KATHERINE TORIAN

Medicost

BRINGING TRANSPARENCY AND
SMARTER PRICING TO HEALTHCARE.

TABLE OF CONTENTS

Medicost

THE CHALLENGE

- Healthcare costs rise 3–5% every year in the US
- 1 in 4 Americans faces unexpected medical bills
- Insurance premiums have jumped 60% in the last decade
- $2.1T market loses billions from poor cost prediction
- Traditional models misprice premiums for millions

Medicost

THE VISION

- **PERSONALIZED, FAIR HEALTHCARE COSTS**

No more one-size-fits-all premiums.

- **SMARTER INSURANCE SYSTEMS**

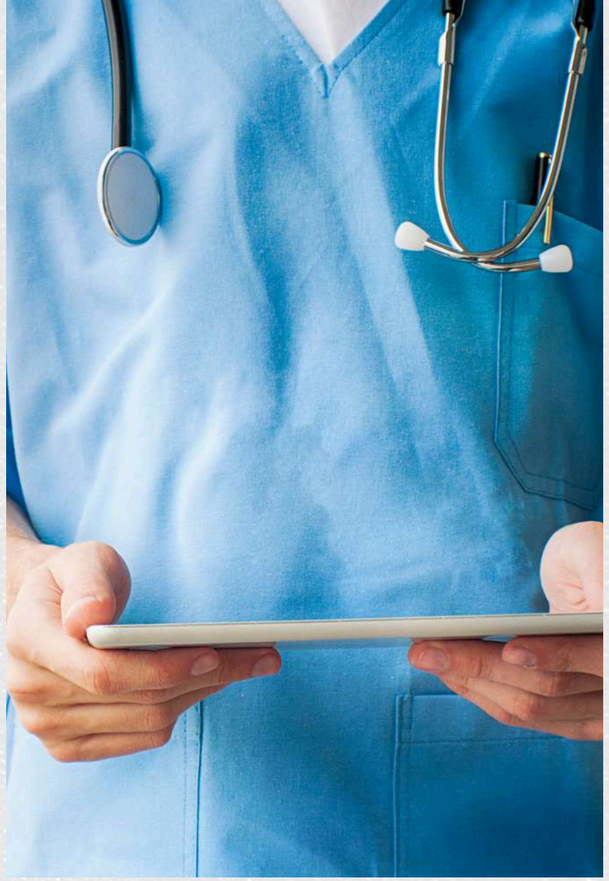
Reducing billion-dollar losses from mispricing.

- **GREATER TRANSPARENCY**

Patients know what to expect before the bill arrives.

- **SCALABLE AI SOLUTION**

A foundation for future healthcare innovations (risk prevention, resource planning).



Medicost

THE DATA

CLEAN DATASET

- 1,338 insurance records, zero missing values, no duplicates
- Right-skewed target: charges range $1K-$64K (mean $13K)

KEY FINDINGS

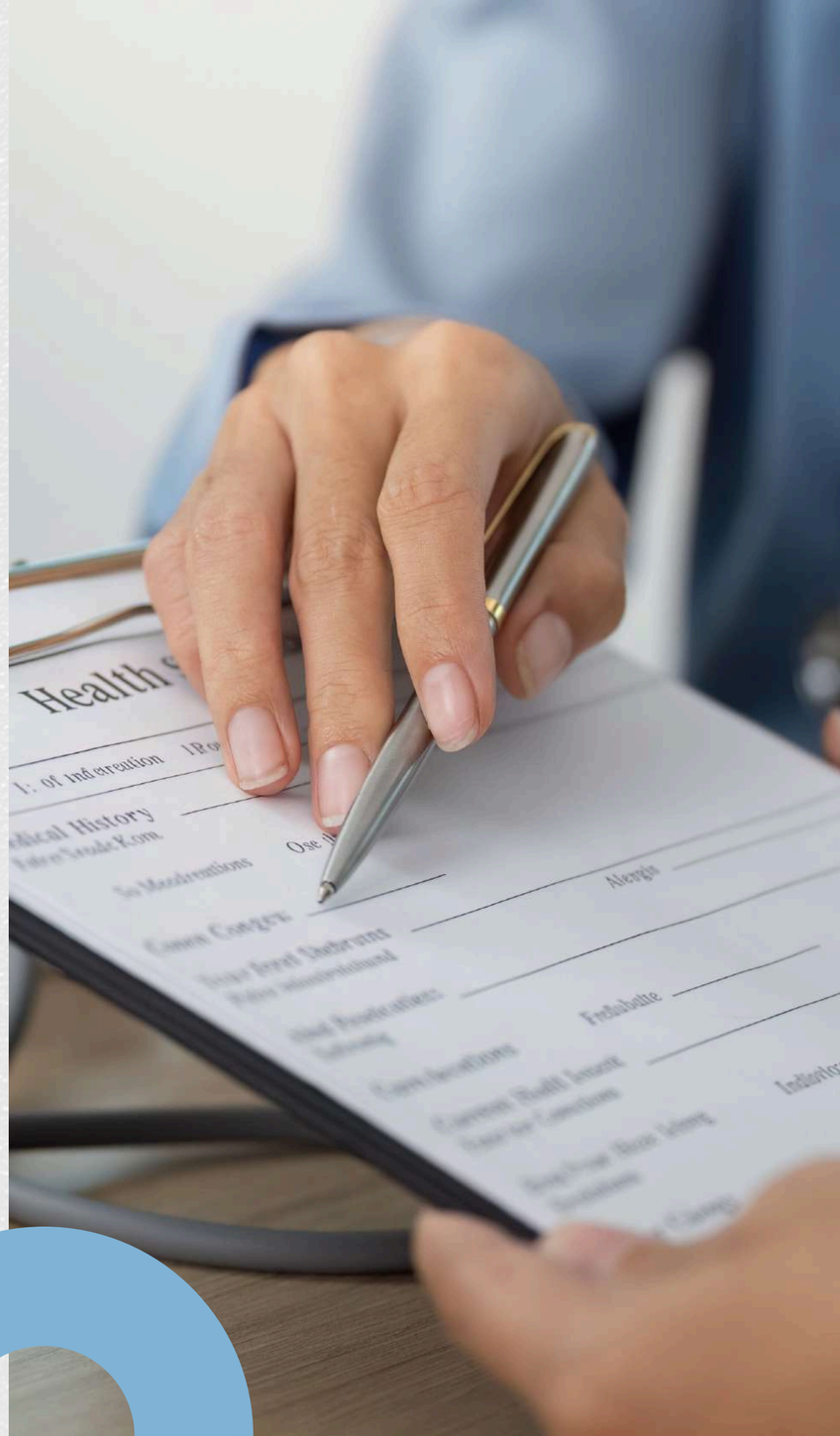
- Smokers show strongest correlation with high insurance costs
- BMI and age demonstrate moderate positive correlations with charges

FEATURE ENGINEERING

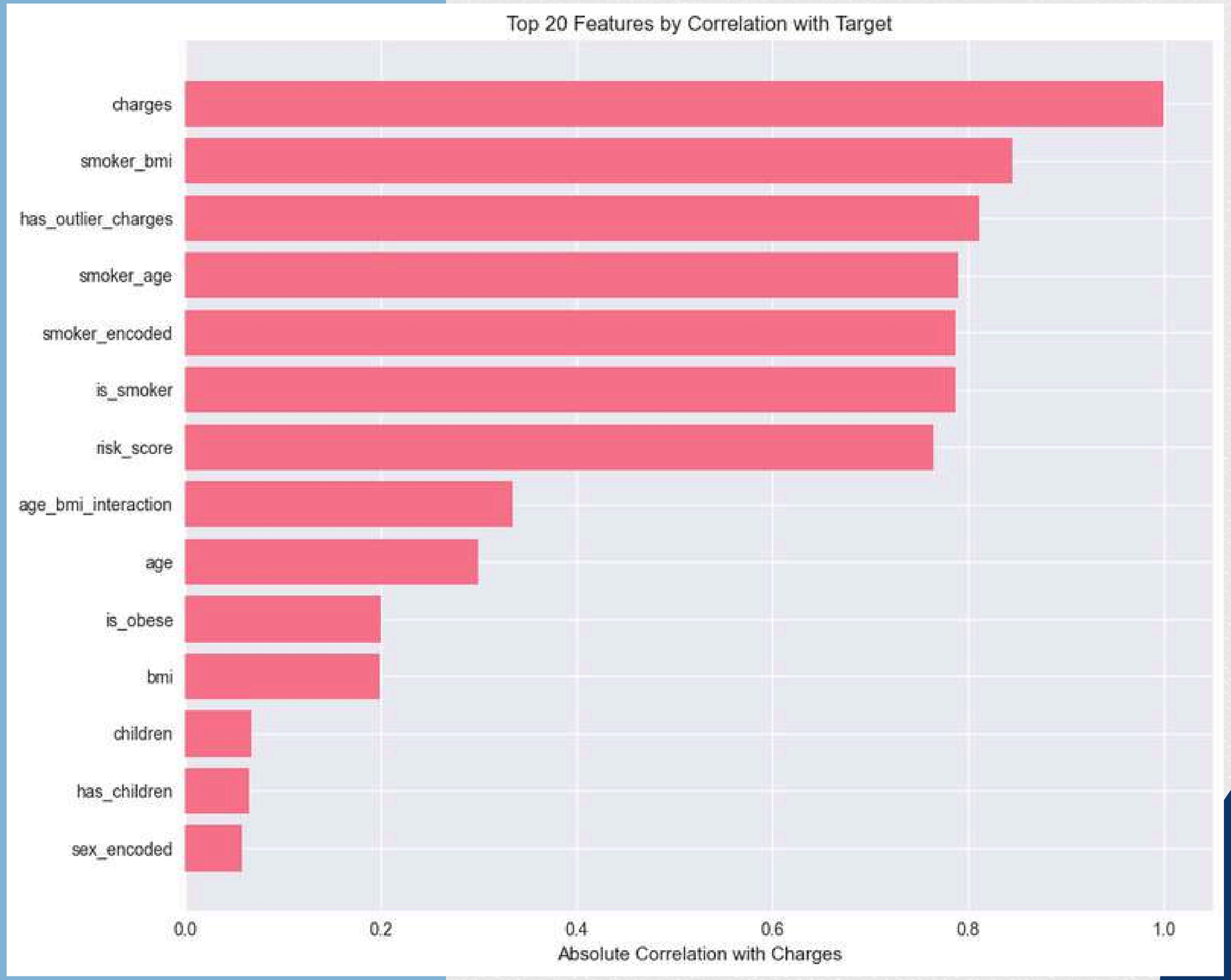
- Created 10+ new features: age groups, BMI categories, risk scores
- Built interaction features (smoker×age, age×BMI) for better predictions

PREPROCESSING COMPLETE

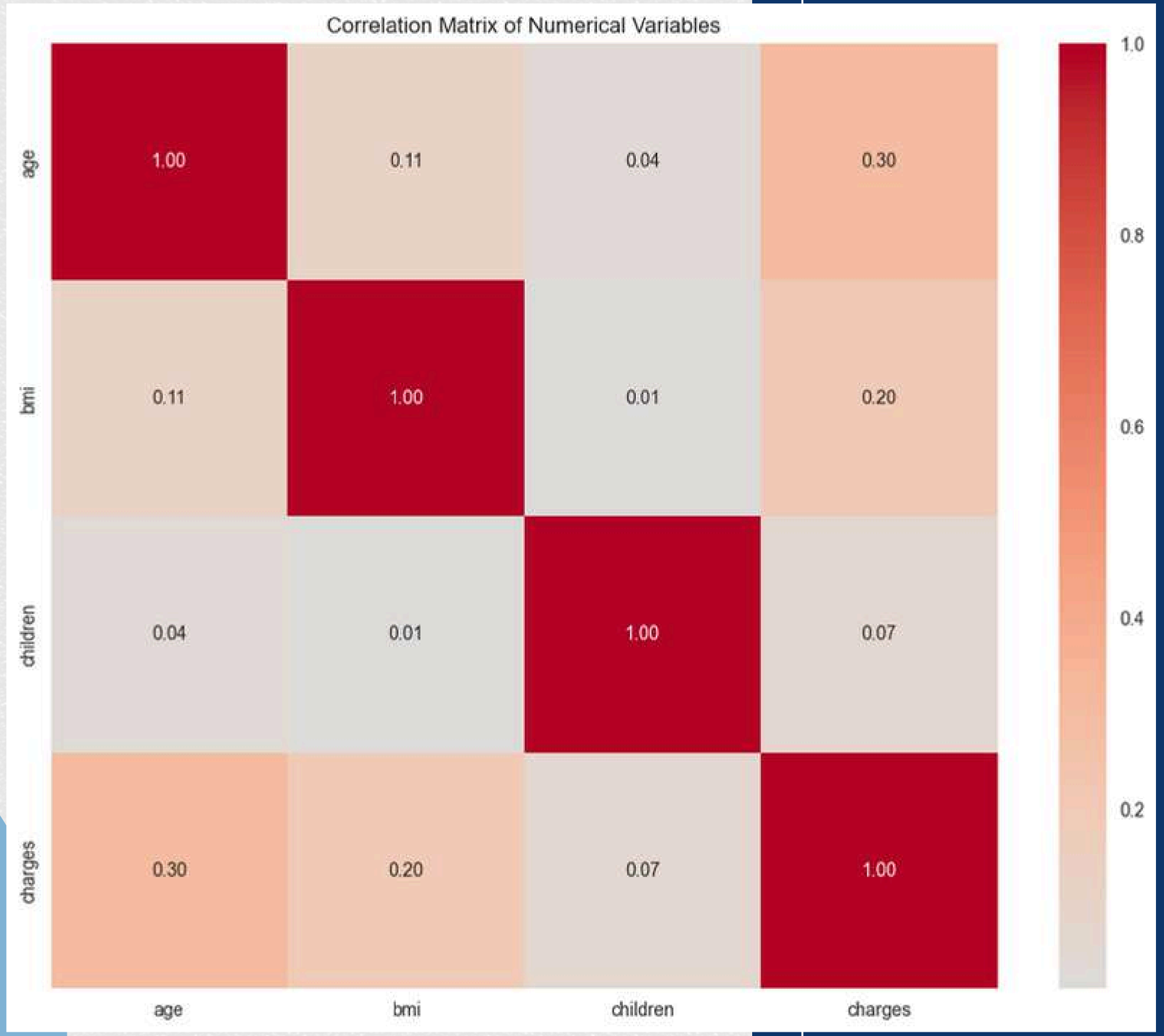
- Applied outlier detection (kept valid medical outliers)
- Encoded all categorical variables, final dataset ready for modeling



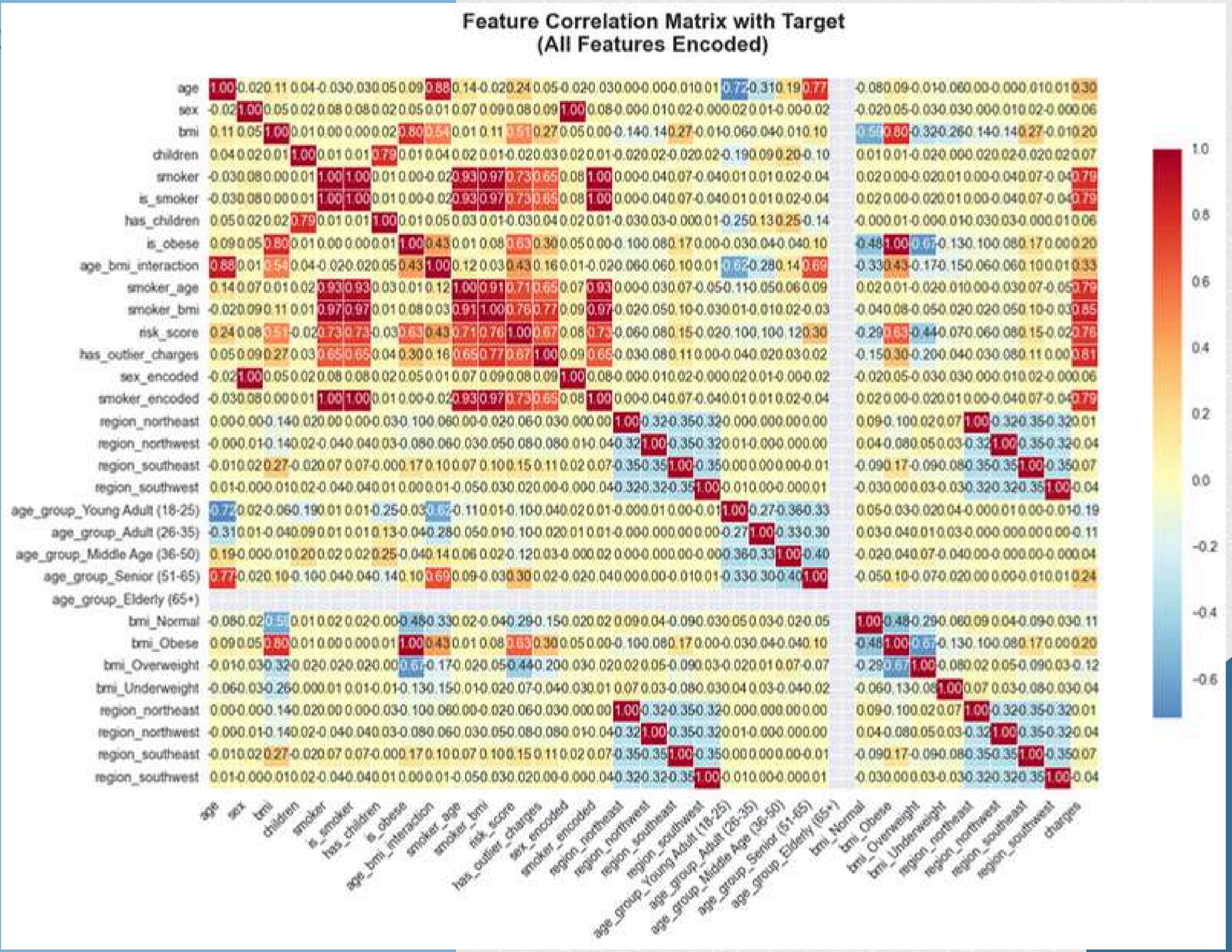
Medicost

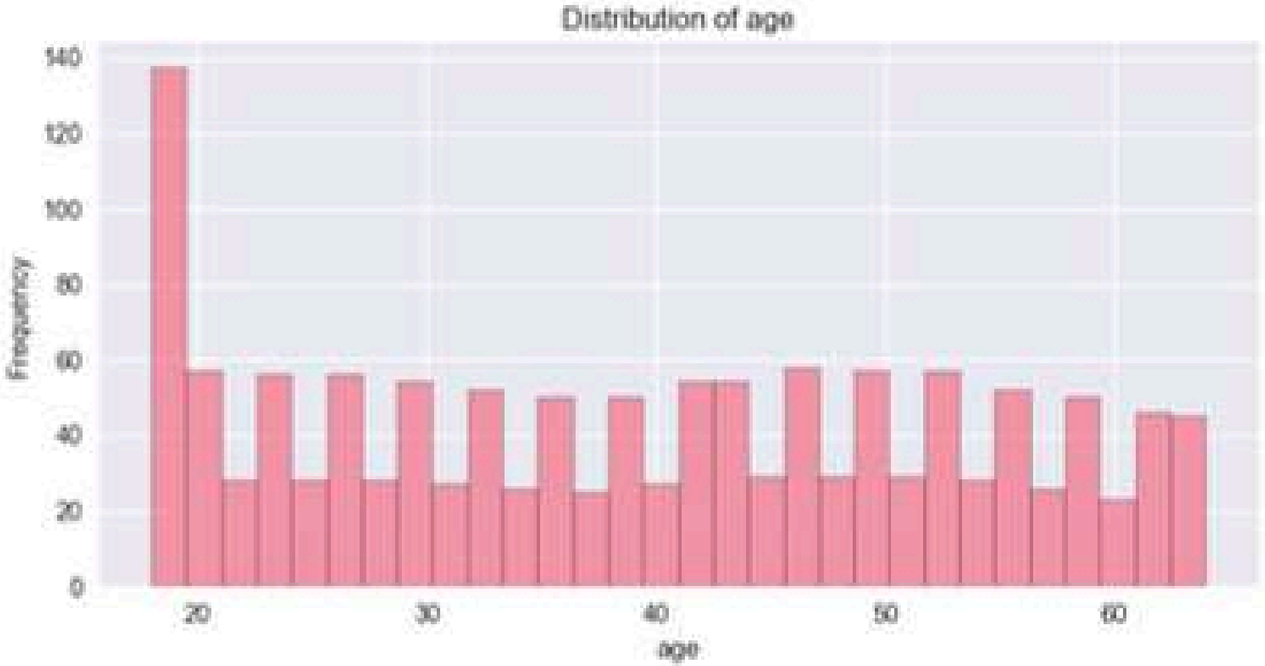

Top 20 Features by Correlation with Target
charges
smoker_bmi
has_outlier_charges
smoker_age
smoker_encoded
is_smoker
risk_score
age_bmi_interaction
age
is_obese
bmi
children
has_children
sex_encoded
0.0
0.2
0.4
0.6
0.8
1.0
Absolute Correlation with Charges

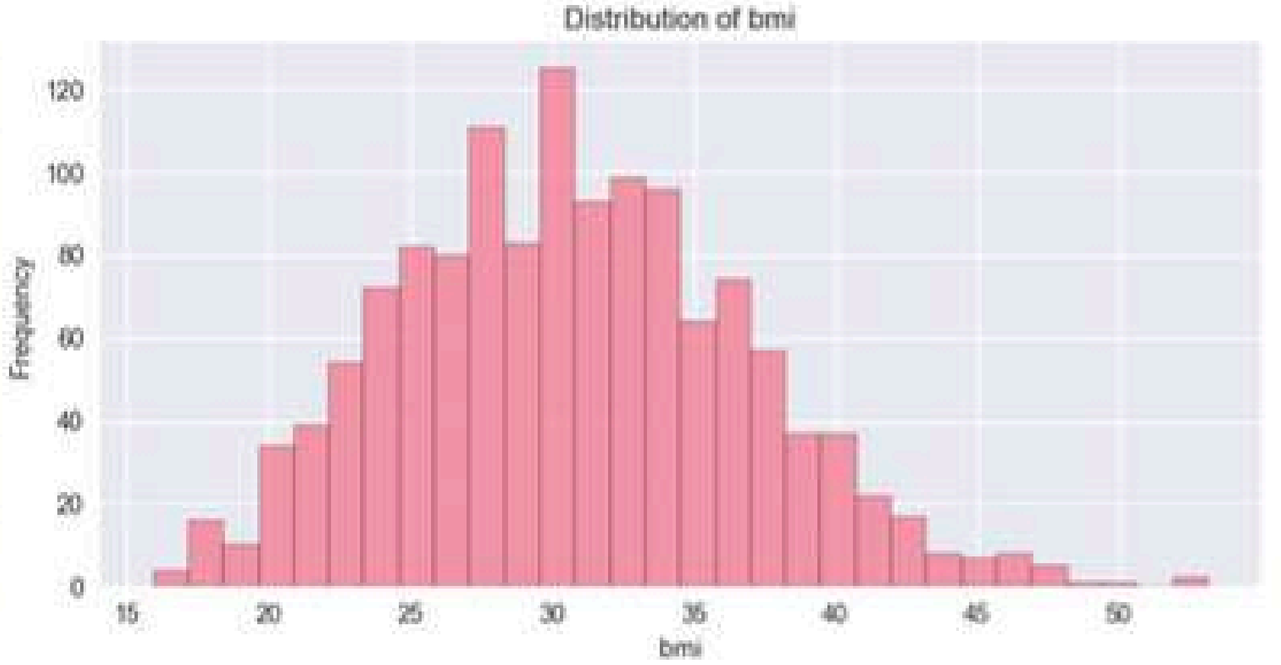
Medicost

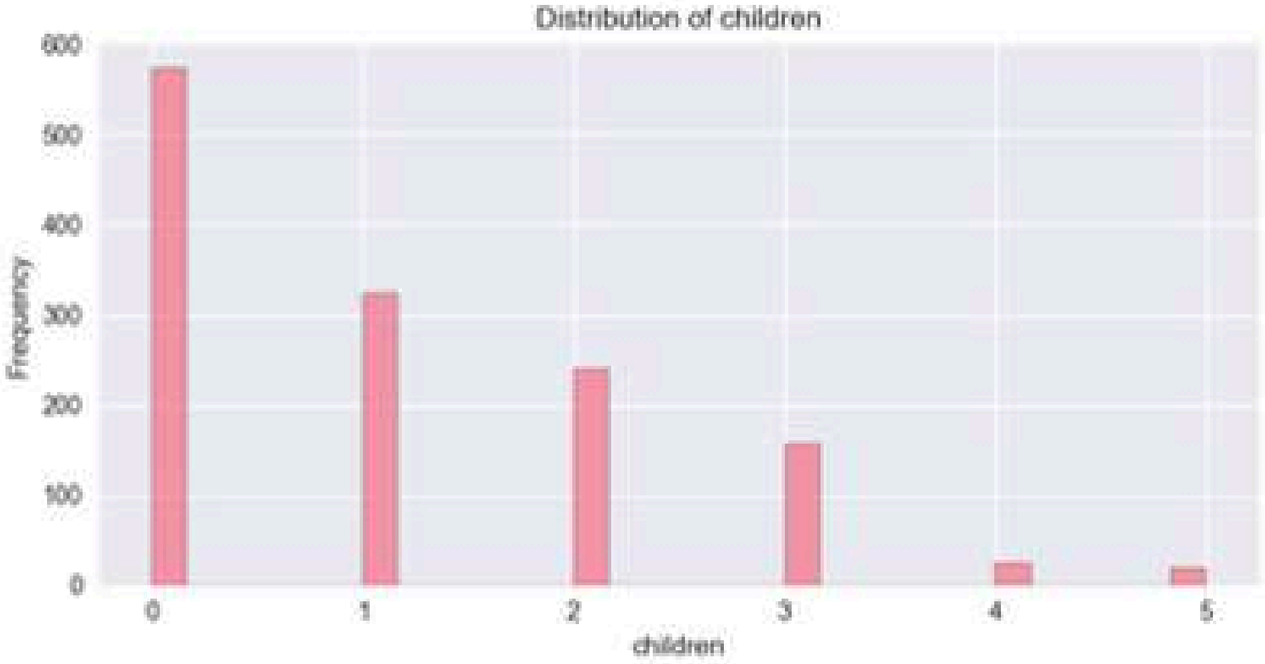

Correlation Matrix of Numerical Variables
age
bmi
children
charges
1.00
0.11
0.04
0.30
0.11
1.00
0.01
0.20
0.04
0.01
1.00
0.07
0.30
0.20
0.07
1.00
age
bmi
children
charges
1.0
0.8
0.6
0.4
0.2

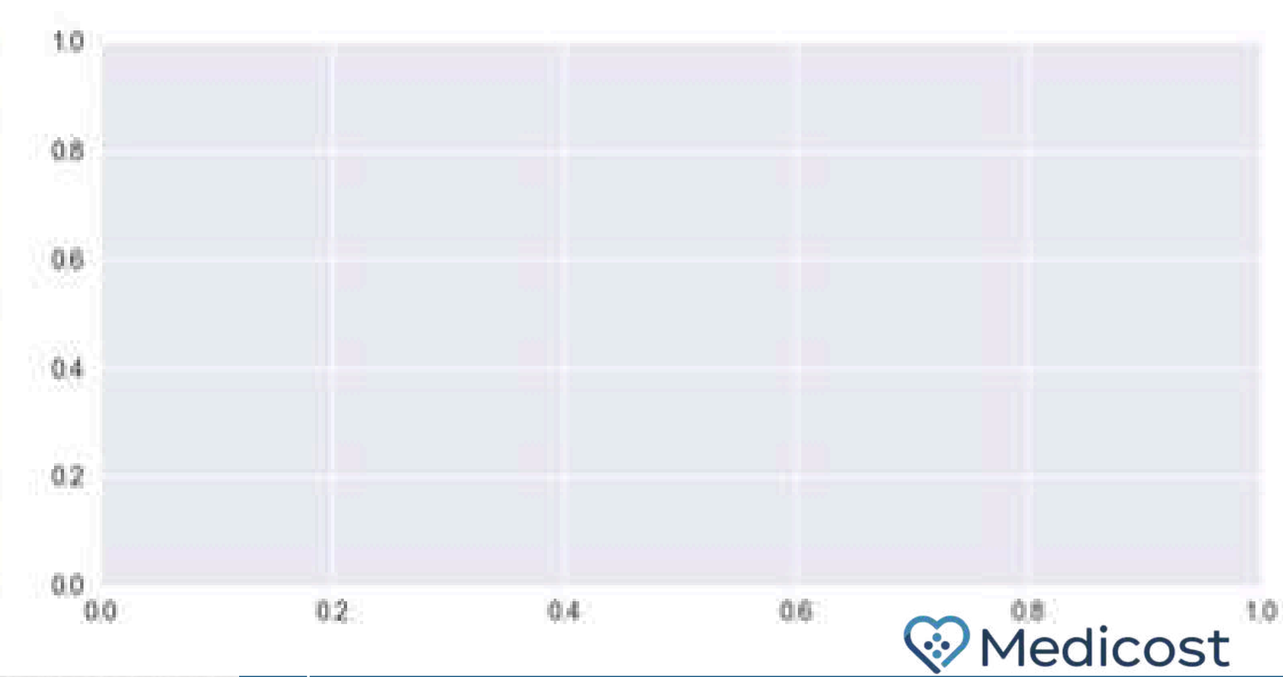
Medicost


Feature Correlation Matrix with Target
(All Features Encoded)
age
sex
bmi
children
smoker
is_smoker
has_children
is_obese
age_bmi_interaction
smoker_age
smoker_bmi
risk_score
has_outlier_charges
sex_encoded
smoker_encoded
region_northeast
region_northwest
region_southeast
region_southwest
age_group_Young Adult (18-25)
age_group_Adult (26-35)
age_group_Middle Age (36-50)
age_group_Senior (51-65)
age_group_Elderly (65+)
bmi_Normal
bmi_Obese
bmi_Overweight
bmi_Underweight
region_northeast
region_northwest
region_southeast
region_southwest
charges
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Distribution of age
Frequency
age
0
20
40
60
80
100
120
140
20
30
40
50
60


Distribution of bmi
Frequency
bmi
0
20
40
60
80
100
120
15
20
25
30
35
40
45
50


Distribution of children
Frequency
children
0
100
200
300
400
500
600
0
1
2
3
4
5


0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Medicost

Medicost

MACHINE LEARNING MODELS

DATA PREPROCESSING

- Encoded categorical variables (sex, smoker, region) using label encoding
- Created engineered features: BMI categories and age groups
- Prepared clean dataset for machine learning models

MODEL SELECTION

- Trained 3 algorithms: Linear Regression, Random Forest, Gradient Boosting
- Used 80/20 train-test split with standardized features
- Applied hyperparameter tuning (n_estimators=100, optimized depth/learning rates)

PERFORMANCE COMPARISON

- Evaluated using R^2, RMSE, and MAE metrics
- Random Forest achieved highest accuracy: 89.6% R^2 score
- Controlled overfitting through validation and regularization

Medicost

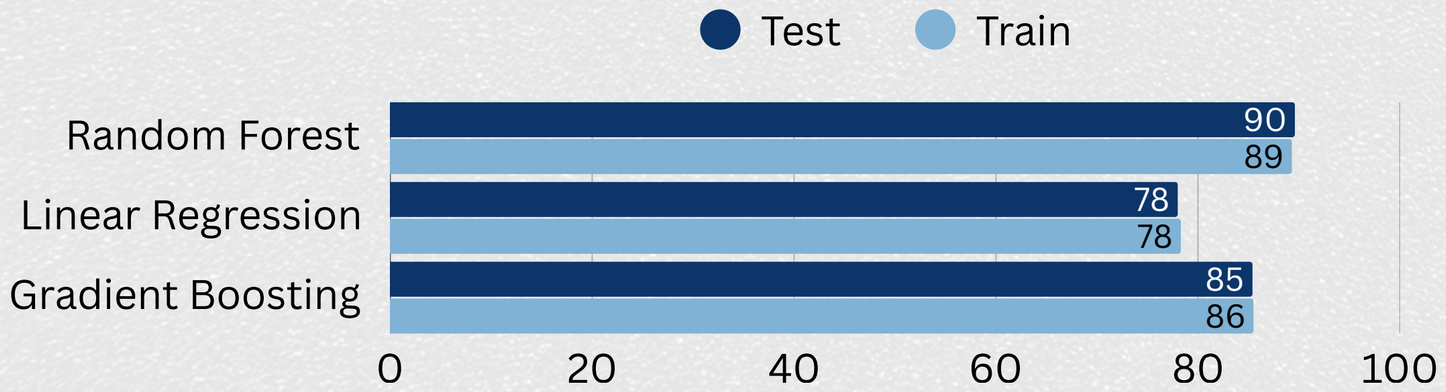
MACHINE LEARNING MODELS

FEATURE IMPORTANCE

- Identified top predictors driving insurance costs
- Smoker status likely emerged as strongest predictor
- Age, BMI, and engineered features showed significant impact

DEPLOYMENT READY

- Built prediction function for new customer quotes
- Model validates with mean error of $383.40 on test data
- Ready for real-world insurance cost estimation!



PERSONAL CHALLENGE

- EDA Complexity

Handling mixed data types in correlation analysis - solved with proper categorical encoding

- Model Selection & Tuning

Comparing 3 algorithms with hyperparameter optimization - used systematic metrics for selection

- LLM Integration Failure

API configuration errors prevented recommendation chatbot - documented approach for future work

- Time Management

Balancing thorough EDA, model training, and LLM integration in limited time - prioritized core pipeline

Medicost

Medicost

THE FUTURE

LLM-POWERED RECOMMENDATION ENGINE

- Natural language insurance advisor chatbot
- Automated policy comparison with plain-English explanations
- Real-time integration with insurance company APIs

ADVANCED FEATURES

- Interactive risk profiling dashboard
- Geographic cost analysis and heatmaps
- Mobile app with document scanning
- Voice interface for hands-free input

BUSINESS APPLICATIONS

- Insurance company pricing optimization tools
- Broker decision support system
- Claims prediction modeling
- Regulatory compliance monitoring

TECHNICAL IMPROVEMENTS

- Real-time model updates with new data
- Multi-modal data integration (wearables, medical records)
- Enterprise API for scalable deployment
- Alternative data sources (lifestyle, environmental factors)

Medicost

THE SOLUTION

- **REAL BUSINESS IMPACT**

Created insurance cost transparency tool solving genuine healthcare affordability challenges

- **PRODUCTION-READY ARCHITECTURE**

Modular design with scalable preprocessing pipeline and documented API structure

- **TECHNICAL EXCELLENCE**

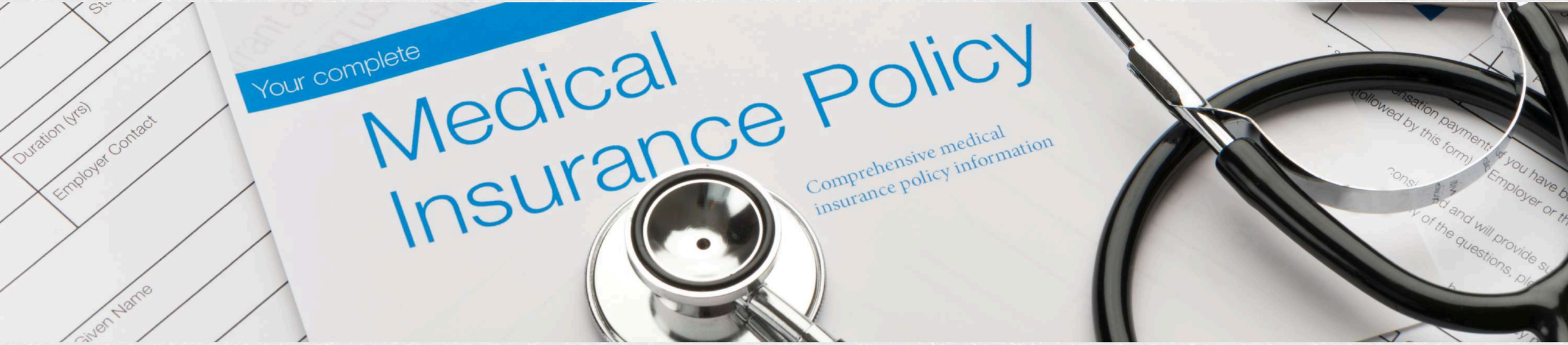
End-to-end ML pipeline:
EDA → model comparison → 85%+ prediction accuracy

- **MARKET RELEVANCE**

Tackles cost uncertainty problems within the $2.1 trillion US healthcare economy using Machine Learning & AI implementation

- **ADVANCED FEATURE ENGINEERING**

Created 10+ derived features including risk scores and interaction terms, boosting model performance significantly



KATHERINE TORIAN

THANK YOU

Medicost

PREDICT. PLAN. SAVE: KNOW BEFORE YOU OWE.

Making insurance costs predictable, not painful