Predicting Healthcare Costs

A Multiple Linear Regression Approach

Kyle Thomson ○ LinkedIn.com/in/kylethomson2 ○ Feb, 2023

An in-depth walkthrough of model-building methodology and EDA can be found at github.com/kylethomson99 | Link

Background and Context

This analysis serves as an example of how data and analytics can be leveraged by both payors and healthcare professionals to better understand individual patient medical costs

Currently the Centers for Medicare & Medicaid Services (CMS) uses a similar, more sophisticated approach to model patient costs | Learn more here

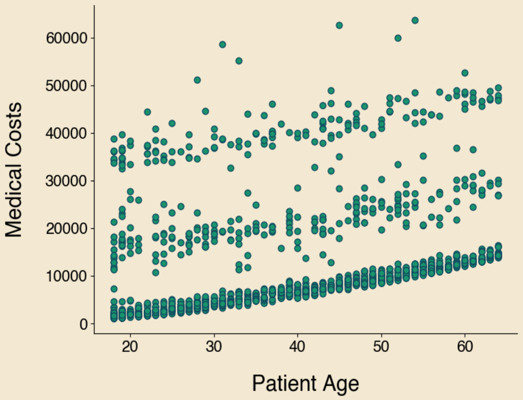
The Model

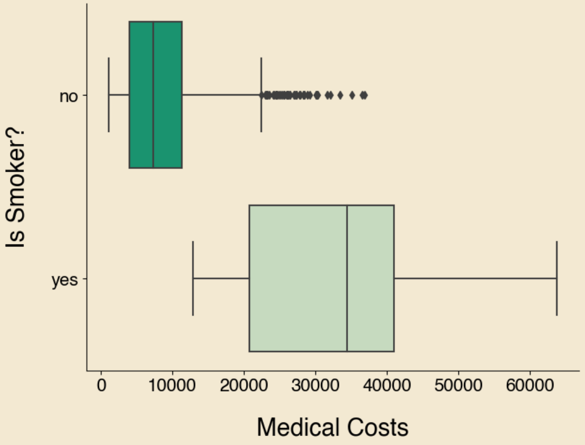
Using a linear regression model, we can predict our response variable (Healthcare Costs) from a given set of demographic patient information.

Certain features, such as smoking are great for identifying individuals with potential for higher medical costs

Exploring our Dataset

Identifying correlations between independent variables such as patient age and medical costs will help increase the accuracy of our model

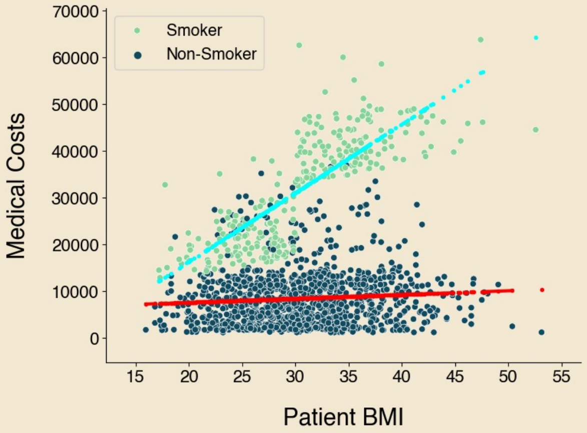




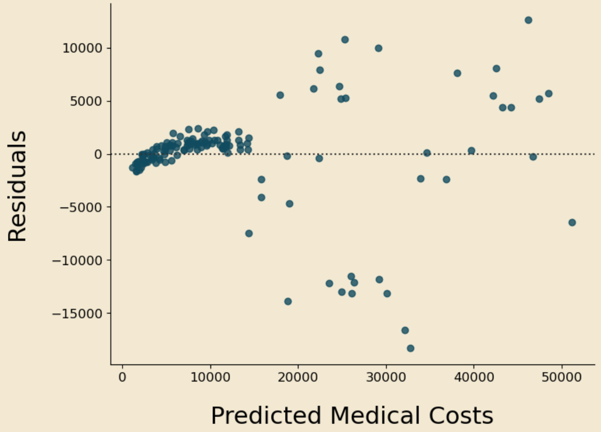
Optimizing our Model

Certain variables have interaction terms which, when introduced to our model will help further increase accuracy

Smokers have a much stronger correlation between BMI and medical costs than nonsmokers



Results



My final model had an R^2 value of 0.84 and was most accurate at predicting costs under $20k

Potential Steps to increase accuracy further

- Train on patients with +$30k charges
- Include more variables to model through additional datasets