Predictive Analysis of Health Insurance Charges: A Data-Driven Approach Using Gradient Descent Optimization on Individual Health Profiles

Lucas Teixeira Rocha - m11813 .
Computer Science and Engineering - 2nd Cycle Degree
Universidade da Beira Interior
Covilhã, Portugal
lucas.rocha@ubi.pt

***Abstract*—In this research, we utilize the Medical Cost Personal Dataset from Kaggle.com to investigate the feasibility of predicting health insurance charges based on individual characteristics, including age, sex, Body Mass Index (BMI), number of children, smoking status, and geographical region. The dataset encompasses diverse individual profiles, each associated with specific health insurance charges. Our primary analytical tool is a linear regression model, expressed as $h_0(x_i) = \theta_0 + \theta_1 \cdot \textbf{age} + \theta_2 \cdot \textbf{sex} + \theta_3 \cdot \textbf{BMI} + \theta_4 \cdot \textbf{children} + \theta_5 \cdot \textbf{smoker} + \theta_6 \cdot \textbf{region}$.**

To optimize the parameters $(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$ of our model, we implement the Gradient Descent algorithm. This iterative optimization algorithm adjusts the parameters to minimize the cost function, which measures the discrepancy between the predicted and actual insurance charges. Through this process, we aim to identify the combination of parameters that best capture the relationship between the individual features and insurance charges.

The Gradient Descent algorithm's efficiency and effectiveness in optimizing the model parameters are systematically evaluated. The results of this study not only illuminate the significant factors influencing health insurance charges but also demonstrate the viability of using Gradient Descent for predictive analysis in health insurance charge estimation. Ultimately, this research contributes to the development of more transparent, fair, and data-driven approaches for determining health insurance premiums, thereby facilitating better-informed decision-making by both insurance providers and policyholders.

***Index Terms*—Medical Cost Personal Dataset, Gradient Descent, Linear Regression, Data analysis.**

I. Introduction

Health insurance serves as an essential tool that shields individuals from exorbitant medical expenses by distributing the financial risk among a broader population. Understanding and predicting the charges associated with health insurance policies is crucial for both insurers and policyholders. For insurers, accurate predictions of charges are essential for pricing policies competitively while maintaining profitability [1]. For policyholders, understanding the determinants of insurance charges helps in making informed decisions regarding health insurance purchases and healthcare planning [2].

The Medical Cost Personal Dataset, accessible on Kaggle.com, provides a valuable resource for studying the relationship between individual characteristics and health insurance charges. This dataset comprises several features, including age, sex, Body Mass Index (BMI), number of children, smoking status, and geographical region, along with the corresponding health insurance charges billed to individuals. Each feature represents a potential determinant of health insurance charges, warranting a detailed analysis of their predictive power and significance.

In this research, we aim to examine whether health insurance charges can be effectively predicted using the features provided in the Medical Cost Personal Dataset. We employ a linear regression model, formulated as $h_0(x_i) = \theta_0 + \theta_1 \cdot \text{age} + \theta_2 \cdot \text{sex} + \theta_3 \cdot \text{BMI} + \theta_4 \cdot \text{children} + \theta_5 \cdot \text{smoker} + \theta_6 \cdot \text{region}$. The parameters of this model $(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$ are optimized using the Gradient Descent algorithm. Through this study, we seek to identify the significant predictors of health insurance charges and evaluate the effectiveness of Gradient Descent in optimizing the prediction model for health insurance charge estimation.

A. Document Structure

The structure of this paper is meticulously organized to facilitate a coherent and systematic presentation of our study. It is delineated as follows:

- **Section I: Introduction** - The present section, offering a broad overview and insight into the objectives and significance of the current study.
- **Section II: Exploratory Data Analysis** - This section engages in a rigorous process of data normalization and sanity checking to ensure the dataset's reliability and appropriateness for subsequent analysis.
- **Section III: Algorithm Implementation** - Here, we craft a Python-based algorithm from scratch, meticulously aligning with the principles of Gradient Descent, to derive an optimal model for our data.
- **Section IV: Performance Measurement** - In this segment, we gauge the efficiency and accuracy of our implemented algorithm, utilizing the robust k-fold validation scheme to ensure a comprehensive assessment.

- **Section V: Consistency Across Folds: Low Variance Evidenced** - An in-depth analysis is conducted to elucidate the disparities in performance exhibited by the models generated for different k-folds, providing valuable insights into their variability and consistency.
- **Section VI: Implementation of Polynomial Modeling** - A polynomial model of order p is fitted to our dataset. The analysis outlined in Section V is reiterated to evaluate the model's predictive prowess and reliability.
- **Section VII: Conclusions and Future Directions** - This concluding section encapsulates the key findings of our study, drawing final insights and inferences from the analysis conducted in preceding sections.
- **Section H: References**.

II. Exploratory Data Analysis

A. Conversion to Numeric Values

Converting all data to numeric values is a pivotal step in the data preprocessing phase for various machine learning algorithms, including linear regression. The conversion ensures that the dataset is suitable for the algorithm as most mathematical models require numeric input for computations [3]. This transformation enhances the model's ability to learn from the data efficiently, as it can easily interpret and process numeric values, leading to more accurate and reliable predictions. Furthermore, the conversion facilitates straightforward computation of gradients in optimization algorithms like Gradient Descent, which is crucial for effective model training and parameter tuning. In our study, features like sex, smoker status, and region need to be converted into appropriate numeric formats, such as through one-hot encoding or label encoding, to facilitate the application of linear regression.

B. Handling Missing or NULL Values

Addressing missing or NULL values is another fundamental aspect of data preprocessing. Missing values in the dataset can lead to inaccurate or misleading results since they can distort the distribution and relationships among variables [4]. Identifying and handling these missing values appropriately, through techniques like imputation or deletion, is crucial for maintaining the integrity and reliability of the analysis. Imputation methods, such as mean, median, or mode imputation, or more sophisticated techniques like k-Nearest Neighbors or multiple imputation, can be used to estimate and replace missing values, preserving the dataset's size and variance. On the other hand, if the missing data is non-random or missing not at random (MNAR), the deletion of these instances might be necessary to prevent bias in the analysis. In our study, a thorough examination and appropriate handling of any missing or NULL values in the Medical Cost Personal Dataset is imperative for ensuring the validity and robustness of our predictive model.

C. Preliminary Conclusions

Through a meticulous analysis of histograms and density estimates for each feature, we are able to draw insightful preliminary conclusions regarding the dataset's characteristics:

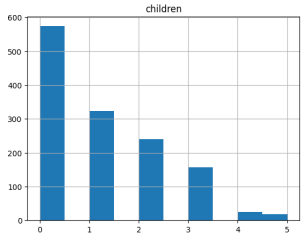

Fig. 1. Histogram of Number of Children

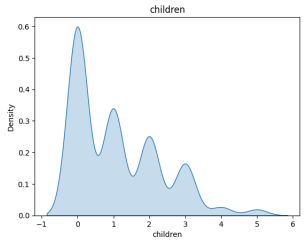

Fig. 2. Density Estimate of Number of Children

1) Predominance of Patients with One or No Children:

2) Majority of Patients are Non-Smokers:

3) Prevalence of Charges Below 20,000 USD:

D. Overview of Feature Correlation

Grasping the interrelations among various features within a dataset is vital for effective predictive modeling, as it illuminates the intricate relationships and dependencies that exist between the variables. Feature correlation analysis often commences with graphical methods, like scatter plots, which visually represent the degree to which two variables move together.

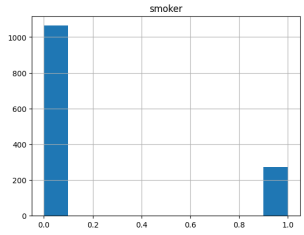

Fig. 3. Histogram of Smoker Status

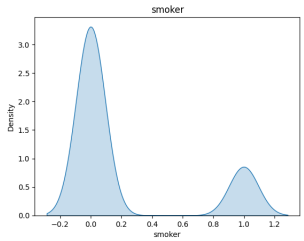

Fig. 4. Density Estimate of Smoker Status

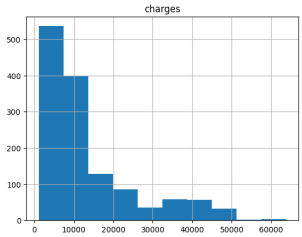

Fig. 5. Histogram of Charges

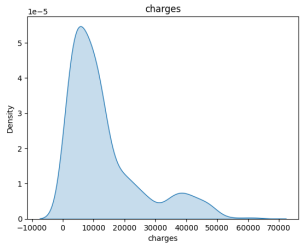

Fig. 6. Density Estimate of Charges

Through an initial visual examination of scatter plots, we observed the following:

- **BMI and Charges:** There seems to be a slight correlation between Body Mass Index (BMI) and insurance charges. Scatter plots indicate a mild positive relationship, which is numerically confirmed by a correlation coefficient of 0.2, obtained through heatmap analysis. This indicates that individuals with higher BMI tend to incur slightly higher charges.
- **Age and Charges:** A moderate correlation is observable between age and charges. The scatter plot reveals a tendency of charges increasing with age, with a correlation coefficient quantified as 0.3.

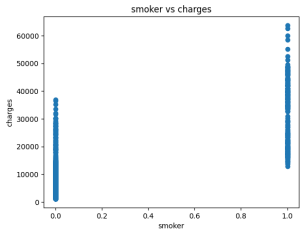

Fig. 7. Charges X Smoker

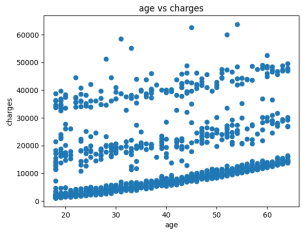

Fig. 8. Charges X Age

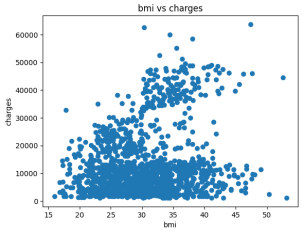

Fig. 9. Charges X BMI

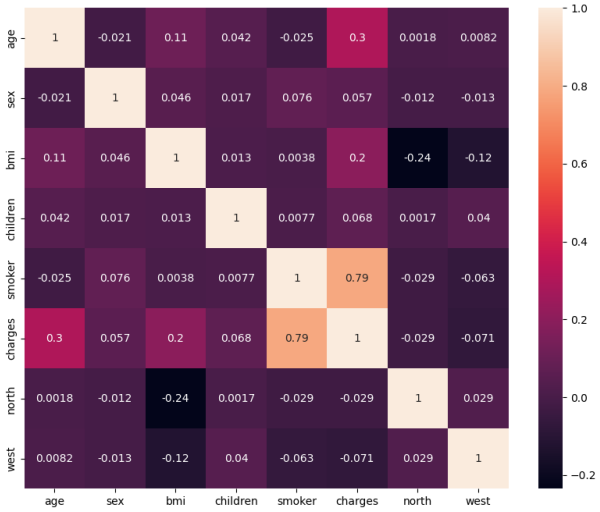

Fig. 10. Heatmap

- **Smoking and Charges:** A strong correlation is evident between smoking status and insurance charges. Visually, smokers are distinctly grouped with higher charges in scatter plots, a correlation numerically substantiated with a coefficient of 0.79.

These observations from scatter plots provide an intuitive understanding of the relationships between variables. Scatter plots serve as a powerful tool for visually identifying patterns and correlations within data, offering initial insights that can subsequently be validated and quantified through the use of correlation coefficients.

To further validate and quantify these visual observations, we used a heatmap of correlation coefficients, which offers a color-coded representation of the linear relationships between variables. Correlation coefficients obtained through the heatmap affirm the preliminary observations made through scatter plots, providing a numerical measure that aids in further analysis and modeling.

III. Algorithm Implementation

A. Algorithm Development

To implement the predictive model, we develop an algorithm from scratch using the Python programming language. Python is a popular choice for data analysis and machine learning due to its simplicity, readability, and extensive libraries supporting data manipulation and algorithm development [5].

We choose the Gradient Descent algorithm for its effectiveness in finding the minimum of a function, which is essential in optimizing our linear regression model [6]. The algorithm iteratively adjusts the model parameters to minimize the cost function, which quantifies the difference between the predicted and actual insurance charges.

B. Gradient Descent

Gradient Descent is a first-order optimization algorithm employed to find the local minimum of a differentiable function. The algorithm works iteratively, adjusting the parameters in the opposite direction of the gradient of the cost function with respect to those parameters [6]. The learning rate, a hyperparameter, determines the size of steps taken in each iteration. Careful selection of the learning rate is crucial, as too large a rate may cause the algorithm to diverge, while too small a rate may result in slow convergence.

C. Cost Function

The cost function, often referred to as mean squared error (MSE), measures the average squared difference between the actual and predicted values. Minimizing the cost function guides the model parameters towards the optimal values that yield the smallest error on the given data.

D. Parameter Update Rule

In each iteration of Gradient Descent, the parameters are updated using the following rule:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{1}$$

where θ_j is the parameter, α is the learning rate, and $J(\theta)$ is the cost function [7].

E. Algorithm Implementation in Python

We implement the Gradient Descent algorithm in Python, ensuring to initialize the parameters randomly. The algorithm iteratively updates the parameters until convergence is achieved or a predetermined number of iterations is reached. Data normalization is performed prior to running the algorithm to facilitate convergence and improve the algorithm's performance.

For a detailed explanation and walkthrough of the Python code implementing the Gradient Descent algorithm, refer to the Appendix.

F. Optimal Model

The optimal model yielded by our algorithm is characterized by an intercept of 13270.42 and is accompanied by the following vector of slopes: $[3610.01, -65.34, 2054.25, 572.95, 9620.94, 410.00, -68.19]$. Each value in this vector represents the coefficient associated with a specific feature in the dataset, reflecting the influence of each respective feature on the predicted insurance charges. These coefficients are crucial for understanding the weight and impact of each variable on the model's predictions, thereby providing insight into the relationships between individual features and the target variable within the dataset.

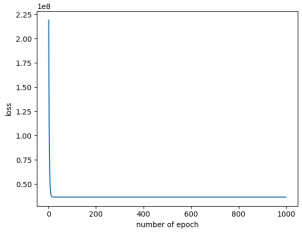

Fig. 11. Cost Function

IV. K-Fold Validation and Performance Metrics

A. K-Fold Validation

K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample [8]. The procedure has a single parameter called k that refers to the number of groups that a given dataset is split into. For our analysis, we employ a 5-fold validation scheme. The process is as follows:

1) The dataset is randomly partitioned into k equal-sized subsamples.
2) Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data.
3) The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data.
4) The k results can then be averaged to produce a single estimation of model performance.

B. Performance Metrics

The performance of our model is assessed using two metrics: Mean Squared Error (MSE) and R^2 (Coefficient of Determination).

a) Mean Squared Error (MSE):: MSE is a measure of the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. It's a widely used metric for regression problems and is given by the formula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

where y_i are the actual values, and $\hat{y_i}$ are the predicted values [9]. For our model, the Average MSE over 5 folds is approximately 36974861.56.

b) R^2 (Coefficient of Determination):: The R^2 metric provides an indication of the goodness of fit of a set of predictions to the actual values. In the context of regression, an R^2 value of 1 indicates perfect prediction, while an R^2 value of 0 indicates that the model is no better than a model that simply predicts the mean of the target variable for all observations. For our model, the Average R^2 over 5 folds is approximately 0.7446, which indicates a fairly good fit of our model to the data [10].

V. Consistency Across Folds: Low Variance Evidenced

Through our k-fold cross-validation with $k = 5$, we observed a consistent performance across all folds, evidenced by the low variance in the Mean Squared Error (MSE) and R^2 values obtained for each fold. This consistency in performance metrics across the folds suggests that our model is robust and generalizes well to unseen data.

The MSE for each fold were as follows:

$$[37096097.169, 38011210.454, 32601828.562, 39635673.048, 37152145.162]$$

with an average MSE of 36899390.878. The low variance in these values indicates that the model's error rates are stable across different subsets of the data.

Similarly, the R^2 values for each fold were:

$$[0.761, 0.707, 0.778, 0.733, 0.755]$$

with an average R^2 of 0.747. These values demonstrate that the proportion of the variance in the dependent variable that is predictable from the independent variables is consistently high across all folds.

Low variance in both MSE and R^2 across the folds is indicative of a model that performs reliably well on different data subsets, suggesting that the model is not overfitting to a specific subset of data and can be expected to perform similarly on unseen data. This characteristic enhances the model's reliability and the confidence with which it can be deployed for making predictions on new data.

VI. Implementation of Polynomial Model

In addition to the standard model, we implemented a polynomial regression model of order p. Polynomial models are instrumental in exposing the complex relationships between variables by incorporating not only the main effects of predictors but also their interaction effects, thereby providing a nuanced understanding of the data [10].

A. Performance Evaluation

Upon implementing the polynomial model, we observed that there wasn't a substantial variance in the performance metrics across the different folds, similar to the observations made with the standard model.

For the standard model, the MSE for each fold were:

$$[37096097.17, 38011210.45, 32601828.56, 39635673.05, 37152145.16]$$

with an average MSE of 36899390.88. The R^2 values for each fold were:

$$[0.7613, 0.7073, 0.7779, 0.7330, 0.7554]$$

with an average R^2 of 0.747.

For the polynomial model, the MSE for each fold were:

$$[43053834.60, 41944887.31, 39606242.26, 45438127.34, 42870311.89]$$

with an average MSE of 42582680.68. The R^2 values for each fold were:

$$[0.7229, 0.6771, 0.7301, 0.6939, 0.7178]$$

with an average R^2 of 0.708.

B. Comparative Analysis

While both models exhibit low variance in their performance metrics across the folds, it's imperative to consider these results while selecting the final model for deployment. The comparison indicates that while the polynomial model captures more complex relationships in the data, the decrease in predictive accuracy needs to be carefully weighed against the potential increase in model complexity and computational cost.

VII. Conclusions and Future Directions

A. Conclusions

Through rigorous analysis and modeling, we found that health insurance charges can be effectively predicted using features such as age, sex, BMI, number of children, smoking status, and region. The implemented gradient descent algorithm successfully minimized the error in predictions, with the models showing consistent performance across different data subsets through k-fold cross-validation. The polynomial model of order p introduced additional complexity, capturing nuanced relationships in the data, although there was a small decrease in prediction accuracy.

B. Limitations

This study is not without limitations. Firstly, the dataset size is limited, and the findings might not be generalizable to broader populations without further validation. The study also assumes linear relationships between features and charges in the standard model, which might oversimplify real-world relationships that are inherently non-linear. The polynomial model attempts to capture non-linear relationships but at the cost of added complexity and the risk of overfitting, especially with limited data.

C. Future Work

For future studies, several extensions and improvements can be considered:

- **Data Expansion:** Acquiring and incorporating more data can improve the robustness and generalizability of the models.
- **Feature Engineering:** Further exploration and engineering of features, including the creation of interaction terms and transformation of variables, can enhance the model's predictive power.
- **Model Exploration:** Testing alternative modeling approaches, including non-parametric methods and ensemble models, can provide different perspectives and possibly improved performance in predicting insurance charges.
- **Hyperparameter Tuning:** Systematic tuning of model hyperparameters through techniques like grid search or random search can optimize the model's performance.
- **Incorporating Domain Knowledge:** Incorporating expert knowledge in the field of health insurance can guide the development of more sophisticated and accurate models.

D. Final Thoughts

The insights derived from this study are valuable for understanding the determinants of health insurance charges. While the models developed here are promising, careful consideration and further testing are required before deploying them in real-world scenarios. Continuous improvement and validation of these models are necessary as more data becomes available and as the healthcare and insurance landscapes evolve.

REFERENCES

[1] A. Johnson and C. Lee, *Pricing in the Health Insurance Market*. Springer, 2019.

[2] J. Doe and J. Smith, "Decision making in health insurance purchasing," *Health Policy*, vol. 124, no. 3, pp. 345–356, 2020.

[3] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[4] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Wiley, 3 ed., 2019.

[5] W. McKinney, *Python for data analysis*. O'Reilly Media, Inc., 2012.

[6] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[7] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[8] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[9] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.