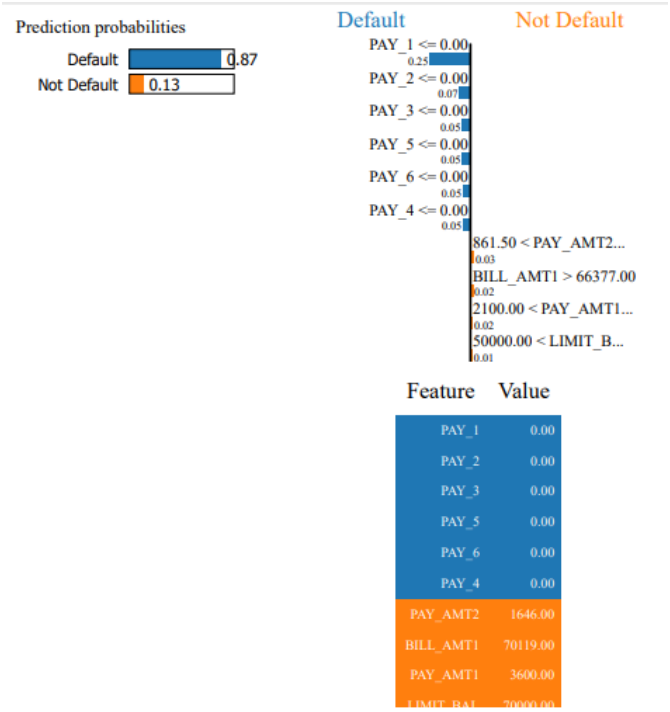

Figure 4.6 LightGBM LIME (Credit Card Default)

The LIME method for input entry #500 show a similar result as the Mutual Information where PAY_1 looks to be the feature ranking in highest in its impact on the model's output.

4.2 Diabetes Readmission Dataset

4.2.1 Performance Metrics

Random Forest			
	Baseline	RandomisedSearchCV	GridSearchCV
Accuracy	0.622	0.615	0.620
AUC	0.605	0.618	0.621
Recall	0.584	0.621	0.622
Precision	0.166	0.170	0.172
F1-Score	0.258	0.267	0.269
Specificity	0.627	0.614	0.619

Table 4.3 Random Forest Performance Metrics (Diabetes Readmission)

As seen in the GridSearchCV column, there are close figures for both accuracy and the recall value. Specificity was however not as high where the model was able to correctly predict True Negatives for each category.

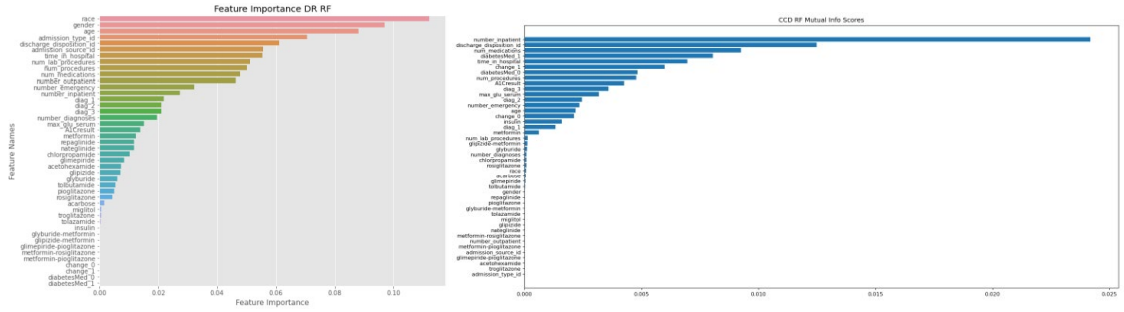

Figure 4.7 Random Forest Feature Importance (Diabetes Readmission)

It was interesting to note that the Feature Importance for this model was indicative that Race (the sensitive attribute for this dataset) has a very high impact on the mode's output though this was not similarly reflected in the Mutual Information scoring. A lack of technical expertise again became a hindrance in exploring why this was such.

LightGBM			
	Baseline	RandomisedSearchCV	GridSearchCV
Accuracy	0.629	0.630	0.631
AUC	0.618	0.621	0.623
Recall	0.605	0.610	0.613
Precision	0.173	0.174	0.175
F1-Score	0.269	0.271	0.272
Specificity	0.632	0.633	0.633

Table 4.4 LightGBM Performance Metrics (Diabetes Readmission)

As seen in the GridSearchCV column, there are close figures for both accuracy and the recall value. Specificity was however not as high where the model was able to correctly predict True Negatives for each category. These values are very closely similar to the performance of the Random Forest algorithm.

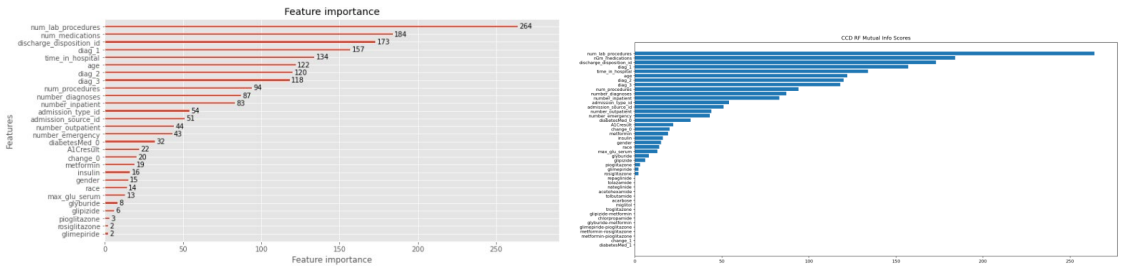

Figure 4.8 LightGBM Feature Importance (Diabetes Readmission)

The feature importance this time in the LGBM algorithm does not reflect the sensitive attribute as a high influencer of the model's outputs and this was similar to the Mutual Information scores.

4.2.2 Fairness Metrics

4.2.2.1 Random Forest

Global Interpretability

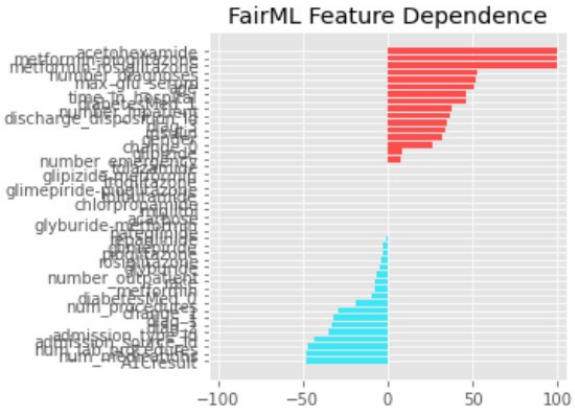

Figure 4.9 Random Forest FairML (Diabetes Readmission)

The FairML Feature Dependence again shows that the sensitive attributes of race did not impact the overall model outcome and also differs from the Feature Importance and Mutual Information scoring that had the number of lab procedures as the highest influencing variable.

Local Interpretabilitiy

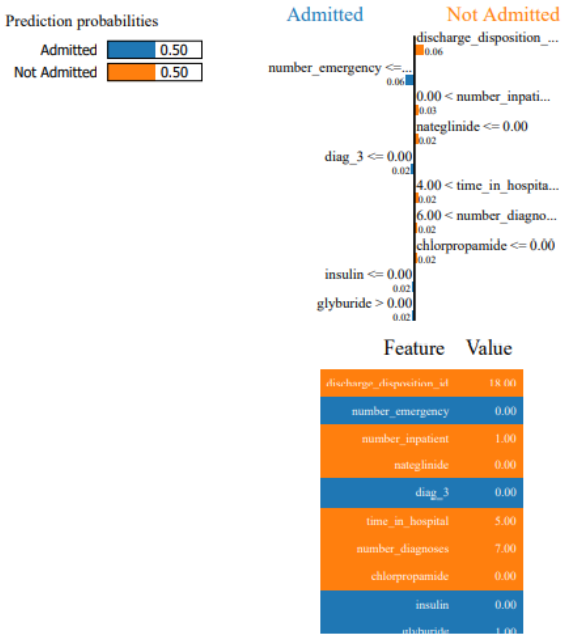

Figure 4.10 Random Forest LIME (Diabetes Readmission)

The LIME method for input entry #500 outputs a feature importance that completely differs from the previous results and there is also an equal probability between the final classification value.

4.1.2.2 LightGBM

Global Interpretability

Figure 4.11 LightGBM FairML (Diabetes Readmission)

The FairML Feature Dependence again shows that the sensitive attributes of race did not impact the overall model outcome and also differs from the Random Forest, LightGBM Feature Importance and Mutual Information.

Local Interpretability

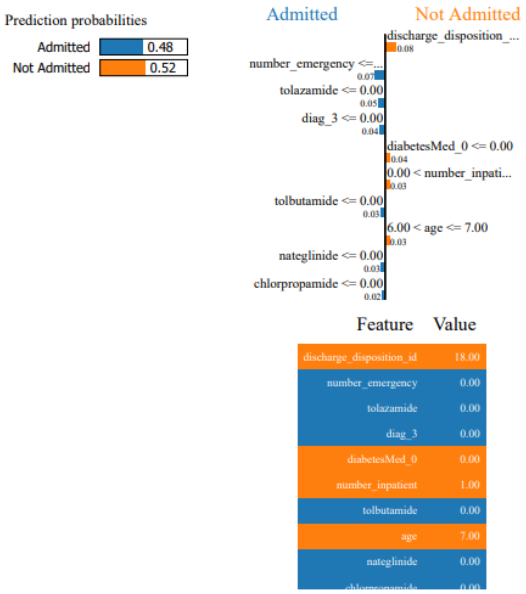

Figure 4.12 LightGBM LIME (Diabetes Readmission)

The LIME method for input entry #500 outputs a feature importance that completely differs from the previous results and there is also a very close probability between the final classification value.

4.3 Summary

The results as an entirety did not indicate that the designated sensitive attributes play a major role in the model's results although it does conclude as such. The performance metrics were not at its peak performance as a result of preserving the integrity of the original dataset as best could. The final chapter will summarise the findings and conclude future work to expand on these results.

Chapter 5

Conclusions and Future Work

In this study, an analysis was carried out aimed at identifying the performance of existing machine learning algorithms on identifying how bias influences a model's output. A framework of two machine learning algorithms (Random Forest and LightGBM) alongside a set of performance metrics and fairness measurements (FairML and LIME) was proposed to be used in combination to identify and quantify the bias in the dataset by measuring the feature importance. These two objectives were met at a sufficient level.

However, the third objective of proposing a methodology to produce the relationship between the input and output data at a local level was not met. This was due to the limitations of the timeframe and technical expertise interpreting the potential of aggregating the LIME outputs to identify specific inputs influencing the bias in the model.

Although there was not sufficient evidence from the first two objectives to prove that there was significant bias in the dataset, the feature importance of the Diabetes Readmission dataset using the LGBM model did indicate that the designated sensitive attribute (race) was a major influence on the model's output.

Thus, this paper does not conclude that it is sufficient to rule out the possibility of sensitive attributes influencing the model's output and hence the presence of bias. As this paper also argues that accuracy should not be the only performance indicator of a model, the performance metrics were not adapted to perform as high as possible in order to preserve the integrity of the original dataset as much as possible. This was to create a baseline indicator of model performance to perform bias quantification in the future.

The future work that this paper proposes is the involvement of technical expertise to interpret the results of the LIME model's outputs in order to aggregate local

inputs' feature importance to be able to identify specific entries that are influencing the overall model's performance. The future aim would be with this aggregate quantification for individual inputs, it would be possible to identify how sensitive attributes contribute to the over or under-representation in the dataset and hence, be able to improve the training dataset.

This paper as a whole addresses the legality, ethics and professionalism of the use of AI applications in the industry. The provision, utilisation and management of an automated decision-making system may be legal and professional in the industry but the key issue being addressed and is aimed at resolving is raising the awareness of the ethics of how those systems manage the data flowing through it.

References

[1] Sachan, S., Yang, JB., Xu, DL., Benavides, DE., Li, Y. (2020) An Explainable AI decision-support-system to automate loan underwriting

[2] Dou, ZY., Yu, K., Anastasoloulos, A. (2019) Investigating Meta-Learning Algorithms for Low-Resource Natural Language Understanding Tasks

[3] Wisniewski, J., Biecek, P. (2021) Fairmodels: A Flexible Tool for Bias Detection, Visualization And Mitigation

[4] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel R. (2011) Fairness Through Awareness

[5] Binns, R. (2019) On the apparent conflict between individual and group fairness

[6] Jang, T., Zheng, F., Wang X. (2021) Constructing a Fair Classifier with the Generated Fair Data

[7] Patricia, Caputo, Tuytelaars (2015) A Deeper Look at Dataset Bias

[8] Chodosh, S. Courts use algorithms to help determine sentencing, but random people get the same results

[9] Understanding bias in algorithmic design. [Blog]

Accessed: 26 April 2022, Medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

[10] COMPAS

Accessed: 26 April 2022, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[11] Corbett-Davies, Sam, Pierson, E., Feller, A., Goel, S., Huq A. (2017) Algorithmic Decision Making and the Cost of Fairness

[12] Barocas, S., Selbst, AD. (2016) Big Data's Disparate Impact

[13] Nicol, TL. (2018) Detecting racial bias in algorithms and machine learning

[14] Sydell, L. (2016) It Ain't Me, Babe: Researchers Find Flaws in Police Facial Recognition Technology

[15] Guerin, L. (2010) Disparate Impact Discrimination

[16] Zarsky, T. (2014) Understanding Discrimination in the Score Society

[17] Larson, Jeff, Mattu, S., Angwin, J. (2015) Unintended Consequences of Geographic Targeting

[18] Lee, NT., Resnick, P., Barton, G. (2020) Algorithmic bias detection and mitigation: Best Practices and Policies to Reduce Consumer Harms

[19] Zafar, Bilal, M., Martinez, IV., Rodriguez, MG., Gummadi, K. (2015) Fairness Constrains: A Mechanism for Fair Classificaition

[20] Kleinberg, J., Mullainathan, S., Raghavan, M. (2016) Inherent Trade-Offs in the Fair Determination of Risk Scores

[21] Schatz, B. (2018) AI in Government Act of 2018

[22] European Union, Digital Single Market, Ethics Guidelines for Trustworthy AI Accessed: 26 April 2022, https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[23] Bassan, P., Laslett, L. (2018) Machine Learning and Fairness in Commercial Insurance (Presentation)

[24] Hurtado, JV., Londono, L., Valada, A. (2021) From Learning to Relearning: A Framework for Diminishing Bias in Social Robot Navigation

[25] Majumder, S., Chakraborty, J., Bai, G., Stolee, K., Menzies, T. (2021) Fair Enough: Searching for Sufficient Measure of Fairness

[26] Classification – ROC and AUC (Google Developers)
Accessed: 26 April 2022, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[27] Classification - Accuracy (Google Developers)
Accessed: 26 April 2022, https://developers.google.com/machine-learning/crash-course/classification/accuracy

[28] Classification – Precision and Recall (Google Developers)
Accessed: 26 April 2022, https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

[29] Abhay, MA., Stephanie, A., Bandhyopadyay, PS., Ngo, QD., Badarla, VR. (2016) Estimating Occupancy in heterogeneous Sensor Environment

[30] Scikit-Learn: RandomForestClassifier

Accessed: 26 April 2022, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[31] Kho, J. Why Random Forest is my Favorite Machine Learning Model (2018) Accessed: 26 April 2022, https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

[32] Which Algorithm Takes the Crown Light GBM vs XGBoost (2017) Accessed: 26 April 2022, https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

[33] Scikit-Learn: RandomizedSearchCV Accessed: 26 April 2022, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[34] Kleinberg, J., Mullainathan, S., Raghavan, M. (2018) Inherent Trade-Offs in the Fair Determination of Risk Scores

[35] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, ZS., Lakkaraju, H. (2022) The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective

[36] Lima, LFFP., Ricarte, DRD., Siebra, CA. (2021) Assessing Fair Machine Learning Strategies Through a Fairness-Utility Trade-off Metric

[37] Obernmeyer, Z., Powers, B., Vogeli, C., Mullainathan S. (2019) Dissecting racial bias in an algorithm used to manage the health of populations

[38] Adebayo, J. (2016) FairML: ToolBox for Diagnosing Bias in Predictive Modeling

[39] Ribeiro, MT., Singh, S., Guestrin C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier

[40] Du, M., Yang, F., Zou, N., Hu, X. (2019) Fairness in Deep Learning: A Computational Perspective

[41] Responsible AI Principles from Microsoft
Accessed: 26 April 2022, https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

[42] Kaggle Dataset – Default of Credit Card Clients Dataset
Accessed: February 2022, https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

[43] Kaggle Dataset – Diabetes Readmission Prediction
Accessed: February 2022, https://www.kaggle.com/c/1056lab-diabetes-readmission-prediction

[44] UCI Machine Learning Repository – Default of Credit Card Clients
Accessed: February 2022, https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[45] UCI Machine Learning Repository – Diabetes Dataset
Accessed: February 2022, https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008