Predicting Medical Costs with Gradient Descent

Carlos Iván Fonseca Mondragón | A01771689

ABSTRACT This report presents the analysis and implementation of a linear regression model for predicting medical costs for an insurance company, based on a dataset with demographic and lifestyle information about its users. The methodology includes data cleaning, categorical variable encoding, outlier treatment, and feature normalization. The model works through an implementation of gradient descent, and its performance is graded using the mean squared error and R^2. This approach yields an R^2 value of 0.71 on the test set, meaning that it can explain 71% of the variance in the costs.

1. INTRODUCTION

The rising costs of healthcare have certainly become a factor in the research and development of models that can help estimate the medical expenses of an individual based on demographic and behavioral variables. Accurate estimations not only help insurance companies optimized their pricing behavior, but also support the public health planning sector. With this context, linear regression provides a basis for modelling a relationship between independent variables (such as age, sex, body bass index, smoking status and region) with a healthcare charge.

This study aims to implement and evaluate a linear regression model for predicting medical costs using a publicly available dataset from the book Machine Learning with R by Brett Lantz, that includes individual-level health and demographic information. The regression model currently presented has been constructed using Python, without the use of any machine learning frameworks, providing a manual application of gradient descent in order to optimize the algorithm. The benefit to this method is a granular view into the learning process, error convergence and parameter adjusting.

Before training, the dataset undergoes several preprocessing steps to improve model performance and reliability. These include the transformation of the BMI variable into categorical classifications, encoding of categorical features into numerical format, winsorization of the target variable to mitigate the impact of outliers, and feature scaling via min-max normalization. The results that this implementation achieves a satisfactory level of prediction accuracy on new data, explaining over 70% of the variance in medical charges.

2. DESCRIBING THE DATASET

"*Medical Cost Personal Datasets"* is a fictional dataset provided in the book Machine Learning with R by Brett Lantz, it contains 1338 columns with the *age, sex, bmi, children, smoking habits, region, and* a *charge* in the insurance. Next, the

columns will be explained in greater detail:

- **age:** The age of the individual, that typically ranges from 18 to 64 years in this dataset. Important because healthcare costs generally increase with age.
- **sex:** The biological sex of the individual: 'male' or 'female'. May correlate with certain medical risks or cost patterns.
- **bmi:** Body Mass Index — a standardized measure of weight relative to height. High BMI often indicates overweight or obesity, which are risk factors for higher medical costs.
- **children:** The number of children covered by the insurance plan.
- **region**: The U.S. geographic region where the person resides: 'northeast', 'northwest', 'southeast', or 'southwest'. Regional cost variations and healthcare access disparities can impact charges.
- **charges:** The total medical insurance charges billed to the individual. This will be the value to predict.

3. DATA TREATMENT

'*insurance.csv*', as the dataset is called, was converted into a Pandas dataframe, which is the first step in transforming the data.

3.1 BMI Transformation

The first manipulation of the data was a change to the *bmi* variable, from continuous into categorical ranges, following standard medical classifications. The categories used were:

- Underweight (< 18.5)
- Normal weight (18.5–24.9)
- Overweight (25.0–29.9)
- Class 1 obesity (30.0–34.9)
- Class 2 obesity (35.0–39.9)
- Class 3 obesity (≥ 40.0)

The resulting bmi_category column replaced the original *bmi* column in the dataset.

3.2 Categorical encoding

Variables *sex, bmi_category, smoker,* and *region* were encoded numerically with the use of dictionaries. For instance, smokers were encoded as 1 for "yes" and 2 for "no", allowing the model to learn cost differences associated with smoking behavior.

3.3 Winsorization of outliers

The *charges* variable was, initially, heavily skewed to the right, with several extreme values. To mitigate the influence of extreme outliers in the charges variable, a winsorization technique was applied at the 5th and 95th percentiles, following robust statistics recommendations (Wilcox, 2012; Tukey, 1962). This technique preserves most of the data's structure, while reducing the data's skewness caused by extreme values.

3.4 Train/Test split

The dataset was split into training (70%) and testing (30%) subsets. This partition ensures that the model is evaluated on unseen data, providing an estimate of its generalization ability.

3.5 Min-max scaling

To ensure all numerical variables contribute equally to the model, min-max scaling was applied to features like age

and children, following common preprocessing practices for gradient-based learning algorithms (Han et al., 2011; Pedregosa et al., 2011).

4. MODEL IMPLEMENTATION

The model calculates a weighted sum of all the input variables, plus a **bias**. These weights (or parameters) start at zero and are updated on each epoch (or run of the gradient descent).

The model improves by looking at how far off its predictions are and correcting itself repeatedly over 3,000 epochs, or until the error is lower than .01. The learning rate (how big each adjustment is) was set to 0.03, which provided a good balance between speed and accuracy.

Linear regression with gradient descent was chosen for these two reasons:

- It's simple and interpretable.
- It shows how much each variable contributes to the final prediction.

Although this model has limitations and doesn't capture complex relationships like a neural network would, it's still valuable when transparency and simplicity are important.

5. TESTING AND RESULTS

After training the model on 70% of the data, its performance was evaluated using the remaining 30% (test set).

The parameters learned during the training run were applied at the test to generate predictions. Since both the input and target variables had been normalized previously using min-max scaling, the predictions were rescaled back to their original scale to enable proper interpretation.

The model's predictive accuracy was first assessed using Mean Squared Error (MSE). This value represents the average squared difference between the predicted and actual costs. Although it is generally not easily interpreted by a human, a lower value generally indicates better performance. The achieved Test MSE was 38031994.34.

To evaluate how well the model explains the variability of the target variable, the R^2 score was computed, with a value of **0.71**. This indicates that approximately 71% of the variance in medical charges is explained by the model. It confirms that features such as the *smoking status* and *bmi* have a large weight when predicting the cost.

A scatter plot was generated to compare predicted and actual charges of all the patients in the test set. Even though the plot shows a good level of alignment between real and predicted values, there is still some divergence in more extreme cases.

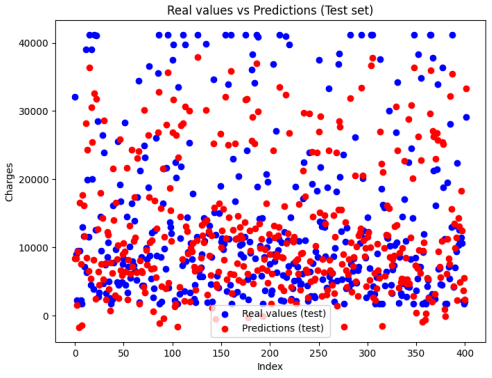

Figure 3: Real values plotted against the models predictions.

The cost function was plotted across all epochs to observe convergence behavior. The curve shows a clear decline in the mean square error over time, proving that the learning rate and epochs were set at appropriate values. The training finished before reaching the full 3,000 iterations, thanks to an early stopping criterion triggered when the training error fell below 0.01.

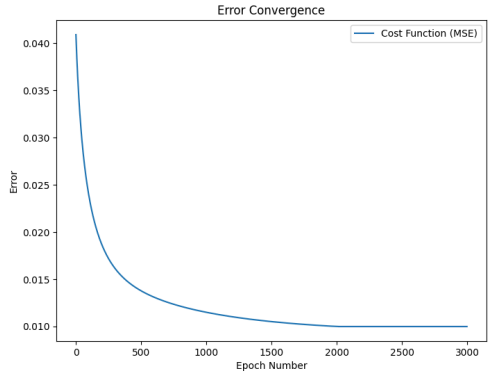

Figure 4: Error convergence plot.

6. CONCLUSION

This project successfully implemented a linear regression model without the use of dedicated frameworks, predicting insurance costs using demographic and lifestyle data. After preprocessing the data with BMI classification, categorical encoding, outlier handling, and normalization, I implemented gradient descent optimization that converged well.

The model achieved an R^2 score of 0.71 on the test set, meaning it explains 71% of the variance in medical costs, which is a good score for such a simple approximation model.

The model shows its limit, since it cannot capture complex relationships that are often present in real world data, like how age and smoking status might contradict each other's effect in calculating a cost.

This project proves that, with a good enough data processing and use of optimization basics, even a simple model like gradient descent can yield meaningful insights.

7. REFERENCES

a. Tukey, J. W. (1962). *The future of data analysis. The Annals of Mathematical Statistics*, 33(1), 1–67.

b. Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques.* Elsevier.