Team 2: Health Insurance

By: Christian Rodriguez, Philip Waymeyer, Savion Ponce, Victor Pham

Overview

- Datasets
- Initial Questions
- Research Process
- Dashboard
- Machine Learning Model
- Conclusion & Recommendations

Datasets

- US Health Insurance
  - Provides an annual culmination of individual medical costs incurred by people with different attributes (e.g. Age, BMI, Children, Sex, Region, Smoker status).
- Indicators of Health Insurance Coverage
  - Provides data on the percentage of different demographics (e.g. Age, Sex, Race, Education, State) at different coverage levels.
- Small Area Health Insurance Estimates (SAHIE)
  - Data on the percentage of selected demographics insured (e.g. Age, Sex, Race, State, County, Income category). Collected by the U.S. Census Bureau.

Initial Questions

1. What factors affect health costs incurred by health insurance customers? (Age, BMI, Children, Region, Sex, Smoker/NonSmoker)
2. How do certain factors affect insurance status?
    a. Education
    b. Location (State)
    c. Sex
3. How does income affect percentage insured?
4. What specific demographics have the highest percentage uninsured?
5. Can we predict the healthcare costs based off the following factors: Age, Body Mass Index, Children, Region, Sex, Smoker/NonSmoker?

Research Process

- Ran data through ETL to facilitate answering initial questions
- Explored data to seek trends in demographics and insurance status
- Created visualizations to display these trends

Dashboard

- Created using Plotly Dash

Machine Learning Model

- Our goal was to create a model predicting medical costs of potential customers
- Used sklearn's LinearRegression as base model using age, BMI, children, sex, region, and smoker status as the independent variables
    - Checked the effect of sex on model
        - Model with sex performed the same as model without sex variable ($R^2 = 0.75$)
        - Removed sex as an input variable to avoid ethical concerns

Machine Learning Model

- Refined model performance by applying polynomials of degree N to the features
  - Checked model performance for multiple values of polynomial degrees from 1 to 8
    - Best model performance came from data with features of polynomial degree 3
    - The new R^2 value was 0.87

Conclusion

- Insurance companies should focus on marketing to uninsured population with low health costs to maximise success and increase potential profit.

Recommendations

- Low-income young men on a national scale and people with a lower level of education are likely to be uninsured.
- Texas, Oklahoma, Georgia have a higher percentage of uninsured people.
- Non-smokers showed a statistically significant lower health costs than smokers, and there was a visible upward trend with age and health costs.
- To get a better idea of the predicted health costs for a potential customer, insert attributes into our predictive model.