Predicting Annual Medical Insurance Costs

Christian Bammann
Department of Electrical and Computer Engineering
University of North Carolina at Charlotte
Charlotte, USA
cbammann@charlotte.edu

Ryan Monroe
Department of Electrical and Computer Engineering
University of North Carolina at Charlotte
Charlotte, USA
rmonro12@charlotte.edu

***Abstract*—This project explores the prediction of annual medical insurance costs using lifestyle and demographic data from the Medical Cost Personal Datasets sourced from Kaggle. The dataset includes 1,338 individual records each with six input features, and one output variable representing the total costs billed to a patient. Three models were implemented and analyzed: Linear Regression (LR), Support Vector Machine (SVM) with an RBF kernel, and a multi-layer Artificial Neural Network (ANN). Results show that the LR model provided a baseline (R² Score = 0.790), the ANN model showed significant improvement (R² Score = 0.861), and the SVM-RBF model achieved the highest accuracy (R² score = 0.872). Feature analysis was performed as well to determine which had the greatest impact on determining an individual's medical expenses. Smoking status and age were identified as the strongest contributing factors, with BMI also having a noticeable effect. These findings highlight the importance of nonlinear models in healthcare cost forecasting, and provide insight into which lifestyle factors have the biggest effect on a patient's medical costs. Python code, analyses, and results are available online on Github (see References).**

***Index Terms*—Linear Regression, Support Vector Machine, Neural Network, Regularization, Machine Learning, Medical Costs, Insurance Premiums, Artificial Intelligence, AI Prediction.**

I. Introduction

This project aims to utilize characteristic lifestyle data such as BMI and smoking history to predict an individual's annual health insurance cost 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. Importantly, the trained models cannot predict external factors. Sudden changes in health habits, such as gaining a large amount of weight or picking up smoking, would alter an individual's expected costs and cannot be accounted for. The same way a young healthy person who experienced a car accident would have a much higher cost than the models are capable of predicting.

II. Approach

This section details the approaches taken for each of the three models and what methodologies were implemented to ensure quality results. Anywhere the term 'charges' is used, it refers to the one-dimensional output of the machine learning model. This is the title of the output feature representing the total medical charges incurred by an individual in a calendar year. The same metrics are used to evaluate each model: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 Score. The most important for this application is the R^2 Score, as it measures how much variance in medical charges the model explains

A. Linear Regression (LR)

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. We elected to apply a log-transform to charges to stabilize variance, standardize the numerical features using a StandardScaler, and encode the categorical features to convert them into numeric values. We then trained the LR model on an 80/20 train/test split and evaluated the model using MSE, MAE, and R² Score. The Linear Regression model was expected to capture general trends and serve as a baseline, but not to model aggressive nonlinear jumps present in the data. Deeper relationships in the data, such as BMI, are not easily determined with this model. However, it is acceptably accurate for predicting charges for healthy individuals, barring the outliers discussed in the introduction.

B. Support Vector Machine (SVM)

For the SVM approach we determined an RBF kernel is best suited for our dataset because it excels at capturing nonlinear relationships making SVM a strong choice for modeling. Again, we 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. The same training splits and evaluation methods were used as the linear regression approach. We expected the SVM model to outperform the LR model due to its ability to model nonlinear relationships. Using this slightly more advanced approach we hoped to gain further insight into the deeper relationships in the data and more accurately predict larger charges. A successful training means the model could actually be viable as a tool for influencing lifestyle changes and encouraging accurate budgeting.

C. Artificial Neural Network (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. However, this increased versatility is highly desirable for our applications. The same methods were taken to stabilize variance, map the features, and scale the dataset. The architecture of the network is the primary and critical step to ensuring the model can feasibly learn the data trends with a sufficiently complex layer arrangement. After which, there is a fine balance in selecting hyperparameters so the model does not overfit to the data and completes training in a reasonable time. For our network, we structured it as follows:

- Five-layer network with three hidden layers and ReLU activation between each
- Layer sizes: 8–128–64–32–1
- Hidden layers have 10% dropout
- Weight Decay = 0.01
- Learning Rate = 0.005 using Adam optimizer
- Epochs = 400

The neural network is useful for its ability to identify the underlying trends in the data due the generalization abilities provided by the multiple layers, which are made more robust by the regularization techniques applied.

III. Dataset and Training Setup

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. All input features are shown in Table I. These features will be manipulated to suit the machine learning models, namely the four possible regions will be split into three separate features. If all three are false, then the fourth region is implied. The PyTorch and Scikit-learn libraries will be instrumental for model construction, training, testing, evaluation, and visualization.

TABLE I
Input Features

Age
Sex
BMI
Children
Smoker
Region

It is important to consider the nature of the dataset and the limitations of data collection. The nuanced history of each individual cannot be represented accurately with six features. Hence, there is inherently a large amount of subjectivity built into the dataset which will make outliers frequent and unpredictable. These anomalies could be explained by unforeseen life events or accidents that are not captured in the features present. Additionally, there are less datapoints for individuals that are more likely to experience higher medical costs which leads to difficulty with training for these outliers.

The models will be trained following the approaches described in Section II. The results and evaluation of each performance are discussed in the sections below.

IV. Results

This section includes the results and graphs for each of the three models' performance.

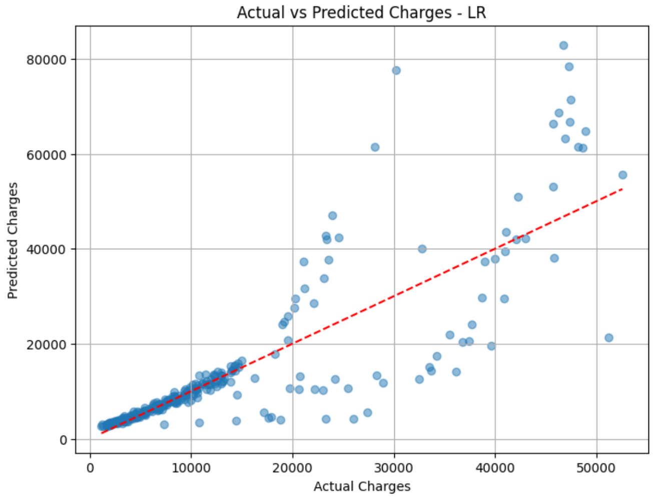

Fig. 1. Linear Regression Model

The Linear Regression model was able to capture the overall relationship, but underestimates mid-high charge individuals. The wide spread shown is a key sign of underfitting.
LR Metrics:

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

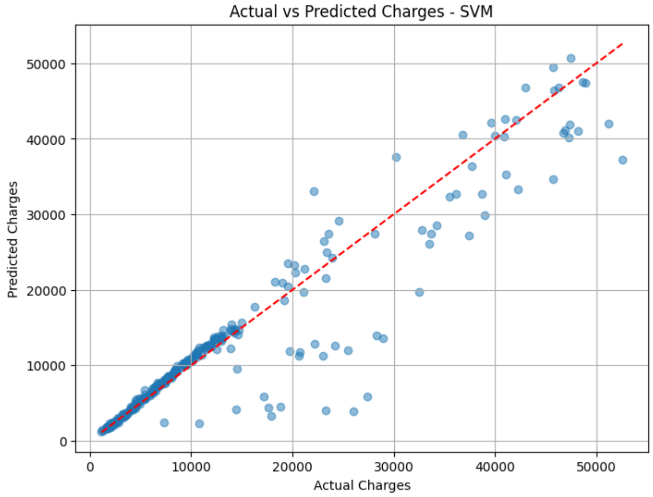

Fig. 2. Support Vector Machine Model

The SVM model captures the nonlinear relationship of the dataset much better than the Linear Regression model. The

error for mid-high charge patients is reduced as well.
SVM Metrics:

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

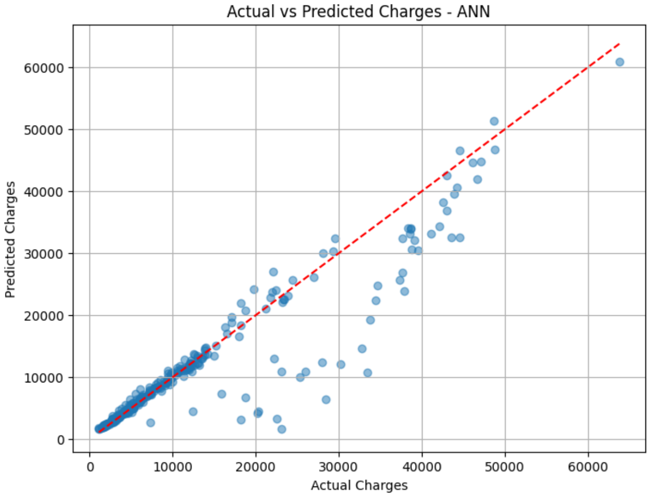

Fig. 3. Artificial Neural Network Model

The ANN model was much better at placing outliers in the mid-high range of the dataset.
ANN Metrics:

- $MSE = 0.125$
- $MAE = 0.194$
- $R^2 = 0.861$

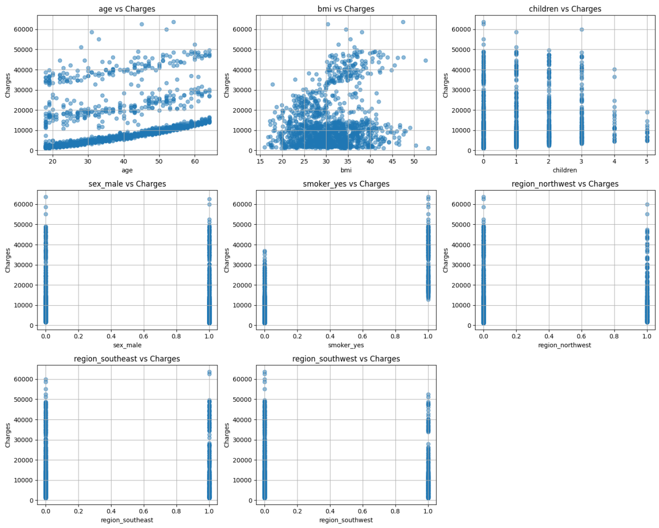

Fig. 4. Visualization of Input Features

From the above plots we can determine major trends that should be analyzed in the predictions. The largest contributors to higher charges are smoking status, age, and BMI. The distinct patterns of the age feature demonstrate the consistent increase of charges with age, as well as the jump in charges between smokers and obese individuals at a given age. A comparison of each model's performance can be seen in Table II.

TABLE II
MODEL PERFORMANCE COMPARISON

Model	MSE	MAE	R^2
LR	0.176	0.263	0.790
SVM	0.107	0.156	0.872
ANN	0.125	0.194	0.861

V. ANALYSIS

A. Linear Regression (LR)

The predicted points follow a general upward trend, showing that the LR model captures the overall relationship of the more standard cases. The model is susceptible to severely underestimating high-charge individuals. The wide spread scatter indicates underfitting for mid to high charges. Linear Regression provides a reasonable baseline, but lacks precision for complex and nonlinear patterns.

B. Support Vector Machine (SVM)

All predicted outputs lie much closer to the diagonal, showcasing the increased prediction accuracy over the LR method. SVM captures nonlinear relationships better and reduces the error for mid to high charge individuals. Greatly improved fit throughout the dataset, but still prone to underestimating high-cost cases. The SVM model delivers substantially better performance than the LR model because of its noticeably higher precision for these predicted values. This can be attributed to the model's capability of learning nonlinear relationships.

C. Artificial Neural Network (ANN)

The ANN was significantly better at grouping data and placing outliers in the mid to high range of the dataset. Even considering the raw evaluation characteristics of the model don't outperform the SVM, the plotted predictions portray a more desirable result. For this application, the models performances are poor if they severely underestimate charges, which the ANN suffers less from. The ANN is still prone to overfitting without implementing the proper regularization techniques. It was time consuming to fine tune hyper-parameters but it was worthwhile to achieve the best performance. The same grouping of data points can be seen near the bottom left of the predictions plot for all three models. These are likely caused by an external factor that is not captured in the dataset, therefore all models struggle to accurately predict the charges.

VI. CONCLUSION

The quickest and easiest way to identify apparent trends in a dataset is to apply a simple linear regression model. However, the results are similar to what can be predicted intuitively. There is an inherent subjectiveness in the dataset that makes accurate predictions difficult, primarily for the higher charges. Therefore, to thoroughly analyze underlying nonlinear patterns, it becomes useful to utilize a neural network. Structuring a neural network comes with its own challenges, such as ensuring it's complex enough to learn features yet still resistant to overfitting and exploding gradients.

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 (obese)
- Children, sex, and region have relatively minor impact on charges compared to the other features

Our machine learning models provide evidence of contributing factors to higher medical bills and insight on what medical costs to expect based on your lifestyle. The trained models could effectively be used to predict the annual medical charges someone can expect to incur, given only their smoker status, BMI, and age. For nonsmokers who are not obese the predictions will be exceptionally accurate, and the results can be used to reinforce lifestyle habits and promote active saving for expected expenses.

References

[1] C. Bammann, *Predicting Annual Medical Insurance Costs*, GitHub repository, Dec. 2025. [Online]. Available: https://github.com/christianbammann/Predicting-Annual-Medical-Insurance-Costs.

[2] M. Choi, *Medical Cost Personal Datasets*, Kaggle, Dec. 2025. [Online]. Available: https://www.kaggle.com/datasets/mirichoi0218/insurance/data.

[3] U. Orji and E. Ukwandu, *Machine Learning For An Explainable Cost Prediction of Medical Insurance*, arXiv preprint, Nov. 2023. [Online]. Available: https://arxiv.org/pdf/2311.14139.