Machine Learning Model Comparison for Insurance Charges Prediction

Problem Statement

This project predicts 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 and tuned via GridSearchCV and RandomizedSearchCV—to determine the best model for accuracy. Evaluation uses RMSE and R^2, with 5-fold cross-validation for robustness. Feature importance from the best model highlights key drivers of charges.

Model Comparison

Six models were trained and evaluated:

- **Linear Regression (Untuned)**:
  - CV RMSE: 6,147.09
  - Test RMSE: 5,796.28
  - Test R^2: 0.78
- **Linear Regression (GridSearchCV)**:
  - Best Parameters: `{'fit_intercept': True}`
  - CV RMSE: 6,147.09
  - Test RMSE: 5,796.28
  - Test R^2: 0.78
- **Linear Regression (RandomizedSearchCV)**:
  - Best Parameters: `{'fit_intercept': True}`
  - CV RMSE: 6,147.09
  - Test RMSE: 5,796.28
  - Test R^2: 0.78
- **Random Forest (Untuned)**:
  - CV RMSE: 4,956.67
  - Test RMSE: 4,577.89
  - Test R^2: 0.87
- **Random Forest (GridSearchCV)**:
  - Best Parameters: `{'max_depth': 10, 'n_estimators': 200}`
  - CV RMSE: 4,920.88
  - Test RMSE: 4,546.05

    - Test R^2: 0.87
- **Random Forest (RandomizedSearchCV)**:
    - Best Parameters: `{'max_depth': 10, 'n_estimators': 100}`
    - CV RMSE: 4,922.55
    - Test RMSE: 4,565.84
    - Test R^2: 0.87

Best Model: Random Forest (GridSearchCV) achieved the lowest Test RMSE (4,546.05) and highest Test R^2 (0.87). Tuning improved Random Forest's Test RMSE by 31.84 from the untuned version (4,577.89 to 4,546.05), while Linear Regression showed no improvement (Test RMSE remained 5,796.28). Random Forest also generalized better, with a CV RMSE of ~4,921 versus ~6,147 for Linear Regression.

Key Insights

1. **Model Performance**:
    - **RMSE Explained**: RMSE reflects the average prediction error in dollars. A Test RMSE of 4,546.05 means predictions deviate by ~$4,546 from actual charges on average.
    - **Context**: Assuming a mean charge of ~$13,270 (based on typical insurance data), the Test RMSE is ~34% of the mean, showing moderate errors. Compared to a naive baseline (RMSE ~$12,110, the standard deviation), this is a significant improvement, but errors remain notable.
    - **Tuning Impact**: Random Forest's Test RMSE dropped by 31.84 after tuning, unlike Linear Regression, which saw no change.
2. **Feature Importance**:
    - For Random Forest (GridSearchCV), `smoker_yes` dominates (importance 0.619), followed by `bmi` (0.211) and `age` (0.133). Features like `sex_male` (0.006) and `region` (<0.005) have minimal impact.
    - **Implication**: Smoking status drives charges most, with BMI and age as secondary factors. Gender and region are less critical for pricing.
3. **Practical Implications**:
    - **Strengths**: The model (R^2 = 0.87) captures 87% of charge variance, making it useful for high-level pricing estimates, especially for smokers.
    - **Limitations**: The $4,546 RMSE suggests inaccuracies, particularly for lower charges (e.g., non-smokers), where this error is a larger proportion of the total.
    - **Business Use**: Insurance firms can prioritize smoking status, BMI, and age in pricing strategies. However, for precise individual quotes, the error margin indicates a need for refinement.

- **Next Steps**: Log-transforming charges, adding interaction terms (e.g., `age * smoker_yes`), or testing models like XGBoost could reduce RMSE further.

Conclusion

Random Forest (GridSearchCV) excels in predicting insurance charges, with smoking as the top cost driver. While effective (Test RMSE: 4,546.05, R^2: 0.87), the moderate error suggests further optimization is needed for precise pricing, especially for lower-cost cases.

google colab

Grok Conversation