Analysis on the Health Data for Medical Insurance Cost

By Team Dot Product

Jin Choi (jkc9890) Shih-Lun Huang (sh7008) Wei-Han Hsu (wh2405)

1. Introduction

In the domain of health insurance, a variety of factors influence a customer's premium price (cost of insurance), and the process of setting premium prices is typically carried out by the pricing and underwriting department [1]. As data is more accessible and easily stored, the healthcare industry generates a large amount of data related to patients, diseases, and diagnoses, but this data has not been properly analyzed, leading to a lack of understanding of its significance and its potential impact on patient healthcare costs. Since there are several factors that can affect the cost of healthcare or insurance, it is important for a variety of stakeholders and health departments to accurately predict individual healthcare expenses using prediction models. The goal of this project is to properly analyze the patient's data to design models to accurately predict the premium price and also observe significant characteristics of the data.

1.1 Data Preparation

The data was retrieved from Kaggle [2], and has a total dimension of 1000 rows and 11 columns. Each row represents a customer of the insurance, and each column represents various features that describe a customer. The features include: age, presence of diabetes, blood pressure issues, transplant history, chronic disease history, height, weight, known allergies, cancer in the family history, number of major surgeries, and the premium price.

Fortunately, the data did not contain any missing values or null values, hence there was no need for handling any missing values. Upon using a box plot to observe any noticeable outliers, we could not detect any presence of outliers in the data. We normalized the non-categorical features for faster computation as well as easier comparison and interpretability. To eliminate any unnecessary features, we conducted LASSO regression, and it turned out that none of the coefficients turned out to be zero, thus we kept all features.

2. Inference

2.1 Question to ask

The motivating question in the hypothesis testing was to identify any relationship between the premium price and age, surgery history, existence of diabetes, and chronic disease history.

2.2 Method of approach

We parsed the data to create two distinct distributions per each of the four features based on the premium price. Then, we conducted the power analysis by setting the desired power to be 0.8, significance level to be .05, finding the effect size of the two distributions, and determining whether we have adequate sample sizes. Then, computed hypothesis testing using the Welch's t-test since the variance between the two groups are different and that we compare scores between two different groups.

2.3 Computation and Analysis

We suspected that age, surgery history, diabetes history, and chronic disease history are all influential in determining the premium price. As for the first step, we separated the data into two age groups, above 50 years old and below 50 years old. The null hypothesis H_o was that there is no difference in premium price among age groups, and the alternate hypothesis H_A was that there is a difference. But before conducting the hypothesis test, we computed the power analysis to ensure that we have enough statistical power to even answer the questions. This process of creating two distributions and conducting power analysis was repeated three more times using other features (surgery history, diabetes history, and chronic disease history). Upon our power analysis, the only distributions that failed to satisfy the power analysis were the distributions with diabetes and without diabetes, thus this analysis was dropped. The required sample size for these distributions were 656, but the two distributions had the sample size of 414 and 572.

Looking at Figure 1, the mean premium price for the two distributions are far apart, without any overlaps of the confidence intervals. After conducting Welch's t-test on distributions of ages above 50 and below 50, the p-value was 5.584×10^{-82}. Since this value is smaller than the alpha value (.05), we reject the null hypothesis and conclude that age group has influence on the premium price. Similarly, the p-values for Welch's t-test on distributions of presence of chronic disease (Figure 2) and surgery history (Figure 3) are 1.73×10^{-13}, 1.6×10^{-11}, respectively. For either case, we reject the null hypothesis since the p-values are far less than the alpha level, thus we conclude that presence of chronic disease and surgery history are influential in the premium price.

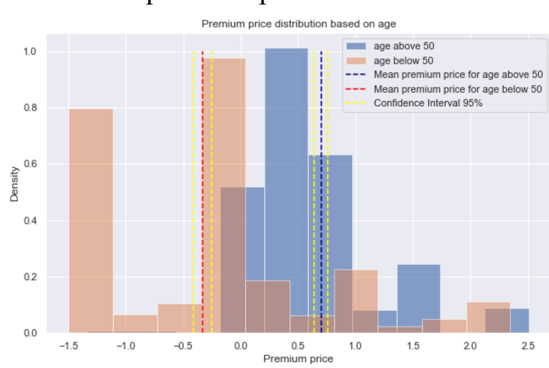

Figure 1

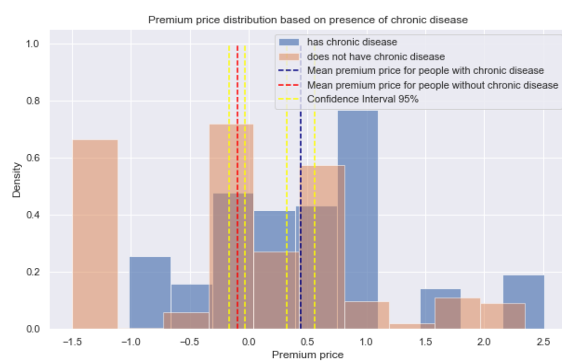

Figure 2

Figure 3

3. Prediction

3.1 Question to ask

For this section, we would like to ask which regularized regression model would perform the best to predict the premium price for the health insurance data.

3.2 Method of approach

We implemented regularized regression models to prevent the "curse of dimensionality" caused by the multivariate yet limited data set. Here, we built the three most popular ones: Lasso Regression, Ridge Regression, and Elastic Net. We then applied GridSearchCV, a type of cross-validation method, to fine-tune each model's parameters. Alpha levels which were between 0.0001 and 20, were tested to find the optimal value.

3.3 Computation and Analysis

Prior to building a model, we first standardized the dataset and initialized all features, excluding "PremiumPrice" as the predictors (986x10) and the "Premium Price" (986x1) as our target. We then fractionated the dataset using an 80/20 split, with the random state as N number:15020304, for generating training and testing data sets. Next, models were further built and optimized by cross-validation. Finally, we computed the coefficient (below table), R^2, and RMSE to evaluate our models as Figure 4.

Model/ Feature	Age	Diabetes	BloodPressure Problems	AnyTransplants	AnyChronic Diseases	Height	Weight	KnownAllergies	HistoryOfCancer InFamily	NumberOfMajor Surgeries
Lasso	0.7335	-0.03391	0.02408	0.29105	0.15041	-0.01952	0.17119	0.02212	0.12559	-0.08325
Ridge	0.73482	-0.04298	0.02086	0.2982	0.15886	-0.01243	0.15299	0.02394	0.11528	-0.07947
Elastic Net	0.6094	0	0	0.23745	0.10957	0	0.09347	0	0.0481	0

In the above chart, the coefficient scores for "Age" indicate a high mathematical relationship with the target, "PremiumPrice"; As for "Height," the scores tell us a low correlation with the target in all three models. The R^2 value tells us that the predictor variables in the models are able to explain about 60% of the premium prices. The RMSE value means the average deviation between the predicted premium price made by the model and the actual price. To improve the model performance, we tried eliminating the variable "Height" since it has the lowest coefficient score. The R^2 and RMSE comparison between the models "with Height" and "without Height" are shown in Figure 4 and Figure 5. The results tell us that there are slight improvements in all three models. Overall, ridge regression has the highest accuracy in predicting Premium Price.

	R^2	RMSE	updated R^2	updated RMSE
Lasso	0.5874	0.6772	0.6021	0.6526
Ridge	0.6018	0.6515	0.6031	0.6519
Elastic Net	0.5953	0.6567	0.6017	0.6515

Figure 4

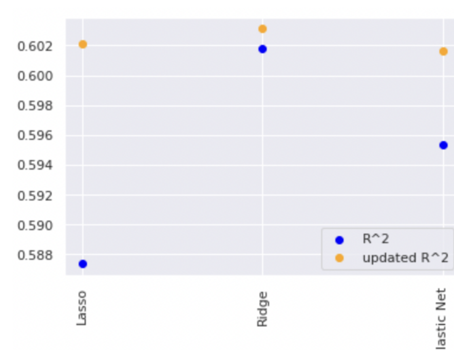

Figure 5

4. Classification

4.1 Question to ask

In the previous parts, we explored the relationship between each feature and how to predict premium prices. In this part, we try to answer a different question. With all the given features in the datasets will we be able to determine a person's diabetes status?

4.2 Method of approach

We start by clustering the dataset with all features excluding 'Diabetes' with the KMeans clustering method to better understand the data. Then to train the classification model, we apply the XGBoost classifier to create the decision tree to determine one's diabetes status.

4.3 Computation and Analysis

In order to better cluster the data, we implement principle component analysis to decrease the dimension of the data. First, we center the data by standardizing features with their means and standard deviations. Next, the Kaiser Criterion approach helps us obtain a systematically optimized number of components by setting a threshold of eigenvalues greater than 1.0 (Figure 6).

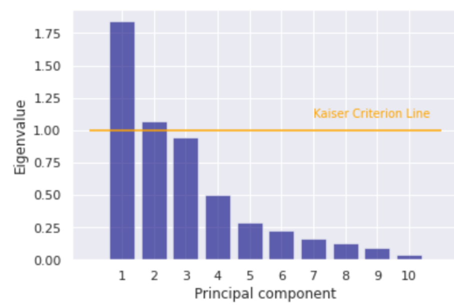

Figure 6

The result shows that the optimal number of principal components is 2, which explains 55.02% of the variance.

After obtaining the PCA-transformed data, Silhouette Score is implemented to compute the optimal number of clusters to apply in the KMeans approach. The result (Figure 7) shows that a clustering model with 2 clusters performs the best with an average silhouette score of 0.405.

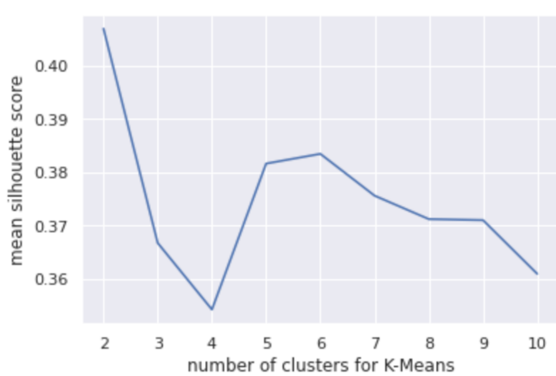

Figure 7

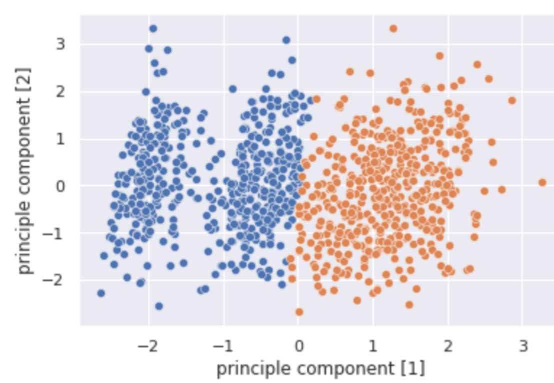

Figure 8

To answer the Diabetes Classification question, we apply the XGBoost method to create a decision tree classifier. XGBoost contains a series of hyperparameters. To reach the optimal model without much knowledge regarding the healthcare and insurance industries, we perform the grid search cross-validation of each hyperparameter according to the recommended intervals. Below we list out the hyperparameters and corresponding values applied to tune the classifier.

```
"max_depth": [3, 4, 5, 7]
"learning_rate": [0.1, 0.01, 0.05]
"gamma": [0, 0.25, 1]
"reg_lambda": [0, 1, 10]
"scale_pos_weight": [1, 3, 5]
"subsample": [0.8]
"colsample_bytree": [0.5]
```

We finalize the model with the following hyperparameters {'colsample_bytree': 0.5, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': 7, 'reg_lambda': 1, 'scale_pos_weight': 1, 'subsample': 0.8} with an area under ROC curve value of 0.6511. The following (Figure 9 and Figure 10) shows the result of the optimized classifier.

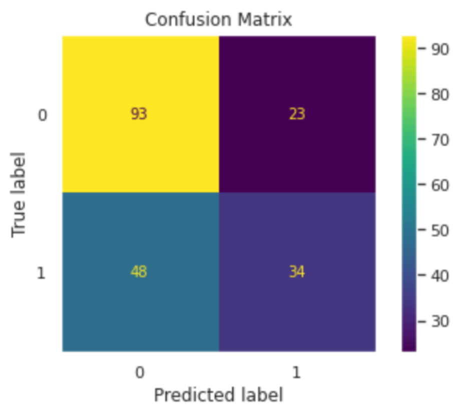

Figure 9

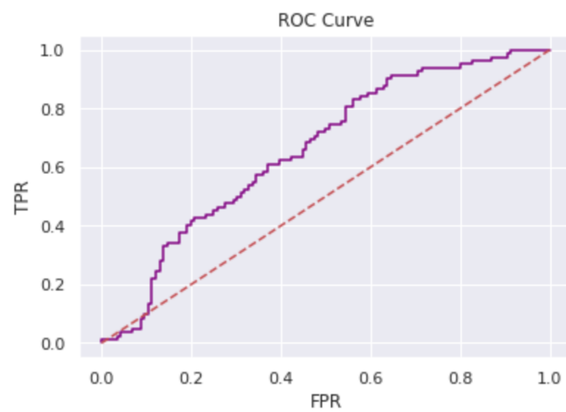

Figure 10

The optimal classifier consists of features with distinct weights shown below.

Feature	Age	Blood Pressure Problems	Any Transplants	Any Chronic Diseases	Hight	Weights	Known Allergies	History of Cancer in Family	Number of Major Surgeries	Premium Price
Weights	0.1230	0.1127	0.0699	0.1198	0.0803	0.0734	0.0790	0.0705	0.1667	0.1047

From the scatter plot obtained by the KMeans clustering, we see that there exist two groups of people determined by all features except Diabetes which may be used to determine whether or not a person has diabetes. The classification model helps put our assumption to test. However, the area under ROC curve value of 0.6511 implies that our model is not robust enough to corroborate our assumption.

5. Conclusion

Upon analyzing the health insurance data, we identified various significant results. First, upon conducting causal inference from hypothesis testing using Welch's t-test, we concluded that age, chronic disease history, and surgery history are factors that affect the insurance premium price. Further, our predictive analysis concluded that out of three possible regularized regression models (Ridge, LASSO, Elastic net), ridge regression resulted in the highest R^2 as well as the lowest RMSE values. In the regression model, the "Age" feature had the highest coefficient value, which meant that age had the closest relationship with the premium price. Lastly, the clustering result shows that there exists two potential groups regarding the status of diabetes, however, the classification model does not perform well on determining diabetes patients.

During the analysis, there were few assumptions that were made which could potentially affect the outcome of our analysis. First, we made the assumption that the relationship between the features and the independent variable is linear, but it is very possible that the relationship is non-linear (i.e. polynomial). Also, we made the assumption that the data retrieved for this project had no biased sampling. Particularly, we assumed that there were no other confounding variables which could have impacted the result. Finally, our assumption that these premium prices are of the same insurance plan draws significant limitations to our prediction as different insurance plans offer various coverages, hence, a distinct underlying pricing model for each plan. To mitigate some of the assumptions, we could gather data using stratified sampling, where we create multiple categories or subgroups in which the confounding variables do not vary much, and survey the patients.

Additional fact about this dataset is that although the "Height" column itself does not present any significant result, linearly combining "Height" and "Weight", thus creating a column "BMI", offers significant results. We created two distributions, heights above mean height and heights below mean height, and analyzed whether the two distributions differ in the mean premium price. The hypothesis test resulted in a p-value higher than the alpha level thus we conclude that height is not an indicator of premium price. However, when we combined "Height" and "Weight" to create a "BMI" feature, the hypothesis test resulted in a p-value less than the alpha level, thus we could conclude that BMI affects the premium price.

Author Contribution:

Jin Choi: Section I, II, V
Wei-Han Hsu: Section I, III, V
Shih-Lun Huang: Section I, IV, V

References:

[1] Insurance premium pricing: https://thismatter.com/money/insurance/rate-making.htm

[2] Kaggle data: https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction