Investigating Bias in Insurance Premium Prediction

Chuan Chen*
cc6580@nyu.edu
Center for Data Science
New York University
New York, NY 10012

Yihang Zhang*
yz2865@nyu.edu
Center for Data Science
New York University
New York, NY 10012

1 Background

This ADS aims to use costumers past medical expenses and their features such as sex, age, bmi, and etc. to predict their future expenses, ultimately to help insurance companies make decisions on premium charges. The purpose of this project is to investigate if there exist bias in this ADS that is favors one group over another unfairly, such as charging a particular group with the same feature values over another group.

2 Input and Output

This data was obtained from the Machine Learning course website (Spring 2017) from Professor Eric Suess at California State University. It has 1338 samples. It contains 6 features, age, sex, bmi, number of children, smoker, region, and 1 target variable, medical expenses, as the output. Typically, if a patient is predicted to have higher medical expenses, the insurance companies will take this into consideration and raise their insurance premiums. The datatype and missing values for each variable can be seen in table 2

variable	datatype	missing values
age	integer	0
sex	binary	0
bmi	float	0
children	integer	0
smoker	binary	0
region	category	0
expenses	float	0

The distribution of numerical features can be seen in figure 1. The distribution of numerical features

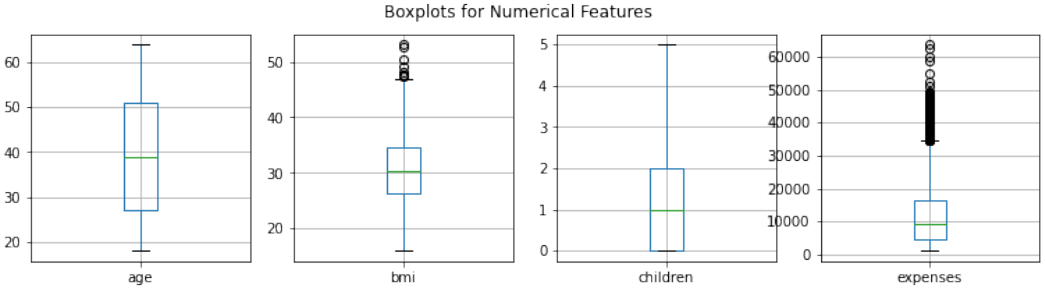

Figure 1: Boxplots for Numerical Features

*Authors are listed alphabetically and contribute equally to the paper.

can be seen in figure 2. The distribution of categorical features can be seen in figure 3. The pairwise

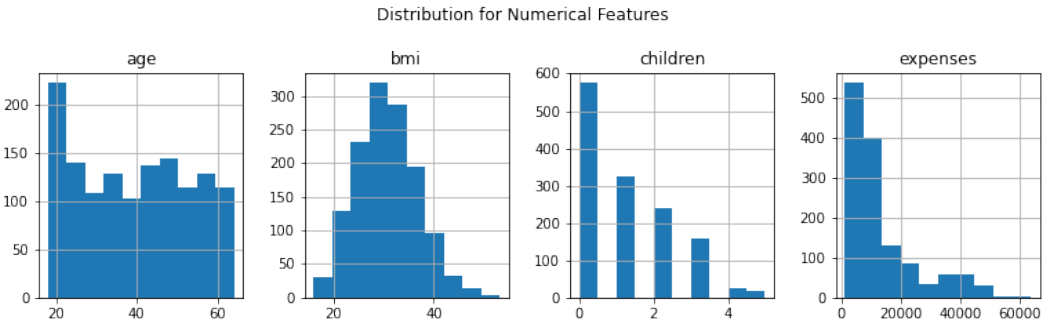

Figure 2: Distribution of Numerical Features

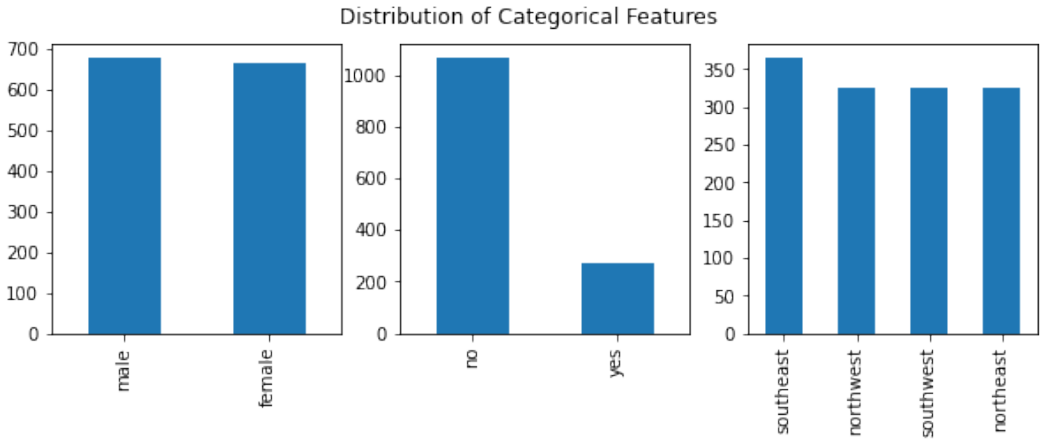

Figure 3: Distribution for Categorical Features

correlation between each feature can shown in figure 4.

The relationship between each feature and the target variable can also investigated and can be seen in figure 5 and 6.

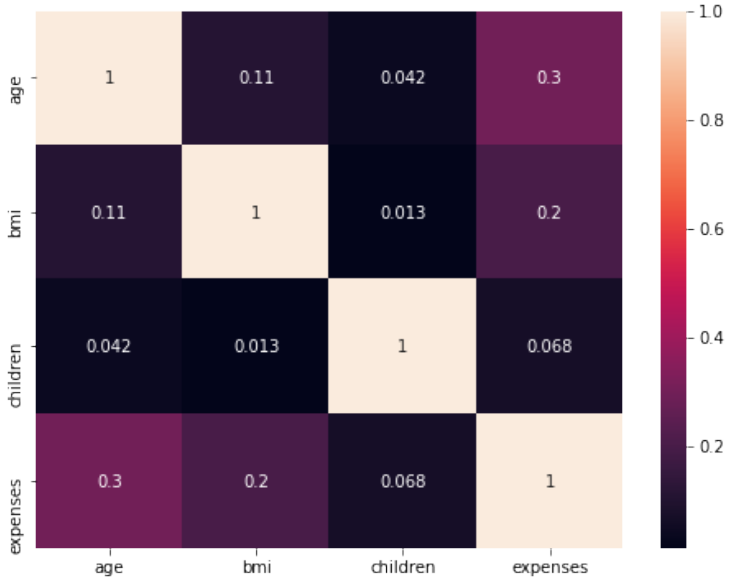

Figure 4: Pairwise Correlation Between Each Feature

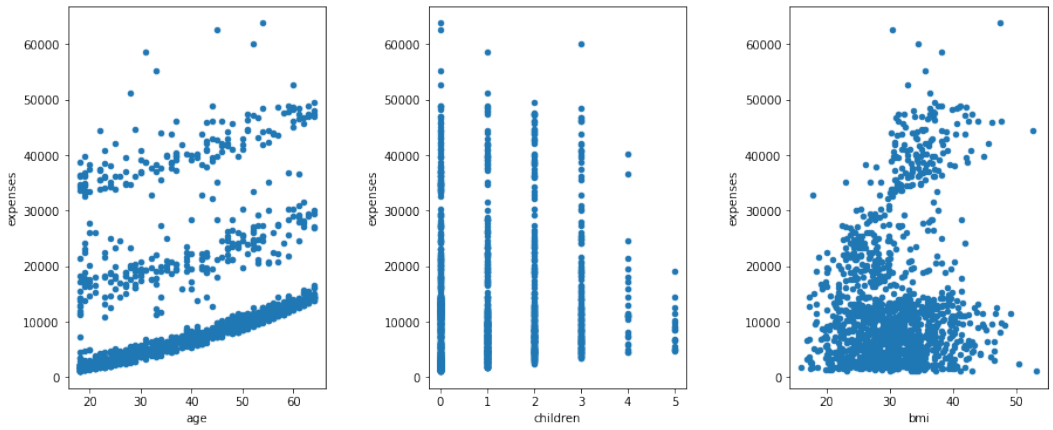

Figure 5: Correlation Between Numerical Features and the Target Variable

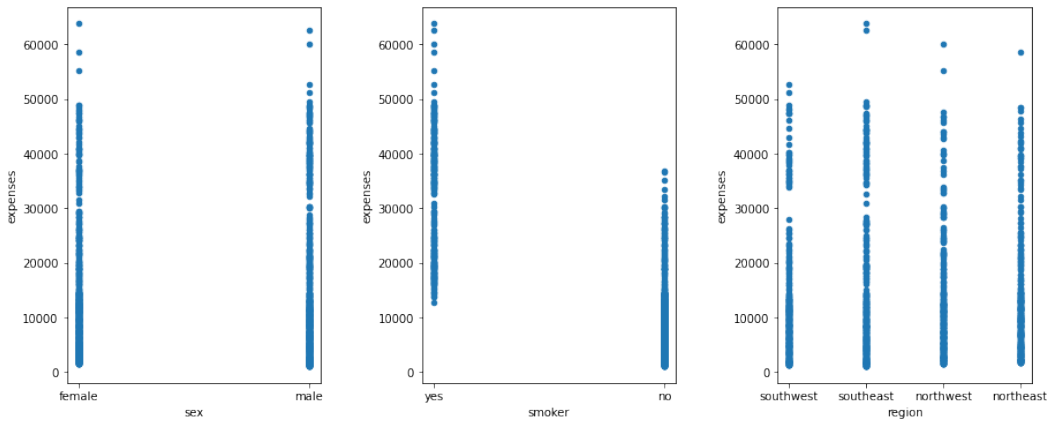

Figure 6: Correlation Between Categorical Features and the Target Variable

3 Implementation and Validation

This data has already been cleaned beforehand, so no additional data cleaning was performed. The categorical features were encoded into integers. Then, the data was split into the train and test sets, and each feature was standardized. The ADS has implemented several different regressors, such as linear, polynomial, support vectors, decision tree, and random forest. It ultimately decided to use random forest as it is the regressor with the highest accuracy on the test set. However, besides accuracy, no the metric were used.

4 Outcomes

We plan to analyze the accuracy of this ADS across different subgroups to see if it contains any bias. We also plan to implement several fairness measures we learned from class, such as mean difference and disparate impact. We plan to also investigate the difference of selection rate, FNR and FPR across groups. We want to analyze the stability of the ADS by adjusting feature values or feature weights slightly and observing the effect it has on the ADS prediction.

5 Summary

Our summary will be written after our thorough analysis in the previous sections.