Predicting Insurance Charges

By Isabel Alvarez and Amanda Murray

What's the problem?

Health insurance can be **expensive**

Before the Affordable Care Act, health insurance companies could charge based on pre-existing conditions and unhealthy habits. It was well known that this was true, but the in and outs what impacted what was a a mystery.

The Data

Our data was obtained from Kaggle (https://www.kaggle.com/mirichoi0218/insurance)
Some of the key features of our data are:

- Age
- Sex
- BMI
- Children
- Smoker
- Region
- Charges (Cost)

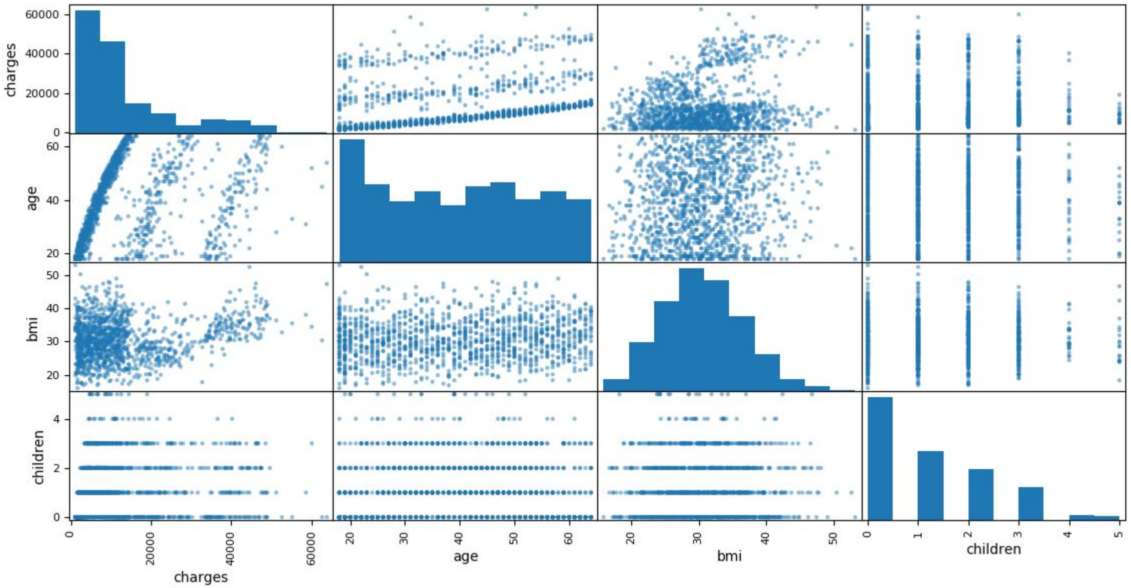
Of those features, we found that some were less useful than others. For example, sex, smoker and region were categorical features, and therefore they offered limited amounts of information.

What are the correlations?

We found these correlations:

Charges	1.000000
Age	0.299008
Bmi	0.198341
Children	0.067998

There’s a very slight correlation between cost and BMI, and a higher correlation between cost and age.



The Model

We initially tried two models: Linear Regression and Random Forest Regression. Originally, our results showed an average error of $7566 for the linear regression and $7549 for the random forest regression. Because the random forest regression provided marginally better results, we decided to fine-tune it further.

Improving our Model

We used a Grid Search to improve our model by testing a grid of various values for hyperparameters for the Random Forest Regression Model since this is the one that provided the best initial results. After this our algorithm was improved to a $7372 error.

The Results

After applying the above modifications we obtained a final error of $7372. As was discussed above, our error diminished at every step in the following manner:

Linear Regression -> $7566

Random Forest Regression -> $7549

Random Forest Regression with Grid Search -> $7372

Learning along the way

When we first started this project, we decided to use the World Happiness data set. By analyzing the data carefully, we discovered early on that the dataset would not be suitable for the task at hand. It turns out that most of the data had been processed beforehand, and therefore obtaining correlations would be trickier than with raw data.

Even though we had to start over with a different dataset, we learned the importance of analyzing the data “as humans”, instead of solely relying on statistical analysis. As we discover, every single step in the process is essential.