MA-691 : Project
Insurance Premium Prediction using Combined Regression Strategy (COBRA)

1. Mohammad Humam Khan - 180123057
2. Kartikeya Singh - 180123021
3. Sourav Goel - 180123058
4. Pragati Ramesh Mahamune - 180123032
5. Milind B Prabhu - 180101091
Instructor : Dr. Arabin Kumar Dey

Submission Date : 19-11-2021

Disclaimer:
This work is for learning purpose only. The work can not be used for publication or as commercial products, etc without instructor's consent.

Website: Our work is deployed on Heroku and is live. Click here to go to Insurance Premium Calculator.

1 Introduction

Insurance Premium Prediction problem deals with the question ”Given a set of characteristics, would we be able to predict an individual's insurance premium?”. This question is very important for health insurance companies because they make profit by collecting higher premiums than the total amount paid to the insured person. The main focus of study is to predict the insurance premium using various factors of an individual such as age, sex, physical/family condition and location. Two very important factors w.r.t. which we analyze the impact on insurance premium are BMI (Body-Mass-Index) and whether the person is a smoker or not? With the help of this study, we signify the importance of healthy lifestyle and how radically it can reduce the insurance premium of the individual. The application that is developed leverages the power of machine learning to tip off people how smoking and maintaining a normal BMI can affect their insurance charges. At the backend, the application uses Combined Regression Strategy (COBRA) [1] algorithm trained on Insurance Premium Prediction

Dataset from Kaggle [dataset] to predict insurance premium of a person given the input parameters. The remaining part of report is organized as follows: In the next section we provide a brief outline of working of COBRA algorithm. Then in section 3, we explain the datset that has been used and in section 4 we discuss the experimental results on the training and test data.

2 Combined Regression Strategy (COBRA)

Suppose we have a training dataset $D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$. D_n is composed of i.i.d. random variables taking their values in $\mathbb{R}^d \times \mathbb{R}$, and distributed as an independent prototype pair $(\mathbf{X}, Y)$ satisfying $\mathbb{E}Y^2 < \infty$ (with the notation $\mathbf{X} = (X_1, \ldots, X_d)$). The space $\mathbb{R}^d$ is equipped with the standard Euclidean metric. The goal is to estimate the regression function $r^\star(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}], \mathbf{x} \in \mathbb{R}^d$, using the data D_n.

The original data set D_n is split into two data sequences $D_k = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_k, Y_k)\}$ and $D_\ell = \{(\mathbf{X}_{k+1}, Y_{k+1}), \ldots, (\mathbf{X}_n, Y_n)\}$, with $\ell = n - k \geq 1$. For convenience, the elements of R_ℓ are renamed $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_\ell, Y_\ell)\}$.

Now, suppose that we are given a collection of $M \geq 1$ competing candidates $r_{k,1}, \ldots, r_{k,M}$ to estimate $r^\star$. These basic estimators called basic machines are assumed to be generated using only the first subsample D_k. These machines can range from linear regression, kernel smoother, SVM, Lasso, naive Bayes, or random forests to neural networks. The essential idea is that these basic machines can be parametric, non-parametric, or semi-parametric, the only constraint being that each of the $r_{k,m}(\mathbf{x}), m = 1, \ldots, M$, is able to provide an estimation of $r^\star(\mathbf{x})$ on the basis of D_k alone. Thus, any collection of model-based or model-free machines are allowed, and the way of combining such a collection is here called the regression collective. The number of basic machines M is fixed.

Given the collection of basic machines $\mathbf{r}_k = (r_{k,1}, \ldots, r_{k,M})$, we define the collective estimator T_n to be

$$T_n(\mathbf{r}_k(\mathbf{x})) = \sum_{i=1}^{\ell} W_{n,i}(\mathbf{x}) Y_i, \quad \mathbf{x} \in \mathbb{R}^d$$

where the random weights $W_{n,i}(\mathbf{x})$ take the form

$$W_{n,i}(\mathbf{x}) = \frac{\mathbf{1}_{\bigcap_{m=1}^{M}} \{| \, r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{x}_i) \, | \leq \varepsilon_\ell\}}{\sum_{j=1}^{\ell} \mathbf{1}_{\bigcap_{m=1}^{M} \left\{ \left| r_{k,m}(\mathbf{x}) - r_{k,m}(\mathbf{x}_j) \right| \leq \varepsilon_\ell \right\}}}$$

Here, ε_ℓ is some positive parameter and, by convention, $0/0 = 0$.

The collective estimator T_n is a local averaging estimator because the predicted value for $r^\star(\mathbf{x})$, that is, the estimated outcome at the query point $\mathbf{x}$, is the unweighted average over those Y_i 's such that $\mathbf{X}_i$ is "close" to the query point. More precisely, for each $\mathbf{X}_i$ in the sample D_ℓ, "close" means that the output at the query point, generated from each basic machine, is within an ε_ℓ - distance of the output generated by the same basic machine at $\mathbf{X}_i$. If a basic machine evaluated at $\mathbf{X}_i$ is close to the basic machine evaluated at the query point $\mathbf{x}$, then the corresponding outcome Y_i is included in the average, and not otherwise.

In this context, ε_ℓ plays the role of a smoothing parameter i.e. in order to retain Y_i, all basic estimators $r_{k,1}, \ldots, r_{k,M}$ have to deliver predictions for the query point $\mathbf{x}$ which are in a ε_ℓ-neighborhood of the predictions $r_{k,1}\left(\mathbf{X}_i\right), \ldots, r_{k,M}\left(\mathbf{X}_i\right)$. The greater the ε_ℓ, the more tolerant the process. Also, the practical performance of T_n strongly relies on an appropriate choice of ε_ℓ. ε_ℓ can be fixed prior to running the algorithm or we can use a data-driven approach to get optimal value of ε_ℓ while running the algorithm itself.

Although T_n is a weighted average of the Y_j 's in D_ℓ only, however, T_n depends on the entire data set D_n, as the rest of the data is used to set up the original machines $r_{k,1}, \ldots, r_{k,M}$. Also, the combined estimator T_n is nonlinear with respect to the basic estimators $r_{k,m}$.

In the definition of the weights as described above, all original estimators are invited to have the same, equally valued opinion on the importance of the observation $\mathbf{X}_i$ (within the range of ε_ℓ) for the corresponding Y_j to be integrated in the combination T_n. However, this unanimity constraint may be relaxed by imposing, for example, that a fixed fraction $\alpha \in \{1/M, 2/M, \ldots, 1\}$ of the machines agrees on the importance of X_i. In that case, the weights take the more sophisticated form:

$$W_{n,i}(\mathbf{x}) = \frac{\mathbf{1}_{\left\{\sum_{m=1}^{M} \mathbf{1}_{\left[\left|r_{k,m}(\mathbf{x})-r_{k,m}(\mathbf{x}_i)\right| \leq \varepsilon_\ell\right\}} \geq M\alpha\right\}}}{\sum_{j=1}^{\ell} \mathbf{1}_{\left\{\sum_{m=1}^{M} \mathbf{1}_{\left|r_{k,m}(\mathbf{x})-r_{k,m}\left(\mathbf{x}_j\right)\right| \leq \varepsilon_\ell \ell} \geq M\alpha\right\}}}$$

3 Dataset

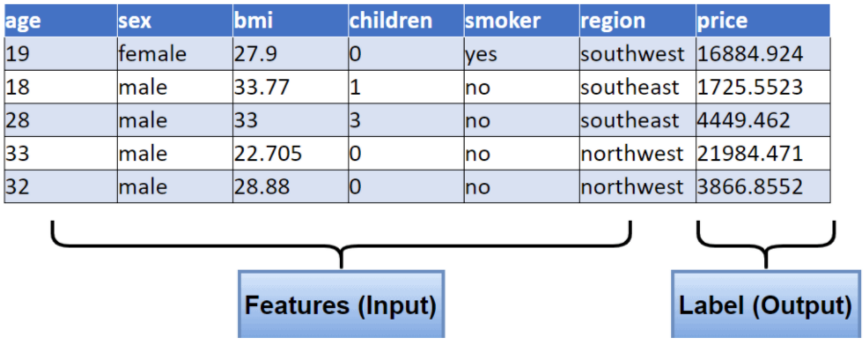
We are using Insurance Premium Prediction Dataset from Kaggle [dataset] to train our model. The dataset contains 1338 observations (rows) and 7 features (columns) per row. There are 4 numerical features (Age, BMI, Number of Children and Expenses) and 3 nominal features (Sex, Smoker, and Region) that were converted into factors with numerical value designated for each level. The data snippet in Figure 1 shows us how the insurance premium price depends and varies with the above mentioned parameters.

Using Exploratory Data Analysis, we can draw following conclusions about the data:-

1. The age of the users availing the insurance ranges from 18 to 64. Therefore the predictions would be more accurate for users aged between these figures.

2. The percentage of data comprising of men is 51% and that of women is 49%. This implies that the dataset used to train is not biased towards any specific gender.

3. The BMI has a bell curve distribution with spike at 27. We can say most of the people are obese according to the dataset and therefore the insurance premium would be a bit higher in general.



age	sex	bmi	children	smoker	region	price
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.88	0	no	northwest	3866.8552

Features (Input)
Label (Output)

Figure 1: Table showing a snippet of dataset

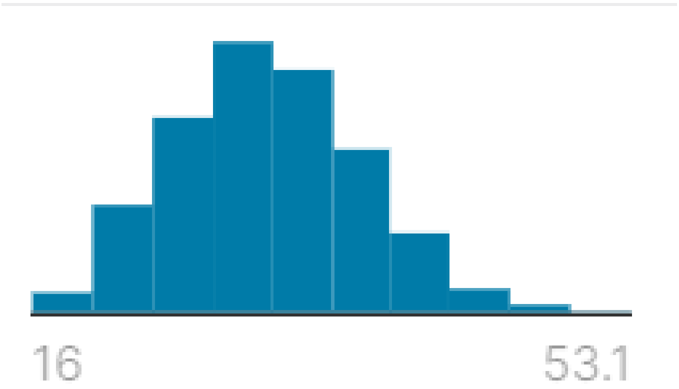


Figure 2: BMI Distribution of dataset

4. It is clearly visible from Exploratory Data Analysis that there is a clear correlation between the insurance premium and whether or not a person is a smoker. Figure 3 below, show the dependence of the insurance premium on the smoking habit of a person. Clearly, a person who smokes is more likely to suffer from serious chronic diseases and therefore should be liable to pay a greater premium.

5. In terms of geography, there is almost equal distribution of people from each of the four regions.

4 Training and Testing the Model

We trained COBRA on the above introduced dataset and then explicitly tested it on a test set to evaluate it's performance. Click here to go to our Github repository. We have used Root-Mean-Squared-Error (RMSE) and Coefficient of Determination (R^2) on test set as metrics to evluate the efficiency of our model. Root Mean Square Error (RMSE) is the standard deviation of the residuals i.e. prediction errors that are a measure of how far from the regression line data points are. R^2 gives

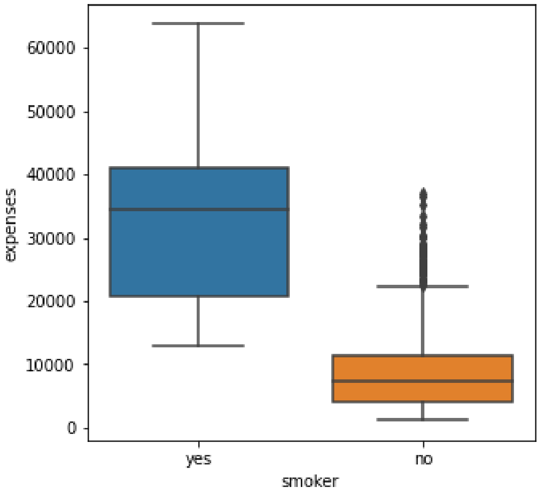


Figure 3: Relationship between insurance premium and whether or not a person is a smoker

an indication of how good a model fits a given dataset. It indicates how close the regression line (i.e the predicted values) is to the actual data values. The R^2 value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

The RMSE is calculated using the following formula -

$$\text{RMSE} = \sqrt{\sum_{k=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}}$$

The R^2 error is calculated using the following formula -

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_{k=1}^{n}(\hat{y_i} - y_i)^2}{\sum_{k=1}^{n}(\bar{y} - y_i)^2}$$

where,
y_i = actual value
$\bar{y}$ = mean of actual values
$\hat{y_i}$ = predicted value

The optimal value of parameter ε_ℓ calculated using grid-search optimization approach is 5122 which is then fixed in the final deployed model for fast prediction of insurance premium. Finally, the RMSE and R^2 values on test data obtained for final deployed model are as follows:

- RMSE = 4840.039
- R^2 = 0.801

5 Deployment

The website is built using a Django framework. We chose this framework because our calculations requires implementation of machine learning algorithms which are efficiently done using Python and since Django is a Python-based back-end framework and also a secure and efficient one, we chose this as our framework. The website is deployed on Heroku. Click here to go to Insurance Premium Calculator.
From the users, we are collecting the following information fields:-

1. Age
2. Sex
3. BMI
4. Children
5. Region
6. Smoker

In the region field we have provided four choices broadly categorizing the geography - South-East, South-West, North-East and North-West. The collected data is then send to our back-end server which calculates the insurance premium or more accurately predicts the insurance premium of the person based on the data they entered using our trained machine learning model. The predicted price is then displayed to the end-user.

References

[1] Benjamin Guedjd James D. Malley Gérard Biau, Aurélie Fischer. Cobra: A combined regression strategy. *Journal of Multivariate Analysis*, 2015.