Predicting Health Insurance Premiums

DSC450 APPLIED DATA SCIENCE

LARA MECHLING, CORRINA HANSON, ISAAC LIEM

Agenda

- Introduction
- Business Problem
- Data
- Methodology
- Results
- Conclusion
- References

Introduction

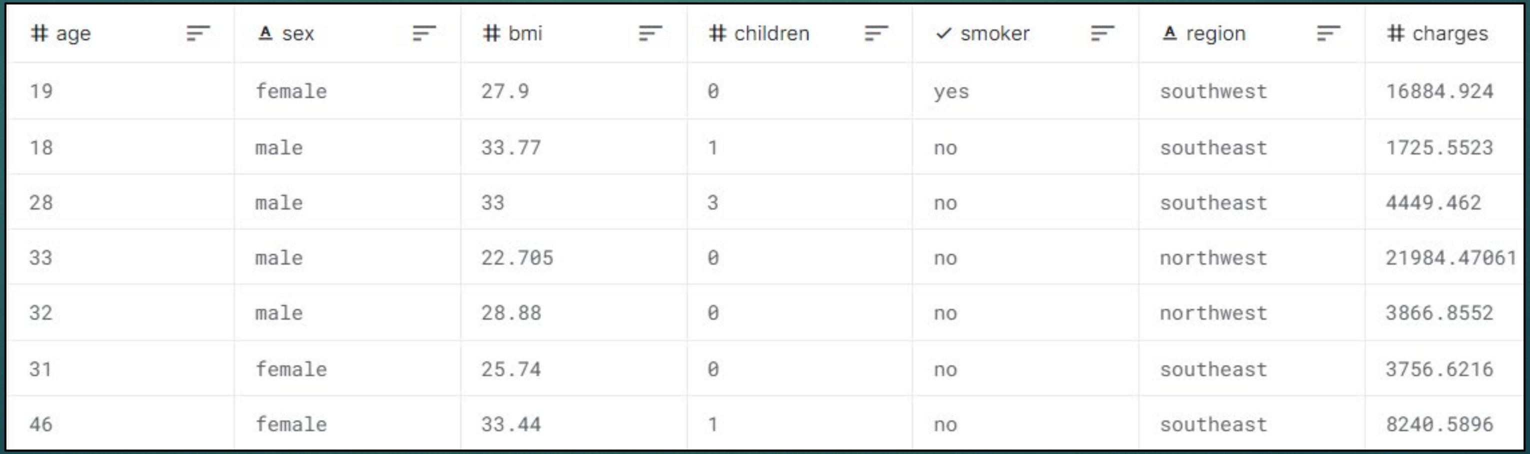
- Health insurance, in 2020, was a thirty-one-billion-dollar industry
- Health insurance is intended to provide protection from extraordinarily high costs of medical care
- Policy holders pay monthly premiums to maintain coverage
- Several factors are used to calculate premiums
- Predictive modeling can be used to determine what someone’s premium will be

Business Problem

- Provide more accurate insurance costs by determining which factors play the strongest roles in determining health insurance premiums

Data

- The data was acquired from Kaggle: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset
- The data consists of1338 row and 7 columns
  - Columns: age, sex, bmi, children, smoker, region, and charges
- The target variable is the charges column

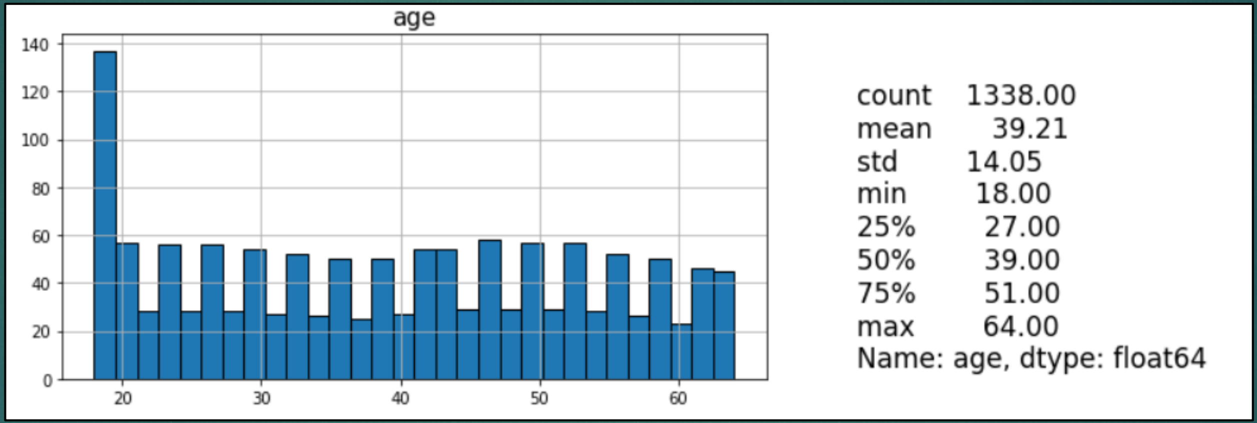


# age	A sex	# bmi	# children	✓ smoker	A region	# charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47061
32	male	28.88	0	no	northwest	3866.8552
31	female	25.74	0	no	southeast	3756.6216
46	female	33.44	1	no	southeast	8240.5896

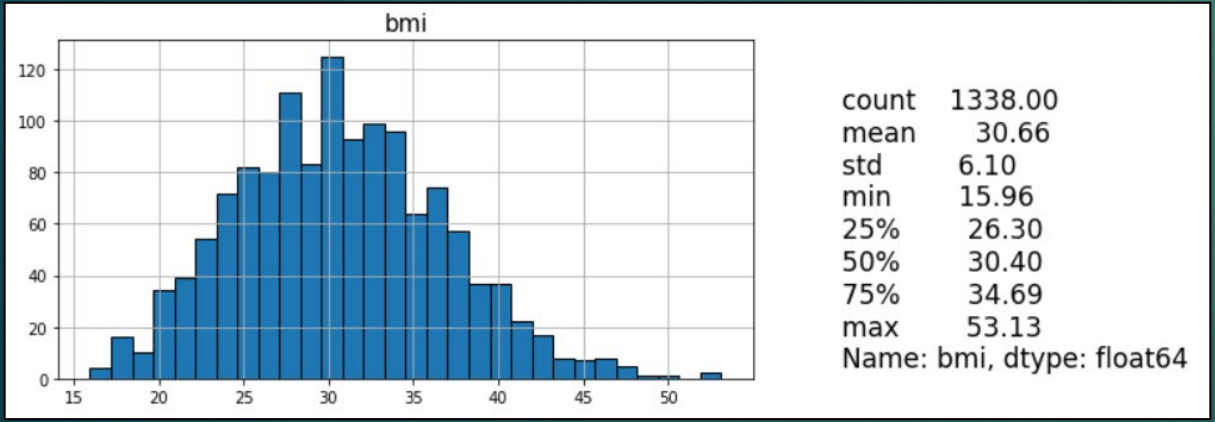
Methodology – Data Preparation

- Data was inspected for missing variables and outliers
- Numerical variables were normalized (scaled) for model building
- Categorical variables were converted to numerical variables
- Exploratory data analysis was performed

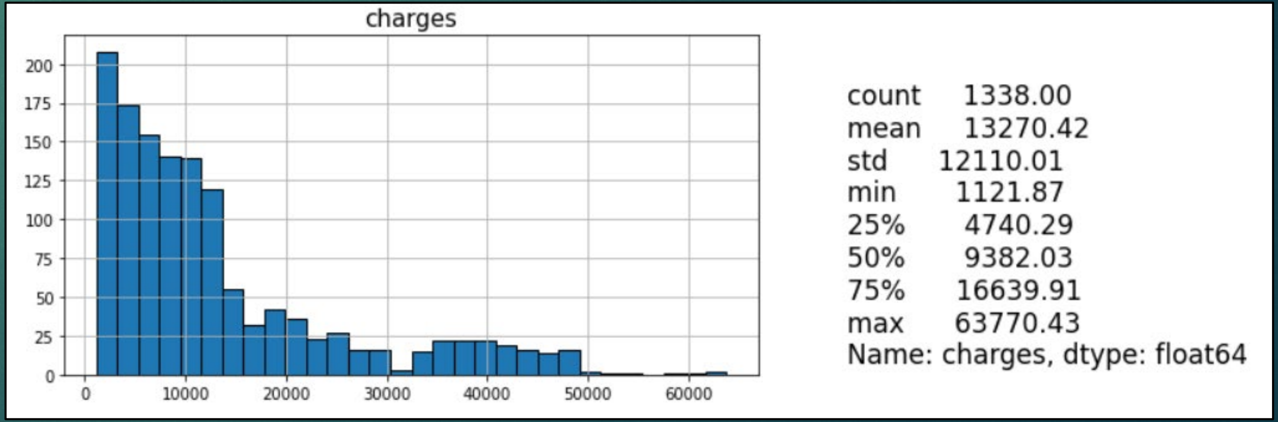
Methodology – EDA
Distribution of Numerical Data



```
count    1338.00
mean       39.21
std        14.05
min        18.00
25%        27.00
50%        39.00
75%        51.00
max        64.00
Name: age, dtype: float64
```



```
count    1338.00
mean       30.66
std         6.10
min        15.96
25%        26.30
50%        30.40
75%        34.69
max        53.13
Name: bmi, dtype: float64
```



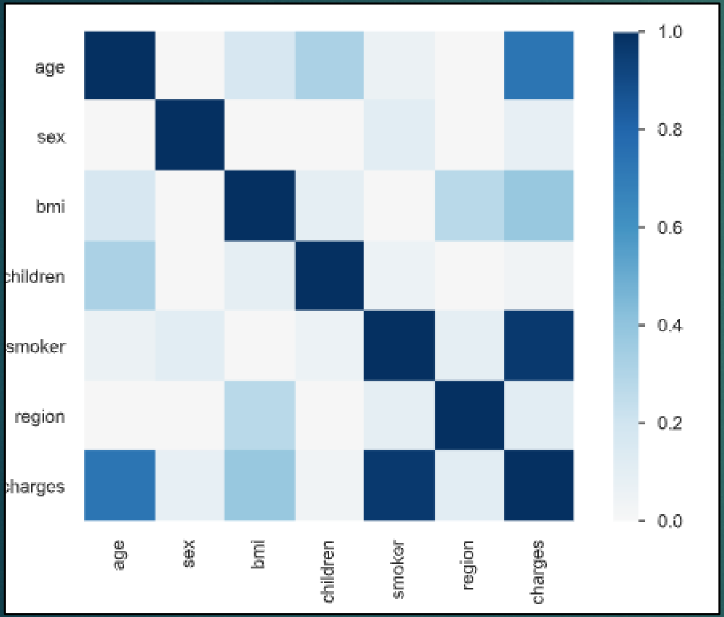
```
count     1338.00
mean     13270.42
std     12110.01
min      1121.87
25%       4740.29
50%       9382.03
75%      16639.91
max      63770.43
Name: charges, dtype: float64
```

Methodology - Regression

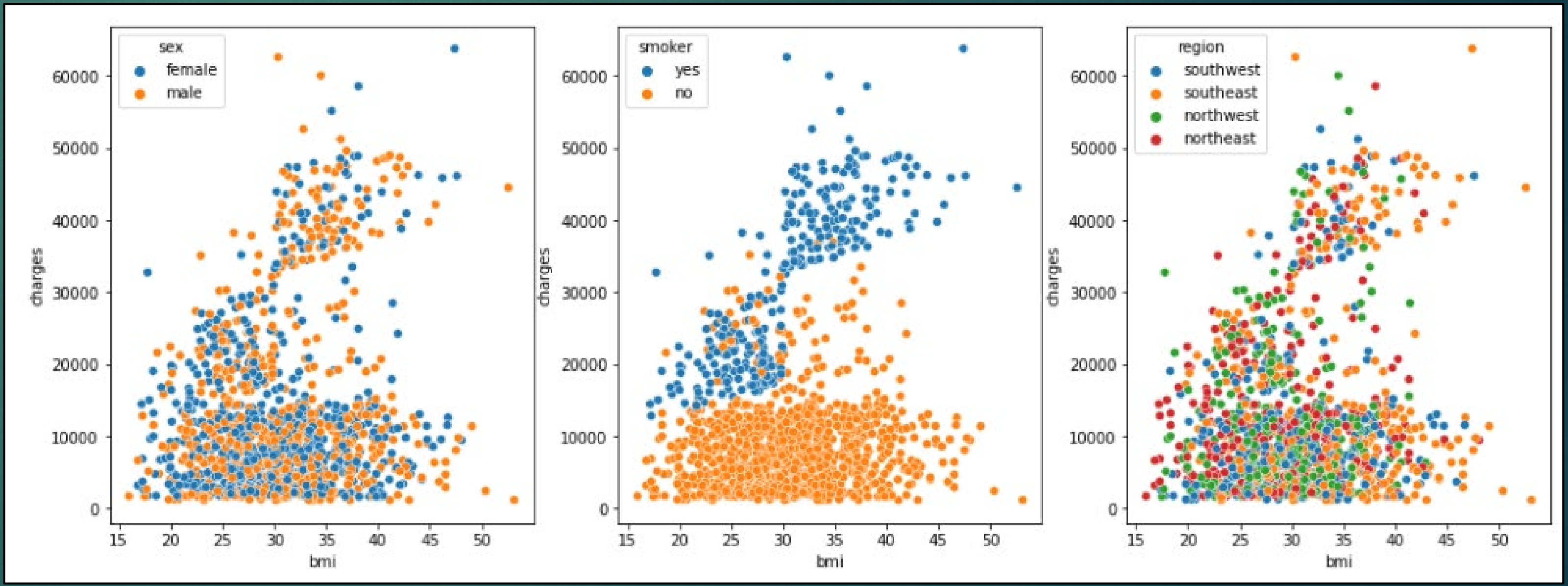
- A regression model was chosen
- y variables were all data points except the target variable
- Test, train, split was employed

Results

- The strength of the relationship between x and y can be seen in the correlation plot below



- Scatter plots to show relationship between charges and bmi/smoking status, bmi/sex, bmi/region



Conclusion

- Many variables are taken into account when calculating health insurance premiums – with some having a greater affect
- Smoking status and age have the highest impact on monthly premiums
- The model preformed with 78% accuracy
- Using an individual's health information to calculate monthly premiums is a more cost-effective approach
- Use voluntary questionnaires to avoid HIPAA violations

References

- Bhardwaj, N., & Anand, R. (2020). Health Insurance Amount Prediction. International Journal of Engineering Research and Technology.
- English, A., & Lewis, J. (2016, March). Privacy Protection in Billing and Insurance Communications. Retrieved from AMA Journal of Ethics: https://journalofethics.ama-assn.org/article/privacy-protection-billing-and-health-insurance-communications/2016-03
- Kaur, T. (2018). Factors Affecting Health Insurance Premiums: Explorative and Predictive Analysis. Retrieved from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dr.lib.iastate.edu/server/api/core/bitstreams/a8729ea4-0ba4-443a-b74d-d5c745470a79/content
- National Association of Insurance Commissioners. (2021). U. S. Health Insurance Industry 2020 Annual Results. National Association of Insurance Commissioners.
- What is Health Insurance Premium? (n.d.). Retrieved from HealthInsurance.org: https://www.healthinsurance.org/glossary/health-insurance-premium/