Mini Project 2

Predicting Medical Insurance Costs

- Goals
- To understand the business value
- To address the business questions of the client
- To come out with models to predict the insurance cost using the data set provided

Presented By Sadvi Sandhya

Predicting Medical Insurance Costs

Business Value

- In order for a health insurance company to make money, it needs to collect more in yearly premiums than it spends on medical care to its beneficiaries.
- Insurance company wants to do risk analysis to design insurance policies and price them for profitability, based on the risk of insuring different groups of customers.

Stake Holders

- Insurance Companies
- Government

Presented By Sadvi Sandhya

Business Questions

- What are the factors influencing the high claim charges ?
- Which group of beneficiaries are at higher risk of large medical expenses?
- Can we predict the cost of health insurance based on factors that influence?

Presented By Sadvi Sandhya

Approach

- Understanding dataset
- Exploratory Data Analysis
- Feature Selection
- Modelling using Linear regression

Presented By Sadvi Sandhya

Dataset

Dataset includes

1,338 examples of beneficiaries

Features indicate the characteristics of the beneficiaries

- *age:*
- *sex:*
- *bmi:*
- *children:*
- *smoker:*

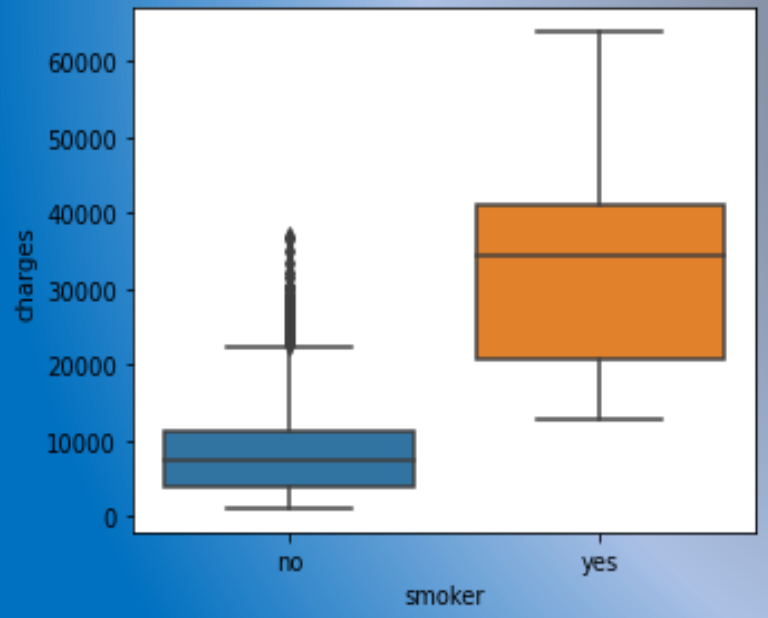
The total medical expenses charged to the plan for the calendar year. The features are:

- charges :

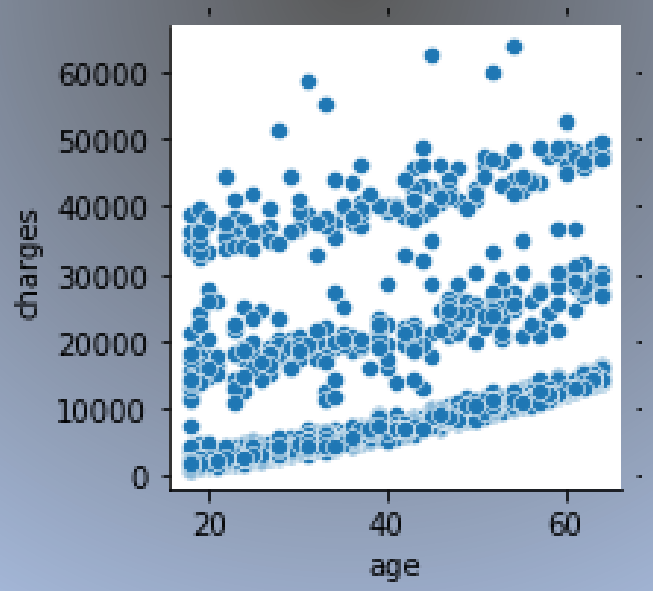
Highlights of the Exploratory Data Analysis

- More number of Smokers Vs Non smokers
- High charges for smokers in comparison to non-smokers.
- As age goes up charges for health insurance also trends up
- Smokers with high Obese have higher charges

Plot of Charge vs Smoker and Non smoker

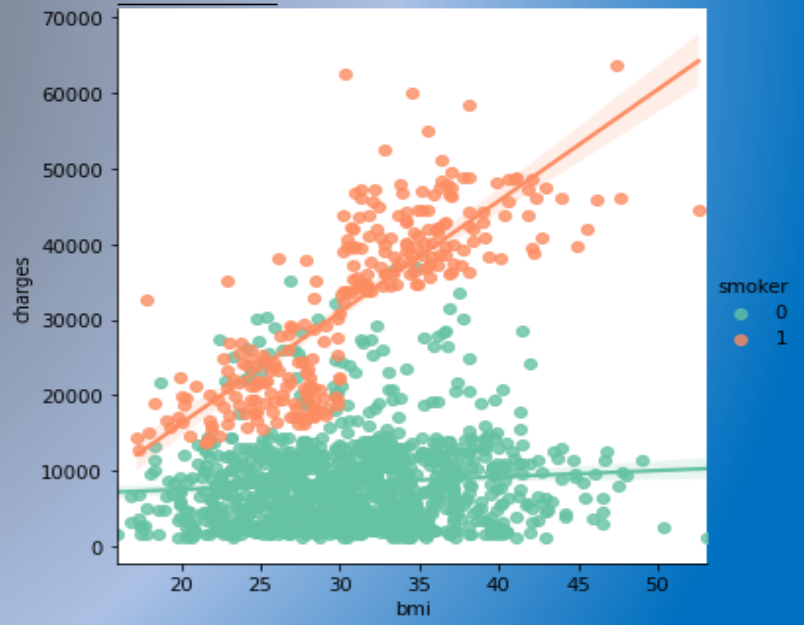


Plot of Charge vs Age



Plot of Charge vs BMI

With different hues for smoker and non smoker



Presented By Sadvi Sandhya

Modelling

- Linear Regression
- Ridge Regression
- Lasso Regression

Metrics

Model	MSE	RMSE	R2_Score
Linear Regression	33136120	5756.39	0.759095
Ridge Regression	33133800	5756.19	0.759112
Lasso Regression	33135920	5756.38	0.759097

The RMSE would suggest that, on average, predictions varied from observed values by an absolute measure of $5756.
The R2_score would suggest that the model has a fit of 75%
The model explains nearly 75 percent of the variation in the dependent variable

Presented By Sadvi Sandhya

Conclusions

- Smokers have higher medical expenses than non-smokers.
- Beneficiaries having BMI >30 together with smoking form a synergistic effect on charge.
- smoking, age and obesity are the factors that contribute the most in the calculation of insurance costs.
- The RMSE would suggest that, on average, predictions varied from observed values by an absolute measure of $5756.

Presented By Sadvi Sandhya