Medical Insurance Cost Premium with Machine Learning

Adam Yang, Guanzhong Wang, Henry Tan, Janice Shen, Jimmy Hong

Introduction

Accurately estimating medical insurance premiums is a critical challenge, as insurers must balance **predictive accuracy** with model **interpretability** to ensure fair and transparent risk assessment. Premium pricing models typically use complex, often nonlinear relationships among demographic, medical, and lifestyle factors. Improving these models is essential because better accuracy helps insurers manage financial risk and allows customers to make more informed decisions about coverage, particularly as healthcare expenditures rise globally.

In this project, we analyze a medical insurance dataset originally released on Kaggle designed as a teaching resource for predictive modeling [1]. The dataset contains 986 observations, each representing a customer who voluntarily provided personal health information. The variables include demographic characteristics such as age, height, and weight, along with a range of binary indicators capturing chronic medical conditions, such as diabetes status, blood pressure problems, transplant history, major surgeries, allergies, and family history of cancer. The goal is to use these predictors to estimate the customer's PremiumPrice, defined as the annual medical insurance premium in Indian Rupees (INR). Although the dataset is synthetic and simplified compared to real actuarial data, it captures a representative set of factors that influence medical risk and therefore insurance costs.

To evaluate and compare different modeling strategies, we employ four widely used supervised learning approaches: Principal Components Regression (PCR), Ridge Regression, Lasso Regression, and Best Subset Selection. Each method offers a distinct way of addressing multicollinearity, high-dimensionality, and variable importance. PCR reduces dimensionality by projecting predictors onto orthogonal components; Ridge regression shrinks coefficients to stabilize estimates in the presence of correlated predictors; Lasso performs variable selection by shrinking some coefficients exactly to zero; and Best Subset Selection searches systematically for the most predictive combination of variables. For fairness and consistency, all models are evaluated using 10-fold cross-validation, which provides a less biased, and more reliable estimate of the model's out-of-sample prediction error.

By comparing the predictive accuracy and selected variables across these methods, our project aims to identify models that **not only perform well** but also **offer interpretable insights** into the key health factors driving insurance premiums. Ultimately, this exercise highlights both the **strengths and limitations** of statistical learning approaches in the context of health-related financial prediction.

Methods

Data Source and Initial Cleaning

We analyzed the "Medicalpremium.csv" dataset from Kaggle, which contains 986 observations of customers who voluntarily provided demographic and health information [1]. The outcome variable, *PremiumPrice,* represents the yearly medical insurance premium in Indian Rupees (INR). The dataset contains ten predictors: age, height, weight, diabetes status, blood pressure problems, transplant history, chronic diseases, known allergies, family history of cancer, and number of major surgeries.

Before building models, we performed a series of preprocessing steps to ensure data quality and consistency. We tried to remove rows with missing values with na.omit(), but found no NA values. The six binary medical history variables (*Diabetes, BloodPressureProblems, AnyTransplants, AnyChronicDiseases, KnownAllergies, HistoryOfCancerInFamily*) were validated to ensure they only contained values 0 or 1; any observations outside this range were discarded. The variable *NumberOfMajorSurgeries* originally included levels 0, 1, 2, and 3. Due to sparsity in level 3, values 2 and 3 were collapsed into a single category ("2"). Hence, there are three different levels in *NumberOfMajorSurgeries*. With 0 as the reference level, *NumberOfMajorSurgeries* is represented with 2 variables, namely *NumberOfMajorSurgeries1* (Number of surgeries = 1) and *NumberOfMajorSurgeries2* (Number of surgeries = 2).

Train/Test Structure and Cross-Validation

To evaluate model performance consistently across all methods, we used 10-fold cross-validation throughout the analysis. This involved splitting the data into ten folds, repeatedly training models on nine folds, and evaluating the model performance with the held-out fold. Performance was assessed using the mean squared error (MSE) averaged across folds, as an estimate of the out-of-sample test error of the models.

Overall Modeling Framework

The goal of this project is to use demographic and health-related variables to predict a customer's annual medical insurance premium. Across all methods, we model the premium as a function of the same set of 11 predictors:

$$\textbf{\textit{PremiumPrice}} \sim \textbf{\textit{Age}} + \textbf{\textit{Height}} + \textbf{\textit{Weight}} + \textbf{\textit{Diabetes}} + \textbf{\textit{BloodPressureProblems}} + \textbf{\textit{AnyTransplants}} + \textbf{\textit{AnyChronicDiseases}} + \textbf{\textit{KnownAllergies}} + \textbf{\textit{HistoryOfCancerInFamily}} + \textbf{\textit{NumberOfMajorSurgeries1}} + \textbf{\textit{NumberOfMajorSurgeries2}}$$

Here, the first three predictors are continuous variables, the next eight are binary indicators, and *NumberOfMajorSurgeries1* and *NumberOfMajorSurgeries2* are dummies representing 1 and 2 surgeries, with 0 as reference levels. These 11 variables collectively

describe a customer's basic health status and medical risk profile and are used consistently across all modeling procedures as the null model.

Principal Components Regression (PCR)

PCR addresses multicollinearity by transforming the original predictors into orthogonal principal components before performing regression. We implemented PCR using the pls package:

pcr(PremiumPrice ~ ., data = data, scale = TRUE, validation = "CV")

Setting validation = "CV" invokes 10-fold cross-validation by default, allowing us to estimate the predictive MSE for models with different numbers of components. The number of components was selected by identifying the one that minimized cross-validated MSE, and a final PCR model was then refitted using that optimal number of components. Because PCR regresses on principal components rather than the raw predictors, interpretability is reduced, but PCR can improve stability when predictors are correlated.

Ridge Regression

Ridge regression imposes an L2 penalty on regression coefficients, shrinking them toward zero without eliminating predictors. This can substantially improve prediction in the presence of multicollinearity. We used the glmnet package with:

cv.glmnet(x, y, alpha = 0)

where alpha = 0 specifies Ridge. The function performed 10-fold cross-validation to select the optimal regularization parameter λ. Continuous predictors (Age, Height, Weight) were standardized to ensure that penalty is applied fairly across predictors. The final Ridge model was refitted using the λ that minimized cross-validated MSE.

Lasso Regression

Lasso regression applies an L1 penalty, shrinking some coefficients exactly to zero and thereby performing variable selection. We constructed a design matrix using model.matrix() to correctly encode categorical variables, and then fit the model using:

cv.glmnet(x, y, alpha = 1)

where alpha = 1 specifies Lasso. The use of cv.glmnet again ensures 10-fold cross-validation for λ selection. After identifying the optimal λ, we extracted the nonzero coefficients to determine which predictors were selected by Lasso. This method provides both prediction accuracy and interpretability by identifying the most important health factors to predict insurance premiums.

Best Subset Selection

Best subset selection systematically searches for all possible subsets of predictors to identify combinations that best explain the response. Because regsubsets() cannot directly handle multi-level factors, we first generated a dummy-encoded matrix:

X <- model.matrix(PremiumPrice ~ ., data = data)

This expanded *NumberOfMajorSurgeries* into two indicator variables (representing levels 1 and 2). We fit models of size 1 through 11 using:

regsubsets(PremiumPrice ~ ., data, nvmax = 11)

Candidate models were chosen using 4 information criterions, namely Bayesian Information Criterion (BIC), Mallows' Cp, Adjusted R^2, and Akaike Information Criterion (AIC). To compare them fairly, each candidate model was refitted using standard linear regression and evaluated with 10-fold cross-validation via cv.glm() from the boot package. The final model chosen was the one with the lowest cross-validated MSE.

Result

Principal Component Regression

PCR was applied to the cleaned MedicalPremium dataset using standardized predictors and 10-fold cross-validation. Models with 1 through 11 components were evaluated. The cross-validated MSE consistently decreased as more components were included, indicating that predictive information is spread across many dimensions rather than captured in only the first few principal components. Because the lowest CV MSE occurred at 11 components, PCR provided no effective dimension reduction for this dataset (Figure 1). The final selected PCR model therefore uses M = 11 components, corresponding to retaining all predictor information, and achieved a 10-fold CV MSE of approximately **14,258,567.**

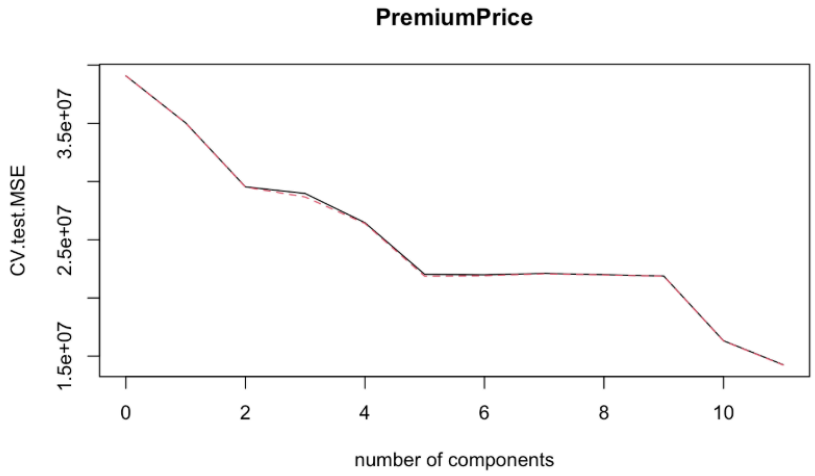

Figure 1: 10-Fold Cross Validation Test MSE against Number of Components for PCR

Because several predictors were stored as factors, R automatically dummy-encoded them prior to component extraction. Binary variables (e.g., *Diabetes*) appear as single dummy variables (e.g., *Diabetes1*), while the three-level *NumberOfMajorSurgeries* factor produces two dummy variables (*NumberOfMajorSurgeries1, NumberOfMajorSurgeries2*). These encoded variables appear in the PCR coefficient output because the regression step is performed on the dummy-expanded predictor matrix. This means that the final coefficients in the output correspond to these individual dummy variables, not the original factors, hence cannot be interpreted directly.

Ridge Regression

Ridge regression was conducted using the glmnet package with an L2 penalty and 10-fold cross-validation to select the tuning parameter λ with minimum CV MSE, which is 435.614763. Using the optimal λ and reconducting ridge regression, we yield a 10-fold CV MSE of approximately **14,367,833** (Figure 2). As the penalty λ increases further (moving right), the error steeply rises, indicating that excessive shrinkage leads to high bias and underfitting.

For this specific problem, Ridge Regression yielded a higher CV MSE compared to PCR. Ridge, by design, performs only variable shrinkage, reducing the magnitude of all coefficients toward zero for stabilization and reducing the variance of the estimates, but never setting them exactly to zero. Because the model retains the coefficients of the original features, it is generally more interpretable than PCR, which relies on latent, less understandable components. However, this stabilization mechanism, which keeps all variables in the model, means Ridge cannot isolate the truly necessary predictors (Ridge does not perform feature selection).

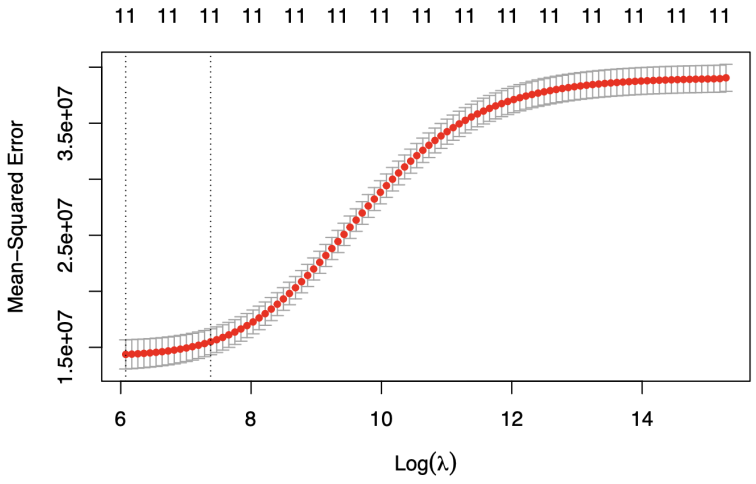

Figure 2: 10-Fold Cross Validation Test MSE against Log(λ) for Ridge Regression

Lasso Regression

Similar to Ridge regression, Lasso regression was conducted using the glmnet package with an L1 penalty and 10-fold cross-validation to select the tuning parameter λ. Cross-validation identified an optimal value of λ≈28.66, yielding a 10-fold CV MSE of approximately **14,221,806** (Figure 3).

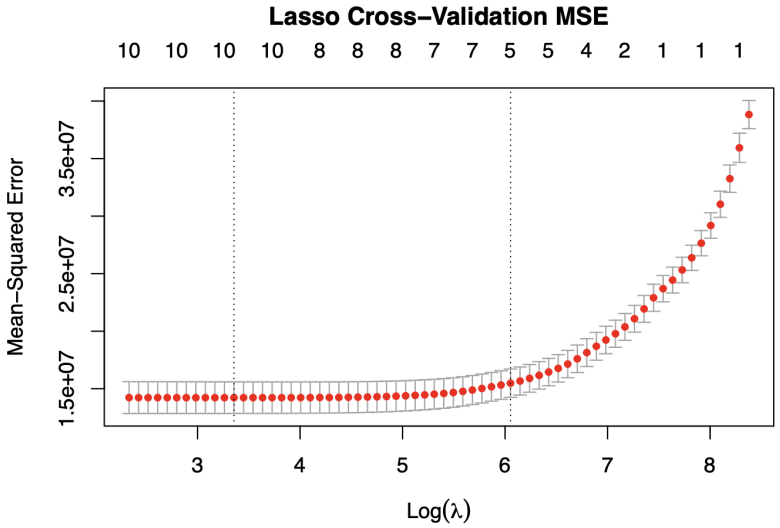

Figure 3: 10-Fold Cross Validation Test MSE against -Log(λ) for Lasso Regression

Unlike Ridge, Lasso performs variable selection by shrinking some coefficients exactly to zero. In this dataset, the Lasso model retained nearly all predictors, including *Age, Height, Weight, Diabetes, Blood Pressure Problems, AnyTransplants, AnyChronicDiseases, HistoryOfCancerInFamily,* and the surgery indicators, indicating their collective importance for predicting premium prices. Only *KnownAllergies* were eliminated, suggesting that allergies might not provide additional predictive value after accounting for the other health factors. The cross-validation curve displayed the typical Lasso pattern: MSE decreased as λ was relaxed, reached a minimum at the selected λ, and increased again when λ became too large due to excessive penalization and underfitting. Overall, Lasso offered strong predictive performance while simplifying the model by removing one uninformative variable.

Best Subset Selection

Best Subset selection was applied to the cleaned MedicalPremium dataset after removing rows with missing or invalid categorical values and converting all binary predictors to factors. The model was fitted using the leaps package with 10-fold cross-validation using the boot package, and all predictors were standardized prior to component extraction. The regsubsets() function (part of the leaps library) was used to perform best subset selection by identifying the best model that contains a given number of predictors with lowest RSS. Eleven models were generated corresponding to 1 to 11 variables. Four candidate models were selected using BIC,

Mallows's Cp, adjusted R squared, and AIC of each model, respectively. Ten-fold cross-validation was applied to each of the candidate models to estimate the mean square error (MSE).

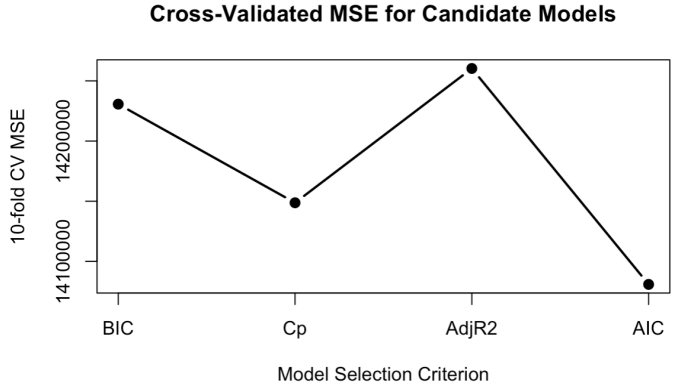

Figure 4. Line graph of 10-fold CV MSE for the four candidate models.

criterion <chr>	size <int>	cv_mse <dbl>
BIC	6	14230680
Cp	6	14148702
AdjR2	7	14260337
AIC	6	14080869

Table 1. Ten-fold CV MSE for the four candidate models and their respective size.

According to Figure 4, the model with six variables and lowest AIC produced the lowest 10-fold CV test MSE. While the CV MSE for the four models were very close to each other in Table 1, three of the four suggested the 6-variable model from the best subset selection. Hence we are more confident to say that the final model is the 6-variable model:

$$PremiumPrice = \beta_0 + \beta_{Age} * Age + \beta_{Tran} I_{Tran} + \beta_{Chr} I_{Chr} + \beta_{Weight} * Weight + \beta_{Cancer} I_{Cancer} + \beta_{Sur2} I_{Sur2}$$

where *PremiumPrice* is yearly medical insurance premium price in Indian Rupees (INR).

- *Age* = Age of the customer.
- *Tran* = 1 if having major organ transplants, 0 otherwise
- *Chr* = 1 if the customer suffers from Chronic ailments like asthma, etc, 0 otherwise.
- *Weight* = Weight of the customer.
- *Cancer* = 1 if any blood relative of the customer has had any form of cancer, 0 otherwise.
- *Sur2* = 1 if the customer had 2 or 3 major surgeries, 0 if the customer had either 0 or 1.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                23439.3      151.1 155.100  < 2e-16 ***
Age                         4673.4      134.6  34.729  < 2e-16 ***
AnyTransplants1             7868.4      518.0  15.188  < 2e-16 ***
AnyChronicDiseases1         2723.0      309.6   8.796  < 2e-16 ***
Weight                      1009.0      118.9   8.483  < 2e-16 ***
HistoryOfCancerInFamily1    2011.2      369.0   5.451 6.35e-08 ***
NumberOfMajorSurgeries2    -1969.8      391.0  -5.037 5.62e-07 ***
```

Table 2. Linear Regression results of the 6-variable model, showing the coefficient estimates of variables

Summary Results

Full Model (Before Shrinkage and Feature Selection = 11 predictors)

PremiumPrice ~ Age + Height + Weight + Diabetes + BloodPressureProblems + AnyTransplants + AnyChronicDiseases + KnownAllergies + HistoryOfCancerInFamily + NumberOfMajorSurgeries1 + NumberOfMajorSurgeries2.

Methods <chr>	Best_Model <chr>	CV_MSE <dbl>	Num_Predictors <chr>
PCR	11 components	14255856	NA
Ridge	lambda = 435.6148	14367833	11
Subset (AIC)	6 predictors	14080869	6
LASSO	lambda = 28.6605	14221806	10

Predictive Model	Predictors Description
PCR	No predictors selected, but form 11 components
Ridge	No predictors are eliminated, include full models
Lasso	Variables eliminated: *KnownAllergies* (total of 10 predictors included)
Best Subset Selection	Variables eliminated: *Height, Diabetes, BloodPressureProblems, KnownAllergies, NumberOfMajorSurgeries1* (total of 6 predictors included)

Conclusion

In this project, we evaluated four predictive modeling approaches, namely PCR, Ridge Regression, Lasso Regression, and Best Subset Selection, to estimate medical insurance premiums from demographic and health characteristics. A major distinction among these models lies in interpretability. PCR has the lowest interpretability, as its principal components cannot be mapped directly to specific health factors, limiting its usefulness for policy or medical decision-making. Ridge stabilizes coefficient estimates and preserves the original variables,

ensuring interpretability, but since it does not perform feature selection, we cannot derive useful information about which predictors are really important and predictive to the insurance premium cost. Lasso provides some interpretability through shrinkage and potential variable elimination, but in our dataset it removed only one predictor, leaving a nearly full model. In contrast, Best Subset Selection offers the highest interpretability and explicitly identifies the smallest group of meaningful predictors. This sparsity makes **Best Subset** particularly valuable when the goal is to understand which health conditions **most strongly** contribute to premium pricing.

In terms of predictive accuracy, Best Subset Selection produced the lowest 10-fold cross-validated MSE, outperforming PCR, Ridge, and Lasso. This outcome agrees with classroom intuition that Best Subset often excels when the true underlying relationship is relatively sparse, when only a subset of predictors truly matters. While Lasso also performs variable selection, its shrinkage penalty can introduce bias when strong predictors are heavily penalized. The final Best Subset model identifies six key predictors: *Age, Weight, Chronic Diseases, Organ Transplants, Family History of Cancer, and having 2 or more major surgeries.* These variables align well with medical understanding of long-term healthcare cost drivers. For example, in Table 2, when holding other factors constant, each additional year of age increases expected yearly premium by approximately 4763 INR, and separately, having undergone a major organ transplant increases premiums by roughly 7868 INR. Such interpretable coefficients offer clear, actionable insight into how individual health characteristics translate into financial risk.

These findings carry important real-world implications. A model like Best Subset Selection, which is both accurate and interpretable, can help insurers design transparent and justifiable pricing strategies, improving consumer trust and regulatory compliance. It can also assist policymakers and healthcare providers in identifying high-risk groups who may benefit from targeted preventive care, subsidy programs, or early intervention to reduce long-term costs. Nonetheless, several limitations must be acknowledged. Cross-validated MSE is sensitive to random fold assignments, meaning different seeds may produce slightly different optimal models. In addition, the dataset is synthetic and simplified, limiting generalizability to real insurance markets. Finally, each modeling approach has intrinsic constraints: PCR sacrifices interpretability, Lasso and Ridge differ in tuning behavior due to their penalty structures, and Best Subset, although effective, can be computationally expensive. To sum up, our interpretations from Best Subset Selection rely on the assumption that the true model is sparse and that this method performs well for our dataset; these conclusions may not generalize to other settings. When sparsity holds, Best Subset can balance predictive accuracy and interpretability, making it a valuable tool for health policy and insurance analytics.

Reference

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