THE UNIVERSITY OF MEMPHIS

Health Insurance Premium at PSI

Vinay Badam, Nikhil Banka

[/100. -15: Presn. -10: Goal/S. -30: LitRev. -15: Case. -30: Dels.]

Goal/Scope

[-2: Popn? -2 Problem Defn? -2 Goal? -4 Typical Q+As?]

- Population
  - Population are the customers who are using Health Insurance premiums at PSI gathered from Kaggle.
  - Considered 1338 observations and 7 predictors.
- Problem Definition
  - Health insurance is a contract between a company and a consumer. The company agrees to pay all or some of the insured person's healthcare costs in return for payment of a monthly premium.
- Goal
  - The goal of this project is to predict the health insurance premiums of customers of PSI six months in advance.
- Typical Q+As
  - Instance: In the pandemic Covid-19, every individual chooses fewer Deductibles by paying high premiums
  - Solution: By taking precautions, maintaining social distancing, and eating healthy food can be prevented from Covid-19 up to some extent. Decrease in premium price.

[1] https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

[2] https://www.kaggle.com/datasets/hhs/health-insurance.

Lit Review

[-5: Tech defs? -12: Findings? -10: Data Corpus? -3: Refs?]

- Tech defs
  - Health insurance or medical insurance is a type of insurance that covers the whole or a part of the risk of a person incurring medical expenses. As with other types of insurance, the risk is shared among many individuals.
- Findings
  - it was able to select the features such as copays for drugs and hospital visits that make sense in determining the price of health insurance premiums
  - After using different machine learning algorithms, I found the Random forest algorithm as the best-performing algorithm for this task.

[3] Authors4 (year). Title, publisher/location, pp.

[1] Authors1 (year). Title, publisher/location, pp.

.. Lit Review

- Data Corpus
  - Dataset is extracted from Kaggle[3], The dataset contains age, sex, location, region, Expenses etc. details regarding healthcare finance
  - The data corpus consists of the data and details about the insurer, their dependents, and their medical costs over a year. The prediction is based on the dataset's other parameters such as sex, BMI, smoker or not, age, region, and location

- References

  - [1] https://www.kaggle.com/code/mariapushkareva/medical-insurance-cost-with-linear-regression
  - [2] https://cs229.stanford.edu/proj2012/Lui-EmployerHealthInsurancePremiumPrediction.pdf
  - [3] https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

Case Study

[-4: ONE cite/ref? -8: Findings? -3: Why Relevant?]

- Insurance Premium prediction
  - Venkat Murali [4] presented a model on insurance premium prediction [4] using linear regression.
  - Here the author used basic linear regression to predict the insurance premium. The author used predictors such as age and smoking to predict the premium. The author of this article used a small dataset from the entire dataset and then created a linear regression training model and fed 30% of the data and then plotted a graph to see the actual versus predicted expenses and r squared value of 0.80.
  - From the above paper, I've considered the entire dataset (1338 records) after analyzing the data and I've found that we need to establish the relationship between different features. From the dataset, I first plotted each variable's significance before creating a model

[4] https://www.kaggle.com/code/venkatmurali/insurance-premium-prediction-linear-regression

Deliverables

[-15: Solution+Code? -5: Ans to Qs? -10: Assessment/Threshold met?]

- Solution
  - The solution will be one python program that predicts the premium with an average relative error of at most 20%.
  - By using the Random forest method, we got an r-squared value of 0.800, which is accurate.

Random Forest Classifier

```
In [56]: from sklearn.ensemble import RandomForestClassifier

In [57]: model_2=RandomForestClassifier(n_jobs=-1, random_state=42)

In [58]: model_2.fit(train_inputs, train_targets)
Out[58]: RandomForestClassifier(n_jobs=-1, random_state=42)

In [59]: %%time
         model_2.score(train_inputs, train_targets)
         CPU times: user 11 s, sys: 41.8 ms, total: 11 s
         Wall time: 2.85 s
Out[59]: 0.9998622440445148

In [60]: from sklearn.metrics import confusion_matrix
         def predict_and_plot_2(inputs, targets,name=''):
             preds=model_2.predict(inputs)
             accuracy = accuracy_score(targets, preds)
             print("Accuracy: {:.2f}%".format(accuracy * 100))
             cf = confusion_matrix(targets, preds, normalize='true')
             plt.figure()
             sns.heatmap(cf, annot=True)
             plt.xlabel('Prediction')
             plt.ylabel('Target')
             plt.title('{} Confusion Matrix'.format(name));
             return preds

In [61]: %%time
         val_preds_2 = predict_and_plot_2(val_inputs, val_targets, 'Validation')

         #
         from sklearn.metrics import r2_score
         score = r2_score(y_test, y_pred)

         score
]: 0.8000184017333828
```

```
lm(formula = Premium)

Residuals:
    Min      1Q  Median      3Q     Max
-37.789  -9.249   0.216   9.551  29.985

Coefficients:
                  Estimate   Std. Error t value
(Intercept)      35.08357164   1.82643394  19.209
sexmale          -0.38484017   0.66040217  -0.583
bmi              -0.13500021   0.05954447  -2.267
children         -0.12910353   0.27455494  -0.470
smokeryes       -24.64966176   1.37781641 -17.890
regionnorthwest   0.24308160   0.94493702   0.257
regionsoutheast  -0.03574700   0.95123560  -0.038
regionsouthwest   0.76280363   0.94924599   0.804
expenses          0.00101069   0.00004682  21.586
                        Pr(>|t|)
(Intercept)     <0.0000000000000002 ***
sexmale                    0.5602
bmi                        0.0235 *
children                   0.6383
smokeryes       <0.0000000000000002 ***
regionnorthwest            0.7970
regionsoutheast            0.9700
regionsouthwest            0.4218
expenses        <0.0000000000000002 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.03 on 1329 degrees of freedom
Multiple R-squared:  0.2718,    Adjusted R-squared:  0.2675
F-statistic: 62.02 on 8 and 1329 DF,  p-value: < 0.00000000000000022
```


[1] Authors3 (year). Title, publisher/location, pp.

[2] Authors2 (year. Title, publisher/location, pp.

.. Deliverables: Running Code

- Ans to Qs?
- Expenses are dependent variables, and it is solved by using a single regression model, based on a predictor expense which is dependent on other variables. The use of a simple regression analysis example will enable you to find out if there exists a relationship between variables.
- The most important dependent variable for the project is Expenses. Which is used in my project. Which is the dependent variable which is dependent on other variables age, sex, and BMI values.
- The Two most important predictors from my data are age, and BMI when considered with other variables these are very important.
- Assessment
- MR model value is high when compared to LR value when independent variables are changed.
- Random forest model has good accuracy results when compared to Logistic regression & Linear regression methods

[1] Authors3 (year). Title, publisher/location, pp.

[2] Authors2 (year. Title, publisher/location, pp.