DS-UA 202: Responsible Data Science
Patricia Guirao, Yunji Sohn and Denise Chow
May 9, 2022

Project Report

Background

In 2018, a graduate student, Mira Choi, hosted "Medical Cost Personal Datasets" on Kaggle, and the purpose of this competition was to use open source developers to create an ADS so that it can accurately predict insurance costs. The ADS we chose to analyze is titled " "Patient Charges || Clustering and Regression", and its purpose is to produce a model that gives an approximation on how much medical charges a patient will incur when receiving treatment. The goal of this specific ADS is to use the variables age, sex, BMI, children, whether the individual is a smoker, and region to predict how much medical care will cost them. Another goal of this ADS is to understand the story behind the patients in this dataset so as to better determine which features should be used in the model to most accurately predict patient charge.

Input

There is only one data file, labeled as insurance.csv, provided. It was uploaded by a GitHub user for Machine Learning purposes. This dataset is in the public domain and is used to predict how much medical charges a patient will incur when receiving treatment. The input data given include (1) age of primary beneficiary, (2) sex, (3) body mass index, (4) number of children covered by health insurance/dependents, (5) whether the beneficiary is a smoker or not, (6) beneficiary's residential area in the US - northeast, southeast southwest or northwest, and (7) individual medical charges by health insurance. The ADS uses all 7 features in its analysis.

1. Data Types:

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

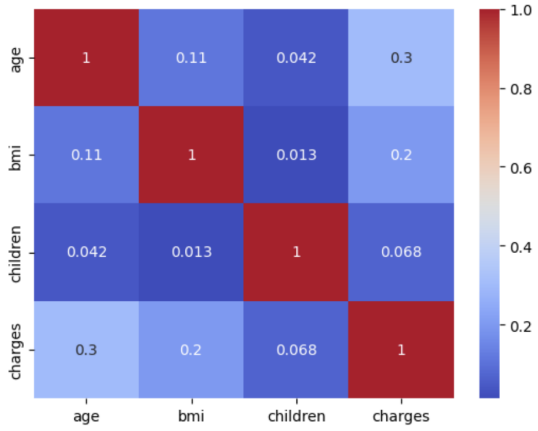
2. Summary Statistics/Data Distribution:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

3. Missing Values - there are no missing values in this data set

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

4. Pairwise Correlations between Continuous Features:



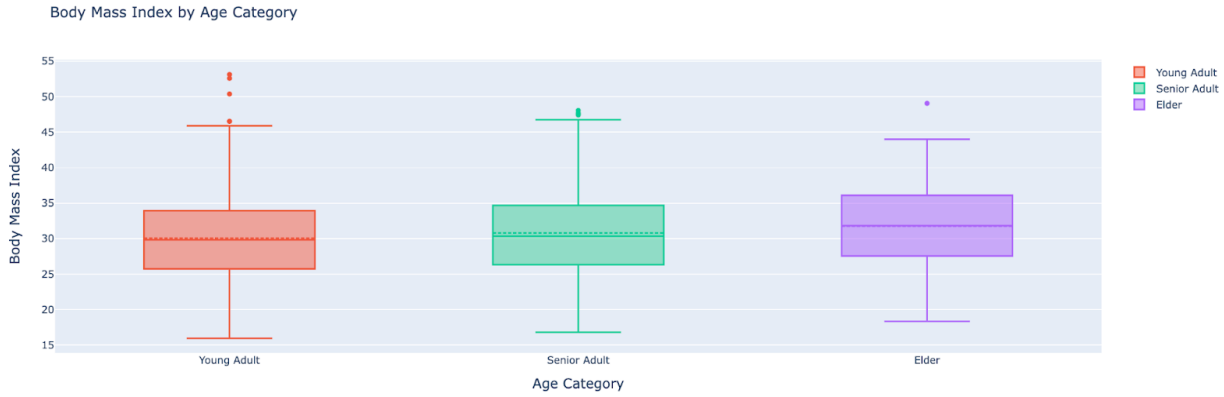
	age	bmi	children	charges
age	1.000000	0.109272	0.042469	0.299008
bmi	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

Output

The output is to predict how much medical charges a patient will incur when receiving treatment, and what factors influence the charge of a specific patient. The ADS uses clustering and multi-regression. The ADS outputs (a) a correlation score between the features to find the relationship between them, (b) outputs p-value to compare the independent categorical variables, (c) outputs a visual distribution of charges to evaluate whether smoking is a big factor for obese people, (d) outputs the most optimal number of clusters using the elbow method to find the optimal number of features, (e) and finally outputs R^2 score to see whether adding features would increase the accuracy or not.

An example of the analyses in the ADS is to find a relationship between BMI and Age. First, they changed age into categorical variables, whereby the young adult category includes anyone between age 18-35, the senior adult category age between 36-55, and the elderly category age above 56. Second, they plotted a normal distribution for BMI and found that most of the BMI frequency is concentrated between 27-33. From the above pairwise correlation matrix, it is also noted that age and charges have a correlation of 0.3, while BMI and charges have a correlation of 0.2. Since the difference in correlation for these two variables with charges is 0.1, age does not have a huge influence on BMI. To further highlight this point, the ADS plotted box plots, which shows for all age categories, the mean and median is similar at around 30-31. The p-value

between these two features is also higher than 0.05, demonstrating that there is no significant change between the age categories when it comes to BMI.



Implementation and Validation

The dataset came already clean and complete as is. To verify this, we checked the data types of each column (and found that each logically matched the column title) and for missing values as well (which we found none).

This system first used unsupervised machine learning methods to identify different groups of patients in our dataset – specifically, the system did manual clustering, K-Means clustering, and hierarchical clustering. Through clustering, we can identify certain kinds of patients that get charged more such as obese smokers. Furthermore, to predict specific charges for each patient, the system fits an OLS Regression to the data, predicting charges based on all the different input variables.

This ADS calculates R-Squared and adjusted R-Squared values to validate the model. Since the R-Squared values are the proportions of the variance in the dependent variable (charges) explained by the predictor variables, the higher, the better. The ADS is validated through trying to maximize the R-Squared values. Therefore, we know that the ADS is meeting its stated goals when the R-Squared value(s) are relatively high.

Outcomes

To determine whether the ADS is fair for different subpopulations, we measured the accuracy and fairness of the ADS by training a baseline random forest model to predict medical charge. We conducted a random forest model for three features in the data provided: age, BMI and sex.

First, we changed the continuous features into categorical variables. For age, we categorized patients into three groups: young adult category include ages 18-35, senior adult category include ages 36-55, and elderly category include ages above 56. For BMI, we categorized patients into four groups: underweight category include BMI less than 18.5, healthy weight category include BMI 18.5-24.9, overweight category include BMI 25-29.9, and obesity category include BMI above 30. For sex, we categorized patients into two groups: male and female.

As noted from the tables below, we found 10 metrics (accuracy, precision, recall, FNR, FPR, FNRD, FPRD, DPR, EOR, and SRD) for overall age, BMI sex. We also found the first 5 metrics for their respective categories/subpopulations (e.g. specific ages ranges like young adults).

Ages:

All Ages	
accuracy	0.932836
precision	0.990909
recall	0.865079
FNR	0.134921
FPR	0.007042
FNRD	1.000000
FPRD	0.333333
DPR	0.000000
EOR	0.000000
SRD	1.000000
dtype: float64	

Young Adults	
accuracy	0.918182
precision	1.000000
recall	0.700000
FNR	0.300000
FPR	0.000000
dtype: float64	

Senior Adults	
accuracy	0.945455
precision	0.980769
recall	0.910714
FNR	0.089286
FPR	0.018519
dtype: float64	

Elderly	
accuracy	1.0
precision	1.0
recall	1.0
FNR	0.0
FPR	0.0
dtype: float64	

Accuracy and precision is relatively high and similar among overall ages, young adults, senior adults and elderly, with values above 0.9. Notably, the elderly population achieved the highest accuracy, precision, and recall rates of 1.0, while having the lowest false negative and false positive rates of 0.0. Recall and false negative rates for overall ages are 0.87 and 0.13 respectively. Senior adults have a recall of 0.91 and false negative rate being 0.089, while young adults exhibit a significantly lower recall of 0.70 and higher false negative rate 0.30. The false positive rate for overall ages is quite low at 0.0070, young adults having an even lower rate of 0.0 and senior adults having a slightly higher rate of 0.019. For overall ages, the false positive rate difference is 0.33. While the false negative rate difference and selection rate difference have a value of 1.0, the equalized odds ratio and equalized odds ratio have a value of 0.0.

BMI: **Note that there were issues with the code for finding 10 metrics for overall BMI. This has been addressed with the professor - we just need to report the first 5 metrics for overall BMI.

All BMI	
accuracy	0.932836
precision	0.991150
recall	0.868217
FNR	0.131783
FPR	0.007194
dtype: float64	

Underweight	
accuracy	1.0
precision	1.0
recall	1.0
FNR	0.0
FPR	0.0
dtype: float64	

Healthy Weight	
accuracy	0.955556
precision	1.000000
recall	0.894737
FNR	0.105263
FPR	0.000000
dtype: float64	

Overweight	
accuracy	0.960526
precision	1.000000
recall	0.909091
FNR	0.090909
FPR	0.000000
dtype: float64	

Obesity	
accuracy	0.923611
precision	0.982456
recall	0.848485
FNR	0.151515
FPR	0.012821
dtype: float64	

In terms of BMI, the accuracy and precision are consistently high for overall BMI, healthy weight, overweight and obesity, with values of around 0.9 and above. The underweight group has the highest values for these variables of 1.0; it also has a recall rate of 1.0. The recall for overall BMI is 0.87, while healthy weight and overweight have a slightly higher value of around 0.9, and obesity has a slightly lower value of 0.85. The false negative rate for overall BMI is 0.13, and the healthy weight and overweight groups have a slightly lower value, while the obesity group has a slightly higher value. The underweight group has a false negative rate of 0.0. The false positive rate for overall BMI is 0.007, and the underweight, healthy weight and overweight groups have a value of 0.0. The obesity group has a false positive rate of 0.013.

Sex:

```
All Genders
accuracy     0.925373
precision    0.973684
recall       0.867188
FNR          0.132812
FPR          0.021429
FNRD         0.030392
FPRD         0.010101
DPR          0.891377
EOR          0.611111
SRD          0.048922
dtype: float64
```

```
Males
accuracy     0.919118
precision    0.979167
recall       0.824561
FNR          0.175439
FPR          0.012658
dtype: float64
```

```
Females
accuracy     0.939850
precision    0.985714
recall       0.907895
FNR          0.092105
FPR          0.017544
dtype: float64
```

For overall genders, females and males, the accuracy and precision are high and relatively with values above 0.9. Recall for overall genders is 0.87, males have a slightly lower value of 0.82 and females have a slightly higher value of 0.91. The false negative rate is 0.13 for overall genders, females have a slightly lower value of 0.09 and males have a slightly higher value of 0.18. The false positive rate is 0.21 for overall genders, while females have a lower value of 0.018 and males have the lowest value of 0.013. The FNRD, FPRD, DPR, EOR and SRD for overall genders are 0.03, 0.01, 0.89, 0.61 and 0.049 respectively.

Analysis:

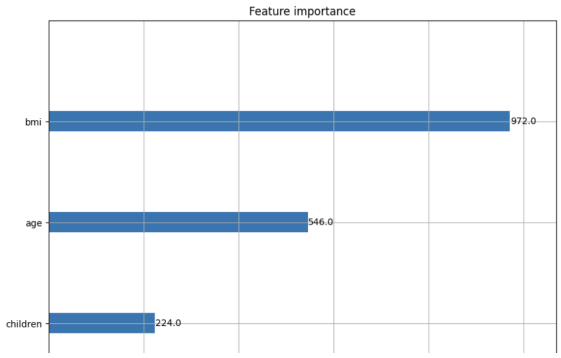
Based on the outcomes above, we can conclude that data produced in the overall data for sex, BMI and gender is relatively consistent and fair with their respective categories. We can also conclude that all three of these features have high accuracy, precision and recall scores. This means that the model is performing well and is able to correctly predict medical charges with a high degree of accuracy. Precision is the proportion of true positives out of all the instances that the model classifies as positive. High precision indicates that the model is able to correctly identify the positive instances with a low rate of false positives. Recall, also known as sensitivity, is the proportion of true positives out of all instances that are actually positive. High recall indicates that the model is able to identify most of the positive instances in the test set.

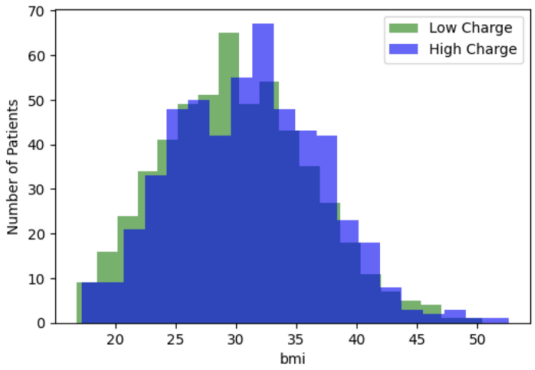
In general, the false negative and false positive rates are relatively low for all three of these features. A low false negative rate means that the model is able to correctly identify most of the positive instances in the test set. A low false positive rate means that the model is able to correctly identify most of the negative instances in the test set. This suggests that the model is performing well and is able to correctly classify instances with a high degree of accuracy, while keeping the rate of incorrect classifications low.

The false negative rate difference and selection rate difference have a value of 1.0 for overall ages. These high scores indicate that the model is very likely to miss positive instances for one demographic group compared to another, which can lead to unfairness in the model's predictions. Demographic parity ratio and equalized odds ratio are both 0.0 for all ages, indicating that the model is achieving perfect fairness and there is no difference in the outcomes or prediction errors between demographic groups. This is contradictory so we have to conduct more analysis to

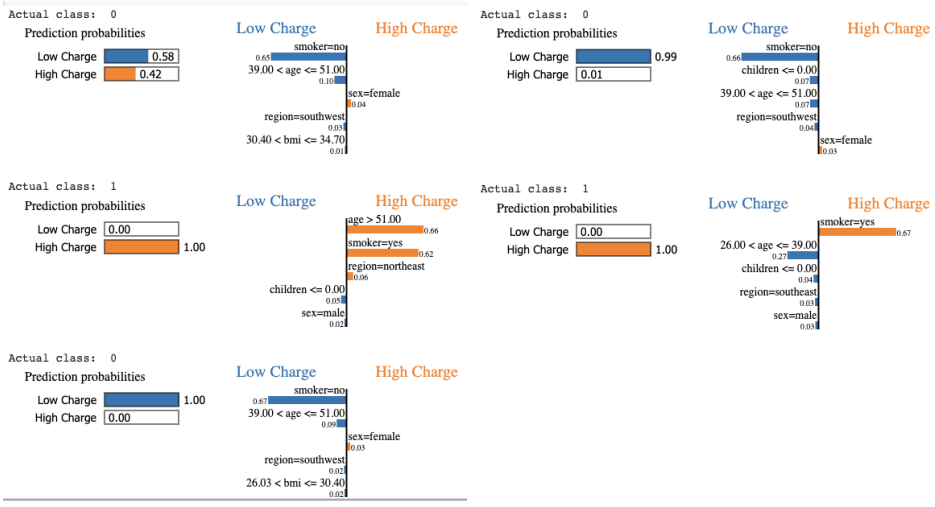
determine any underlying biases that make the model produce these results. On the other hand, FNRD, FPRD and SRD are a lot lower for all genders, EOR is slightly higher and DPR is significantly higher.

Next, we used LIME to predict which features are most influential in the prediction of medical charges, and this is useful for identifying potential biases in the model. From the graph below, we see that BMI has significantly the highest importance, then age and then children. Since BMI is the most influential factor of medical charges, we analyzed it more. From the histogram below, people with a BMI between 32-35 are charged the highest, whereas people with a BMI between 27-30 are charged the lowest.





Lastly, we looked at individual cases, and based on the prediction probabilities below, age and smoking are features that predict high charge the most, and the smoker feature predict low charge the most.



Summary

To answer this question, it is important to address that the purpose of this Kaggle competition is for Machine Learning purposes, so the data provided might not accurately reflect the real-world.

If it were not for learning purposes but rather accurately predicting real-world medical charges, this data would not be appropriate as it was uploaded from a random GitHub user, and thus might not be reliable. This data is seemingly too perfect as it has zero missing values, which is near impossible if it were to reflect the real-world. Moreover, if it were to truly reflect the real-world, the data should include more features and patient records as there are definitely more variables that could potentially influence medical charges.

However, since this Kaggle competition is for Machine Learning purposes, this data is appropriate for this ADS, which serves to predict patients' medical charges. Based on six different features, it shows how patients' charges could be different. While there are very few variables, we believe it contains the necessary ones that could affect medical charges the most and thus could be used to predict the outcome. However, including the type of health insurance coverage of the patient, the type of chronic medical condition(s) like diabetes or asthma, and the complexity of the medical procedure/treatment received would be useful in accurately predicting medical charges. Going more in depth with the first example, patients who have private health insurance plans may have different coverage limits or co-payments compared to those who are covered only by government-funded health insurance such as Medicare. Patients who do not have any health insurance coverage may have to pay out-of-pocket for medical services, which can be a significant burden on their finances. Therefore, while this data was appropriate for this ADS as it was for Machine Learning purposes, it would not be appropriate if it was predicting real-world medical charges as the data does not accurately reflect society and more features should be provided to make better predictions.

We found that the implementation lacked the necessary robustness, accuracy, and fairness to ensure reliable results, and therefore, we made adjustments and included additional models. First, to enhance the system's robustness, we recommend testing it with a diverse set of inputs and simulating errors to gauge its behavior under unexpected conditions. As only seven features were provided with no missing values in the data, the inclusion of random errors in the input data and seeing how they affect patients' charges could improve its robustness. Regarding accuracy, the ADS relied on R-squared to validate the model, but we believe that testing it with a more comprehensive range of scenarios and edge cases could expose any potential biases or inaccuracies in the system. Validation techniques such as cross-validation or bootstrapping could help determine the system's stability and reliability across various data subsets. Although the ADS compared output for different demographic groups and conducted OLS, we suggest using more fairness metrics like disparate impact, statistical parity difference, or equal opportunity difference to improve its overall equity. By doing so, we can ensure that the system does not perpetuate any discrimination or bias and treat all patients fairly.

There are three major groups of stakeholders that could find these measures appropriate. First, patients are directly impacted by medical charges and are responsible for paying for their healthcare expenses. They are the main decision-maker in terms of their treatment plans and medical costs. By ensuring that the charges they were charged are fair, patients can trust the

healthcare system and be assured that they're receiving fair and equitable treatment. Second, healthcare providers, such as hospitals and clinics, play a significant role in determining the cost of medical services. They set the prices for their services. By accurately predicting patients' medical charges, they can better manage their finances and allocate resources for their patients' needs. Third, the government provides government programs/coverage, such as Medicare, for eligible individuals and sets reimbursement rates for medical services. They may also regulate healthcare pricing. Moreover, they can identify areas where improvements are needed and address the policies.+

We would not be very comfortable deploying this ADS in the public sector or in the industry as we feel that it did not provide enough testing to ensure robustness, accuracy, and fairness. Although the system did different types of clustering to identify certain kinds of patients that get charged more, as well as fitting an OLS Regression to the data, we feel that it needs to do more to ensure equity. For example, it could include random errors in the input data to ensure that the outcomes will remain similar, if not the same, despite these additions. Real-world data and conditions are often not this clean and straightforward – therefore, having those conditions simulated in the ADS and then further accounted for to assure maintenance of accuracy and fairness would first be necessary before considering the deployment of this ADS for real-world use.

Regarding data collection, we suggest adding more variables to the dataset such as the medical history of patients and the severity of their illness or injury to improve the accuracy of medical charge predictions. Patients with a history of chronic conditions or previous medical procedures may require ongoing treatment or monitoring, leading to higher costs. Moreover, the severity of the illness or injury is crucial as more severe cases usually require more extensive treatment and monitoring, resulting in higher charges. For example, patients with major fractures may require surgery, hospitalization, and ongoing physical therapy, which are more expensive than patients with minor fractures. Additionally, as discussed in part (a), including information about the type of health insurance could also aid in predicting medical charges since private health insurance plans, government-funded health insurance programs, and no insurance coverage can have a significant impact on the overall cost of medical care.

To improve data processing and to ensure that the ADS can be used in the real-world and not just for Machine Learning purposes, data cleaning should be implemented. This includes dealing with irrelevant data, such as duplicates, missing values, and outliers, and transforming the data into a format that can be used by the ADS. Another improvement would be to use more validation techniques like cross-validation or bootstrapping to validate the ADS to ensure that it is accurate and fair. It is also important to use multiple metrics to evaluate the ADS, such as false negative and positive rates, precision, recall and more. Improvements that can be made to the data analysis methodology of the ADS would be feature selection; it is important carefully the features that are most important to train the model, which we did through LIME. Another improvement would be to have more performance metrics, which we included. While the ADS

gave us the R-Squared and adjusted R-Squared, it is important to consider other performance metrics such as precision, recall, and F1 score, which provide a more nuanced view of the performance of the model. Our group has used LIME and added 10 metrics for overall features and 5 metrics for categories of each of these features to improve the data analysis.