Data Analytics Capstone Topic: Medical Insurance Cost Prediction.

(With focus on Fairness, Interpretability and Data Augmentation)

Final Presentation

Name: Ketan Bhatt

Mentor’s name: Dr Sagarika Borah

Executive summary

- **Objective and scope:**
  - **Goal:** Build an interpretable and transparent regression model for predicting medical insurance costs.
  - **Data:** Uses demographic, health, and synthetic lifestyle features.
  - **Focus:** Develop fair and explainable models with ethical mitigation strategies.
  - **Evaluation:** Assess subgroup fairness, especially across gender and regions, using XGBoost.
  - **Analysis:** Study the effect of synthetic features on model performance and interpretability.
- **Importance:** Rising healthcare costs and risk differences among people make it important to predict expenses fairly and accurately, so insurance and financial plans can stay affordable and accessible for everyone.
- **Data:** Utilized a deduplicated, encoded dataset of 1,337 insured individuals expanded with engineered features (interaction terms, bins, risk flags) and synthetic clinical/lifestyle variables (e.g., blood pressure, cholesterol, lifestyle habits etc.), growing the dataset to millions of synthetic records to represent a real world problem for robust modelling.
- **Final Results:** Achieved significant predictive accuracy gains with XGBoost as well as fairness controlled as well as custom Sub-group models, improving R^2 from 0.85 to 0.98 (and other metrics also); fairness improved while and interpretability remained consistent with mitigation techniques and SHAP analysis confirming stable key drivers and reduced subgroup bias across all demographics.
- **Usage:** Delivered a scalable, fairness-aware, explainable set of model training approach applicable to real-world insurance settings, ensuring equitable insurance pricing, regulatory compliance, and enhanced trust for insurers and insured parties alike.

Gap Analysis

Study / Source	Achievements & Focus	Limitations / Gaps	Way Forward / Addressed by this Project
Kaggle Dataset Analysis (2022)	Baseline linear regression on demographic/health features	Lacks fairness audits, interpretability, documentation	Extend with SHAP, fairness diagnostics, audit logs
IEEE Medical Charges ML (2023)	XGBoost with feature engineering for accuracy	No SHAP interpretability or fairness audits	Integrate SHAP-based feature attribution, fairness
Nature Digital Medicine on Fairness (2022)	Fairness metrics, bias mitigation techniques	No modular fairness audits for insurance cost pipeline	Modularize audits for insurance pricing context
arXiv SHAP in Healthcare (2023)	SHAP feature importance visualization	No integration into fairness-auditable pipelines	Embed SHAP in audit logs & fairness overlays
Springer Traditional Regression (2021)	Linear/logistic for premium pricing	No fairness or interpretability features	Modernize with interpretable ML and fairness diagnostics
ScienceDirect Ensemble Models (2022)	Random Forest and boosting for prediction	No SHAP, fairness checks, or stakeholder-ready outputs	Add interpretability and fairness layers

Gap Analysis (contd.)

Study / Source	Achievements & Focus	Limitations / Gaps	Way Forward / Addressed by this Project
MDPI Synthetic Feature Engineering (2023)	Synthetic features to improve generalization	No SHAP or fairness diagnostics	Combine synthetic features with SHAP and fairness
OsloMet Regression Metrics (2022)	Compared RMSE, MAE, R^2 for healthcare models	No fairness or interpretability coverage	Extend to include fairness and SHAP
ACM Gender Bias in Insurance ML (2023)	Identified gender bias, proposed fairness metrics	No modular audit or stakeholder narratives	Build audit modules with markdown-native transparency
arXiv SHAP vs LIME Comparison (2023)	Compared interpretability techniques, found SHAP consistent	No embedding in fairness-auditable models	Operationalize SHAP within fairness/audit frameworks

Research Questions

- **The identified Research Questions are:**

1) How can we develop an interpretable, transparent and comprehensive regression model for predicting medical insurance costs using demographic, health, and synthetic lifestyle data?
2) How can we systematically assess, mitigate, and validate fairness in medical insurance cost prediction models across demographic subgroups using interpretable metrics and ethical remediation strategies?
3) To what extent does the XGBoost model exhibit demographic fairness across gender and regional groups when predicting insurance costs?
4) How do synthetic lifestyle features and contextual enhancements affect predictive performance and interpretability?

Research questions – Null and Alternative Hypothesis

Hypothesis	Description	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
H1a	Synthetic features improve model accuracy (performance)	Synthetic features do **not** improve model accuracy	Synthetic features **improve** model accuracy
H1b	Synthetic features improve model interpretability	Synthetic features do **not** improve model interpretability	Synthetic features **improve** model interpretability
H2	XGBoost exhibits subgroup bias beyond global metrics	No significant subgroup bias beyond global metrics	Significant subgroup bias exists beyond global metrics
H3	Mitigation improves fairness without hurting overall performance	Mitigation does not improve fairness or harms performance	Mitigation improves fairness without lowering performance
H4	SHAP feature importance remains stable with added contextual data	No significant change in SHAP importance across stages	Significant change in SHAP importance reflecting interpretability stability

Statistical Methods Used: Paired t-tests (ttest_rel) applied to cross-validation fold results and subgroup metrics: RMSE, MAE, R^2 for accuracy and fairness; Mean Absolute SHAP values for interpretability.
Significance level: $\alpha=0.05$*

*: The significance level (alpha) was set to 5% because it is a conventional threshold in non-life-critical medical research that balances the risk of false positives (Type I error) with detecting meaningful effects, providing reasonable confidence in results without requiring overly stringent criteria.

Data description – background

- Data: (Data source: https://gts.ai/dataset-download/medical-insurance-cost-prediction/)
(alternate source: https://www.kaggle.com/datasets/geethamgampa/medical-insurance-cost-prediction)
- It is a Medical insurance cost data collated from U.S. demographics, featuring attributes such as age, BMI, smoking status, number of children, region, and sex.
- The data contained 2772 records, which fully complies the requirement of 568 records as per the 95% confidence interval calculations (since the nature of Research Questions are related and requires a common analysis a lot, the Confidence Interval calculation remains the same for all RQs)
- While, processing the data, we realised that it contains huge no. of duplicate records (1435 of 2772), which were removed, leaving 1337 records. This still complies the CI requirement of 568 records for the given variation in the target feature (the CI method).

However, this data is too simplistic and requires feature engineering & synthetic data augmentation so that it:

- Improves model accuracy and generalization: Feature engineering creates meaningful variables from raw data, helping models capture complex healthcare patterns and make better predictions.
- Enhances fairness and interpretability: Transforming and augmenting features reduces bias and provides clearer insights, which are critical in medical decision contexts.
- Increases dataset diversity and robustness: Synthetic data augmentation expands patient profiles, enabling models to learn from varied scenarios and perform reliably in real-world settings.

Data description – Original Data

- The data included following features:
- **age:** Age in years of the policyholder (full number)
- **sex:** Gender of the policyholder (Male/Female)
- **bmi:** A ratio indicating body fat based on weight and height (represents Body Mass Index)
- **children:** Number of dependents insured under the policy
- **smoker:** Smoking status (Yes/No)
- **region:** The geographical area of residence (northeast, northwest, southeast, southwest)
- **charges:** Medical expenses billed to the individual – Target attribute

Please find the details of the feature engineering in the next slide

Data description – Feature Engineered data

Purpose of Feature Engineering: Feature engineering refines raw data into structured, meaningful variables that improve model learning, predictive accuracy, and support analyses such as fairness and interpretability.

Category	Feature	Feature Meaning	Method of Calculation	How it Adds Value
Feature Binning	age_bin	Age grouped into discrete risk categories	Categorize age into five bins (e.g., 0-18, 19-35, etc.)	Captures non-linear age effects; aids subgroup analysis
	bmi_bin	BMI grouped into obesity and risk levels	Turn BMI into five bins: extreme thinness(0), underweight(1), normal(2), overweight(3), obese(4).	Models distinct health risk classes more effectively
Feature Interaction	age__bmi	Combined effect of age and BMI	Multiply normalized age and BMI	Captures compounded risk effects for age-BMI interaction
	age__smoker	Joint impact of age and smoking status	Multiply age by binary smoker flag	Enhances modeling of amplified risk in older smokers
	bmi__smoker	Joint impact of bmi and smoking status	Multiply bmi by binary smoker flag	Enhances modeling of amplified risk in heavy weight smokers
Behavioural Flag	is_high_risk	Flag identifying high-risk individuals	Smoker, Age>60, BMI>35	Focuses model learning on high-cost patient groups
	is_young_smoker	Flag for young individuals who smoke	Smoker, Age<30	Enables focused risk stratification for early onset risks

Data description – Synthetic Data Augmentation

Augmentation Strategy and Purpose: We use a custom augmentation class to generate balanced, realistic synthetic features that reflect real-world insurance risk profiles. This approach expands dataset diversity and addresses subgroup representation, enhancing model learning, fairness, and robustness.

Strategy

- Custom class creates synthetic clinical and lifestyle features (e.g., blood pressure, lifestyle scores).
- Stratified augmentation ensures balanced representation across age, BMI, smoker status, and synthetic features.
- Adds synthetic features like sum insured and activity metrics to simulate real-world insurance scenarios.
- Ensure that we create these synthetic features independent of current or other synthetic features.

Purpose

- Increase dataset size and feature diversity for better learning of complex risk patterns.
- Improve model generalization and reduce overfitting with broader patient profiles.
- Support fairness analysis by ensuring minority and high-risk subgroup representation.
- Build actionable, trustworthy models for insurance pricing and risk stratification.

Data description – Synthetic Data Augmentation

Category and Feature wise explanation of Synthetic feature based data augmentation:

Category	Feature	Value Range / Levels	Step Size	Impact on Charges (USD)	Rationale
Health Indicators	systolic_bp_syn	0 = Normal, 1 = Elevated, 2 = Hypertensive	1	0 → 1500 → 4000	Higher blood pressure increases cost
	cholesterol_syn	0 = Normal, 1 = Borderline, 2 = High	1	0 → 1000 → 3000	Elevated cholesterol adds risk
	fasting_blood_sugar_syn	0 = Normal, 1 = Prediabetic, 2 = Diabetic	1	0 → 1500 → 3500	Diabetes significantly raises cost
Policy Related	sum_insured_syn	USD 100,000 to USD 1,000,000	300,000	4% of insured amount	Direct cost proportional to coverage

Data description – Synthetic Data Augmentation (contd.)

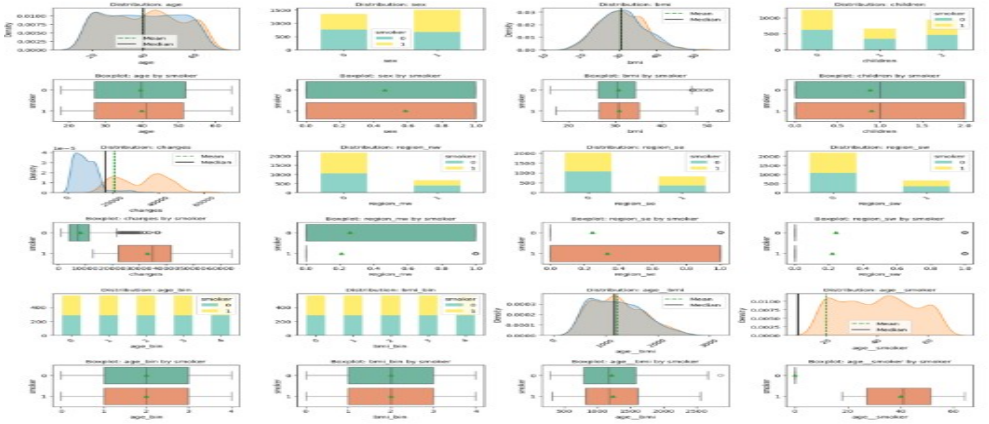
Category	Feature	Value Range / Levels	Step Size	Impact on Charges (USD)	Rationale
Lifestyle Factors	health_habits_syn	0 = Poor, 1 = Healthy	1	2000 → 0	Composite of steps/day, gym, exercise, calories burnt, etc.
	work_stress_syn	0 = Low, 1 = High	1	0 → 2000	Higher work stress increases health risk
	sleep_quality_syn	0 = Poor, 1 = Good	1	1800 → 0	Better sleep reduces health risk
	alcohol_consumption_syn	0 = None, 1 = Moderate, 2 = Heavy	1	0 → 1000 → 3000	Higher alcohol use increases cost

Exploratory Data Analysis

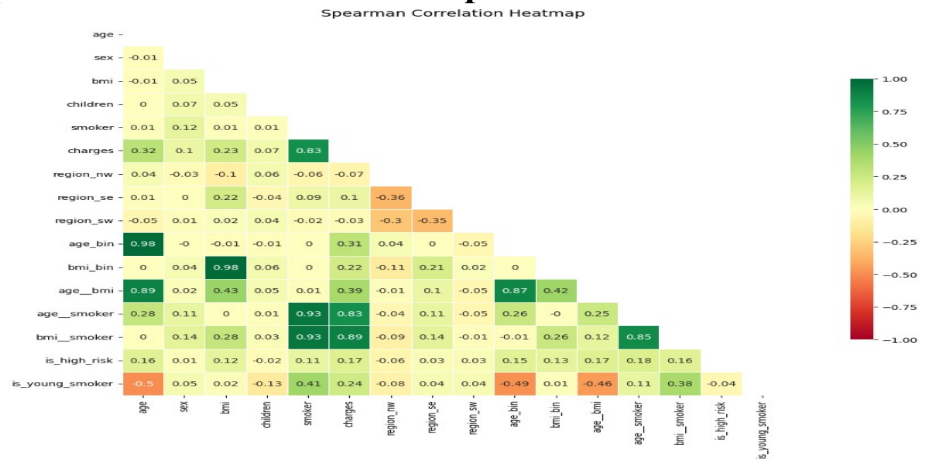
Enhanced Descriptive Statistical report

	age	sex	bmi	children	smoker	charges	region_nw	region_se	region_sw	age_bin	bmi_bin	age_bmi	age_smoker	bmi_smoker	is_high_risk	is_young_smoker
count	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000	2850.000
mean	39.819	0.525	30.734	0.896	0.500	20592.966	0.235	0.292	0.231	2.000	2.000	1223.203	19.998	15.399	0.011	0.142
std	14.274	0.499	6.039	0.870	0.500	15186.987	0.424	0.455	0.422	1.414	1.414	501.702	22.457	15.975	0.105	0.350
min	17.911	0.000	15.983	0.000	0.000	1127.264	0.000	0.000	0.000	0.000	0.000	286.054	0.000	0.000	0.000	0.000
25%	27.030	0.000	26.407	0.000	0.000	7247.734	0.000	0.000	0.000	1.000	1.000	803.887	0.000	0.000	0.000	0.000
50%	40.015	1.000	30.558	1.000	0.500	16642.526	0.000	0.000	0.000	2.000	2.000	1173.097	8.955	8.579	0.000	0.000
75%	51.789	1.000	34.800	2.000	1.000	34957.163	0.000	1.000	0.000	3.000	3.000	1592.894	40.917	30.654	0.000	0.000
max	64.314	1.000	52.718	2.000	1.000	63963.461	1.000	1.000	1.000	4.000	4.000	2857.136	64.299	52.647	1.000	1.000
cv	0.358	0.951	0.196	0.971	1.000	0.737	1.804	1.559	1.824	0.707	0.707	0.410	1.123	1.037	9.386	2.454
IQR	24.758	1.000	8.393	2.000	1.000	27709.429	0.000	1.000	0.000	2.000	2.000	789.007	40.917	30.654	0.000	0.000
skew	-0.041	-2.852	0.088	-0.357	0.000	0.780	1.663	1.924	1.645	0.000	0.000	0.300	1.475	1.281	0.320	1.223
upper_lim	57.676	1.000	36.921	2.500	1.250	44114.482	0.000	1.500	0.000	3.500	3.500	1802.792	56.898	41.692	0.000	0.000
lower_lim	20.538	-0.500	24.332	-0.500	-0.250	2550.338	0.000	0.000	0.000	0.500	0.500	619.282	-4.478	-4.289	0.000	0.000
outlier_spread	1.249	0.667	2.918	0.667	0.667	1.512	inf	0.667	inf	1.333	1.333	2.172	1.048	1.145	inf	inf

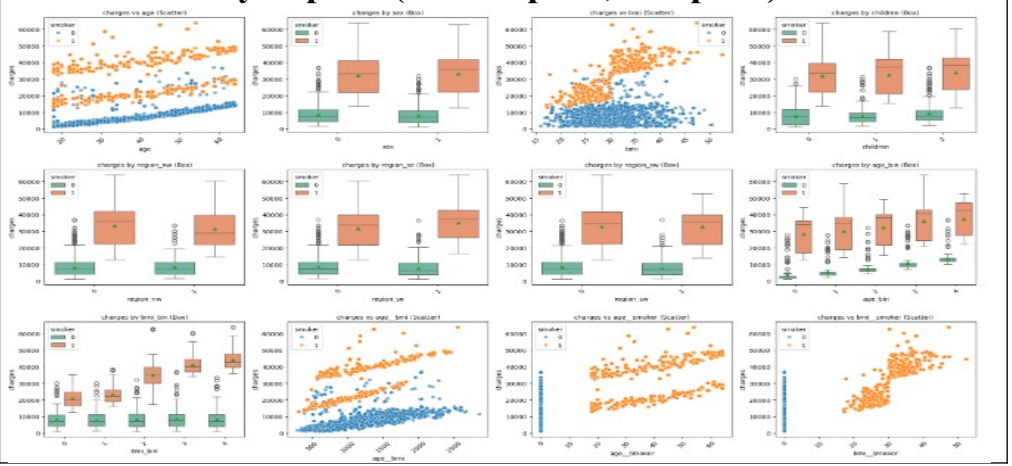
Univariate Analysis plots (KDE plots, count plots, box plots)



Spearman Correlation Heatmap



Bivariate Analysis plots (Scatter plots, box plots)



Exploratory Data Analysis: key findings

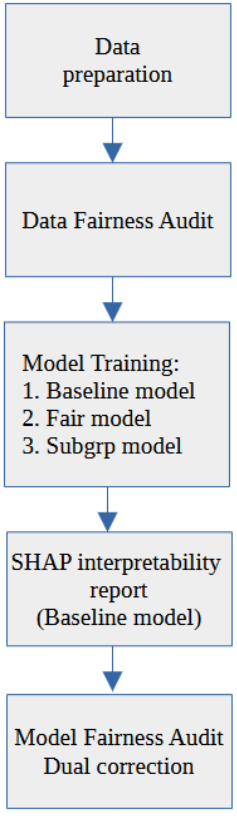
Stage 1: Original cleaned, encoded data	Stage 2: Feature Engineered, up-sampled data	Stage 3: Synthetic data augmentation
General Observations:		
- Size: 1377 rows x 9 columns - Dataset balanced in 'sex' and 'region' distribution. - 'smoker' minority (~20%). - 'charges' heavily right-skewed with high-cost outliers. - 'bmi' slightly right-skewed; with some outliers. - children concentrated at 0–2. - age evenly spread, no strong skew. - Regions roughly balanced.	- Size: 2850 rows x 16 columns - Little larger dataset (n=2850) with engineered features: (age_bin, bmi_bin, age__bmi, age__smoker, bmi__smoker, is_high_risk, is_young_smoker). - smoker proportion now 50% (balanced). - charges still right-skewed but mean higher (~20.6k). - is_high_risk rare (~1.1%). - is_young_smoker ~14%. - Age/BMI bins evenly distributed across categories.	- Age evenly spread; BMI slightly right-skewed - Regions roughly balanced across all dataset stages - Engineered features added in Stage 2 and 3 - Stage 3 features includes synthetic health and lifestyle metrics
Correlation / Influence on Target ('charges'):		
- Strongest positive correlation: smoker (~0.66 Spearman). - Moderate positive: age (~0.53), bmi (~0.20). - Negative: region_se and region_sw (~-0.35). - Bivariate plots show smokers have much higher charges across all subgroups. - Age and BMI both amplify smoker effect on charges.	- smoker still dominant driver of charges. - age and bmi retain positive influence. - Engineered interactions (age__smoker, bmi__smoker) show stronger correlation with charges than base features. - is_high_risk and is_young_smoker show distinct high charge clusters. Have low correlation with target. - age/bmi bins reveal stepwise increase in charges with higher bins, especially for smokers.	- Smoker status remains the second strongest positive driver of charges, the first one being the sum_insured_syn. - Age and BMI show consistent positive influence - Interaction terms (age__smoker, bmi__smoker) amplify predictive signal - Stage 3 lifestyle metrics may add marginal lift - Negative influence from certain regional categories persists

Exploratory Data Analysis: key findings (contd.)

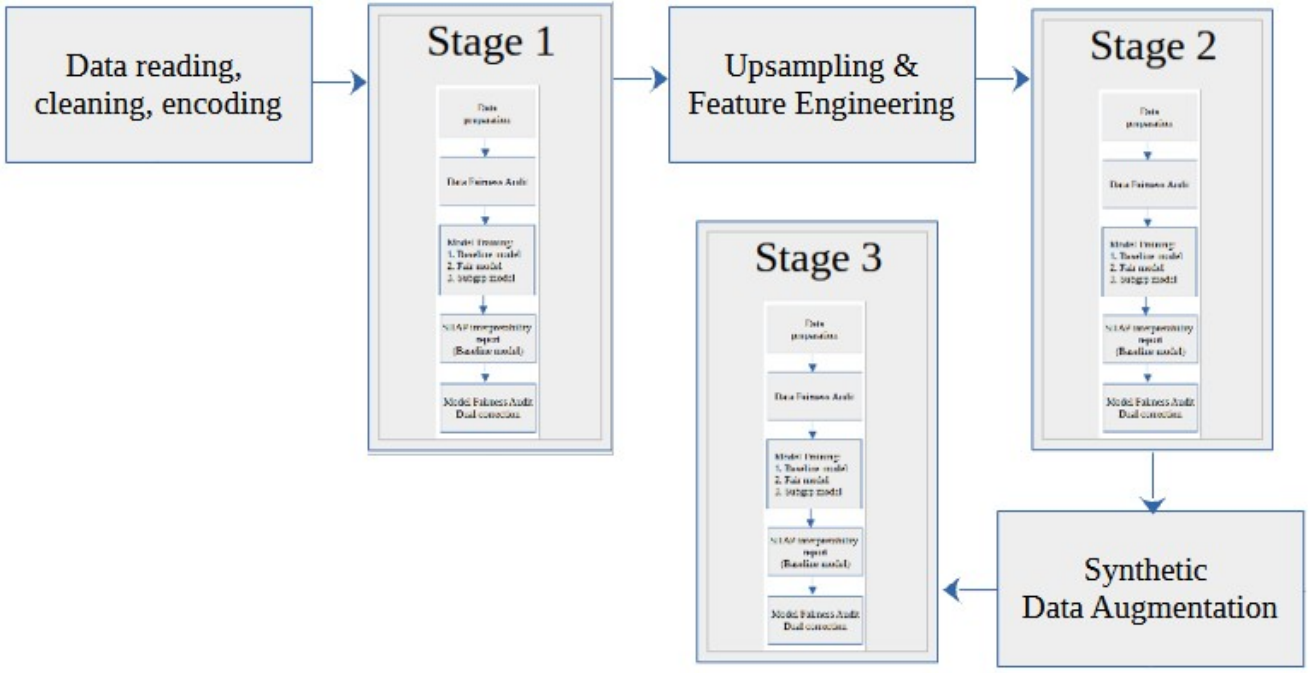
Stage 1: Original cleaned, encoded data	Stage 2: Feature Engineered, up-sampled data	Stage 3: Synthetic data augmentation
Multicollinearity:		
- Low collinearity among most base features. - Some mild correlation between age and charges, BMI and smoker. - Region dummies mutually exclusive but not collinear with other predictors.	- Engineered features introduce higher correlations: • age ↔ age_bin (perfect mapping). • BMI ↔ bmi_bin (perfect mapping). • Interaction terms (age__smoker, bmi__smoker) correlated with both components. - Binary risk flags (is_high_risk, is_young_smoker) partially collinear with smoker and age/BMI bins. - Still no severe collinearity among independent base features.	- Base features show low collinearity across all stages - Region dummies mutually exclusive, no severe correlation - Perfect mapping between bins and original variables - Interaction terms correlated with both component features - As per our strategy, Stage 3 synthetic metrics do not correlate with other features
Recommended actions ahead:		
- Based on the high correlation with target, maintain smoker as key predictor but prepare fairness checks due to group imbalance. - Consider binning of age and bmi for improving predictability of terget variable. - Explore interaction terms between smoker and age/BMI in Stage 2. - Validate outlier influence on model coefficients. - For training the model with fairness, consider balancing the smoker subgroups (0, 1) through up-sampling.	- Consider dimensionality reduction or selective inclusion for interaction terms to avoid redundancy. - Based on the correlation with target, consider dropping is_high_risk, is_young_smoker and few more engineered features subject to their SHAP values. - Use stratified modelling or subgroup analysis for sensitive features. - Investigate non-linear effects of age/bmi bins on charges. - Prepare fairness metrics for balanced smoker dataset to ensure no subgroup over-penalisation.	

Architecture diagram/Workflow

Stage workflow



Complete workflow



Model building: Our Model strategy

Our layered modeling pipeline consists of baseline XGBoost, Fairlearn-constrained model, and customized Hybrid subgroup models, sequentially addressing accuracy and fairness challenges in insurance cost prediction.

- The baseline XGBoost model establishes a performance benchmark without bias interventions, helping to identify initial predictive power and potential fairness issues.
- Fairlearn-constrained modeling applies global fairness constraints on the baseline model (using Python's Fairlearn library function), via in-processing algorithms, aiming to reduce bias across sensitive attributes while maintaining strong predictive accuracy.
- The custom made hybrid subgroup model class segments data by 'smoker' status, training specialized models per subgroup (smoker=0, smoker=1) to enhance both predictive precision and subgroup fairness.
- SHAP-based interpretability reveals key feature influences, supporting transparent and model independent explanations critical to ethical deployment and stakeholder trust.
- Model regularization uses strong L1 (reg_alpha=0.5) and L2 (reg_lambda=1.0) penalties, controlled tree depth(8), and subsampling(4) to prevent overfitting and improve generalizability.

Model building – Our fairness strategy

Phase	Strategy	Why	Impact / How it Helped
Before Training	Fairness audit	Detects existing biases in terms of dataset imbalances, influence of subgroup on target and extent of skew in the data	Identifies bias patterns and guides further bias mitigation steps by highlighting subgroup disparities clearly
	Upsampling sensitive groups	Mitigates class imbalance for minority groups	Helps reduce bias in training data, ensure better model training where each sensitive feature subgroup is given equal opportunity to equally contribute during training, leading to fairer subgroup performance
	Stratified train-test split	Maintains subgroup distribution in splits	Ensures fair evaluation and prevents leakage of subgroup bias
During Training	Fairlearn fairness-constrained model with Exponentiated Gradient	Uses Exponentiated Gradient to iteratively enforce fairness constraints with minimal accuracy loss	Maintains accuracy while reducing bias across sensitive features
	Custom Hybrid Subgroup Model	Specialized models per subgroup improve fairness	Achieves better subgroup-specific metrics (e.g., smoker subgroup RMSE reduced)
Post Training	Dual correction (Class: TrippleFactorCorrection)	Adjusts residual bias at global and subgroup levels based on gap between y & predicted y and subgroupR2 & Mean of subgroup R2	Lowers subgroup MAE by 20-50, stabilizes error variance, maintains overall prediction quality

Stage 1-> 2: Only Feature Engineering, only up-sampling or both

Scenario	Model	MSE	MAE	R^2	MAE CV (%)	MSE CV (%)	R^2 CV (%)
Original data	Baseline	18,633,840	2,644.95	0.85	2.83	15.26	6.59
	Fair	18,398,500	2,702.29	0.84	3.13	13.16	6.19
	Subgroup	18,870,450	2,581.14	0.84	3.01	11.71	6.07
Only feature-engineered, no upsampling	Baseline	17,961,410	2,445.86	0.84	5.34	13.34	6.07
	Fair	17,203,070	2,374.38	0.85	5.68	14.68	5.99
	Subgroup	19,010,640	2,447.72	0.83	1.24	13.14	6.47
Only upsampling, no feature engineering	Baseline	14,556,000	2,323.45	0.93	1.99	13.32	0.96
	Fair	14,284,070	2,297.26	0.93	0.28	12.83	0.90
	Subgroup	15,054,190	2,305.80	0.93	5.11	6.02	0.44
Both feature engineering and upsampling	Baseline	7,462,335	1,251.79	0.97	1.61	9.94	0.56
	Fair	7,596,041	1,273.24	0.97	0.64	18.34	0.86
	Subgroup	8,189,787	1,343.66	0.96	4.45	23.21	1.10

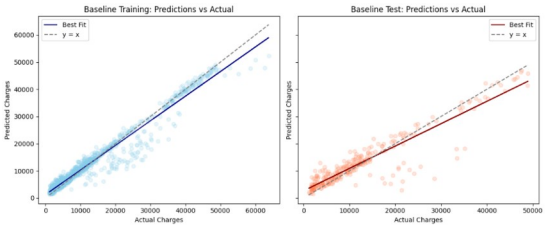
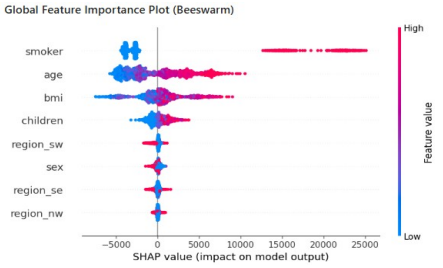
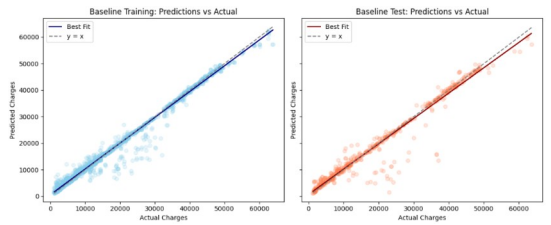
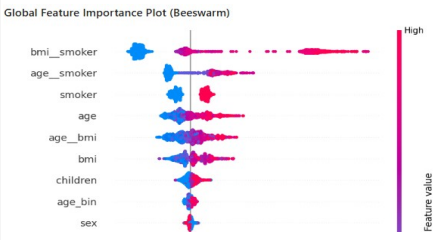
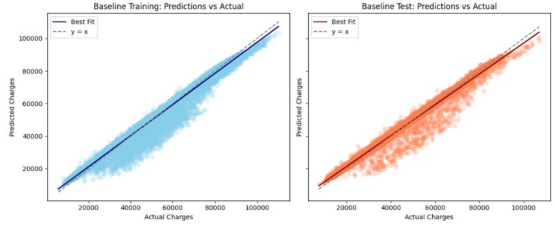
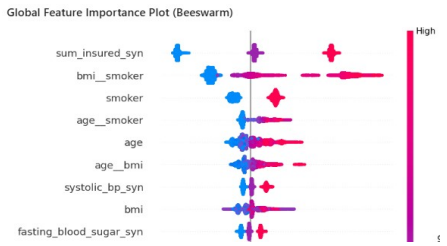
Key take-aways:

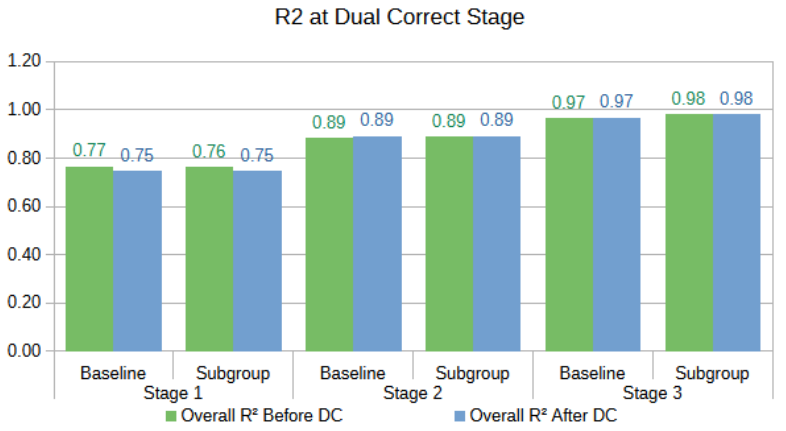
- Combining feature engineering and upsampling improves model accuracy by enhancing feature representation and balancing the data distribution.
- Applying fairness constraints like Fairlearn reduces prediction disparities across subgroups, ensuring more equitable model outcomes.
- The best practice is to integrate feature engineering, upsampling, and fairness-aware modeling together for optimal performance and ethical deployment.

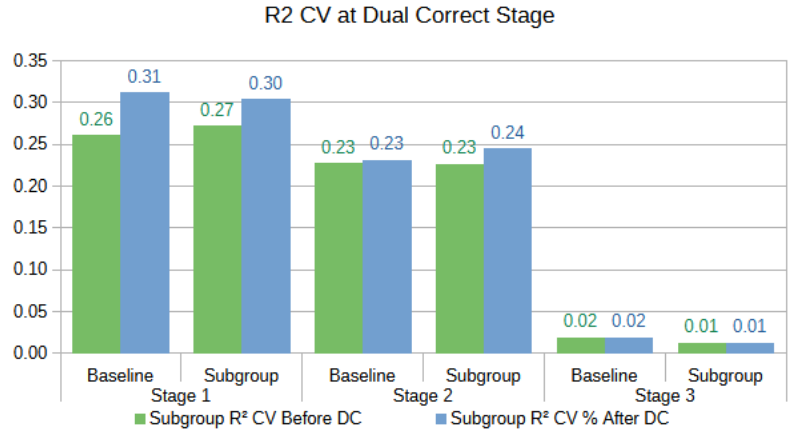
Results: Step by step findings

Stg	Data Preparation	Data Fairness Audit	Model Training	Interpretability Analysis	Model fairness audit/correction
1	- Read the data (2772x7). - Drop duplicates (1337x7). - Label encode sex, smoker. - One-hot encode region (1337x9).	- children show subgroup imbalance – may consider clubbing subgroups 3,4,5 into 2. - High smoker cost disparity with median change > $25K! - Regional group skewness.	- Baseline XGBoost Test R2: 0.85. - Fairness constraint model: similar R2, RMSE, slightly higher MAE. - Hybrid Subgroup Model: Maintains performance as baseline, improved performance for smoker subgroups. - All models show sign of minimal over-fit.	SHAP analysis reveals: - smoker: the highest feature influence of 45.6% - age is the next with 19.6% - bmi is next to follow with 14% - children, region, sex has lower influence on the target - charges	- Overall model accuracy metrics (RMSE, MAE, R^2) degrade slightly after dual correction as expected. - Dual correction mainly recalibrates biases at group levels but does not drastically alter the global fit metrics. - The fairness audit on subgroup model, is showing stabilized or slightly improved consistency.
2	- Merge children subgroup 3,4,5 into 2. Now 2 means 2+ - Upsample dataset for balancing smoker, age_bin, bmi_bin attributes (2850x9) - Added features explained earlier (2850x16)	- The smoker and bmi_bin, age_bin features are perfectly balanced. - Children and sex show moderate imbalance. - region features continue to show similar imbalance as stage 1. - Median charges for smokers continue to show a large difference.	- Baseline model shows good test performance (R2=0.95). - Fairlearn-constrained model maintains similar accuracy with slightly higher RMSE and MAE as fairness trade-off. - Hybrid subgroup models, tailored to smokers/non-smokers, maintains test R2 while achieving better subgroup fairness. (for smoker subgroup, achieves R2 of 1.0)	- bmi__smoker: top importance of 46.6%. age_smoker is the next one with 13.9% - so smoker continues to dominate. - age__bmi interactions at 9.2% shows continuation of influence of age & bmi. - First 3 most important features are Engineered. - age_bin, bmi_bin, is_high_risk, is_young_smoker – low influence. Can drop before next stage	- Dual correction led to slight decreases in overall and subgroup MAE, indicating modest error improvement. - Error variability (stddev) mostly decreased across groups, showing more consistent predictions. - Group-wise fairness metrics shifted minimally, reflecting stable and balanced subgroup performance. - Overall R^2 metrics showed a small increase, at group-level post-correction.
3	- Drop age_bin, bmi_bin, is_high_risk, is_young_smoker - Define synthetic feature dictionary. - Instantiate and run our own DataAugmentor to generate synthetic augmented data (16.6 million x 21). - Taken 50K random samples to run analysis with available memory.	- Subgroup distributions for synthetic features are well balanced. - Significant differences in median target across key sensitive groups. - No features show high skewness. - Data Augmentation did not disturb earlier features' balance levels. - Target medians indicate strong influence of smoker status and sum insured.	- Baseline and fairness-constrained models show strong overall test R2 (~0.97). - Fairlearn model slightly improves RMSE and MAE compared to baseline. - Hybrid subgroup model improves subgroup-specific accuracy, especially for smokers (R2=0.99). - All models maintain robust fit and controlled errors across sensitive groups.	- New feature sum_insured_syn is top feature with 22.5% importance. - The bmi__smoker interaction remains influential with 20.7%, against Stage 2's 30.6%. - The smoker feature importance increased to 8.9% from Stage 2's 7.6%, showing more direct influence. - children, sex and region has the lowest influence, indirectly suggesting model's fairness towards these.	- Dual correction improved overall MAE by about 50 units. - Group-wise fairness metrics changed minimally, indicating stable subgroup performance post-correction. - Slight decreases in error variability suggest improved prediction consistency for sensitive groups. - Overall R^2 remained stable with marginal improvement, confirming balanced fairness and accuracy.

Results Summary – Model Performance, Explainability, Fairness Audit

Stage	Baseline Model Performance (y_pred vs y-true)	Model explainability (SHAP values)
1	Training R2:0.95, Test R2: 0.85 	
2	Training R2: 0.99, Test R2: 0.95 	
3	Training R2: 0.98, Test R2: 0.97 	





Research questions - Test results

We have carried out these tests stage wise:

- Stage 1(S1): With original cleaned, encoded data;
- Stage 2(S2) : With Feature engineered data with upsampling of few sensitive attributes; and
- Stage 3 (S3): After data augmentation with Synthetic features.

Please find below summary of findings of these Hypothesis tests*:

Hypothesis	Test Outcome	t-statistic (approx.)	p-value (approx.)	Interpretation
H1a	Significant improvement in RMSE & R^2 with synthetic features; MAE improvement less clear	8.89 (S1->S2), 5.11 (S2->S3) for RMSE; 16.30 (S1->S2), -1.60 (S2->S3) for MAE	0.0009 (S1->S2), 0.0069 (S2->S3) for RMSE; 0.0001 (S1->S2), 0.18 (S2->S3) for MAE	Synthetic features improve accuracy significantly, especially between stages 1 and 2; MAE plateaus afterward.
H1b	No significant change in SHAP feature importance across stages	$t \sim 1.57$	$p > 0.05$	Interpretability remains stable and consistent despite addition of synthetic features and complexity.
H2	Significant subgroup bias detected beyond global metrics	Various values (e.g., sex t=13.35 p=0.00018 for MAE)	$p < 0.05$ for many subgroups	XGBoost exhibits subgroup bias; fairness issues revealed in subgroups despite global accuracy.
H3	Mitigation improves subgroup fairness without accuracy loss	Various values (mixed, some t>3, p<0.05)	Mostly < 0.05	Fairness mitigation via SubGrp+DC consistently improves fairness with no harm to overall accuracy.
H4	SHAP importance scores stable across feature stages	$t \sim 1.5$	$p > 0.05$	Feature importance (SHAP) stability maintained, supporting trustworthy model interpretability.

**: The complete hypothesis testing has been done independent of the main code.*

Implementation Strategy

- Developed modular code with custom classes/methods (TrainModel, FairnessAuditor, DualFactorCorrection, SHAPAnalyzer, DataAugmentor), and functions(explore.py, FairnessAnalysis.py, transformation.py) for easy execution of the stage pipeline for exploratory data analysis, model creation, training, fairness auditing, and interpretability analysis etc.
- Integrated Fairlearn's Exponentiated Gradient algorithm for in-processing fairness mitigation applied during model training.
- Introduced a Hybrid Subgroup Model class to train specialized sub-models on sensitive groups (e.g., smokers) to enhance subgroup fairness and accuracy.
- Incorporated dual correction using class TrippleFactorCorrection linear regression to post-process and adjust for residual global and subgroup biases in predictions.

User benefit matrix

Benefit Category	Benefit Description	Insurance Providers	Policy Analysts & Regulators	Consumer Advocates
Risk-Based Pricing & Transparency	- Adjust premiums by risk factors	✓		
	- Enable clear understanding of charges	✓	✓	✓
Fairness & Equity Assurance	- Ensure predictions treat all groups fairly	✓	✓	✓
	- Audit compliance	✓	✓	
Model Interpretability & Trust	- Explain model decisions with SHAP to build stakeholder trust	✓	✓	✓
Fraud & Anomaly Detection	- Detect suspicious claims or data inconsistencies	✓		
Policy Simulation & Insights	- Evaluate impact of policy changes, assess public health and equity		✓	✓
Personalized Consumer Guidance	- Provide feedback on lifestyle effects and plan comparison			✓
Community & Regulatory Support	- Tools for education, oversight, and fair pricing advocacy		✓	✓

Conclusion

Please find below the final conclusion of the Capstone project:

- High Predictive & Ethical Integrity: The final pipeline delivers robust predictive accuracy for insurance cost prediction while maintaining fairness across gender and regions, as confirmed by comprehensive audits and FairLearn evaluations.

- Proven Fairness Remediation: Strategies like DualCorrect and stratified data augmentation, implemented during the capstone, consistently reduced disparities and balanced subgroup outcomes, directly addressing the research questions on fairness mitigation.

- Interpretability & Transparency: SHAP analyses demonstrated meaningful, equitable contributions of synthetic and contextual features, enhancing both model interpretability and stakeholder trust.

- Aligned with Stakeholder Goals: The integrated approach fulfills requirements for ethical modeling, transparency, and actionable insights, validating the effectiveness of fairness corrections and explainability methods.

- Capstone Extensions: Recent work strengthened fairness validation and clarified the impact of synthetic lifestyle features, reinforcing model reliability and supporting informed insurance decisions.

Future work

Building on these findings, future work can focus on three key directions:

Production deployment with Automated monitoring and alerts

- Why: Implementing production deployment transforms the insurance prediction model from a research prototype into a real-world tool, with real data ensuring continuous operational value for stakeholders.
- What: Train these models with real data and deploy the prediction pipeline in production and establish automated flagging systems to detect model drift or emerging biases in production, ensuring early intervention if fairness or performance declines.

Robustness Testing

- Why: Insurance prediction models may encounter rare, adversarial, or out-of-distribution inputs that could challenge fairness and accuracy. Proactively testing these cases is crucial for understanding and mitigating vulnerabilities that aren't visible with standard evaluation.
- What: Introduce adversarial testing workflows (such as generating subtly modified or synthetic "edge case" samples) and out-of-distribution scenario analyses to systematically audit model stability and fairness under stress. This helps ensure reliable operation even in atypical or intentionally manipulated data settings.

Causal Fairness Exploration

- Why: Observing disparities does not always reveal their underlying causes; correlation-based fairness metrics may conflate genuine risk factors with structural bias. Causal analysis can disentangle model behaviors due to data characteristics from those resulting in unfair treatment.
- What: Incorporate causal inference tools (e.g., causal diagrams, structural equation modeling) to differentiate between direct, indirect, and spurious causes of group disparities. This enables the design of interventions targeted at true sources of unfairness, improving ethical remediation.

Stakeholder Feedback Loop

- Why: Maintaining trust and utility requires continuously involving domain experts in fairness auditing and model refinement. Interactive interfaces empower stakeholders to annotate edge cases, review explainability summaries, and directly influence remediation strategies.
- What: Develop visual dashboards and markdown-native summaries to facilitate expert review, annotation, and feedback collection. Use this input for iterative fairness adjustments and to prioritize real-world relevance in ongoing model updates.

References

1. Shi, J. (2013). Efficiency in plan choice with risk adjustment and premium discrimination in health insurance exchanges. Harvard University. URL: https://scholar.harvard.edu/files/shi/files/efficiency_in_plan_choice_in_exchanges_julie_shi.pdf

2. Le, D.H. (2024). Enhancing medical insurance pricing prediction with SHAP-XGBoost for informed decision-making. In: From Smart City to Smart Factory for Sustainable Future. Springer. URL: https://link.springer.com/chapter/10.1007/978-3-031-65656-9_32

3. Miyachi, T. (2022). A comparison of “fairness” in insurance underwriting in Japan and the U.S. Harvard Program on U.S.-Japan Relations. URL: https://us-japan.wcfia.harvard.edu/resource/22-mtmiyachitomokapdf

4. Hickey, J.M., Di Stefano, P.G. & Vasileiou, V. (2020). Fairness by explicability and adversarial SHAP learning. arXiv preprint, arXiv:2003.05330. URL: https://arxiv.org/abs/2003.05330

5. Jalali, M.S., DiGennaro, C., Guitar, A., Lew, K. & Rahmandad, H. (2020). Evolution and reproducibility of simulation modeling in health policy over half a century. Harvard Medical School. URL: https://scholar.harvard.edu/files/jalali/files/simulation_modeling_in_health_policy.pdf

6. Nohara, Y., Matsumoto, K. & Soejima, H. (2022). Explanation of machine learning models using SHAP and application for real hospital data. Computer Methods and Programs in Biomedicine, 214, 106584. URL: https://doi.org/10.1016/j.cmpb.2021.106584

7. Mihirette, S. & Tan, Q. (2022). SHAP algorithm for healthcare data classification. In: Hybrid Artificial Intelligent Systems. Springer. URL: https://link.springer.com/chapter/10.1007/978-3-031-15471-3_31

8. Lundberg, S. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. URL: https://arxiv.org/abs/1705.07874

9. Rosenberg, D. (2002). Deregulating insurance subrogation: Towards an ex ante market in tort claims. Harvard Law School, Olin Center Discussion Paper No. 395. URL: http://law.harvard.edu/programs/olin_center/papers/pdf/395.pdf

References (contd.)

10. Briones, F. (2025). Using machine learning and SHAP to predict ER wait times and walkouts in healthcare settings. GitHub repository. URL: https://github.com/fritzbriones/machine-learning-er-healthcare

11. Bird, S., Hutchinson, B., Kenthapadi, K., Kiciman, E. & Mitchell, M. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI systems. Microsoft Research. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

12. Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In: Proceedings of the 8th Innovations in Theoretical Computer Science Conference. URL: https://arxiv.org/abs/1609.05807

13. Pearl, J. (2009). Causality: Models, reasoning, and inference. Cambridge University Press. URL: https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B

14. American Academy of Actuaries. (2024). Drivers of 2025 health insurance premium changes. https://www.actuary.org/sites/default/files/2024-08/health-brief-2025-premium-changes.pdf

15. Deloitte Insights. (2025). 2025 US health care outlook. https://www.deloitte.com/us/en/insights/industry/health-care/life-sciences-and-health-care-industry-outlooks/2025-us-health-care-executive-outlook.html

16. Health System Tracker. (2024). How much and why ACA Marketplace premiums are going up in 2025. https://www.healthsystemtracker.org/brief/how-much-and-why-aca-marketplace-premiums-are-going-up-in-2025/

17. PwC. (2025). Medical cost trend: Behind the numbers. https://www.pwc.com/us/en/industries/health-industries/library/behind-the-numbers.html

Any questions please