IJERT
ISSN : 2278 - 0181

THESIS

Thesis ID : IJERTTH0026

Applying Machine Learning to Address Gender Bias In Life Insurance Pricing

Moulinath Chakrabarty

Liverpool John Moores University, Liverpool, UK

Published By

International Journal of Engineering Research and Technology (www.ijert.org)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

APPLYING MACHINE LEARNING TO ADDRESS GENDER BIAS IN LIFE INSURANCE PRICING

MOULINATH CHAKRABARTY

FINAL THESIS REPORT

FEBRUARY 2023

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

TABLE OF CONTENTS

Published by www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

DEDICATION

This research is being dedicated to the pursuit of Knowledge.

Published by www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

ACKNOWLEDGEMENTS

The researcher acknowledges thesis supervisor Mr. Sanchit Aggarwal, entire faculty/supervisory panel at Institute of Information Technology, Bangalore, India and Liverpool John Moores University, Liverpool, UK for extending their knowledge, experience, and guidance towards the successful completion of this research work.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

ABSTRACT

Insurance is a vital part of how society mitigates risks. It has a huge influence on modern society both in terms of risk protection as well as financial wellness. It is very important to ensure the right pricing approach to attract and retain customers, thereby maintaining socio-economic balance, and keeping the industry financially viable as well.

Insurance industry has historically harbored many forms of bias such as racial bias, gender bias, financial bias etc. In order to ensure the holistic value provided by Insurance, it is needed to address any gender bias including gender definition, to provide discrimination-free pricing, as pricing is a vital cog in the wheel of Insurance. On detailed review of existing research literature, traces of gender bias were re-confirmed, and a critical gap was identified in the form of no provision for gender non-conformity. Given the tremendous pervasiveness of machine learning in Insurance, this research work focused on exploring the identification of a machine learning approach that can address this bias. It was also decided to adopt US Life Insurance as a use case for this exploration given the dominant role US plays in global Insurance and the share Life Insurance occupies within US Insurance sector.

This research work first reviewed a broad cross-section of existing research literature in this space to re-confirm gender bias including gender non-conformity, in the global Insurance industry including US Life Insurance. It was also identified that even though there are many machine learning approaches used in Insurance and many focusing on pricing as well, there seems to be a gap in the form of there being no machine learning approach providing a direct and optimal approach to address gender bias in Insurance pricing and allowing for gender non-conformity to be addressed with high attention. Based on the review of machine learning approaches currently being used/propagated, it was decided to pursue regression models as the approach for this research study. The plan followed by the research work was to build, train and test a set of regression models on a US Life Insurance dataset retrieved from public domain as well as 2 synthetic datasets – one addressing gender percentage adjustment and the other addressing gender non-conformity.

Based on the validation of the models on the 3 datasets and analysis of the evaluation metrics, Gradient Boosting Regression was identified as the machine learning model that yielded the most optimal results on the 3 datasets. This model achieved the Regression score of 0.8418 and the RMSE value of 4245.9821. Significantly, when run against the synthetic dataset specifically created to include gender non-conformity into the pricing framework, it produced the least number of errors between predicted and actual values of the predictor variable. This was truly a critical contribution made by this research work in the context of inclusivity.

Given the tremendous impact of Insurance on humanity and the dominant share US Life Insurance has in the global Insurance industry, the potential of the Gradient Boosting Regression model to be able to address gender

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

bias specifically gender non-conformity in the pricing framework is quite significant. If this model can be further validated against larger datasets and productionized in US Life Insurance industry and then expanded at a global Insurance level, it can lead to a great degree of fairness and inclusivity ensuring high customer satisfaction and profitable industry growth.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

LIST OF TABLES

LIST OF FIGURES

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

LIST OF ABBREVIATIONS

Abbreviation	Full Form
ABI	Association of British Insurers
ACC	Average of Accuracy
ADASYN	Adaptive Synthetic Sampling
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
ANS	Adaptive Neighbor Synthetic
ARR	Adjusted Risk Ratio
AUROC	Area under ROC Curve
BLSMOTE	Borderline SMOTE
BMI	Body Mass Index
CAGR	Compound Annual Growth Rate
CANN	Combined Actuarial Neural Net
CNN	Convolutional Neural Networks
DBSMOTE	Density-Based SMOTE
EC	European Commission
ECJ	European Court of Justice
EDR	Expected Deviation Ratio
EWL	European Women's League
EDA	Exploratory Data Analysis
FNN	Feedforward Neural Network
GAM	Generalized Addictive Model
GB	Gradient Boosting
GLM	Generalized Linear Model
KNN	K-Nearest Neighbor

Published by www.ijert.org

LDA	Latent Dirichlet Allocation
LDS	Linear Dynamical Systems
LJMU	Liverpool John Moores University
LMT	Logistic Model Tree
LSMC	Least-Squares Monte Carlo
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCC	Matthews Correlation Coefficient
ME	Mean Error
MPE	Mean Percentage Error
MSE	Mean-Squared Error
NHANES	National Health and Nutrition Examination Survey
NN	Neural Network
PAYD	Pay-as-you-drive
PCA	Principal Component Analysis
PCC	Probability of Correct Classification
PHI	Private Health Insurance
PNN	Probabilistic Neural Network
PoV	Point of View
PRC	Precision-Recall Curve
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
RSLS	Relocating SLS
SA	Simulated Annealing
SHI	Social Health Insurance
SLS	Safe-Level SMOTE
SMOTE	Synthetic Minority Oversampling Technique
SMOTE-ENN	SMOTE-Edited Nearest Neighbors
SMR	Standardized Mortality Ratio

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

SVM	Support Vector Machine
TNB	Transgender and Non-Binary
UBI	Usage-Based Insurance
USTS	United States Transgender Survey
XGBoost	Extreme Gradient Boosting

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

CHAPTER 1: INTRODUCTION

1.1. BACKGROUND OF THE STUDY

Insurance Industry:

Insurance industry has a huge role to play in the society. Given how humanity has progressed through many stages of evolution, one thing has stood out and that is the uncertainty of life. While humanity needs to accept the uncertainty of life, there needs to be a way to handle the risk caused by such uncertainties. Risks can be avoided, accepted, transferred, or mitigated. Insurance helps with mitigation of risk. Insurance influences our lives in many ways:

- Insurance provides life cover through Life Insurance, guaranteed income through Annuity, retirement benefits through Retirement products, auto accident cover through Auto Insurance, home protection through Home Insurance, Specialty Insurance like boats and fleets, newer insurance like drone
- The impact of Insurance on global economy can be appreciated per the data below:
  "The global insurance market grew from $5,376.92 billion in 2021 to $5,838.43 billion in 2022 at a compound annual growth rate (CAGR) of 8.6%." (Insurance Global Market Report 2022 by the Business Research Company).
  "Insurance Market size was valued at USD 4.47 Trillion in 2020 and is projected to reach USD 224.34 Trillion by 2028, growing at a CAGR of 63.13% from 2021 to 2028." (Verified Market Research, 2022)

FIGURE 1: INSURANCE MARKET SIZE AND FORECAST

- Insurance grows the capital market and generates positive financial flows – Pilijan, Cogoljevic and Pilijan (2015)
- Insurance leads to increasing general savings rate and decreasing level of unnecessary precautionary savings - Liedtke (2007)

Published by www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

On a closer look at US Insurance industry:

- Per Wikipedia, "…of the $6.861 trillion of global direct premiums written worldwide in 2021, $2.719 trillion (39.6%) were written in the United States.", which establishes the leading position played by US in the global Insurance industry.

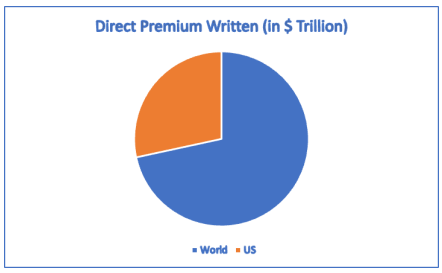


FIGURE 2: US SHARE OF GLOBAL INSURANCE INDUSTRY

- US is the largest Insurance industry in the world and Life Insurance takes the largest share of that. Hence, it is very important to focus on US Life Insurance to solve specific problems which can then be globalized. "As of 2021, the U.S. Insurance industry is worth $1.4 trillion (in written net premiums). Of that $1.4 trillion, life and annuity insurers accounted for 52%, while property and casualty approximated the remaining 48%." – Flynn (2022)

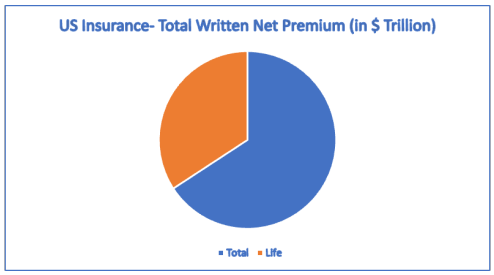


FIGURE 3: LIFE INSURANCE SHARE OF US INSURANCE INDUSTRY

Unfortunately, Insurance industry has some existing gaps and biases :

- Pension savings gap (more than USD 100 trillion) , cyber protection gap (around USD 400 billion) , healthcare protection gaps (out-of-pocket expenses about 2 % of global GDP) – Schanz (2018)
- Cognitive bias leading to poor renewal rate (7% in Nicaragua and 4% in India) - Platteau and Ontiveros (2021)
- Racial bias leading to overcharging for drivers in minority neighborhoods (30% more) – Lee (2022)
- Gender bias leading to skewed pricing for Auto Insurance (43 out of 51 US states charge women more for car insurance than men) - Bodine (2022)
- Gender rating costs women approximately USD 1 billion per year – Garrett et al (2012)

- Gender definition bias leading to lack of coverage and life hazards for specific individuals - Cohen et al (2019)

Pricing is a very critical component of Insurance, as the policy applicants will broadly accept/reject Insurance products, based on pricing. Per Carannante et al (2022), the industry (Life Insurance in particular) is going through profitability issues in the aftermath of the Covid-19 pandemic due to social and economic volatility globally. Life Insurance companies are having to adjust premiums and, in this situation, if prospective policy applicants shy away from applying because of bias-induced pricing, the companies and therefore the industry will have serious implications. Thus it is very important to have the right framework for offering discrimination-free pricing for Life Insurance products in US. Additionally, it is critical to ensure inclusivity in Insurance, and Pricing is the right starting point for that.

Machine learning in Insurance:

Insurance sits on petabytes of data. As part of digital transformation evolution, one of the initiatives to provide more personalized customer service has been to bring in machine learning and address customer segmentation strategies. Per Paruchuri (2020), Insurance is also using big data in driving analytics from unstructured data accumulated over many years in the form of scanned documents, paper documents etc. Blier-Wong et al (2020) rightly point out that one big opportunity for effective leverage of machine learning in Insurance context is the pricing (applicable across Insurance categories) or ratemaking (Property & Casualty Insurance e.g. Auto, Home etc.) where actuarial science can leverage predictive modelling based on past data about policy applicants. Machine learning can enhance all facets of Insurance such as marketing, sales, new business, underwriting, rating/premium, policy administration.

Per Rawat et al (2021), machine learning is being used across multiple use cases in Insurance industry, including: Neural networks and probabilistic neural networks for claim prediction in export credit finance, naïve bayes in claim analysis, multiple linear regression for risk prediction, feature selection techniques for aggregate Auto Insurance data analysis, random forest for Auto Insurance fraud detection, multivariate decision trees for predictive analysis of claims, gradient boosting for Insurance lost cost modeling, simulated annealing (SA) and Walsh analysis for feature selection using SVM as underlying classifier for insolvency prediction, bayesian multilayer neural network for Insurance claim fraud detection etc.

Given the specific criticality of Pricing in the Insurance industry, number of statistical and machine learning approaches are/can be considered in the industry per Jain: Poisson-Gamma GLM's, Tweedie compound Poisson GLM's and artificial neural networks. While Jain compared the models based on test data MSE, AIC, risk premium ratio analysis etc. , there seemed to be no clear choice in terms of the bias removal.

Now, as machine learning algorithms will primarily act upon the data/principles fed into them, any bias that exists in the data or the defining principles will in turn permeate into the output of the algorithms. As an example

pointed out by Matar (2022), in the City of Los Angeles, Department of Water and Power v. Manhart, the Supreme Court considered a discriminatory pension system wherein female employees were having to contribute more than males even when the benefits were the same – this was only because females were deemed to have higher life expectancy. In this case, since the data/policy itself was biased, any machine learning algorithm based on such biased data even with the intent of introducing predictive analytics, would only be able to generate discriminatory pricing/contribution output only.

In Summary:

Firstly, Insurance has a huge impact on the global economy. US (Life Insurance in particular) has a major share of the global Insurance industry. Also, within Insurance life-cycle, pricing is a very sensitive and critical component to be able to drive up business and deliver at a profitable rate. Secondly, gender bias is one of the leading biases present in the industry and permeates into pricing too. Thirdly, while machine learning is being used across multiple use cases in Insurance, there is a clear need to review existing algorithms/models in terms of efficacy to treat gender bias in pricing. Thus, considering all these points, there is a specific opportunity to further review existing work done in the field of gender bias in Insurance pricing including the use of machine learning, focus on US Life Insurance and leverage the learnings/outcomes at a global level.

1.2 PROBLEM STATEMENT

Context:

Per the statistics already highlighted, it is evident that Insurance plays a major socio-economic role in US as well as across the globe. There are, though, biases playing a part in the industry. Gender bias seems to be existing as well. Also, artificial intelligence and machine learning are being leveraged for multiple use cases in Insurance, including contract pricing.

Issues to be addressed by research:

- Confirm gender bias in global Insurance industry
- Confirm Insurance industry's treatment of gender non-conformity and fluidity
- Confirm impact of gender bias on pricing of Insurance contracts
- Review the usage of machine learning in Insurance with focus on pricing frameworks
- Suggest an optimal-fit machine learning algorithm to address gender bias in pricing of Insurance contracts

Relevance of the research:

Based on reviews of existing literature, gender bias seems to be a source of discrimination in Insurance including pricing. With the growing role of Insurance in the society as well as the pressure Insurance industry is under from a profitability perspective in the post-Covid world, Insurance industry has to grow its business more aggressively. With that being the target, Insurance cannot be biased towards a particular gender (there are other biases as well, the detailed review of which is outside the scope of this research) and demotivate members of a particular gender, as well as personas who do not want to conform to any specific gender, from signing up for Insurance or put them at a financial disadvantage. Additionally, given the immense potential of machine learning in transforming how Insurance can provide personalized services to the policy applicants/owners, the power of machine learning should be leveraged further in providing a fair pricing to everybody.

Proposed Objectives:

Research Objective #1: Review how gender bias is being addressed in US Life Insurance

- Review existing research work to confirm the existence of gender bias in global Insurance industry
- Review existing research work to specifically analyze the existence of gender bias in US Life Insurance industry, given the major share US Insurance has in global Insurance, and Life Insurance being a major share of US Insurance (as mentioned in section 1.1)
- Review relevant legal rulings and the corresponding impact on existence/treatment of gender bias e.g. ECJ ruling and ensuing European Commission guidelines from Case C- 236/09 Test-Achats (Rebert and Hoyweghen (2015)) issued with respect to removing gender-based discrimination in Insurance pricing
- Review how the industry is handling non-conformity to gender definition and gender fluidity

Research Objective #2: Suggest machine learning framework to remove gender bias in US Life Insurance pricing

- Review existing research work on the usage of machine learning in Insurance – globally including US
- Review research work on the existing/proposed usage of machine learning for addressing gender bias in Insurance pricing – globally including US. Reviews to be done of algorithms/models used, target and achieved outcomes, and any gaps w.r.t efficacy of handling gender bias
- Suggest an optimal machine learning approach to address gender bias in US Life Insurance, an approach that can be potentially extrapolated globally

1.3 SCOPE OF THE STUDY

The study has considered the following action items as in-scope:

Firstly, it will review how gender bias is being addressed in global Insurance industry including US Life Insurance industry. As part of that, it will review a general behavior of the industry towards gender bias and any bias

towards gender non-conformity. Key legal and government rulings implemented will be reviewed for outcomes in effectively addressing any such bias. Given the dominant share US has of the overall Insurance sector globally and the majority of US Insurance being Life Insurance, US Life Insurance will be considered as a key use case. Secondly, the study will focus on the efficacy of machine learning in addressing gender bias within pricing, which happens to be a key component of the Insurance life-cycle. As part of this specific item of focus, the study will broadly review existing as well as proposed machine learning frameworks and models in the context of Insurance pricing as to how those have been able to address gender bias. Finally, the study will propose an optimal machine learning algorithm to address gender bias while focusing on how to include gender non-conformity in the framework.

All other biases existing in Insurance industry, as well as review and refinement of machine learning approaches to address those biases, will not be included in the scope of this research.

1.4 SIGNIFICANCE OF THE STUDY

Insurance industry has a huge responsibility towards society. We rely on Insurance to mitigate identified risks, be prepared for unknown risks, and adopt financial planning. Thus, Insurance needs to be fair and neutral in the best interest of the policyowners. There cannot be any bias based on prejudices around religion, race, gender, and economic background. This is becoming even more significant given how Insurance industry is trying to adopt machine learning techniques. This is because production-ready machine learning algorithms will have to be trained with historical data as well as inbuilt logic, and if data and logic are carrying bias, then the algorithms will also produce biased results.

This study specifically focuses on

A. Gender bias from a US Life Insurance pricing context (as established in section 1.1, US Life Insurance has a significant share of the turnover of the global Insurance industry, thereby this particular sector is a very impactful use case in the context of this research work)
B. Effective usage of machine learning to handle gender bias from the pricing perspective

Inclusivity is a very critical component for humanity. Thus, in Insurance, we cannot have any bias towards gender, defined or undefined, when it comes to how products will be offered or priced. Insurance needs to be completely welcoming of personas who do not want to conform to any defined gender. This study, as it not only focuses on ensuring pricing to be free of gender bias, as well as how to ensure inclusivity of gender non-conformance from a pricing standpoint, is very critical in making sure that global Insurance industry becomes more transparent and inclusive. Also, given how pervasive machine learning has become throughout Insurance, the study is also very strategic from the perspective of using the unlimited potential of machine learning towards ensuring that the industry is not only free from gender bias, but even for all other biases, as the framework to be proposed will be able to expand into pricing frameworks used for other types of Insurance such as Personal Insurance, Commercial Insurance, Specialty Insurance etc. Adopting US Life Insurance as a use case for this study is tactically important

given the financial impact made by this particular sector. The proposed machine learning framework will also be extensible at a global level.

1.5 STRUCTURE OF THE STUDY

The study broadly encompasses 3 chapters:

a. Chapter 1- Introduction. In this chapter, the key sections covered are–
Section 1.1- background of the study, Section 1.2- problem statement under the purview of the study, Section 1.3- scope of the study, Section 1.4- significance of the study, Section 1.5- structure of the study.

b. Chapter 2- Literature review. In this chapter, the key sections covered are-
Section 2.1- introduction to the literature review process, Section 2.2- review of the approach towards gender bias, Section 2.3- review of the industry's specific treatment of gender non-conformity and gender fluidity, Section 2.4- review of the industry's currently implemented/proposed machine learning frameworks/models to address gender bias in pricing, Section 2.5- summary of the literature review.

c. Chapter 3- Research methodology. In this chapter, the key sections covered are-
Section 3.1 - introduction to research methodology, Section 3.2- research process followed, including sub-Section 3.2.1- data selection, sub-section 3.2.2- data pre-processing, sub-section 3.2.3- data visualization, sub-section 3.2.4- solution design, sub-section 3.2.5- model validation, sub-section 3.2.6- model evaluation, sub-section 3.2.7- inferences, sub-section 3.2.8- summary and next steps.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

As part of the research work, existing research literature has been reviewed on two aspects:

A. How does the global Insurance industry approach gender bias, including gender non-conformity and gender fluidity from a contract pricing perspective?

The study has covered the observations w.r.t Insurance industry approaches followed in North America, Europe, Africa, Latin America, and Asia. The study has also covered a broad spectrum across Insurance sectors such as Auto, Health and Life Insurance- the intent being to first understand the broader approach before delving into

the specific area of interest in Life Insurance in US. The study has reviewed the industry approach towards gender bias as well as, has looked at how gender non-conformity, gender fluidity and gender affirmation procedures are being treated by the industry.

B. How have the machine learning approaches, in general, considered a treatment of gender bias into contract pricing frameworks?

In addition to reviewing multiple research work for analysis of the existence of gender bias in global Insurance industry from a contract pricing perspective, this research paper has also reviewed how the Insurance industry is using machine learning in the context of contract pricing and how that is influencing the treatment of gender bias from a pricing perspective. Review has been done of some broader theory discourses as well as studies focusing on specific machine learning approaches.

2.2 REVIEW OF THE APPROACH TOWARDS GENDER BIAS

The study of the German Health Insurance industry by Huang and Salm (2019) has pointed out that the unisex mandate, formalized to inhibit the gender as a determining factor for Insurance premiums, leads to a premium disparity across genders. This also leads to a risk variance across the Social Health Insurance (SHI) and Private Health Insurance (PHI) sectors. Fusco and Porrini (2020) have done a detailed study of the effect of the ban on using gender as a pricing determinant for Italian Auto Insurance. As part of their study, they have analyzed data before and after the ban for a period between 2011 and 2014, in order to measure the impact of the gender variable on the Auto Insurance premiums. They have considered 7 variables including gender and ran 4 regression models to study the effect of gender discrimination on pre-ban prices, post-ban prices, prices on entire period of data and prices on entire period of data with consideration of a dummy variable to assess the impact of the ban. For all the 4 regression models, they have studied the adjusted R-squared values. Based on their study, one of their observations has been that there is an impact of the gender ban on pricing that has to be analyzed in the broader context of the variable framework used for pricing. Arelly and Montserrat (2013) have done a study of the gender effect on mortality as part of Insurance pricing framework. As part of their study, they have applied different mortality models like the Lee-Carter model, the Renshaw-Haberman model, the Currie model and the Cairns, Blake and Dowd model on general Mexico population data for the period 1990-2009 and have compared the results with unisex mortality tables used by Mexican Insurance companies. They have also inspected their findings against the Brass-type relational model on Swiss population data. Their final conclusion has been that the mortality tables used by the Insurance companies are favoring male population. This has been based on their observation that the main gap between male and female population in the context of mortality risk for the Mexican population occurs in the age interval of 20 to 50, which is also where the highest

policyholder percentage occurs. Abaichi (2018) has done a study of the factors influencing Life Insurance pricing in Kenya, considering Icea Lion Life Assurance Company as a case in point. The company personnel sample population have taken a survey to first determine if demographic factors do impact insurance pricing. A five-point Likert scale has been leveraged to capture the information. The study has brought out an observation that there is an impact on pricing caused by the gender, and this is truer in ordinary life products than group life or credit-type products. The study though has seemed to suggest that at a broader level, gender is used in a more complex way for pricing Auto Insurance than Life Insurance, thereby possibly underlining the regional specificity of the Kenyan Insurance market. Rebert and Hoyweghen (2015) have examined the ECJ (European Court of Justice) ruling of 2012 in the Test-Achats case regarding the usage of gender in Insurance pricing. They have analyzed documents from EC (European Commission) and ECJ; groups concerned with equality such as EWL (European Women's League) and leading Insurance companies of Europe, primarily focusing on gender in the context of Life Insurance in Europe. One of their primary findings has been that the ECJ rulings of 2012 seem to be in harmony with the circumstantial permission provided to gender-based practices in Insurance- this has clearly showcased that there has been a loophole in the rulings based on which, Insurers can work around the gender nuance, and rather than effecting fundamental changes to how gender is treated in underwriting, merely minimizes the causal effect of such an approach. They have also suggested that it is unfair to apply group-level data on individual applicants which is further skewing the framework/outcome away from certain genders. Their final observation has been to contextualize gender tables with lifestyle factors thereby moving the needle away from a category bias to a broader set of impact-making choices. Brodolini, Calafa and Bonardi (2011) have brought out some critical perspectives in their study of the use of gender in Insurance pricing- A. There is disparity on the implementation of the ECJ rulings of 2012 given how one set of European countries like Denmark, France, Holland and others use unisex data while another set of European countries like Belgium, Germany, Italy and others use gender-related data; B. Term Life Insurances seem to be favoring women over men while private pensions schemes seem to be doing the opposite; C. Defined-benefits schemes seem to be using gender as an actuarial factor on an immediate basis while defined-contributions schemes seem to be waiting for the payout point. Thus, their overall findings have leaned towards gaps in the way gender influences Life Insurance and Retirement savings instruments from a pricing/payout perspective. Lindholm et al (2022) have inspected the proxy discrimination or indirect discrimination from the perspective of protected variables like gender. Their approach has been to develop a statistical model that would be free from discrimination by developing the response variable (price) from the non-protected variables. They have developed a regression model to come up with a best-estimate price for a randomly chosen Insurance policy from an Insurance portfolio, and this price has been conditionally distributed based on vectors representing non-protected as well as protected variables. While they have first illustrated that the default fairness through unawareness concept is inherently while indirectly discriminatory because there is statistical dependence between the protected and unprotected variables, their own model has relied on the response variable (price) being fully described by non-protected variables and not much additional information being dependent on protected variables like gender. However, when they have subjected their model to the group fairness axioms, their model has not satisfied the independence axiom, the

separation axiom, and the sufficiency axioms, as for all of those axioms, the inter-dependence of the protected and non-protected variables has come out. They have also suggested that even some of the non-protected covariates e.g. waist-to-hip ratio, as recommended by the European Commission are also gender-dependent, thereby introducing the question of how to pre-process non-protected variables in the context of indirect relation with protected variables. Bereketoglu (2022) has undertaken a study of the impact of gender-based rating on Health Insurance pricing across EU and Asia. The approach has been to take a medical cost dataset from Spain and apply bayesian regression models through PyMC3 Markov Chain Monte Carlo algorithm on variable correlation as well as apply hierarchical models on multiple geographical regions. The overall conclusion of the study has been that while gender seems to have a lower effect on the pricing directly on non-US regions (which is, per the study, contradicting findings from various other studies in this field), the combined consideration of gender and region in a multivariate Cholesky covariance model, some impact on pricing can be observed. The study has suggested further deep-dive into similar data however at a larger and more complex scale to analyze the outcome further. Chan (2014) has studied the impact of gender-neutral pricing on the UK Life Insurance industry. The focus of the study has been on how term assurance policy premiums were affected based on gender. Chan's study has highlighted the impact that different genders would have to face post the adjustments made by the UK Life Insurers to accommodate the mandates under the ECJ rulings of 2012. Chan has found out that for two companies with different male-female compositions in the term assurance product portfolio, pre-ruling and post-ruling prices went up by less than 10% for males and up to 18% for females. Chan has studied the Rothschild-Stiglitz adverse selection model in the context of how a higher-risk gender can seek more insurance under the EJC-mandated unisex pricing system thereby resulting in a cascading effect on the demand-provision cycle. Davenport et al (2019) has professed that it may not be a significant problem to have gender bias as part of Insurance pricing which again is conflicting with some of the other research work in this field.

2.3 Review of the treatment of gender non-conformity and gender fluidity

Mackenzie (2019) has highlighted numerous examples and patterns of how gender conversion has been mis-treated by Insurance industry in terms of coverage, pricing, risk rating and even denial. With a growing percentage of human population wanting to explore and recognize their personal preferences, this attitude of discrepancy and discrimination does a major disservice to humanity in its evolving form. Cohen et al (2019), in their study, have highlighted the significant inconsistencies in how US Insurance companies are determining criteria for gender-affirming medical procedures. This is a social problem that is taking on considerable proportions. Bakko and Katari (2020) have conducted a study of Insurance denials for TNB (Transgender and Non-Binary) individuals. Based on USTS (United States Transgender Survey) data covering 27715 participants, they have applied multivariate logistic regressions and calculated ARR (Adjusted Risk Ratios) to analyze 8 different forms of denials by type of Insurance. Bivariate results have revealed significant correlation between

Insurance type and 5 of the denial categories- a key finding was that 55.8% of respondents who underwent gender-transition surgeries, were denied Insurance coverage. Multivariate results have demonstrated a key finding that self-identification as transgender led to a greater likelihood of experiencing denials from Insurance companies. Thus, the study has clearly highlighted that the self-identification of the gender leads to significant bias on behalf of the Insurance companies.

2.4 MACHINE LEARNING FRAMEWORKS AND MODELS TO HANDLE GENDER BIAS IN PRICING

Lindholm et al (2021) have advocated an actuarial approach to formulate discrimination-free prices for Insurance contracts. In the process, they have found that omitting discriminatory information may indirectly lead to price discrimination caused by the inter-dependency between variables. They have observed that in order to do a holistic assessment of discrimination-free pricing, the framework needs to access all discriminatory traits, all of which e.g. religious beliefs or gender conformity, may not be available, thereby rendering the framework somewhat sub-optimal. Their study though, has focused more on validating the proposed actuarial model, and less on how to ensure a gender-neutral pricing approach. Barry and Charpentier (2022) have documented multiple forms of bias that had historically existed in US Life Insurance industry, prior to introduction of machine learning, including race, gender, zip code, credit rating, ethnicity. They have classified biases into three types: Type 1 biases being purely mistakes or prejudice-driven; Type 2 biases reflecting on a statistical reality; however variables are non-causal; Type 3 biases reflecting on a statistical reality; however variables are causal. They then described how the use of machine learning addressed the Type 1 and type 2 biases and examined how big data can approach Type 3 biases. However, their study, while providing perspectives on how to address fairness in the light of personalization promoted by big data, has not directly proposed a machine learning algorithm to look at how discrimination-free pricing can be arrived at. Kuo and Lipton (2020) have also published their research work on explaining machine learning models in Insurance pricing. They have based their framework on how to interpret machine learning models used in Insurance pricing, which can be used by actuaries to align pricing models with actuarial standards of practice like ASOP 41. However, they have not proposed any direct approach to ensure discrimination-free pricing. Corbett-Davies and Goel (2018) have discussed about the importance of balancing fairness algorithms with real societal impacts and decoupling statistical models from policy interventions. However, while they have highlighted anti-classification fairness criterion to recommend the exclusion of a protected attribute like gender from decision matrix, they have not suggested a direct approach to remove gender bias from Insurance pricing. Davenport (2017) has explained how the inherent bias in data influences the efficacy of the algorithms and even though the study has touched on gender bias existing in Insurance, has not suggested an approach to address it. Lockhart (2022) has touched upon gender bias in Insurance pricing; however, the study has not offered a direct approach to mitigate it. Loi and Christen (2021) have analyzed the degree of trade-offs between removal of indiscriminate discrimination and predictive accuracy from an Insurance pricing perspective. They have discussed multiple types of trade-offs e.g., fairness and accuracy, multiple fairness requirements etc. They have broadly recommended that big data and machine

learning, if not used sensitively, can bring in elements of discrimination and some moral/ethical hazards. Their study has not recommended any machine learning algorithm to treat gender bias in Life Insurance. The ABI/Oxera report (2010) has provided some commentaries on the gender consideration in UK Insurance industry overall. The report has covered motor Insurance, private medical Insurance, term Life Insurance and Pension/Annuity products from the perspective of how gender is used as one of the determinants in Insurance pricing, and the observations have included: A. Gender along with behavioral patterns influence pricing of motor Insurance; B. Gender influence is limited in private medical Insurance; C. Gender in co-relation with mortality data is considered for pricing term Life Insurance as well as Pension/Annuity products. The report has argued that a gender ban can have some counter-intuitive impacts on the industry e.g. first-order redistributive impacts, product redesign impact and second-order market impacts, and has provided some suggestions on how to address these impacts in the four types of Insurance mentioned above. Overall, the report has not suggested any direct machine learning approach to address gender bias though. Cather (2021) has provided a broad discourse on how discrimination is being and can be handled in Insurance pricing frameworks. The study has covered how Aristotelian equality (similar entities should be treated likewise, while dissimilar entities need to be treated differently) is being violated by sometimes offering the same premium to different risk profiles, how the ban on using certain protected variables for pricing needs to be considered in the backdrop of industry impact caused by any such ban, and how Aristotelian equality needs to be ensured in specific Insurance categories like Personal Annuity. However, the study has not provided any clear direction on how machine learning algorithms can be reviewed in totality to determine an optimal approach for pricing Life/Annuity products while ensuring fairness from discrimination. In summary, while there are multiple frameworks that have been propagated, there does not seem to be a framework proposed to directly address the gender bias in Life Insurance pricing.

Krah, Nikolic and Korn (2018) have published a detailed study on how Least Squares Monte Carlo method can be used by Insurance companies to apply proxy modeling for calculating solvency capital requirements. They have used this method in order to apply approximation techniques for dealing with funds valuations ranging in the order of hundreds of millions to obtain full loss distributions. The method has been used by them to deal with complex requirements such as applying discounted average for European options and applying least-squares method for calculating the optimal exercise boundary for US options in a backward way. Complex as this might be, this has not proposed any direct approach to treat any bias as part of the modeling. Chancel et al (2022) have come up with a study of machine learning techniques used in risk modeling in Insurance sector. Risk modeling is particularly important for designing Insurance products for mortality/longevity, disability, critical illness, and long-term care. Interestingly, as part of their pre-processing step on US NHANES (National Health and Nutrition Examination Survey) data , they have used pseudo data tables for discretizing the data and computing risk exposures. They have suggested few specific metrics for model evaluation which will be more relevant from a survival modeling perspective, e.g. SMR (standardized mortality ratio), concordance index, Brier score, exposure weighted AUC etc. They have also explained the potential usage of several machine learning models covering discrete modeling (binomial regression, Poisson regression, random forest, LightGBM, XGBoost (, logistic GAM, CatBoost) and time-to-event modeling (Cox, Cox-Net, Cox Tree, Cox XGBoost (extreme gradient boosting),

survival tree, random survival forest). However, even though the study has broadly covered multiple machine learning models that can be used in Insurance, it has not directly suggested a machine learning approach to address any gender bias in Insurance pricing. Kotb and Ming (2021) have conducted a detailed study on using SMOTE (synthetic minority oversampling technique) models to analyze Insurance premium/price calculation in the context of renewals. As part of their study, their objective has been to document how machine learning classifiers like logistics regression, SVM (support vector machine), random forest, AdaBoost, XGBoost, neural networks etc. generate sub-optimal results if used on imbalanced Insurance data directly, whereas those classifiers will work better on Insurance data balanced by SMOTE techniques. They have taken an Insurance dataset from Egypt with 4 categorical variables and 6 continuous variables, and as part of the data preparation, they have applied feature scaling and one-hot encoding. Given the premium/price categories of defaults and non-defaults were uneven, they have applied SMOTE family methods including SMOTE, ADASYN (adaptive synthetic sampling), BLSMOTE (Borderline-SMOTE), DBSMOTE (density-Based SMOTE), ANS (adaptive neighbor synthetic), SLS (safe-level SMOTE), RSLS (relocating SLS), and hybrid techniques like SMOTE-ENN (SMOTE-edited nearest neighbors and SMOTE-Tomek. They have run the machine learning classifiers (as mentioned above) on the original imbalanced data as well the data balanced by using the SMOTE techniques, and have compared the results by inspecting the standard evaluation metrics like accuracy, sensitivity, specificity, AUC, and additionally they have also run statistical significance tests like ANOVA test and Friedman test. Their overall finding has been that the SVM technique with the SMOTE-TOMEK module has produced the best result (e.g. sensitivity improved from 3.2% to 83.84%). However, their study has not effectively addressed the problem of gender bias in pricing and has been a broader solution at the overall pricing default level. Matar (2022) has conducted a study to implement bootstrapping method on a German credit dataset and has established that the model is imbibing unfairness existing in the base data. However the study has not proposed any direct approach to address gender bias. Blier-Wong et al (2020) have published a detailed study on how machine learning is being used for pricing and reserving purposes in Property&Casualty Insurance pricing. They have highlighted machine learning techniques discussed by other researchers for a priori pricing including SVM, GAM (generalized additive models), GLM (generalized linear models), gradient boosting, multivariate decision trees; and techniques used by other researchers for a posteriori pricing e.g. regression tree credibility model. They have also highlighted neural models discussed by other researchers for Insurance pricing e.g. CANN (combined actuarial neural net), and specifically for telematics pricing e.g. K-means and CNN (convolutional neural networks). While their work has been detailed in terms of machine learning approaches used in pricing and reserving, they have not suggested a direct approach to address gender bias in Insurance pricing particularly Life Insurance. They have mentioned about anticlassification to remove the protected attributes which basically lead to proxy modelling, which is an aspect this current research work has already observed from the study done by Lindholm et al (2020). Henckaerts (2021) also has reviewed the machine learning approaches used in Insurance industry, with specific focus on telematics and UBI (usage-based Insurance). The research work has covered white-box approaches such as GLM (generalized regression models) as well as black-box models such as GBM (gradient boosting machines) and has argued that while the black-box approach provides advantages from statistical and economic perspectives, it

provides challenges from an interpretability perspective, which is important given the regulatory nature of the Insurance industry. As an optimal approach, the study has suggested using model-agnostic interpretable data-driven surrogate models to ensure performance as well as interpretability. There has been a reference to the study done by Ayuso, Perez-Marin and Guillen (2016) which has suggested that PAYD (pay-as-you-drive) framework puts the focus on behavioral patterns rather than any discriminatory variables in the context of pricing. However, both the studies conducted by Henckaerts (2021) and Ayuso, Perez-Marin and Guillen (2016) have not specifically addressed the gender discrimination aspect of Life Insurance pricing. Per Mosley and Wenman (2022), explainable AI will be critical to balance the regulatory/transparency mandates in Insurance sector, along with bringing in machine learning algorithms to eliminate rating bias. Their study has focused on model fairness and model de-biasing and has adopted a GLM on a French Auto Insurance dataset to explain how de-biasing needs to go through the steps of fair pre-processing, fair in-processing and fair post-processing. While their study has suggested a possibility that re-weighting technique renders the model more effectively de-biased in terms of removal of a protected variable, the effectiveness of their solution is somewhat in question in the absence of an original statistical measure of the bias. Their framework therefore has not addressed the gender bias problem directly. Shimao and Huang (2022) have conducted a study on the cross-effects on policyholder welfare brought about by policy regulations and fairness practices driven by machine learning algorithms. They have primarily used GLM and XGBoost for cost modeling based on certain fairness criteria- accountability, unawareness, demographic parity, and sensitivity. In parallel, they have considered the pricing regulations such as accountable pricing, price optimization ban, pricing with demographic parity and pricing with actuarial group fairness. They have run the models on a French Insurance dataset and have considered gender as the protected attribute. One key finding has been that there is an impact on pricing caused by gender gap, and another key finding has been the observed impact on gender welfare brought about by using fairness algorithms in an inefficient way. Zhou, Marecek and Shorten (2021) have conducted a study on treating biased data. Their study has followed an objective of treating subgroup fairness and instant fairness for LDS (linear dynamical systems). Their approach, among other considerations, has illustrated how gender bias is handled in Annuity payout systems through a single series trajectory for male and female annuitants in terms of mortality rates being considered as a single LDS. However, their proposed solution to the fairness bias problem in terms of using convexification has not finally illustrated how the gender bias treatment is being able to achieve effectiveness or not. Grari et al (2020) have published their study on using adversarial learning techniques to treat the fairness biases- demographic parity and equalized odds. While pre-processing and post-processing have been observed by them to act on the input or the output of a trained predictor, they have focused on penalizing during in-processing through adversarial learning. They have run their model on French motor insurance datasets and have observed that fairness increases as evident through metrics such as ACC (average of accuracy), EDR (expected deviation ratio), MSE (average of Mean Squared Error) etc. However, this study has focused more on pricing based on predictive claim likelihood which, in turn, considers gender indirectly contributing to behavioral patterns such as aggressiveness, color of cars etc. – there has been no direct approach suggested to address gender bias for Life Insurance pricing. Frees and Huang (2021) have provided a discourse on discrimination that

broadly exists in the Insurance industry (they have also discussed the discrimination brought about as an aftermath of the Covid-19 pandemic whereby, statistics influenced by the pandemic impact have seemingly impacted the industry from a proxy discrimination perspective; as well as they have discussed the discriminatory perspective of genetic testing in Life Insurance). As a proposed solution, their research has suggested adopting linear model strategies by omitting protected variables e.g. gender (however they have also pointed out that Insurers, in that case, may very likely seek to incorporate proxies in that case, which may again bring about proxy discrimination). As part of their overview of machine learning approaches to address discrimination, they have suggested (like Lindholm et al (2022), Chancel et al (2022), Mosley and Wenman (2022), Grari et al (2020)) pre-processing, in-processing and post-processing. However, their study has broadly addressed how actuaries, as part of Insurance pricing, need to balance between discrimination removal and broader connotations; so their study also has not suggested a direct machine learning approach to address gender bias per se. Lindholm et al (2022) have proposed a multi-task network approach for treating Insurance pricing of discrimination. They have used a FNN (feed-forward neural networks) to calculate the best-estimate price and then assume a suitable pricing distribution to calculate discrimination-free Insurance pricing by breaking the link between discriminatory covariates and non-discriminatory co-variates. Their study has also showed that depending on the availability of the discriminatory variables, the accuracy is higher for a multi-task vis-à-vis a plain FNN. So, the efficacy of the model is still dependent on the availability of the discriminatory information.

2.5 SUMMARY

Based on the literature review, findings can be summarized as:

- There is a gender bias existing in the global Insurance industry
    - Europe - studies done in UK, Germany and Italy corroborate this
    - Latin America - study done in Mexico corroborates this
    - Africa - study done in Kenya corroborates this
    - Broader studies done across EU and Asia corroborate this
- Specifically the industry has significant maturity to attain when it comes to treating gender fluidity, gender non-conformity and gender-affirmation. Self-identification of gender leads to bias from Insurance companies for coverage as well as pricing.
- Multiple machine learning frameworks/models have either been used in the industry or have been proposed by researchers
    - LSMC for proxy modelling
    - Classifiers like XGBoost, AdaBoost etc. complemented by SMOTE models
    - White-box approaches such as GLM and Black-Box approaches e.g. GBM
    - Neural networks like CNN and FNN
    - Model de-biasing through pre-processing, in-processing and post-processing

- LDS
- Adversarial learning

- While these approaches are technically sound and achieve broad objectives from the perspective of addressing multiple forms of discrimination in the Insurance industry, there is a need of a machine learning approach to directly address gender bias from a pricing perspective, which will address inclusivity by allowing gender non-conformity to be considered by the framework
- It is also important to balance the efficacy of any existing/proposed model with the broader impact to the industry as well an optimal alignment with policy regulations

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

CHAPTER 3: RESEARCH METHODOLOGY

3.1 INTRODUCTION

Based on the review of existing research literature on machine learning models used for treating gender bias in Insurance pricing, some observations that are significant in terms of determination of the research methodology to be followed by this study are:

Fusco and Porrini (2020) have applied regression models on Italian Auto Insurance database to establish that there is an impact of gender bias. However, there has been no solution proposed for treating the bias as such. Bereketoglu (2022) have applied bayesian regression models on Spanish medical cost dataset and there has been no clear solution proposed for directly treating gender bias. Bakko and Katari (2020) have applied multivariate logistics regressions on US Transgender data and have confirmed that self-identification of gender leads to bias, however, there has been no solution recommended. Krah, Nikolic and Korn (2018) have applied LSMC method for applying proxy modelling to address solvency capital requirements in EU as well as US markets, however there has been no solution provided for treating gender bias. Chancel et al (2022) have applied several discrete modeling and time-to-event modelling techniques on US NHANES data with an objective of investigating mortality data for risk modelling as part of Life Insurance pricing. However, this study has been more focused on mortality analysis and not on treating gender bias as part of Life Insurance pricing. Kotb and Ming (2021) have taken an Egypt Insurance dataset and have applied SMOTE techniques to illustrate the benefits of balancing originally imbalanced Insurance data in order to optimize the outcome of machine learning classifiers. However, their focus has been more on policy renewals and not on treating gender bias. Matar (2022) have applied bootstrapping method on a German credit dataset and have established the existence of bias. However, there has been no recommendation provided to directly treat gender bias. Blier-Wong et al (2020) have conducted a broad study of machine learning algorithms (neural networks, SVM etc.) which are used, or can be used for rate-making and reserving in US Property&Casualty Insurance, there has been no recommendation provided for treating gender bias in pricing. Henckaerts (2021) has highlighted the importance of considering white-box approaches such as regression models given the transparency and interpretability requirements of the global Insurance industry. The study has focused more on behavioral pattern analysis though. Zhou, Marecek and Shorten (2021) have applied LDS on Annuity payout systems to approach gender bias by considering mortality rates as a single series (gender-agnostic). However, there has been no clear evidence of the outcome being effective. Grari et al (2020) have applied adversarial learning techniques on French motor Insurance data and have observed that fairness increased by applying penalization during in-processing. However, there has been no direct approach recommended to treat gender bias in Life Insurance pricing. Lindholm et al (2022) have applied FNN to suggest best-estimate Insurance pricing, however their study has been tightly bound to the availability of discriminatory variables.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

In summary, while multiple machine learning techniques have been used on different aspects of Insurance pricing with a focus on bias treatment, there are 2 unexplored areas based on a broad review of the existing research literature in this space:

- Need for a machine learning approach to directly treat gender bias in Insurance pricing; take US Life Insurance pricing as a use case (given the major share US has in global Insurance industry and the greater share Life Insurance has within US insurance) and based on the efficacy of the model, the approach can eventually be extrapolated at a global level (not considered within the scope of this research though)
- Explore regression as a machine learning technique (given the gaps seem to be existing with the other approaches) to solve the gender bias gap including gender non-conformity

3.2 RESEARCH PROCESS

This study followed a step-by-step research process as detailed below.

3.2.1 DATA SELECTION

Insurance pricing is essentially about calculating the premium structure for policy applicants based on a risk assessment done for them. This risk assessment leads to underwriting of the risk and a rate/premium calculation. There are number of factors taken into consideration:

- Age
- Type of coverage e.g. for Auto, it can be liability, collision, comprehensive, medical payments, uninsured/underinsured motorist coverage, for Life insurance it can be term insurance, whole life insurance etc.
- Amount of coverage
- Personal information e.g. personal medical history, family medical history, job type, smoking habits, zip code
- Actuarial life tables consisting of mortality information

Given it is difficult to access production data sets from Insurance companies because of the sensitive nature of the data, a dataset from public domain was chosen which has a lesser number of columns, however these specific

Published by www.ijert.org

variables should be able to help validate the approach for ensuring discrimination-free pricing in context of gender bias removal.

The dataset used had been accessed from https://www.kaggle.com/code/mariapushkareva/medical-insurance-cost-with-linear-regression/data

It was confirmed that this dataset is available on the public domain and therefore, there is no violation of data privacy.

Dataset snapshot is given below:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.855

FIGURE 4: DATASET SNAPSHOT

The dataset had 1338 rows and 7 columns. The information captured by the columns are:

Table 1: Dataset description

Serial Number	Column Name	Significance	Value type
1	Age	Age of the participant	Integer value
2	Sex	Gender	Object (male/female)
3	BMI (body mass index)	Index of good/bad health, calculated in terms of weight in relation to height (e.g. kg/sq.m)	Float
4	Children	Number of children	Integer value
5	Smoker	Whether smokes or does not smoke	Object (yes/no)

6	Region	US geographical region based on zip code	Object (Northeast, Northwest, Southeast, Southwest)
7	Charges	Charges to be paid to insurer	Integer value (this is the target variable)

The dataset was downloaded on MacOS into a personal folder to be accessed for the model development. A Python code shell was created, and all required libraries were imported.

3.2.2 DATA PRE-PROCESSING

The dataset was checked for statistical compliance and the results were found to be:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150

FIGURE 5: STATISTICAL VIEW OF DATASET

The dataset was checked for missing data and the results were found to be:

```
age missing values:0
sex missing values:0
bmi missing values:0
children missing values:0
```

FIGURE 6: MISSING VALUES VIEW OF DATASET

So, the dataset did not have any missing values.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

The dataset was also checked for any outliers and the results were found to be:

```
age
 percentage= 0.0
 number=0
 mean= 39.20702541106129
 min=18
 max=64
bmi
 percentage= 0.672645739910314
 number=9
 mean= 30.663396860986538
 min=15.96
 max=53.13
children
 percentage= 0.0
 number=0
 mean= 1.0949177877429
 min=0
 max=5
charges
 percentage= 10.388639760837071
 number=139
 mean= 13270.422265141257
 min=1121.8739
 max=63770.42801
```

FIGURE 7: OUTLIER VIEW OF DATASET

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

So, the dataset did not have too many outliers. The range was a little more for charges, which was to be expected based on the number of children.

3.2.3 DATA VISUALIZATION

Distribution of charges on source of variables was inspected

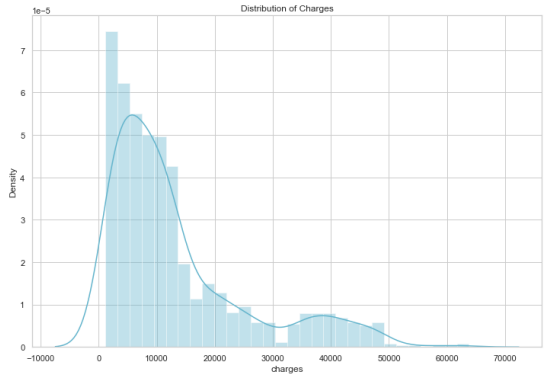


FIGURE 8: DISTRIBUTION OF CHARGES ON SOURCE VARIABLE

Evidently the distribution was found to be right-skewed, and this was because of the charges having a different data range than the source variables. In order to obtain a more normalized view, natural log was applied.

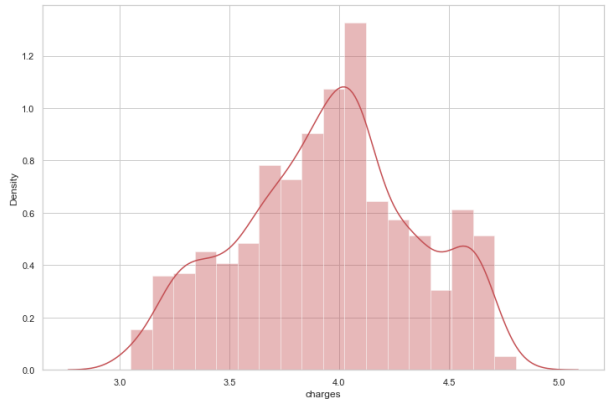


FIGURE 9: NORMALIZED VIEW OF DISTRIBUTION OF CHARGES ON SOURCE VARIABLES

Published by www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

Next, the distribution of charges against the source variables was reviewed:

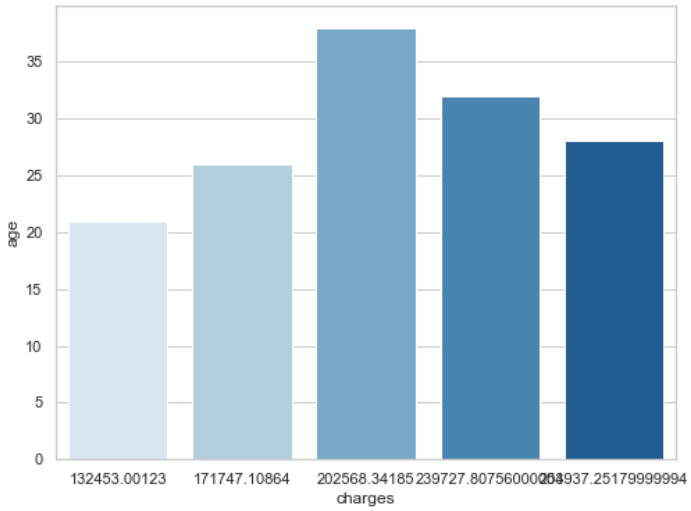


FIGURE 10: CHARGES BY AGE

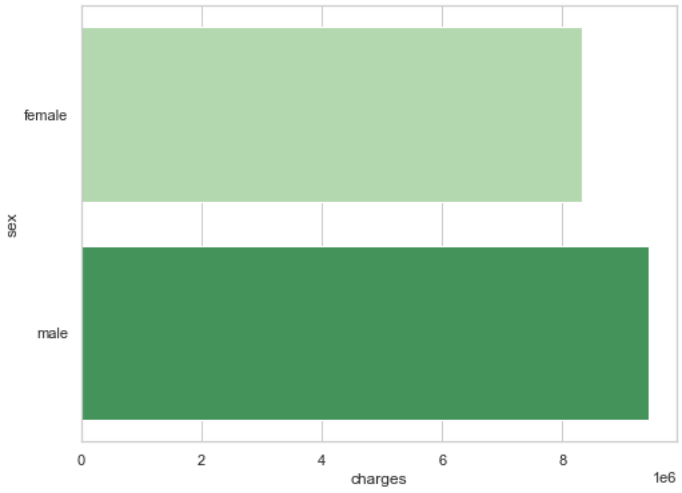


FIGURE 11: CHARGES BY GENDER

Published by www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

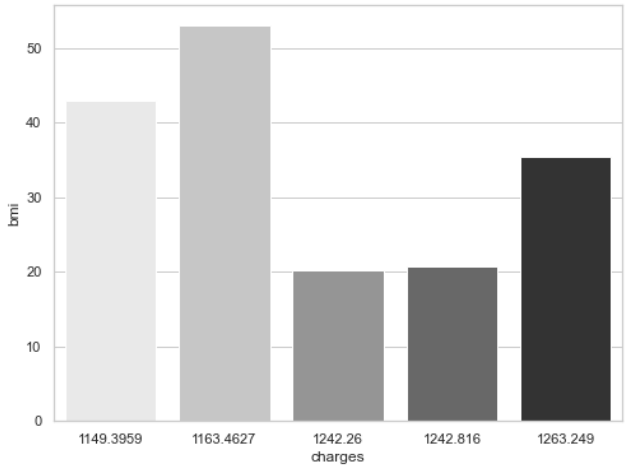


FIGURE 12: CHARGES BY BMI

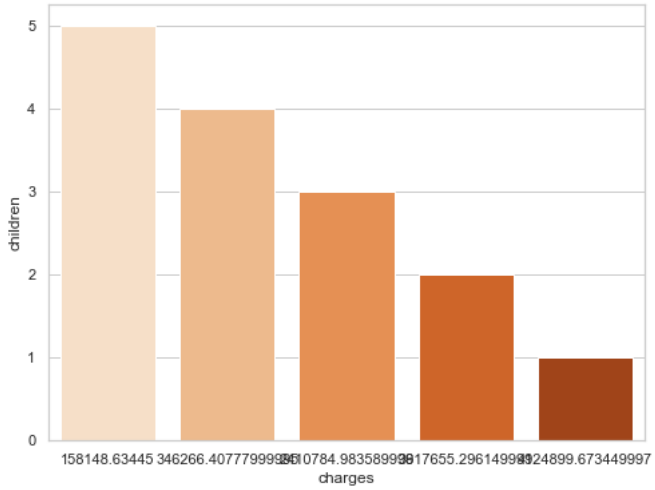


FIGURE 13: CHARGES BY CHILDREN

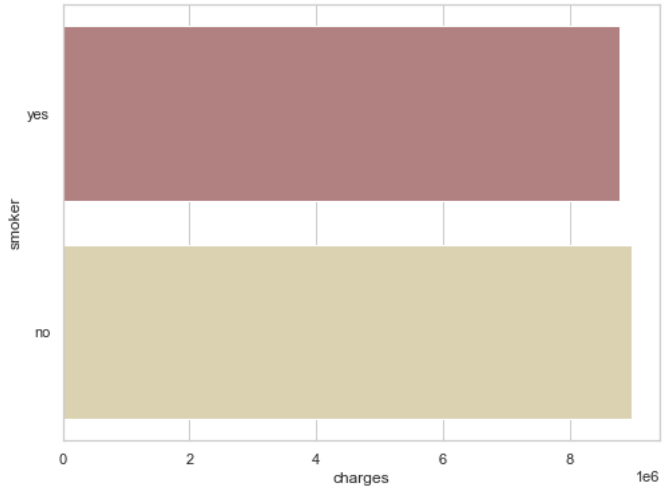


FIGURE 14: CHARGES BY SMOKING STATUS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

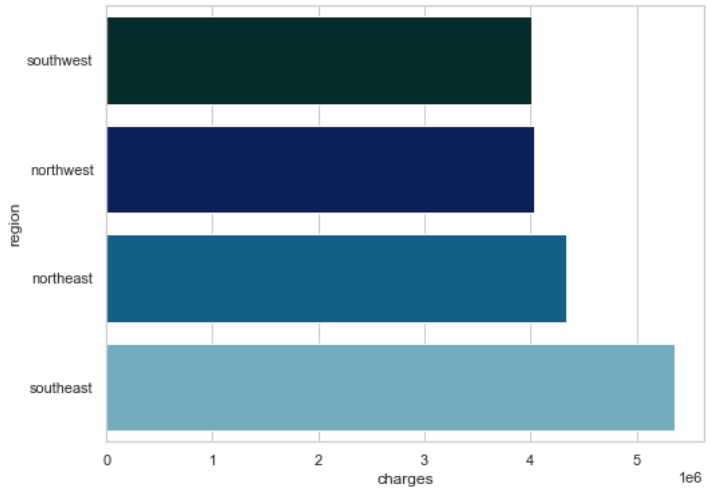


FIGURE 15: CHARGES BY REGION

Observations:

- Charges by age: Post normalization and usage of natural log, the charges variables were distributed against age in a more normalized way, with highest occurrence of charges appearing around the mean of the age band (approximately 39)
- Charges by gender: Males seemed to be paying about 20% premium more than females (in total)
- Charges by BMI : Highest charges was around mean BMI (approximately 30). This was more because the data possibly being evenly distributed across a general population sample, and the data getting normalized around mean BMI
- Charges by children : Sum of charges was highest in the cases where number of children was 1 or 2. This was because the mean is between 1 and 2 for the 'number of children' variable and with the data being normalized, the sum of charges was highest in that range
- Charges by smoking : Evenly distributed
- Charges by region : Charges were the highest for southeast region (about 20-25% more compared to the other regions). This was due to a mix of other variables e.g age, gender, smoking status coming together and influencing the regions.

Next, gender and charges were mapped w.r.t each other considering some source variables

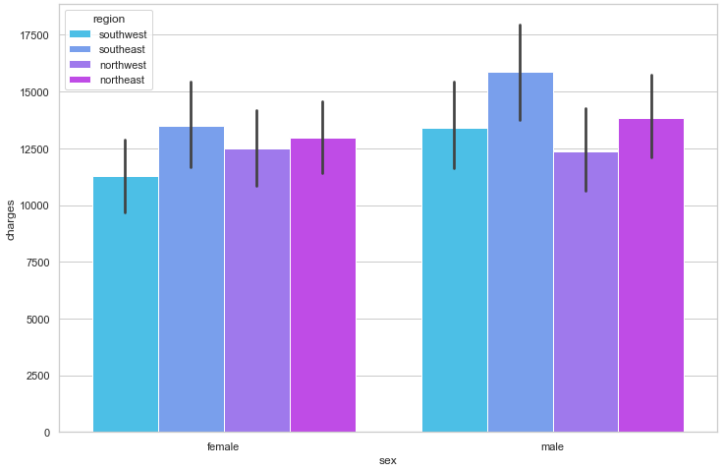


FIGURE 16: GENDER-CHARGES DISTRIBUTION FOR REGION

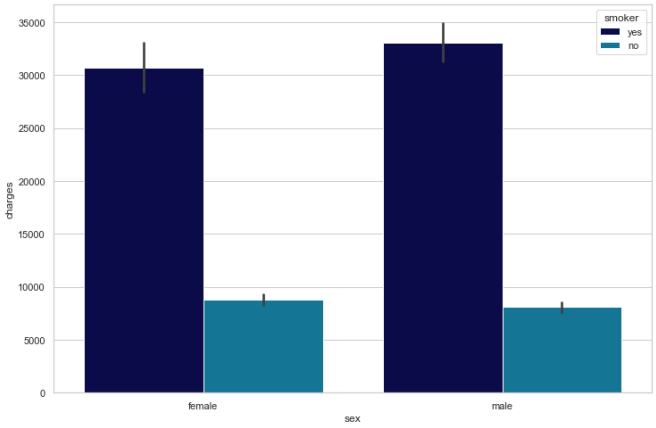


FIGURE 17: GENDER-CHARGES DISTRIBUTION FOR SMOKING STATUS

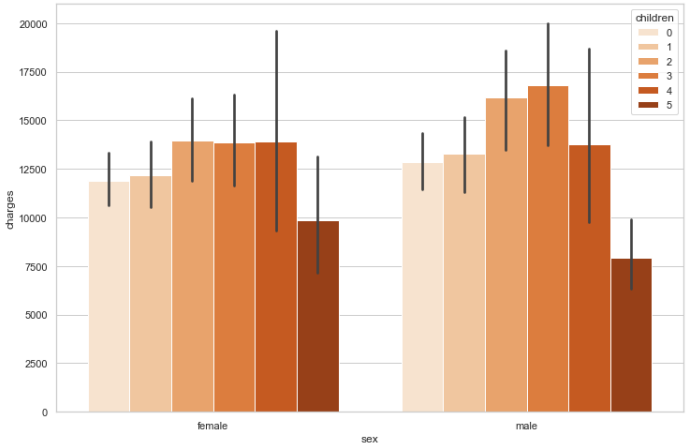


FIGURE 18: GENDER-CHARGES DISTRIBUTION FOR NUMBER OF CHILDREN

Observations:

- Southeast region had the highest charges for both males and females. This was in alignment with the previous set of observations.

- Smokers were found to pay the highest charges for both genders, with male smokers paying more. This pointed out two things: A. The correlation between smoking status and charges needed to be looked into, and B. The pricing algorithm penalizes males for smoking more as compared to females. The reason for that would need to be found out through further analysis of morbidity variables, which was outside the scope of this research.
- For number of children from 0 to 3, males were identified to be paying more than females, and for 4 and 5, it was the opposite. This would have required more in-depth of study of correlation between gender and number of children, which was outside the scope of this research.

Next, charges were analyzed by age, BMI, and children according to the gender factor

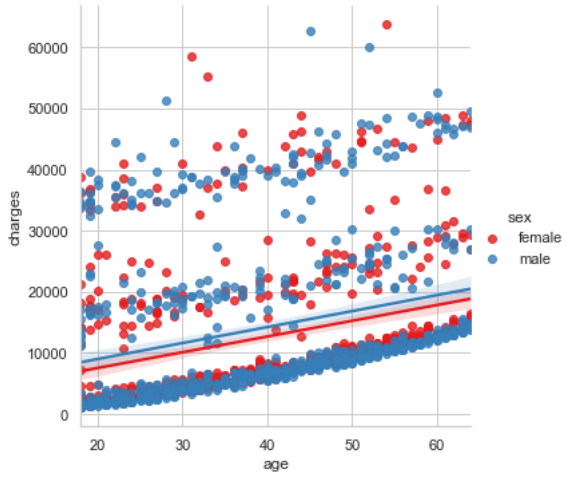


FIGURE 19: CHARGES-AGE DISTRIBUTION BY GENDER

FIGURE 20: CHARGES-BMI DISTRIBUTION BY GENDER

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

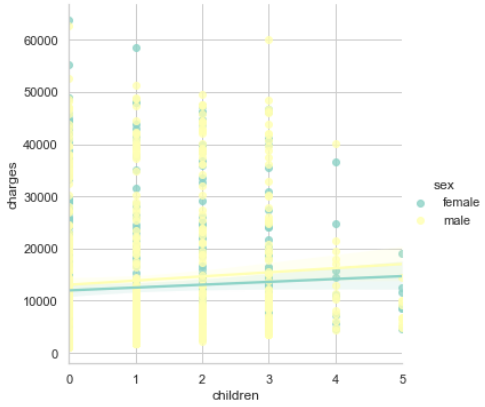


FIGURE 21: CHARGES-NUMBER OF CHILDREN DISTRIBUTION BY GENDER

Observations:

- Charges were more for males than females as age, BMI and number of children increased
- The reasons for that higher proportion for age and BMI increase would be tied into a deeper analysis of actuarial life tables for mortality and morbidity analysis, which were outside the scope of this research. The study of the correlation between the gender and the number of children was also outside the scope of this research

Next, correlation was determined among the numeric variables

FIGURE 22: CORRELATION AMONG NUMERIC VARIABLES

Observations:

- Correlation seemed to be positive among the numeric variables, however weak
- Highest co-relation was between charges and age and is 0.3, which is not significant

Next, pairplots was run for the numeric variables against gender.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

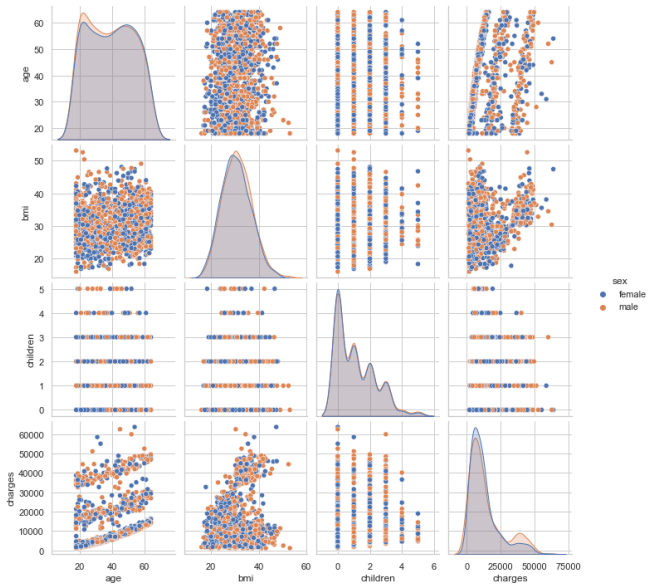


FIGURE 23: PAIRPLOTS FOR NUMERIC VARIABLES AGAINST GENDER

Observation:

- Charges were more for males as compared to females, with increase in age, BMI and number of children – this aligned with an earlier observation as well

Next, scatterplots were run for the source variables and charges, against gender.

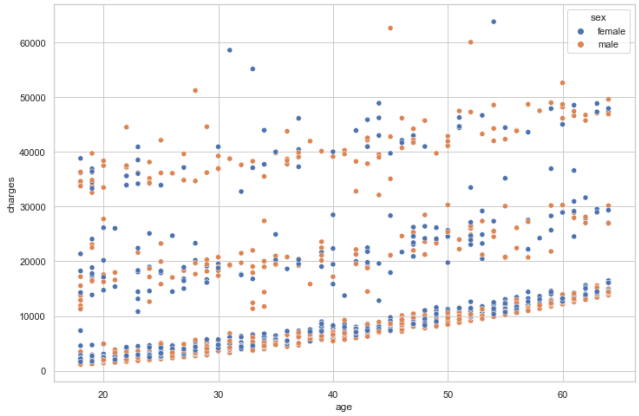


FIGURE 24: SCATTERPLOT FOR CHARGES AND AGE, AGAINST GENDER

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

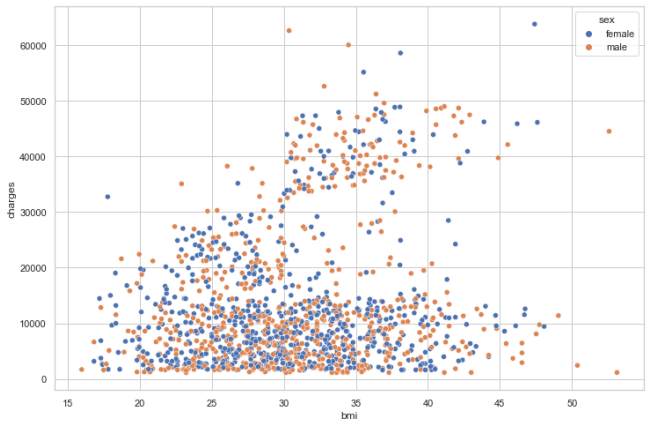


FIGURE 25: SCATTERPLOT FOR CHARGES AND BMI, AGAINST GENDER

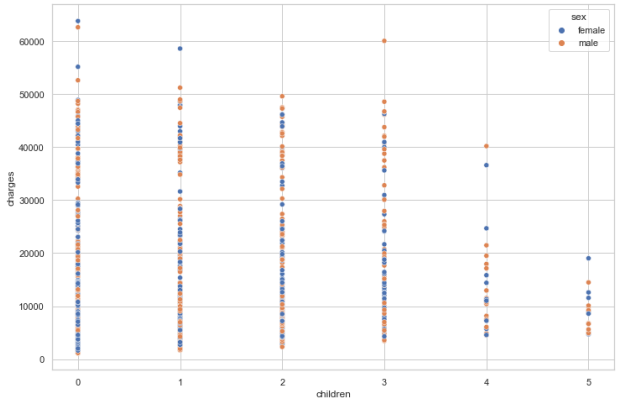


FIGURE 26: SCATTERPLOT FOR CHARGES AND NUMBER OF CHILDREN, AGAINST GENDER

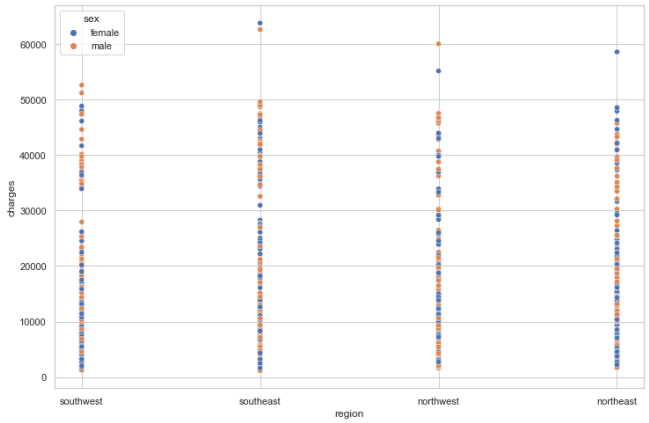


FIGURE 27: SCATTERPLOT FOR CHARGES AND REGION, AGAINST GENDER

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

FIGURE 28: SCATTERPLOT FOR CHARGES AND SMOKING STATUS, AGAINST GENDER

Observations:

- Charges-age, against gender: There were 3 bands, with some empty areas in between. As the range of charges increases, the males seemed to be paying more and more compared to females. So, the empty areas between the bands might have been due to some product (Insurance policies) characteristics while the male-female charges gap with increase in age aligned with earlier observations
- Charges-BMI, against gender: There was one horizontal band which was more centered around BMI of 30 (which was also the mean from earlier observation) – this implied people with average health (average BMI) were paying average charges, while people with worsening health (increasing BMI) were paying more and within that trend, males are paying more
- Charges-number of children, against gender: Charges were more even for the bands with number of children being up to 3, and then dropping. This required more analysis into the composition of larger families, their family medical histories etc.) which was outside the scope of this research
- Charges-region, against gender: Southeast region had more charges as aligned with earlier observations. A deep-dive into region characteristics was outside the scope of this research
- Charges-smoking status, against gender: Smokers were paying more charges, and male smokers were paying more in the 45000-52000 charges range (that specific analysis was outside the scope of this research)

Next, categorical variables were converted into numerics, and co-relation was checked again.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

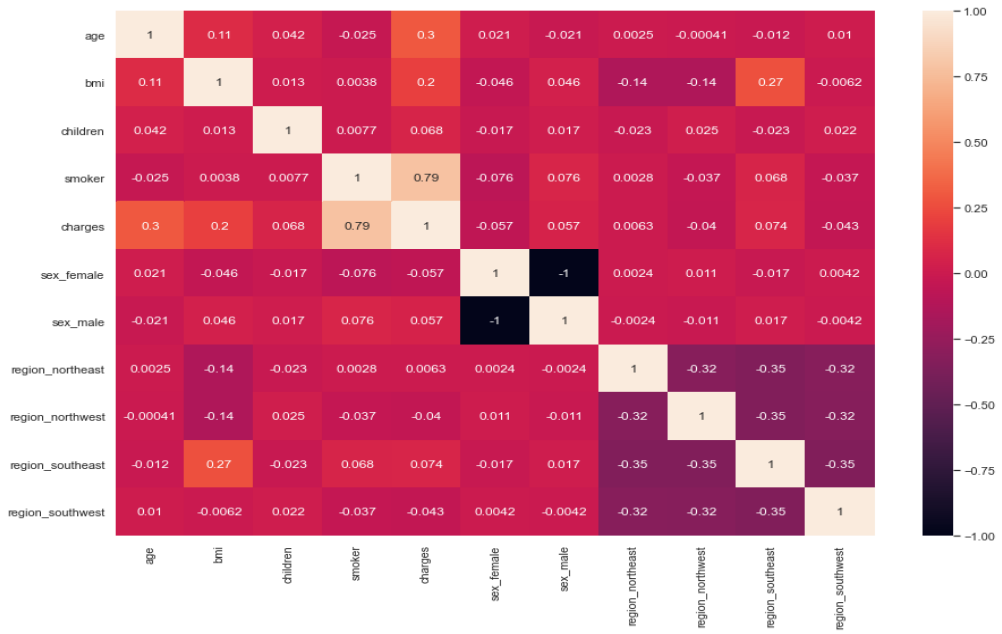


FIGURE 29: CORRELATION AMONG NUMERIC VARIABLES POST CONVERSION OF CATEGORICAL

Observations:

- Direct correlation between gender, age, BMI, number of children, region(some correlation with southeast region) and charges was not very significant
- Direct correlation between smoking status and charges was most prominent based on the current data

In summary:

- Smoking status had the highest direct impact on charges
- Direct impact of gender on charges was not very significant
- With increase in age, BMI and number of children, charges levied on males were more than females
- The gender field did not have any entry for individuals who did not want to be defined by a specific gender or had undergone gender change procedures- this is a very critical gap from inclusivity perspective

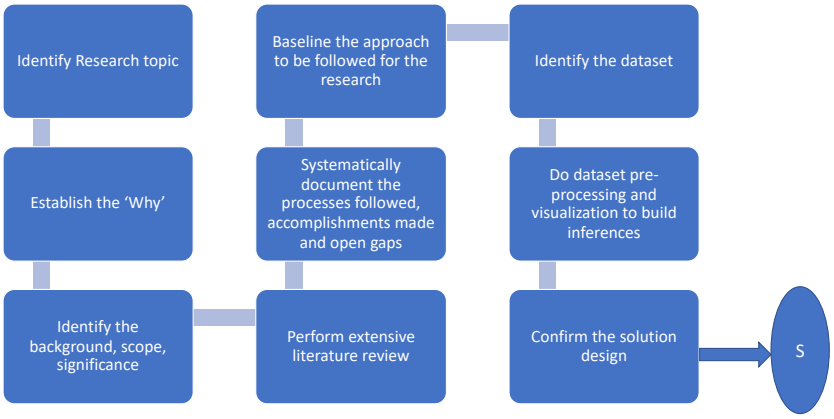
Given the relatively equal distribution of gender in the as-is dataset which bars the opportunity to inspect the impact of a higher-percentage gender on the pricing ,as well as the as-is dataset having no inclusivity (as evident by the lack of allowance for individuals who do not want to conform to a particular gender), it was determined to expand the solution design to accommodate these two gaps by introducing 2 synthetic datasets. Section 3.2.4 covers the synthetic datasets in detail in terms of their composition, their comparison with the as-is dataset and the results obtained by running the regression models against them.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

3.2.4 SOLUTION DESIGN

Based on the inferences from the data pre-processing and visualization, it was determined that the as-is dataset has an equal distribution of male and female gender. Moreover, the dataset does not have any category for personas who follow gender non-conformity or gender fluidity. Given the focus of this research work on suggesting an optimal machine learning algorithm which will not be discriminating pricing on the basis of gender as well as ensure inclusivity for personas who do not want to be known by any specific gender, it was decided to create 2 synthetic datasets- one with an adjusted male-female percentage and the other including a new gender category for personas who follow gender non-conformity and gender fluidity. The as-is dataset was downloaded onto notepad and 2 variants were manually created and edited to align with the proposed compositions of the 2 synthetic datasets. In summary, the solution design followed the following steps:

A. Multiple regression models were trained and tested on the as-is dataset; and it was noted which model produced best results
B. A synthetic form of the as-is dataset was created by adjusting the male-female distribution and then the same models as from (A) were trained and tested on this synthetic dataset no. 1; and it was noted which model produced best results
C. Another synthetic form of the as-is dataset was created by including a 3rd type of gender and calling it 'non-conforming'; the same models as per (A) and (B) were trained and tested on this synthetic dataset no. 2; and it was noted which model produced best results
D. Evaluation metrics for all of (A), (B) and (C) were compared, and the best model was determined



Published by www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

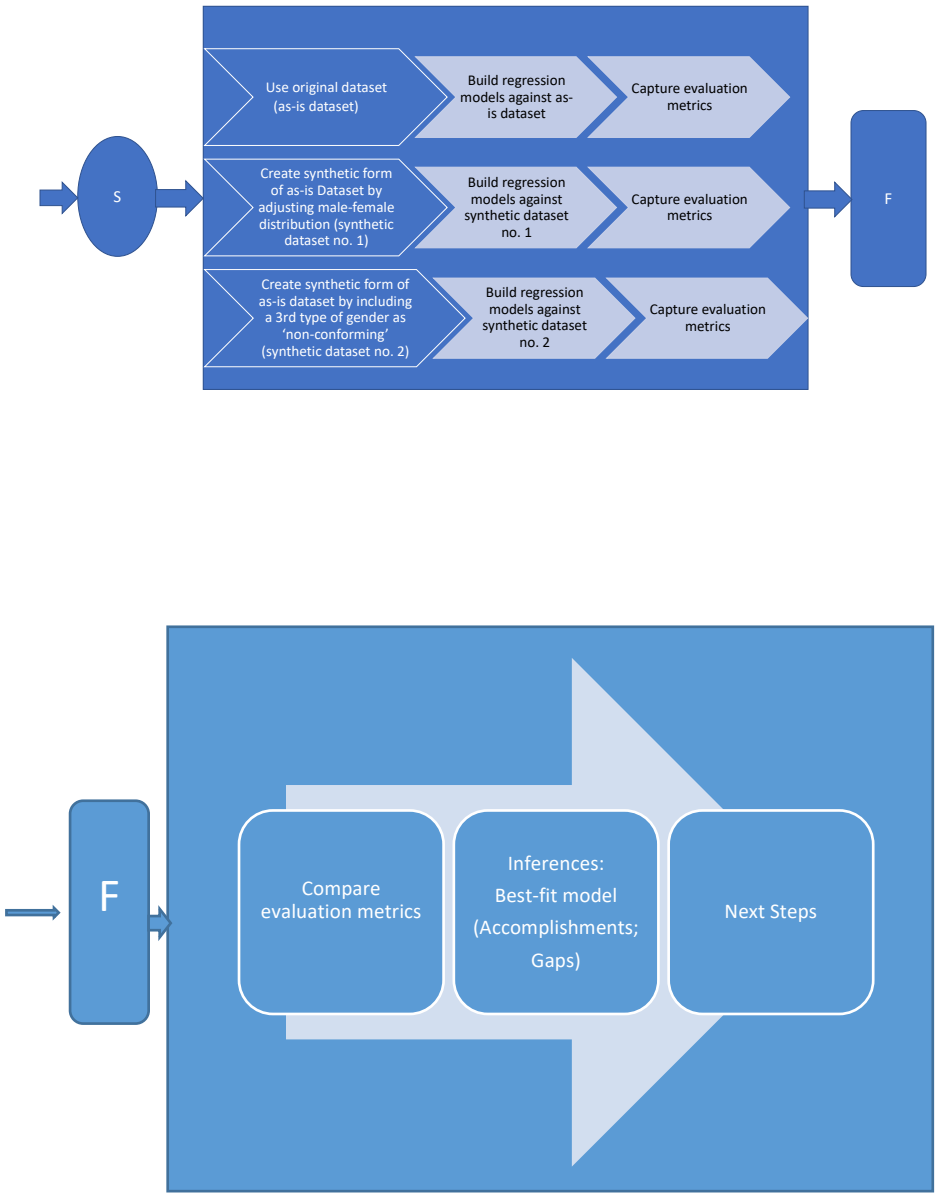


FIGURE 30: RESEARCH METHODOLOGY FLOW DIAGRAM

3.2.5 MODEL VALIDATION

The following regression models were used on the datasets:

1. Multiple linear regression

   Per Vadapalli (2022), this is a type of regression model where a relationship is established between two or more independent variables and the corresponding dependent variables. This is represented by

Published by www.ijert.org

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon_i;$$

where,

- y represents the predicted value of the dependent variable y
- β0 represents the intercept
- β1X1 represents the regression coefficient of the first variable
- βnXn represents the regression coefficient of the last variable
- X represents the independent variable
- ∈ represents the error

2. Ridge regression

Per Knoldus.com, ridge regression is a regularization technique used for feature selection using a shrinkage method for penalization. While lasso regression takes the magnitude of the co-efficients, ridge regression takes the square. This is represented by

$$L_{hridge}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m} w_j \hat{\beta}_j^2.$$

where,

- Alpha (α) is the penalty value denoting the amount of shrinkage
- yi is the predicted value of the dependent variable y
- xi is the value of the independent variable x
- Lambda is the tuning parameter

3. Lasso regression

Per Knoldus.com, lasso regression, like ridge regression, is a regularization technique used for feature selection using a shrinkage method for penalization. This is represented by

$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m} |\hat{\beta}_j|.$$

where,

- Alpha (α) is the penalty value denoting the amount of shrinkage
- yi is the predicted value of the dependent variable y
- xi is the value of the independent variable x
- Lambda is the tuning parameter

4. Polynomial regression

Per javatpoint.com, this is a type of regression algorithm that models the relationship between a dependent variable (y) and independent variable (x) as the n-degree polynomial. This is represented by

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_2 x_1^3 + \ldots\ldots b_n x_1^n ;$$

where,

- y is the predicted value of the dependent variable y
- b_0 is the intercept
- $b_1 x_1$ is the regression coefficient of the first variable

- $b_n x_1^n$ is the regression coefficient of the last variable in the n-degree

5. Random forest regression (RFR)

Per Mwiti (2023), random forest is an ensemble of decision trees.

- Each tree is created from different row samples, and at each such node created, different feature samples are selected for splitting
- Individual predictions are derived from each such tree
- Averaging of predictions is done to produce a result

6. Gradient boosting regression (GBR)

Per Masui (2022), gradient boosting is a variant of ensemble method where multiple weak models can be created and combined to get better performance as a whole. This is represented by

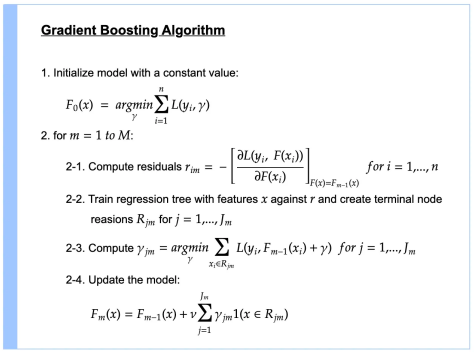


FIGURE 31: GRADIENT BOOSTING ALGORITHM

where,

- F_o is initial constant value prediction
- L is the loss function
- y is the value of the dependent variable y
- M is the number of iterations
- $r_i m$ is residuals calculated by taking a derivative of the loss function with respect to the previous prediction of constant value
- j represents a terminal node on the tree, m denotes the tree index, and capital J denoted the total number of nodes

7. Decision tree regression (DTR)

Per Gurucharan (2020), decision tree follows supervised learning in a tree-structured process.

- There are 3 types of nodes- root node is the initial node, interior nodes are the features and the branches are the decision rules, and leaf nodes represent the outcome
- Any particular data point is run through the entire tree through true/false method till the leaf node
- The final prediction is determined as the average of the values of the dependent variable in that leaf node

8. K-nearest neighbor regression (KNNR)

Per Muhajir (2019), KNN algorithm stores all available cases and predicts a numerical target based on measures of similarity such as distance functions. KNN, when used for regression, calculates the average of numerical target of the K nearest neighbors. Mathematically, the distance functions used in regression for continuous variables are expressed as:

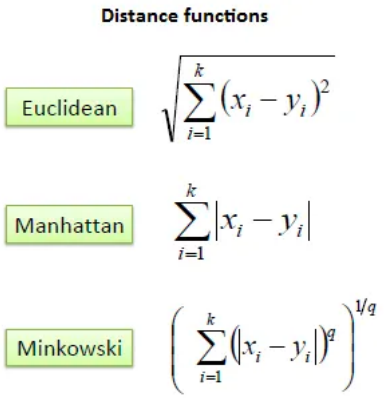


FIGURE 32: KNNR ALGORITHM

The evaluation metrics used for comparing the models were:

A. Regression score

Per Ogunbiyi (2022), R2 score is an effective metric for determining regression scores, or accuracy of regression models. It is a statistical measure that shows how well the model is performing in terms of matching its predictions, on a scale of 0 to 1. Mathematically, it is expressed as

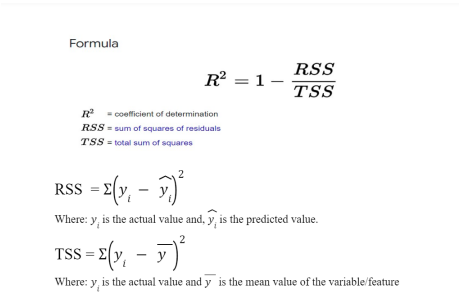


FIGURE 33: FORMULA FOR R2 SCORE

B. Mean Absolute Error (MAE)

Per Brownlee (2021), MAE is useful because the units of the error match the units of the predicted value. Changes in MAE are linear. MAE also does not give different ranges of weights to different errors and scores increase linearly as errors increase. MAE is calculated as the average value of absolute error values. Mathematically, it is expressed as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

MAE = 1 / N * sum for i to N abs(y_i – yhat_i), where, y_i represents the i'th expected value in the dataset, yhat_i represents the i'th predicted value and abs() is the absolute function for the error values.

C. Mean Squared Error (MSE)

Per Brownlee (2021), MSE is an important loss function for algorithms which can be optimized using least squares method on regression problems. It is calculated as the mean/average of the squares of the differences between predicted and actual values. Mathematically, it is expressed as MSE = 1 / N * sum for i to N (y_i – yhat_i)^2, where, y_i represents the i'th expected value in the dataset and yhat_i represents the i'th predicted value.

D. Root Mean Squared Error (RMSE)

Per Brownlee (2021), RMSE is an extension of MSE. It is calculated by determining the square root of MSE, which means that the square root of the error is calculated, and units remain the same. Thus, MSE loss is looked at for training the regression model while RMSE is used to evaluate model performance. Mathematically, RMSE is expressed as RMSE = sqrt(MSE), which also implies, RMSE = sqrt(1 / N * sum for i to N (y_i – yhat_i)^2), where, y_i represents the i'th expected value in the dataset and yhat_i represents the i'th predicted value.

Python code was used for training and testing the regression models on the 3 datasets. The algorithm for all the 3 variations was:

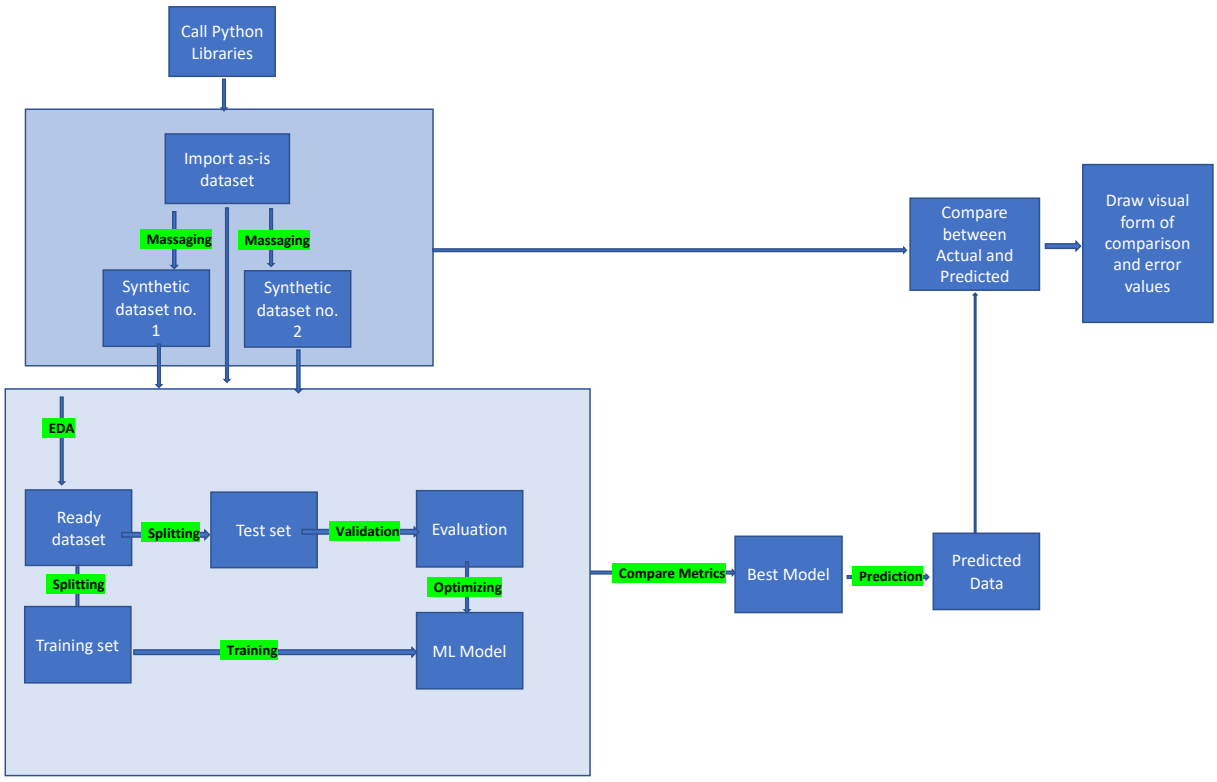


FIGURE 34: ALGORITHM FOR REGRESSION MODELS

General steps of the model train/test process were:

Published by www.ijert.org

Pre-processing: Charges was selected as the predictor variable y, while the rest of the dataset was selected as the dependent variable set x. The model fitting was done on x_train and y_train (train=70%; test=30%).

Hyperparameter tuning: To tune hyperparameters for the non-linear models, Grid search was used for RFR, DTR, kNNR and GBR with cv as 5 and n_jobs as 1, and multi-threading was used.

Evaluation Metrics calculations: MAE was calculated by running the metrics.mean_absolute_error function on the predicted and actual values for y_test. MSE was calculated by running the metrics.mean_squared_error function on the predicted and actual values for y_test. RMSE was calculated by running the np.sqrt function on the MSE value. Regression score was calculated by running the r2_score function on the predicted and actual values for y_test.

Steps, specific to each regression model, of the model train/test process were:

Multiple linear regression:

- The LinearRegression function was used from the sklearn.linear-model library
- The predict variables y_train_pred and y_test_pred were calculated by running the Lin_reg.predict function on the x_train and x-test data cuts

Ridge regression:

- The Ridge function was used from the sklearn.linear-model library
- alpha was chosen as 0.1
- The predict variables y_train_pred and y_test_pred were calculated by running the Ridge.predict function on the x_train and x-test data cuts

Lasso regression:

- The Lasso function was used from the sklearn.linear-model library
- alpha was chosen as 0.001; max-iter was chosen as 100; tol was chosen as 0.0001; selection was chosen as 'cyclic'
- The predict variables y_train_pred and y_test_pred were calculated by running the Lasso.predict function on the x_train and x-test data cuts

Polynomial regression:

- The PolynomialFeatures function was used from the sklearn.preprocessing library
- Degree of 4 was chosen
- The predict variables y_train_pred and y_test_pred were calculated by running the Pol-reg.predict function on the x_train and x-test data cuts

Random forest regression:

- The predict variables y_pred and y_pred1 were calculated by running the reg_RFR.predict function on the x_train and x-test data cuts

Decision tree regression:

- The predict variables y_pred and y_pred1 were calculated by running the reg_DTR.predict function on the x_train and x-test data cuts

k-nearest neighbor regression:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

- The predict variables y_pred and y_pred1 were calculated by running the reg_KNNR.predict function on the x_train and x-test data cuts

Gradient boosting regression:

- The predict variables y_pred and y_pred1 were calculated by running the reg_GBR.predict function on the x_train and x-test data cuts

3.2.5.2 Model Development (Build & Train)

A. Build and train models on as-is dataset.

The regression models were trained and tested on the as-is dataset, and the evaluation metrics were documented.

B. Build and train models on a synthetic dataset created by adjusting the male-female % from the original dataset.

The female% was increased to 83% in a new synthetic dataset (synthetic dataset no. 1). Originally it was about 50%, with the other 50% being male. No other change was made to the dataset. Some key aspects of this dataset from data pre-processing and data visualization (comparing against data pre-processing and data visualization for the original dataset): Statistical view remains the same. Missing values view remains the same. Outlier view remains the same. Distribution of charges on source variables remains the same, for both original and normalized views. Charges by Age remains the same.

Charges by Gender view has changed.

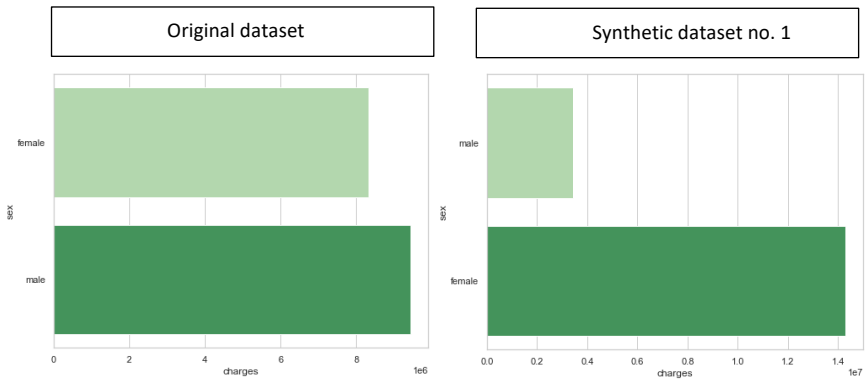


FIGURE 35: CHARGES BY GENDER - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

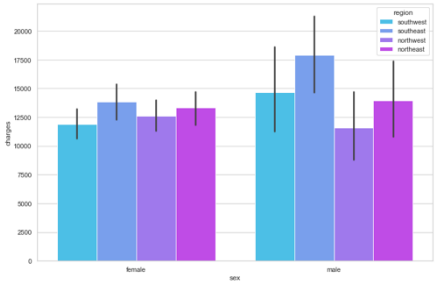
This is due to the adjusted male-female%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

Charges by BMI remains the same. Charges by Children remain the same. Charges by Smoking status remains the same. Charges by Region remains the same.

There are some changes to the Gender-Charges distribution by Region.

Original dataset



Synthetic dataset no. 1

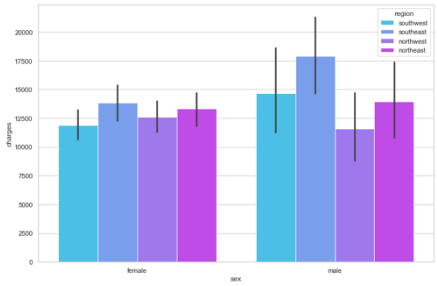
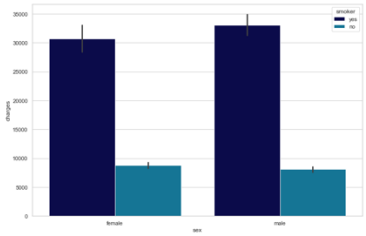


FIGURE 36: GENDER-CHARGES DISTRIBUTION FOR REGION - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

This is due to the adjusted male-female%.

There are some changes to the Gender - Charges distribution by Smoking status.

Original dataset



Synthetic dataset no. 1

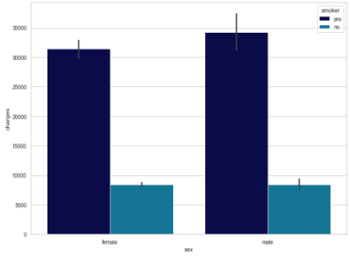
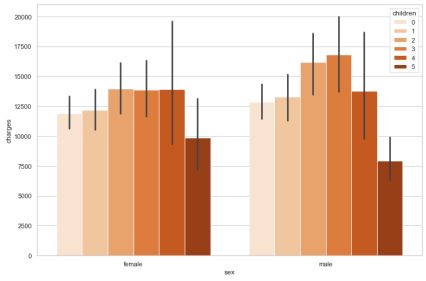


FIGURE 37: GENDER-CHARGES DISTRIBUTION FOR SMOKING STATUS - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

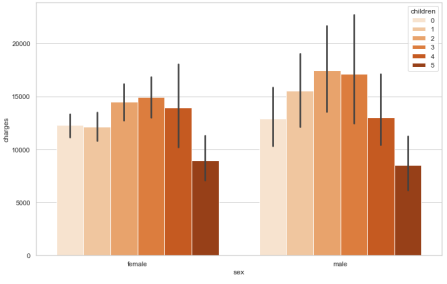
This is due to the adjusted male-female%.

There are some changes to the Gender - Charges distribution by Number of Children.

Original dataset



Synthetic dataset no. 1



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

FIGURE 38: GENDER-CHARGES DISTRIBUTION FOR NUMBER OF CHILDREN - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

This is due to the adjusted male-female%.

There are some changes to the Charges – Age distribution by Gender.

Original dataset	Synthetic dataset no. 1

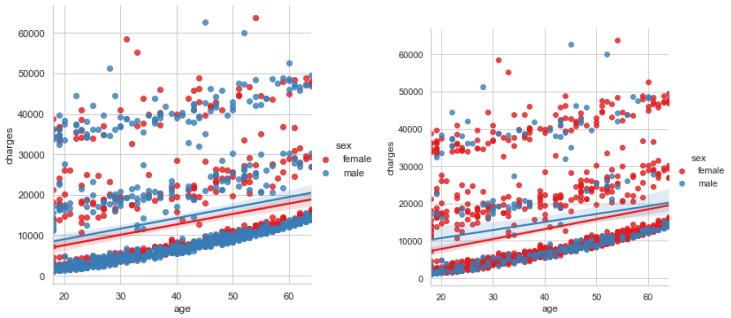


FIGURE 39: CHARGES-AGE DISTRIBUTION FOR GENDER - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

The distribution is heavier towards females in the view of the synthetic dataset no. 1 because of the adjusted male-female%.

There are some changes to the Charges – BMI distribution by Gender.

Original dataset	Synthetic dataset no. 1

FIGURE 40: CHARGES-BMI DISTRIBUTION FOR GENDER - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

The distribution is heavier towards females in the view of the synthetic dataset no. 1 because of the adjusted male-female%.

There are some changes to the Charges – Number of Children distribution by Gender.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

Original dataset | Synthetic dataset no. 1

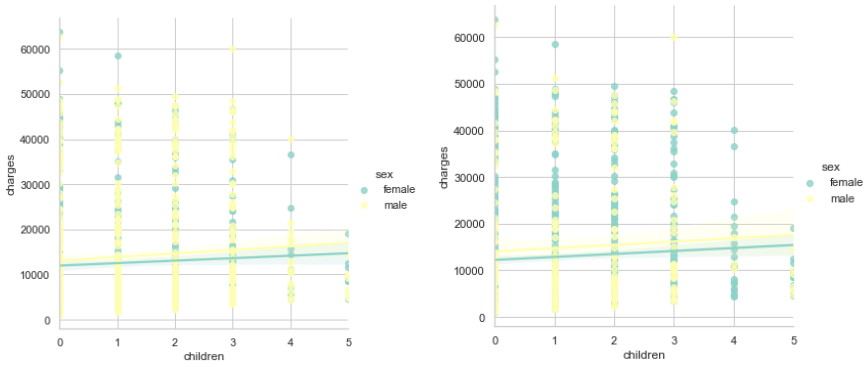


FIGURE 41: CHARGES-NUMBER OF CHILDREN DISTRIBUTION FOR GENDER - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

The distribution is heavier towards females in the view of the synthetic dataset no. 1 because of the adjusted male-female%.

No changes to the correlation among Age, BMI, Charges, and Number of Children.

There are some changes to the pairplot view of the numeric variables against gender.

Original dataset | Synthetic dataset no. 1

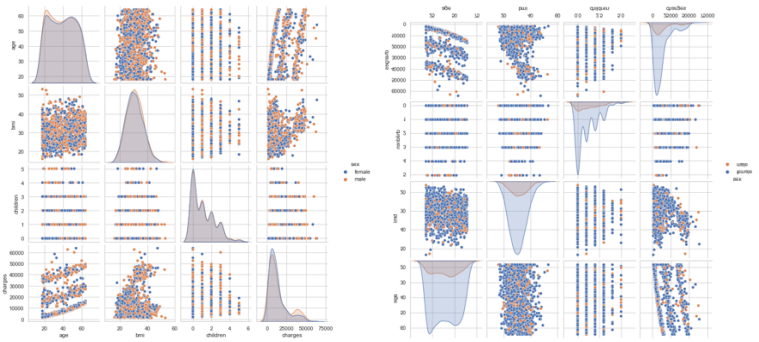


FIGURE 42: PAIRPLOTS FOR NUMERIC VARIABLES AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

The distribution is heavier towards females in the view of the synthetic dataset no. 1 because of the adjusted male-female%.

There are some changes to the scatterplot view for source variables and Charges, against Gender

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

Original dataset

Synthetic dataset no. 1

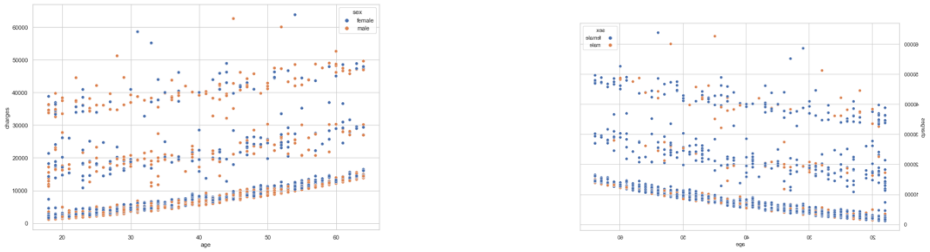


FIGURE 43: SCATTERPLOT FOR CHARGES AND AGE, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

Original dataset

Synthetic dataset no. 1

FIGURE 44: SCATTERPLOT FOR CHARGES AND BMI, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

Original dataset

Synthetic dataset no. 1

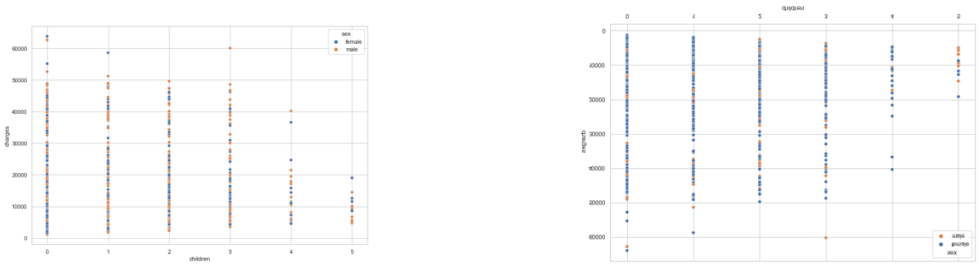


FIGURE 45: SCATTERPLOT FOR CHARGES AND NUMBER OF CHILDREN, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

Original dataset

Synthetic dataset no. 1

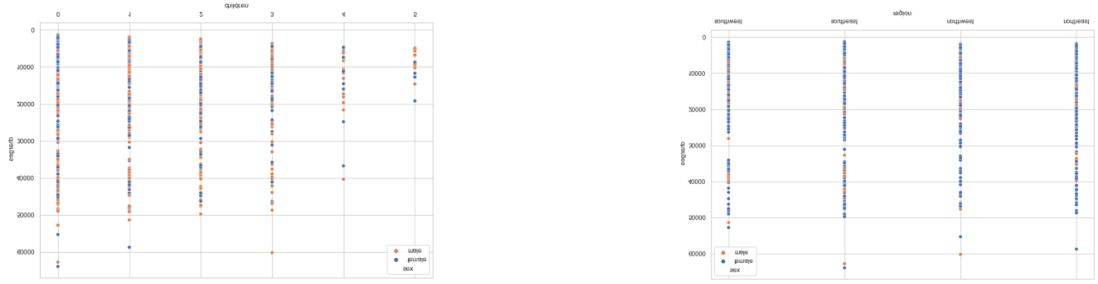


FIGURE 46: SCATTERPLOT FOR CHARGES AND REGION, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

Original dataset

Synthetic dataset no. 1

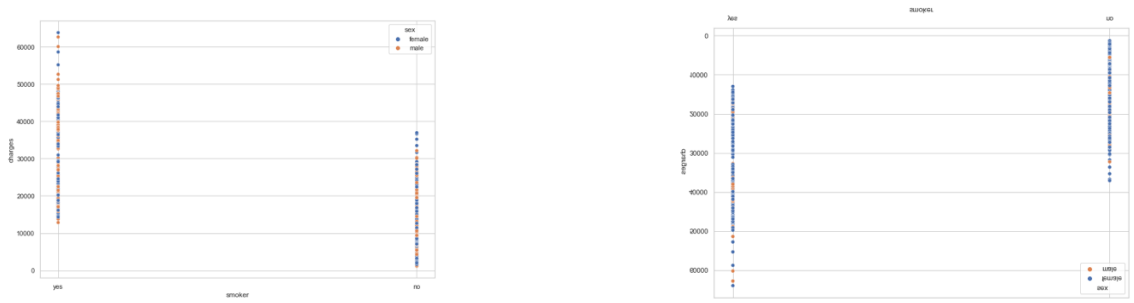


FIGURE 47: SCATTERPLOT FOR CHARGES AND SMOKING STATUS, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

For all these scatterplots, the observations as against the original dataset remain the same, however the distribution is heavier towards females in the view of the synthetic dataset no. 1 because of the adjusted male-female%.

Post conversion of categorical variables to numeric variables, and rechecking of correlation among numerics, very minor changes are there which do not merit any deep-dive.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

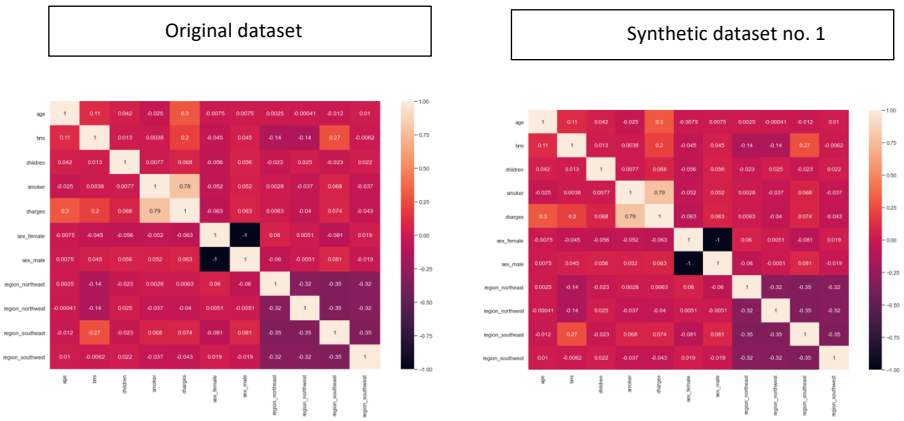


FIGURE 48: CORRELATION AMONG NUMERIC VARIABLES POST CONVERSION OF CATEGORICAL - COMPARISON BETWEEN ORIGINAL DATASET AND SYNTHETIC DATASET NO. 1

Finally, the regression models were trained and tested on the synthetic dataset no. 1, and the evaluation metrics were documented.

C. Build and train models on a synthetic dataset created by adjusting the male-female % from the original dataset.

A new synthetic dataset (synthetic dataset no. 2) was created by introducing a new gender type ('Non-conforming'). The new gender distribution % were : Female 36% ; Male 35%; Non-conforming 29%.
No other changes were made to the dataset. Some key aspects of this dataset from data pre-processing and data visualization (comparing against data pre-processing and data visualization for the original dataset and synthetic dataset no. 2): Statistical view remains the same. Missing values view remains the same. Outlier view remains the same. Distribution of charges on source variables remains the same, for both original and normalized views. Charges by Age remains the same.

Charges by Gender have changed.

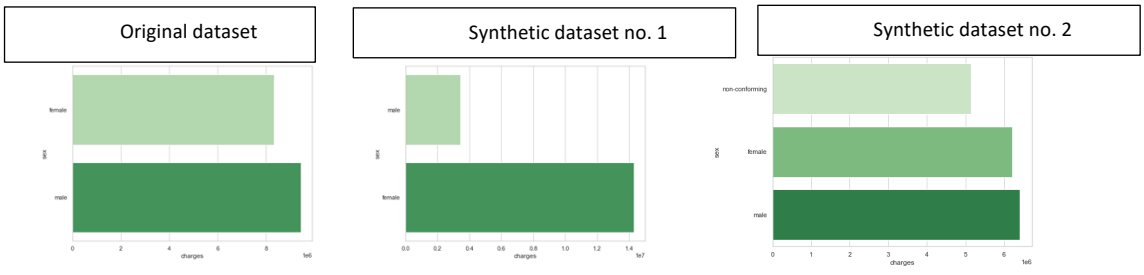


FIGURE 49: CHARGES BY GENDER - COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

This is due to the changes in gender% across the 3 datasets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

Charges by BMI remains the same. Charges by Children remain the same. Charges by Smoking status remains the same. Charges by Region remains the same.

There are some changes to the Gender-Charges distribution by Region.

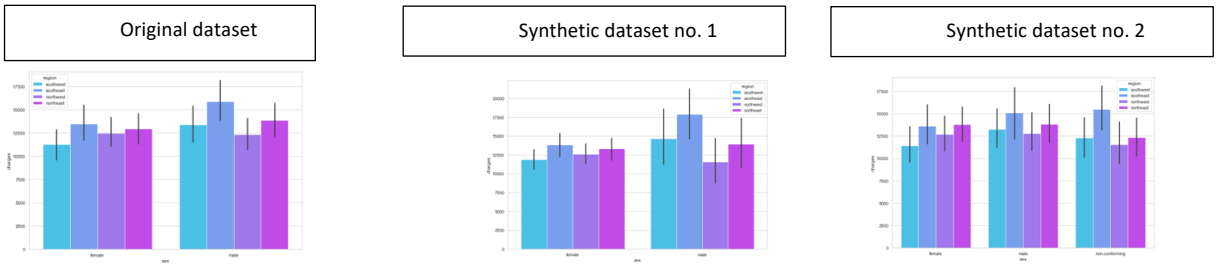


FIGURE 50: GENDER-CHARGES DISTRIBUTION FOR REGION - COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

This is due to the changes in gender% across the 3 datasets.

There are some changes to the Gender - Charges distribution by Smoking status.

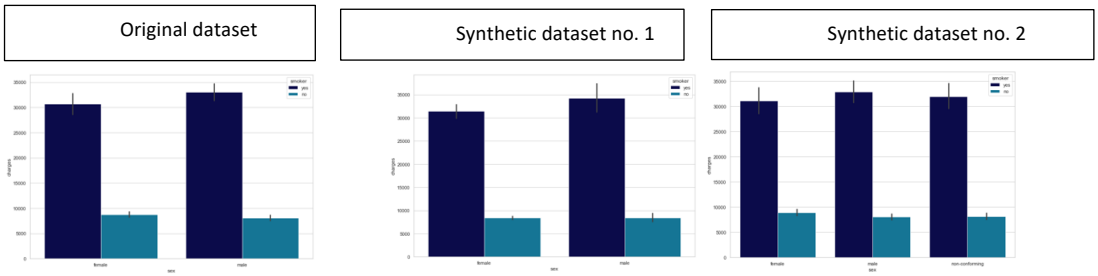


FIGURE 51: GENDER-CHARGES DISTRIBUTION FOR SMOKING STATUS - COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

This is due to the changes in gender% across the 3 datasets.

There are some changes to the Gender – Number of Children distribution by Smoking status.

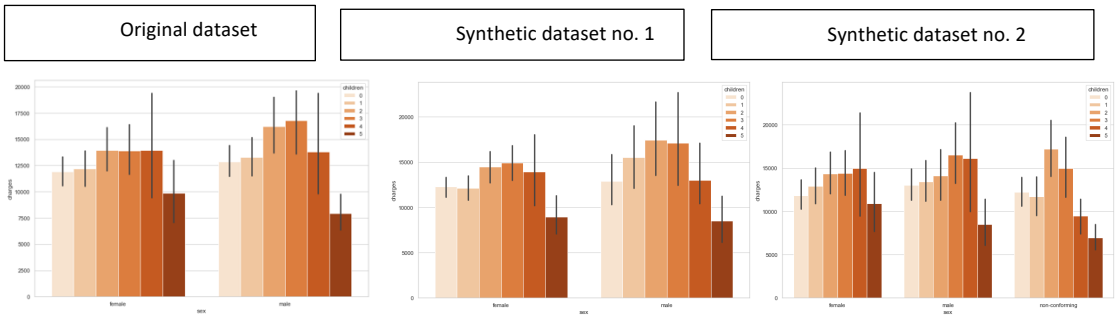


FIGURE 52: GENDER-CHARGES DISTRIBUTION FOR NUMBER OF CHILDREN - COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

This is due to the changes in gender% across the 3 datasets.

There are some changes to the Charges – Age distribution by Gender.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

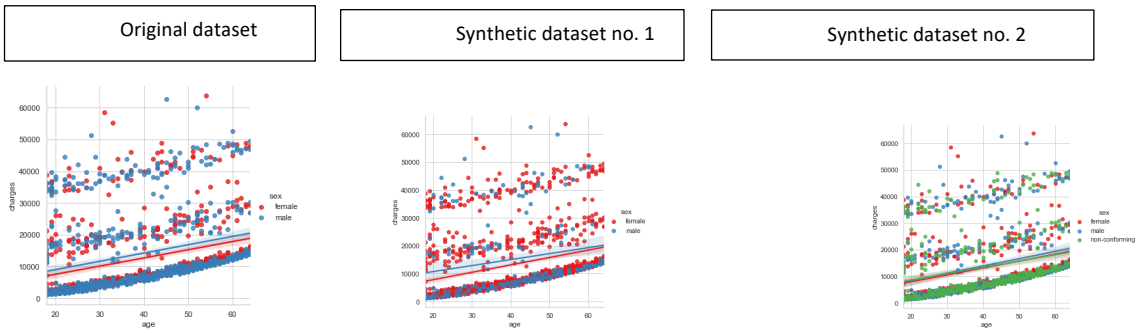


FIGURE 53: CHARGES-AGE DISTRIBUTION FOR GENDER - COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

The distribution is heavier towards non-conforming in the view of the synthetic dataset no. 1 because of the adjusted gender%.

There are some changes to the Charges – BMI distribution by Gender.

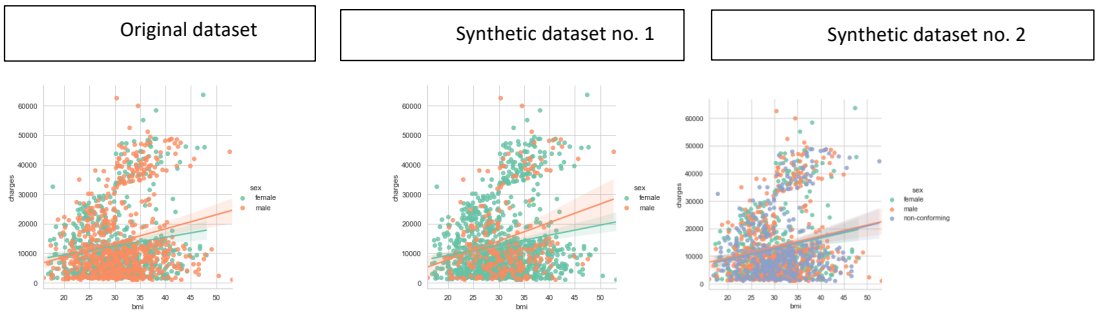


FIGURE 54: CHARGES-BMI DISTRIBUTION FOR GENDER - COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

The distribution is heavier towards non-conforming in the view of the synthetic dataset no. 2 below BMI of 35. Analysis of any health factors is outside the scope of this study.

There are some changes to the Charges – Number of Children distribution by Gender.

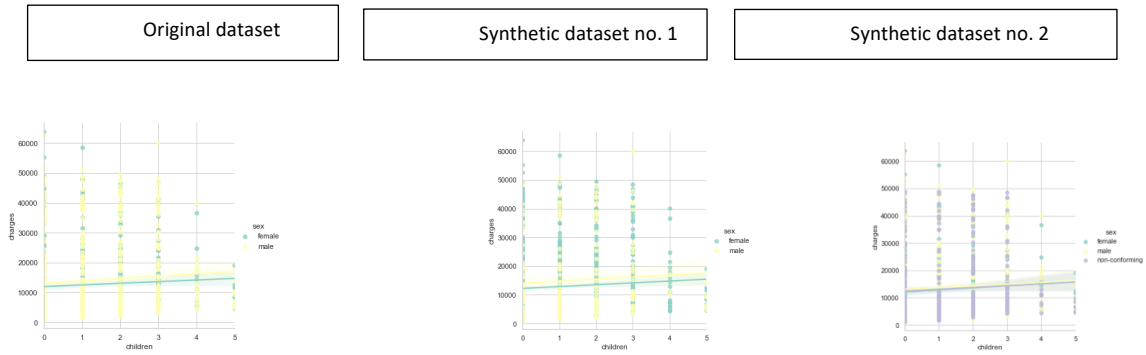


FIGURE 55: CHARGES-NUMBER OF CHILDREN DISTRIBUTION FOR GENDER - COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

Charges are lower for non-confirming for upto 3 children. The deep-dive into family medical history is outside the scope of this study.

No changes to the correlation among Age, BMI, Charges, and Number of Children.

There are some changes to the pairplot view of the numeric variables against gender.

Original dataset	Synthetic dataset no. 1	Synthetic dataset no. 2

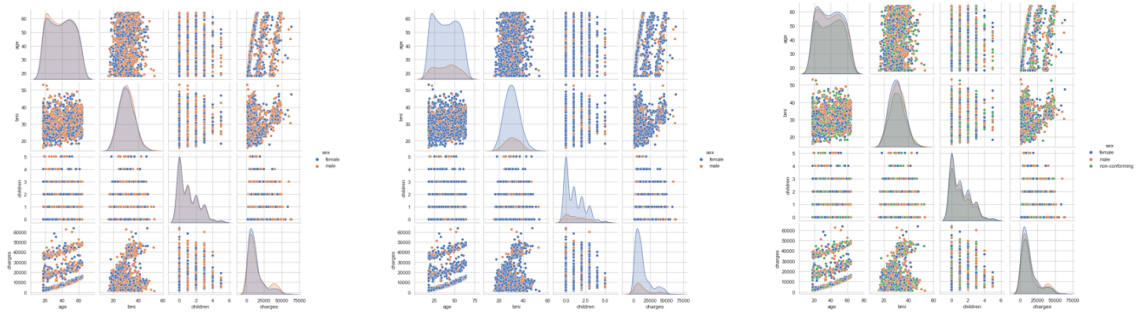


FIGURE 56: PAIRPLOTS FOR NUMERIC VARIABLES AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

The distribution is mostly even for the genders across the 3 datasets.

There are some changes to the scatterplot view for source variables and Charges, against Gender

Original dataset	Synthetic dataset no. 1	Synthetic dataset no. 2

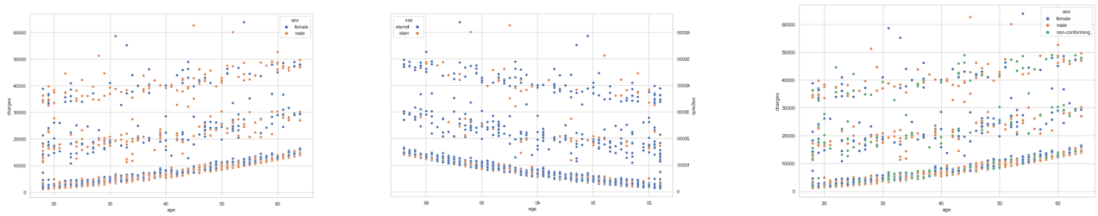


FIGURE 57: SCATTERPLOT FOR CHARGES AND AGE, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

Original dataset	Synthetic dataset no. 1	Synthetic dataset no. 2

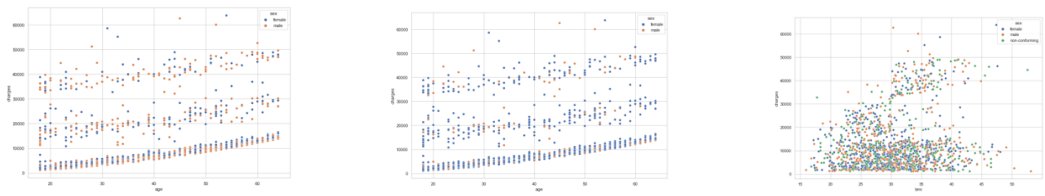


FIGURE 58: SCATTERPLOT FOR CHARGES AND BMI, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

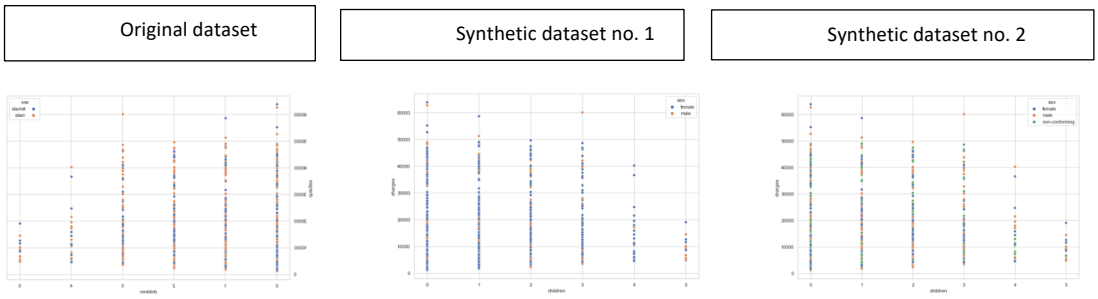


FIGURE 59: SCATTERPLOT FOR CHARGES AND NUMBER OF CHILDREN, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

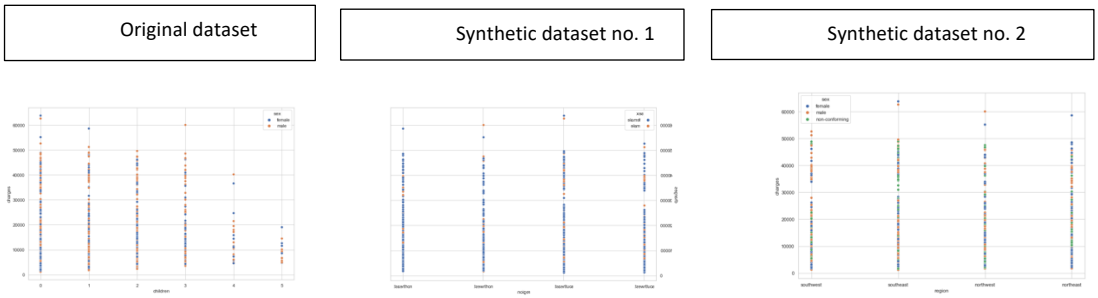


FIGURE 60: SCATTERPLOT FOR CHARGES AND REGION, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

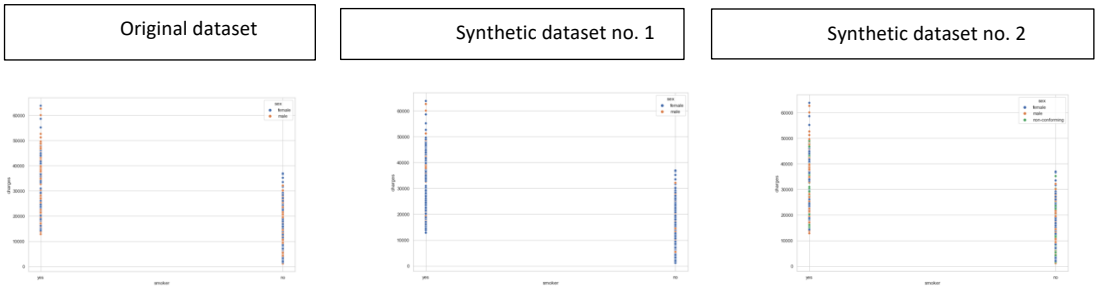


FIGURE 61: SCATTERPLOT FOR CHARGES AND SMOKING STATUS, AGAINST GENDER – COMPARISON BETWEEN ORIGINAL DATASET, SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

For all these scatterplots, the observations as against the original dataset remain the same, also the distribution seems to be even for the genders across the 3 datasets.

Post conversion of categorical variables to numeric variables, and rechecking of correlation among numerics, very minor changes are there which do not merit any deep-dive.

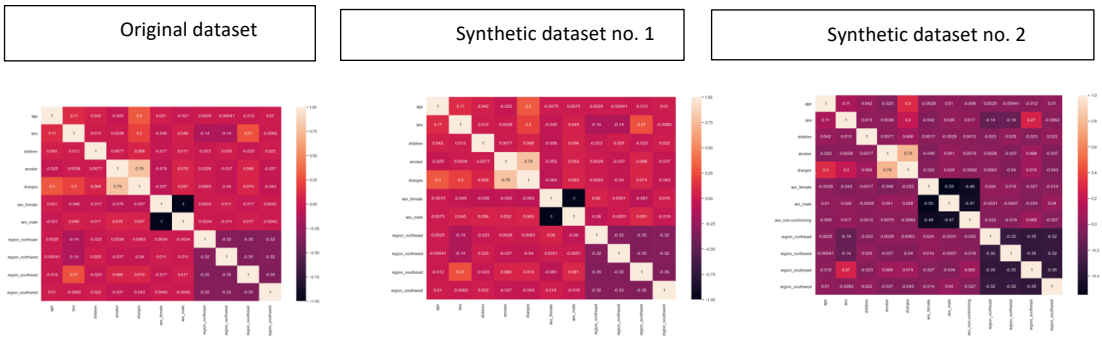


FIGURE 62: CORRELATION AMONG NUMERIC VARIABLES POST CONVERSION OF CATEGORICAL - COMPARISON BETWEEN ORIGINAL DATASET , SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

Finally, the regression models were trained and tested on the synthetic dataset no. 2, and the evaluation metrics were documented.

3.2.6 MODEL EVALUATION

The evaluation metrics post build, train and test runs for all the 3 datasets were captured as below:

TABLE 2: EVALUATION METRICS

Dataset	Model	Regression score	MAE ytest	MSE ytest	RMSE ytest
As-is dataset	Linear regression	0.7909	4011,4496	33342497,83	5774,2963
	Ridge regression	0.7909	4012,9746	33349434,74	5774,8969
	Lasso regression	0.7909	4011,4504	33342501,03	5774,2966
	Polynomial regression	0.7315	4184,7242	42803903,71	6542,4692
	Gradient boosting regression	0.8418	2506,5859	18028364,31	4245,9821
	Decision tree regression	0.8236	4392,559	43308572,44	6580,9249
	Random forest regression	0.8321	2608,3337	17736722,46	4211,4988
	K-nearest neighbor regression	0.1492	8916,9056	138460831,7	11766,9381
Synthetic dataset no. 1	Linear regression	0.7908	4032,8233	33355475,62	5775,4199
	Ridge regression	0.7908	4034,3141	33362462,01	5776,0248
	Lasso regression	0.7908	4032,8233	33355477,62	5775,4201
	Polynomial regression	0.6952	4193,6643	48602115,95	6971,5218
	Gradient boosting regression	0.8413	2598,3517	18322368,21	4280,4636
	Decision tree regression	0.8236	4218,0185	37132778,62	6093,6671
	Random forest regression	0.8340	2668,497	18423738,74	4292,2883
	K-nearest neighbor regression	0.1579	8911,8597	138669645,7	11775,8076
Synthetic dataset no. 2	Linear regression	0.7917	3991,4941	33218689,93	5763,5657
	Ridge regression	0.7916	3992,9588	33225667,48	5764,171
	Lasso regression	0.7917	3991,4946	33218692,19	5763,5659
	Polynomial regression	0.7149	4302,8532	45460635,04	6742,4502
	Gradient boosting regression	0.8395	2531,0095	18076551,77	4251,6528
	Decision tree regression	0.8236	4234,2599	39796394,95	6308,4384
	Random forest regression	0.8200	3030,0714	20517401,14	4529,6138
	K-nearest neighbor regression	0.1477	8941,8122	139128225,8	11795,2628

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

3.2.7 INFERENCES

2 key attributes were considered for analyzing the evaluation metrics:

- Highest regression score
- Lowest RMSE

From an analysis of the models, it was evident that GBR, DTR and RTR had the best values for regression score and RMSE. On further analyzing these values for those two models:

It was observed that GBR had the highest regression score for all the datasets. GBR also had the lowest RMSE for 2 out of the 3 datasets while RFR had the lowest RMSE for the other dataset.

Thus, it was concluded that GBR could be considered as the optimal regression model for the problem statement.

Based on GBR, the following steps were performed:

a. Predicted and actual values for the predictor variable Charges were calculated
b. Visualized form of the comparison between predicted values and actual values for the predictor variable Charges, was drawn
c. Visualized form for the Error values was drawn

Results:

A. Predictor value comparison :

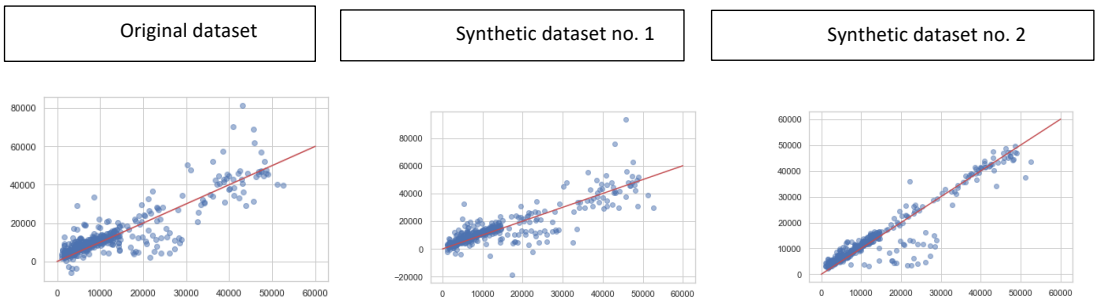


FIGURE 63: PREDICTED VALUES AND ACTUAL VALUES FOR THE PREDICTOR VARIABLE CHARGES - COMPARISON BETWEEN ORIGINAL DATASET , SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

B. Error value comparison:

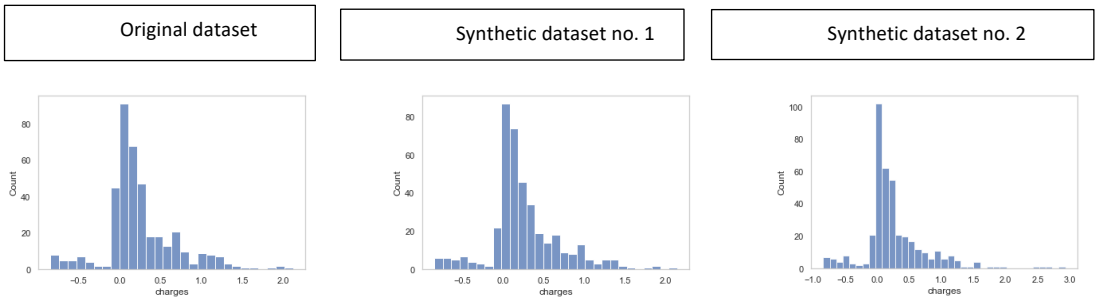


FIGURE 64: ERROR VALUES FOR THE MODELS - COMPARISON BETWEEN ORIGINAL DATASET , SYNTHETIC DATASET NO. 1 AND SYNTHETIC DATASET NO. 2

In addition to the most optimal-performance model, GBR also produced the least number of errors for synthetic dataset no. 2. This was a very critical observation given that this was the dataset used by this research work to introduce the 'non-conforming' gender in the Life Insurance pricing framework.

3.2.8 SUMMARY AND NEXT STEPS

This research work was intended to take a broad view of any gender bias existing in the US Life Insurance industry from a pricing perspective and suggest a machine learning approach to address any such bias. The idea behind this approach was to consider US Life Insurance as a use case for the approach considering the major share US Insurance has within global Insurance industry and the leading share Life Insurance has within US Insurance.

In order to achieve these objectives, the research first established the immense influence of Insurance on modern lives. It was observed that Insurance provides life cover, guaranteed income, retirement benefits, auto damage cover, home protection, specialty asset protection, protection for pets, protection for newer technological advances such as drones, capital market growth, general savings rate increase. Post-Covid, the global Insurance has seen a CAGR in excess of 8%. It was also observed that US is the largest Insurance industry in the world and within US, Life Insurance is a leader. Therefore, it was deemed fit that any fundamental problem can be addressed leveraging US Life Insurance as a prototype and thereafter, the solution can be adopted at a global level, considering the geo-specific criteria existing in Insurance industry globally. It was also established that in spite of the growing influence of Insurance and many advancements made by the industry, there still exist traces of systemic bias such as gender bias, racial bias, gender definition bias etc. which are unfortunate instances of discrimination and need to be tackled with urgency. Thus, the first objective of this research work was reiterated as taking a look at global Insurance industry including US Life insurance and reconfirm gender bias from a pricing perspective. Secondly, it was observed as to how pervasive machine learning has become in Insurance and how it is being used across several use cases in the industry. However, machine learning can only work with the greatest degree of fairness when underlying data/principles are unbiased. Therefore, the second objective of this

research work was confirmed as suggesting a machine learning approach to optimally address gender bias in US Life Insurance pricing framework, with a future target of extrapolating that approach to a global level. A broad study was performed of existing research literature in this field. The review covered how the global Insurance industry including US Life Insurance, approaches gender bias including gender non-conformity, and how existing machine learning approaches have considered the treatment of gender bias into the pricing frameworks.

Based on the literature review, this research work was able to establish:

- There is a gender bias existing in the global Insurance industry, based on studies done in Europe (UK, Germany, Italy), Latin America (Mexico), Africa (Kenya), EU and Asia
- The industry has significant maturity to attain when it comes to treating gender fluidity, gender non-conformity and gender-affirmation. Self-identification of gender leads to bias from Insurance companies for coverage as well as pricing. This is a very critical gap identified by this research work as needing urgent and sensitive attention
- Multiple machine learning frameworks/models are either used in the industry or are proposed by researchers
    - LSMC for proxy modelling
    - Classifiers XGBoost, AdaBoost etc. complemented by SMOTE models
    - White-box approaches such as GLM and Black-Box approaches e.g. GBM
    - Neural networks like CNN and FNN
    - Model de-biasing through pre-processing, in-processing and post-processing
    - LDS
    - Adversarial learning
- While these approaches seem to be technically sound and achieve broad objectives from the perspective of addressing multiple forms of discrimination in the Insurance industry, there seems to be a gap around a machine learning approach to directly address gender bias including gender definition from a pricing perspective

Subsequently, this research work was motivated to explore regression as a machine learning technique (given the gaps seem to be existing with the other approaches) to solve the gender bias including gender definition from perspective of ensuring discrimination-free pricing in US Life Insurance.

As part of this approach, a public dataset was chosen with a set of parameters optimally critical to pricing in US Life Insurance. Given the progress made by Python in machine learning, Python was decided as the coding language. Once the dataset was retrieved, adequate pre-processing was performed including statistical view, missing values view, outliers view etc. On finding that the dataset was of good quality, significant amount of EDA was performed on the dataset to arrive at inferences around the existing data. Some pertinent observations around gender were:

- Males seem to be broadly paying about 20% more premium than females

- There is no inclusivity in the pricing framework for individuals who do not want to be tied down by a specific gender definition- this is a very critical gap

The solution plan for this research work was therefore decided as:

A. To train and test multiple regression models on the as-is dataset; note which model gives best results
B. To create a synthetic form of the dataset by adjusting the male-female distribution and then train and test the same models as were done in (A); note which model gives best results
C. To create another synthetic form of the dataset by including a 3rd type of gender and calling it 'non-conforming'; train and test the same models as per (A) and (B); note which model gives best results
D. To compare evaluation metrics for all of (A), (B) and (C) and determine the best model

The regression models identified for this approach were:

1. Multiple linear regression
2. Ridge regression
3. Lasso regression
4. Polynomial regression
5. Random forest regression
6. Gradient boosting regression
7. Decision tree regression
8. K-nearest neighbor regression

It was also decided that for all the regression models, train: test split would be 70:30 and the evaluation metrics would be regression score, MAE, MSE and RMSE. Also, for the regression model producing the most optimal values for the evaluation metrics (highest regression score, and lowest RMSE value), predicted and actual values would be compared.

Post the completion of the build/train/test and metrics analysis on all the 3 datasets, GBR was identified as the regression model producing the most optimal values for regression score and RMSE, which were the 2 evaluation metrics identified for final recommendation. Furthermore, on comparison of the predicted and actual values for the predicted variable Charges, GBR produced the least number of errors on the synthetic dataset no. 2. This was a key accomplishment as this was the dataset used by this research work to address the critical gap around the pricing framework not addressing the aspect of gender non-conformity, definition, and fluidity.

The next steps identified out of this research are suggested as:

- Use a larger dataset with more pricing parameters for further fine-tuning of the model. Given the scarcity of US Life Insurance datasets in public domain, it may be necessary to adopt a synthetic data generation approach
- Partner with a US Life Insurance carrier to explore productionizing the model
- Expand the implementation of the model to a global Life Insurance level by customizing the model to region-specific criteria in alignment with policies

REFERENCES

1. Pilijan.I , Cogoljevic. D and Pilijan T (2015). 'Role of insurance companies in financial market', International Review , 2015, br. 1-2, str. 94-102 , pp.2. Available at https://scindeks.ceon.rs/article.aspx?artid=2217-97391502094P (Accessed: 23 October 2022)

2. The Business Research Company (2022). 'Insurance Global Market Report 2022'. Available at https://www.thebusinessresearchcompany.com/report/insurance-global-market-report#:~:text=This%20insurance%20research%20report%20delivers,(CAGR)%20of%208.6%25. (Accessed: 25 December 2022)

3. Verified Market Research (2022). 'Global Insurance Market Size By Type (Life Insurance, Non-Life Insurance), By Organization Size (Large Enterprises, Small and Medium-sized Enterprises (SMEs)), By Geographic Scope And Forecast'. Available at https://www.verifiedmarketresearch.com/product/insurance-market/. (Accessed: 25 February 2023)

4. Flynn, J (2022). '20+ Interesting U.S. Insurance Industry Statistics [2022]: Insurance Facts in 2022'. Available at https://www.zippia.com/advice/insurance-industry-statistics/ (Accessed: 25 December 2022)

5. Liedtke, P (2007). 'What's Insurance to a Modern Economy', *The Geneva Papers on Risk and Insurance - Issues and Practice* , **Volume 32**, pages 211–221 (2007), pp. 217. Available at https://link.springer.com/article/10.1057/palgrave.gpp.2510128 (Accessed: 23 October 2022)

6. Schanz, K-U (2018). 'Understanding and Addressing Global Insurance Protection Gaps'. Available at https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/research_brief_-_global_insurance_protection_gaps.pdf (Accessed: 23 October 2022)

7. Platteau, JP and Ontiveros DU (2013). 'Understanding and information failures in insurance: Evidence from India', Institute for Advanced Development Studies (INESAD), La Paz , Development Research Working Paper Series No. 07/2013, pp. 4. Available at https://www.econstor.eu/handle/10419/106345 (Accessed: 23 October 2022)

8. Lee, R (2022). 'AI can perpetuate racial bias in insurance underwriting'. Available at https://money.yahoo.com/ai-perpetuates-bias-insurance-132122338.html (Accessed: 25 December 2022)

9. Bodine, R (2022). 'Male vs. Female Auto Insurance Rates [New Study + Surprise Results]'. Available at https://www.autoinsurance.org/man-vs-woman-car-insurance/ (Accessed: 25 December 2022)

10. Garrett, D et al (National Women's Law Center) (2012). 'Turning to Fairness – Insurance discrimination against women today and the Affordable Care Act'. Available at https://nwlc.org/wp-content/uploads/2015/08/nwlc_2012_turningtofairness_report.pdf (Accessed: 25 December 2022)

11. Cohen, WA et al (2019). 'Navigating Insurance policies in the United States for gender-affirming surgery', PRS Global Open 2019 Dec; 7(12): e2564, pp 1.
doi: 10.1097/GOX.0000000000002564

12. Wikipedia. 'Insurance in the United States'. Available at https://en.wikipedia.org/wiki/Insurance_in_the_United_States#cite_note-fio-1 (Accessed: 26 February 2023)

13. Rawat, S et al (2021). 'Application of machine learning and data visualization techniques for decision support in the insurance sector'. Available at https://reader.elsevier.com/reader/sd/pii/S2667096821000057?token=6C28D99838523C244F98E209252247D871E2D79DF6F10FF91E07792A8713F9387164586F1098A5AB9418F989F651BFA5&originRegion=eu-west-1&originCreation=20221225045832 (Accessed: 25 December 2022)

14. Carannante, M et al (2022) 'Disruption of Life Insurance Profitability in the Aftermath of the COVID-19 Pandemic', mdpi/Journals/Risks/Volume 10/Issue 2.
doi: https://doi.org/10.3390/risks10020040

15. Jain, N. 'Towards Machine Learning: Alternative Methods for Insurance Pricing – Poisson-Gamma GLM's, Tweedie GLM's and Artificial Neural Networks'. Available at https://www.actuaries.org.uk/system/files/field/document/F7%20Navarun%20Jain.pdf (Accessed: 25 December 2022)

16. Shimao, H and Huang, F (2022). 'Welfare implications of fairness and accountability for Insurance pricing', *UNSW Business School Research Paper Forthcoming*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4225159 (Accessed: 24 October 2022)

17. Bakko and Katari (2020). 'Transgender-Related Insurance Denials as Barrier to Transgender Healthcare: Differences in Experience by Insurance Type', pubmed.gov, J Gen Intern Med/2020 Jun/35(6):1693-1700.
doi: 0.1007/s11606-020-05724-2

18. Rebert, L and Hoyweghen IV (2015). 'The right to underwrite gender: The goods & services directive and the politics of Insurance pricing', Amsterdam University Press - A-Z Publications- Tijdschrift voor Genderstudies, Volume 18, Issue 4, Dec 2015, p. 413 – 431, pp 413. doi: https://doi.org/10.5117/TVGN2015.4.REBE.

19. Brodolini, F and Calafa, L and Bonardi, O (2011). 'The Use of Gender in Insurance Pricing', Think Tank/European Parliament/Research/Advanced search. Available at https://www.europarl.europa.eu/thinktank/en/document/IPOL-FEMM_NT(2011)453214 (Accessed: 30 December 2022)

20. Paruchuri, H (2020). 'The impact of Machine Learning on the future of Insurance industry', American Journal of Trade & Policy – Archives/Vol. 7 No. 3(2020): September-December Issue/ Policy and Practice Reviews, pp 86.
doi: https://doi.org/10.18034/ajtp.v7i3.537

21. Blier-Wong, C (2020). 'Machine Learning in P&C Insurance: A Review for Pricing and Reserving', mdpi / Journals /Risks / Volume 9 / Issue 1 / 10.3390/risks9010004, pp4.
doi: https://doi.org/10.3390/risks9010004

22. Matar, R (2022). 'Testing fairness in Insurance', University of Quebec in Montreal - Archipelago /Faculty of Science/Department of Mathematics/pp. 22. Available at https://archipel.uqam.ca/15451/1/M17534.pdf (Accessed: 24 October 2022)

23. Lindholm, M et al (2022). 'A Discussion of Discrimination and Fairness in Insurance Pricing', arxiv/Computer Science/Machine Learning/pp.8.
doi: https://doi.org/10.48550/arXiv.2209.00858

24. Lindholm, M et al (2021). 'Discrimination-Free Insurance Pricing', Cambridge University Press / Journals /ASTIN Bulletin: The Journal of the IAA/pp.32. Available at

https://www.cambridge.org/core/journals/astin-bulletin-journal-of-the-iaa/article/discriminationfree-insurance-pricing/ED25C4053690E56050F437B8DF2AD117 (Accessed on 24 October 2022)

25. Barry, L and Charpentier A (2022). 'The Fairness of Machine Learning in Insurance: New Rags for an Old Man?', arxiv.org/ General Economics (econ.GN); Computers and Society (cs.CY). doi: https://doi.org/10.48550/arXiv.2205.08112

26. Huang, S and Salm, M (2019). 'The effect of a ban on gender-based pricing on risk selection in the German health insurance market', Wiley Online Library / Health Economics /Volume 29, Issue 1/pp. 12-13. doi: https://doi.org/10.1002/hec.3958

27. Bereketoglu, AB (2022). 'Gender Rating & Regional Effect on Insurance Price', Research Square/pp7. doi: https://doi.org/10.21203/rs.3.rs-1942934/v1

28. Abachi, J (2018). 'Factors that influence pricing of Life Insurance products: A case study of Icea Lion Life Assurance Company', USIU-A Digital Repository Home / These and Dissertations / Chandaria School of Business/ pp50. Available at http://41.204.183.105/handle/11732/3944 (Accessed: 25 October 2022)

29. Chan, CY (2014). 'The Impact of Gender-neutral Pricing on the Life Insurance Industry', scor.com (Cass Business School, City University London)/ pp.48. Available at https://www.scor.com/sites/default/files/2015_uk_chancho-yeung.pdf (Accessed: 25 October 2022)

30. Davenport et al (2019). 'How artificial intelligence will change the future of marketing', link.springer.com, Journal of the Academy of Marketing Science volume 48, pages24–42. Available at https://link.springer.com/article/10.1007/s11747-019-00696-0 (Accessed: 25 October 2022)

31. Cather, D (2021). 'Insurance Pricing Discrimination and Aristotelian Equality: An Application to Minority Annuity Pricing), ssrn/ Journal of Insurance Regulation 2021/pp. 26/27. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3923078(Accessed: 25 October 2022)

32. Loi, M and Christen, M (2021). 'Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector', SpringerLink / Open Access / Philosophy and Technology 34, 967-992(2021)/pp 24. Available at https://link.springer.com/article/10.1007/s13347-021-00444-9 (Accessed: 26 October 2022)

33. ABI/Oxera (2010). 'The Use of Gender in Insurance Pricing', https://www.oxera.com/wp-content/uploads/media/oxera_library/the-use-of-gender-in-insurance-pricing.pdf.

34. Grari, V et al (2020). 'A fair pricing model via adversarial learning', arxiv / Statistics / Machine Learning/ pp.16.
doi: https://doi.org/10.48550/arXiv.2202.12008

35. Zhou, Q, Marecek, J and Shorten, RN (2021). 'Fairness in Forecasting and Learning Linear Dynamical System', AAAI / Archives / Vol. 35 No.12: AAAI-21 Technical Tracks 12 / AAAI Technical Track on Machine Learning V/ pp. 7.
doi: https://doi.org/10.1609/aaai.v35i12.17328

36. Davenport, JH (2017). 'The debate about "algorithms"', University of Bath / Mathematics Today / no. August, / pp 9.
Available at https://purehost.bath.ac.uk/ws/portalfiles/portal/156541073/JHDforMathTodayPostRW.pdf (Accessed: 25 October 2022)

37. Arelly, O and Montserrat, G (2013). 'A Comparison between General Population Mortality and Life Tables for Insurance in Mexico under Gender Proportion Inequality', Revista de Métodos Cuantitativos para la Economía y la Empresa, vol. 16, diciembre-, 2013, pp 65. Available at https://www.redalyc.org/pdf/2331/233129568003.pdf (Accessed: 26 October 2022)

38. Mosley, R and Wenman, R (2022). 'Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance', Casualty Actuarial Society / CAS Research Paper Series on Race and Insurance Pricing / pp 22. Available at https://www.casact.org/sites/default/files/2022-03/Research-Paper_Methods-for-Quantifying-Discriminatory-Effects.pdf (Accessed: 26 October 2022)

39. Lindholm, M et al (2022). 'A multi-task network approach for calculating discrimination-free Insurance price', Arxiv / Computer Science / Machine Learning / pp 22-23.

doi: https://doi.org/10.48550/arXiv.2207.02799

40. Henckaerts, R (2021). 'Insurance pricing in the era of machine learning and telematics technology', lirias.kuleuven.be / Faculty of Economics and Business / pp 103. Available at scholar.google.com (Accessed: 26 October 2020)

41. Frees, EW and Huang F (2021). 'The Discriminating (Pricing) Actuary', Taylor&Francis Online / All Journals / North American Actuarial Journal / List of Issues / Latest Articles / pp 21. doi: https://doi.org/10.1080/10920277.2021.1951296

42. Kotb, MH and Ming, R (2021). 'Comparing SMOTE Family Techniques in Predicting Insurance Premium Defaulting using Machine Learning Models', Proquest / International Journal of Advanced Computer Science and Applications / West Yorkshire / Vol. 12 Iss.9 / pp 1-10. doi: 10.14569/IJACSA.2021.0120970

43. Chancel, A et al (2022). 'Applying Machine Learning to Life Insurance: Some knowledge sharing to muster it', arxiv / Statistics / Machine Learning / pp 12-51. doi: https://doi.org/10.48550/arXiv.2209.02057

44. Krah, A-S and Nikolic Z and Korn R (2018). 'A Least-Squares Monte Carlo Framework in Proxy Modeling of Life Insurance Companies', mdpi / Journals / Risks / Volume 6 / Issue 2 / pp 1-24. doi: https://doi.org/10.3390/risks6020062

45. Mackenzie, L (2019) 'Digitizing difference: Fraudulence, Gender non-conformity and Data', University Digital Conservancy Home / University of Minnesota Twin Cities / Dissertations and Theses / Dissertations / pp 120. Available at https://conservancy.umn.edu/handle/11299/202920 (Accessed: 26 October 2022)

46. Vadapalli, P (2022). '6 Types of Regression Models in Machine Learning You Should Know About'. Available at https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/#:~:text=The%20two%20major%20types%20of,regression%20and%20multiple%20linear%20regression. (Accessed: 7th February 2023)

47. Knoldus.com (2023). 'Lasso and Ridge Regression'. Available at https://blog.knoldus.com/lasso-and-ridge-regression/. (Accessed: 8th February 2023)

48. Javatpoint.com(2023).'ML Polynomial Regression. Available at https://www.javatpoint.com/machine-learning-polynomial-regression. (Accessed: 8th February 2023)

49. Mwiti, D (2023). 'Random Forest Regression: When Does It Fail and Why?". Available at https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why. (Accessed: 8th February 2023)

50. Masui, T(2022). 'All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression'. Available at https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502. (Accessed: 7th February 2023)

51. Gurucharan, MK (2020). 'Machine Learning Basics: Decision Tree Regression'. Avaiilable at https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda#:~:text=Decision%20Tree%20is%20one%20of,with%20three%20types%20of%20nodes. (Accessed: 8th February 2023)

52. Muhajir, I (2019). 'K-Neighbors Regression Analysis in Python'. Available at https://medium.com/analytics-vidhya/k-neighbors-regression-analysis-in-python-61532d56d8e4. (Accessed: 8th February 2023)

53. Ogunbiyi, I (2022). 'Top Evaluation Metrics for Regression problems in Machine Learning'. Available at https://www.freecodecamp.org/news/evaluation-metrics-for-regression-problems-machine-learning/. (Accessed: 8th February 2023).

54. Brownlee,J (2021). 'Regression Metrics for Machine Learning'. Available at https://machinelearningmastery.com/regression-metrics-for-machine-learning/#:~:text=As%20such%2C%20it%20may%20be,y_i%20%E2%80%93%20yhat_i)%5E2. (Accessed: 8th February 2023).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

APPENDIX A: RESEARCH PLAN

TABLE 3: RESEARCH PLAN

Research Paper				
Activity	**START DATE**	**DUE DATE**	**% COMPLETE**	**DONE**
Literature review	**11/03/2022**	**11/25/2022**	**100%**	**Done**
Summarize gaps	**11/28/2022**	**12/02/2022**	**100%**	**Done**
Identify solution approach	**12/05/2022**	**12/16/2022**	**100%**	**Done**
Access dataset and complete data pre-processing and visualization	**12/19/2022**	**12/23/2022**	**100%**	**Done**
Prepare and submit Interim report	**12/26/2022**	**01/04/2023**	**100%**	**Done**
Build , Train and Test machine learning models	**01/05/2023**	**02/03/2023**	**100%**	**Done**
Model evaluation and finalization	**02/06/2023**	**02/10/2023**	**100%**	**Done**
Prepare and submit final research Paper and video presentation	**02/13/2023**	**03/01/2023**	**100%**	**Done**

TABLE 4: RISKS AND MITIGATIONS

Risk	**Impact**	**Mitigation**	**Status**
Relevant research papers not available/accessible	Lack of reference work	Prepare ground-up PoV	Adequate availability of existing research work; review completed
Dataset not available	Lack of real data	Synthetic data preparation	Dataset available
Solution approaches not feasible	Change of direction	Identification of new solution approach	Original solution approach slightly tweaked by introducing synthetic datasets to test concept
Aggressive timelines for submission	Failure to submit on time	Diligent work and constant calibration of timelines	On track

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181

APPENDIX B: REQUIREMENTS RESOURCES

A. Access to internet
B. Access to research paper repositories
C. LJMU e-Library
D. Access to the dataset to be used for the model evaluation
E. MacBook Air with macOS 11.2.3 (Apple M1; GPU;8 Core; 8 GB Memory)
F. Python3.8.2
G. Anaconda-Navigator ($PKG_VERSION)
H. Jupyter Notebook 6.3.0
I. Python libraries
    a. Pandas
    b. Numpy
    c. Seaborn
    d. Matplotlib
    e. Sklearn
J. Microsoft Word
K. Microsoft Excel
L. Microsoft PowerPoint