Medical Charges Prediction and Analysis

About the Dataset

This dataset is dedicated to the cost of treatment of different patients. The cost of treatment depends on many factors: diagnosis, type of clinic, city of residence, age and so on. There is no data on the diagnosis of patients. But other information is available that can help us to make a conclusion about the health of patients and practice regression analysis.

**Columns**

* **age**: age of primary beneficiary
* **sex**: insurance contractor gender, female, male
* **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
* **children**: Number of children covered by health insurance / Number of dependents
* **smoker**: Smoking
* **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
* **charges**: Individual medical costs billed by health insurance

Use case tasks (for directions purpose only):

* How is the distribution of various properties based on gender?
* Perform hypothesis testing to check whether having children more children results in the fact that people smoke less?
* Which variable do you think effect the charges the most and why (prove it).
* What can you say about the difference in distribution of charges for smokers and non smokers.
* How is BMI affecting the distribution of charges.
* Plot the region wise distribution of charges based on other variables and draw the insights.
* Make a well performing model predicting the charges based on the various other variables.

System design tasks (for directions purpose only):

* Design system architecture to deploy this Regression ML Model in production
* How do you perform canary build?
* What should be the strategy for ML Model Monitoring?
* How do you perform load and stress testing?
* How do you track, monitor and audit ML training?
* Design framework for continuous delivery and automation of machine learning tasks.