Title: Medical Insurance Cost Prediction Capstone

Subtitle: Data-Driven Analysis of Health Risks & Pricing

Presenter: ADENIYI ADEOLUWA MOFOLUWADARA

Date: February 2026

Agenda

- Dataset overview
- Problem statement
- Key EDA insights
- Modeling approach
- Model performance comparison
- Final model selection
- Business or real-world interpretation

Problem Statement

The Problem: Insurance companies struggle to set fair premiums. Undercharging leads to losses; overcharging leads to customer churn.

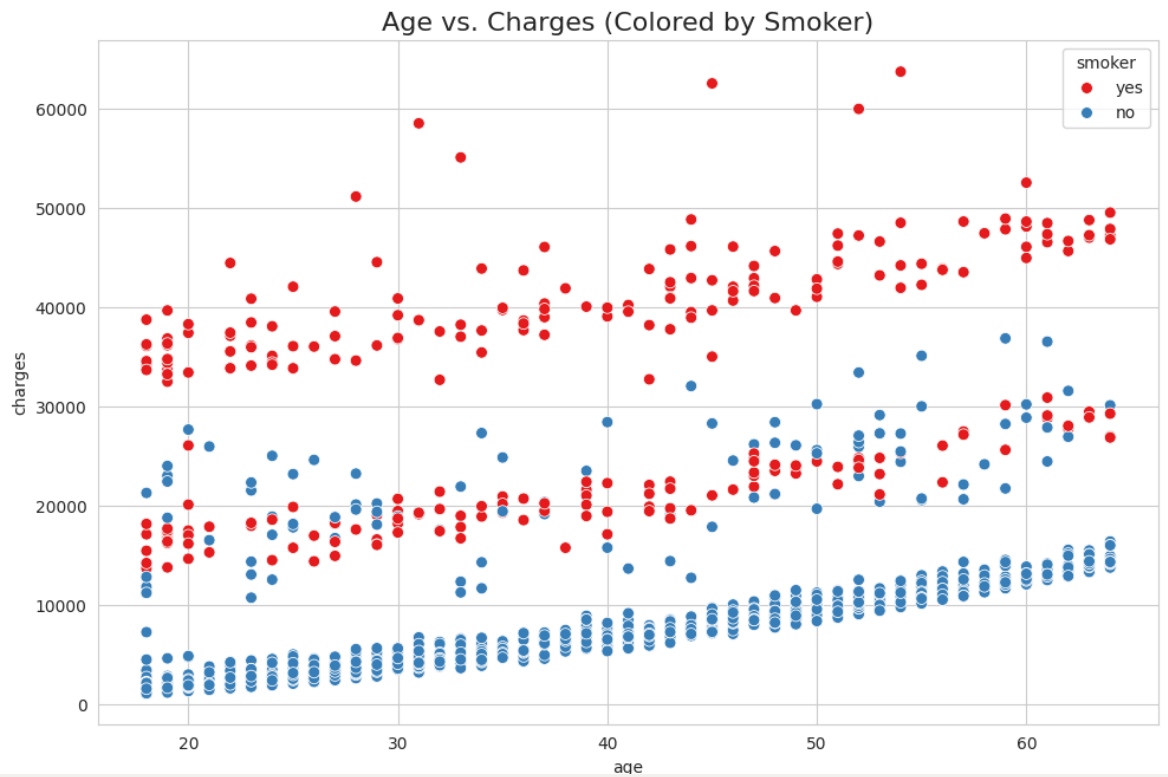
My Goal: Build a machine learning solution to accurately predict individual medical costs based on patient data.

Approach: I treated this as both a Regression problem (predicting exact price) and a Classification problem (identifying high-risk patients).

Dataset Overview

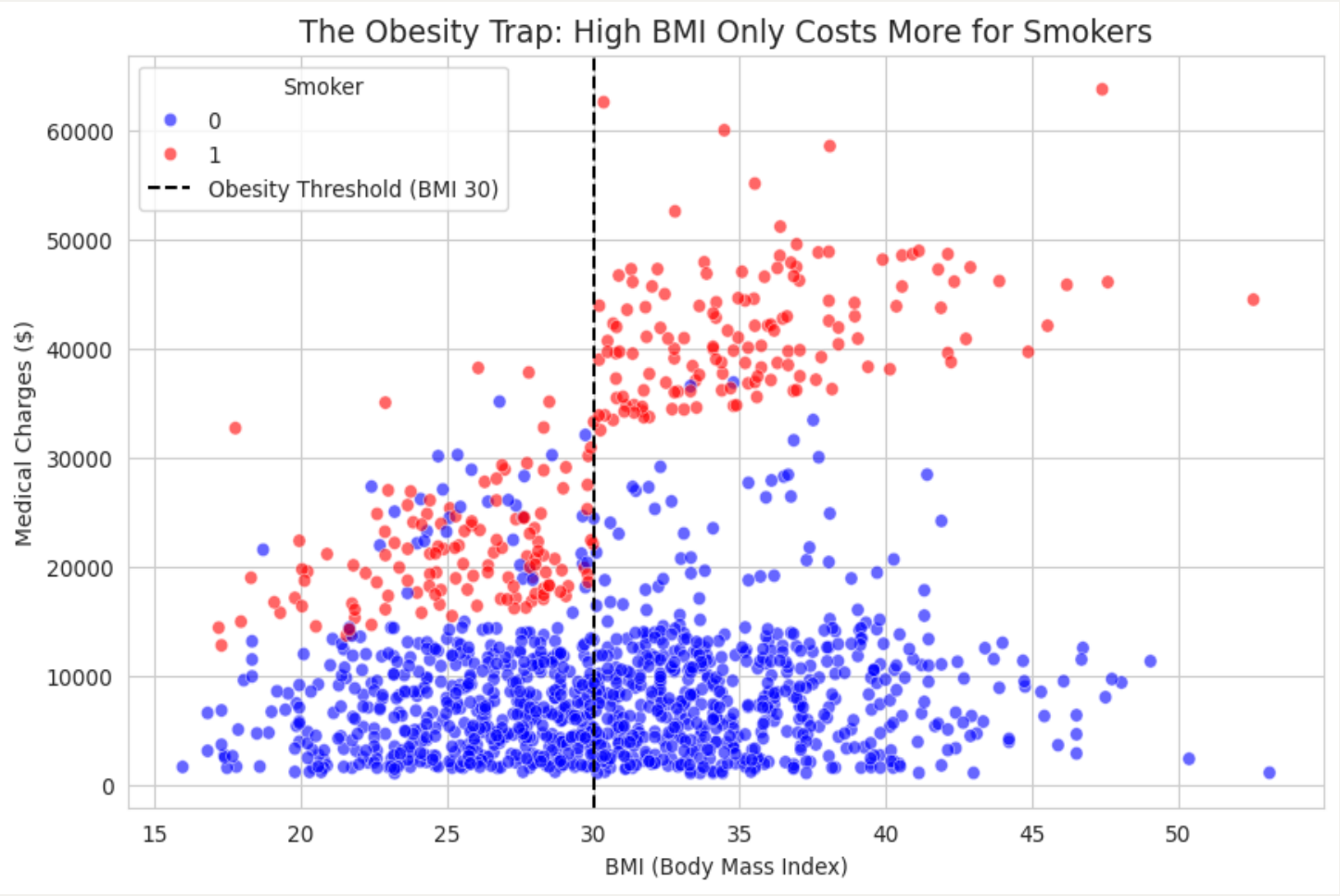
- Source: Medical Cost Personal Dataset (1,338 records).
- Key Features: Age, Sex, BMI, Children, Smoker, Region.
- Target: charges (The dollar amount billed).
- Data Quality: No missing values. 50/50 gender balance.

Key Insight #1 (The "3 Stripes")



- Finding: Smoking is the #1 driver of cost.
- Evidence: The chart shows three distinct "stripes" of cost. The red dots (smokers) are consistently in a much higher bracket than blue dots (non-smokers).

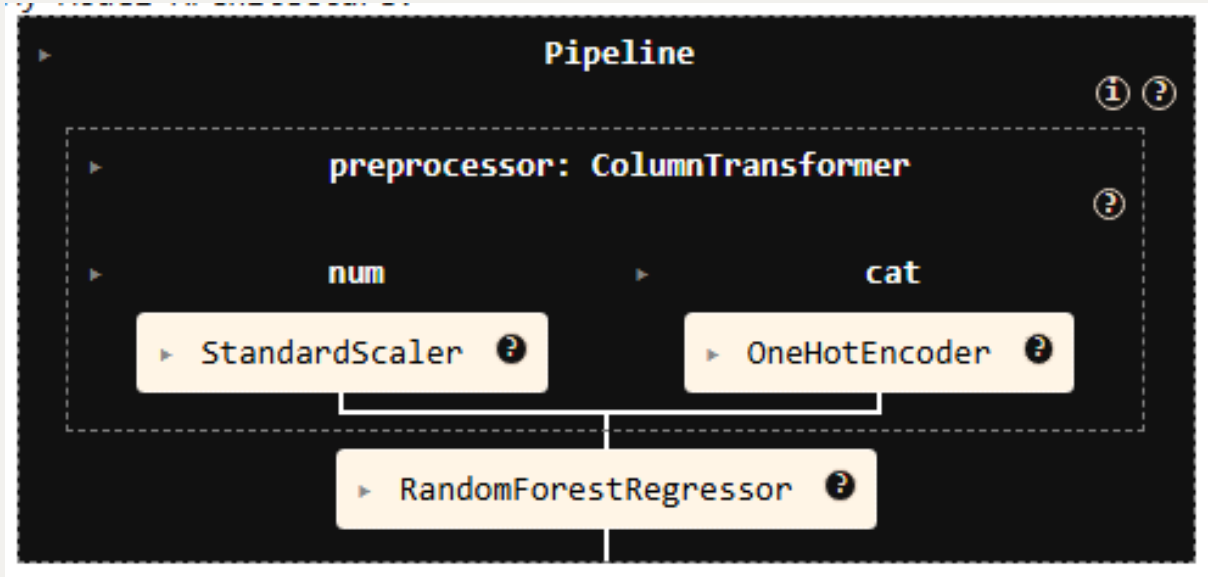
Key Insight #2 (The "Obesity Trap")



Finding: High BMI (Obesity) is financially dangerous only if the patient is also a smoker.

Evidence: For non-smokers, being overweight has a minimal impact on immediate medical costs.

Modeling Strategy (The Pipeline)



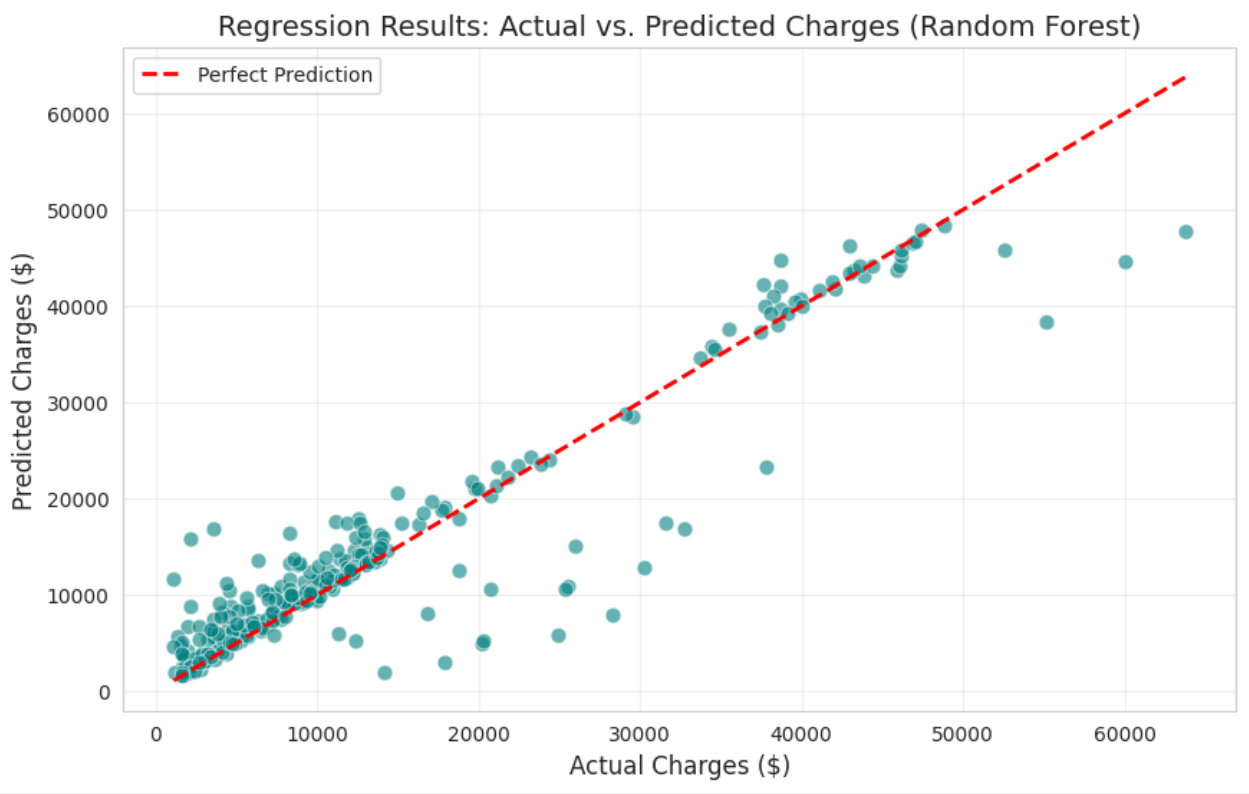
Method: I used a Scikit-Learn Pipeline to prevent data leakage and ensure robust preprocessing.

Architecture:

Preprocessing: StandardScaler (for Age/BMI) + OneHotEncoder (for Region/Sex).

Model: Random Forest Regressor (chosen for its ability to handle complex interactions).

Model Performance



Regression Results:

R^2 Score: ~87% (My model explains 87% of price variation).

MAE: Significantly lower error than the linear baseline.

Classification Results:

Accuracy: ~94% (High precision in identifying high-risk patients).

Validation: Verified using 5-Fold Cross-Validation to ensure stability.

Business Recommendations

1. **Targeted Pricing**: Create a specific premium tier for "Obese Smokers" as they drive the majority of extreme costs.

2. **ROI on Cessation**: Invest in smoking cessation programs. Converting a smoker to a non-smoker yields the highest financial return.

3. **Automated Triage**: Use my Classification Model to automatically flag high-risk applications for manual review

Conclusion

Summary: I successfully transformed raw data into a predictive engine that outperforms standard baselines.

Next Steps: Test the model on newer data and consider regional cost adjustments.