Machine Learning Model Comparison for Insurance Charges Prediction

Problem Statement

This project aims to predict insurance charges using the `insurance.csv` dataset (1,338 records) with features like `age`, `BMI`, `children`, `sex`, `smoker`, and `region`. We compare Linear Regression and Random Forest models (untuned, GridSearchCV, RandomizedSearchCV) to find the best model, using Mean Squared Error (MSE) and R^2 as metrics, with 5-fold cross-validation for robustness. The feature importance of the best model is analyzed to identify key drivers of charges.

Model Comparison

We evaluated six models, with results as follows:

- **Linear Regression (Untuned)**: Test MSE: 33,596,920, Test R^2: 0.784, CV MSE: 37,947,891
- **Linear Regression (GridSearchCV)**: Test MSE: 33,596,920, Test R^2: 0.784, CV MSE: 37,947,891 (Best Parameters: `fit_intercept=True`)
- **Linear Regression (RandomizedSearchCV)**: Test MSE: 33,596,920, Test R^2: 0.784, CV MSE: 37,947,891 (Best Parameters: `fit_intercept=True`)
- **Random Forest (Untuned)**: Test MSE: 20,957,080, Test R^2: 0.865, CV MSE: 24,441,731
- **Random Forest (GridSearchCV)**: Test MSE: 20,666,560, Test R^2: 0.867, CV MSE: 24,441,731 (Best Parameters: `max_depth=10`, `n_estimators=200`)
- **Random Forest (RandomizedSearchCV)**: Test MSE: 20,846,880, Test R^2: 0.866, CV MSE: 24,458,311 (Best Parameters: `max_depth=10`, `n_estimators=100`)

Random Forest (GridSearchCV) performed best, with the lowest Test MSE (20,666,560) and highest Test R^2 (0.867). Tuning slightly improved the Random Forest's Test MSE by ~290,520 compared to the untuned model. Linear Regression showed no improvement from tuning, as the default `fit_intercept=True` was optimal. Random Forest also generalized better, with a CV MSE of ~24.4M versus Linear Regression's ~37.9M.

Key Insights

1. **Model Performance**: Random Forest outperformed Linear Regression (Test R^2: 0.867 vs. 0.784), better capturing non-linear relationships in the data.
2. **Feature Importance**: For Random Forest (GridSearchCV), `smoker_yes` was the most important feature (importance 0.619), followed by `bmi` (0.211) and `age` (0.133). Features like `sex_male` (0.006) and regional indicators (<0.005) had minimal impact.
3. **Practical Implications**: Smoking status is the primary driver of insurance charges, with BMI and age also significant. Insurance pricing models should prioritize these factors over gender or region.

This analysis confirms Random Forest’s effectiveness for predicting insurance charges, with smoking as the key cost driver.