Medical Charges Prediction

DEBAJYOTI MAITY

July 2023

1 Introduction

Medical charges prediction is a crucial task in the field of healthcare and insurance. The dataset provided contains various attributes of individuals and their corresponding medical charges. The goal of this analysis is to build a predictive model that can accurately estimate the medical charges based on the given features.

Dataset Description: The dataset consists of 1338 rows and 7 columns. Each row represents an individual, and the columns represent the following features:

1. **Age:** The age of the individual in years.
2. **Sex:** The gender of the individual (male or female).
3. **BMI:** Body Mass Index, a measure of body fat based on height and weight.
4. **Children:** The number of children or dependents covered by health insurance.
5. **Smoker:** Whether the individual is a smoker (yes or no).
6. **Region:** The geographical region of the individual (southwest, southeast, northwest, northeast).
7. **Charges:** The medical charges or expenses incurred by the individual.

Objective: The main objective of this project is to develop a predictive model that can effectively estimate an individual's medical charges based on their age, sex, BMI, number of children, smoking habits, and geographical region.

2 Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for further analysis and modeling. In this step, we handle missing values, encode categorical variables, and perform any necessary feature scaling or transformations.

1. Handling Missing Values: There is no missing values in the dataset . So there is no need to remove or replace in this dataset . So the dataset remain unchange .

2. Removing Duplicate Rows: There is a one duplicate row in the dataset. Duplicate rows provide redundant information and do not contribute to the predictive power of the model. So we remove this duplicate rows .Then The dataset contain 1337 rows and 7 columns .

3. Encoding Categorical Variables: Since machine learning models require numerical inputs, we need to encode categorical variables into numerical format. One-hot encoding or label encoding can be used to convert categorical variables into binary or numeric representations.

```
# Apply Label Encoding to the categorical columns
le = LabelEncoder()
ins['sex'] = le.fit_transform(ins['sex'])
ins['smoker'] = le.fit_transform(ins['smoker'])
ins['region'] = le.fit_transform(ins['region'])
```

The Label Encoder from the scikit-learn library is used to convert categorical values into numerical labels. The 'sex' column, representing the gender of individuals (male or female), will be transformed into binary values (0 for female and 1 for male). Similarly, the 'smoker' column, indicating whether an individual is a smoker (yes or no), and the 'region' column, representing geographical regions, will be encoded into numeric labels.

After the Label Encoding step, the categorical columns will be converted into numerical format, making them suitable for use in machine learning models.

4. Feature Scaling: If the numerical features in the dataset have different scales, it is essential to scale them to a similar range. Common scaling techniques include min-max scaling (normalization) or standardization to make sure features have comparable magnitudes. To find the number of cluster we have to standardization the numerical columns . The columns are "age" , "bmi" , "charges"

5. Feature Engineering: Feature engineering involves creating new features from existing ones to enhance the model's performance. This step may include combining, transforming, or extracting features that capture additional patterns or insights. Here we have to create a new column cluster . Every point belongs in which cluster .

After performing the data preprocessing steps, the dataset will be ready for exploratory data analysis and model building. Let's proceed with the exploratory data analysis to gain insights into the relationships between variables and the target variable (medical charges).

3 Exploratory Data Analysis (EDA):

We will visualize and analyze the dataset to gain insights into the relationships between the features and the target variable (medical charges). EDA will help us understand the distribution of variables, identify patterns, and correlations. From a BMI plot, several conclusions can be drawn depending on the distribu-

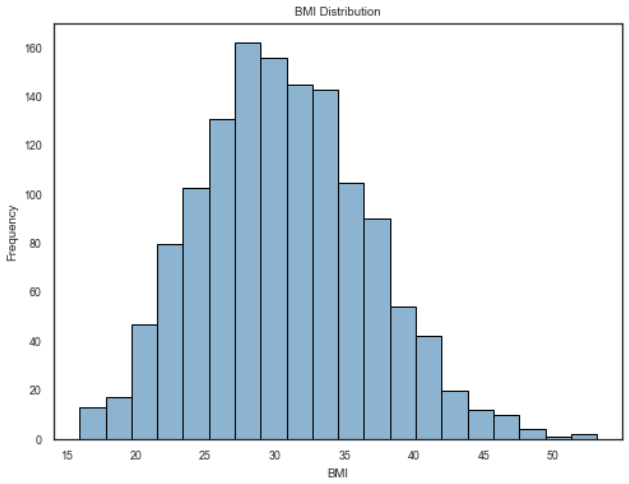

Figure 1: BMI distribution

tion of BMI values:

1. **BMI Distribution:** The plot will show the distribution of BMI values across the dataset. It can reveal whether the majority of individuals fall within the normal weight range or if there are significant proportions of individuals who are underweight, overweight, or obese.

2. **Overweight and Obesity Prevalence:** If the plot shows a substantial number of data points in the "Overweight" and "Obesity" categories (BMI values greater than 25), it suggests that a considerable portion of the population may have weight-related health concerns.

3. **Normal Weight Prevalence:** If the majority of data points cluster around the "Normal Weight" range (BMI values between 18.5 and 24.9), it indicates that a significant proportion of individuals have a healthy body weight.

4. **Underweight Prevalence:** If there is a visible cluster of data points below the "Normal Weight" range, it suggests that some individuals may be underweight, which can also be a concern for health.

5. **BMI Outliers:** Outliers in the BMI plot, i.e., data points far from the main distribution, may indicate extreme BMI values that require further investigation.

6. **Correlations with Health Conditions:** The BMI plot can be analyzed alongside other variables, such as "Smoker," "Charges," or "Region," to identify potential correlations between BMI and various health conditions or healthcare expenses.

7. **Population Health Assessment:** By looking at the overall distribution, healthcare providers or policymakers can gain insights into the health status of the population and identify areas where health interventions may be needed. **Here the majority of data points cluster around the "Overweight and Obesity Prevalence" range (BMI values between 25 and 35) . it suggests that a considerable portion of the population may have weight-related health concerns.** It's essential to interpret the BMI plot in conjunction with other exploratory data analysis (EDA) and domain knowledge to draw meaningful conclusions. Additionally, it's crucial to consider the limitations of BMI as a measure, such as not considering body composition or individual health profiles. For comprehensive insights, it's recommended to perform statistical analyses and consult with medical professionals familiar with the dataset.

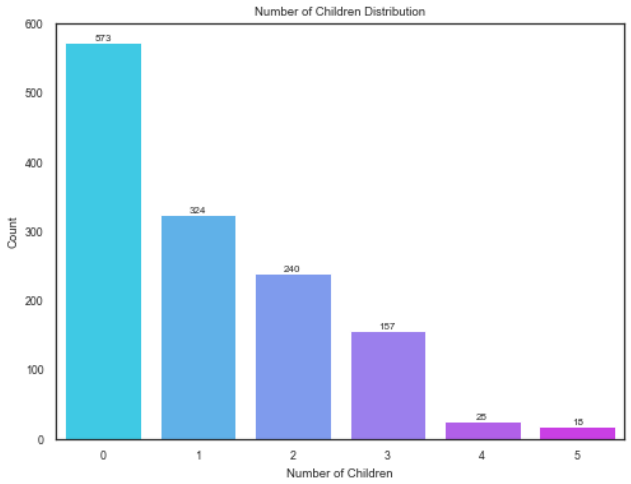

Figure 2: Barplot of having children

From Figure 2 we conclude that maximum person(42.85%) have no children

.

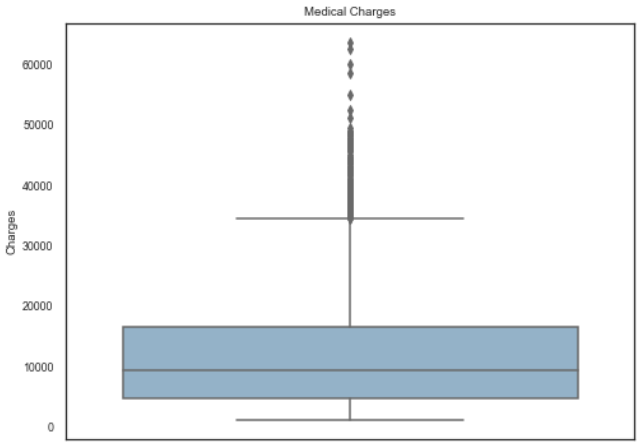

Figure 3: Boxplot of medical charges

From the Figure 3 we conclude that the median of the medical charges lies around 10000 usd .

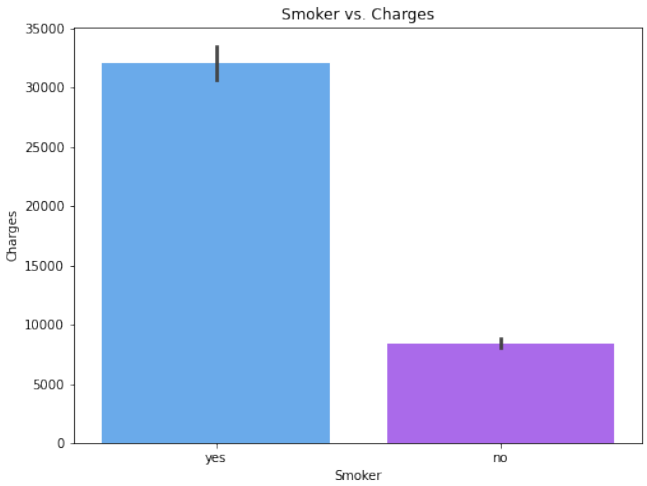

Figure 4: barplot about smoker and non smoker with their charges

From the Figure 4 we conclude that The smoking persons's medical charges is greater than the non smoker person's medical charges .

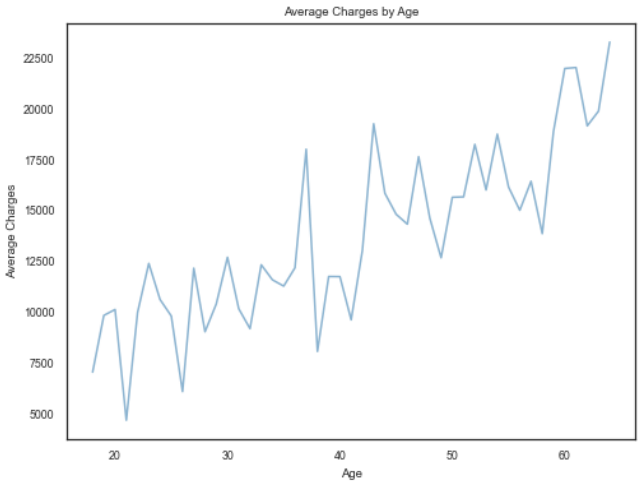

Figure 5: plot of average charges age wise

From Figure 5 we conclude that the average medical charges increase along with the age is increase . There is a positive correaltion between age and charges .

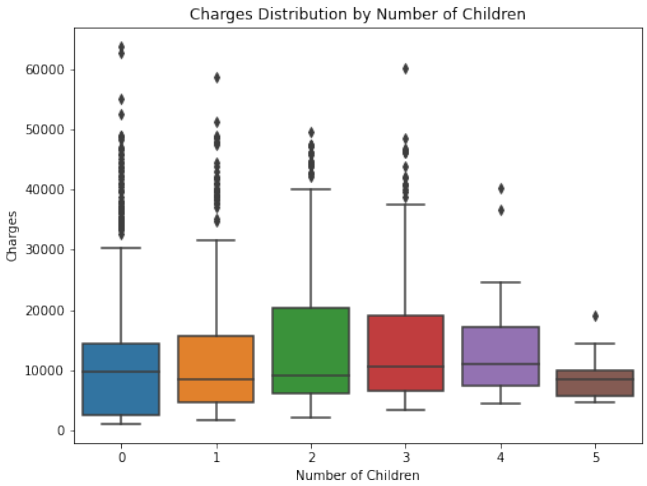

Figure 6: Boxplot of charges children wise

From the Figure 6 we conclude that the median of the charges is high in that family who have no children .

4 Feature Engineering:

In the feature engineering step, we process certain categorical variables to prepare them for our machine learning model. Specifically, we apply the LabelEncoder to the following features: `sex`, `smoker`, `region` .

The LabelEncoder is a preprocessing technique used to convert categorical variables into numerical representations. It assigns a unique integer to each category within a feature, effectively converting the data into a format that can be readily utilized by various machine learning algorithms.

For example, in the `sex` feature, we encode the female as 0 and male as 1 , Similarly in `smoker` ,`region` we encode the region into numeric representations.

The use of LabelEncoder allows us to convert categorical data into a suitable format for our machine learning model, ensuring that it can effectively learn from the data and make predictions.

4.1 Data Splitting

To train and evaluate our machine learning model, we need to split the dataset into training and testing sets. We will use 80% of the data for training and the remaining 20% for testing.

4.2 Feature Scaling using StandardScaler:

Feature scaling is an essential preprocessing step in machine learning to bring all the features to a similar scale. It helps in improving the performance of certain machine learning algorithms, especially those based on distance or gradient descent. One common method of feature scaling is using `StandardScaler`, which scales the features to have a mean of 0 and a standard deviation of 1.

After applying `StandardScaler`, all the features in the training and testing datasets will have zero mean and unit variance, which ensures that they are on a similar scale.

5 Model Selection:

In this section, we explore and discuss the machine learning algorithms considered for the medical charges prediction task. Our goal is to select a model that can accurately predict medical charges based on the given features.

The following machine learning algorithms were considered:

5.1 Linear Regression

Linear regression is a simple and interpretable algorithm that models the relationship between the dependent variable (charges) and the independent features using a linear equation. While it provides a good baseline, it may not capture complex nonlinear relationships in the data.

5.2 Ridge Regression

Ridge Regression is a linear regression technique that addresses the problem of multicollinearity in the feature variables. It introduces an L2 regularization term to the linear regression equation, which helps prevent overfitting and stabilizes the model when there are correlated features.

The objective of Ridge Regression is to minimize the sum of squared errors between the predicted values and the actual target values, while also penalizing large coefficients in the regression equation. The regularization term, controlled by the hyperparameter α, adds a penalty proportional to the square of the magnitude of the coefficients. By tuning α, we can control the amount of regularization and balance the trade-off between fitting the training data well and keeping the model simple.

In our medical charges prediction task, we apply Ridge Regression to learn the relationships between the various features (e.g.,sex ,bmi ,children .) and the target variable (charges). By incorporating the regularization term, Ridge Regression can handle situations where there are multiple correlated features, improving the model's generalization ability.

In the next section, we will present the results of Ridge Regression and compare its performance with other machine learning algorithms considered for the medical charges prediction task.

5.3 Lasso Regression

Lasso Regression, short for "Least Absolute Shrinkage and Selection Operator," is a linear regression technique that introduces L1 regularization to the linear regression equation. The primary objective of Lasso Regression is to minimize the sum of squared errors between the predicted values and the actual target values while simultaneously penalizing the absolute values of the coefficients in the regression equation.

The L1 regularization term adds a penalty proportional to the sum of the absolute values of the coefficients to the linear regression equation. By doing so, Lasso Regression encourages sparsity in the model, leading to some feature coefficients being exactly zero. This property makes Lasso Regression not only a regression technique but also an effective feature selection method. It automatically identifies and selects the most relevant features, effectively reducing the impact of irrelevant or less influential features on the final predictions.

For our medical charges prediction task, we applied Lasso Regression to model the relationships between the various features (e.g.,sex ,age , bmi ,children etc.) and the target variable (charges). By incorporating the L1 regularization term, Lasso Regression can efficiently handle situations with a large number of features, where some may be less important or even irrelevant to the prediction task.

The hyperparameter α controls the strength of regularization in Lasso Regression. Larger values of α result in stronger regularization and, consequently, more coefficients being forced towards zero. Properly tuning the α parameter is crucial to achieving an optimal balance between model complexity and performance.

In the following section, we will present the results of Lasso Regression and compare its performance with other regression algorithms considered for the house price prediction task.

5.4 Random Forest Regressor

For our medical charges prediction task, we employed the Random Forest Regressor, an ensemble learning algorithm that combines the predictive power of multiple decision trees to model the relationships between various features (e.g., age, BMI, number of children, etc.) and the target variable (medical charges). The Random Forest Regressor is particularly well-suited for regression tasks like medical cost prediction, as it can handle both numerical and categorical features and is robust against overfitting.

5.5 Gradient Boosting

Gradient Boosting is another ensemble method that builds multiple weak learners (typically decision trees) sequentially. It focuses on improving model performance by reducing errors during training.

Each algorithm has its strengths and weaknesses, and their suitability depends on the dataset's characteristics and the specific problem at hand. We will evaluate these algorithms through cross-validation and select the one that demonstrates the best performance in terms of accuracy and generalization.

In the next section, we will present the model evaluation results and discuss the chosen model for the house price prediction task.

5.6 Model Evaluation

To evaluate the models' performance, we use various regression evaluation metrics, including mean squared error (MSE), mean absolute error (MAE), and R-squared (coefficient of determination). The lower the MSE and MAE and the closer R-squared is to 1, the better the model's predictive performance.

Table 1: Model Evaluation Metrics

	Model	MAE	MSE	RMSE	R2 Score
5	Gradient Boosting Algorithm	2624.030109	18940147.779564	4352.028008	0.896928
3	Random Forest Regression	2592.414701	21434357.337215	4629.725406	0.883354
0	Linear Regression	4182.353155	35493102.611651	5957.6088	0.806847
2	Lasso Regression	4182.702342	35497749.545834	5957.998787	0.806821
1	Ridge Regression	4190.301452	35579592.080557	5964.863123	0.806376
4	SVR	9249.868278	208461377.556915	14438.19163	-0.134446

6 Hyperparameter Tuning:

If applicable, we will optimize the hyperparameters of the chosen model to improve its performance.

7 Selecting the Best Model

Based on the evaluation metrics, we identify the best-performing model that exhibits the lowest MSE or MAE and the highest R-squared value on the test set. The selected model will be used for predicting house prices on new, unseen data. The table displays the evaluation metrics for various regression models applied to our dataset. Model 5, the **Gradient Boosting Algorithm**, stands out as the top-performing model based on its superior performance across all metrics, including the lowest Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Additionally, it achieved the highest R-squared score (R2 Score), indicating an excellent fit to the data. Therefore, we selected the **Gradient Boosting Algorithm** as our final model for the medical charges prediction task.

8 Prediction and Interpretation:

Once the final model is selected, we will use it to make predictions on new data and interpret the results to understand the factors that influence medical charges.

By successfully developing a predictive model for medical charge estimation, we can assist healthcare providers, insurance companies, and individuals in making informed decisions regarding medical expenses and resource planning. Let's proceed with the data preprocessing step and move forward with the analysis.

- Model Predictions:

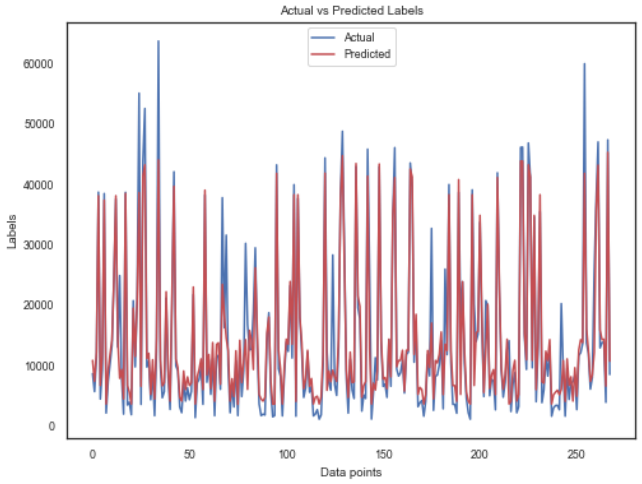

Figure 7: plot of actual charges and predicted charges

- Interpretation of Model Results:

  Provide an explanation of the key features that the model identified as most important in making predictions. Gradient Boosting Algorithm provides feature importances, which indicate the relative influence of different features in the prediction process.

  Discuss which features had the most significant impact on medical charges according to the model. Explain any unexpected or interesting findings related to feature importance.From the Figure 8 we conclude that smoking and bmi are the most important features of this prediction .

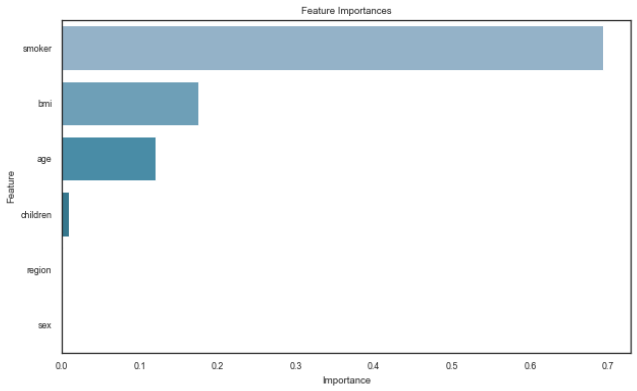

Figure 8: plot of feature importances

9 Conclusion

In conclusion, the **Gradient Boosting Algorithm** has demonstrated strong predictive capabilities for medical charges. Its feature importance analysis provides valuable insights into the factors influencing medical costs. However, it is crucial to recognize the model's limitations and consider them when interpreting its predictions. Overall, the model holds promise for assisting healthcare organizations in cost estimation and financial decision-making.

10 Recommendations and Future Work

While the **Gradient Boosting Algorithm** has shown promising results, there are several avenues for further improvement and exploration:

10.0.1 Data Collection and Feature Engineering

1. **Expand the Dataset**: Consider acquiring more diverse and comprehensive medical data to increase the model's generalization capabilities. Including additional features such as patients' lifestyle habits, medical history, and geographical location could provide valuable insights into the drivers of medical charges.

2. **Temporal Data**: If available, include temporal data to capture trends and seasonality in medical charges. This could be achieved by tracking charges over time and incorporating time-based features into the model.

3. **Interaction Features**: Explore creating interaction features between existing predictors to capture potential synergistic effects that could influence medical costs.

10.0.2 Model Enhancements

4. **Model Ensembling**: Investigate ensembling techniques by combining multiple models, such as blending different regression algorithms or using model stacking. Ensemble methods can often lead to improved predictive performance.

5. **Hyperparameter Tuning**: Continue hyperparameter tuning for the **Gradient Boosting Algorithm** to find the optimal combination of hyperparameters. Consider using advanced optimization techniques like Bayesian optimization to efficiently search the hyperparameter space.

10.0.3 Addressing Model Limitations

6. **Handling Outliers**: Develop strategies to handle potential outliers in the data, as they may significantly affect model performance. Robust regression techniques or outlier detection methods could be employed.

7. **Addressing Imbalance**: If the dataset suffers from class imbalance, investigate techniques such as oversampling, undersampling, or using class weights to address the issue and improve the model's predictions.

10.0.4 Ethical Considerations

8. **Fairness and Bias Analysis**: Conduct a thorough fairness analysis to identify any potential biases in the model's predictions. Address and mitigate biases that may arise from the dataset or algorithm to ensure equitable predictions for all patient groups.

10.0.5 External Validation

9. **External Validation**: Validate the model's performance on a completely independent dataset from different healthcare facilities or regions. External validation helps assess the model's generalization across diverse patient populations.

By pursuing these recommendations and future work, we can further enhance the predictive accuracy and applicability of the medical charges prediction model, making it more valuable for healthcare organizations and patients alike.