Medical Insurance Cost Prediction (Regression)

This dataset is brought from kaggle and called Medical Cost Personal Datasets which has insurer information which might help us predict the insurance cost in the US

It has about 1338 row and 7 columns which are the following :

- Age : Age of primary beneficiary
- Sex : Insurance contractor gender, female, male
- BMI : Body mass index
- Children : Number of children covered by health insurance / Number of dependents
- Smoker : Whether a smoker or not
- Region : The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- Charges : Individual medical costs billed by health insurance

Notably the data is being split into 80% for training and 20% testing

Dataset Link : https://www.kaggle.com/datasets/mirichoi0218/insurance

Taissir Boukrouba - 22084758

Preprocessing Techniques Employed

Standard Scaler :

StandardScaler is a technique used to preprocess numerical data in machine learning. It standardizes features by centering them around a mean of 0 and scaling them to have a standard deviation of 1

Use : Because our values had varying ranges like *Age (18-64) , Children (0-5)* and other features. Leaving it as such will cause some problems like dominance of features with larger scales , distance-based algorithms (like SVM) would be affected and also some optimisation techniques used in training models (like gradient descent) can struggle with features on vastly different scales and this would cause a bad overall model performance

Label Encoder :

Label encoder is a technique used in machine learning to convert categorical data (like text labels) into numerical values. It's a simple and efficient way to prepare categorical data for use in machine learning models that typically require numerical inputs

Use : We have 2 categorical features which are *Sex (Male/Female)* and *Smoker (Yes/No)* and ML models can't directly understand and process text labels like "Male", "Female", or "Yes", "No". They require numerical features to perform calculations and learn patterns. Also it’s unclear for distance-based algorithms (like SVM) to calculate distance when variables are not numerical.

Fan, C., Chen, M

Preprocessing Techniques Employed

Log Transformation :

Log transformation is a data preprocessing technique applied to a feature (variable) to address issues like skewed distributions and improve the performance of machine learning models

Use : When doing the box plot , we have found that our target variable *charges* had an upper extreme (max) of about 35K USD but we found values that goes beyond that reaching about 64K USD (these were outliers). Applying the log transformation will help compress these large values into smaller ones reducing the gap between our normal values and the outliers and also get rid of the skewness of the distribution cause by them

The log transformation was applied solely to the training target variable, which represented medical charges. After obtaining predictions, we performed an anti-logarithmic transformation to convert the predicted values back to the original price scale for error evaluation. This allows us to assess the model's accuracy in terms of actual medical charges

Other Techniques :

To guarantee a clean dataset for modeling, I went beyond the mentioned techniques. Quality checks confirmed the absence of null values, a crucial step to avoid data inconsistencies. Additionally, I reviewed the cardinality (number of unique categories) within categorical features. This ensured a reasonable number of categories for machine learning models, potentially avoiding the need for further encoding techniques. Because the dataset was clean of null values and there were not features with very high or very low cardinality , their techniques (like imputation or dropping) were not used.

Fan, C., Chen, M

ML Architecture

Linear Regression : its architecture is a simple, single-layer model with a straight line representing the relationship between features (independent variables) and the target variable (dependent variable). It estimates a single weight for each feature to determine the slope and intercept of the line that best fits the training data

SVR: its architecture is more complex, can be linear or non-linear depending on the chosen kernel function (e.g., linear kernel for linear regression-like behavior, non-linear kernels like RBF for capturing complex relationships)**.** It focuses on finding a hyperplane (decision boundary) with the maximum margin between the data points and the margins (defined by a cost function). *Sabzekar, M.*

MLP Regressor: Its architecture is Inspired by biological neural networks, consisting of an input layer, one or more hidden layers with activation functions (introduce non-linearity), and an output layer. It uses backpropagation algorithm to iteratively adjust the weights and biases in each layer to minimize the error between the predicted and actual target values. *Murtagh, F*

RFR: it's an ensemble method that combines predictions from multiple decision trees, each built on a random subset of features and data points (bootstrapping).Each decision tree partitions the data space based on feature values to make predictions. The final prediction is the average (regression) . *Breiman, L*

Hyperparameter Tuning Using Grid-Search :

In machine learning, GridSearchCV from Scikit-learn automates hyperparameter tuning. It tests various combinations of hyperparameter settings and uses cross-validation to pick the one that optimizes a chosen performance metric.

Dufour, J.-M. and Neves, J

Summary Table of Chosen Parameters

Before choosing the best parameters , I did a **GridSearchCV** with 5 Folds for each model

01	Linear Regression	There were no parameters chosen for this one
02	Random Forest Regressor	• N_estimators : 100 • Min_samples_split : 10 • Max_depth : 8
03	Support Vector Regressor	• kernel: 'rbf' • 'epsilon': 0.1 • 'C': 1
04	Multi Layer Perceptron	• 'alpha': 5.623 • 'hidden_layer_sizes': (100,) • 'solver': 'lbfgs'

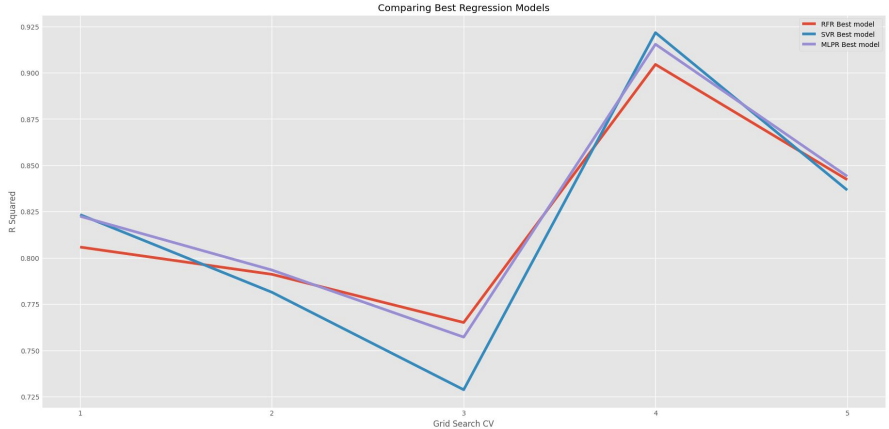
Results and critical evaluation of the models

Here we are using mainly R-squared to evaluate the models performance but we are going to also consider Mean Absolute Error because the aforementioned is more prone to overfitting and MAE is more interpretable to the real data .

Linear Regression serves as a good baseline model due to its simplicity and interpretability. It allows you to establish a fundamental understanding of the linear relationships between features and the target variable (medical prices) which is the reason i chose it as a dummy model. However, medical costs are likely influenced by various features that might have non-linear relationships. LR, by design, can only capture linear relationships. This limitation explains its lower R-squared (0.4) and high MAE (4.7K $)

Comparing Advanced Models :

While the baseline LR model established a starting point, the more sophisticated models (SVR, RFR, MLP) exhibited demonstrably superior performance in predicting medical prices :



Results and critical evaluation of the models

Top Performer:

Random Forest Regressor being an ensemble model clearly achieved the highest R-squared score (0.89) and lowest MAE score (1944) and , indicating its predictions deviated from the actual medical prices by an average of 1944$ on the chosen normal scale (anti-log). Considering our charges ranges from 1.1K $ to 64K $ this is a very good performance which suggests RFR performs best at capturing the central tendency of medical prices.

Close Contenders:

Support Vector Regressor (SVR) and Multi-Layer Perceptron (MLP) follow closely with R-Squared scores of 0.88 and 0.87 but the real difference is shown when comparing their Mean Absolute Error of 2022$ and 2378$ indicating an average deviation of about 2 - 2.3K $ from the actual prices. These models also provided good accuracy close to the top performer.

Overall :

These 3 Models did way better than our base model (Linear Regression) which was obvious since the correlations were very low in the correlation matrix indicating that the relationship between variables was non-linear which explained the bad performance of LR

RFR offered more interpretability compared to SVR (especially with complex kernels) and MLP (black box nature). This can be crucial when understanding the factors influencing price predictions (which is our case). In our case we also had a lot of features (7) which explained the robustness of RFR to high dimensionality compared to the other models. Finally , considering the small performance gaps between the three models, it reinforces the value of grid search in optimizing model parameters.

Limitations

Loss of Precision :

Performing two transformations (log and anti-log) can introduce slight numerical errors that might not be significant for other data types, but can be impactful when dealing with regression data which require careful precision like medical prices, where even small differences can matter

Reality of Medical Costs:

Medical prices often exhibit a right skew, meaning there are a few very expensive procedures and a larger number of less expensive ones. This skewness can pose problems for machine learning models.

Higher Variance Risk:

With only 1338 rows, this dataset might be considered relatively small for robust statistical analysis, especially for complex models. Smaller datasets can lead to higher variance in the model's performance and potentially unreliable results.

Generalizability Concerns:

A small dataset might not adequately capture the full range of variations present in the real population. The model trained on this data might not generalize well to unseen data points.

Potential Improvements

Deeper GridSearch :

While grid search has been valuable for hyperparameter tuning, exploring a wider range of parameters could further optimize the models and potentially squeeze out even better performance in medical price prediction.

Feature Engineering:

We might be able to create new features from existing ones (e.g., age groups, BMI categories) to capture more complex relationships.

Better model selection:

Linear regression, while interpretable, struggles with non-linear relationships in medical prices. On the other hand, complex models like SVM, MLP, and RFR might be prone to overfitting with limited data. To offer some balance we can use Lasso and Ridge (L1,L2 Regularisation techniques)

Combined Performances:

While both SVR, MLP, and RFR achieved good results in predicting medical prices, leveraging an ensemble learning approach has the potential to unlock even greater accuracy. Ensemble methods combine the predictions from multiple models, essentially creating a more robust results that leverages the strengths of each individual model.

*Also it's better to avoid **Distance-based Algorithms** like SVR which performed arguably well on this dataset , but it using a distance to measure error can be problematic when the features are not fit to scaling methods.*

References

1. Sabzekar, M. and Hasheminejad, S.M.H. (2021). Robust Regression Using Support Vector Regressions. *Chaos, Solitons & Fractals*, 144, p.110738. doi:https://doi.org/10.1016/j.chaos.2021.110738.
2. Fan, C., Chen, M., Wang, X., Wang, J. and Huang, B. (2021). A Review on Data Preprocessing Techniques toward Efficient and Reliable Knowledge Discovery from Building Operational Data. *Frontiers in Energy Research*, [online] 9. doi:https://doi.org/10.3389/fenrg.2021.652801.
3. Murtagh, F. (1991). Multilayer Perceptrons for Classification and Regression. *Neurocomputing*, 2(5-6), pp.183–197. doi:https://doi.org/10.1016/0925-2312(91)90023-5.
4. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (2017). *Classification and Regression Trees*. Routledge. doi:https://doi.org/10.1201/9781315139470.
5. Dufour, J.-M. and Neves, J. (2019). Finite-sample Inference and Nonstandard Asymptotics with Monte Carlo Tests and R. *Handbook of Statistics*, pp.3–31. doi:https://doi.org/10.1016/bs.host.2019.05.001.