MIS S381N Group 27: Data Science Programming project

Qingye Ding
Rianna Patel
Shehzad Ali
Trevor Moos
Yashpreet Kaur

Project Goal:

Description:
The project is aimed at forecasting individual medical costs billed by health insurance providers based on six predictors. A brief summary of the variables are given below:

Variable	Variable description
Age	Age of primary beneficiary
Sex	Insurance contractor gender, female, male
BMI	Body mass index, a height-to-weight ratio (kg / m ^ 2)
Children	Number of dependents
Smoker	Does the individual smoke?
Region	The beneficiary's residential area in the US: northeast, southeast, southwest, northwest.
Charges	Individual medical costs billed by health insurance providers

The dataset consists of 1,337 individual data points with no blanks or erroneous values for any of the variables.

Importance of problem:

The global health & medical insurance market is expected to grow from $385.24 billion in 2020 to $390.54 billion in 2021 . The market is expected to reach $653.4 billion in 2025 . The healthcare industry in the US contributes to approx. 18% of the total GDP of the US. The United States health insurance industry continued its tremendous growth trend as it experienced a significant increase in net earnings to $31 billion and an

increase in the profit margin to 3.8% in 2020 compared to net earnings of $22 billion and a profit margin of 3% in 2019. Additionally, enrollment for health insurance in the US increased 4% (9 million) to 240 million in 2020 compared to 2019. The average national cost for health insurance in the USA is $456 for an individual and $1,152 for a family per month.

The health & medical insurance market is one of the most prominent markets in the US economy and there is increased importance of the insurance industry in the post-COVID era. Forecasting individual medical costs billed by health insurance providers could help both the insurance seekers and the insurance providers to evaluate their cost and revenue streams respectively.

Exploratory Analysis:

We classified the continuous variables BMI, age, and medical charges. The CDC has four categories for BMI and we put our data into the same buckets. Age was divided into the following five sets: 18-29, 30-39, 40-49, 50-59, and 60-65. When grouping medical charges, we noticed that around the 70% mark of sorted charges, there was an exponential increase. Therefore, we grouped the top 30% of charges as high-cost individuals and the bottom 70% as normal-cost individuals.

First, we analyzed the relationship between each variable and charges individually. The variables that instantly seemed significant were age, BMI, and smoking. Healthy and overweight individuals paid a similar amount in medical charges, but there was a 41% jump in medical costs for a BMI obese individual compared to a BMI healthy individual. We were not surprised to see medical costs increase as you get older, but it was interesting to learn that being obese/overweight exponentially increases your medical costs as you age. Refer to figure 1 in the appendix for a chart displaying this jump. Next, we turned to analyze the smokers in our dataset. Twenty percent of our dataset smoked. These individuals paid around fourfold what a non-smoker paid in medical charges. Figure 2 in the appendix is a growth that displays the massive increase in medical costs for smokers and non-smokers by BMI group. Smokers who are underweight, healthy, or overweight pay around triple of what their non-smoking counterparts paid. When all else is held equal, obese individuals who smoke spent 369% more than non-smokers. Smoking has an immense impact on medical costs. The difference in average medical charges by region was interesting until we noticed that this was more reflective of the demographics of each area, such as the number of smokers.

Solutions and Insights:

Method:

We ran regression models to predict the exact value of medical costs for an individual, and classifier models to predict if a consumer belongs to a high-cost or normal-cost group.

Linear Regression:

We created 3 linear regression models: with all variables, all important variables, and with partial important variables, respectively.

Model 1 with all features has an AIC of 18960.68. The R-squared value is 0.755. Observing p-values, we identified the following important features: smoker, age, bmi, and children. The result implies that Model 1 contains irrelevant variables such as sex and region. Based on this finding, we created a second model, which only includes important features.

Model 2 contains only important features: smoker, age, bmi, and children. AIC is 18959.63, and R-squared value is 0.753. Model 2 achieved the lowest AIC so we believe it is the best model.

We also conducted an experiment of removing a statistically significant feature, children. This model uses only 3 predictors--age, bmi, and smoker. Removing children reduced the R-squared value of model 3 to 0.749, increased AIC to 18968.59. The increase in AIC indicates that the variable 'children' plays an important role in the accuracy of the regression model.

All three models are very similar, but we chose model 2 as the best model because it is the simplest.

Logistic Regression:

With crosstab matrices we found that age didn't have a significant effect on the cost of an individual except for ages 60-65, which were more likely to be high cost. This aspect wasn't surprising as many individuals develop health issues with age. Additionally, both females and males are approximately equally as likely to be high or low cost, showing that sex does not have a significant affect. Our analysis of BMIs, number of children, and region of residency led us to see that these variables have a weak relation to the cost of an individual. Contrastingly, we found that 99.9% low cost individuals are non-smokers, whereas 66.7% of high cost individuals are smokers. This is not surprising since it is widely known that smoking causes many detrimental illnesses, such as cancer and heart disease. After splitting the data set into training and testing sets, we created a model to fit the classifier with 92.5% accuracy. The variable weights given from this model show that smoking and the 60-65 age category have the highest positive weights of 5.8 and 1.97, respectively, whereas the 30-39 and 40-49 age categories have the highest negative weights of -0.92 and -0.63, respectively.

Random Forest Regressor:
The random forest regressor uses four features on each split. The most important variable is smoker, followed by BMI and age. The RMSE for this model is 4,689. The RMSE for RF is lower than the RMSE for linear regression, making it the best regression model.

Random Forest Classifier:
The random forest classifier was run on all of the variables after the bucketing for BMI and Age was applied. The model was able to predict with about 90% accuracy and found 'Smoker' to be the most important variable by far. In figure 3, we can see how dramatically more important 'Smoker' is at predicting than any other variable. It is surprising to see how factors like Age 60-65 and Obese BMI pale in importance compared to 'Smoking'. Every other factor has less than a 5% importance, and Overweight, Underweight, and all of the age categories from 18-49 have a less than 1% importance. It is extremely surprising to see age matter so little, and Overweight to also be so unimportant.

Naive-Bayes:
This classifier uses every variable and possible answer as a predictor. For example, our formula includes children_0,children_1, children_2, children_3, children_4, and children_5 as predictors. The classifier has a 90.7% test accuracy rate which is second only to the logistic regression. It was clear to us that being a smoker was the most impactful factor when predicting medical cost classification, but the smoking_yes importance variable was almost four times the size of the second most important variable. The regions were not an important feature in the Naive-Bayes classifier.

Conclusion and Discussion:

The best regression model is the Random Forest because it has the lowest RMSE of 4,689. The best classifier model is the Logistic Regression model with a test accuracy of 92.5%. The insurance company can refer to the Random Forest regression model while trying to estimate their revenue streams. They could use the Logistic Regression classification model during customer segmentation phase to come up with appropriate insurance-marketing strategies for each segment. Our finding is that people who typically have high medical costs are older, fall into the BMI obese category, and smoke.

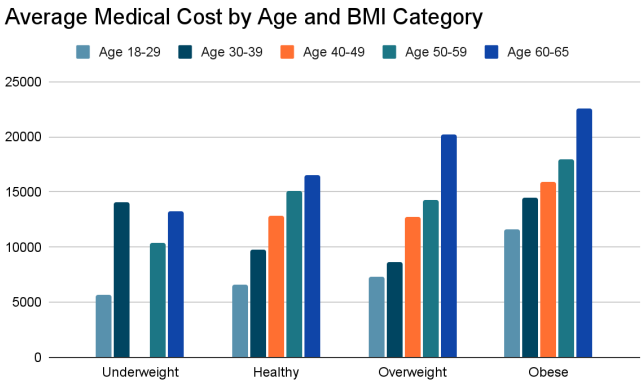
Appendix:
Table Summary:

Table 1:

	Regression Models		
	Linear Regression	KNN	Random Forest
RMSE	6,253	9,758	4,689

Table 2:

	Classification Models		
	Naive Bayes	Random Forest	Logistic Regression
Test Accuracy	90.7%	89.9%	92.5%



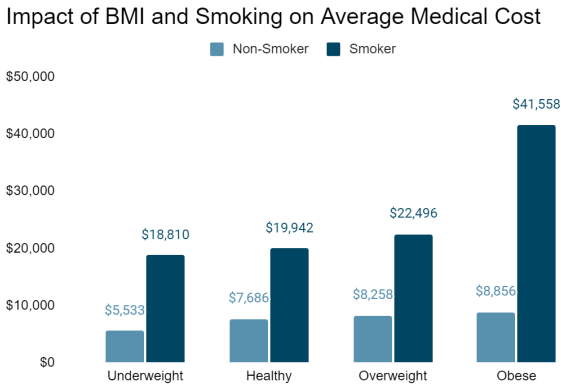
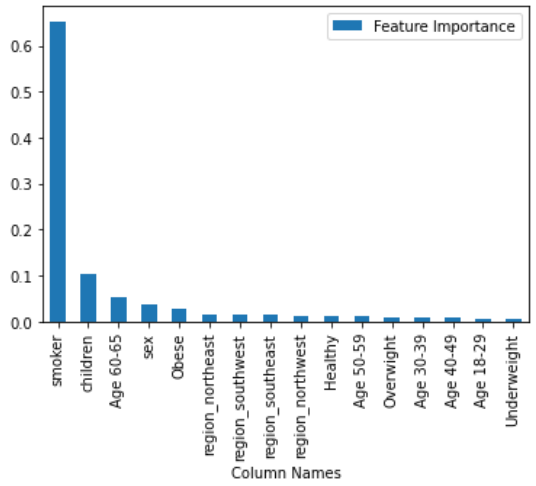


Figure 1: Obesity/Overweight individuals have a much larger jump in medical costs as they age	Figure 2: Smoking triples medical costs for underweight, healthy, and overweight individuals, but increases by 4.69x for obese individuals.



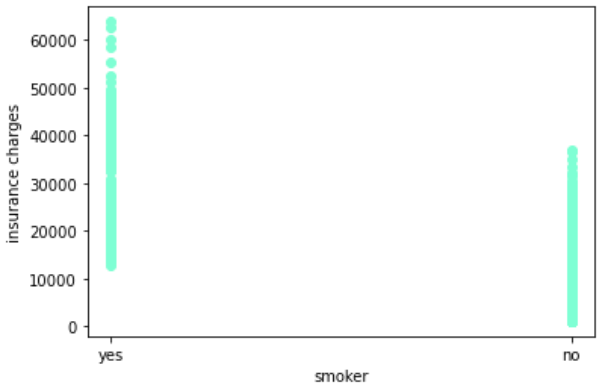
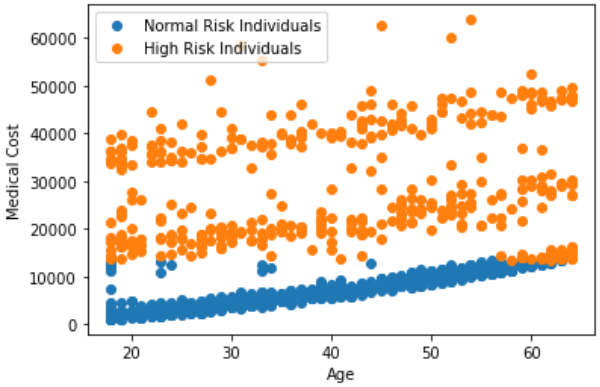


Figure 3: Variable importance graph from Random Forest Classifier	Figure 4: Smokers have higher medical costs than non-smokers



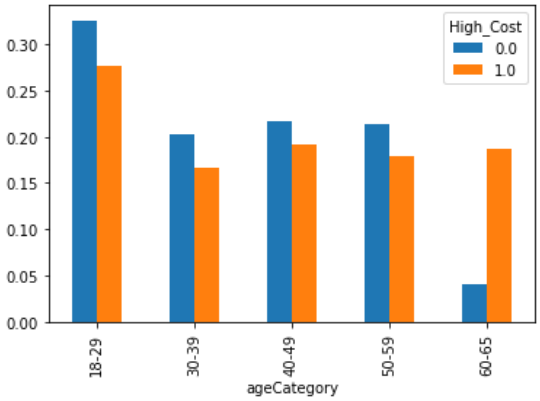
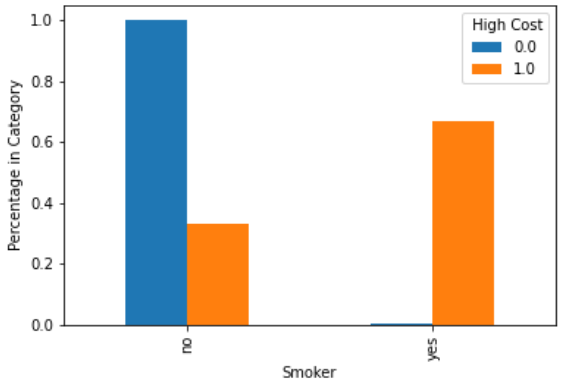


Figure 5: Medical costs increase as you get older and a large percentage of people over the age of 65 are high cost individuals.	Figure 6: A bar graph showing the percentage breakdown of each age group in either high or normal cost category.



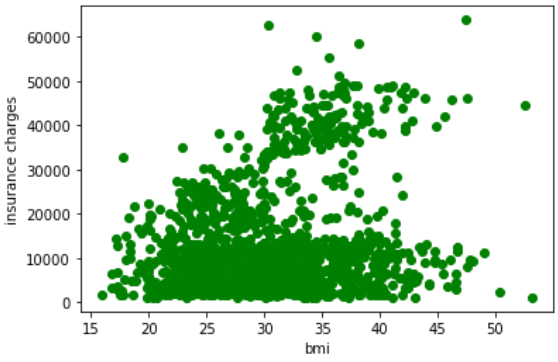
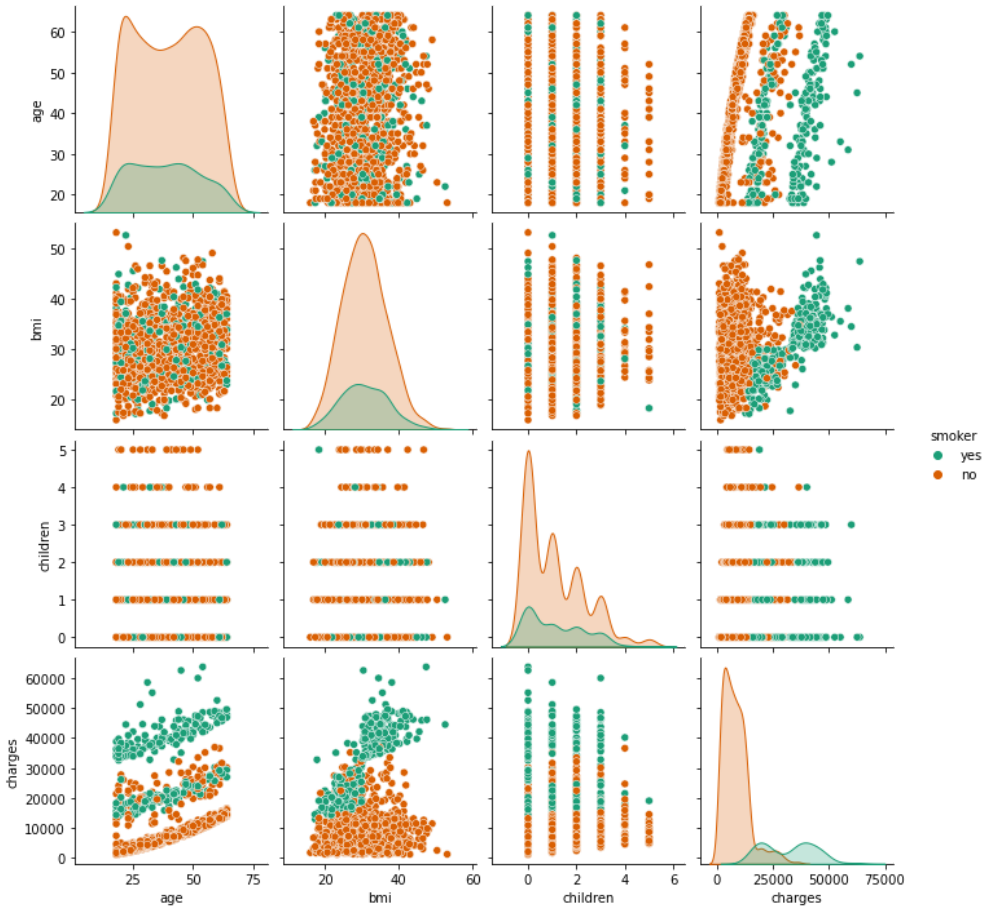
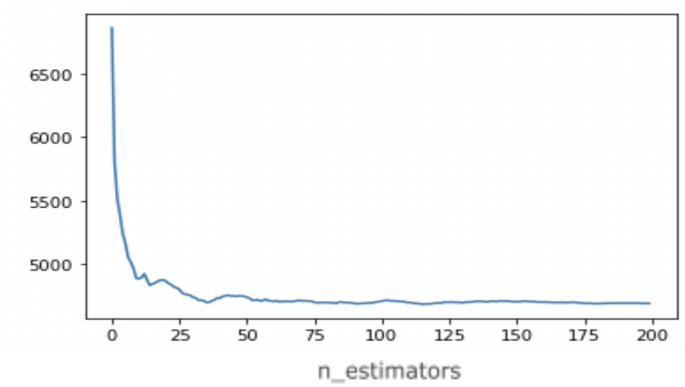


Figure 7: Not all high-cost individuals smoke, but almost every individual who smokes is a high-cost individual	Figure 8: There are people with high BMIs who do not have high medical costs. These are the people who do not smoke. This displays the impact of smoking as your BMI increases.

Figure 9: This is a pair plot that allows you to see the relationship between the important variables and the impact of smoking on all of them.





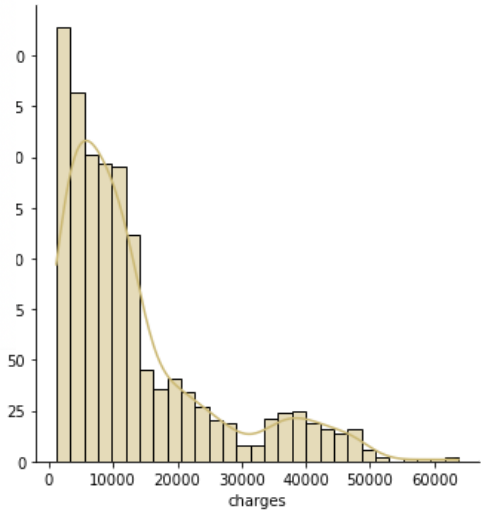


Figure 10: This graph from the random forest regression allows us to pick the number of trees by comparing the number of trees versus the RMSE. After 50 there is very minimal change in RMSE.	Figure 11: This is a histogram of the medical charges. It has a long tail and this chart influenced us to choose the 70-30 method for categorization instead of a qcut. The 70% mark of charges is at the 15,000 mark so you can see how wide the top 30% is versus the bottom 70%.