Predicting Annual Medical Insurance Costs

Christian Bammann and Ryan Monroe
Department of Electrical and Computer Engineering
University of North Carolina at Charlotte

Introduction and Motivation

This project utilizes characteristic lifestyle data such as BMI and smoking history to predict an individual's health insurance cost for a calendar year using machine learning models.

This will provide insight into which attributes have the strongest effect on healthcare expenses, help improve cost forecasting for insurance providers, and highlight how lifestyle choices impact medical costs.

Dataset and Training Setup

Dataset

The dataset utilized for this project is the **Medical Cost Personal Datasets** from Kaggle, originally sourced from the book *Machine Learning with R* by Brett Lantz.

It contains **1,338** individual records, each with six features: `age`, `sex`, `BMI`, `children`, `smoker`, and `region`. There is one output variable, `charges`, that represents the total medical cost billed by insurance.

Training Setup

The model will be trained using three methodologies: **Linear Regression (LR)**, **Support Vector Machine (SVM)**, and **Artificial Neural Network (ANN)**, implemented with PyTorch.

Each model's performance will be evaluated using the **MSE, MAE** and $\mathbf{R^2}$ score.

Approach - Linear Regression (LR)

Why Linear Regression?

Linear Regression is a simple and easily interpretable baseline model. It helps understand linear relationships and provides a foundation for comparing against more complex models.

Preprocessing

Applied a log-transform to *charges* to stabilize variance, standardized numerical features using a StandardScaler, and encoded categorical features to convert them into numeric values.

Training Setup

Trained the LR model on an 80/20 train/test split and evaluated the model using MSE, MAE, and R^2 Score.

Expectation

Expected the Linear Regression model to capture general trends and serve as a baseline, but not to model aggressive nonlinear jumps present in the data.

Approach - Support Vector Machine (SVM)

Why SVM?

SVM with an RBF kernel is well suited for our dataset and excels at capturing nonlinear relationships. This makes SVM a strong choice for modeling.

Preprocessing

Applied a log-transform to *charges* to stabilize variance, standardized numerical features using a StandardScaler, and encoded categorical features to convert them into numeric values.

Training Setup

Trained the SVM model using the RBF kernel using a 80/20 train/test split. Evaluated the model using MSE, MAE, and R^2 Score.

Expectation

Expected the SVM model to outperform the LR model due to its ability to model nonlinear relationships.

Approach - Artificial Neural Network (ANN)

Why ANN?

Artificial Neural Networks have the capability to learn complex patterns in datasets. This can come at the expense of longer and more computationally intense training.

Preprocessing

Applied a log-transform to *charges* to stabilize variance, standardized numerical features using a StandardScaler, and encoded categorical features to convert them into numeric values.

Training Setup

80/20 train/test split
4-layer network w/ ReLU activation
8 inputs (32-16-8) 1 output
Hidden layers have 10% dropout
200 epochs, 0.005 learning rate

Expectation

The neural network should be able to identify and better predict the underlying trends in the data due the generalization abilities of the multiple layers.

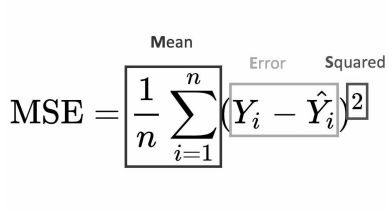
Evaluation Metrics

MSE

Mean Squared Error (MSE)

MSE is a measure of the average squared difference between and predicted and actual values.

Lower MSE = predictions are closer to the actual values.

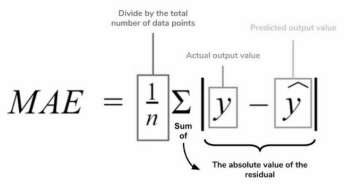


MAE

Mean Absolute Error (MAE)

MAE is a measure of the average magnitude of difference between and predicted and actual values.

Lower MAE = smaller average error.



R²

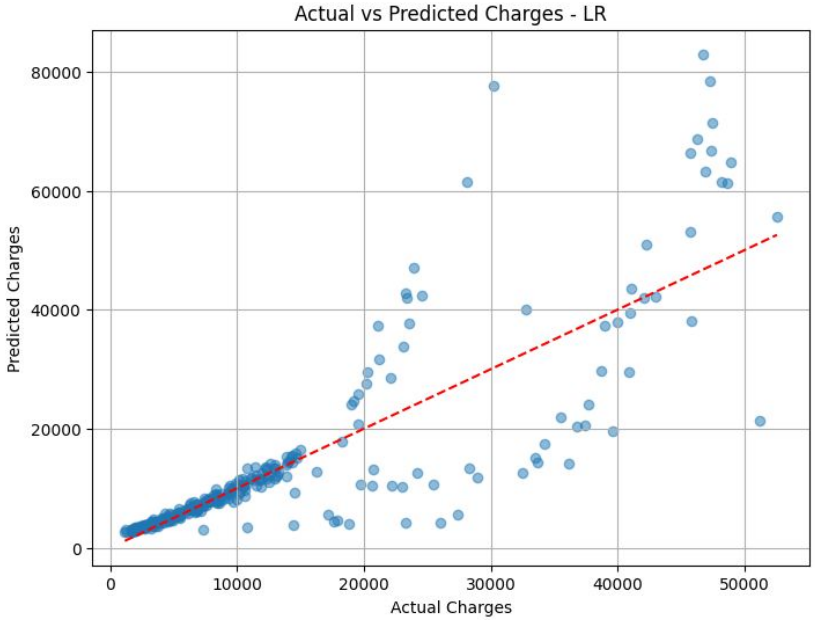
R² Score

R^2 Score (coefficient of determination) is the proportion of variance in the target explained by the model.

R^2 = 1.0 = perfect fit.

$$R^2 = 1 - \frac{\Sigma_i^N (y_i - \hat{y}_i)^2}{\Sigma_i^N (y_i - \bar{y})^2}$$

Actual vs. Predicted Charges – Linear Regression



Analysis

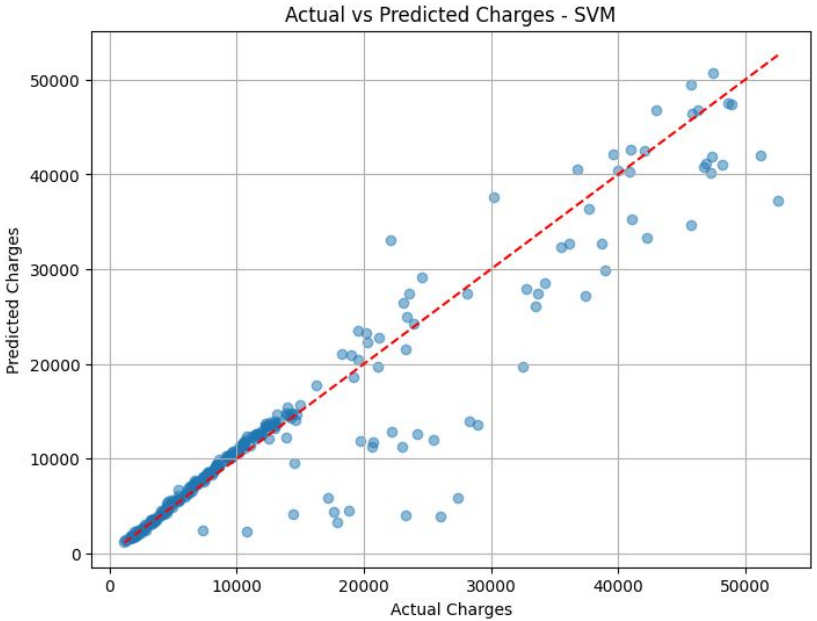
- Points follow general upward trend, showing LR model captures the overall relationship
- The model underestimates high-charge patients
- The wide spread scatter indicates underfitting for mid–high *charges*

Conclusion

- Linear Regression provides a reasonable baseline, but lacks accuracy for complex and nonlinear patterns.

MSE: 0.176 | MAE: 0.263 | R^2: 0.790

Actual vs. Predicted Charges – SVM



Analysis

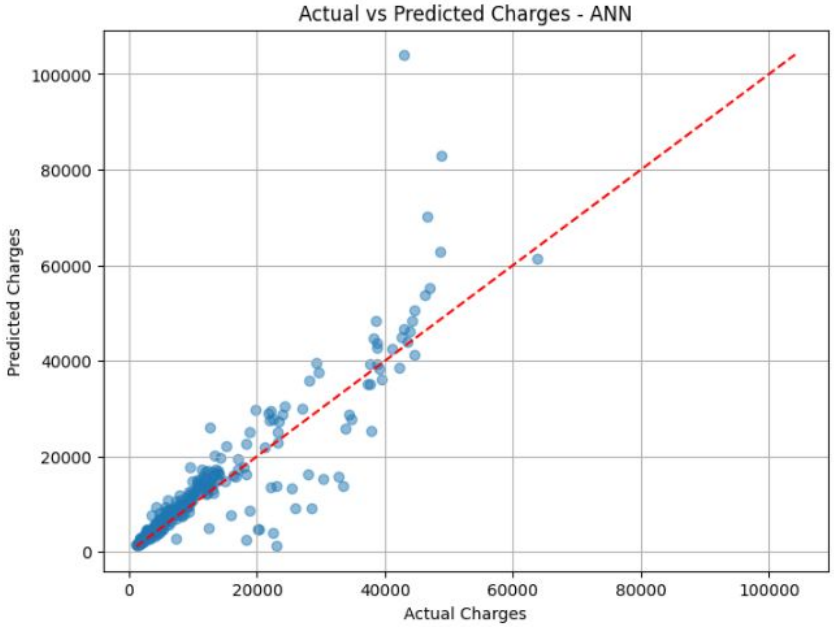
- Points lie much closer to the diagonal, showing much better prediction accuracy over LR
- SVM captures nonlinear relationships better and reduces error for mid–high ***charge*** patients
- Greatly improved fit, but still underestimates high-cost cases

Conclusion

- SVM model delivers substantially better performance than the Linear Regression model
- Provides noticeably higher accuracy for this dataset due to the ability to model nonlinear patterns

MSE: 0.107 | MAE: 0.156 | R^2: 0.872

Actual vs. Predicted Charges – ANN



Analysis

- The ANN was better at grouping data in the upper range of the data set
- Struggled to place outliers without carefully choosing training parameters
- Prone to overfitting without implementing dropout

Conclusion

- Time consuming to fine tune hyperparameters but is worthwhile for the quality
- Still prone to underpredicting larger targets

MSE: 0.160 | MAE: 0.257 | R^2: 0.822

Results

Model	Δ$	MSE	MAE	R^2 Score
Linear Regression	$3,150	0.176	0.263	0.790
Support Vector Machine (SVM)	$1,890	0.107	0.156	0.872
Artificial Neural Network (ANN)	$3,084	0.160	0.257	0.822

Interpretation and Analysis

The Artificial Neural Network (ANN) model performed *well*. ANN requires larger amount of data and more tuning.

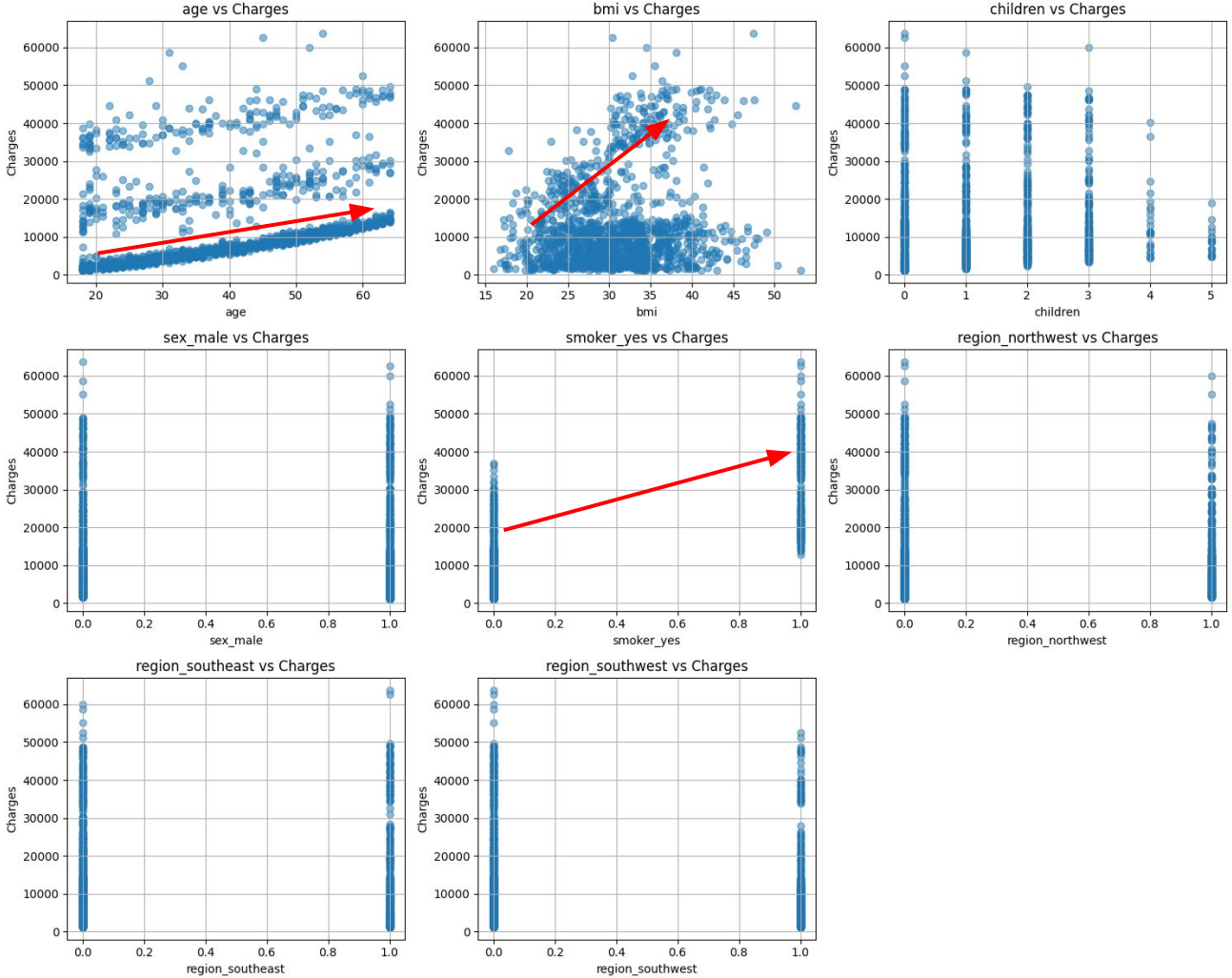
- MSE = 0.160
- MAE = 0.257
- R^2 = 0.822

The Support Vector Machine (SVM) model performed the *best*. SVM models excel on small nonlinear datasets.

- MSE = 0.107
- MAE = 0.156
- R^2 = 0.872

The Linear Regression (LR) model performed the *worst*. LR struggles with nonlinear data relationships.

- MSE = 0.176
- MSE = 0.263
- R^2 = 0.790



Key Trends:

- Age
- BMI
- Smoker

Features vs. Charges

Features vs. Charges – Analysis

After analyzing each feature, it became apparent that *smoker_yes* had the biggest effect on *charges*, followed by *age*.

Key Insights

- Individuals who smoked had significantly higher charges than those who do not smoke
- Age is the second strongest factor, with older individuals tending to have higher charges
- Higher BMI often corresponds to higher charges, especially when greater than 30
- Children, sex, and region have relatively minor influence on charges compared to the features above

Lessons Learned

The quickest and easiest way to identify apparent trends in a dataset is to apply a simple linear regression model. However, the results are nothing more than can be predicted intuitively. There is an inherent subjectiveness in the dataset that makes accurate predictions difficult.

Therefore, to thoroughly analyze underlying patterns it becomes necessary to utilize a neural network. Structuring a neural network comes with it's own challenges, such as ensuring it's complex enough to learn features yet still resistant to overfitting and exploding gradients.

Our models provide evidence of contributing factors to higher medical bills and insight on what medical costs to expect based on your lifestyle.

Thank You!