RDS Project Report

Ke Yang (ky630) and Zane Dennis (zdd210)

1. Background

The automated decision system (ADS) we chose to study is a regression system explicitly designed to predict total health insurance charges for a given person. However, the ostensible eventual purpose of the system would be for an insurance company to identify the risk of insuring a given applicant. The company would then use the risk calculated by the ADS to determine how much to charge the applicant, or whether to deny them coverage altogether.

This usage creates an ecosystem with two primary stakeholders: the insurance company and the applicant(s). The company, whose ultimate goal is to make a profit, does so by gaining customers and charging them more in premiums than the customer gets charged in medical charges. If they charge their customers too little in premiums, they lose money by paying out more in medical charges than they gain. If they charge their customers or potential customers too much, they risk losing those customers altogether to competitors offering cheaper rates. Thus, the company wants an accurate ADS so that they can offer cheap, but not too cheap, rates to their applicants. The applicants, meanwhile, have the goal of getting insurance for the cheapest rate possible. As such, they want an accurate ADS so that the company will not charge them an unduly high rate (though they benefit from errors made in the other direction).

2. Input and output

This data was used as an example dataset in the book *Machine Learning with R* by Brett Lantz and is now public domain. Beyond that, its origin is unclear. While it is not known whether the data was artificially generated or compiled from a real-world source, we can at least see that the data consists of people living in the United States (as can be inferred from the description of the "region" feature) and it appears to represent a hypothetical database of existing customers of a health insurance company (as can be inferred from the knowledge of the "charges" target variable).

The data contains six features in addition to the target variable: age, sex, bmi, children, smoker, region, and charges (the target variable). 'Age," "sex," and "bmi" all refer to the applicant; "age" is an integer in years (18-64), "sex" is a binary feature ("male" or "female"), and "bmi" is a continuous value rounded to two decimal places. "Children" refers to the number of children the applicant claims as a dependent, not to how many children the applicant has had. It ranges from 0 to 5. The exact definition of "smoker" is not specified, so possible meanings may or may not include current smoking habits, a history of smoking, or even something else like a history including secondhand smoke. In the data, it is only a binary feature ("yes" or "no"). "Region," referring to the applicant's domicile, designates four geographical quadrants within the United States: "northeast, "northwest," "southeast," and "southwest." "Charges" is a continuous value

ranging from approximately 11,000 to 64,000, and it denotes the total medical costs billed by that customer to the insurance company. All categorical features are stored as strings.

There are no missing values in any feature. Full distributions of each feature are displayed in Figure 1 and Figure 2 below.

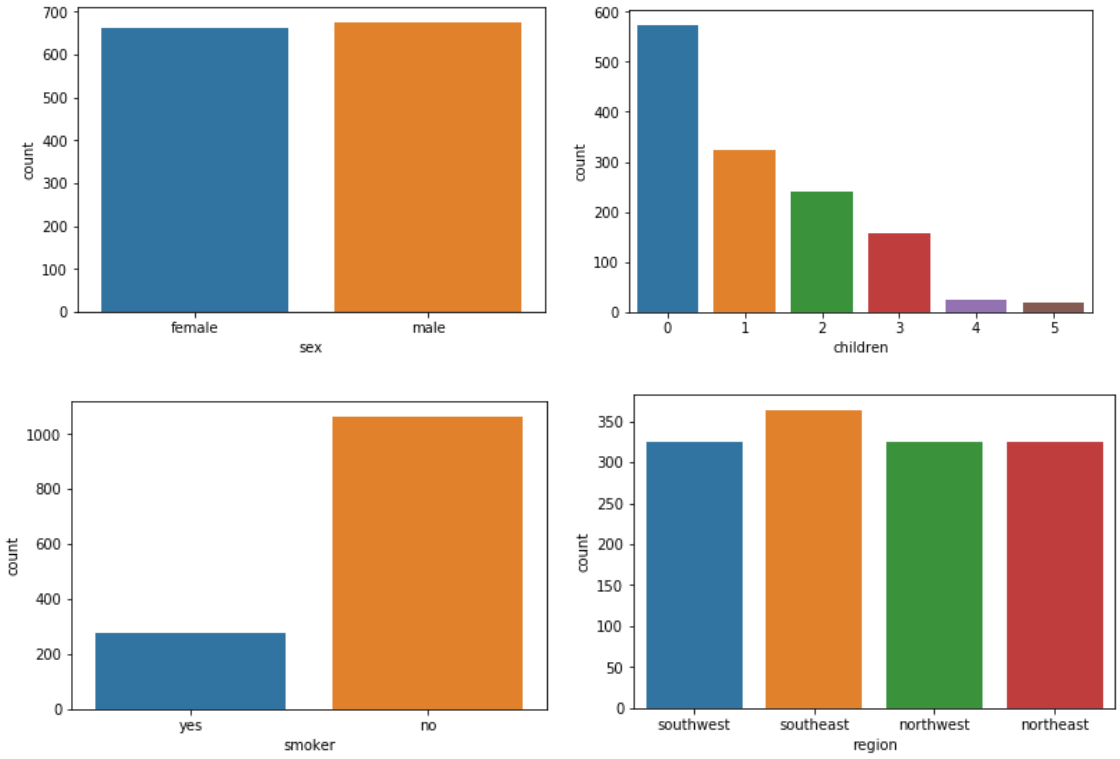


Figure 1. Distributions of the categorical features

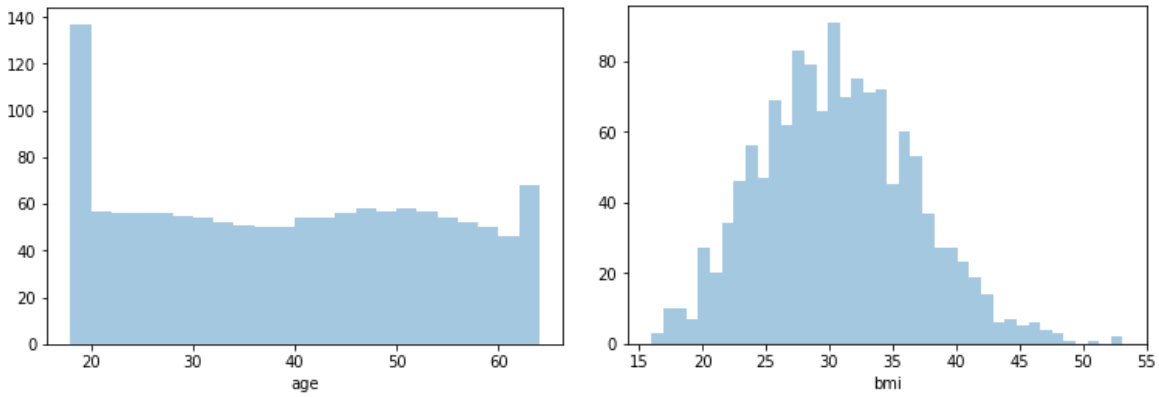


Figure 2. Distributions of the continuous features

Men and women are approximately equally represented in the data, as are the geographical regions (southeast is slightly more common than the rest but not drastically so). Age is approximately uniform within the interval with two exceptions: a small spike in the top bin and a massive spike in the first bin. The range of ages is due to 18 representing adulthood and 65

representing the age at which Americans become eligible for Medicare (government insurance), at which point they would use Medicare instead of private insurance. The prevalence of younger Americans in the dataset helps explain the distribution of children, as having no children is most common (with each additional dependent child less likely). Smokers are much less common than non-smokers, and BMI is approximately normally distributed around a mean of 30, with a slight right skew.

Figure 3 below shows the distribution of charges for each feature. In these graphs, “age” and “bmi” have been binarized into young/old and non-obese/obese. Those at least 25 years old were designated as old, and those with a BMI of at least 30 were designated as obese. Figure 4 shows pairwise correlations between attributes.

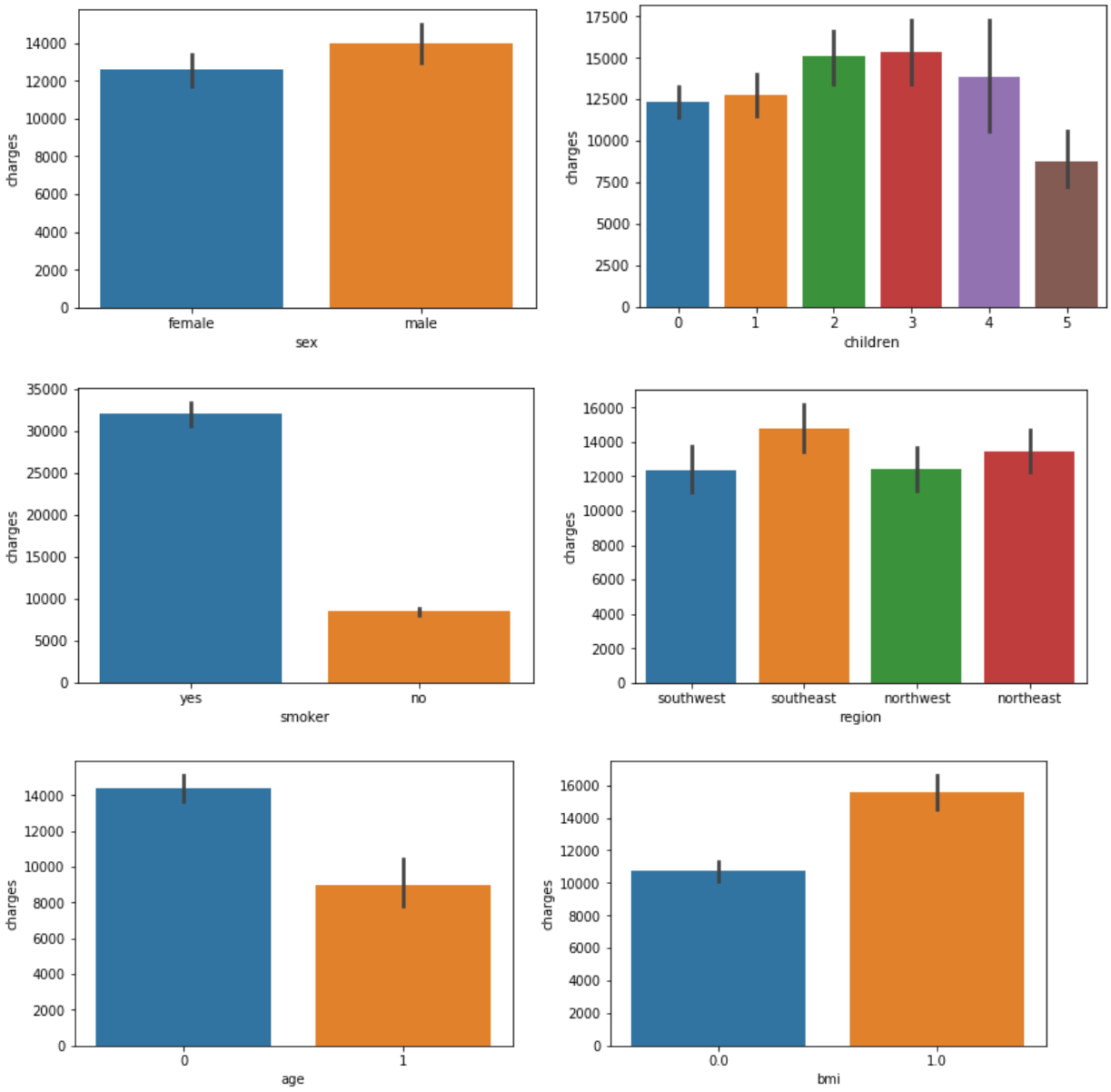


Figure 3. Distributions of the target variable for each feature

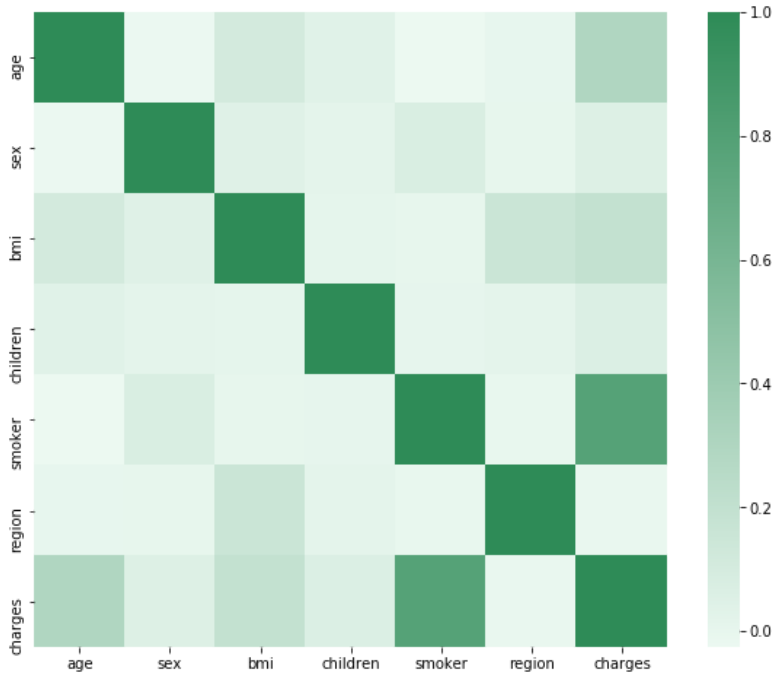


Figure 4. Correlations between all pairs of attributes

We found that there were few notable correlations between attributes in our dataset. The most prominent correlation is that between “smoker” and “charges,” an intuitive one considering the well-known relationship between smoking and lung cancer. The next strongest appears to be between “age” and “charges,” which once again makes perfect sense, as a wide variety of expensive health conditions become more common as humans age. It is almost surprising this correlation isn’t stronger. Similarly, there appears to be a possible weak correlation between “bmi” and “charges” that one might expect to be much stronger. The rest of the correlation map appears to be mostly statistical noise.

Our ADS uses three different regression models that output the predicted charges, in dollars, an input applicant would generate as a customer. All threes predictions are outputted separately, and investigation of differences in predictions on the same input is left to the user. These predicted charges can be understood as the level of financial risk the insurance company would take on by offering insurance to that applicant. As such, the predicted charges would inform the premium rates the company offers the applicant, or whether the company accepts the applicant at all.

3. Implementation and validation

The ADS’ preprocessing consists entirely of encoding the categorical variables (“sex,” “smoker,” and “region”) as integers. No data cleaning was necessary, as data exploration revealed pre-cleaned input data (no null or otherwise unexpected or inappropriate values).

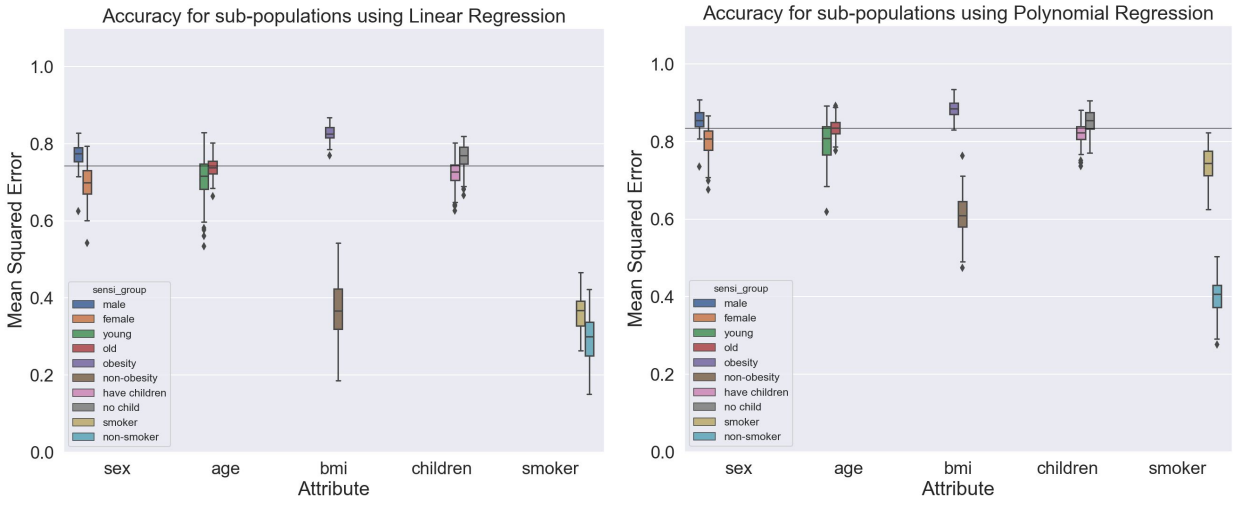
The ADS's implementation consists of a simple linear regression, a linear regression with polynomial features of degree 2, and a random forest regression with 100 estimators and mean squared error (MSE) as the criterion. Each regressor was validated using the r^2 value of the predictions against the correct charges. Validation was done using a separate test set.

4. Outcomes

4.1 Accuracy

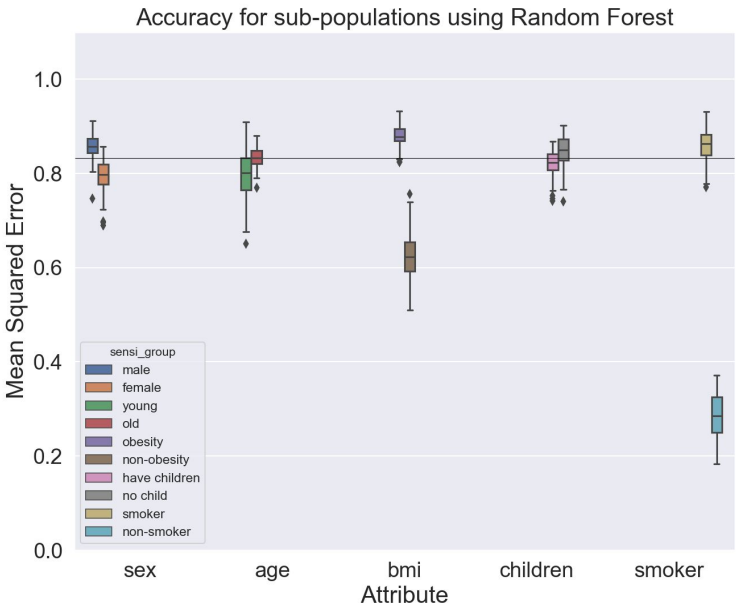
Since our ADS uses regression models, we measured the accuracy by mean squared error. We designated each model as such: Linear Regression (LR), Polynomial Linear Regression (PR), and Random Forest (RF).

For the two numerical attributes: age and bmi, we discretize them into binary attributes by a threshold. We label people with age less than 25 as young and old otherwise. We label people with bmi greater or equal than 30 as obese and non-obese otherwise based on general BMI's definition. We consider 10 subpopulations defined by the attributes age, bmi, sex, children, and smoker. We show the accuracy of each subpopulation in Figure 5 below.



a) Linear Regression

b) Polynomial Regression



c) Random Forest

Figure 5. Accuracy for subpopulations

We can observe from Figure 5 that, among the three regression models, Random Forest has the highest accuracy in most subpopulations (with non-smokers being the notable exception). Note that the overall accuracies for the models (aggregated from 100 different training and test splits) are around 80% (marked by the black lines in Figure 5), specifically, 74.2% for Linear Regression, 83.5% for Polynomial Regression, and 83.3% for Random Forest. Though Polynomial Regression model has the highest overall aggregated accuracy, Random Forest model is more accurate considering the accuracy across all ten subpopulations. This is because Random Forest uses a bootstrapping strategy to improve the model's accuracy.

The accuracy of the four subpopulations defined by the "bmi" and "smoker" attributes differ the most. This accuracy difference can be observed in all 3 regression models. All three regression models are more accurate for smokers and people with obese than the corresponding other groups. We notice that the accuracy for obese people, males, smokers, and people with no children are the subpopulations that have accuracy greater than the overall accuracy in two regression models (LR and RF).

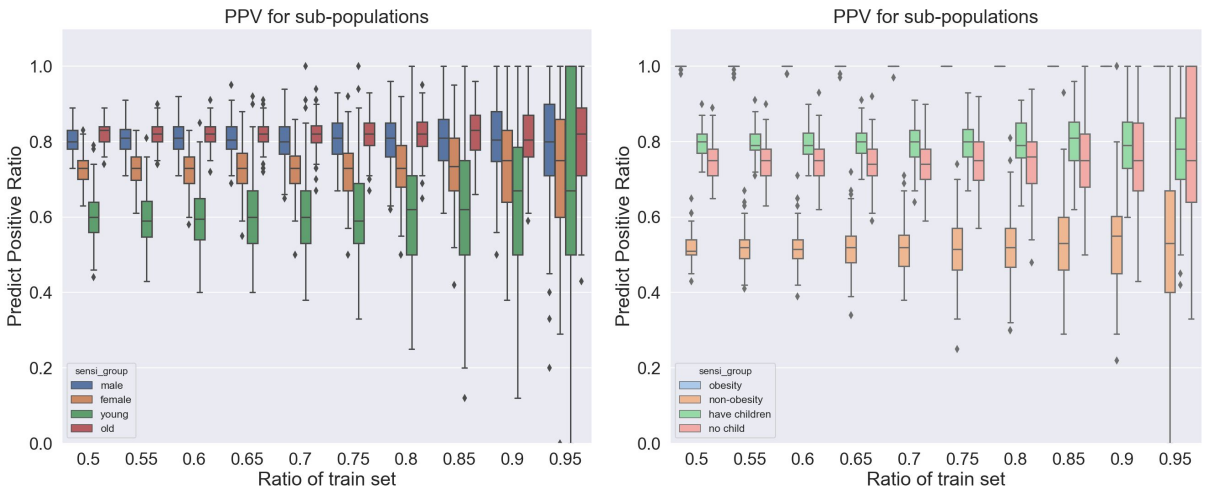
Based on the value distributions shown in Figure 1 and 2, we notice that the subpopulations defined by age and smoker are not balanced. Most people in the input data do not smoke (79.5%) and are relatively old by the age threshold 25 (79.2%). However, the accuracy for smokers is much higher than for non-smokers. In a more balanced pair of distributions, obese (52.8%) and non-obese (47.2%), the accuracy for obese from all 3 regression models is greater than for non-obese, especially in the LR model.

4.2 Fairness

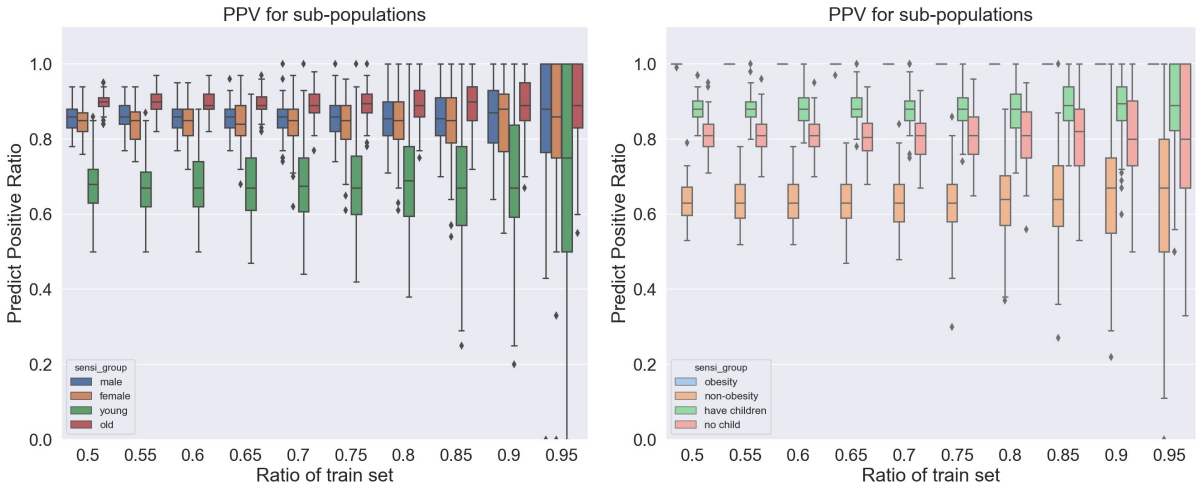
In this section, we consider the group fairness, which we define as predictive parity, across the ten subpopulations. To do this analysis, we visualize Positive Predictive Value (PPV) for each subpopulation. Since the subpopulations with respect to “bmi” and “smoker” are not balanced in our input data, we believe that PPV is a good fit to perform group fairness analysis for our data and ADS. As one of the calibrated predictive measures, PPV takes the population size into account. Additionally, it holds a greater importance to both stakeholders of the system. The insurance company wants to be certain an applicant tagged as high-risk is actually high-risk because they do not want to risk losing the customer altogether to another company by offering an incorrectly high premium. Meanwhile, the customer wants to be certain that they are actually high risk if tagged as such so that they are not overpaying for unnecessarily expensive insurance.

As we observed in Figure 4, insurance charges are highly correlated with smoking status We ignore the “smoker” attribute in this fairness analysis, since most of the people with positive predicted charges are smokers. We compare the PPV for all other eight populations defining by sex, age, bmi, and children.

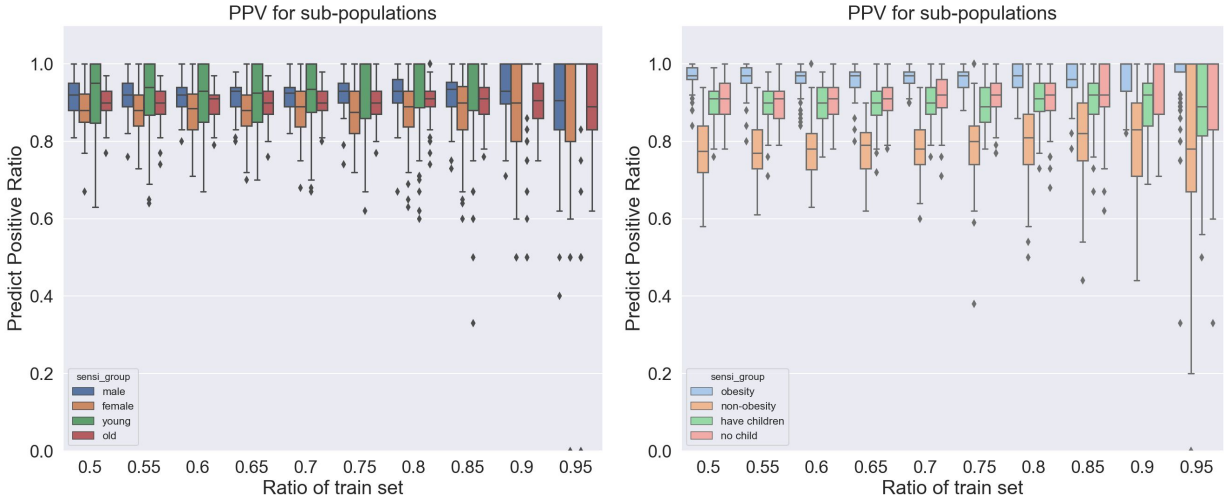
We visualize the PPV distributions of all eight subpopulations with different train-test splits in the following Figure 6.



a) Linear Regression



b) Polynomial Regression



c) Random Forest

Figure 6. PPV distributions for all subpopulations

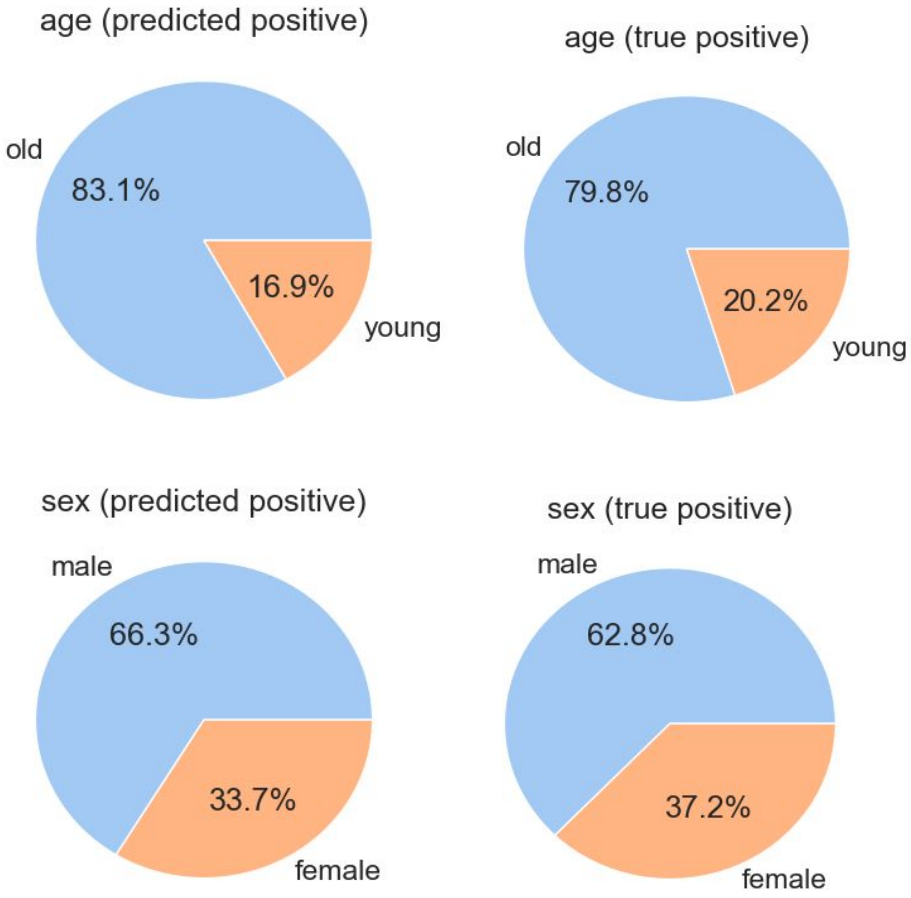
In the “age” attribute, the two subpopulations are not balanced in the input data with 80% of people falling in the older group. In Figure 6, we notice that PPV of young and old subpopulations have greater difference in Linear Regression and Polynomial Regression models. However, the PPV for young and old are more balanced in the results of Random Forest model (Figure 6c). Since we know that Random Forest model uses bootstrapping aggregation, we can guess that the predictions of charges for young and old subpopulations are sensitive to the “age” attribute. A regression model using an aggregation strategy, like Random Forest, might be preferred for this data.

Let us consider the “bmi” attribute next. The subpopulations for obese and non-obese are more balanced in the input data. However, the PPV for obese is much more greater than PPV for non-obese in all 3 regression models. Even in Random Forest, the difference of PPV between

two subpopulations is noticeable. We can therefore conclude that bmi impacts the positive prediction of insurance charges.

4.3 Diversity

Diversity component shows the representations of the ten subpopulations in the predicted and true positive sets of input data. Specifically, we want to compare the representation of the positive instances in the predicted results for all the subpopulations. Figure 7 shows the results of the Random Forest model. For brevity, we do not show the results from the Linear and Polynomial Regression models in the report. Please refer to “Fairness_Diversity_Analysis.ipynb” for those results.



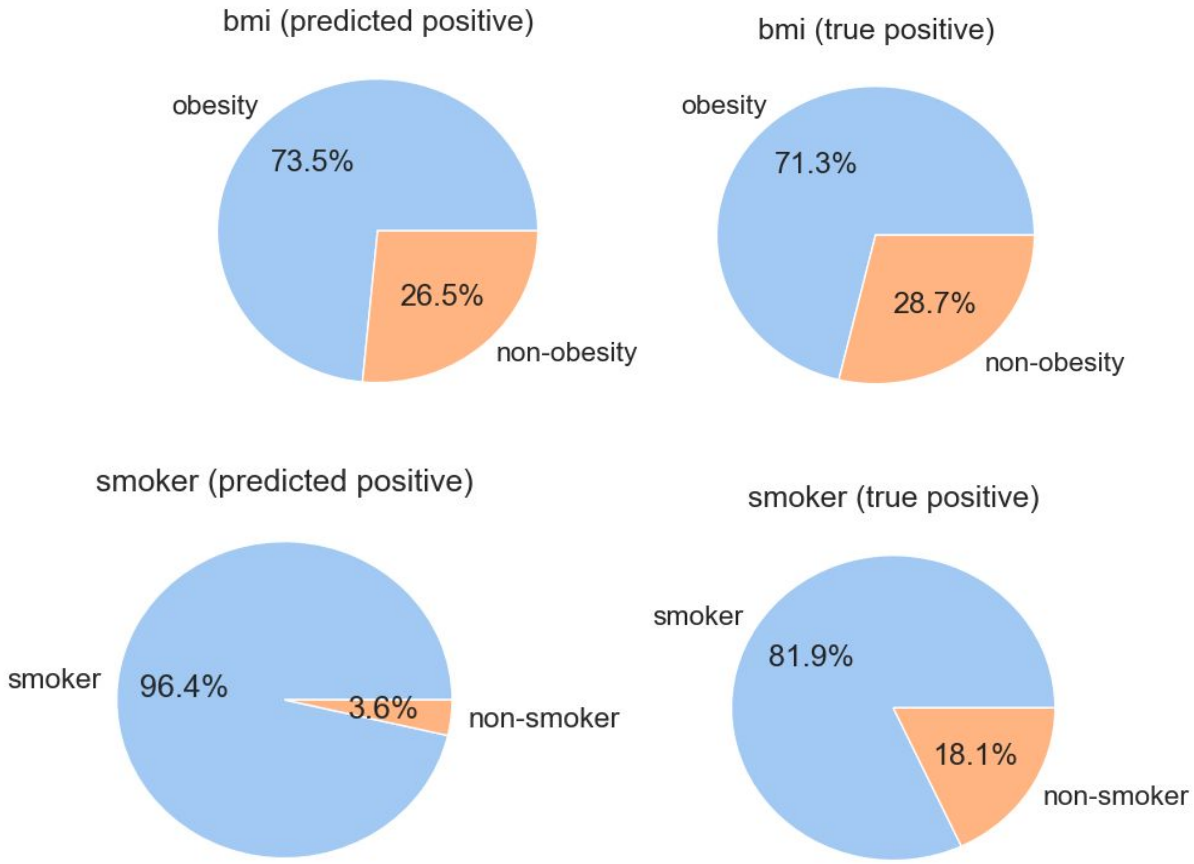
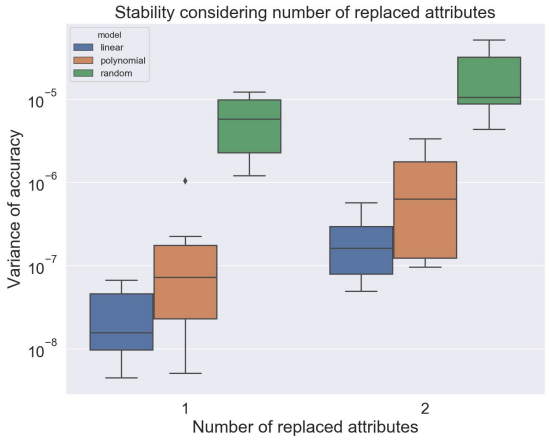


Figure 7. Diversity of 4 attributes using Random Forest model

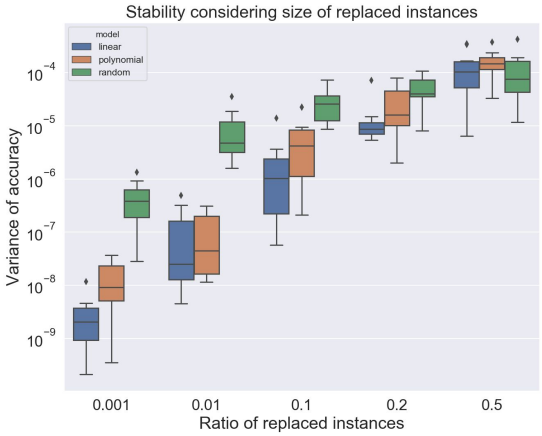
4.4 Stability or Robustness

In this section, we consider the stability or robustness of our ADS given the uncertainty of the input data. Specifically, the uncertainty of the input data is captured by small fluctuations in the training set. For brevity, we focus on the Random Forest model in this section since the Random Forest model is known to have high variance and tend to overfit. However, our methodology can be used to quantify the stability of any ADS.

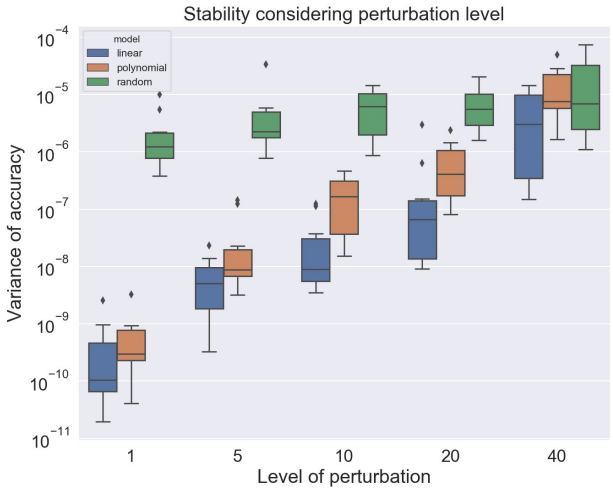
The small fluctuations of the input data are done by changing the value of numerical attributes by a small amount in the training set. We only experiment on numerical attributes in this section in order to quantify the impact of the level of perturbation. We include the perturbation of the categorical attributes in the next section of individual-level explanation. We analyze the impact of three parameters: the number of attributes being tuned, the size of tuned instances, and the level of perturbation, i.e., how much change is done on each instance. We quantify the stability of the ADS by the variance of the ADS' accuracy (mean squared error) with regards to the small fluctuations in the training set. Figure 8 below shows the results of tuning 3 parameters for 3 regression models.



a) Number of replaced attributes



b) Ratio of replaced instances



c) Level of perturbation

Figure 8. Stability of Random Forest regression model

As we expected, the number of replaced attributes in the training data affects the results of regression models more than the other two parameters. We can also observe that, for all 3 parameters, Random Forest model is relatively less robust considering the small fluctuations in the training set. We expect to see this result since we know Random Forest model uses bootstrap aggregation. Linear Regression and Polynomial Regression models have a stable results comparing to Random Forest model considering change of input. However, the variance of all 3 regression models for the fluctuations of input data is low.

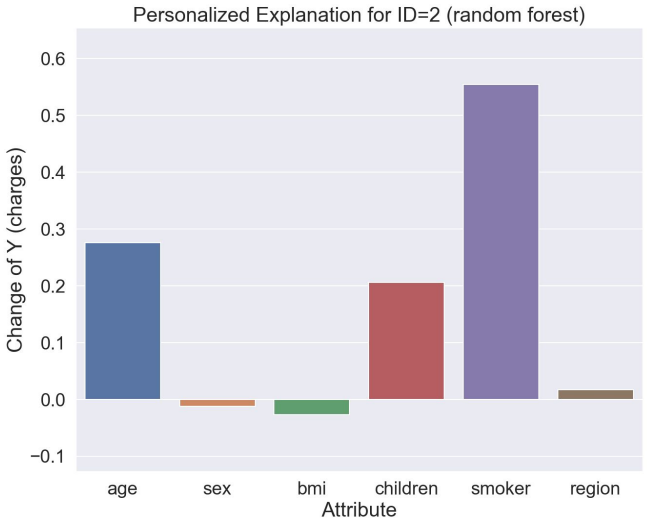
4.5 Explanations of individual predictions

We also developed an individual-level explanation component to quantify the impact of attributes toward an individual's prediction, using the Unary-QII methodology covered in "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with

Learning Systems" by Datta, Sen, and Zick in 2016. Since there are few correlations in our input data, we did not use the correlated QII method. We aimed to discover the most influential attribute toward a positive prediction for each individual and show the distribution of influence of all the attributes for a specific individual. We here show the explanations of two specific predictions in the input data in Figure 9.

```
cur_individual_idx = 2
data = pd.read_csv("insurance.csv")
data.loc[cur_individual_idx, :]
```

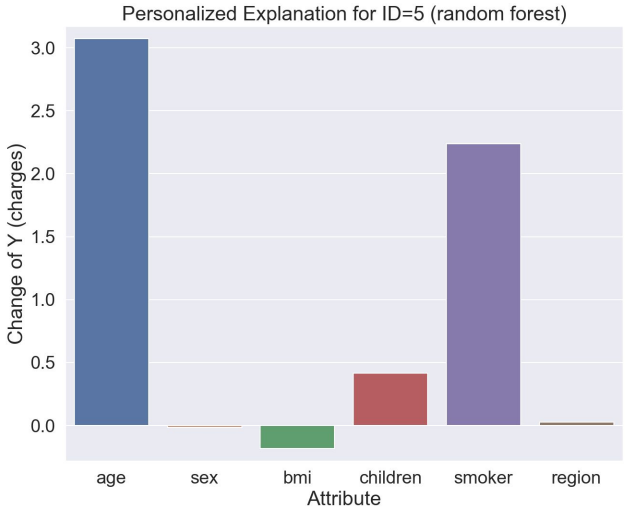
```
age                28
sex              male
bmi                33
children            3
smoker             no
region      southeast
charges       4449.46
Name: 2, dtype: object
```



a) Individual 2

```
cur_individual_idx = 5
data = pd.read_csv("insurance.csv")
data.loc[cur_individual_idx, :]
```

```
age                31
sex            female
bmi             25.74
children            0
smoker             no
region      southeast
charges       3756.62
Name: 5, dtype: object
```



b) Individual 5

Figure 9. The input data and explanations for two individuals

We can observe from the above figure that the most influential attribute toward a positive prediction for individual 2 is his smoking status, with age and children also contributing. For individual 5, the most influential attribute is age, with smoking status also contributing. From

these two examples, it appears that both smoking status and age play a significant role in the ADS' predictions.

5. Summary

The data appears mostly appropriate for this ADS, as it either is or at least imitates a set of customers of an insurance company. However, one possible problem might be that, if the company has had strict health standards on their insurance applications in the past, the dataset might not properly represent the unhealthiest end of applicants. This might cause the system to underestimate an extremely unhealthy person and thus lose the company money by accepting an applicant who should not be accepted.

The ADS' Random Forest model, chosen as generally the best of the three, is very accurate overall and in most subpopulations but fails with respect to the non-obese and non-smoker populations. When comparing PPV, it is reasonably fair across all subpopulations (with the possible exception of non-obese), but its discrepancies in accuracy are somewhat unfair to non-obese and non-smokers, who may be charged inaccurate rates. In examining diversity, the Random Forest model appears to perform quite well on the sensitive attributes of sex and age.

To decide whether or not to deploy this system at the company, we would need to first examine the current application decision process and audit it for accuracy, fairness, and diversity. While this system is imperfect, it is good enough that it very well may be an improvement over the current one. Its high accuracy, fairness, and diversity measures are promising.

It is impossible to critique the data collection methodology without knowledge of that methodology. With respect to the data analysis, we would recommend exploring more models that might fit the data better and attempting hyperparameter optimization.