Predicting Insurance Costs: Analysis and Insights

SDS Mini-Datathon Team

National University of Singapore

October 20, 2025

Problem Framing & Objective

- **Challenge**: Predict medical insurance charges using demographic and lifestyle data
- **Goal**: Build accurate regression models and identify key cost drivers
- **Importance**: Fair pricing, risk assessment, healthcare insights
- **Dataset**: 1,338 records with age, sex, BMI, children, smoker, region, charges

Exploratory Data Analysis

- Data quality: No missing values, clean dataset
- Key statistics: Charges range \$1,122–\$63,771 (mean \$13,270)
- Distribution: Right-skewed charges, normal BMI/age

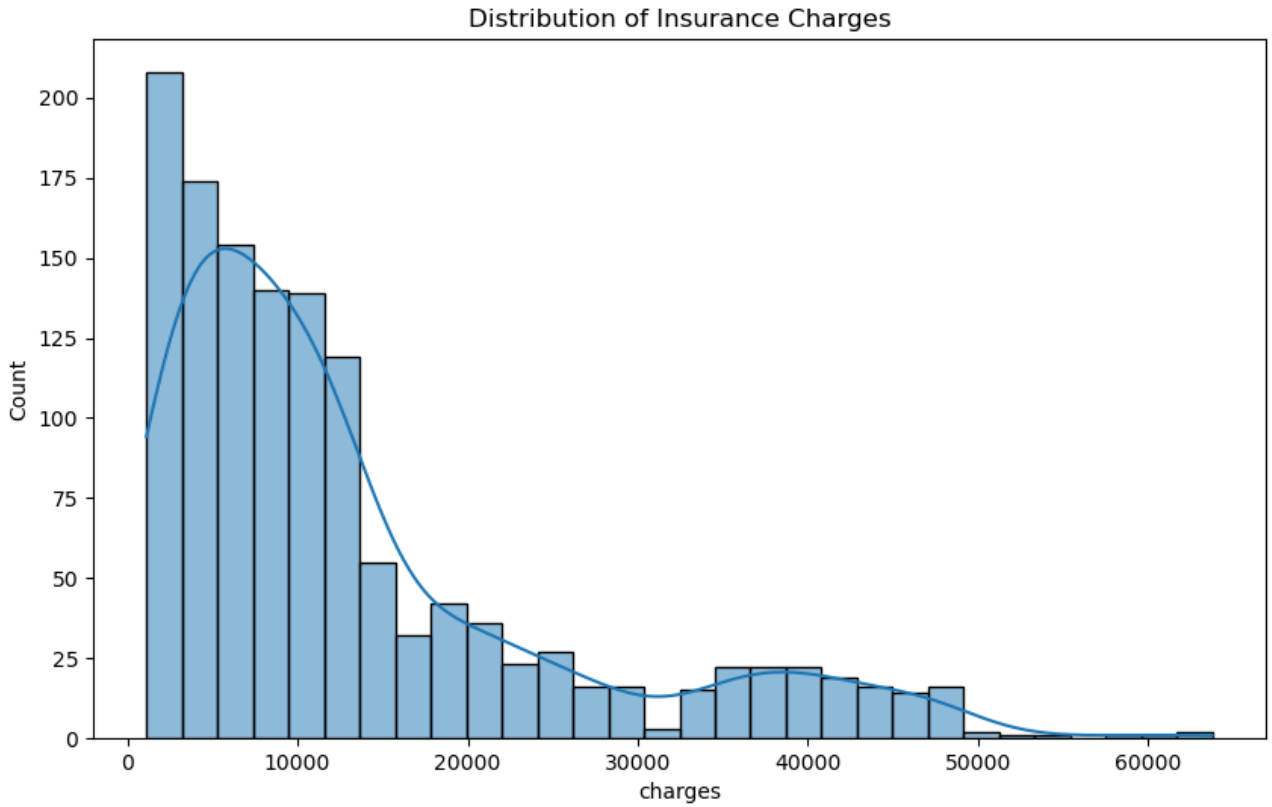

Figure: Distribution of Insurance Charges

Exploratory Data Analysis (Cont.)

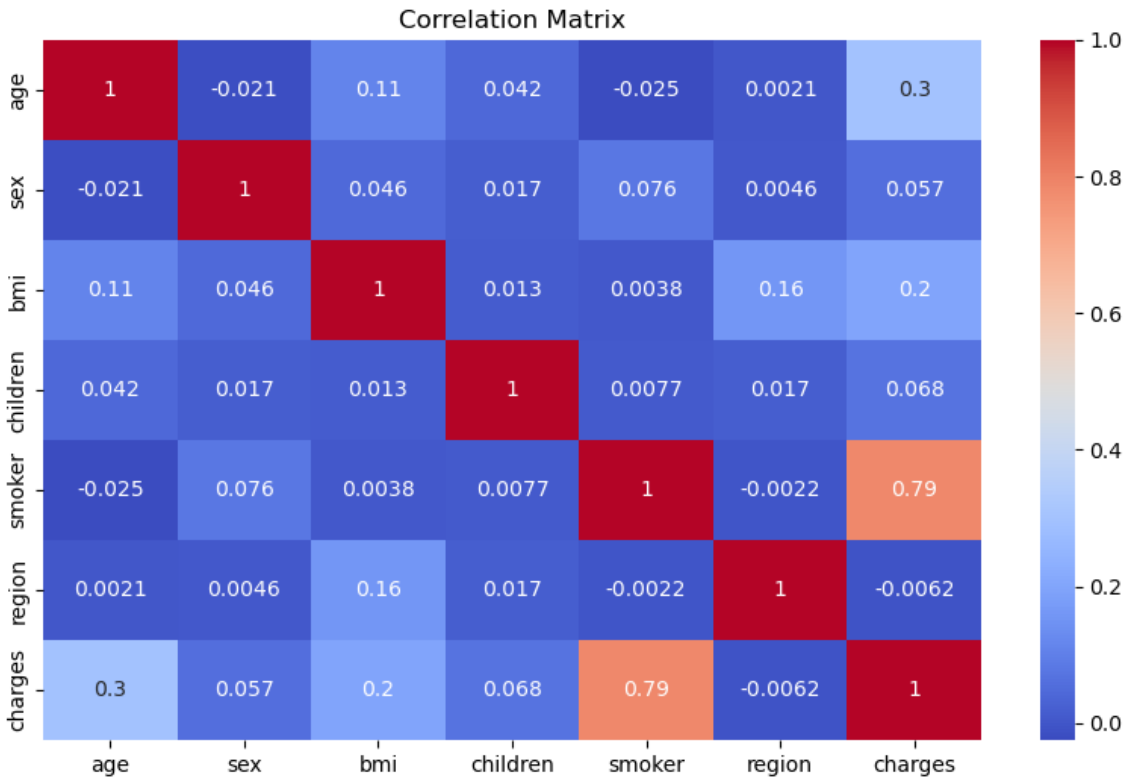

Figure: Correlation Matrix

- Strong correlations: smoker (+0.79), age (+0.30), BMI (+0.20)
- Regional variations in charges

Regression & Modeling Approach

- **Baseline Models**:
  - Linear Regression: Interpretable, assumes linearity
  - Decision Tree: Handles non-linearities, prone to overfitting
- **Advanced Model**:
  - Random Forest: Ensemble of trees, robust and accurate
  - Hyperparameter tuning with GridSearchCV
- **Evaluation**: R^2, RMSE, MAE; Cross-validation for stability

Key Findings & Visualizations

Model	R^2	RMSE	MAE
Linear Regression	0.78	5,794	4,128
Decision Tree	0.72	6,636	3,028
Random Forest	0.86	4,603	2,550

Table: Model Performance Comparison

- Random Forest best performer (R^2=0.86)
- Cross-validation confirms stability (CV R^2: 0.825 ± 0.085)

Feature Impact Analysis

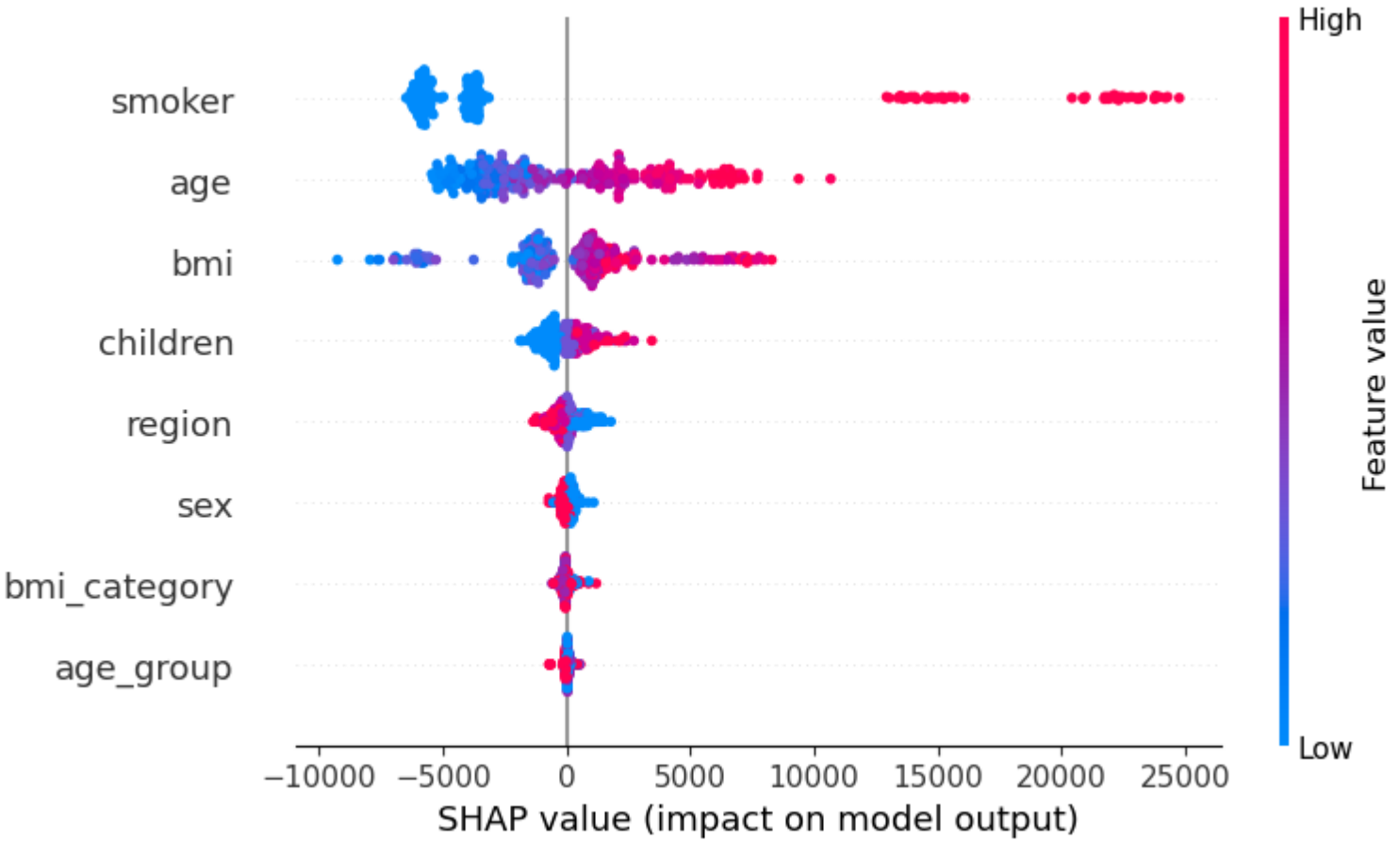

Figure: SHAP Summary Plot

- Top factors: Smoking (+20k+), Age (+200/year), BMI (+500/unit)
- SHAP provides individual prediction explanations

Fairness Analysis

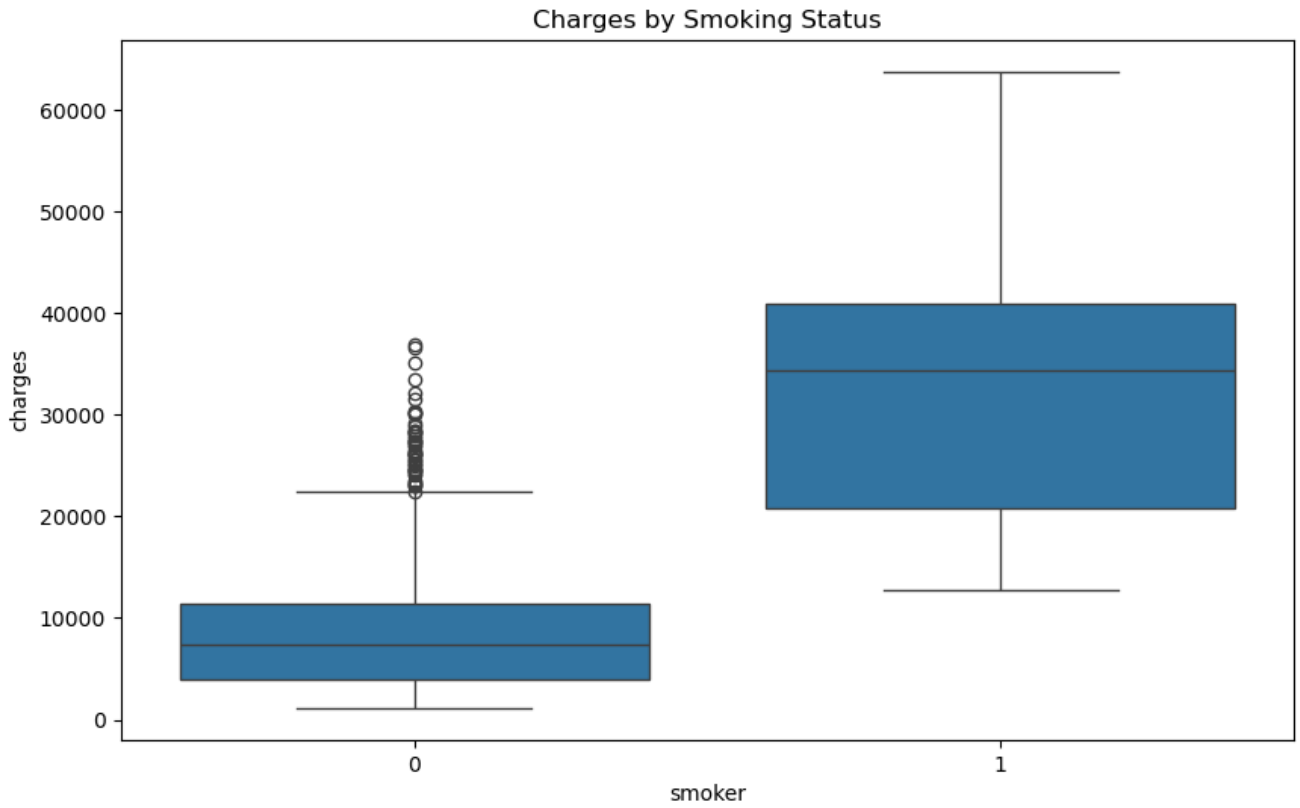

Figure: Charges by Smoking Status

- Gender gap: Males $1,387 higher (minimal)
- Regional differences: Southeast 19% higher

Practical Recommendations

- **For Insurers**:
  - Use Random Forest for premium calculation
  - Target wellness programs for smokers/high BMI
- **Policy Implications**:
  - Fair pricing based on verifiable risks
  - Monitor regional healthcare access disparities
- **Future Work**:
  - Collect longitudinal data, medical history
  - Implement advanced models (neural networks)

Difficulties Faced & Solutions

- **Dataset Limitations**:
  - Small size (1,338 records) limits generalizability
  - Missing features: medical history, genetics, lifestyle details
  - Static data: no temporal health changes
- **Technical Challenges**:
  - SHAP library version conflicts: Resolved by using compatible API
  - Model overfitting: Addressed with cross-validation and tuning
  - Interpretability: Enhanced with SHAP beyond standard methods
- **Solutions**:
  - Feature engineering for better segmentation
  - Rigorous validation and error analysis
  - Innovative explainability techniques

Conclusion

- Successfully predicted 86% of insurance cost variance
- Smoking is the dominant factor, followed by age and BMI
- Models are fair and practical for real-world use
- Future: Larger datasets and advanced AI for better accuracy

Thank you for your attention!