Michael Gonzalez
DSC 690 Winter 2020
Professor Williams

Can insurance providers change rates base on your data?

Abstract

This project will look at anonymous health insurance data from a Kaggle competition. This data will be used in developing a predictive model to see if it possible to find features that can either raise or lower insurance coverage rates.

Background

The selected dataset comes from the Kaggle competition website that includes health insurance data that removes the identities of people who gave their consent on being included in this dataset. Below is a link to the dataset: https://www.kaggle.com/bmarco/health-insurance-data

I want to find out how it could be possible to predict new insurance rates based on features in this dataset? When finding these features who should be responsible to make decisions on insurance rates? Analyzing this dataset will give me a better experience developing prediction models. This type of predictive model could be useful in predicting accurate insurance rates. The only issue is with this dataset is that it has a limited amount of data. This will hinder the scalability of this predictive model in the future of this project.

The Walk Through

Starting with the dataset from the Kaggle website that was based on health insurance data, which was in a CSV format. I decided to use the Pandas package to read the dataset. I have also used the Seaborn, Matplotlib and Numpy packages to create data visualizations that will help in answering my research questions. Going back to the dataset, I want to talk about it and give you the reader a better understanding of it. This dataset has seven variables and 1,338 rows of data that includes Age, Sex, BMI, Children, Smoker, Region, and Charges. This dataset is great to use for developing a model that will be able to predict accurate insurance rates.

Data Attributes and Dummy Variable Encoding:

The following code was used to get the information about the data frames in terms of the different columns present in this dataset.

```
In [5]: # Checking the data type of the attributes

        df.dtypes
```

```
Out[5]: age           int64
        sex          object
        bmi         float64
        children      int64
        smoker       object
        region       object
        charges     float64
        dtype: object
```

I am going to use dummy variable encoding to numerical values and I will display the first five rows.

	age	bmi	children	charges	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
0	19	27.900	0	16884.92400	0	1	0	0	1
1	18	33.770	1	1725.55230	1	0	0	1	0
2	28	33.000	3	4449.46200	1	0	0	1	0
3	33	22.705	0	21984.47061	1	0	1	0	0
4	32	28.880	0	3866.85520	1	0	1	0	0

Missing Data Check:

One final thing is to check for any missing data. It is very clear to see that there is no number of any null values. It is time to proceed with the model fitting process of this project.

```
In [25]: #Checking for missing values in the dataframe

         ins.isnull().sum()
```

```
Out[25]: AGE         0
         SEX         0
         BMI         0
         CHILDREN    0
         SMOKER      0
         REGION      0
         CHARGES     0
         dtype: int64
```

Model Fitting

Given this will be a prediction problem, all the chosen models are going to regression models. The first model is a multiple linear regression. The next model well be a polynomial regression to see if I will be able to see any patterns that did not show up in the first model. The third model is a random forest regression model.

Splitting the data for the three models:

Extracting variables into the x and y

In [4]:
```
# Extracting independent variables into X
# Extracting the dependent variable into Y

X = df.iloc[:, [0,1,2,4,5,6,7]].values
y = df.iloc[:, 3].values
```

Splitting data into the trainning and testing sets

In [6]:
```
# Splitting the dataset into the Training set and Test set
import sklearn
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

Model Fitting (Multiple Linear Regression)

Fitting data to MLR and the accuracy rate:

Fitting Multiple Linear Regression on the trainning set

In [10]:
```
#Setting up Multiple Linear Regression to the Training set
import sklearn
from sklearn.linear_model import LinearRegression

regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

Out[10]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Extracting predictions and checking R squared values

In [11]:
```
# Extracting the predictions
y_pred = regressor.predict(X_test)

# Checking r square value
import sklearn
from sklearn.metrics import r2_score
score = r2_score(y_test, y_pred)
score
```

Out[11]:
```
0.7984171871612084
```

The accuracy was about 79.8% using the multiple linear regression model. This shows that the accuracy rate is a good start. This is a good baseline to compare with the other two predictive models.

Model Fitting (Polynomial Regression)

Fitting data to PR and the accuracy rate:

Polynomial Regression Model

In [21]:
```
#Making a polynomial Regeression Model
import sklearn
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression

poly_reg = PolynomialFeatures(degree = 4)
X_train_poly = poly_reg.fit_transform(X_train)
lin_reg_2 = LinearRegression()
lin_reg_2.fit(X_train_poly, y_train)
```

Out[21]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Finding the R squared values of the y_test and y_pred

In [24]:
```
#Finding the R squared values of y_test and y_pred
import sklearn
from sklearn.metrics import r2_score

score = r2_score(y_test, y_pred)
score
```

Out[24]:
```
0.8536939725842647
```

Comparing this model with the multiple linear regression model, the accuracy is higher than the latter. This improvement is only about 5% , but it is a clear distinction between the two models.

Model Fitting (Random Forest Regression)

Fitting data to RFR and the accuracy rate:

Random Forest Regression Model

In [7]:
```
#Making Random Forest Regression fit to the dataset
import sklearn
from sklearn.ensemble import RandomForestRegressor

regressor = RandomForestRegressor(n_estimators = 3, random_state = 0)
regressor.fit(X_train, y_train)
```

Out[7]:
```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=3, n_jobs=1,
           oob_score=False, random_state=0, verbose=0, warm_start=False)
```

Finding the R squared values of the y_test and y_pred

In [9]:
```
#Finding the R squared values of y_test and y_pred
import sklearn
from sklearn.metrics import r2_score

score = r2_score(y_test, y_pred)
score
```

Out[9]:
```
0.8483823111786635
```

From the accuracy of the random forest model, it is like the polynomial regression model. Since these two models had higher accuracy rates, when compared with the other predictive models. I am confident in using either the polynomial regression or the random forest predictive model.

Conclusion

The focus of this project was to see if it is possible to use health insurance data to predict insurance rates based on features. Developing and testing three predictive models, the accuracy of the models was good. Only two of the three models had higher accuracy rates. The Polynomial Regression model was the best model that has had the highest accuracy rate. The Polynomial Regression model's accuracy was 85% and

it will be able to predict accurate insurance rates. The accuracy rate could have been higher if the dataset had more data. Other than that, this is a great outcome for this project.

References

[1] Kirchner, T. How machine learning is revolutionizing insurance pricing at AXA. November 15, 2018. https://digital.hbs.edu/platform-rctom/submission/predicting-your-casualties-how-machine-learning-is-revolutionizing-insurance-pricing-at-axa/ (Information on predicting insurance rates)

[2] Sato, K. Using machine learning for insurance pricing. March 29, 2017. https://cloud.google.com/blog/products/gcp/using-machine-learning-for-insurance-pricing-optimization(Information on using machine learning to predict insurance pricing)

[3]Spedicato, G, Dutang, C. & Petrini, L. Machine Learning Methods to Perform Pricing Optimization. N.D. https://www.variancejournal.org/articlespress/articles/Machine-Spedicato.pdf (Article on machine learning methods)

[4] DataRobot. Insurance Pricing. N. D. https://www.datarobot.com/use-cases/insurance-pricing/ (Article on insurance pricing)

[5] Prianda Galih, B. Prediction of Health Insurance Costs with Linear Regression using Python. December 24, 2018. https://medium.com/@BAYUGALIH/prediction-of-health-insurance-costs-with-linear-regression-8fd95a905a40 (Article on health insurance prediction with python)

[6] Forbes. Can AI Cure What Ails Health Insurance? February 11, 2019. https://www.forbes.com/sites/insights-intelai/2019/02/11/can-ai-cure-what-ails-health-insurance/?sh=3b86ee952d59 (Article on using AI on health insurance)

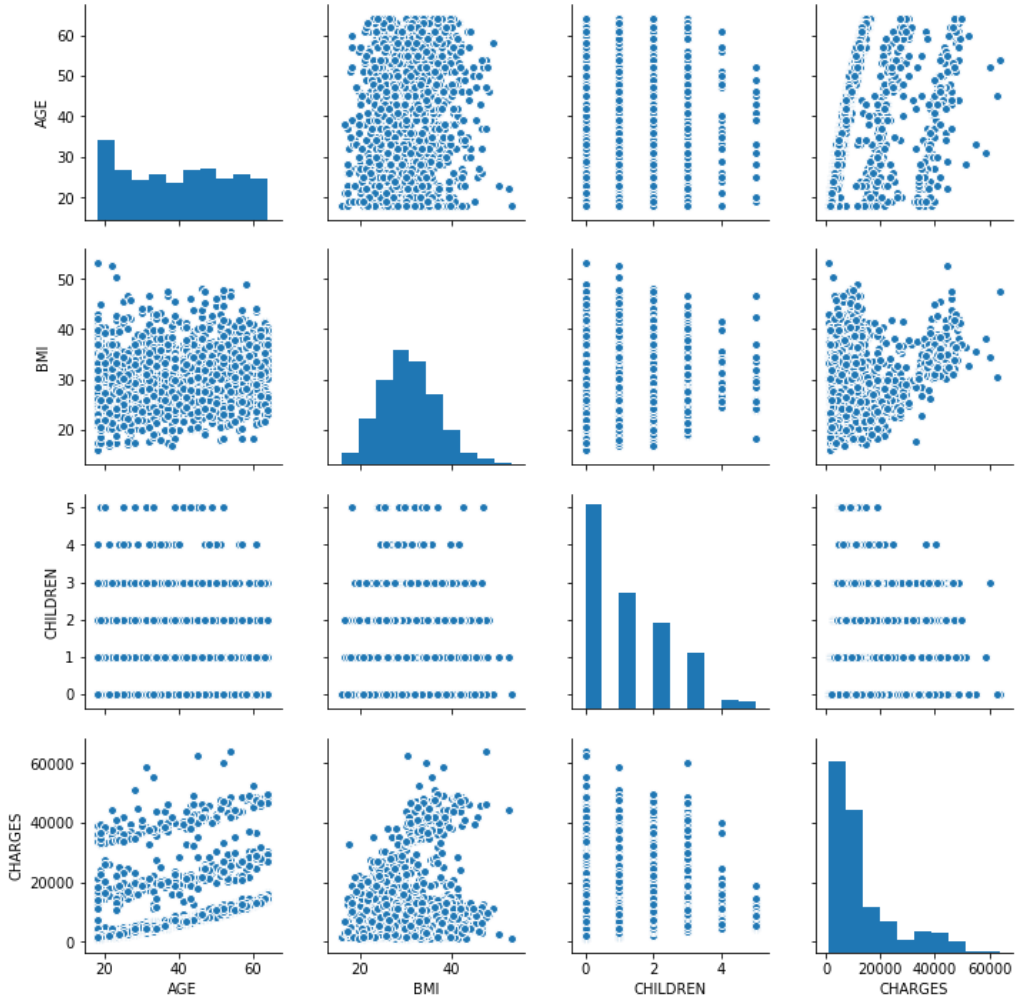
[7] CalHealthNet. Playing fortune teller with California health insurance. N. D. https://www.calhealth.net/Predicting_Future_of_California_health_insurance.html (Article on California health insurance predictions)

[8] Yohn, A. 11 Ways Predictive Analytics in Insurance Will Shape the Industry in 2021. N. D. https://www.duckcreek.com/blog/predictive-analyitics-reshaping-insurance-industry/ (Article on predictive analytics in health insurance)

[9] National Conference of State Legislatures. Health Insurance: Premiums and Increases. December 04, 2018. https://www.ncsl.org/research/health/health-insurance-premiums.aspx (Article on health insurance rates)

[10] Allen, M. Health Insurers Are Vacuuming Up Details About You. July 17, 2018. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates (Article on health insurers use of machine learning)

Appendix



This is a pair plot of my exploratory data analysis process of the health insurance data. This was a preliminary step in looking for any trends or outliers. This pair plot only displays a few of the variables like age, BMI, children, and charges.

10 Questions

1. What made you want to do a project on this topic?
2. Were there any issues finding any research materials?
3. Do you think health insurance companies are collecting too much information?
4. Do you think the predictive models could be used on larger datasets?
5. Did you have to make any adjustments to your dataset for the insights?
6. Why is the dataset anonymous?
7. Did you find any correlations with any of the dataset's variables?
8. Why are insurance companies looking for ways to charge more for customers?
9. Did you find any other uses for these predictive models?
10. Are there any areas of your project that you would like to improve upon?