Medical Insurance Costs Forecast using Linear Regression

Lucius Filho 96123
DETI
University of Aveiro
Aveiro, Portugal
luciusviniciusf@ua.pt

Tomé Carvalho 97939
DETI
University of Aveiro
Aveiro, Portugal
tomecarvalho@ua.pt

***Abstract*—The chosen dataset was adapted from Machine Learning with R by Brett Lantz [1]. The health care system in the United States is largely privatized, in contrast with most European countries, which largely feature universal health care systems. Consequently, American citizens often have to take into account insurance costs. Our main goal in this project is to explore which factors influence insurance costs for patients, by performing data analysis to search for patterns and interpreting them, and to fit a regression model to estimate insurance charges based on the factors.**

***Index Terms*—health care, smoking, bmi, obesity, age, region, children, machine learning, regression, data analysis, linear, random forest, ridge, lasso, polynomial**

I. Introduction

In 2019, the U.S. spent about $11,100 per person on health care — the highest health care cost per capita across the OECD. For comparison, Switzerland was the second highest-spending country with about $7,700 in health care expenses per capita, while the average for wealthy OECD countries, excluding the U.S., was only $5,500. The U.S. thus spends a disproportionate amount on health care. [2]

Given the high health care costs and private-centered systems, U.S. citizens must carefully consider the costs of medical insurance, which implies being aware of the factors that lead to increases in expenses. Our goal is to shed a light on the impact of those present in the dataset and fit a model to be able to predict the costs.

II. Dataset

A. Dataset description

The dataset used is available on Kaggle [3], as well as GitHub [4]. It comprises 1338 rows of data, with the following columns:

- age: Age of the insured
- sex: Biological sex of the insured
- bmi: Body mass index (BMI) of the insured
- children: Number of children/dependents covered by health insurance
- smoker: Whether the insured smokes tobacco
- region: Insured's residential area in the US (Northeast, Southeast, Southwest, Northwest)
- charges: Individual medical costs billed by health insurance (in U.S. dollars)

B. Preprocessing

Initially, the data was not suitable for analysis, as there were features with string values: *sex*, *smoker* and *region*.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Fig. 1. Original data format

To overcome this issue, we applied label encoding.

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520

Fig. 2. Encoded data format

Additionally, we tested the use of One Hot Encoding, for the "region" feature. However, as the results we obtained were marginally worse, it was not kept in the final version.

C. Feature distribution

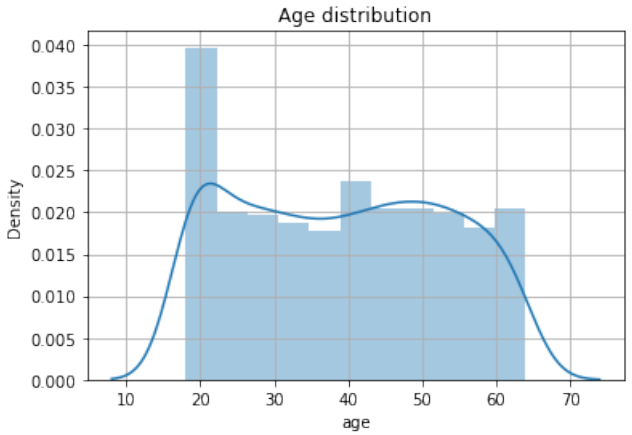

Fig. 3. Age distribution

Individuals around 20 years old have a higher prevalence in our dataset than other age groups. From that range to 64, the maximum, the distribution is more even.

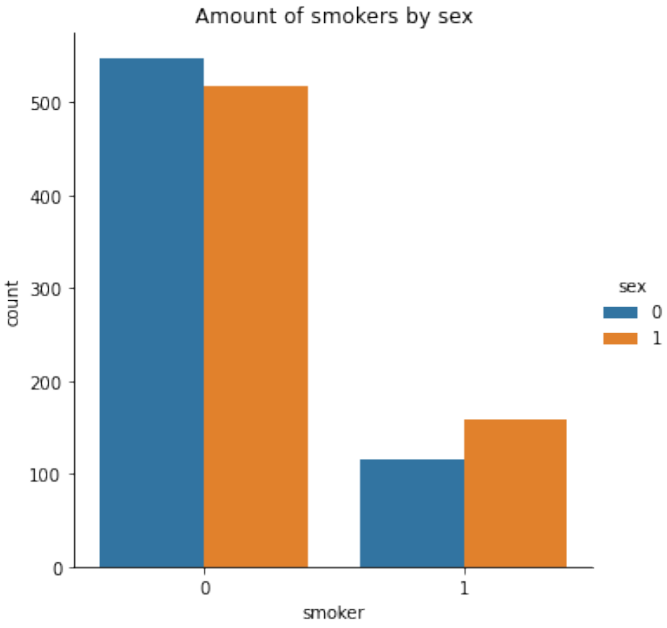

Fig. 4. Amount of smokers by sex (0: female, 1: male)

The amount of non-smokers is significantly more than that of smokers. There is a greater amount of male smokers than female smokers. This may lead to males having higher insurance costs overall, since smoking greatly increases charges, as we will verify later.

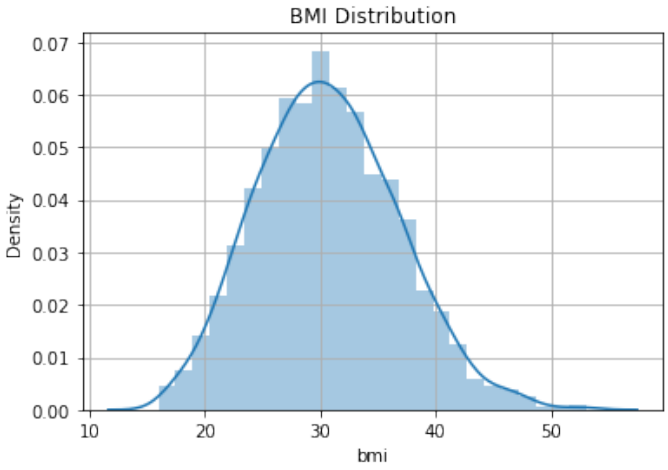

Fig. 5. BMI distribution

The average BMI of individuals in our dataset is around 30kg/m^2. Major adult BMI classifications are underweight (under 18.5), normal weight (18.5 to 24.9), overweight (25 to 29.9), and obese (30 or more). This means roughly half of our dataset is obese. In constrast, very few are underweight.

D. Correlations

Before investigating any further, the possibility of correlation between different features must be investigated.

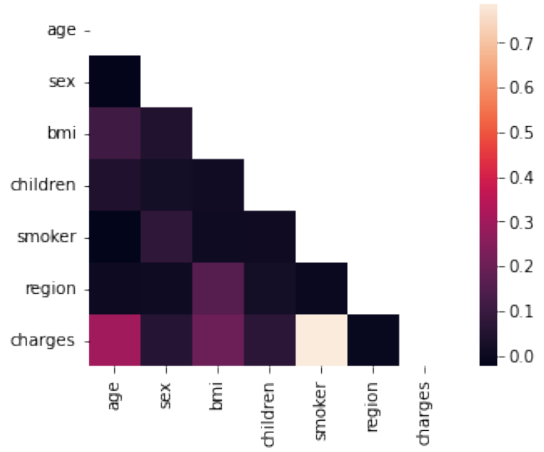

Fig. 6. Graph of correlations between different features and charges.

```
region      -0.006208
sex          0.057292
children     0.067998
bmi          0.198341
age          0.299008
smoker       0.787251
charges      1.000000
Name: charges, dtype: float64
```

Fig. 7. Correlation values between different features and charges.

Through the analysis of Figs. 6 and 7, it is inferred that ***smoker* is the feature most responsible** for affecting the costs. *bmi* and *age* are important as well.

An interesting detail is the correlation between *region* and *bmi*.

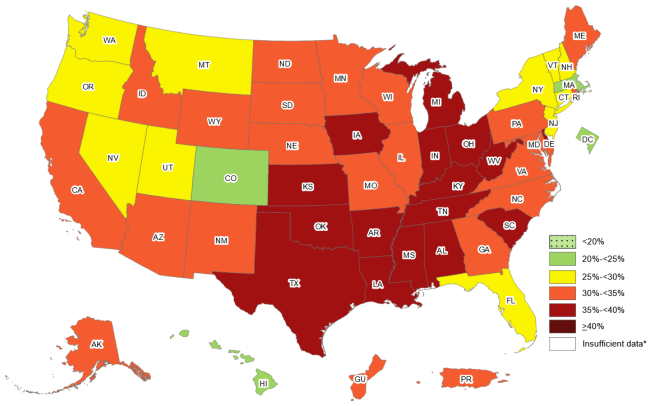

Fig. 8. Prevalence of obesity in the U.S. (CDC, 2020)

According to Fig. 8, there is a higher prevalence of obesity among the Southeastern states. This is a plausible explanation for the aforementioned correlation. Through our experiment with OHE, we found this to be correct. Individuals from the Southeast did, in fact, have a higher BMI on average. The region itself is not significantly correlated with insurance charges.

We can also conclude that *children* is rather unimportant.

E. Analysis of the impact of smoking

Firstly, to demonstrate the smoking consequence on the costs, it is necessary to analyze Fig. 9 below:

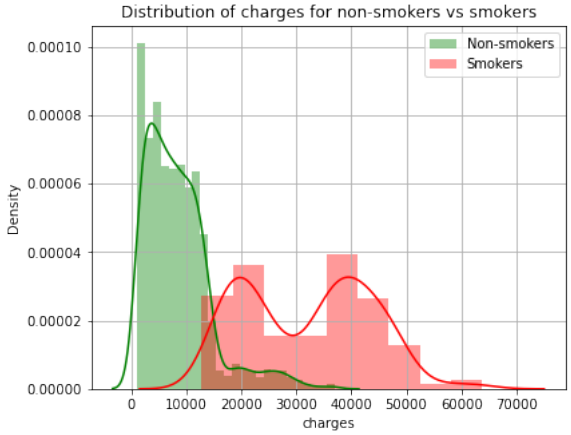

Fig. 9. Charges distribution between smokers and non-smokers

In the graph above, the upper range of charges for **non-smokers** overlaps with the lower range for **smokers**, evidencing that smoking leads to a tremendous increase in insurance costs.

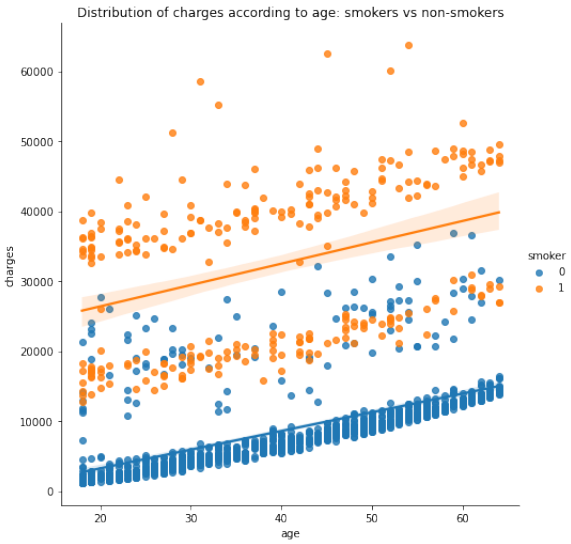

Fig. 10. Charges according to age: smokers (1) vs non-smokers (0)

Fig 10 evidences that, independently of age, smokers are usually charged considerably more than non-smokers. Furthermore, there seem to be two groups of smokers, one with even higher costs. The reason for this will be made clear in the next subsection, II-F.

F. Analysis of the impact of obesity

The following graphs only separate ranges into obese and not obese, as we assessed that the values for overweight and underweight were close to those of normal weight.

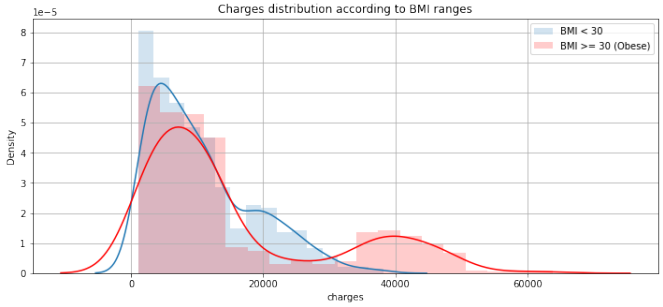

Fig. 11. Charges distribution according to obesity

From Fig. 11, it can be concluded that, while a good amount of obese individuals have close charges to non-obese ones, there is a significant amount of the former whose costs are higher than nearly all of non-obese individuals', even exceeding $40,000.

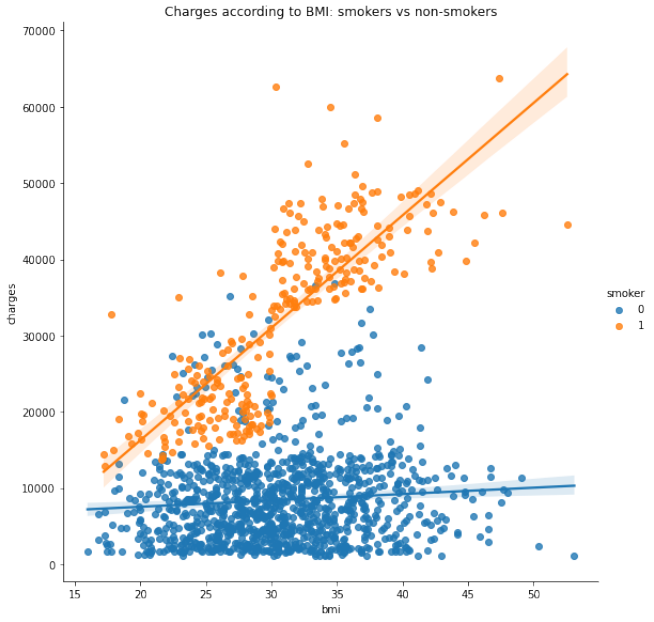

Fig. 12. Charges according to BMI: smokers vs non-smokers

Fig. 12 shows that, while costs already tend to rise as BMI increases past obesity, the issue is exacerbated when combined with smoking, as there is a dramatic increase in charges for smokers after the obesity threshold.

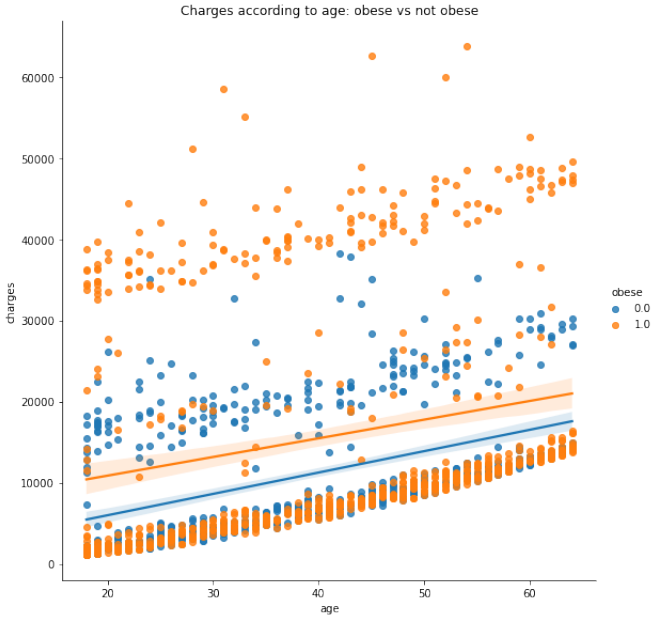

Fig. 13. Charges according to age: obese (1) vs not obese (0)

Fig. 13 evidences that obesity may lead to considerable increases in insurance costs even at younger ages.

III. Regression Model

A. Types of regression

Since the goal is to predict the charges value, the use of a Regression Model is necessary. We applied the following types of regression:

- **Linear Regression** [5]: fits a linear model with coefficients w = (w1, . . . , wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.
- **Random Forest Regression** [6]: a random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- **Ridge Regression** [7]: linear least squares with l2 regularization; minimizes the objective function: $||y - X_w||_2^2 + alpha * ||w||_2^2$; solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm.
- **Lasso Regression** [8]: linear Model trained with L1 prior as regularizer (aka the Lasso); its optimization objective is $(1/(2 * n_{samples})) * ||y - X_w||_2^2 + alpha * ||w||_1$.
- **Polynomial Regression** [9]: a new feature matrix consisting of all polynomial combinations of the features with degree less than or equal to the specified degree is generated; linear regression is applied on the new matrix.

B. Cross-validation

Since our dataset did not split its data into training and testing data, we opted to use cross-validation, a resampling

method that uses different data portions to train and test a model on different iterations, represented in Fig. 14, to calculate the score for each regression.

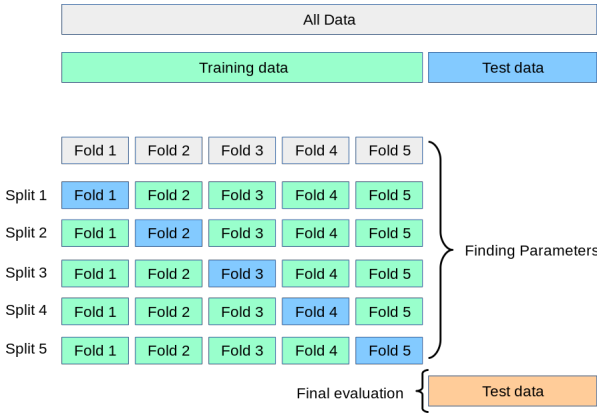

Fig. 14. Cross-validation [10]

C. Test/train split: MSE and R2

Next, we used a test/train split to calculate the Mean Squared Error (MSE) and Coefficient of Determination (R2) of the training and testing data, respectively represented in Tables II and III. The **Training** set contains 892 examples, while the **Test** set contains 446 rows.

TABLE I
CROSS-VALIDATION RESULTS

Regression Type	Cross-validation Score
Linear Regression	0.744
Random Forest Regression	0.832
Ridge Regression	0.744
Lasso Regression	0.744
Polynomial Regression	0.779

TABLE II
REGRESSION TYPE RESULTS WITH TRAINING

Regression Type	MSE	R2
Linear Regression	38188846.246	0.733
Random Forest Regression	3921867.301	0.973
Ridge Regression	3921867.301	0.973
Lasso Regression	38188846.648	0.733
Polynomial Regression	24581739.662	0.828

TABLE III
REGRESSION TYPE RESULTS WITH TESTING

Regression Type	MSE	R2
Linear Regression	32239880.653	0.795
Random Forest Regression	20890254.844	0.867
Ridge Regression	32260213.477	0.795
Lasso Regression	32240053.736	0.795
Polynomial Regression	18117605.544	0.885

D. Graphical representation of results

In addition to the tables, Figs. 15, 16, 17, 18, 19 plot the predicted value in the x axis and the difference between it and the expected value in the y axis, graphically representing each regression's results.

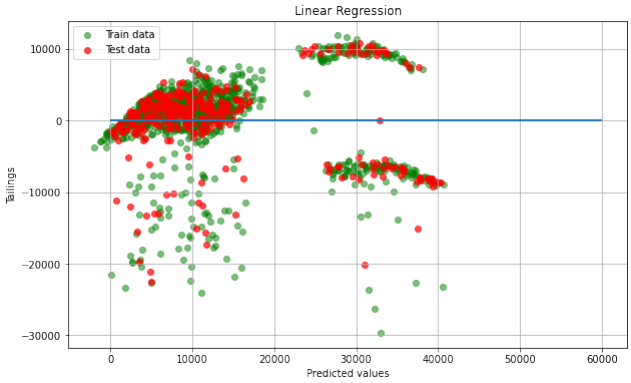

Fig. 15. Linear Regression results.

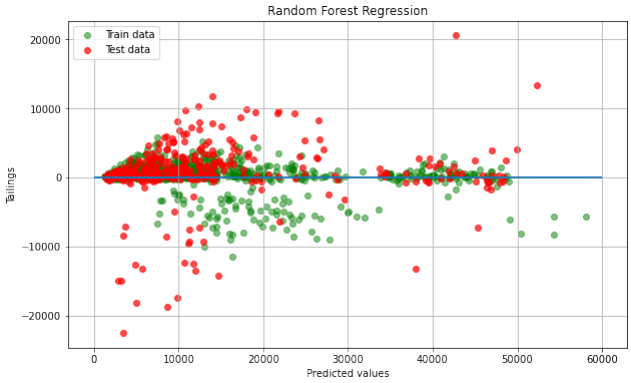

Fig. 16. Random Forest Regression results.

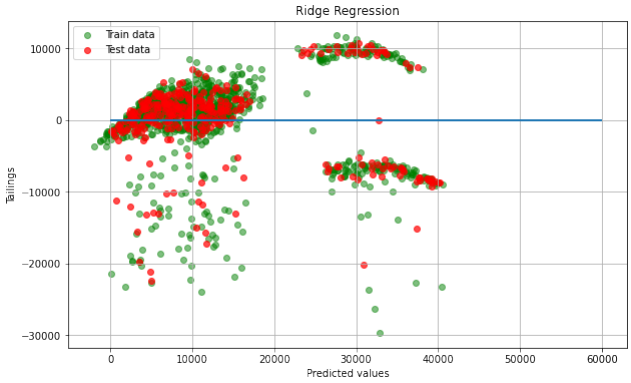

Fig. 17. Ridge Regression results.

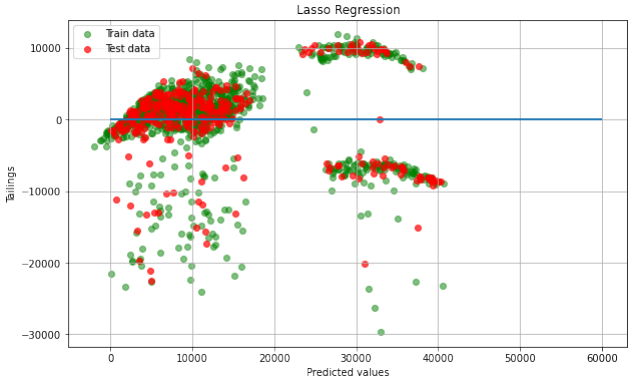

Fig. 18. Lasso Regression results.

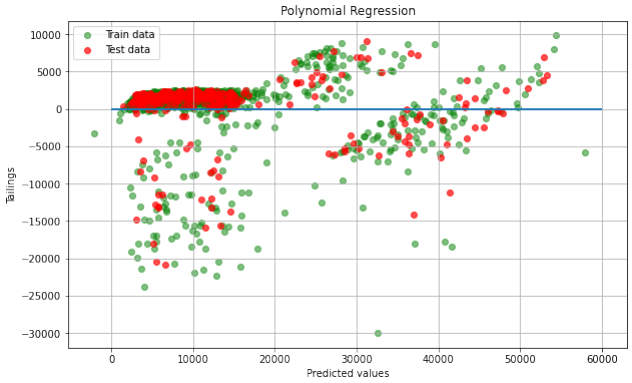

Fig. 19. Polynomial Regression results.

Through the graphs above, we can see the Linear, Ridge and Lasso regressions pale in comparison to Polynomial and Random Forest, as their points are more dispersed than those of the latter two.

Random Forest Regression exhibits more accurate predictions for higher charges. However, significant dispersion at lower costs makes it less precise in comparison to Polynomial Regression.

E. Feature Importances in the Random Forest Regression

The Random Forest Regressor allows the visualization of the feature importances, computed as the mean and standard deviation of accumulation of the impurity decrease within each tree, as in Fig. 20

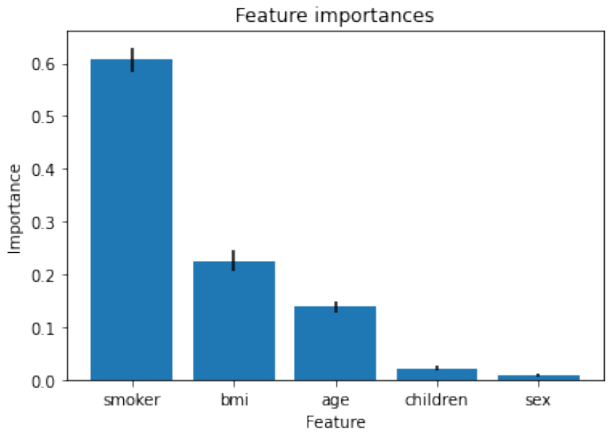

Fig. 20. Importance of Features

As expected, the three features with the most importance were, in descending order: *smoker*, *bmi* and *age*. The considerable difference between *smoker* and all the remaining features is to note, restating the large impact of smoking on medical insurance costs.

IV. Conclusion

As seen previously, the Polynomial Regression was the most appropriate one for low charge predictions, while Random Forest Regression was more suitable for higher values.

When it comes to the impact of the features, it was concluded that smoking is by far the most impactful one, greatly increasing the costs. BMI and age were also quite relevant. In contrast, the region, number of children and sex had negligible impact.

We were able to accomplish the goals of our project successfully. Namely, analyzing the dataset, applying and comparing multiple types of regression, as well as drawing conclusions from the results obtained.

The work done sheds a light on the main factors (and combinations thereof) that lead to increased insurance costs in the United States.

V. Division of Labor

Both students collaborated an equal amount through online meetings for synchronous development, in order to achieve collective responsibility over the entirety of the project.

References

[1] Lantz, Brett (2013). *Machine Learning with R*. Packt Publishing
[2] How Does the U.S. Healthcare System Compare to Other Countries?
[3] Kaggle Dataset
[4] GitHub Dataset
[5] Sklearn - Linear Regression
[6] Sklearn - Random Forest Regressor
[7] Sklearn - Ridge
[8] Sklearn - Lasso
[9] Sklearn - Polynomial Features
[10] Sklearn - Cross Validation