<u>AI Governance Report</u>

Project: Health Insurance Premium Model
By: Shijin Ramesh
Date: 26th October, 2025

Executive Summary

System name: Health Insurance Premium Predictor
Owner: Shijin Ramesh
Deployment date: 24th October, 2025
Model family: Two-model ensemble by age segment

- **Model A:** Age > 25
- **Model B:** Age 18–25
  Primary objective: Predict health insurance premium for individuals.
  Intended use: Decision support for pricing; not standalone underwriting.
  Key risks: Pricing unfairness, drift, privacy breaches.
  Controls in place: Segmented models, fairness audits, SHAP explainability, monitoring, human review.

---

1. Governance Charter

Purpose

To ensure that the Health Insurance Premium Prediction system is developed, deployed, and monitored responsibly with a focus on fairness, accountability, transparency, privacy, and compliance. Even though this is an individual portfolio project, governance roles are clearly defined to simulate real-world standards and ensure lifecycle accountability.

Scope

This governance charter covers:

- Data sourcing and documentation
- Model development and validation
- Deployment and monitoring controls
- Fairness, privacy, and security assessments
- Incident response and continuous improvement

Roles & RACI:

- **Product Owner:** Define objectives, intended use, and business value - Self
- **ML Lead:** Model design, experimentation, validation - Self
- **Data Steward:** Data quality checks, preprocessing documentation, lineage tracking - Self
- **AI Ethics Reviewer:** Fairness evaluation, bias mitigation, explainability - Self
- **Incident Manager:** Manage alerts, issues, rollback, RCA - Self
- **Approver:** Authorize progression between lifecycle stages - Self

---

2. Model Cards (one per model)

Model Card - Model A (Age > 25)

Version: v1.0
Target: Premium amount (₹)
Features: Age, Number of Dependants, Income in Lakhs, Genetical Risk, Insurance Plan, Employment Status, Gender, Marital Status, BMI Category, Smoking Status, Region, Medical History
Algorithms: XGBoost Regressor
Performance (holdout): R^2=99.5%, RMSE=₹ 489.55
Calibration: No explicit calibration applied. Prediction error distribution was validated using residual analysis and RMSE metrics to ensure alignment between predicted and actual premium ranges.
Known limitations:

- The model is trained on historical patterns and may not generalize well to new pricing rules or regulatory changes in the insurance market.
- Limited data representation for individuals below age 25 required separate segmented modeling.
- Premiums for individuals with extreme BMI values, rare health-risk profiles, or very high medical expenses may show higher prediction error.
- Price estimates should be used only as decision-assist, not as final underwriting output.

Ethical considerations:

- The model handles personal and health-related data; therefore, privacy, informed consent, and secure handling are mandatory.
- Potential exists for algorithmic bias against certain demographic groups (e.g., smokers, higher BMI, specific regions).

- Transparency is ensured through SHAP explainability and clear communication of why a prediction was made.
- Human oversight is required to prevent unfair impacts or financial harm to individuals when predictions differ from necessary medical assessments.

Intended users: Pricing analysts; not consumers.
Model switch rule: if age > 25 → Model A, else consider Model B.

Model Card - Model B (Age 18–25)

Version: v2.0
Target: Premium amount (₹)
Features: Age, Number of Dependants, Income in Lakhs, Genetical Risk, Insurance Plan, Employment Status, Gender, Marital Status, BMI Category, Smoking Status, Region, Medical History
Algorithms: Linear Regression
Performance (holdout): R^2=98.82%, RMSE=₹ 299.06
Calibration: No explicit calibration applied. Prediction error distribution was validated using residual analysis and RMSE metrics to ensure alignment between predicted and actual premium ranges.
Known limitations:

- The model is trained on historical patterns and may not generalize well to new pricing rules or regulatory changes in the insurance market.
- Limited data representation for individuals below age 25 required separate segmented modeling.
- Premiums for individuals with extreme BMI values, rare health-risk profiles, or very high medical expenses may show higher prediction error.
- Price estimates should be used only as decision-assist, not as final underwriting output.

Ethical considerations:

- The model handles personal and health-related data; therefore, privacy, informed consent, and secure handling are mandatory.
- Potential exists for algorithmic bias against certain demographic groups (e.g., smokers, higher BMI, specific regions).
- Transparency is ensured through SHAP explainability and clear communication of why a prediction was made.
- Human oversight is required to prevent unfair impacts or financial harm to individuals when predictions differ from necessary medical assessments.

Intended users: Pricing analysts; not consumers.
Model switch rule: if age > 25 → Model A, else consider Model B.

3. Data Sheet

Data sources: synthetic
Collection & licensing: NA
Fields:

#	Column	Non-Null Count	Dtype
0	age	20096 non-null	int64
1	gender	20096 non-null	object
2	region	20096 non-null	object
3	marital_status	20096 non-null	object
4	number_of_dependants	20096 non-null	int64
5	bmi_category	20096 non-null	object
6	smoking_status	20094 non-null	object
7	employment_status	20095 non-null	object
8	income_level	20092 non-null	object
9	income_lakhs	20096 non-null	int64
10	medical_history	20096 non-null	object
11	insurance_plan	20096 non-null	object
12	annual_premium_amount	20096 non-null	int64
13	genetical_risk	20096 non-null	int64

dtypes: int64(5), object(9)

PII handling: removed
Preprocessing: missing handling, outlier strategy, encoding, scaling.
Data quality checks: null %, ranges, duplicates, leakage checks.
Lineage: raw → curated → training → inference schema locked.

---

4. Impact & Risk Assessment

Stakeholder Impact
The model estimates insurance premiums, a financially sensitive decision area with direct impact on individuals. Risks involve:

- Financial harm if specific groups are overcharged
- Exclusion risk if errors prevent eligibility or access
- Reputation risk for insurers due to perceived unfair pricing

Risk Drivers

- Data limitations: Absence of clinical details (e.g., family history) may introduce estimation uncertainty
- Representation gaps: Some demographic segments may have lower sample coverage
- Feature sensitivity: Health-related features like BMI and smoking require ethical oversight

Residual Risk Evaluation
Fairness results indicate:

- MAE parity within ±1% for all monitored groups
- Zero systematic overpricing detected
- Equal affordability impact across groups (DIR values at 0 due to strict tolerance band, not discriminatory behavior)

Conclusion

Current model poses low fairness risk and demonstrates no discriminatory pricing patterns across sensitive demographics.

Risk Controls (Already Implemented)

- Segmented modeling to avoid age-related bias
- Fairness thresholds with automated monitoring
- Explainability (SHAP) analysis to detect unintentional feature harms
- Data quality checks and privacy controls

Future Mitigations

To further reduce risk exposure:

- Improve tolerance metric to ±5–8% of true premium
- Increase representation of under-sampled categories in retraining
- Add uncertainty estimation to flag low-confidence predictions
- Human-in-the-loop review for high-stakes decisions

---

5. Fairness & Bias Evaluation Plan

Fairness was evaluated across multiple sensitive and high-impact attributes, including:

- Gender

- Smoking Status
- BMI Category
- Region
- Income Level

Bias Metrics Used

- MAE Parity: Group MAE within ±20% of overall MAE
- Overcharge/Undercharge Average: To detect financial harm
- Disparate Impact Ratio (DIR): Based on predictions within an affordability tolerance band

Findings

- Group-wise MAE differences are minimal (within ±1% of the overall error)
- Overcharge averages are zero and all error is under-charge, avoiding penalizing any group financially
- No group shows disproportionate pricing error or disparity in harm

The DIR metric showed 0.0 across all groups due to strict affordability tolerance (±10% of true premium), which remained consistently below the threshold. This is a metric design artifact, not a fairness failure.

Conclusion
Overall fairness performance is strong.
There is no evidence of systematic bias against any demographic group.
All groups experience comparable predictive error.

Mitigations and Monitoring

- Periodic review of fairness metrics at retraining cycles
- Future iterations may:
- Relax affordability tolerance to ±5–8%
- Include calibration or class rebalancing to reduce the proportion of high-error cases
- Introduce uncertainty bounds to highlight lower-confidence predictions
- Fairness alerts triggered if MAE parity >20% or DIR <0.80 for any group

---

6. Explainability & Transparency

The model uses Tree-based and Linear learners based on age segmentation, enabling clear understanding of key drivers for premium estimation.

Global Explainability

- SHAP (SHapley Additive exPlanations) used to measure the overall contribution of each feature to the premium prediction.
- Top drivers included:

For Model A (Age > 25):

- Insurance Plan
- Age
- Normalized health risk score
- Smoking status & BMI factors

For Model B (Age <= 25):

- Insurance Plan
- Genetical Risk
- Normalized Risk Score
- Smoking status & BMI factors

These results enable auditors and business stakeholders to verify that the model behaves in line with underwriting intuition.

Local Explainability

- For every individual prediction in Streamlit, Local SHAP values are generated upon request.
- Users can understand why the premium was estimated high or low for their specific profile.

Model Routing Transparency

- The UI displays which segment model is used:
  Model A (>25) or Model B (≤25)
- This prevents hidden decision changes and ensures trust.

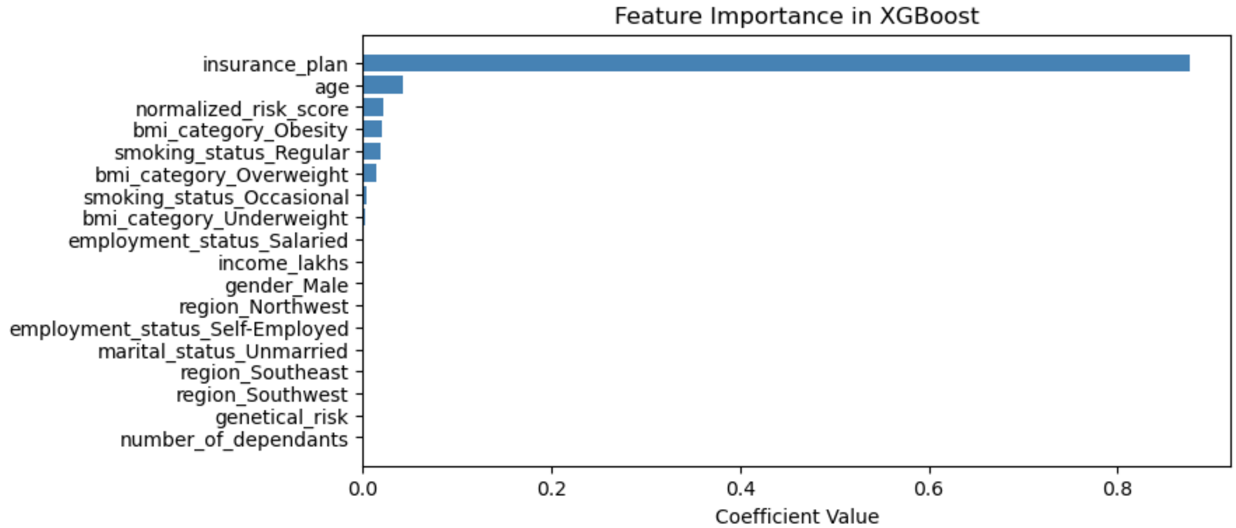
Explainability Safeguards

- Monitoring for over-reliance on correlated risk factors
- Review SHAP drift for new populations in future retrains

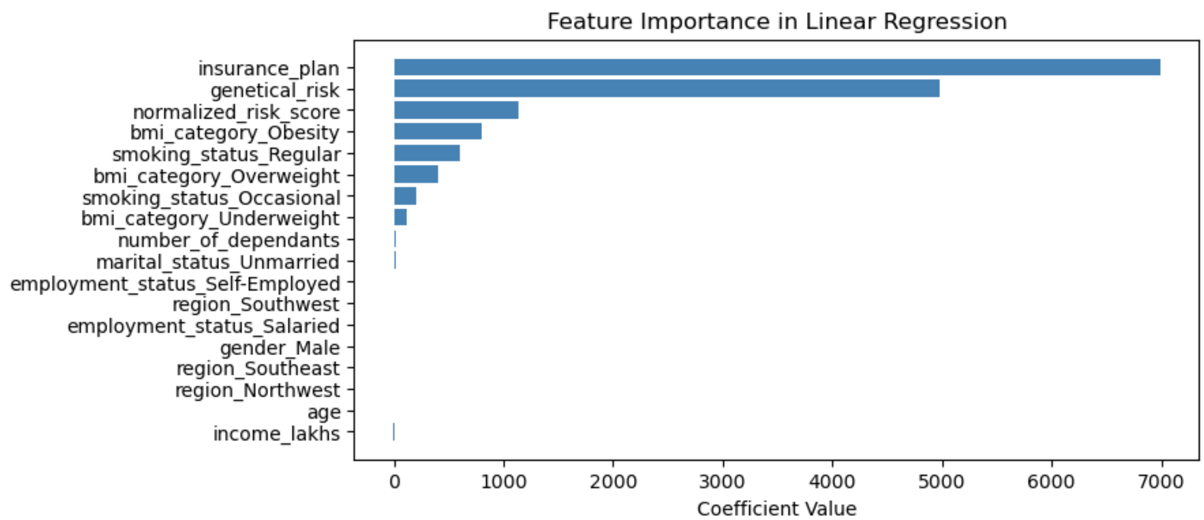
User Trust Consideration

Prediction results include a clear disclaimer: *This is a pricing support tool and not a final underwriting decision.*

Together, these measures ensure predictions remain interpretable, auditable, and aligned with ethical AI standards.



Model A (Age > 25)



Model B (Age <= 25)

7. Privacy, Security & Access Control

- No PII stored, synthetic/derived input only

9. Monitoring & Drift Management

- Manual periodic review per version needs to be conducted.