CS 4641 Final Report

Introduction/Background

Medical insurance premiums can vary massively based on health and demographic data. Insurance companies calculate based on factors such as age, smoking status, and prior conditions, as the cost of healthcare will often rise with various vulnerabilities. While the specific algorithms are unique to each provider, we can use data analysis techniques to estimate different individuals' health care costs based on their demographics. For this project, we decided to use three supervised training methods: a linear regression model [2], a random forest regressor [3], and a neural network [5]. These three methods were chosen because our dataset seeked to predict a continuous value and came with labelled data that we could use as training data.

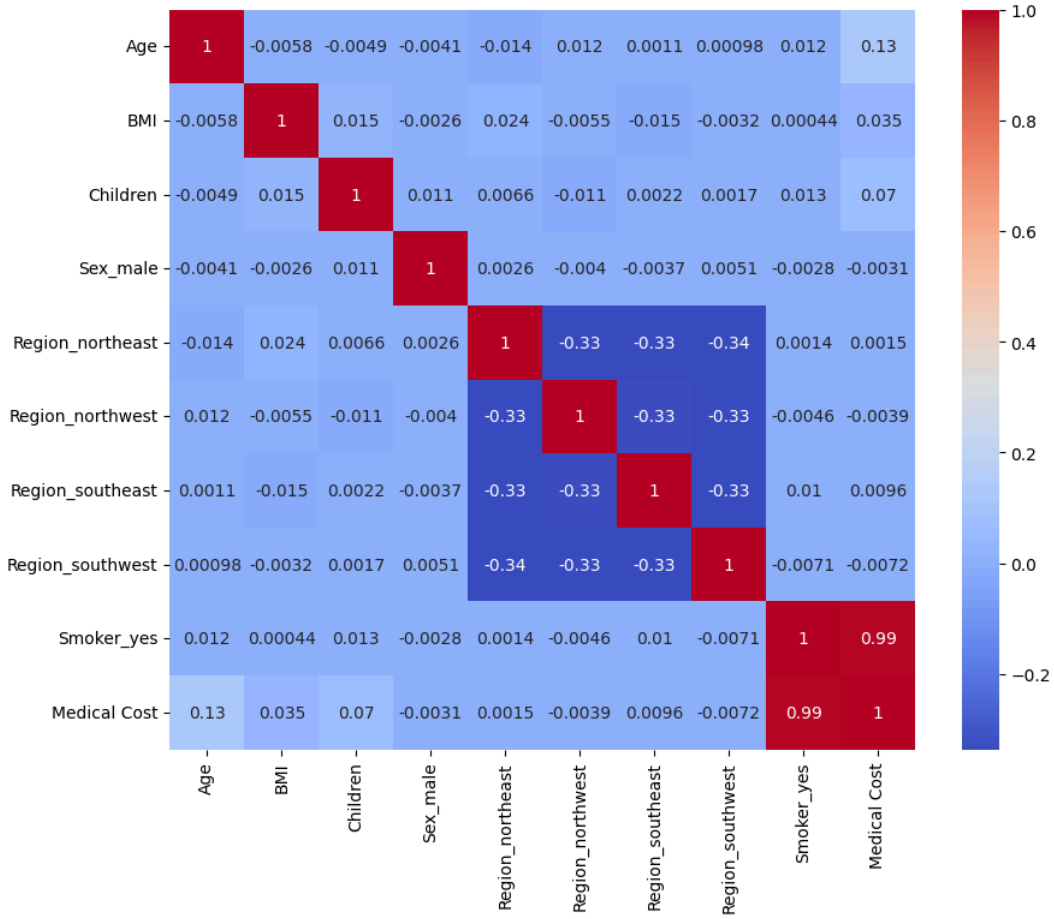
Problem Definition

Healthcare insurance costs have been rising due to factors like aging populations, chronic illnesses, and unequal access to care. For healthcare providers and insurers, predicting an individual's healthcare costs based on health and demographic data is vital for resource allocation, insurance planning, and intervention strategies. This project focuses on using healthcare data, including medical history, health metrics (blood pressure, BMI, etc...), and demographic factors (age, gender, socioeconomic status, etc...), to build a predictive model for estimating healthcare costs. The motivation stems from the need for more efficient resource management, fair premium adjustments, and early interventions for high-risk patients. Additionally, predictive models can help highlight healthcare disparities, guiding policymakers in addressing inequities. By leveraging machine learning, this project aims to bring to light how different socioeconomic factors can affect an individual’s healthcare insurance costs.

Methods

Dataset Description

There are seven features in our dataset: age, sex, BMI, children, smoker, region, and medical cost. The dataset's data was compiled with anonymous patients' medical costs and data from 2010-2020. Additionally, there are 10,000 complete data points in the dataset. From our preliminary exploration of the dataset, we discovered that our dataset is highly correlated on smoking, which is revealed in the correlation matrix below.



Data Preprocessing

We began our implementation with preprocessing the data through one hot encoding, which creates a separate boolean column for each potential category and doesn't imply any numerical relationship like label encoding [1]. We performed one-hot encoding on all of the categorical data, including sex, region, and smoker. This helped to convert the dataset into something that linear regression can use. We also checked for any null values in the dataset to impute but found none, so we did not have to remove any incomplete data from the dataset. Finally, we ended by normalizing the input features to a consistent scale using sklearn's MinMaxScaler. This is helpful for models that are sensitive to variance, like neural networks.

Supervised Learning Models

Moving on to the actual models, we began by implementing a linear regression model [2]. Since we are trying to predict a continuous value, linear regression is an easy and fast way to identify a relationship between the features. The discriminative nature of the model allows for quick direct estimation and a clear equation to help in codifying a hypothesized relationship. We used scikit-learn to quickly perform the regression and pandas to store and process the data. This proved to be an effective method in our testing and other sources have obtained useful results from this model [4].

Next, we created a random forest regressor [3]. This differs from a classifier in that we are predicting a continuous value instead of a discrete value. Random forest regressors work by creating many trees to predict any specific patient's medical cost. By having many trees that randomly draw from the features of the dataset, this helps to combat overfitting and any heavy dependence on a single feature. We were able

to implement this regression using scikit-learn, and this was a good model for our data.

Finally, we also created a neural network consisting of three fully connected layers and trained the model on the dataset [5]. Neural networks are trained to predict medical costs by running the inputted data through the network and backpropagating the losses through each neuron to update their weights so that they can learn how to make a better prediction. After 50 epochs of updates, we have fully trained our new model. We implemented this neural network using tensorflow and it proved to be a good model for our data.

Results and Discussion

Quantitative Metrics

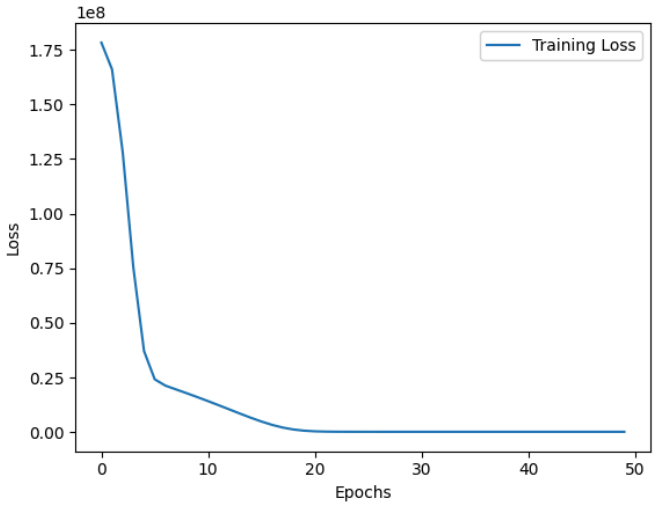
We used three quantitative metrics to quantify our model and subsequent results: the mean absolute error, the root mean squared error, and the R-squared (Coefficient of Determination) value. Here are the results that we obtained from our testing:

Model Type	Mean Absolute Error	RMSE	R-squared value
Linear Regression	251.097	290.074	0.99772
Random Forest	266.829	317.305	0.99727
Neural Network	251.889	291.810	0.99769

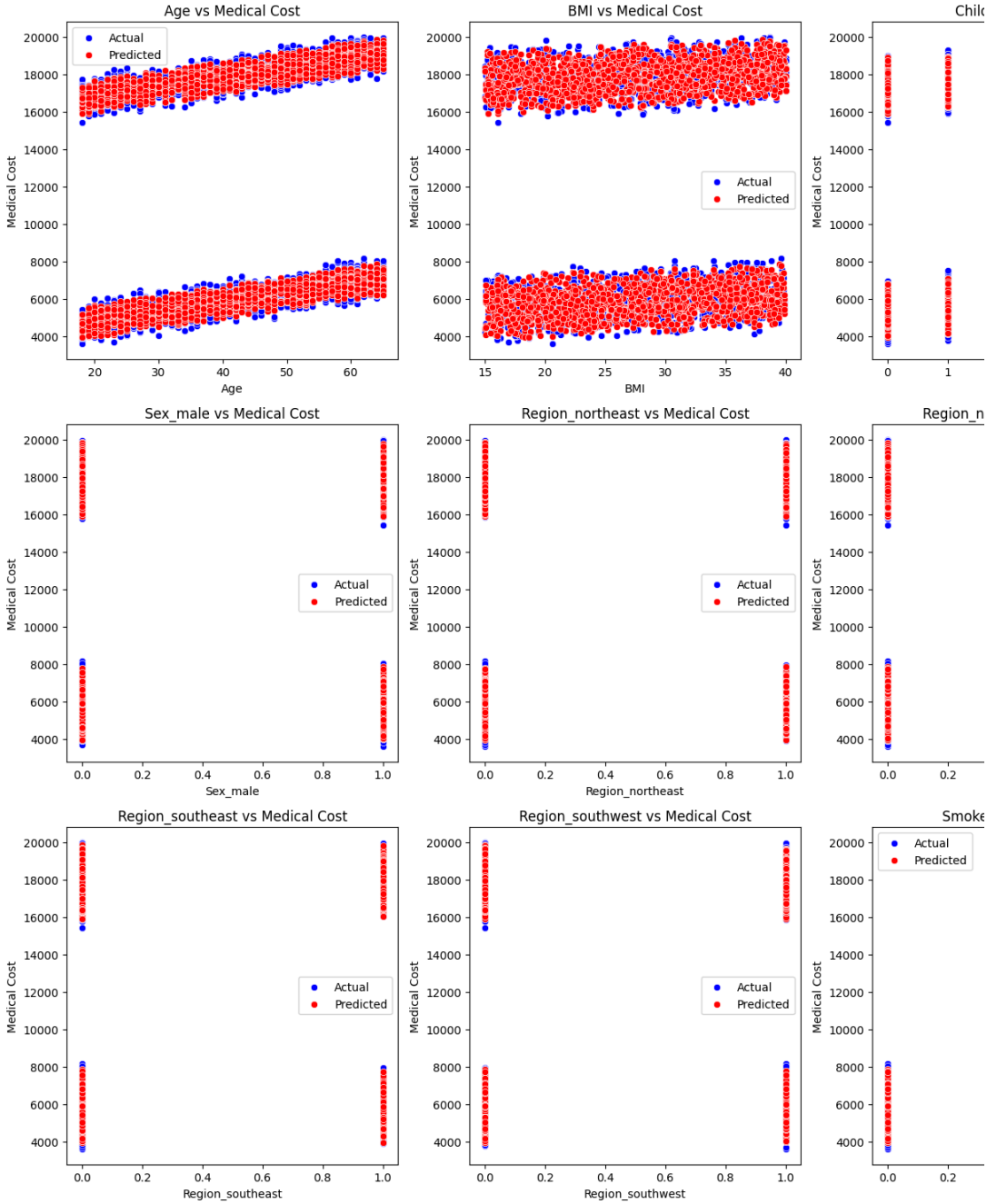
Surprisingly, the linear regression model, which we anticipated would perform the worst, ended up with the lowest error. The neural network performed almost exactly the same (albeit very slightly worse) and the random forest performed noticeably worse than the other two. However, the mean absolute error in average cost was relatively good as the medical costs are in the thousands or tens of thousands, so a difference of $251 or $266 is not that bad. Looking back on our results, the reason the linear regression model worked so well was because the correlation between features and the overall medical cost seemed to be very linearly coordinated. One potential reason the random forest regression model did not work as well was because by far the strongest indicator of medical cost was whether someone was a smoker or not. By attempting to prevent overfitting by training the trees in the random forest on only a select few random features, it is possible some of those trees did not include the smoker trait, which was a major indicator of medical cost. Furthermore, there is a chance that our neural network overfitted on the training data, as training loss plateaued near epoch 20, so any later epochs might have caused our neural network to overfit on the training data. By overfitting, our neural network would have performed worse on the validation data, thus leading to potentially worse error scores.

Visualization

Neural Network Training Loss:



Model Prediction Visualization:



This set of graphs represents the test dataset compared to the predictions of the linear regression model, across all of the features. As you can see, there are primarily two bands to the actual medical costs: around $4,000 to $7,000 and $16,000 to $19,000. Our models generally did a good job at predicting within these bands. Furthermore, by far the best predictor of medical prices was whether one was a smoker or not, followed by one's age.

Analysis

Our project aims to provide individuals with a good estimate of their healthcare insurance cost so they can determine for themselves the credibility of various insurance providers' claims. Writing our project proposal, we wished to achieve an accurate model without overfitting on the training data, aiming for a target mean absolute error of below $500, which would be reasonable for predicting healthcare insurance prices which started in the thousands. Given the results from our linear regression model, we were able to achieve our goals. We believe that our model performed well because there was an underlying linear relationship between the features and medical costs, that there was little noise in the dataset, and the features were not highly correlated with each other.

References

[1] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," Journal of Big Data, vol. 7, no. 28, pp. 1-41, 2020. doi: https://doi.org/10.1186/s40537-020-00305-w.

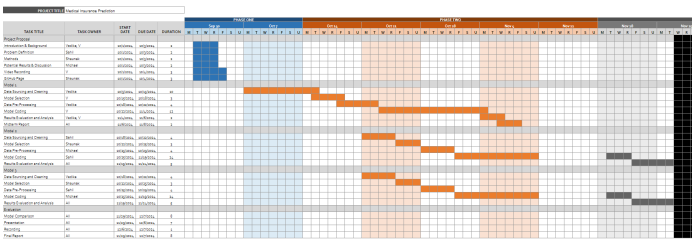
[2] A. Schneider, G. Hommel, and M. Blettner, "Linear regression analysis: part 14 of a series on evaluation of scientific publications," Deutsches Arzteblatt International, vol. 107, no. 44, pp. 776–782, 2010. doi: https://doi.org/10.3238/arztebl.2010.0776.

[3] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324.

[4] Y. Hu, “Comparison and Analysis of the Effectiveness of Linear Regression, Decision Tree, and Random Forest Models for Health Insurance Premium Forecasting,” Advances in economics, management and political sciences, vol. 79, no. 1, pp. 347–353, Apr. 2024, doi: https://doi.org/10.54254/2754-1169/79/20241754.

[5] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural Networks, vol. 61, no. 61, pp. 85–117, Jan. 2015, doi: https://doi.org/10.1016/j.neunet.2014.09.003

Gantt Chart



Contribution Table

Name	Contributions
Sahil	Video Presentation, Problem Definition, Quantitative Metrics
Michael	Random Forest, Introduction & Background, Data Preprocessing
Shaunak	Linear Regression, Neural Network, Visualization, Github Page
V	Video Presentation, Analysis