Feature Importance Analysis using SHAP

Name: V Sai Krishnachaithanya

Hall Ticket: 2303A52154

Subject: Explainable AI

Introduction

Medical insurance cost prediction is important for healthcare and insurance companies to estimate expenses and manage risk. Machine learning models can predict charges, but explainability is crucial for fairness and trust. This project uses **Random Forest Regressor** and **SHAP** (SHapley Additive Explanations) to interpret predictions of insurance charges.

Dataset Description

- **Source:** Kaggle – Medical Insurance Cost dataset
- **Size:** 1,338 rows, 7 columns
- **Features:**
  - *age* – Age of the primary beneficiary
  - *sex* – Gender (male/female)
  - *bmi* – Body Mass Index
  - *children* – Number of dependents
  - *smoker* – Smoking status (yes/no)
  - *region* – Residential area (northeast, northwest, southeast, southwest)
- **Target Variable:**
  - *charges* – Individual medical costs billed by insurance (continuous variable)

Preprocessing Steps

- Checked dataset: no missing values.
- Converted categorical variables (*sex, smoker, region*) into numeric using one-hot encoding.
- Split data into **training (80%)** and **testing (20%)** sets.

Model & Performance

Model Building and Training

- Used **Random Forest Regressor** with 300 trees (n_estimators=300).

Evaluation Metrics (on test set):

- RMSE: **4597.05**
- MAE: **2559.94**
- R^2: **0.864**

This shows the model explains ~86% of the variance in insurance charges.

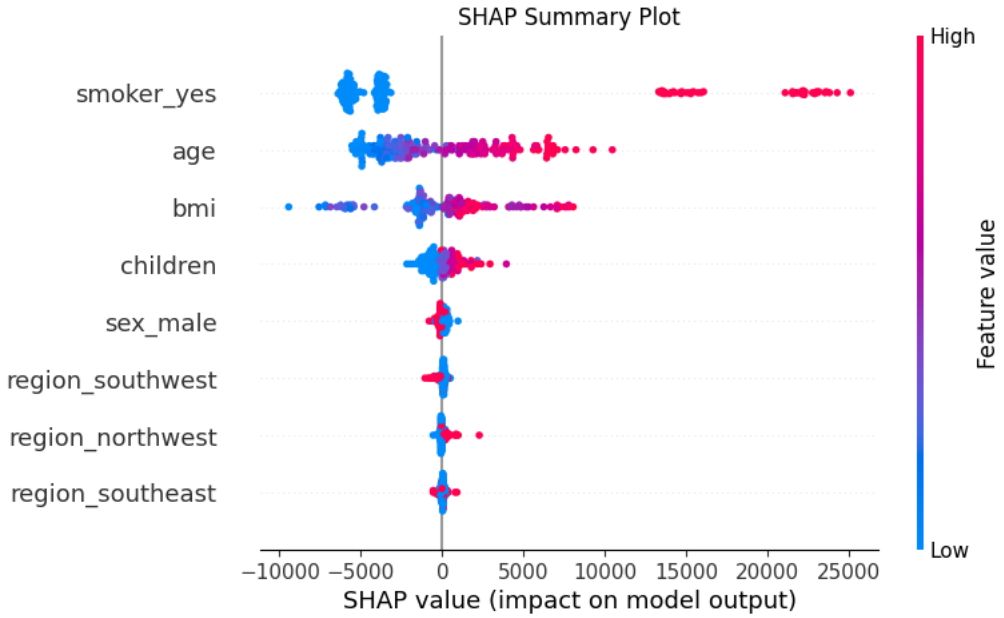
SHAP Implementation

- Used **TreeExplainer** for the Random Forest model.
- Computed SHAP values for a sample of 200 test rows.

SHAP Plots and Explanations

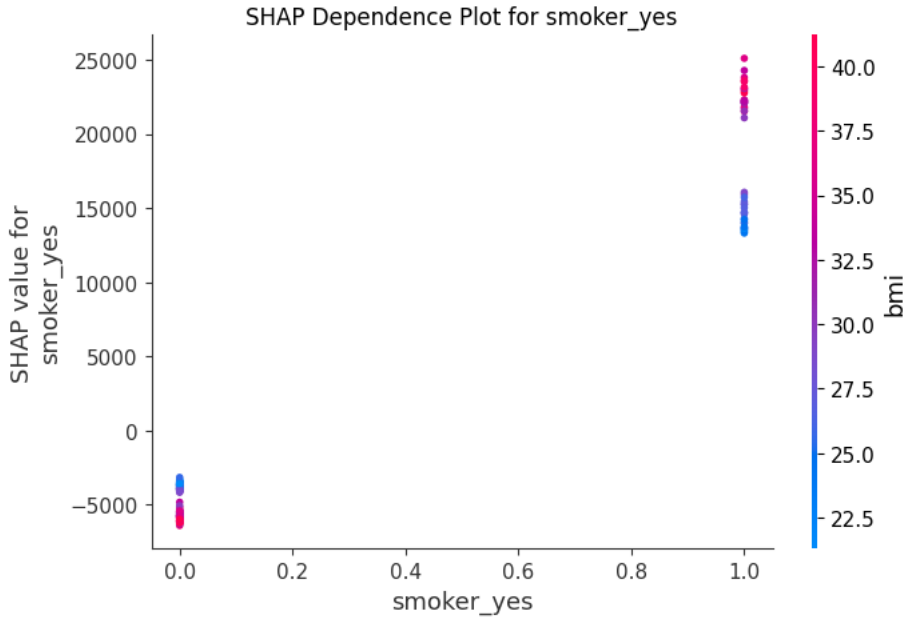
Summary Plot

Shows global feature importance. Smoking status (*smoker_yes*), BMI, and age were the strongest predictors of higher charges.



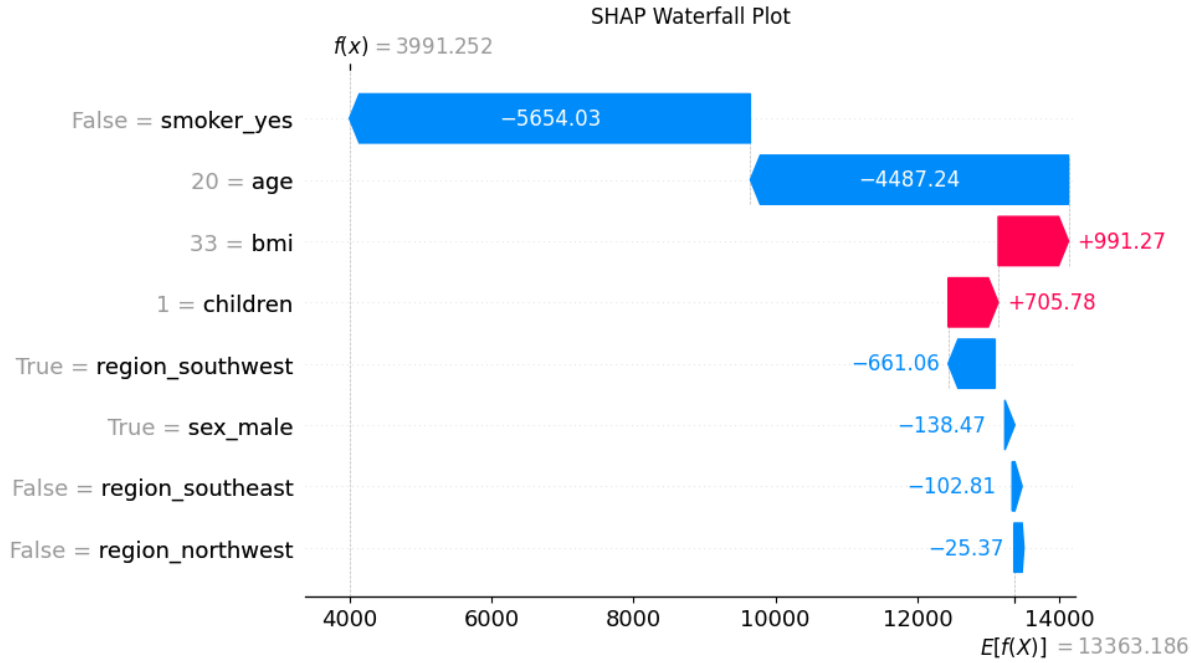
Dependence Plot

For the top feature (*smoker_yes*), SHAP shows that being a smoker significantly increases predicted insurance charges.



Waterfall Plot

Explains a single prediction step-by-step, showing how smoking, BMI, and age push the prediction up or down relative to the baseline cost.



Feature Importances (Model-Based)

Random Forest's built-in feature importance values confirm that smoking, BMI, and age are the most critical drivers of insurance costs.

Result Interpretation

Top 5 Most Influential Features (by SHAP):

1. Smoker status (smoker_yes)
2. BMI
3. Age
4. Region (southeast)
5. Number of children
6. **Comparison with Model Feature Importance:**

- SHAP and Random Forest feature importances are consistent.
- Smoking, BMI, and age dominate predictions.

Domain Meaningfulness:

- Smokers are known to have higher medical costs.
- Higher BMI correlates with obesity-related health issues.
- Age naturally increases health risks.

Conclusion

SHAP provided transparent interpretation of the Random Forest model for insurance cost prediction. Results highlight smoking, BMI, and age as the strongest predictors, aligning with medical knowledge.

Future improvements could include:

- Testing other regression models (XGBoost, LightGBM)
- Applying hyperparameter tuning for higher accuracy
- Using more healthcare-related features if available